Comparative genomics and population genomics of the wild relatives of maize

Inaugural dissertation

for the award of the doctorate
from the Faculty of Mathematics and Natural Sciences
of the Heinrich Heine University Düsseldorf

submitted by
Joseph Atemia
from Nakuru, Kenya

Düsseldorf, January 2025

From the Department/Institute of Biological Data Science
of Heinrich Heine University Düsseldorf

Printed by permission of the Faculty of Mathematics and Natural Sciences of
Heinrich Heine University Düsseldorf

Examiners:
1. Prof. Dr. Björn Usadel
2. Prof. Dr. Benjamin Stich

Date of the oral examination:

DECLARATION

I declare under oath that I have produced my thesis independently and without any undue assistance by third parties under consideration of the 'Principles for the Safeguarding of Good Scientific Practice at Heinrich Heine University Düsseldorf. The dissertation has not yet been submitted in the submitted form or in a similar form by any other institution. I also declare that I have only submitted this dissertation in this and no other doctoral procedure and that this doctoral procedure was not preceded by a failed doctoral procedure.

Jülich, on 08.01.2025

Joseph Atemia

The results presented in this dissertation have been published in the following original publication or are summarized in the submitted manuscript included.

Joseph Atemia[†], **Ana Wegier, Diana María Rivera-Rodríguez, Alicia Mastretta-Yanes, Nancy Gálvez-Reyes, Lino de la Cruz-Larios, José de Jésus Sánchez-González, and Asis Hallab** (2024). From Habitat to Genotype: The Complex Interplay of Climate, Phenotypes, and Taxonomy in Teosinte

[†]First author

Constantin Eiteneuer[†], **David Velasco**[†], **Joseph Atemia**[†], **Dan Wang, Rainer Schwacke, Vanessa Wahl, Andrea Schrader, Julia J. Reimer, Sven Fahrner, Roland Pieruschka, Ulrich Schurr, Björn Usadel and Asis Hallab** (2022). GXP: Analyze and plot plant omics data in web browsers

[†]Share first author

CONTENTS

SUMMARY

Teosintes, the wild relatives of maize, are vital for global food security due to their rich genetic diversity and potential for crop improvement in response to climate change. This thesis delves into the genetic basis of teosinte's adaptation to diverse climatic environments and introduces a bioinformatic tool for advancing plant omics research. Through a comprehensive genome-wide association study (GWAS) of 3,455 individuals from 276 teosinte populations across Latin America, we analyzed environmental, morphological, and genomic data, identifying key Single Nucleotide Polymorphisms (SNPs) associated with climate resilience and morphological traits. These results reveal significant genetic differentiation shaped by local environmental conditions, highlighting the critical need for conservation strategies that preserve teosinte's adaptive potential. The study emphasizes the importance of both *in situ* and *ex situ* conservation to safeguard genetic diversity for future maize breeding. Complementing this ecological and genomic investigation, we developed Gene Expression Plotter (GXP), a web-based tool for visualizing and analyzing (Ribonucleic acid sequencing) RNAseq and metabolomics data. GXP eliminates the need for custom installations or manual programming, allowing researchers to directly explore omics data in their browsers. This tool accelerates the interpretation of transcriptomic results, making it particularly useful for projects where understanding gene expression linked to adaptive traits is crucial.

ZUSAMMENFASSUNG

Teosintes, die wilden Verwandten des Mais, sind aufgrund ihrer reichen genetischen Vielfalt und ihres Potenzials zur Züchtung von Maisvarianten, die den Herausforderungen des Klimawandels widerstehen können, von entscheidender Bedeutung für die globale Nahrungsmittelsicherheit. Diese Arbeit befasst sich mit den genetischen Grundlagen der Anpassung von Teosintes an unterschiedliche Umgebungen und stellt ein bioinformatisches Werkzeug zur Datenverwaltung, interaktiven Visualisierung und Analyse der Omics-Forschung, insbesondere der Transkriptomik, vor. Im Rahmen einer umfassenden genomweiten Assoziationsstudie (GWAS) von 3.455 Teosinte-Akzessionen aus 276 Teosinte-Populationen aus ganz Zentralamerika analysierten wir Umwelt-, morphologische und genomische Daten und identifizierten wichtige Single Nucleotide Polymorphisms (SNPs), die mit Klimaresistenz und morphologischen Merkmalen in Zusammenhang stehen. Diese Ergebnisse zeigen eine signifikante genetische Differenzierung, die durch die lokalen Umweltbedingungen geprägt ist, und unterstreichen den dringenden Bedarf an Schutzstrategien, die das Anpassungspotenzial von Teosintes bewahren. Die Studie betont die Bedeutung sowohl der *In situ*- als auch der *Ex situ*-Erhaltung, um die genetische Vielfalt für die zukünftige Maiszüchtung zu sichern. Ergänzend zu dieser ökologischen und genomischen Untersuchung haben wir Gene Expression Plotter (GXP) entwickelt, ein webbasiertes Tool zur Visualisierung und Analyse von Transkriptom- (RNAseq) und Metabolomics-Daten. GXP macht benutzerdefinierte Installationen oder manuelle Programmierung überflüssig und ermöglicht es Forschern, Omics-Daten direkt im Web-Browser Browsern zu verwalten, visualisieren und interaktiv zu untersuchen. Dieses Tool beschleunigt die Interpretation transkriptomischer Ergebnisse und ist daher besonders nützlich für Projekte, bei denen das Verständnis der Genexpression im Zusammenhang mit adaptiven Merkmalen von entscheidender Bedeutung ist.

List of Abbreviations

ABS	Access and Benefit Sharing
ANOVA	Analysis of Variance
CBD	Convention on Biological Diversity
CIMMYT	International Maize and Wheat Improvement Center
CWR	Crop Wild Relatives
DE	Differential Expression
DEG	Differentially Expressed Gene
DEGES	DEG Elimination Strategy
FC	Fold Changes
FDR	False Discovery Rate
Fst	Fixation Index
GBF	Global Biodiversity Framework
GSEA	Gene Set Enrichment Analysis
GXP	Gene Expression Plotter
HISAT	Hierarchical Indexing for Spliced Alignment of Transcripts
ICN	International Union for Conservation of Nature
KEGG	Kyoto Encyclopedia of Genes and Genomes
MDS	Multi-dimensional Scaling
ORA	Over-Representation Analysis
PCA	Principal Component Analysis
PCs	Principal Components
PPI	Protein-Protein Interaction
RLE	Relative Log Expression
scRNA-Seq	Single-cell RNA sequencing
SNP	Single Nucleotide Polymorphism
TCC	Tag Count Comparison
TMM	Trimmed Mean of M values
USDA-ARS	United States Department of Agriculture - Agricultural Research Service
WGCNA	Weighted Gene Correlation Network Analysis
ZeaGBS	Zea Genotyping-by-Sequencing

CHAPTER 1

General introduction

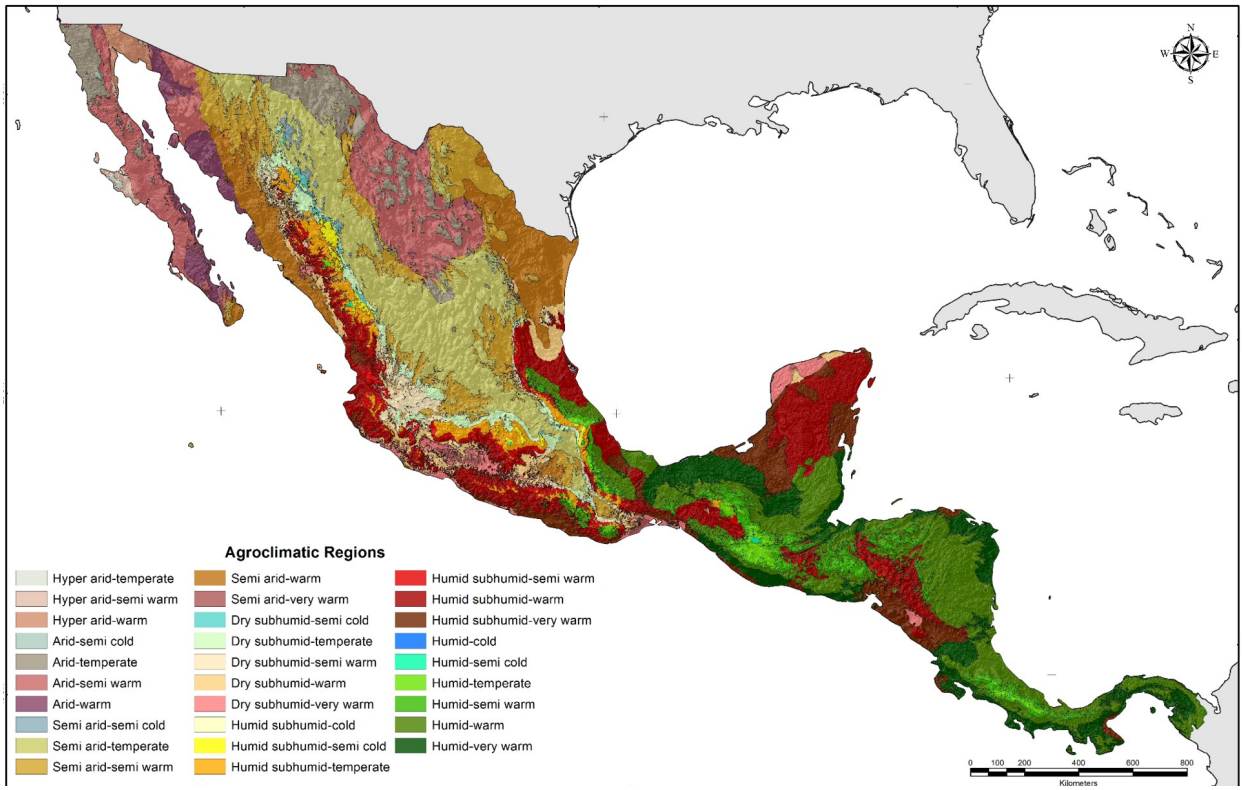
Taxonomy and Distribution of Teosinte: A Diverse Genetic Reservoir

Teosinte, a group of wild grasses in the family *Poaceae* and genus *Zea*, is native to Meso-America and categorized into two main sections: *Luxuriantes* and *Zea*. The section *Zea* includes several wild relatives of maize, with three recognized subspecies of *Zea mays*: *Zea mays* ssp. *mexicana* (*mexicana* hereafter), *Zea mays* ssp. *parviglumis* (*parviglumis* hereafter), and *Zea mays* ssp. *huehuetenangensis*, the latter being endemic to western Guatemala (Hufford et al., 2012; Rivera-Rodríguez et al., 2019; Sánchez González et al., 2018). *Parviglumis* and *mexicana* are particularly important as they represent the primary ancestral taxa from which maize was domesticated, with evidence pointing to a process involving introgression between the two species (Hufford et al., 2012; Moreno-Letelier et al., 2020; Yang et al., 2023).

Mexicana occupies a broad geographic range across Mexico, thriving in the cooler, drier highlands of northern and central Mexico at altitudes of 1600 to 2700 meters, while *parviglumis* is found at lower, warmer elevations (<1800 meters) in southwestern Mexico, where the species adapts to moderately wet conditions (Figure 1). The *mexicana* and *parviglumis* taxa generally maintain distinct geographic ranges, except in the eastern Balsas River Basin, where there is evidence of gene flow, suggesting the presence of a hybrid zone or possibly an ancestral gene pool for both subspecies (Hufford et al., 2012; Moreno-Letelier et al., 2020; Sánchez González et al., 2018). In addition to *Zea mays*, the genus *Zea* includes several species classified under Section *Luxuriantes*, which are discussed in detail in the second section of the thesis.

Figure 1

A.


Agroclimatic Regions
Hyper arid-temperate
Hyper arid-semi warm
Hyper arid-warm
Arid-semi cold
Arid-temperate
Arid-semi warm
Arid-warm
Semi arid-semi cold
Semi arid-temperate
Semi arid-semi warm
Semi arid-warm
Semi arid-very warm
Dry subhumid-semi cold
Dry subhumid-temperate
Dry subhumid-semi warm
Dry subhumid-warm
Dry subhumid-very warm
Humid subhumid-cold
Humid subhumid-semi cold
Humid subhumid-temperate
Humid subhumid-semi warm
Humid subhumid-warm
Humid subhumid-very warm
Humid-cold
Humid-semi cold
Humid-temperate
Humid-semi warm
Humid-warm
Humid-very warm
0 100 200 400 600 800
Kilometers

B.

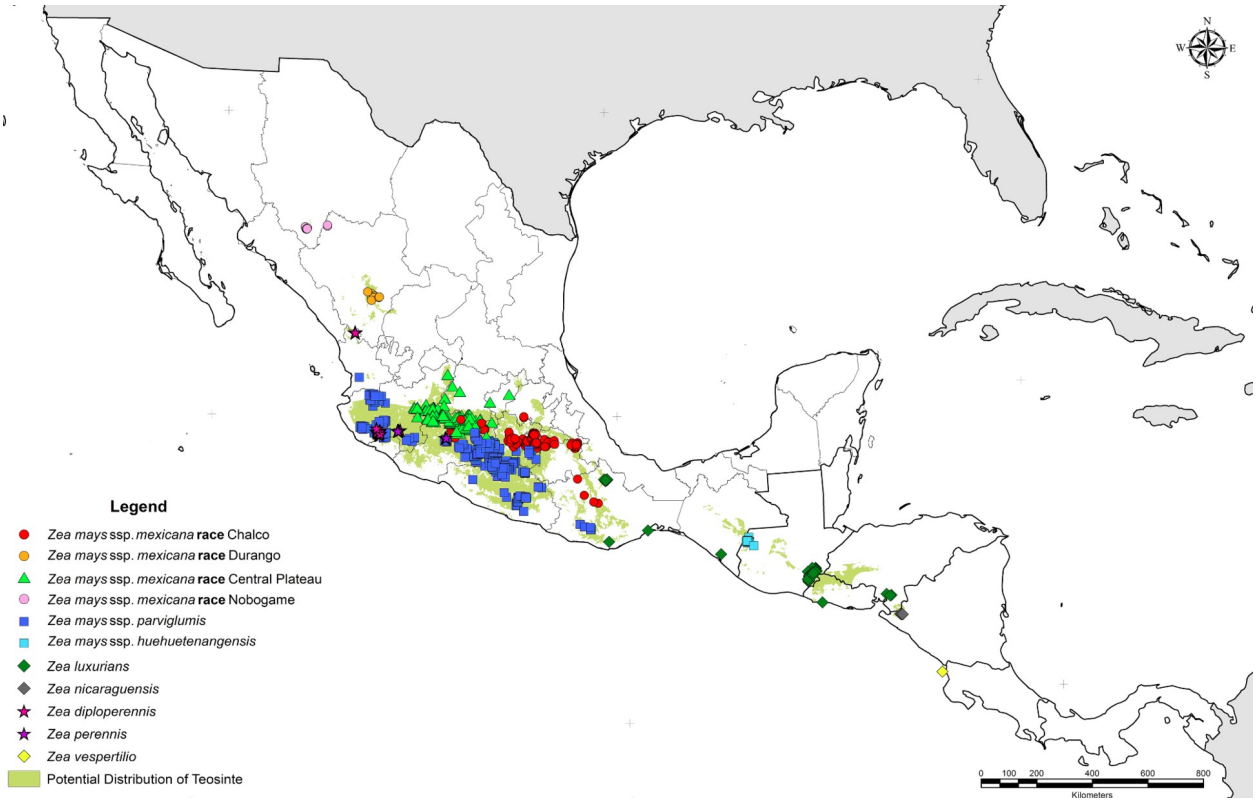

Legend
Zea mays ssp. mexicana race Chalco
Zea mays ssp. mexicana race Durango
Zea mays ssp. mexicana race Central Plateau
Zea mays ssp. mexicana race Nobogame
Zea mays ssp. parviglumis
Zea mays ssp. huehuetenangensis
Zea luxurians
Zea nicaraguensis
Zea diploperennis
Zea perennis
Zea vespertilio
Potential Distribution of Teosinte
0 100 200 400 600 800
Kilometers

Figure 1. Geographic and environmental diversity of sampled teosinte habitats (Reproduced with permission from Sánchez González et al., 2018). Panel **A** highlights the wide range of climatic and agro-climatic regions represented in the study, including arid, temperate, and humid environments. Panel **B** Sampling locations span diverse landscapes, from coastal plains to mountainous regions, reflecting variation in elevation, growing seasons, and soil conditions. This comprehensive sampling captures the ecological breadth of teosinte's natural habitats.

Genomic Structure and Domestication of Maize

The genome sizes of maize and teosinte are similar, approximately 2.3 gigabases (Gb), yet they exhibit substantial structural and genetic diversity, largely driven by transposable elements. For example, retrotransposons make up over 80% of the maize genome, significantly contributing to its structural complexity and adaptability (Sanmiguel and Bennetzen 1998). Despite this structural diversity, key genomic regions and conserved genes have maintained sufficient similarity to enable the domestication of maize from teosinte through selective breeding and evolutionary processes over the past ~9,000 years in southwestern Mexico (Matsuoka et al. 2002).

Maize domestication involved major phenotypic transformations, including reduced branching, enlarged ears, and softer kernels, enabling easier harvesting and storage. Five key domestication genes are identified in maize: *tb1* (teosinte branched 1), *tga1* (teosinte glume architecture 1), *ra1* (*ramosa1*), zfl2 (*Zea floricaula leafy2*), and *zagl1* (*Zea* agamous-like 1) (Vigouroux et al. 2002; Doebley 2004; Sigmon and Vollbrecht 2010). These genes control significant morphological changes, such as increased apical dominance (tb1) and reduced shattering (sh1), each promoting traits beneficial for agriculture. The tb1 gene, for example, enhances apical dominance by reducing lateral branching, which is critical for the single-stalk growth habit typical of maize (Doebley et al. 1995). Understanding these genetic controls allows researchers to trace the domestication pathways from teosinte to maize, linking genomic changes to observable agronomic traits.

Selective sweeps during maize domestication significantly impacted regions controlling traits advantageous to early agriculture. Notable is the tb1 gene's "hopscotch" transposon insertion that increases its expression and thus reduces branching, facilitating the evolution from the highly branched teosinte to the single-stalk maize phenotype (Studer et al. 2011). Similarly, the tga1 gene underwent a key mutation that altered its protein sequence, reducing the hardness of the fruit case surrounding teosinte kernels. This change allowed the kernels to become exposed and easier to harvest, a crucial adaptation for early agricultural use and a defining trait in the domestication of maize (H. Wang et al., 2005). Additional structural variations, such as copy number variations (CNVs) and presence-absence

variations (PAVs), have been identified between teosinte and maize, contributing to phenotypic diversity (Schnable et al. 2009; Swanson-Wagner et al. 2010). Importantly, these variations are not randomly distributed across the genome but are often concentrated in regions associated with domestication-related traits, reflecting targeted selective pressures during the domestication process (Studer et al., 2011; H. Wang et al., 2005). This genetic divergence highlights the selective pressures exerted on teosinte and demonstrates how human-mediated selection, alongside genetic diversity shaped by natural adaptation, contributed to the development of the highly adaptable maize genome used in agriculture today (Hufford et al., 2013; Wright et al., 2005).

Teosinte has contributed valuable nutritional traits to maize, which have been selectively enhanced through domestication. This includes genes associated with kernel size, oil content, and protein quality, which are critical for the nutritional and economic value of maize (S. A. Flint-Garcia et al., 2009; Karn et al., 2017; Zavala-López et al., 2018). Maize's evolution from teosinte has led to a remarkable increase in kernel size and number, which are advantageous for yield. Additionally, teosinte's genetic diversity is valuable for fortifying maize with bioavailable micronutrients, essential for food security. This genetic foundation supports ongoing efforts to breed maize varieties with improved nutritional content, which is crucial for addressing malnutrition in maize-dependent regions.

Teosinte as a Genetic Resource: Enhancing Maize Resilience through Diverse Adaptations

Teosinte’s rich genetic diversity and adaptation to a broad range of Mesoamerican environments make it an invaluable resource for maize improvement. Despite its importance, modern maize breeding practices aimed at enhancing specific traits like yield, disease resistance, and kernel size have inadvertently reduced genetic diversity in the maize genome, leading to a narrowing of the gene pool and an increase in the proportion of rare alleles (Jiao et al., 2012). As maize’s ancestral species, teosinte has preserved extensive genetic diversity through millennia of natural selection in diverse climatic and agroecological conditions, spanning environments from arid to temperate, humid regions, mountainous terrains, and varying soil types across Mesoamerica (Sánchez González et al., 2018). This variation in teosinte climatic growing conditions has led to genetic adaptations responding to environmental pressures such as drought, soil nutrient levels, and temperature fluctuations (Chen et al., 2022; Pyhäjärvi et al., 2013). The genetic adaptations provide traits that can enhance maize's resilience to stresses aggravated by climate change, such as extreme

weather events, shifting pest and pathogen patterns, and fluctuations in water availability (Amusan et al., 2008).

Beyond its role in agricultural resilience, teosinte's rich genetic diversity offers a unique window into the evolutionary processes and genetic mechanisms that drive complex traits in the genus *Zea*. Research on teosinte has provided key insights into the genetic foundations of traits such as flowering time, plant architecture, and kernel composition traits shaped by both natural selection and the processes of domestication (S. Flint-Garcia, 2017). For example, the introgression of the circadian clock gene *ZmPRR37a* from *Zea mays* ssp. *mexicana* into maize has been shown to likely enhance adaptation to higher latitudes (Chen et al., 2022; Yang et al., 2023). Such findings not only aid in conserving crop wild relatives' biodiversity but also bolster breeding programs focused on sustaining maize production across diverse and shifting environments.

Challenges in Conserving Teosinte's Genetic Diversity

Preserving teosintes' genetic diversity presents both scientific and logistical challenges, especially as natural teosintes' populations face pressures from habitat loss, climate change, and agricultural expansion. To maintain the genetic variability of teosinte, a dual approach involving both *in situ* and *ex situ* conservation is needed. *In situ* conservation, which protects teosinte within its natural habitat, allows for the ongoing evolution of these populations in response to local environmental pressures, which in turn contributes to the persistence of adaptive traits relevant to maize crop improvement (Maxted et al., 2010, 2012). However, *in situ* efforts require coordinated habitat management and local community engagement to protect critical areas from land-use changes and degradation (FAO, 2020). Complementary *ex situ* conservation methods, such as seed banking, botanical gardens, and germplasm repositories, provide an essential safeguard against the loss of genetic diversity in natural populations (Maxted, 2013). These methods are also critical for research and breeding programs, enabling controlled access to diverse genetic materials and facilitating studies on genetic and phenotypic variation. Four major germplasm banks CIMMYT, INIFAP, the University of Guadalajara, and USDA-ARS (U.S. Department of Agriculture-Agricultural Research Service) hold some of the largest and most accessible collections of teosinte germplasm. These collections represent some of the most extensive and widely utilized resources for conservation and research. However, further collection missions are needed to address the underrepresentation of several species, including *Zea diploperennis*, *Zea perennis*, and wild *Zea mays ssp. huehuetenanguensis* populations from Guatemala and northwestern Mexico (Goettsch et al., 2021; Guzzon et al., 2021; Sánchez González et al., 2018)

These conservation efforts align with international biodiversity targets, including the Kunming-Montreal Global Biodiversity Framework, which underscores the importance of preserving crop wild relatives and fostering collaboration across countries for biodiversity conservation (CBD, 2022). The Framework advocates for integrated approaches that blend *in situ* and *ex situ* methods, recognizing that crop wild relatives like teosinte are not only valuable resources for future breeding efforts but also keystones for maintaining ecosystem balance. Achieving this balance demands sustained funding, policy support, and scientific engagement to navigate the challenges associated with genetic resource management (Maxted & Kell, 2009).

Despite these efforts, the conservation of teosintes' genetic diversity remains precarious. National and international policies need to further prioritize the protection of wild relatives, integrating them into agricultural and environmental agendas to secure these genetic resources for the future (FAO, 2020; Maxted et al., 2013). Given the rapid pace of environmental change, establishing and maintaining dynamic conservation systems for teosintes is essential to preserve its evolutionary potential and to support ongoing improvements in maize breeding.

Omics Tools and Their Role in Plant Research

Modern advancements in life sciences have led to a massive increase in quantitative data across diverse omics fields, including genomics, transcriptomics, proteomics, epigenomics, metabolomics, and many more. Technologies like RNA sequencing (RNA-seq) produce large volumes of gene expression data, offering crucial insights into gene regulation and metabolic pathways across various conditions and treatments (Stark et al., 2019; Z. Wang et al., 2009). To facilitate meaningful interpretation, researchers often employ bioinformatics pipelines for data preprocessing, which are often outsourced to specialized laboratories or core facilities (Conesa et al., 2016). Once data is generated, however, researchers face the challenge of analyzing and visualizing complex datasets to uncover biological patterns and generate testable hypotheses. These steps typically include statistical analyses such as principal component analysis (PCA), hierarchical clustering, and enrichment analysis to identify overrepresented functional categories among differentially expressed genes or metabolites (Langfelder & Horvath, 2008; Subramanian et al., 2005).

To address these challenges, various international and national efforts have significantly enhanced bioinformatics capabilities. Initiatives like de.NBI (the German Network for Bioinformatics Infrastructure) and NFDI (National Research Data Infrastructure) provide robust platforms and resources for secure and comprehensive omics data analysis. Similarly,

ELIXIR, a pan-European infrastructure for life sciences data, has played a pivotal role in integrating bioinformatics resources, facilitating access to high-quality tools, and ensuring secure data sharing across borders (Drysdale et al., 2020). These efforts complement specialized tools for RNA-seq and metabolomics data, which often require programming skills or involve sharing potentially sensitive data on third-party platforms (Conesa et al., 2016; Eiteneuer et al., 2022; Love et al., 2014). This limitation can pose challenges for researchers, particularly those who need streamlined, secure, and user-friendly solutions for downstream analysis of omics data. In response, graphical user interface (GUI) based tools have been developed to offer user-friendly solutions, allowing researchers to perform clustering, enrichment analyses, and visualization without extensive coding requirements (Patro et al., 2017; Soneson & Robinson, 2018). However, many existing tools focus narrowly on specific organisms or data types, which can limit their applicability across diverse research contexts, especially when working with integrated datasets from multiple omics sources (Hernández-de-Diego et al., 2018; Julkowska et al., 2019).

In the context of plant research, where integrated omics data hold transformative potential, we developed GXP (Gene Expression Plotter), a versatile and secure tool to enhance our teosinte comparative genomics study (Eiteneuer et al., 2022). Initially designed to support and streamline our ongoing transcriptomic data analysis, GXP is instrumental in facilitating the exploration of gene expression patterns and providing intuitive data visualization. GXP is a user-friendly tool that allows researchers to visualize and analyze omics data directly in their web browser. This tool accelerates the analysis of complex datasets, making it easier to identify key genes and biological patterns. Addressing the need for user-friendly, integrative tools, GXP enables in-depth analysis across diverse omics types directly on researchers' devices, safeguarding both data privacy and reproducibility. With robust capabilities for visualization, statistical analysis, and secure data handling, GXP is an ideal analysis platform especially where comprehensive, cross-omics pipelines are essential. Beyond simplifying omics data analysis, GXP aligns with the increasing emphasis on transparency and reproducibility in scientific research, fostering collaborative and reproducible practices (Peng, 2011; Wilkinson et al., 2016).

Thesis Outline

The research described in this thesis, Comparative Genomics and Population Genetics of the Wild Relatives of Maize, investigates the genetic diversity and adaptive potential of teosintes to support breeding and conservation strategies for maize. Through GWAS and population structure analysis, this study identifies unique genomic regions associated with climate-driven adaptations, revealing the high genetic diversity within teosinte

populations and their distinct adaptations to specific environmental conditions. These findings highlight the importance of conserving teosinte populations in their natural habitats to preserve their valuable adaptive traits. Additionally, this thesis presents GXP, a novel web-based tool developed for the efficient visualization and exploration of plant omics data, which complements the analytical approaches used here and facilitates future research on complex adaptive traits.

Chapter 2 investigates the genetic diversity and population structure of teosinte, analyzing adaptations across distinct environments to understand their implications for conservation and breeding. Our findings reveal that teosinte populations exhibit high genetic diversity, with unique adaptations to specific environmental conditions, including distinct genetic loci associated with similar climates. This diversity underscores teosinte's potential as a resource for breeding climate-resilient maize and highlights the need to prioritize the preservation of all distinct populations in their natural habitats to safeguard their adaptive traits and genetic diversity for future use.

Chapter 3 presents the GXP tool for visualizing and analyzing plant omics data, showcasing its application to teosinte datasets in a web-based format. GXP enables efficient, high-throughput exploration of omics data, facilitating insights into complex adaptive traits across populations and environments.

Chapter 4 synthesizes findings from previous chapters, discussing their implications for crop improvement and conservation. This chapter contextualizes the adaptive SNPs and genomic regions identified in teosinte, emphasizing their potential for developing climate-resilient maize varieties. It also examines the challenges of genetic linkage and introgression, especially when transferring traits associated with environmental adaptation. The chapter concludes with recommendations to prioritize conservation efforts for distinct teosinte populations and highlights strategies for leveraging teosintes' unique genetic diversity to enhance maize resilience in response to global climate change.

CHAPTER 2

From Habitat to Genotype: The Complex Interplay of Climate, Phenotypes, and Taxonomy in Teosinte

This manuscript was submitted to Molecular Ecology in December 2024.

Authors:
***Joseph Atemia**, Ana Wegier, Diana María Rivera-Rodríguez, Alicia Mastretta-Yanes, Nancy Gálvez-Reyes, Lino de la Cruz-Larios, José de Jésus Sánchez-González, and Asis Hallab*[#]

#Corresponding author: a.hallab@fz-juelich.de

Contribution: First author

Joseph Atemia performed the data analyses. Ana Wegier, Alicia Mastretta-Yanes, and Asis Hallab designed the research and supervised the project. **Joseph Atemia**, Ana Wegier, Alicia Mastretta-Yanes, and Asis Hallab wrote the manuscript. Nancy Gálvez-Reyes, Lino de la Cruz-Larios, José de Jesús Sánchez-González, and Diana María Rivera-Rodríguez contributed to data gathering and validation, provided critical feedback on the GWAS analyses, and assisted with the interpretation of results. All authors reviewed and approved the final manuscript.

Abstract

Teosintes, the wild relatives of maize, exhibit a wide ecogeographic distribution across Mexico and Central America, spanning starkly varying precipitation and temperatures. Understanding the genetic basis of teosinte's adaptation to such conditions is crucial for its *in situ* conservation. We present findings from a study of 3,455 individuals across 276 teosinte populations, encompassing all known taxa except *Zea vespertillio*. Environmental and

morphological data, along with genotype data comprising 33,929 SNPs, were analyzed to elucidate the genetic population structure, ecological adaptation, and candidate genes associated with various climatic factors and phenotypic traits. Our results revealed distinct genetic and phenotypic adaptations within teosinte populations, shaped by the climate conditions of their habitats. Genome-wide association studies (GWAS) identified significant SNPs associated with morphological traits and environmental factors, elucidating adaptive mechanisms in teosinte evolution. Comparative analysis with maize literature on GWAS on SNPs found to be associated with agronomically important maize phenotypes highlighted both shared and unique genetic variants between teosinte and maize. Furthermore, protein function annotated to marker loci regions revealed the multifaceted nature of adaptive strategies in teosinte, indicating different potentially adaptive loci, even between populations growing in similar environmental conditions. Recognizing this diversity is important for teosinte conservation, as pointed out by new international frameworks, and for its management considering teosinte's gene flow with maize, potential transgenic flow, and the risk of new weeds emergence. Our research underscores the importance of studying the genetic diversity of crop wild relatives at the population level within centers of origin.

Introduction

Human survival relies heavily on plants, with sugarcane, maize, wheat, and rice contributing significantly to global food production. These four crops account for a substantial share of global crop production and provide around 50-60% of the world's direct caloric intake (FAOSTAT, 2023). Making informed decisions regarding the conservation of genetic diversity in these species, and their crop wild relatives (CWR) species, which are species phylogenetically related to crops, including their ancestral and closely related species (Maxted et al., 2006), is critically influenced by the countries where these plants are distributed. These countries are often located in the crop's centers of origin and domestication. Given the importance of genetic diversity for food security, the Convention on Biological Diversity (CBD) has progressively expanded its focus on the conservation of cultivated plants and their wild relatives. The Cartagena Protocol initially emphasized safeguarding the centers of origin, diversity, and domestication of cultivated plants from the presence of genetically modified organisms. However, it was with Aichi Target 13 (https://www.cbd.int/sp/targets/) that the explicit conservation of the genetic diversity of cultivated plants and their wild relatives became a defined objective, with a primary focus on *ex situ* conservation. This commitment has been further advanced in the Kunming-Montreal Global Biodiversity Framework (GBF), where genetic diversity within and between populations is to be maintained, including in crop wild

relative species (CBD, 2022; Hoban et al., 2024). Importantly, this highlights the need for *in situ* conservation actions at the population level. Focusing on populations is valid because extinction begins with the loss of populations, and losing populations results in the loss of local adaptations that may not exist elsewhere in the species distribution, thereby reducing the species' overall adaptive potential (Hoban et al., 2024). For crop wild relatives distributed in their centers of origin and domestication, *in situ* conservation presents additional challenges due to gene flow with their domesticated counterparts, increasing pressures from intensive agriculture, and interactions with other species within the agrobiodiversity system, all of which create a complex system that simultaneously benefits and threatens genetic diversity (Castañeda-Álvarez et al., 2016; Chen et al., 2015; Goettsch et al., 2021). Population-level studies that focus on the distribution of adaptive genetic diversity are crucial for monitoring and protecting CWR genetic diversity, but such studies are more commonly conducted in the context of *ex situ* conservation and plant breeding applications than for *in situ* conservation and public policy development (Dempewolf et al., 2017; Mastretta-Yanes et al., 2018; 2024).

CWR species richness is highest in the centers of origin of their related crops and nearby biogeographic regions (Castañeda-Álvarez et al., 2016). Mexico and the countries comprising ancient Mesoamerica are hotspots for CWR of some of the world's most important crops, including beans, chili peppers, pumpkins, and maize (Hummer & Hancock, 2015; Maxted & Vincent, 2021). The closest wild relatives of Maize (*Zea mays* ssp. *mays*) comprise nine taxa within the genus *Zea*, known as teosintes (CONABIO, 2011; Rivera-Rodríguez et al., 2023), which can hybridize readily among themselves and with maize (Rojas-Barrera et al., 2019). The *Zea mays* includes three subspecies: *Zea mays ssp. huehuetenangensis* from Guatemala, as well as *Zea mays* ssp. *parviglumis* (hereafter *parviglumis*) and *Zea mays* ssp. *mexicana* (hereafter *mexicana*), which have diverse geographic distributions in Mexico (Rivera-Rodríguez et al., 2019). Within them, *parviglumis* and *mexicana* represent the ancestral taxa from which maize was domesticated, through a process involving introgression (Hufford et al., 2012; Moreno-Letelier et al., 2020; Yang et al., 2023). The remaining teosintes species belong to Section *Luxuriantes*, which is characterized by fragmented distributions with isolated populations. These species include *Z. diploperennis* and *Z. perennis*, which are distributed in Mexican tropical mountains, and *Z. luxurians*, native to southeastern Guatemala, Honduras, El Salvador, and southern México. Two other species, *Z. nicaraguensis,* and *Z. vespertilio* are restricted to Nicaragua and Costa Rica, respectively (Iltis & Benz, 2000; Laurito, 2013; Rivera-Rodríguez et al., 2023; Sanchez Gonzalez et al., 1998; Sánchez González et al., 2018). Mexico is also home to the greatest diversity of maize landraces (Perales & Golicher, 2014), which are still cultivated on approximately 4 million ha under campesino agriculture (Bellon et al., 2018). Modern maize varieties were also introduced to Mexico

between 1946 and 1960 and are widely cultivated today (Rojas-Barrera et al., 2019). Altogether, Mexico presents a complex scenario where gene flow in all directions between maize and teosintes (Rojas-Barrera et al., 2019). Consequently, genetic changes in either wild or domesticated populations can potentially transfer to the other and vice versa. Besides making *in situ* conservation more challenging from a biological perspective, this also has significant legal, political, economic, and social implications, especially when gene flow involves genetically modified maize (Acevedo et al., 2011; Agapito-Tenfen & Wickson, 2018).

CWR hold greater genetic diversity than their domesticated counterparts because they did not undergo the domestication bottleneck and continue to undergo natural selection in diverse habitats (Maxted et al., 2006). For instance, only around 10-20% of the genetic diversity present in wild rice species was retained in cultivated rice (Zhu et al., 2007), while in sunflowers, this is around 40-50% (A. Liu & Burke, 2006). From an ethical perspective, like other wild species, CWR have an inherent right to exist and be conserved (M. Alexander, 2013). Beyond this, their importance has been historically underscored through their contributions to improving cultivated crops (Dempewolf et al., 2017). Perhaps the most famous case is how a wild potato from Mexico (*Solanum demissum*), helped to confer domesticated potatoes (*S. tuberosum*) resistance to Phytophthora infestans, the pathogen that caused the Irish Potato Famine in the 1840s and continues to decimate European and U.S. potato cultivars (Salaman, 1937; Śliwka & Zimnoch-Guzowska, 2013). Modern examples of incorporating CWR into breeding include tomatoes (Rick & Chetelat, 1995; Yan et al., 2018) and teosintes, where for instance, gray leaf spot resistance was assessed in BC4S2 near-isogenic lines with genetic introgressions from the teosinte *parviglumis* in a maize B73 background (Lennon et al., 2016).

The recognition of CWR (Crop Wild Relatives) as critical sources of adaptive diversity for crop breeding has made them key targets for seed bank collections for over a century (Curry, 2022; Maxted et al., 2010). This recognition spurred the development of international frameworks, such as Aichi Target 13 in 2010, which explicitly aimed to preserve CWR genetic diversity and implement strategies to minimize genetic erosion (https://www.cbd.int/sp/targets/). Similarly, the International Treaty on Plant Genetic Resources for Food and Agriculture emphasized the importance of CWR conservation (Maxted et al., 2010). In tandem with conservation efforts, the extensive use of foreign genetic resources in crop breeding highlighted the need for equitable benefit-sharing, leading to the establishment of the Nagoya Protocol (Kohsaka, 2012; Maxted et al., 2010). These developments collectively drove increased efforts in *ex situ* conservation within gene banks and, to a lesser extent, inspired countries to adopt *in situ* conservation strategies e.g. (De La Torre S. et al., 2018; Perrino & Perrino, 2020; Satori et al., 2022). Historically, CWR

conservation strategies have focused on species-level representation e.g. (Bioversity International & University of Birmingham, 2017; Castañeda-Álvarez et al., 2016), with fewer initiatives addressing ecogeographic representation within species distributions e.g. (Parra-Quijano et al., 2012; Tobón-Niedfeldt et al., 2022). While these approaches have contributed to safeguarding genetic diversity in alignment with Aichi Target 13, measuring progress remained a challenge. The Global Biodiversity Framework (GBF), adopted in 2022, addresses this gap by introducing explicit indicators to track genetic diversity. The mandatory headline indicator A.4 measures the proportion of populations with sufficient effective population size (*Ne* > 500) to retain genetic diversity, while the complementary PM indicator provides additional support in estimating progress toward A.4.

Re-targeting genetic diversity conservation efforts toward *in situ* strategies, as recommended by the GBF, is particularly important for CWR. The loss of natural habitats due to land-use change and the expansion of intensive agriculture, in contrast to traditional agriculture where CWR can be tolerated and even encouraged, has contributed to the decline of CWR populations (Casas et al., 2007; Zamora-Tavares et al., 2015). Consequently, in Mesoamerica, approximately 35% of CWR species are classified as endangered by the IUCN Red List (Goettsch et al., 2021). Among teosintes, six out of nine taxa are at some level of extinction risk (Goettsch et al., 2021). This includes all assessed species of the section Luxuriantes, as well as both *mexicana* races from northern Mexico, namely Durango and Nobogame. The only taxa categorised as Least Concern are the central Mexican *mexicana* races (Mesa Central and Chalco) and *parviglumis*. However, this assessment assumes that the former taxa are homogeneous units, which is not supported by genetic evidence. A recent study on the genetic population structure of teosintes involving 33,929 SNPs and 3,604 individuals of all taxa except *Z. vespertilio*, revealed that there is indeed considerable genetic diversity within the taxonomic groups (Rivera-Rodríguez et al., 2023). Subspecies *parviglumis* comprises 13 genetic clusters, some of them forming in highly isolated locations. Subspecies *mexicana* comprises five genetic clusters, which do not fully correspond to the race subdivision; for instance, the Mesa Central race includes two genetic clusters distributed along a West-East gradient (Rivera-Rodríguez et al., 2023). Additionally, the study revealed significant differences in the levels of genetic diversity within populations across all seven taxa, with many populations showing lower genetic variation than expected, even within *parviglumis* and *mexicana* (Rivera-Rodríguez et al., 2023). Thus, even taxa categorized as Least Concern by the IUCN Red List may experience a loss of genetic diversity. A pilot study assessing GBF genetic diversity indicators in >900 species confirms this pattern. Among Least Concern taxa in the study, 117 out of 292 (40%) had a *Ne* 500 indicator value of less than 0.5,

meaning fewer than half of their populations are large enough to avoid significant loss of genetic diversity by genetic drift (Mastretta-Yanes et al., 2024).

The teosintes *parviglumis, mexicana, Z. luxurians, Z. perennis,* and *Z. diploperennis* were included in the above-mentioned study. All of them showed low values of the *Ne* 500 indicator, except for *parviglumis*, which was classified as data deficient for that metric. However, the PM indicator is more positive because it shows that all assessed teosintes still maintain all their known populations, except for *Z. luxurians*, which was collected from a locality in Oaxaca in 1842 but has not been found there in recent surveys (Sánchez González et al., 2018). This is interesting news because teosintes are distributed across a wide range of very different ecological conditions (Sánchez González et al., 2018). Therefore, it is reasonable to assume that if most populations still exist, they would retain the range of local adaptations that could potentially occur within teosintes. Previous studies on some *mexicana* and *parviglumis* populations have shown that indeed local adaptation plays an important role, with an adaptive divergence between taxa occurring despite gene flow (Aguirre-Liguori et al., 2019) and populations located at the limit of the species' ecological niches show increased genomic signatures of local adaptation to new unsuitable environments (Aguirre-Liguori et al., 2017). However, previous genetic studies for conservation planning of teosintes have focused on neutral genetic structure (Rivera-Rodríguez et al., 2023), thereby overlooking the distribution of such local adaptations. Since the distribution of neutral and adaptive genetic variation should be integrated to optimize conservation efforts (Funk et al., 2012), this integrated approach is crucial for teosintes. This would allow designating and monitoring conservation units that truly encompass the larger plethora of "genetic options" that teosintes offer to cope with environmental change themselves and to help maize breeding for the wide range of agricultural systems, including traditional ones. Paradoxically, understanding the distribution of teosintes' adaptive genetic diversity is also needed because they can pose a threat to agriculture: teosintes have already been established as invasive weeds in Spain and France, thriving under climatic conditions outside their native range and indicating rapid genetic evolution (Corre et al., 2020).

In this study, we contribute to teosintes' conservation genetics planning and monitoring by investigating the associations between 19 morphological traits, 246 environmental variables, and corresponding genotype variations among 3,455 teosinte individuals across all taxa, except *Z. vespertilio.* For each taxon, we first explore the range of genetic, morphological, and environmental differences at the population level. Next, we follow a genome-wide association study (GWAS) approach (Uffelmann et al., 2021) to identify associations between candidate loci and both morphological traits and environmental variables, identifying protein function annotation in associated loci region candidate genes.

Finally, we compare the candidate genes found in teosintes against those previously reported for maize, to examine whether teosintes' adaptive variation is found in their domesticated counterpart. Our findings highlight the diverse teosinte adaptation signatures, emphasizing the urgent need for enhanced conservation strategies that incorporate *in situ* approaches and recognize maize and teosintes as a wild-domesticated-system within their center of origin.

Materials and methods

Teosinte sampling

To identify the genetic loci associated with the diversity of teosinte populations, we used a panel of 3,455 individuals from 276 populations of teosinte taxa. These individuals were collected between 1968 and 2007, with both the median and mode collection dates falling in 2007. The panel encompasses all known species, subspecies, and races of teosintes, except *Zea vespertillio* (Rivera-Rodríguez et al., 2023; Sánchez González et al., 2018).

Environmental data

The extensive collection of 3,455 individuals allowed us to capture the wide climatic variability within the ecogeographic distribution of teosinte. We incorporated a comprehensive set of 237 climatic variables, assessed at the collection sites of these individuals (see TableS3 for the full list of variables). The characterization of these environmental factors was based on data obtained from the National Environmental Information System (NEIS), as detailed in previous studies (Rivera-Rodríguez et al., 2019; Sánchez González et al., 2018). The 237 climatic variables include conditions such as temperature, precipitation, humidity, solar radiation, and photoperiod length, among others, most of them are crucial for understanding the ecogeographic distribution and environmental adaptations of teosinte, as previously modeled by (Sánchez González et al., 2018). For further context, see Table 1 from Sánchez González et al. (2018), which provides a detailed overview of the variables analyzed.

Phenotypic data

The climatic data were integrated with phenotypic data. Eighteen morphological and physiological traits were characterized for the 3,455 teosinte individuals included in this study, as reported in a previous work (Rivera-Rodríguez et al., 2019). In that study, plants were grown under controlled greenhouse conditions at the Centro Universitario de Ciencias Biológicas y Agropecuarias (CUCBA) in Nextipac, Jalisco, Mexico (20°45′N, 103°31′W), to ensure consistency in environmental factors influencing growth and development. This

controlled environment, maintained from June 6, 2014, to June 17, 2015, allowed for precise quantification of morphological traits, minimizing the variability caused by environmental factors (Rivera-Rodríguez et al., 2019). The traits assessed included structural features such as plant height, leaf width, spikelet width, and total surface area; flowering time-related traits such as heat units to silk and heat units to pollen shed; and other traits essential for understanding the phenotypic diversity and potential adaptive significance within teosinte (see TableS3 for the full list of variables).

Subsequently, we assessed the similarity of teosinte habitat climatic conditions for pairs of teosinte populations. This was calculated as the overlap or *relative range intersection* (RRI) of values measured for a given climate factor for two populations *A* and *B* as:

$$RRI(A,B) = \frac{min(max(A), max(B)) - max(min(A), min(B))}{max(A \cup B) - min(A \cup B)}$$

Note that negative RRI values are set to zero.

Genotype data

In our study, we utilized a dataset comprising 33,929 single nucleotide polymorphisms (SNPs) identified in 3,455 teosinte individuals using genotyping by sequencing (GBS). This dataset was originally generated as part of a previous investigation by (Rivera-Rodríguez et al., 2023) and was processed using the Tassel5GBS Production Pipeline (ZeaGBSv2.7 Production TOPM v2.7) (Glaubitz et al., 2014; Rivera-Rodríguez et al., 2023). The pipeline reference incorporates genotypic information from a comprehensive collection of over 60,000 maize samples, ensuring a robust representation of genetic variation within the maize species. Rodríguez et al. (2023) describe the implementation and quality control processes in detail.

Genetic diversity and population structure

To assess genetic diversity, fixation statistics (Fst) were computed from the SNPs dataset using the hierfstat R package v0.5-11 (Goudet, 2020), providing insights into the degree of genetic differentiation among teosinte populations. Furthermore, principal component analysis (PCA) was conducted on both genotype data and phenotype-environmental data to elucidate patterns of variation and clustering among teosinte individuals. The PCA computation was conducted using the *prcomp* function in R, and results

were visualized with the *factorextra* R package v1.0.7 (Kassambara, 2016). This approach allowed for a comprehensive exploration of genetic variability and population structure and its relationship with phenotypic and environmental factors. This was also conducted to uncover the potential correlations between these parameters allowing for insights into the adaptive responses of teosinte populations to varying ecological conditions. Moreover, admixture analysis results previously conducted by Rodríguez et al. (2023) using ADMIXTURE software (Alexander et al., 2009), were incorporated into our methodology. These results further clarified the population structure, the extent of genetic admixture, and the relationships among teosinte individuals.

GWAS and candidate gene mapping

To explore the genetic relationships between environmental variables and morphological traits across all teosinte taxa in our investigation we performed association testing using the Fixed and Random Model Circulating Probability Unification (FarmCPU) and Multi-Locus Mixed Model (MLMM) models available in GAPIT v3 (Liu et al., 2016; Segura et al., 2012; J. Wang & Zhang, 2021). These models were specifically selected for their ability to address confounding factors inherent in Genome-Wide Association Studies (GWAS), such as population structure and genetic relatedness, which can result in false associations if not properly accounted for.

FarmCPU addresses confounding by iteratively separating fixed and random effect models. The fixed effect model tests markers in the presence of cofactors (associated markers) identified during iterations, avoiding the confounding associated with traditional kinship-based methods (Liu et al., 2016; J. Wang & Zhang, 2021). Meanwhile, the random effect model uses a maximum likelihood approach to iteratively select associated markers, reducing the risk of overfitting seen in stepwise regression. This iterative process enhances statistical power, enabling FarmCPU to efficiently detect true marker-trait associations while controlling for population structure and relatedness (Liu et al., 2016; J. Wang & Zhang, 2021). MLMM builds upon traditional single-locus methods by iteratively incorporating multiple loci through forward-backward stepwise regression (Segura et al., 2012). This process allows MLMM to more effectively capture the polygenic nature of complex traits and refine the genetic architecture that single-locus models may overlook. MLMM incorporates kinship matrices to model genetic relatedness among individuals, using the kinship matrix as a variance-covariance matrix, while cofactors (associated markers) are iteratively adjusted during the regression process (Segura et al., 2012). The kinship matrix was computed using the VanRaden algorithm (VanRaden, 2008; J. Wang & Zhang, 2021). The method estimates additive genetic relationships by deriving a matrix that quantifies the proportion of shared

alleles among individuals allowing control for the background genetic similarity that can confound marker-trait associations. By leveraging this iterative model selection approach, MLMM identifies the most informative loci, reducing the likelihood of false positives (Segura et al., 2012). To account for population structure, we included principal components (PCs) as covariates in the models, as population stratification can cause false associations between markers and traits. The PCs summarize genetic variation and capture the underlying population substructure. Using GAPIT's built-in functions, we employed forward model selection using the Bayesian Information Criterion (BIC), to determine the optimal number of PCs for each phenotype (J. Wang & Zhang, 2021). Finally, to address the risk of false positives from multiple testing, we applied the Bonferroni correction to set a stringent significance threshold. This method adjusts the p-value threshold by dividing the alpha level by the number of independent tests(Sedgwick, 2012).

The methods described in this section were implemented using the GAPIT v3, which offers an integrated framework for efficient and robust GWAS (J. Wang & Zhang, 2021).

Candidate genes identification and protein function annotation

To identify candidate genes, all genetic loci with single nucleotide polymorphisms (SNPs) that were found to have a significant association with teosinte phenotypic traits or habitat climatic growing conditions were physically identified by mapping the GWAS significant SNPs onto the *Zea mays* Zm-B73-REFERENCE-NAM-5.0 reference genome (Woodhouse et al., 2021). Genes found within the 50 kb window upstream and downstream (±50 kb) flanking regions of the significant SNPs' positions were considered candidate genes. The functional annotation of these candidate genes was performed using Mercator4 and prot-scriber, enabling the identification of gene functions and biological pathways (Bolger et al., 2021; Jayakodi et al., 2023; Schwacke et al., 2019; Woodhouse et al., 2021).

Comparative analysis of Teosinte GWAS SNPs with maize literature GWAS SNPs

We leveraged data from the Atlas database to elucidate the overlap between significant SNPs identified in our Teosinte GWAS and those reported in existing maize literature (Tian et al., 2020). The Atlas database aggregates GWAS SNPs from 133 studies, compiled and remapped to the B73_v4 reference genome by the National Genomics Data Center at the Chinese Academy of Sciences (Tian et al., 2020). To ensure compatibility, we accessed the compiled GWAS SNP data through the Maize Genetics and Genomics Database (MaizeGDB) where the SNP coordinates from B73 RefGen_v2, v3, and v4 were previously remapped to

B73 RefGen_v5 by aligning 100-bp flanking sequences from RefGen_v4 SNPs (retrieved using bedtools) to RefGen_v5, ensuring a minimum threshold of 98% coverage and 98% sequence identity (Woodhouse et al., 2021). The SNPs were systematically compared to enable a match within a ±50 kb window to account for genomic variability and ensure a comprehensive assessment of SNP overlap between the Teosinte GWAS and the maize literature GWAS datasets. A SNP matrix was constructed and subsequently used to generate distance matrices based on angular distance, Hamming distance, and Euclidean distance. These distance matrices were then used to perform hierarchical clustering analysis, the resulting clusters were visualized using the Interactive Tree of Life (iTOL) tool (Letunic & Bork, 2024).

Results

Morphological and environmental differences between Teosinte populations

Genetic, morphological and environmental differences within teosinte species and between teosinte populations were assessed to elucidate the relationship among taxonomy, genetic population structure, habitat climatic conditions, and teosinte phenotypic traits. Correlation analysis between habitat climatic factors and phenotypic traits revealed a strong relationship, indicating a high degree of association between these variables (Table S1). Hierarchical clustering based on pairwise correlations identified eleven highly correlated groups (Figure 1).

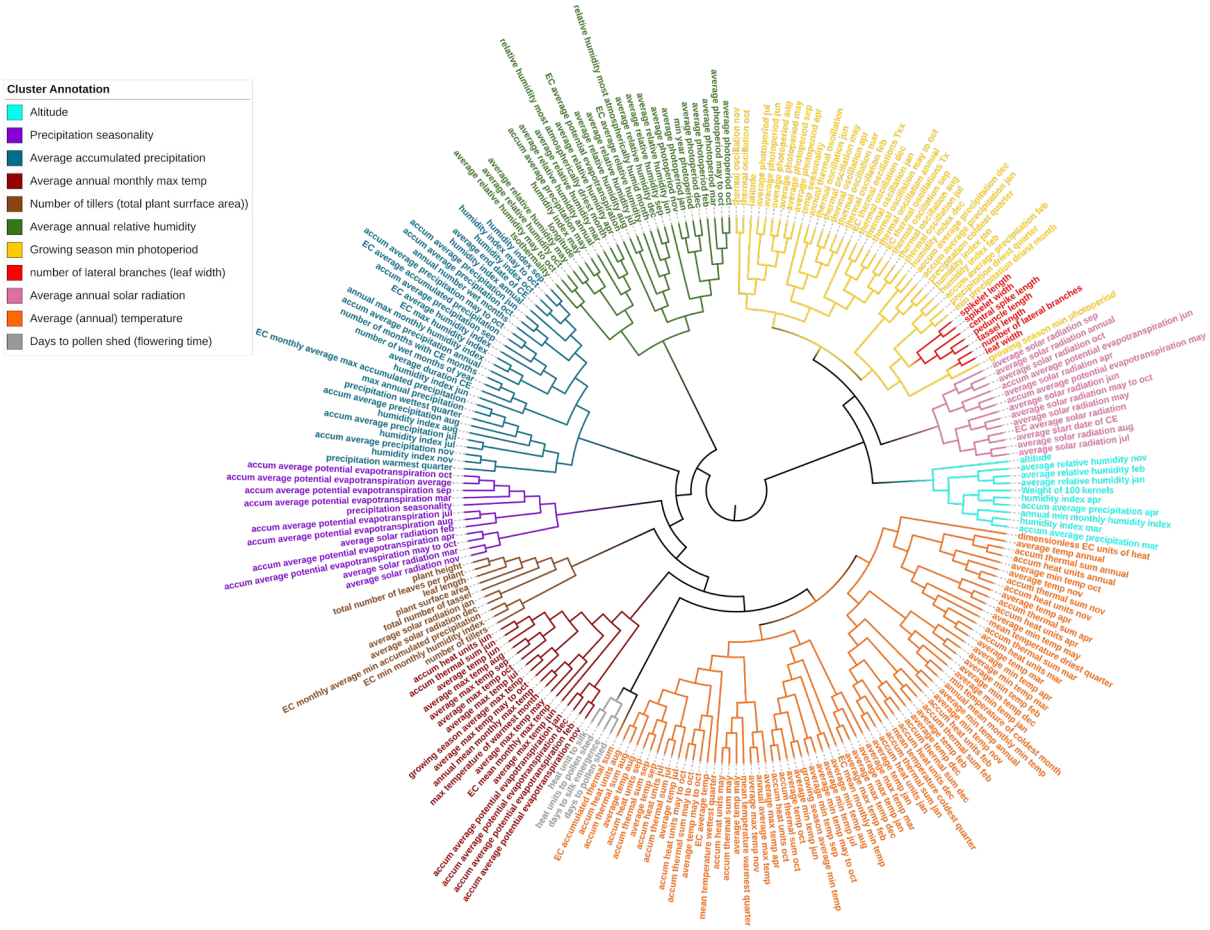


Figure 1. The circular dendrogram illustrates clusters derived from a correlation analysis of environmental factors and morphological traits, highlighting their interrelationships. Eleven distinct clusters are identified, each represented by a unique color and labeled based on the variable that contributes most significantly to the variation within that cluster. Each branch represents a factor or trait, grouped according to their correlation, with closely related variables forming cohesive clustered. To improve visual clarity and focus on the cluster groupings, branch lengths are not scaled. The variable with the highest contribution to variation within each cluster is used for labeling, characterizing the highly correlated variables associated with each cluster, as indicated in the key.

As shown in the biplot in Figure 2, phenotypic traits related to total plant photosynthetic surface exhibited strong correlations with humidity and precipitation, while reproductive organ dimension traits were correlated with temperature and soil moisture. These findings suggest adaptations for efficient reproduction processes under specific environmental conditions. The variations in climatic habitat conditions align well with teosinte taxonomy, as evidenced in Figure 2, where individual taxa exhibit distinct ranges of climatic growing conditions. Notably, some overlap in habitat conditions was observed between *parviglumis* and *mexicana* where aside from differences in cluster groups of temperature and flowering time factors, the remaining climatic factor groups shared a mean relative value range intersection (RRI; see

Methods) of 40% to 75% (Figure 3a). This distribution of climatic conditions underscores distinct ecological adaptations within teosinte taxa. For instance, *mexicana* is adapted to environments with low annual average temperatures (12.3 to 20.5 °C), low to moderate annual average precipitation (451 to 1321 mm), and moderate to high altitudes (1500 to 2990 m). In contrast, *parviglumis* prefers regions with medium to higher temperatures (17.1 to 28.3 °C), medium to high precipitation levels (1115 to 1431 mm), and moderate to medium altitude ranges of 143 to 1960 m (Figure 3b).

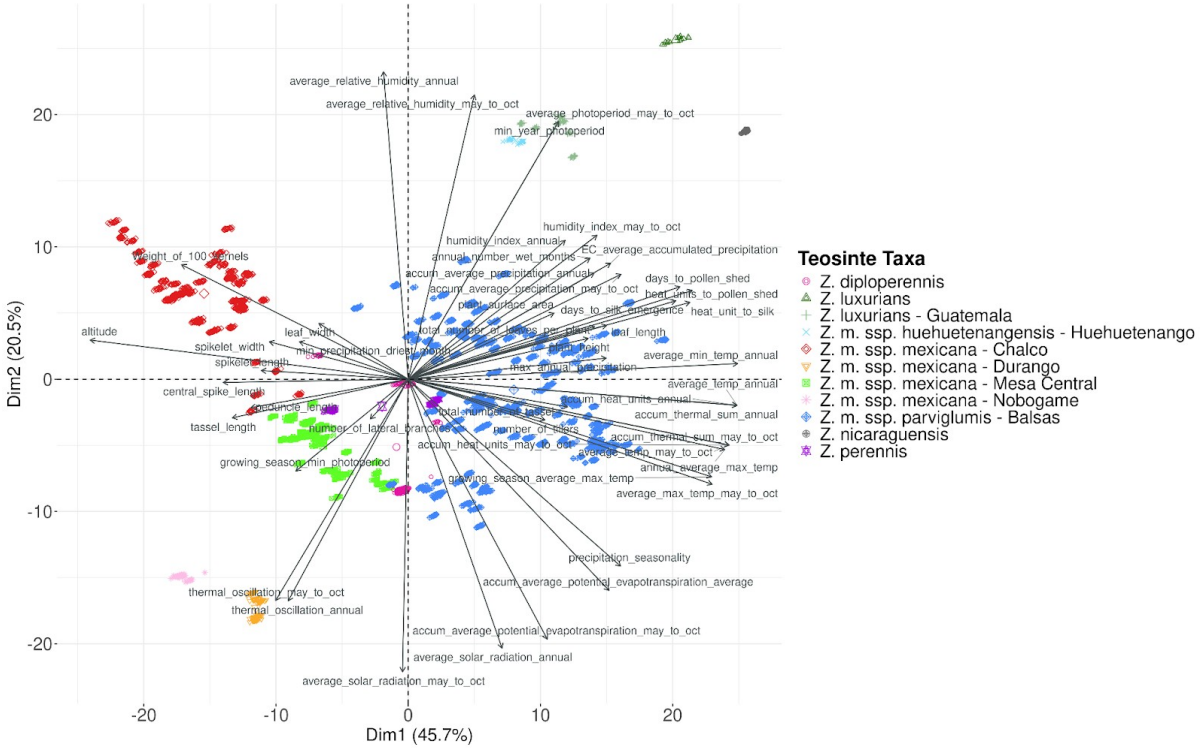
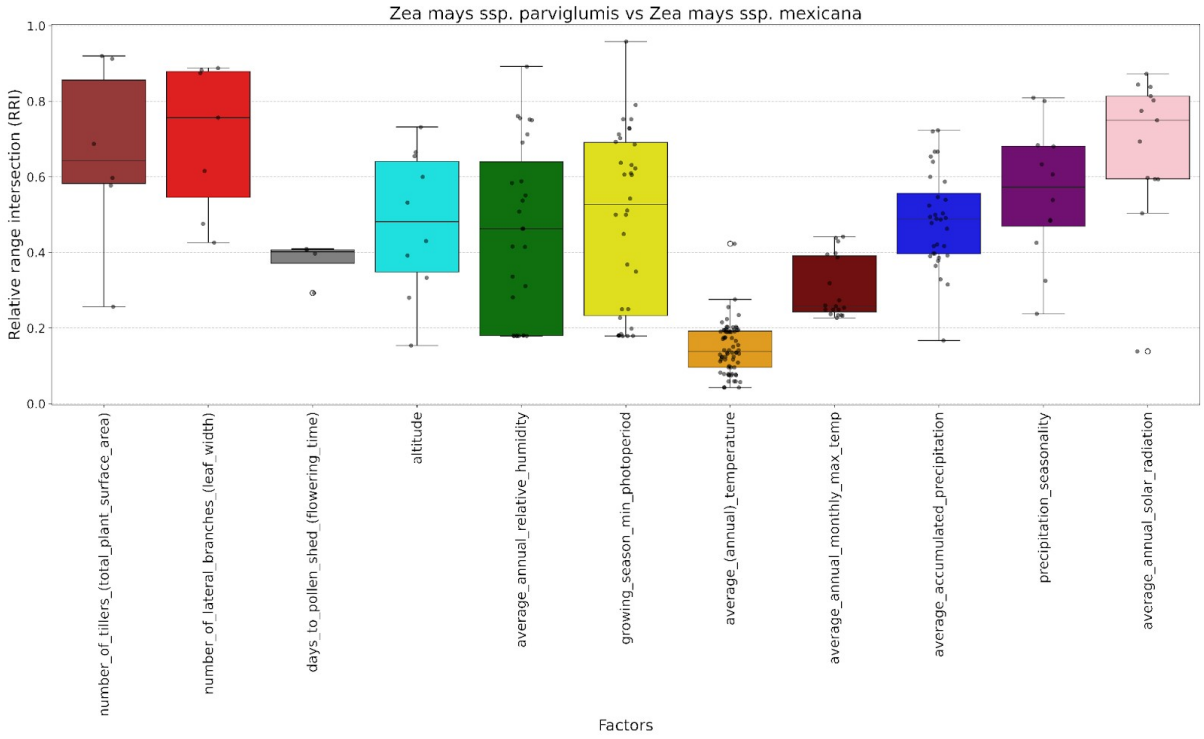


Figure 2. Principal Component Analysis (PCA) biplot of morphological and climatic data from 3,604 teosinte individuals. The first two principal components (PCs) explain 67.2% of the total variation. Each point represents an individual, colored according to its taxonomic group. The biplot reveals that teosinte taxonomy is largely aligned with distinct climatic growing conditions and phenotypic traits, although some overlap occurs, particularly between *mexicana* and *parviglumis*. Arrows represent key climatic and phenotypic variables: the direction indicates their influence on the first two PCs, while their length reflects the magnitude of this influence. Traits such as plant height, surface area, and flowering time, show strong correlations with climatic factors like humidity. Altitude is mainly associated with the weight of 100 kernels, while solar radiation impacts tillering, as observed in the formation of clonal side plants. This analysis highlights the relationship between climatic conditions and teosinte adaptation across different taxa.

a)



b)

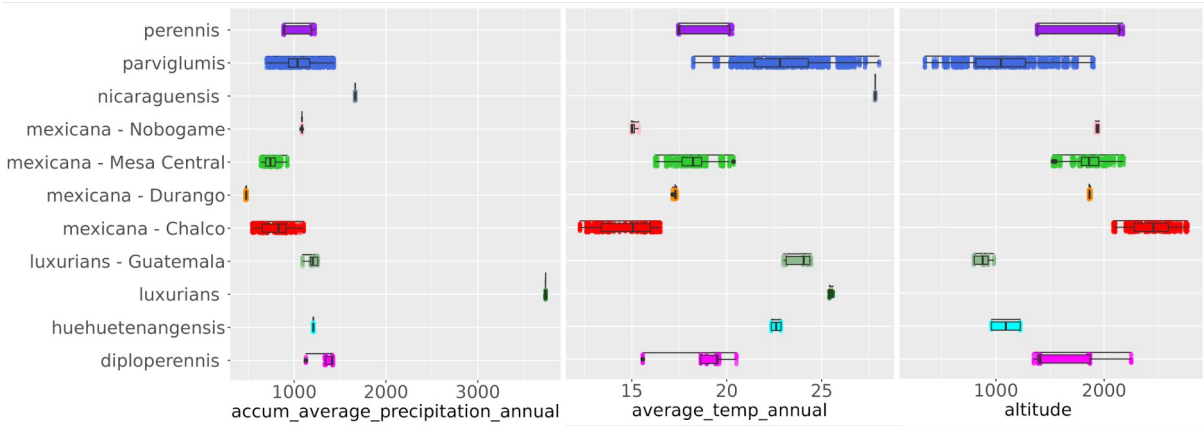


Figure 3. A. The figure illustrates the overlap of climatic conditions between teosinte subspecies *parviglumis* and *mexicana*, determined by Relative Range Intersection (RRI; see Methods) across 11 factor groups derived from the 255 environmental factors and morphological traits presented in Figure 1. The factor groups for temperature and flowering time exhibit distinct clustering patterns with minimal overlap. In contrast, the remaining factor groups demonstrate varying degrees of overlap, with RRI values ranging from 40% to 75%. This visualization effectively highlights the differences in the contributions of environmental and morphological factors between the two subspecies. **B.** Distribution plots illustrating the values of three crucial habitat climatic and geographic conditions for approximately four thousand teosinte individuals. The left y-axis categorizes the seven teosinte taxa, including four landraces or sub-populations of *mexicana*. The climatic and geographic factors shown are precipitation seasonality (left), average annual temperature (center), and altitude (right). The plot demonstrates how

teosinte taxa are adapted to specific climatic and geographic conditions. While taxa like the *Zea luxurians* and *Zea mays* ssp. *huehuetenangensis* are extreme specialists, others exhibit more opportunistic growth in varying conditions. Notably, there is an overlap in habitat conditions between *mexicana* and *parviglumis*, highlighting shared environmental adaptations.

Population Structure and Genetic Diversity

The key result of the teosinte genetic variation analysis is that the grouping of teosinte individuals is mirrored in teosinte taxonomic classification at the species level (i.e., not race level), as shown in Supplementary Figure S1. This confirms that the teosinte genetic population structure supports the current teosinte taxonomy. Furthermore, the genetic clusters revealed a high degree of genetic continuity within *parviglumis* and *mexicana*, suggesting complex population structure and high genetic diversity within these taxa. Interestingly, some overlap was observed between the genetic variation of *mexicana* and *parviglumis*. These findings are confirmed by the ADMIXTURE analysis (Figure 4, Supplementary Figure S2), which further demonstrates that teosinte taxonomy is underpinned by distinct genetic population structures for each taxon. Some taxa consist of several sub-populations, while others are more homogenous populations. The ADMIXTURE analysis identified 28 genetic clusters (k=28) as the best fit for the data, based on the lowest cross-validation error. Interestingly, some populations exhibiting admixture were predominantly located in close geographic neighborhoods (Figure 4), indicating localized gene flow.

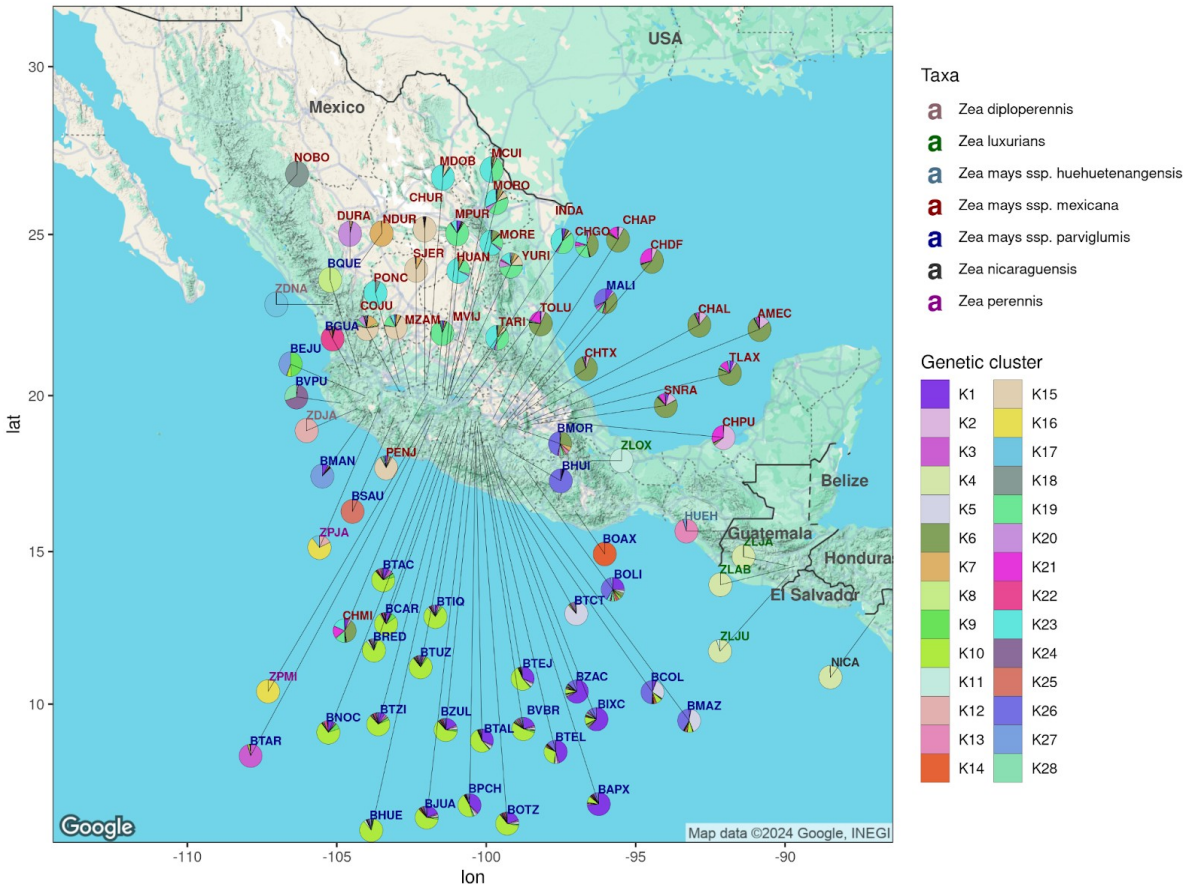


Figure 4. This figure illustrates the genetic structure and geographic distribution of teosinte populations across Mesoamerica. Each population is labeled with its four-letter abbreviation and is represented by a pie chart that visualizes the population's genetic structure. The pie charts show the average proportion of individuals' genomes assigned to 28 colored genetic clusters, which correspond to inferred ancestral populations (see legend). These clusters were identified through an ADMIXTURE analysis of approximately 3,604 teosinte individuals. Supplementary Figure 2 presents the individual admixture profiles for each individual. The geographic distribution closely reflects the genetic structure, with highly admixed populations concentrated in central México. These populations predominantly belong to *parviglumis* from regions such as Guerrero (BAPX, BCOL, BIXC, BMAZ, BOLI, BTCT, BTEL, BZAC), Michoacán (BCAR, BHUE, BJUA, BNOC, BRED, BTAC, BTAR, BTIQ, BTUZ, BTZI, CHGO, CHMI, CHUR, COJU, INDA, MORE, MORO, MPUR, MVIJ, PENJ, TARI), Jalisco (BEJU, BGUA, BJUA, BMAN, BVPU, PONC, SJER), and México (BOTZ, BPCH, BTAL, BTEJ, BVBR, MALI). Exceptions include BOAX (Oaxaca), BSAU (Jalisco), and BQUE (Jalisco). Similarly, *mexicana* is represented in populations from México (AMEC, CHAL, CHAP, CHTX, TOLU), Distrito Federal (CHDF), Michoacán (CHGO, CHMI, CHUR), Puebla (CHPU, SNRA, TLAX), Durango (DURA, NDUR), and Tlaxcala (TLAX), with exceptions being NOBO (Chihuahua) and NDUR (Durango). Other taxa include *Zea mays* ssp. *huehuetenangensis*, found in Huehuetenango (HUEH); *Zea diploperennis* from Jalisco (ZDJA) and Nayarit (ZDNA); *Zea luxurians*, distributed in Chiquimula and Jutiapa (ZLAB), Jalapa (ZLJA), and

Jutiapa (ZLJU); *Zea perennis* from Jalisco (ZPJA) and Michoacán (ZPMI); and *Zea nicaraguensis* from Chinandega (NICA).

The pairwise fixation index analysis revealed notable patterns of genetic differentiation among teosinte populations (Table 1). For instance, the comparison between *mexicana* and *parviglumis* yielded a low Fst value of 0.0647, suggesting a moderate level of genetic similarity, potential ongoing gene flow and shared adaptations to related environments. *mexicana* exhibited relatively low Fst values when compared to populations of other taxas, indicating lesser overall genetic differentiation. Its Fst values ranged from 0.2435 to 0.3858. In contrast, *Zea luxurians* and *Zea nicaraguensis* displayed higher Fst values overall, signifying greater genetic differentiation from other taxa, with values ranging from 0.3236 to 0.4879, while *Zea nicaraguensis* exhibited relatively high Fst values compared to other taxa, signifying substantial genetic differentiation within this population.

Table 1: Pairwise Fixation Index (Fst) Values Among Teosinte Taxa

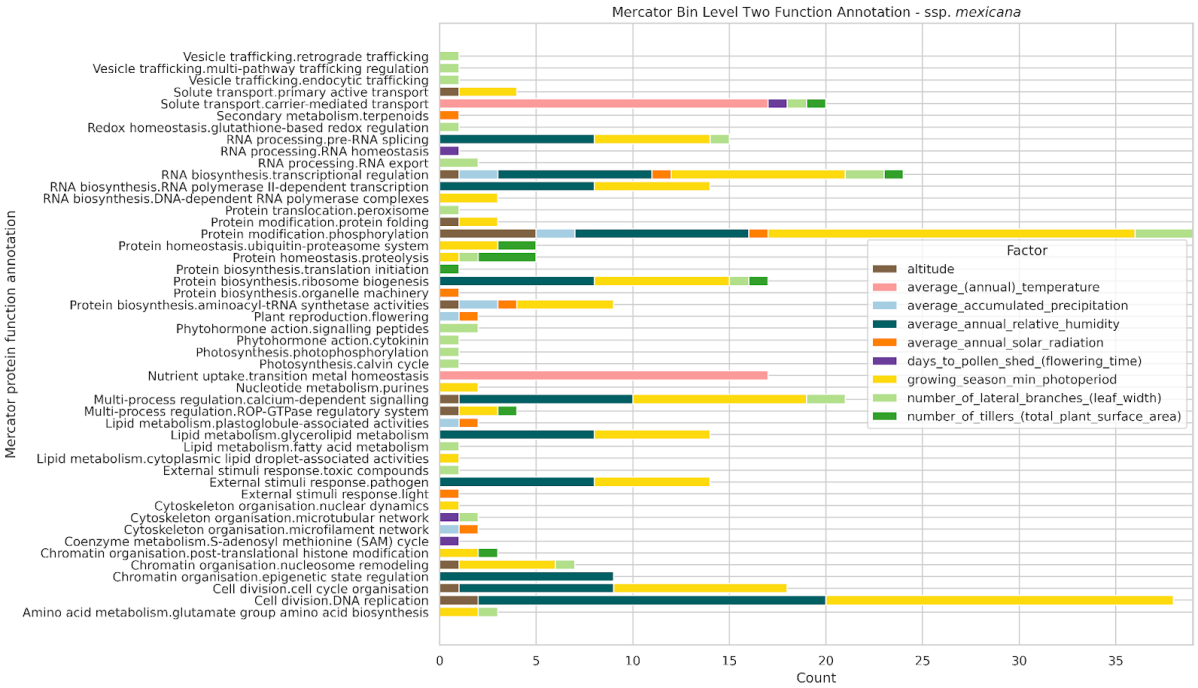
	parviglumis	mexicana	diploperennis	huehuetenangensis	luxurians	nicaraguensis
mexicana	0.0647					
diploperennis	0.2929	0.2951				
huehuetenangensis	0.2441	0.2435	0.4172			
luxurians	0.3236	0.3453	0.4023	0.4387		
nicaraguensis	0.3678	0.3858	0.4500	0.4879	0.1540	
perennis	0.2395	0.2454	0.1958	0.3879	0.3902	0.4419

Local adaptations within *mexicana* and *parviglumis*

In our investigation of local adaptations within teosinte taxa, we conducted GWAS to identify significant SNPs associated with various climatic factors and phenotypic traits. The standard quality control with QQ-plots of expected versus observed p-values revealed that the GWAS results for *mexicana* and *parviglumis* were the only ones with robust and trustworthy results. In contrast, the GWAS results for the other teosinte taxa were less reliable, likely due to smaller sample sizes. For *mexicana*, we identified 100 significant SNPs associated with teosinte habitat climatic factors and phenotypic traits. Upon mapping these SNPs to the B73_v5 reference genome, we identified 306 protein-coding genes located within a ±50 kb window of the candidate SNPs (Woodhouse et al., 2021). Functional annotation of these

genes using Mercator4 resulted in the assignment of 145 annotations of molecular functions in the form of Mapman Bins (Bolger et al., 2021; Schwacke et al., 2019). We found that 95 Mapman Bins at the hierarchical depth two were found to be enriched (Fisher's exact test, adjusted p-values < *0.003*), indicating potential associations with unique adaptations (Figure 5a). Similarly, for *parviglumis*, we identified 89 significant SNPs associated with climatic factors and phenotypic traits. Mapping these SNPs to the reference genome revealed 267 protein-coding genes within the vicinity of the candidate SNP markers. Functional annotation of these genes resulted in 125 Mercator protein functional annotations, with 85 of them found to be enriched (Fisher's exact test, adjusted p-values < *0.002*) at the hierarchical depth two (Figure 5b).

A)



B)

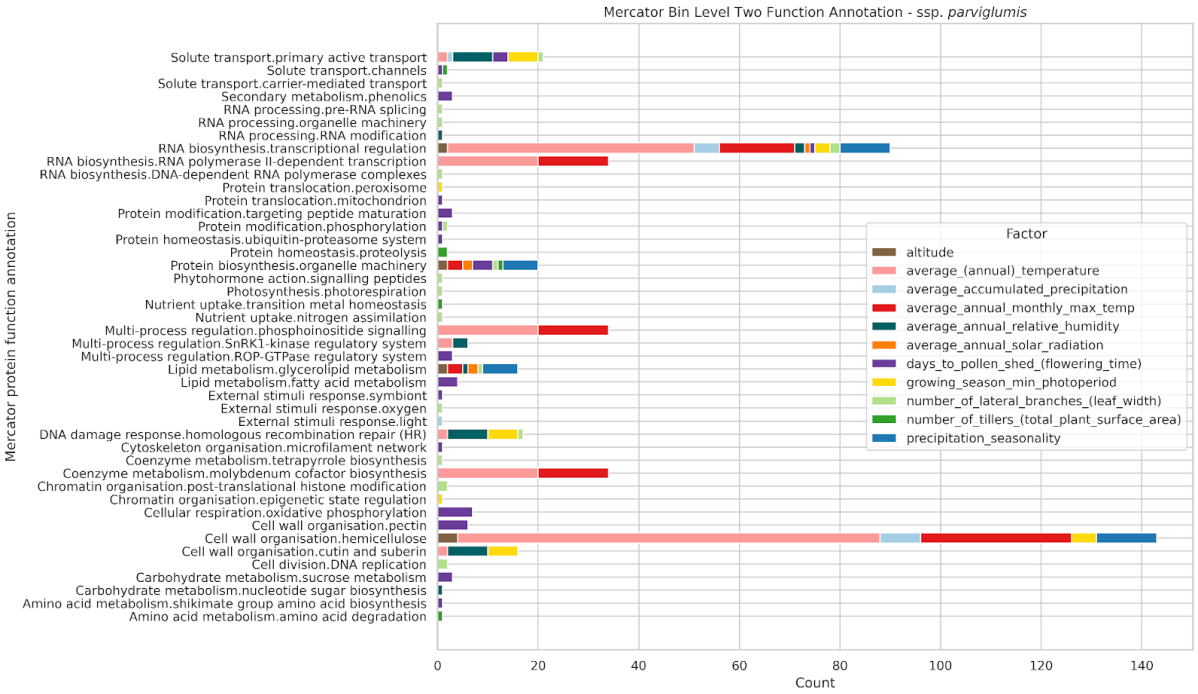


Figure 5. The figure illustrates the enriched molecular protein functions linked to protein-coding genes located within a ±50 kb proximity of GWAS loci for *mexicana* (Figure 5A) and *parviglumis* (Figure 5B). Molecular function annotations are displayed on the y-axis, while the count of protein functions related to groups of climatic and phenotypic factors is represented on the x-axis. The color coding indicates the specific group of climatic and phenotypic factors associated with each respective molecular function (see Figure 1). In Figure 5A for *mexicana*, temperature-related factors show significant associations with enriched functions in nutrient uptake, while precipitation and altitude-related factors are linked to enriched protein functions involved in cell division. Conversely, in Figure 5B for *parviglumis*, temperature-related factors are associated with protein functions related to glycerolipid metabolism.

Additionally, precipitation, temperature, and humidity-related factors are associated with proteins involved in light stimulus-response and cutin and suberin cell wall organization.

Distinct genetic and functional adaptations between *mexicana* and *parviglumis* in response to environmental factors

The comparison of *mexicana* and *parviglumis* using Jaccard Similarity Indices of SNPs associated with climatic factors and phenotypic traits revealed minimal overlap (see supplemental Figure S3a). Specifically, there were almost no shared significant SNPs between the two taxa. Further analysis of the associated loci revealed varying degrees of molecular function similarity between *mexicana* and *parviglumis*, depending on the climatic factors considered. The highest Jaccard Similarity index (25%) was observed for factors highly correlated with average accumulated precipitation (Figure 1), suggesting potential shared adaptations to these climatic habitat factors. However, no similarity in enriched molecular functions was observed for precipitation seasonality, days to pollen shed (flowering time), and average annual temperature factors (supplemental Figure S3b). These results imply distinct genetic and functional adaptations of *mexicana* and *parviglumis* populations to their respective environments.

We investigated the protein functions associated with the morphological traits and environmental factors of *parviglumis* and *mexicana*, identifying several proteins potentially involved in adaptation and environmental responses. In *parviglumis*, key findings indicate that average annual relative humidity is associated with damage response pathways (homologous recombination repair), solute transport (primary active transport), and cell wall organization (cutin and suberin). Additionally, carbohydrate metabolism, particularly nucleotide sugar biosynthesis, and multi-process regulation via the SnRK1-kinase regulatory system were also prominent. Altitude was associated with candidate genes with annotations related to protein biosynthesis (organelle machinery), lipid metabolism (glycerolipid metabolism), cell wall organization (hemicellulose), and reactive oxygen-induced signaling in redox homeostasis. The growing season's minimum photoperiod in *parviglumis* was linked to candidate genes with functional annotations related to damage response, chromatin organization (epigenetic state regulation), and coenzyme metabolism (chlorophyll metabolism). The analysis of average annual solar radiation and precipitation seasonality identified associations with lipid metabolism, specifically glycerolipid metabolism. Average accumulated precipitation influenced solute transport (primary active transport) and cell wall organization, particularly hemicellulose, while also affecting chromatin organization and lipid metabolism. Days to pollen shed (flowering time) showed links to solute transport and protein biosynthesis, alongside various metabolic processes including carbohydrate metabolism and cellular respiration.

Morphological traits, such as the number of lateral branches and tillers, were associated with nutrient uptake (nitrogen assimilation), protein biosynthesis, and cell division.

In *mexicana*, functional annotations reveal distinct responses to environmental variables and morphological traits. Significant findings indicate that average annual relative humidity is associated with chromatin organization (nucleosome remodeling), pathogen response, pre-R splicing, and multi-process regulation (calcium-dependent signaling). Altitude influences pathways related to solute transport (primary active transport), protein biosynthesis (aminoacyl-tRNA synthetase activities), and cell division. The growing season's minimum photoperiod affects chromatin organization (post-translational histone modification), lipid metabolism, and multi-process regulation, including R processing and nuclear dynamics. Morphological traits, such as the number of lateral branches and tillers, were linked to protein biosynthesis (ribosome biogenesis) and vesicle trafficking. Furthermore, annotations related to average annual solar radiation, precipitation seasonality, and average accumulated precipitation, highlight pathways involved in flowering and secondary metabolism. Other notable aspects include solute transport and nutrient uptake related to transition metal homeostasis, particularly for temperature variables (Supplementary Table S2; Supplementary Table S3).

Figure 6 provides insights into the population structure, habitat climate, genetic diversity, and their associations in *mexicana*. The environmental factors play a pivotal role in shaping the genetic landscape, with distinct races clustering into discrete clades based on their adaptive responses to various environmental conditions. In other words, the grouping of *mexicana* individuals according to habitat climatic growing conditions is mirrored in population genetics and GWAS results, effectively positioning the *mexicana* races into distinct clades. These corresponding results are indicative of niche specialization among populations and races. Additionally, and not in contradiction to the former results, the presence of admixing populations within *mexicana* provides compelling evidence of high gene flow and a complex population structure (Figure 6, third panel). In *mexicana,* no population appears to dominate a certain range of habitat climatic conditions, rather different populations grow in similar climatic conditions (Figure 6, second panel). This suggests distinct adaptation strategies for different *mexicana* populations to similar habitat climatic conditions. The Nobogame and Durango races exhibit a unique genetic profile that sets them apart from other *mexicana* populations. This genetic distinctiveness aligns with the geographic isolation of their respective ranges. Prior studies have observed this pattern, reporting elevated pairwise Fst values for Nobogame and Durango, indicative of higher genetic differentiation relative to other *mexicana* races (Rivera-Rodríguez et al., 2023). Supporting this observation, the allele states plot for significant GWAS SNP markers shows a high degree of homozygosity within these

races (Figure 6, bottom panel). These findings highlight the dynamic interplay between genetic drift, correlated with geographic isolation, population structure, and environmental adaptation within *mexicana*, shedding light on the intricate evolutionary processes shaping this subspecies' genetic diversity.

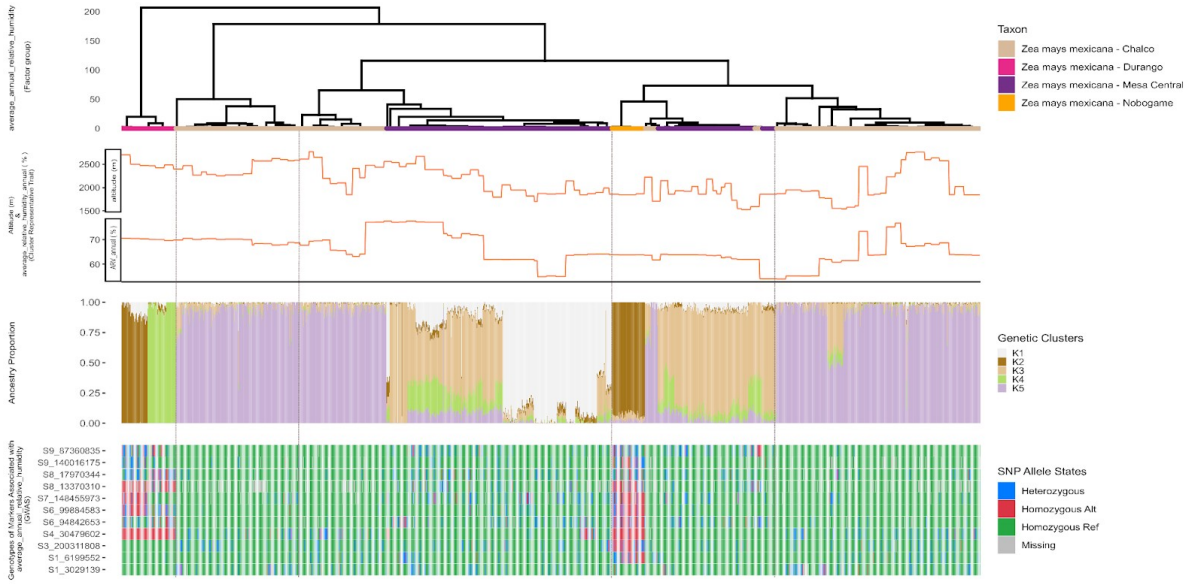


Figure 6. Integrated visualization of habitat climatic growing conditions (top two panels), genetic structure, and allele types of GWAS markers (bottom panel). The respective values of *mexicana* individuals are sorted by individuals (x-axis). In detail, the top panel shows the dendrogram generated by correlation-based clustering of habitat climatic growing conditions of the group of "average annual humidity" for all *mexicana* individuals (see figure 1 for details grouping). The second panel focuses on *mexicana* individuals' altitude (top curve) and average relative humidity (bottom curve) values assessed for each respective individual. The third panel shows the genetic makeup as produced by our ADMIXTURE analysis for each individual and five assumed population clusters. Finally, the fourth and bottom panel displays the genotype of *mexicana* individuals at GWAS SNP marker results (fourth bottom panel). The top panel shows the result of Euclidean distance-based hierarchical clustering on average annual relative humidity measured at the collection sites of spp. *mexicana* individuals. The dendrogram leaves are colored according to their taxonomy. The second panel shows the corresponding collection site values of the habitat climatic factors altitude and annual mean monthly maximum temperature, for comparison. Panel three shows the ADMIXTURE analysis results, that is the genetic markup of each respective individual corresponding to the other panels. Finally, the fourth bottom panel shows the genotypes each *mexicana* individual has at eleven GWAS markers that are significantly associated with habitat climatic factors from the "average annual humidity" group (see top panel and figure 1). Allele types are colored: green for the reference allele, red alternate allele, blue for heterozygous individuals and grey for missing data. The left column of this panel shows the identifiers of the eleven GWAS SNP marker.

Comparative analysis of Teosinte GWAS SNPs with maize literature GWAS SNPs

In our GWAS analyses, we identified SNP markers significantly associated with the habitat climatic conditions and phenotypic traits of the teosinte subspecies *mexicana* and *parviglumis*. We then compared these findings with SNP markers linked to maize agronomic traits, spanning 58 traits from the literature. This comparison assessed the relatedness of teosinte climatic factors, teosinte phenotypic traits, and maize phenotypic traits (hereafter referred to as "factors") based on shared SNP markers within a ±50 kb window (Table S4).

For *mexicana*, 47 GWAS SNPs matched those found in maize, while 32 SNPs were novel. Similarities were observed in traits such as plant height, grain weight, disease resistance, and various biochemical compositions. Similarly, *parviglumis* had 39 SNPs that matched maize markers, with 29 being novel. The overlap included 53 maize traits, particularly those related to grain weight, root development, biochemical composition, and disease resistance. Differences were observed between the two subspecies, with *mexicana* showing SNP associations with 30 factors compared to 25 factors associated with *parviglumis*. However, commonalities were also identified, with 29 traits sharing SNP markers between both subspecies, including key agronomic and physiological features like grain weight, plant height, disease resistance, and root development.

To further explore these relationships, we calculated pairwise distances between all analyzed climatic and phenotypic factors based on shared and non-shared SNP markers (allowing a ±50 kb window) and performed hierarchical clustering. The resulting dendrogram (Figure 7) was compared with the clustering of teosinte factors (Figure 1).

Near the root of the dendrogram (Figure 7A), we found that teosinte climatic conditions such as growing season length, growing season evapotranspiration, winter precipitation, and latitude share SNP markers with maize traits related to yield (e.g., cob weight, 100-grain weight), photosynthetic efficiency (e.g., leaf number, plant height), and flowering time (e.g., days to silking).

Further down the dendrogram, another cluster (Figure 7B) revealed that teosinte climatic traits, such as relative humidity and photoperiod are genetically linked with maize traits, including salt tolerance, mercury accumulation, and leaf orientation. A subsequent cluster (Figure 7D) grouped teosinte growing conditions, including average and maximum temperatures, with maize root traits (Figure 7C), such as lignin content, root surface area, and stem diameter. Interestingly, this cluster also included teosinte traits like leaf dimensions, plant surface area, and tiller number, suggesting that climatic conditions related to geographic longitude are associated with root, plant surface, and yield traits.

Finally, deeper in the dendrogram (Figure 7E), we found that maize traits such as root fork number, chlorophyll content, and grain oil content shared SNP markers with teosinte climatic conditions, including winter evapotranspiration and minimum temperatures.

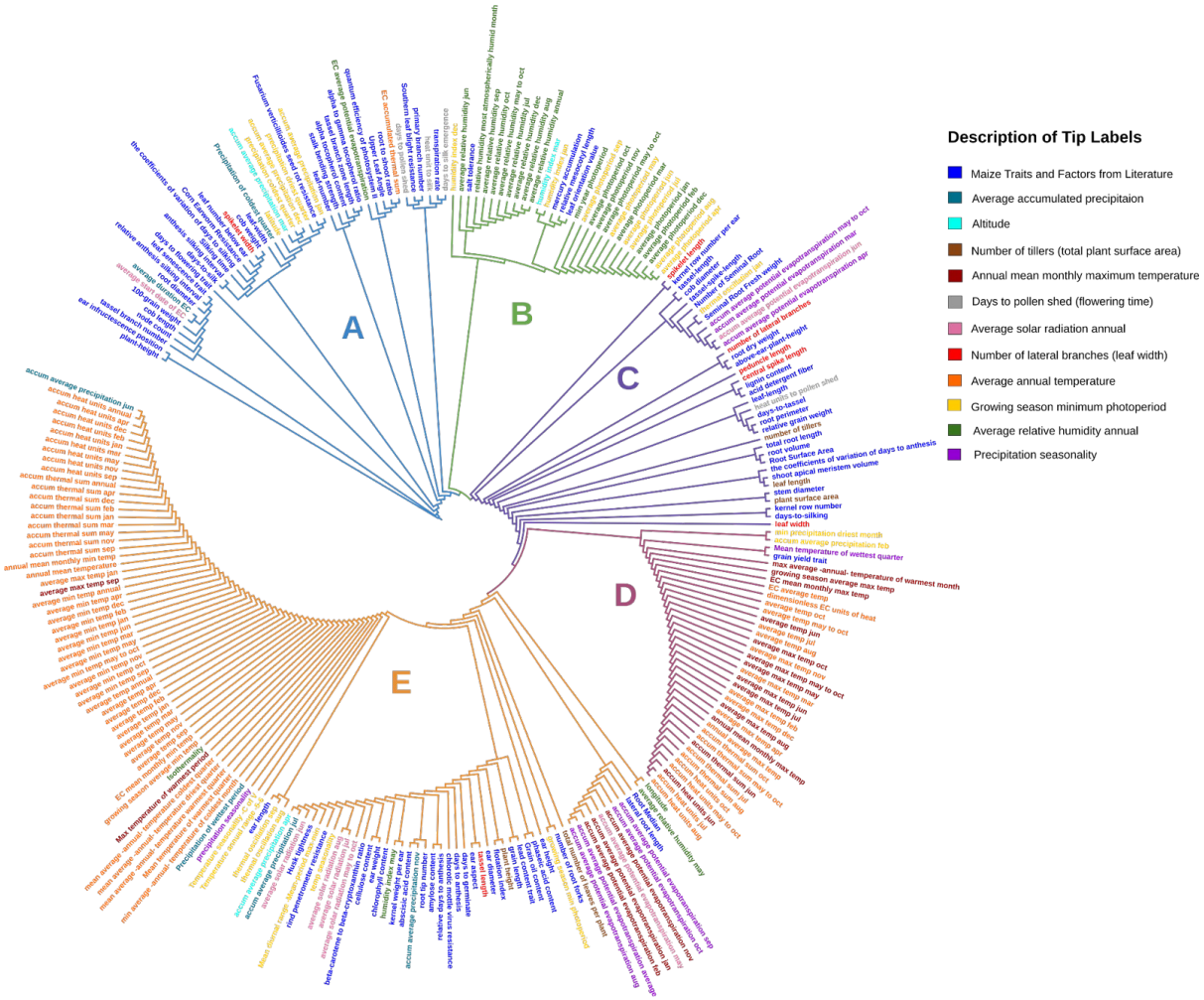


Figure 7. The figure shows hierarchical clustering of maize phenotypic traits such as plant height, disease resistance, and root development with teosinte environmental factors and morphological traits, based on shared SNP markers within a ±50 kb distance. Pairwise distances between climatic and phenotypic factors reveal distinct clusters, highlighting genetic links between maize traits and the environmental conditions in which teosinte grows. (A) Near the base of the dendrogram, SNP markers shared between maize yield traits (e.g., cob weight, 100-grain weight), photosynthetic efficiency (e.g., leaf number, plant height), and flowering time (e.g., days to silking) are genetically associated with teosinte climatic factors such as growing season length, evapotranspiration, winter precipitation, and latitude. (B) Another cluster shows genetic connections between maize traits related to stress tolerance (e.g., salt tolerance, mercury accumulation), leaf orientation, and teosinte climatic factors such as relative humidity and photoperiod. (C) A further cluster groups teosinte climatic conditions, such as average and maximum temperatures, with maize root traits like lignin content, root surface area, and stem diameter. This cluster also includes teosinte traits like leaf dimensions, plant surface area, and tiller number, suggesting a genetic link to geographic longitude. (D) This cluster reveals relationships

between maize root traits and climatic factors like temperature extremes and geographic features, indicating environmental influences on plant architecture. (E) In the deepest part of the dendrogram, maize traits such as root fork number, chlorophyll content, and grain oil content are genetically linked to teosinte environmental conditions like winter evapotranspiration and minimum temperatures.

Discussion

We performed an extensive population genetics and genome-wide association study (GWAS) on teosinte species to explore their genetic diversity, population structure, and adaptation to climatic conditions, as well as their relationship with phenotypic traits in both teosinte and maize. The analysis provided valuable insights into the potential adaptive mechanisms and environmental responses of teosinte populations. To our knowledge, GWAS has not previously been employed in the study of wild populations within a conservation context. Our findings highlight the utility of GWAS in uncovering genetic adaptations to environmental conditions and comparing these adaptations across different populations. This approach underscores the potential of GWAS as a powerful tool for analyzing genetic variation within ecological research. A key finding is that teosinte populations have developed divergent adaptive strategies despite experiencing similar environmental conditions, supporting the importance of the PM indicator of the GBF for monitoring the conservation of genetic diversity.

Adaptive genetic diversity is differentially distributed across species, populations, and environments.

Our results show that the teosinte taxonomy largely reflects genetic population structure and habitat climatic conditions showing a strong agreement between those measures. This conclusion is supported by both principal component analysis, which identifies the key axes of variation in the teosinte genetic data, and admixture analysis, which provides independent validation of these patterns.

Teosinte’s adaptation to diverse climatic environments, spanning multiple taxa, is reflected in high levels of genetic variation across all sampled populations, including all known teosinte populations in Mexico. Phenotypic traits also vary significantly among taxa and these can be categorized into traits related to total plant surface, reproductive organ dimensions, flowering time, and weight of 100 kernels. Notably, the weight of 100 kernels, while correlated with yield, is distinct from it, as yield is also influenced by the number of ears per plant. In this, the total plant surface is correlated with humidity traits (see Figures 2 and 3).

Our findings indicate that altitude plays a crucial role in influencing teosinte yield, suggesting that elevation impacts the species' adaptation strategies to different ecological

niches. Additionally, humidity and rainfall were also identified as key factors shaping teosinte yield, while temperature showed a significant correlation with flowering time, emphasizing the importance of temperature adaptation mechanisms in teosinte's reproductive cycle.

Genome-wide association studies revealed specific candidate genes associated with adaptive traits in *parviglumis*, highlighting pathways likely supporting survival in warmer, moderately wet regions. Notably, genes related to cutin and suberin biosynthesis suggest mechanisms for mitigating heat stress, while genes regulating flowering time indicate localized adaptations to regional growing seasons. Membrane fluidity, regulated by glycerolipid synthesis, suggests resilience to temperature variability by maintaining cellular integrity. Associations between relative humidity and genes involved in homologous recombination repair and solute transport suggest that *parviglumis* has evolved robust responses to moisture fluctuations, with primary active transport genes enhancing nutrient and water uptake essential in consistently humid environments. Furthermore, photoperiod-related genes involved in chromatin organization and damage response indicate *parviglumis*' ability to fine-tune developmental processes to changing daylight conditions, suggesting an epigenetically regulated adaptation to photoperiod shifts. Morphological traits, including increased lateral branches and tillers, correlate with nutrient uptake and protein biosynthesis pathways, underscoring a strategy geared toward resource acquisition and enhanced competitive growth in ecologically diverse habitats.

For *mexicana* races (Chalco, Durango, Mesa Central, and Nobogame), which spans colder, high-altitude environments, associations with relative humidity highlight genes involved in chromatin remodeling, pathogen response, and calcium-dependent signaling, suggesting strategies for rapid adaptation to fluctuating humidity and pathogen pressures. Genes involved in pre-RNA splicing and multi-process regulation support fine-tuned gene expression in response to environmental changes, potentially enhancing resilience in variable conditions. High-altitude adaptations include candidate genes for primary active transport and aminoacyl-tRNA synthetase activities, indicating enhanced nutrient uptake and protein synthesis critical for growth under nutrient-limited conditions. Notably, introgression from *mexicana* into modern maize has been linked to increased protein content (L. Wang et al., 2008), highlighting the potential of these adaptations to contribute valuable traits to cultivated maize. Photoperiod-sensitive pathways (Durango and Nobogame are the highest north latitude populations) affecting chromatin organization and lipid metabolism highlight mechanisms that may regulate flowering and biomass allocation in response to light exposure, with histone modifications likely influencing gene expression critical for developmental processes. Morphological traits such as increased branching and tiller numbers, align with protein biosynthesis and vesicle trafficking pathways, suggesting that growth and developmental adjustments in *mexicana* are

both environmentally responsive and genetically modulated. Finally, environmental factors, including solar radiation, precipitation seasonality, and total precipitation, were associated with flowering time and secondary metabolism pathways, underscoring how seasonal and climatic signals integrate with physiological processes for reproduction and resource management. Transition metal homeostasis, particularly under temperature influences, reflects the specialized adaptations of *mexicana* to its high-altitude, variable environment, revealing a sophisticated balance between environmental adaptation and physiological robustness.

However, because the samples partially are from the early 1970s, we cannot be sure that the measured genetic structure reflects the current status. Additionally, phenotypic traits were measured under controlled greenhouse conditions in central Mexico. While this approach ensured consistency in environmental factors influencing growth and development, it may not fully capture the variation in traits that would be expressed under natural field conditions. This is particularly relevant for taxa growing at different altitudes, where environmental pressures such as water availability, temperature fluctuations, and soil conditions may significantly influence phenotypic traits. Consequently, the results observed under greenhouse conditions should be interpreted with caution, as they might differ from those in more variable and stress-prone field environments. Future studies could build on these findings by evaluating phenotypic traits under field conditions, to provide a more comprehensive understanding of their ecological and adaptive significance, as greater genetic and phenotypic diversity likely exists *in situ*. Interestingly, some taxa and populations share similar climatic habitats but exhibit distinct genetic makeups (admixture) and different SNP patterns.

Additionally, the results of our identification of protein functions being linked to teosinte habitat climatic conditions and phenotypic traits have to be considered with care and require experimental verification. This is because several factors in the analysis can provide sources of bias. One key factor is the identification of genetic loci associated with either climatic factors or phenotypic traits. In this, quality control of SNPs plays an important role which discards loci showing a high degree of correlation within a given linkage disequilibrium window. Thus, not all SNPs are considered in the GWAS calculations, as they introduce a bias regarding the correct localization of significant associations. Additionally, protein-coding genes are identified within a ±50 kb vicinity of GWAS markers, further contributing to the uncertainty of genetic localization. Nonetheless, the important key result here is that different teosinte taxa and populations show distinct genetic variation associated with teosinte habitat climatic conditions and phenotypes, and this associated variation appears to be at least partially adaptive as we do find distinct protein functions in the vicinity of the GWAS markers.

All populations matter: implications for conservation and management.

Our findings underscore the importance of conserving wild teosinte populations for several reasons. First, the genetic structure of teosinte populations is unique to each subspecies and corresponds closely with teosinte taxonomy. The significant genetic distinctiveness of teosinte species, including their major populations, is independently confirmed by our Admixture and principal component analyses (PCA). Furthermore, the PCA of teosinte habitat climatic conditions and phenotypic traits supports teosinte taxonomy and correlates with population structure, suggesting that the distribution of teosinte has driven substantial genetic diversification. Second, our GWAS reveals that even when *parviglumis* and *mexicana* populations share similar environmental conditions, their evolutionary trajectories have resulted in distinct adaptations to these climates. Third, this is corroborated by our protein function analysis of genetic markers, which shows that different molecular functions are associated with these adaptations. Finally, our comparison of published maize genetic markers associated with key agronomic traits indicates that many desirable crop characteristics are influenced by crucial environmental factors in teosintes. These findings underscore the evolutionary and agronomic significance of teosinte and emphasize the need to preserve its genetic diversity for future crop improvement and ecological resilience.

Consequently, each teosinte population harbors unique genetic diversity that cannot be found elsewhere. This underscores the critical importance of the “Populations Maintained” indicator within the GBF. The loss of any population would erase its unique evolutionary history and local adaptations, thereby narrowing the genetic diversity available to the species for adapting to environmental changes. (Hoban et al., 2024; Mastretta-Yanes et al., 2024). Practically, this also has direct implications for *in situ* and *ex situ* conservation strategies, namely: conservation activities and monitoring must be done at the population level, not the species.

As demonstrated by our results, unique environmental conditions and adaptive evolution shape the distinct genetic structures of teosinte populations. Therefore, conservation efforts should prioritize *in situ* preservation to maintain this adaptive genetic diversity. In other words, preserving all teosinte populations in their natural habitats is essential to safeguard this genetic diversity. While *ex situ* conservation can be valuable for safeguarding genetic material against large-scale catastrophes (e.g., a hurricane destroying an entire population). However, re-introductions to the wild should consider our findings regarding population structure and the presence of distinct adaptive variation, despite similar environmental conditions. Additionally, our results show that while *ex situ* conservation can serve as a complementary approach, truly

maintaining genetic diversity requires recognizing it as the outcome of a continuous evolutionary process.

Importantly, maize-teosinte hybridization has occurred for thousands of years in present-day Mexico. Initially, this hybridization likely occurred at relatively low rates (Ellstrand et al., 2007) due to genetic or phenological incompatibilities. However, the expansion of the agricultural frontier over the last century has significantly increased the proportion of teosinte populations now in contact with maize, including both considering local landraces under traditional management and high-yielding varieties grown under intensive conditions (Rojas-Barrera et al., 2019). This generates complex dynamics for gene flow and selection pressures, with important implications for teosintes' genetic diversity conservation and maize agriculture. On one hand, the much larger population size of the crops compared to their wild counterparts can lead to gene flow, causing genetic assimilation, wherein crop genes replace wild ones, and demographic swamping, wherein hybrids (crop-wild) are less fertile than their wild parents, ultimately causing wild populations to shrink (Haygood et al 2003). This is of particular concern in geographic areas with relatively small teosinte populations, where intensive agriculture has expanded in recent decades (e.g., Jalisco). On the other hand, teosintes can acquire resistance genes against pesticides, potentially leading to the emergence of superweeds, as has already been observed in Spain (EFSA et al., 2022). In this context, understanding and mapping the distribution of genetic diversity between teosinte populations can help predict which environmental conditions could be invaded by weedy teosintes, especially in the absence of proper management practices and in the face of climate change.

Conclusion

In conclusion, our study underscores the intricate relationship between teosinte's genetic population structure and its habitat environmental conditions, which drives the development of adaptive variation across different populations. This highlights the importance of *in situ* conservation, recognizing teosinte as a vital source of genetic diversity, not only for conservation purposes but also in the context of new weeds emergence. Understanding crops and their wild relatives as part of a complex wild-domesticated continuum is crucial, especially in centers of origin. Exploring how teosinte diversity can be used to develop sustainable agricultural practices is beyond the scope of this article, but such efforts must start with recognizing and safeguarding genetic diversity in countries that are centers of crop origin and domestication.

Acknowledgment

This research was supported by funding from the German Federal Ministry of Education and Research (BMBF) under project number 031B0921. This research is part of the Programa de Monitoreo de Teocintles undertaken by the Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO).

Supplemental Information

Table S1. Pairwise correlation matrix of teosinte individuals for habitat, climate, and phenotype variables. This table presents the correlation coefficients between pairs of variables, with values ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation), and 0 indicating no correlation. Correlations closer to 1 or -1 suggest a stronger linear relationship between variables. The table includes variables related to environmental factors (e.g., temperature, precipitation, altitude) and phenotypic traits (e.g., plant height, number of tillers weight of 100 kernels) observed across teosinte populations providing insights into the interplay between environmental conditions and phenotypic expression in teosinte individuals.
TableS1_pairwise_correlation_habitat_climate_phenotypes
https://docs.google.com/spreadsheets/d/
1X4Ai8CosqS6L9EWnKPDxQlwDnXEiKdxbQ9GAO3GsxqM/edit?usp=sharing

Table S2. The table presents functional categories linked to traits and environmental factors across various teosinte taxa, revealing key biological processes and pathways involved in their adaptation and development. For example, for the trait number of lateral branches (leaf width), *parviglumis* is highly associated with processes such as photosynthesis (e.g., Calvin cycle, photophosphorylation) and protein modification (e.g., phosphorylation), alongside pathways for nitrogen assimilation and cell division. Meanwhile, *mexicana* (all races) shows significant associations with chromatin organization, protein homeostasis (e.g., proteolysis, ribosome biogenesis), and responses to external stimuli like toxic compounds. For the environmental factor of precipitation seasonality, *parviglumis* is associated with processes like lipid metabolism (e.g., glycerolipid metabolism) and transcriptional regulation, while *mexicana* (all races) displays functional categories related to chromatin remodeling and vesicle trafficking. In response to growing season minimum photoperiod, *parviglumis* activates damage response pathways (e.g., homologous recombination repair) and solute transport, while *mexicana* Mesa Central demonstrates enrichment in pathways like protein homeostasis (e.g., ubiquitin-proteasome system) and photosynthesis.

TableS2_unique function annotaions per factor assocaited with respective the teosinte taxa
https://docs.google.com/spreadsheets/d/16kj4lR6PUTrQOS7-xu_NvxJBJ_6OPvrDnGnJrWbytLY/edit?usp=sharing

Table S3. The table provides detailed functional annotations for SNPs linked to specific traits and environmental factors in teosinte populations. The columns include SNP identifiers, distances to the nearest gene, minor allele frequencies (MAF), associated protein functions, and pathways. It also contains a comprehensive list of climate and morphological variables used in the study, along with their units and detailed definitions.

TableS3_teosinte_gwas_snps_and_annotations
https://docs.google.com/spreadsheets/u/0/d/15a53qAVd7rptsML9sUbEVAShjpZREGMVrEu9OxMIJ5o/edit

Table S4. Comparison of SNP markers significantly associated with habitat climatic conditions and phenotypic traits of *mexicana* and *parviglumis*, with maize agronomic traits. This table presents the shared SNP markers between teosinte climatic factors, phenotypic traits, and 58 maize agronomic traits from the literature. The SNPs were identified within a ±50 kb distance and highlight the genetic overlap between environmental adaptation in teosinte and maize crop traits.

TableS4_snps_comparison_teosinte_maize_traits_50kb
(https://docs.google.com/spreadsheets/d/1ihaFLl3CAN4l45zzAZXd_55bGt8hhr7GmpVpl--YVAk/edit?usp=sharing)

Figure S1

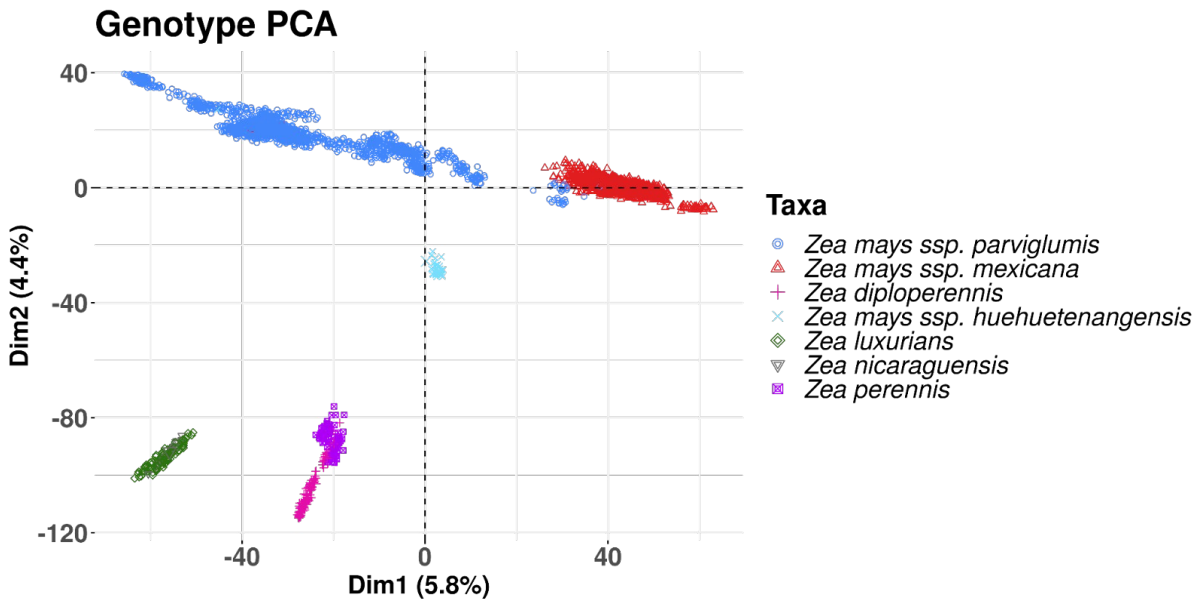


Figure S1. PCA plot illustrating genetic relationships among teosinte taxa based on genotype data, with the first two principal components (PC1 and PC2) explaining 10.2% of the total genetic variation. Individuals are color-coded by taxonomic classification. *Parviglumis* (blue circles) and *mexicana* (red triangles) show a broad spread along both axes, with *parviglumis* extending further along PC2. *Zea diploperennis* (pink crosses) and *Zea perennis* (purple squares) are positioned in the lower-left quadrant. *Zea luxurians* (green diamonds) cluster tightly in the negative PC2 space, alongside *Zea nicaraguensis* (grey inverted triangles). *Zea mays* ssp. *huehuetenangensis* (cyan 'x') clusters distinctly in the center of the plot. These clustering patterns reflect the genetic variation distribution across the different teosinte taxa, highlighting their distinct genetic backgrounds.

Figure S2

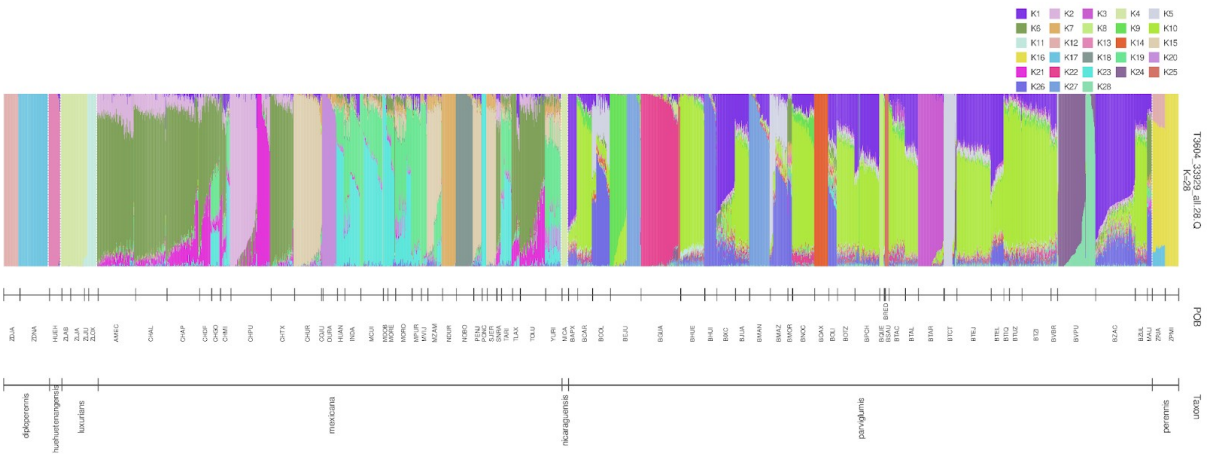
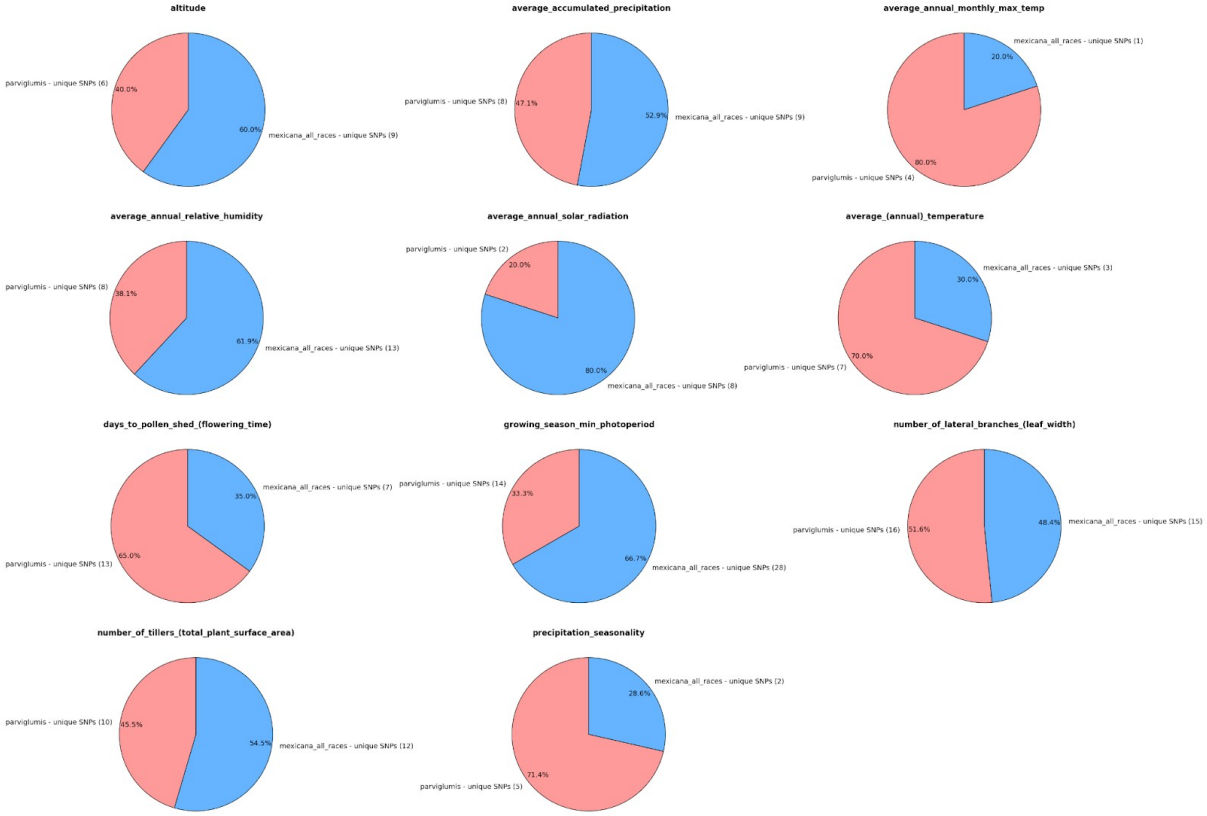


Figure S2. Admixture analysis of 3,604 teosinte individuals reveals 28 distinct genetic clusters (*K*1–*K*28), each represented by a unique color. Each thin bar along the x-axis corresponds to an individual, with the colors within each bar indicating the proportion of the genome derived from each cluster. Based on cross-validation, the number of clusters (K=28) was determined as the best fit to explain population structure. *Parviglumis* and *mexicana* exhibit the greatest genetic diversity and population sizes, with multiple subpopulations identified within each taxon. Individuals are grouped by population and sorted by taxonomic classification, offering a comprehensive visualization of genetic structure and admixture. Notable admixture patterns within *parviglumis* and *mexicana* reveal substantial genetic diversity and shared ancestry both within and across populations. Other taxa, including *Zea mays* ssp. *huehuetenangensis*, *Zea diploperennis*, *Zea luxurians*, and *Zea perennis*, display distinct genetic signatures corresponding to their geographic distributions, yet also show varying degrees of admixture (see Figure 4). This analysis underscores the complex genetic structure and admixture patterns that define the wild relatives of maize across diverse ecological regions.

Figure S3

a)


altitude
parviglumis - unique SNPs (6)
40.0%
60.0%
mexicana_all_races - unique SNPs (9)
average_accumulated_precipitation
parviglumis - unique SNPs (8)
47.1%
52.9%
mexicana_all_races - unique SNPs (9)
average_annual_monthly_max_temp
mexicana_all_races - unique SNPs (1)
20.0%
80.0%
parviglumis - unique SNPs (4)
average_annual_relative_humidity
parviglumis - unique SNPs (8)
38.1%
61.9%
mexicana_all_races - unique SNPs (13)
average_annual_solar_radiation
parviglumis - unique SNPs (2)
20.0%
80.0%
mexicana_all_races - unique SNPs (8)
average_(annual)_temperature
mexicana_all_races - unique SNPs (3)
30.0%
70.0%
parviglumis - unique SNPs (7)
days_to_pollen_shed_(flowering_time)
mexicana_all_races - unique SNPs (7)
35.0%
65.0%
parviglumis - unique SNPs (13)
growing_season_min_photoperiod
parviglumis - unique SNPs (14)
33.3%
66.7%
mexicana_all_races - unique SNPs (28)
number_of_lateral_branches_(leaf_width)
parviglumis - unique SNPs (16)
51.6%
48.4%
mexicana_all_races - unique SNPs (15)
number_of_tillers_(total_plant_surface_area)
parviglumis - unique SNPs (10)
45.5%
54.5%
mexicana_all_races - unique SNPs (12)
precipitation_seasonality
mexicana_all_races - unique SNPs (2)
28.6%
71.4%
parviglumis - unique SNPs (5)

b)

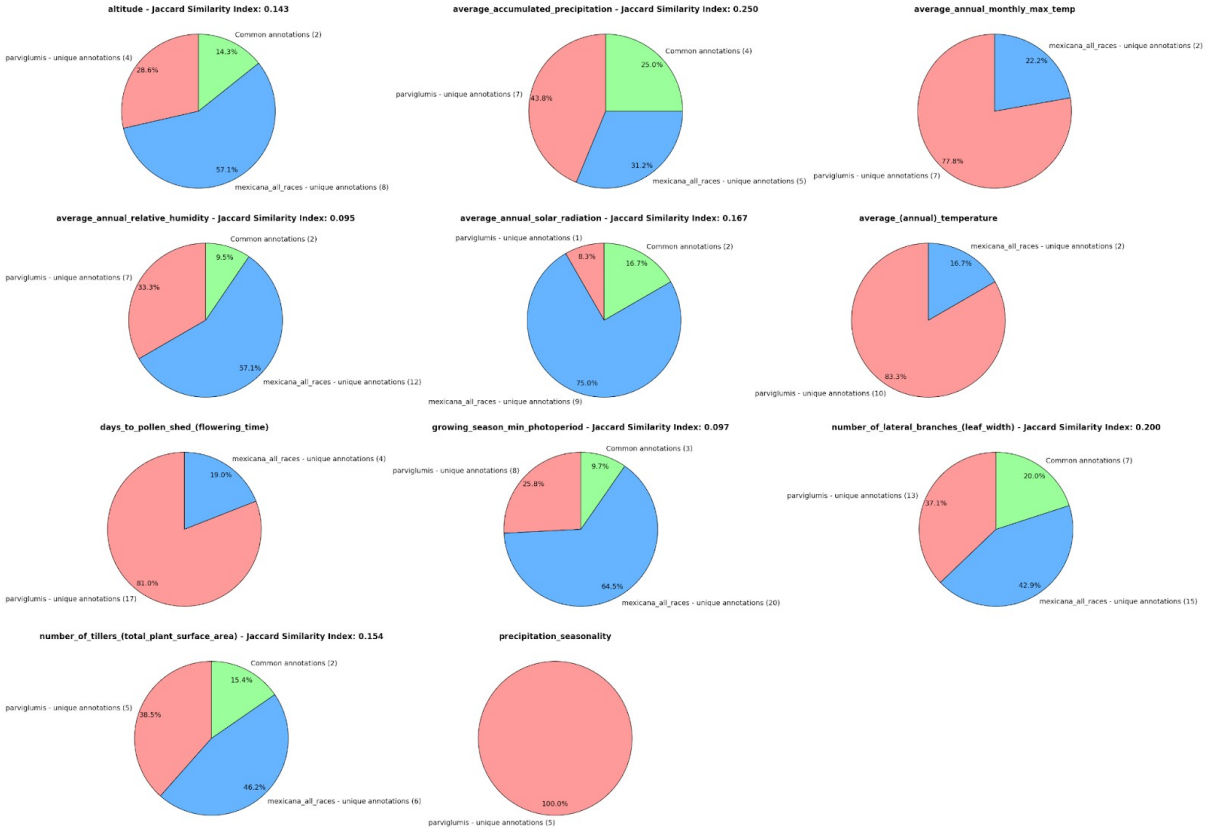

altitude - Jaccard Similarity Index: 0.143
Common annotations (2)
parviglumis - unique annotations (4)
14.3%
28.6%
57.1%
mexicana_all_races - unique annotations (8)
average_accumulated_precipitation - Jaccard Similarity Index: 0.250
Common annotations (4)
parviglumis - unique annotations (7)
25.0%
43.8%
31.2%
mexicana_all_races - unique annotations (5)
average_annual_monthly_max_temp
mexicana_all_races - unique annotations (2)
22.2%
77.8%
parviglumis - unique annotations (7)
average_annual_relative_humidity - Jaccard Similarity Index: 0.095
Common annotations (2)
parviglumis - unique annotations (7)
9.5%
33.3%
57.1%
mexicana_all_races - unique annotations (12)
average_annual_solar_radiation - Jaccard Similarity Index: 0.167
parviglumis - unique annotations (1)
Common annotations (2)
8.3%
16.7%
75.0%
mexicana_all_races - unique annotations (9)
average_(annual)_temperature
mexicana_all_races - unique annotations (2)
16.7%
83.3%
parviglumis - unique annotations (10)
days_to_pollen_shed_(flowering_time)
mexicana_all_races - unique annotations (4)
19.0%
81.0%
parviglumis - unique annotations (17)
growing_season_min_photoperiod - Jaccard Similarity Index: 0.097
Common annotations (3)
parviglumis - unique annotations (8)
9.7%
25.8%
64.5%
mexicana_all_races - unique annotations (20)
number_of_lateral_branches_(leaf_width) - Jaccard Similarity Index: 0.200
Common annotations (7)
parviglumis - unique annotations (13)
20.0%
37.1%
42.9%
mexicana_all_races - unique annotations (15)
number_of_tillers_(total_plant_surface_area) - Jaccard Similarity Index: 0.154
Common annotations (2)
parviglumis - unique annotations (5)
15.4%
38.5%
46.2%
mexicana_all_races - unique annotations (6)
precipitation_seasonality
100.0%
parviglumis - unique annotations (5)

Figure S3. Comparative GWAS results for *mexicana* and *parviglumis.* For each factor group (see Figure 1), pie charts illustrate the number of associated markers (Figure S3a) and candidate gene protein functions (Figure S3b). Associations unique to *mexicana* are shown in blue, those unique to *parviglumis* in red, and shared associations in green. The Jaccard Similarity Index is displayed next to factor labels where values exceed 0, highlighting the similarity between taxa. Panel S3a shows no shared SNPs between the two taxa, while Panel S3b indicates that certain protein functions are shared across specific factor groups.

CHAPTER 3

GXP: Analyze and plot plant omics data in web browsers

This manuscript was published in the MDPI Plants Journal in March 2022.

Authors:
Constantin Eiteneuer†, David Velasco†, Joseph Atemia†, Dan Wang, Rainer Schwacke, Vanessa Wahl, Andrea Schrader, Julia J. Reimer, Sven Fahrner, Roland Pieruschka, Ulrich Schurr, Björn Usadel and Asis Hallab

#Corresponding author: a.hallab@fz-juelich.de

Contribution: †Shared first author

Asis Hallab and Björn Usadel conceptualized the project. Asis Hallab led the software design of the Gene Expression Plotter (GXP) browser-based application. **Joseph Atemia**, Constantin Eiteneuer, David Velasco, and Dan Wang programmed the application software and carried out software tests. Constantin Eiteneuer and David Velasco designed the graphical interface. Björn Usadel provided scientific guidance, particularly regarding the methods to be applied, and feedback on the user interface. Björn Usadel and Rainer Schwacke implemented a proof-of-concept Javascript software implementation of the MapMan visualizations. Andrea Schrader provided an extensive review through testing and editing the GXP user manual and delivered detailed user feedback. Vanessa Wahl iteratively used and tested GXP in an RNAseq project, contributed to feature design, and provided extensive user feedback. Julia J. Reimer provided preprocessed biological data. Roland Pieruschka, Sven Fahrner, and Ulrich Schurr provided project administration, integrated and applied GXP into ongoing scientific research projects, and gave user feedback. Asis Hallab, **Joseph Atemia**, Constantin Eiteneuer, and Björn Usadel wrote the manuscript with the help of all authors.

plants

MDPI

Technical Note

GXP: Analyze and Plot Plant Omics Data in Web Browsers

Constantin Eiteneuer [1,†], David Velasco [2,†], Joseph Atemia [3,†], Dan Wang [1], Rainer Schwacke [3], Vanessa Wahl [4], Andrea Schrader [5], Julia J. Reimer [5,6], Sven Fahrner [1], Roland Pieruschka [1], Ulrich Schurr [1], Björn Usadel [3] and Asis Hallab [3,*]

1 IBG-2 Plant Sciences, Forschungszentrum Jülich, 52428 Jülich, Germany; c.eiteneuer@fz-juelich.de (C.E.); d.wang@fz-juelich.de (D.W.); s.fahrner@fz-juelich.de (S.F.); r.pieruschka@fz-juelich.de (R.P.); u.schurr@fz-juelich.de (U.S.)
2 Faculty of Natural Sciences, Norges Teknisk-Naturvitenskapelige Universitet, 7034 Trondheim, Norway; davidve@stud.ntnu.no
3 IBG-4 Bioinformatics, Forschungszentrum Jülich, 52428 Jülich, Germany; j.atemia@fz-juelich.de (J.A.); r.schwacke@fz-juelich.de (R.S.); b.usadel@fz-juelich.de (B.U.)
4 Max Planck Institute for Molecular Plant Physiology, 14476 Potsdam, Germany; vwahl@mpimp-golm.mpg.de
5 Institute for Biology I, RWTH Aachen University, 52062 Aachen, Germany; schrader@bio1.rwth-aachen.de (A.S.); julia.reimer@hs-emden-leer.de (J.J.R.)
6 Faculty of Technology, University of Applied Science Emden/Leer, Molecular Biosciences, 26723 Emden, Germany
* Correspondence: a.hallab@fz-juelich.de
† These authors contributed equally to this work.

Abstract: Next-generation sequencing and metabolomics have become very cost and work efficient and are integrated into an ever-growing number of life science research projects. Typically, established software pipelines analyze raw data and produce quantitative data informing about gene expression or concentrations of metabolites. These results need to be visualized and further analyzed in order to support scientific hypothesis building and identification of underlying biological patterns. Some of these tools already exist, but require installation or manual programming. We developed "Gene Expression Plotter" (GXP), an RNAseq and Metabolomics data visualization and analysis tool entirely running in the user's web browser, thus not needing any custom installation, manual programming or uploading of confidential data to third party servers. Consequently, upon receiving the bioinformatic raw data analysis of RNAseq or other omics results, GXP immediately enables the user to interact with the data according to biological questions by performing knowledge-driven, in-depth data analyses and candidate identification via visualization and data exploration. Thereby, GXP can support and accelerate complex interdisciplinary omics projects and downstream analyses. GXP offers an easy way to publish data, plots, and analysis results either as a simple exported file or as a custom website. GXP is freely available on GitHub (see introduction)

Keywords: RNA sequencing; metabolomics; data visualization; overrepresentation analysis; correlation; cluster analysis; principal component analysis; scientific plotting; Mapman; Mercator

check for updates

Citation: Eiteneuer, C.; Velasco, D.; Atemia, J.; Wang, D.; Schwacke, R.; Wahl, V.; Schrader, A.; Reimer, J.J.; Fahrner, S.; Pieruschka, R.; et al. GXP: Analyze and Plot Plant Omics Data in Web Browsers. *Plants* **2022**, *11*, 745. https://doi.org/10.3390/plants11060745

Academic Editors: Ji Huang and Yufeng Wu

Received: 11 January 2022
Accepted: 1 March 2022
Published: 11 March 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Modern life science research projects often produce quantitative data, for example to quantify gene expression or metabolite concentration in tissue samples. Many well-established tools exist that carry out the wet lab and bioinformatics procedures to produce such count data from the samples. Not rarely, these pipelines are carried out by third party laboratories. In the subsequent step, the life scientist, having ordered such RNAseq or other Omics experiments, needs to investigate these count data to form scientific hypotheses and identify underlying biological patterns. Typically, this involves plotting the quantitative data, carrying out Principal Component Analyses and correlation-based hierarchical

clustering to elucidate differences between experimental conditions. Genetic or metabolic responses to the tested experimental conditions and treatments often are summarized by identification of enriched traits within significantly up- or down-regulated genes or metabolites of interest. These steps often require manual programming, installation of software, or sending potentially confidential data to webservers for analysis. Gene Expression Plotter (GXP) minimizes these requirements by enabling the user to load count data, generate a variety of informative plots, and carry out typical clustering, principal component and overrepresentation analyses, without the need to write any code, install any software, or send the data to third party servers. Furthermore, GXP enables the user to save all loaded data along with the work done, including generated plots and carried out analysis. With this feature the user can not only save the current work to continue at a later time, but also share data, plots, and analysis results with others, simply by sending the exported GXP database file. Naturally, such a file can be published for example in the form of an article's supplement thus enabling readers to directly obtain the data, see plots and analysis results and even carry out their own subsequent investigations.

Gene Expression Plotter consumes two types of input tables, which can be prepared with standard spreadsheet programs, e.g., Microsoft Excel, or can be generated directly by the bioinformaticians producing the quantitative data (for details see results Section 2.1). In short, in the input quantifications table, each row represents quantitative values assessed for a single gene or metabolite, and columns correspond to the different samples for which these quantifications were assessed. In addition to the pure quantifications data table a free format information table can be provided. In it, too, rows correspond to the genes or metabolites, as they appear in the quantifications data table, and columns can provide any free text, categorical, or numeric information the user wants to load into GXP. Such free format information typically comprises knowledge about molecular gene function, for example in the form of Mapman4 Bin annotations [1–3], InterPro conserved protein domains [4], or Gene Ontology terms [5]. Additionally, in this table, the user can provide information, e.g., in the form of logarithmic fold change values, quantifying how much the expression of a particular gene changes when contrasting two selected experimental conditions, e.g., control versus stress treatment. Or, in case metabolites have already been quantified, information about their respective chemical properties and involved pathways can be provided. Both of this optional categorical and numerical data can later be used in GXP's enrichment analyses or can be displayed in the captions of generated plots, here comprising free text information, too. Thus, GXP has been developed to consume generic input data, making it a highly versatile tool, particularly in the context of overrepresentation analysis.

To produce, analyze, visualize, and publish quantitative omics data many tools exist. Among others, they provide means to plot the data, carry out clustering, and conduct principal component and overrepresentation analyses. A number of these tools specialize in RNAseq analysis [6–21], most of which consume the raw gene expression count data produced by standard gene expression quantifiers [22–25] and enable the user to identify differentially expressed genes [6–9,11–13,15,16,20,21,26,27] and review the results in form of comprehensive reports and/or plots [6–13,15,16,18–21,26–28]. Some [7–14,18–20] are implemented as an R / Shiny [29,30] application or use other forms of graphical user interfaces (GUI) [15,16,21,27]. These GUI tools allow the user (i) to either execute the tool installed locally on their computer and/or use it on a public web-server and (ii), by means of the GUI, eliminate the need to program plots and analyses manually, with few exceptions [6,28] which require some manual coding to make use of the extended provided functionality. By means of integration of curated published data some tools offer specific analyses, e.g., providing high-confidence insights into molecular gene function. One of these, GENAVi [12], enables the identification of differentially expressed genes (DEGs) in human or mouse RNAseq data by contrasting input with published data. OnestopRNAseq [16] is another example and offers several useful analyses by integration of curated public data from several model animal organisms. In addition, Plant Physiospace [31] enables the user to compare differential gene expression data with curated signatures to identify similar

genetic responses investigated in other already-published studies. Other tools focus on metabolomics [32–34] or, integrating RNAseq and Metabolomics data, the identification of genotype–phenotype relationships [35]. Some of the introduced tools [7,13,28] offer the elucidation of interactive plots where, e.g., hovering with the mouse over data points in a plot summarizing differential gene expression opens another plot illustrating the expression counts of the particular gene represented by the hovered-over data point. For a more detailed review of the above tools, see supplemental Text S1. In comparison with the above tools, GXP has a "downstream focus" on the visualization of quantitative omics data and subsequent clustering, principal component, and overrepresentation analysis. In this context, the key features of GXP are that (i) it does not require manual programming, nor installation of particular software, (ii) it can thus be used on a simple tablet or even smartphone, and (iii) GXP is versatile, in that it can consume any quantitative omics data stemming from RNAseq or metabolomics analyses. This genericity especially includes any additional arbitrary information on the quantified entities, that is, either genes or metabolites. As explained above, this particularly enables the user to carry out overrepresentation analysis on any, numerical or categorical, data the user provides (see results sections "2.1 Handling input and output data" and "2.5 Overrepresentation (enrichment) analysis" for more details). Furthermore, (iv) GXP is the first, immediately accessible, mature implementation of the popular Mapman tool that visually combines quantitative omics data analysis results with diagrams of metabolic pathways and other processes (see results section "2.4 MapMan web browser plots" for more details). Finally, (v) GXP ensures complete data safety. To explain this, consider that even though GXP is deployed on a webserver, once it has been loaded into the user's browser, it runs there completely independent of that server. This form of implementation is called a "single page application" (SPA) in which the webserver is little more than a file system delivering GXP to the user's browser. Among other things the implementation as a SPA implies that at no time is any data sent to any server. All analyses are carried out on the user's computer in the used web-browser and all data remains exactly there.

Gene Expression Plotter is freely available on GitHub (https://usadellab.github.io/GeneExpressionPlots/; accession date 11 January 2022). Its code has been released as open source (https://github.com/usadellab/GeneExpressionPlots; accession date 11 January 2022) under the GNU public license, version 3. All functions carrying out the above-described analyses are tested with automatic software unit tests to ensure correct calculations. As mentioned, GXP provides means to export all loaded data, generated plots, and carried out analyses results into a single file, dubbed the "GXP database". Such a GXP database can be used for publication, e.g., in the form of an article supplement. Additionally, GXP can easily be copied and this copy can be published including custom quantitative omics data, plots, and carried-out analyses. Such a copy can be made available, e.g., on GitHub free of charge or any other webserver. The GXP manual has detailed instructions on a simple procedure to set up such a custom copy of GXP and includes screenshots on seven easy steps that only require a GitHub account and a web browser.

Thus, our new tool, Gene Expression Plotter, enables the end-user to visualize and analyze quantitative data, typically taken from RNAseq or metabolomics analyses, identify similarity between experimental conditions by cluster and principal component analysis, generate visual summaries of genetic or metabolic responses, identify overrepresented transcripts or metabolite characteristics, and even use GXP to publish the data along with plots and analysis results.

2. Results

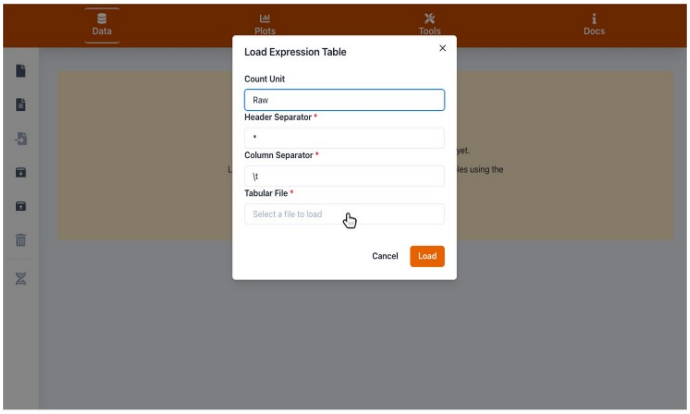
2.1. Handling Input and Output Data

With the aim of providing a single suite in which a user can visualize and analyze quantitative omics data and also share and publish the output, we programmed "Gene-Expression-Plotter" (GXP). GXP is available freely with a GPL license on GitHub: https://github.com/usadellab/GeneExpressionPlots (accession date 11 January 2022).

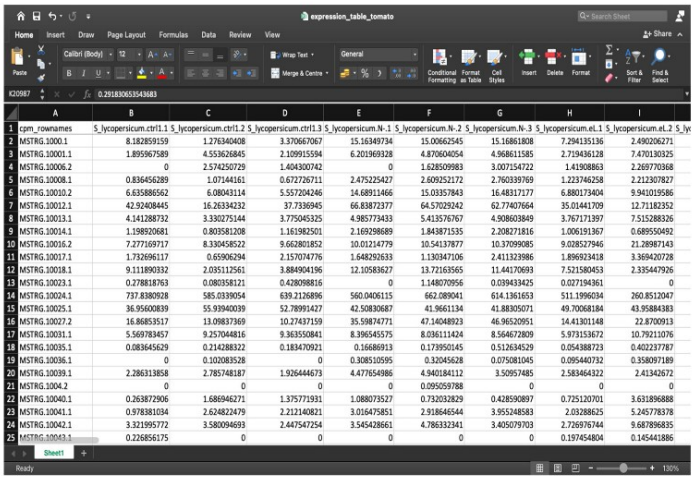
GXP is provided with a comprehensive online documentation and a manual (see menu "Docs"). To use GXP, the user is first asked to import quantitative RNA-seq or metabolomics data in a tabular format. Each row should represent, e.g., a single quantified transcript. The transcript identifiers are typically provided in the first column. All following columns should contain the transcript quantifications for each respective genotype, replicate or treatment as specified by the column names (Figure 1a,b). GXP is made aware of the statistical factors differentiating in the experiment the respective biological replicates. This is encoded directly in the respective column names in the input quantification table. Here, as one studied factor the user can for example specify the time after an experimental treatment at which a sample was obtained or the type of stress treatment a sample was exposed to. Such a factor is then used by GXP to position points on the x-axis in plots. We consequently dub such factors "x-axis factors". In the example data included in GXP, a column name positioning data points over the "ctrl1" x-axis tick (Figure 2) would be, for example, "S_lycopersicum.ctrl1.1" or "S_lycopersicum.ctrl1.2". Note that the x-axis factor is identified by its location between the first and second "." in the column name. The user can select another character, e.g., "*" instead of ".". Strictly speaking, GXP accepts a single x-axis factor that can have multiple values. In the example data these can be "ctrl1" (control) and the stress treatments "cold" (chilling temperature), "eL" (extended light), and "N-" (nitrogen deficiency). Thus, x-axis factors join (in typical RNAseq experiments three) biological replicates that were subjected to the same experimental treatment into a single bin. Such x-axis bins are subsequently used to calculate y-axis error bars for data points representing several joined replicates (see Section 2.2 and Figure 2). Another type of factor can optionally be introduced and is used to compare for example biological species, genotypes, or different treatments. A good example for such a type of factor can be the comparison of a wild (*Solanum pennellii*) versus a domesticated (*S. lycopersicum*) tomato species. We dub such factors that group several biological conditions "group factors". Note that these group factors are not used to generate tick labels on the x-axis, but rather imply two plots showing, e.g., gene expression in the wild type and domesticated species side-by-side (Figure 2a,b). A user can specify as many group factors as were investigated in a respective research project. Additionally, in the case of group factors, GXP is made aware of these simply through the column names of the counts table input. In the example data the column names "S_lycopersicum.ctrl1.1" and "S_pennellii.ctrl1.3" indicate the group factor "species" with values *S. lycopersicum* and *S. pennellii*, respectively. Note that the group factors are also identified by their position in the column names, appearing before the x-axis factor and separated by the "." character. As mentioned before, the user can specify any other character as separator, e.g., "*".

Optionally, GXP can use any extra information associated with specific quantified transcripts or metabolites. This is done by loading a separate table in which each row corresponds to a single quantified entity (transcript or metabolite), and in which each column holds additional generic information (Figure 1c,d). Examples of such additional generic information include ontological annotations informing about the molecular function of proteins, e.g., terms from the Mapman4 framework [3,36], from the Gene Ontology (GO) project [5], or from KEGG pathways [37], or differential expression between contrasted conditions, e.g., cold stress treatment versus control conditions. Information about differential expression can be provided in the form of logarithmic fold change of gene expression and/or adjusted *p*-values used to identify significant changes in gene expression. In fact, in this optional information table, the user can provide any information in the form of columns that contain either free text, numerical values, e.g., chemical properties, such as hydrophobicity, of metabolites, or categorical annotations about molecular protein function, similar to the terms obtained from Mapman4 ontology [36], the GO [5], the KEGG [37], or InterPro [4]. Note that GXP offers tools to help the user obtain and import molecular gene function annotations in the form of Mapman Bins for his data in the respective "Mapman functional annotations" section of the "Tools" menu. Importantly numeric and categorical information can be used to carry out subsequent overrepresentation analysis, while all

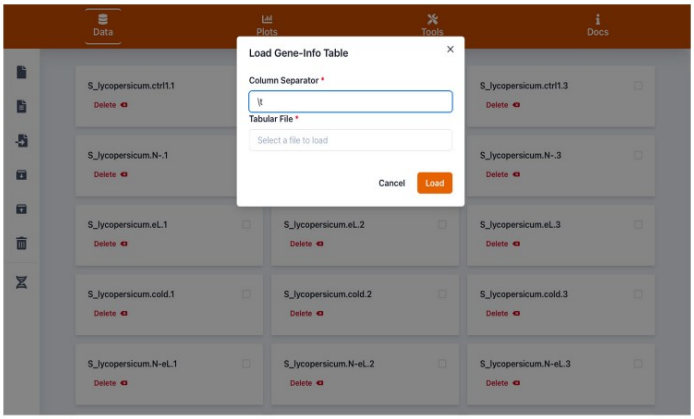
information is used in the gene browser to provide the user with a rich database about the studied transcripts or metabolites (Section 2.1.1 and Figure 1e).



(a)

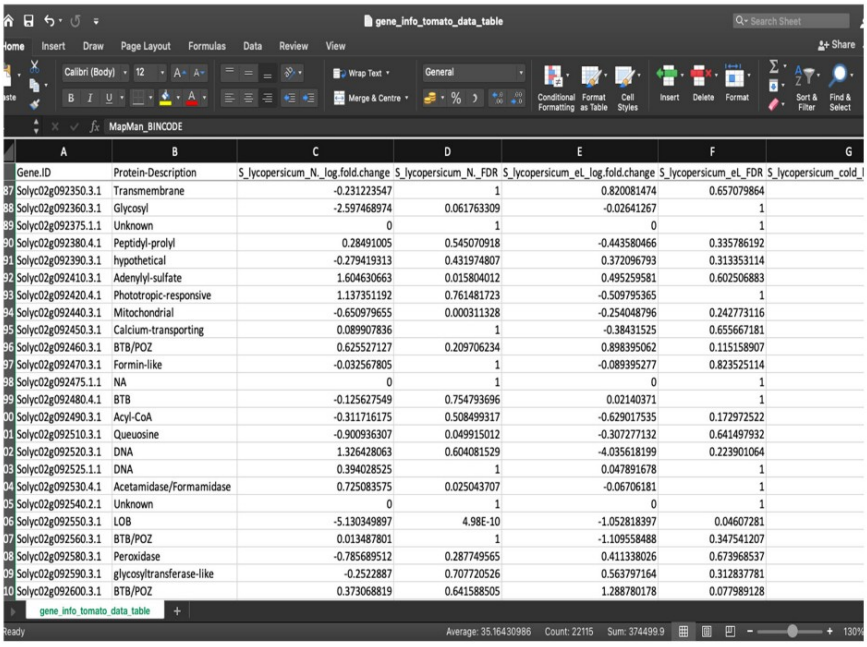


(b)



(c)

Figure 1. *Cont.*



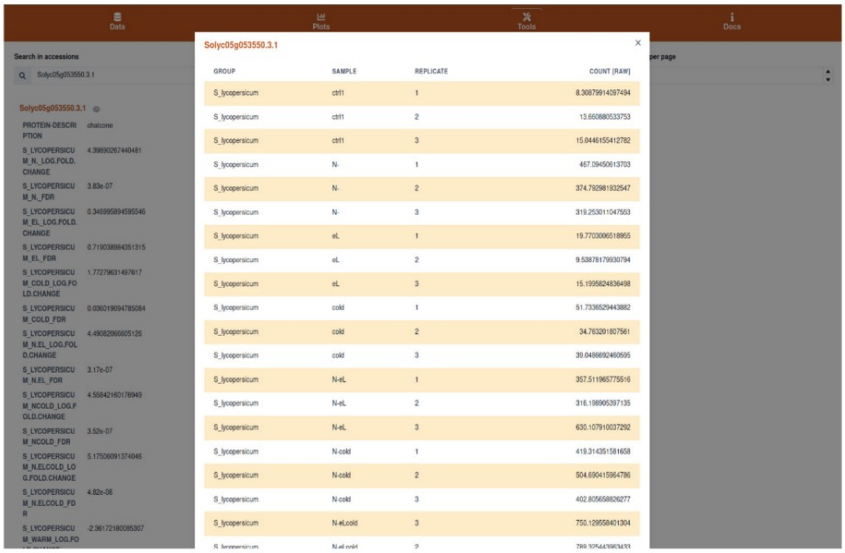
gene_info_tomato_data_table

	A	B	C	D	E	F	G
	Gene.ID	Protein-Description	S_lycopersicum_N._log.fold.change	S_lycopersicum_N._FDR	S_lycopersicum_eL_log.fold.change	S_lycopersicum_eL_FDR	S_lycopersicum_cold_l
87	Solyc02g092350.3.1	Transmembrane	-0.231223547	1	0.820081474	0.657079864	
88	Solyc02g092360.3.1	Glycosyl	-2.597468974	0.061763309	-0.02641267	1	
89	Solyc02g092375.1.1	Unknown	0	1	0	1	
90	Solyc02g092380.4.1	Peptidyl-prolyl	0.28491005	0.545070918	-0.443580466	0.335786192	
91	Solyc02g092390.3.1	hypothetical	-0.279419313	0.431974807	0.372096793	0.313353114	
92	Solyc02g092410.3.1	Adenylyl-sulfate	1.604630663	0.015804012	0.495259581	0.602506883	
93	Solyc02g092420.4.1	Phototropic-responsive	1.137351192	0.761481723	-0.509795365	1	
94	Solyc02g092440.3.1	Mitochondrial	-0.650979655	0.000311328	-0.254048796	0.242773116	
95	Solyc02g092450.3.1	Calcium-transporting	0.089907836	1	-0.38431525	0.655667181	
96	Solyc02g092460.3.1	BTB/POZ	0.625527127	0.209706234	0.898395062	0.115158907	
97	Solyc02g092470.3.1	Formin-like	-0.032567805	1	-0.089395277	0.823525114	
98	Solyc02g092475.1.1	NA	0	1	0	1	
99	Solyc02g092480.4.1	BTB	-0.125627549	0.754793696	0.02140371	1	
00	Solyc02g092490.3.1	Acyl-CoA	-0.311716175	0.508499317	-0.629017535	0.172972522	
01	Solyc02g092510.3.1	Queuosine	-0.900936307	0.049915012	-0.307277132	0.641497932	
02	Solyc02g092520.3.1	DNA	1.326428063	0.604081529	-4.035618199	0.223901064	
03	Solyc02g092525.1.1	DNA	0.394028525	1	0.047891678	1	
04	Solyc02g092530.4.1	Acetamidase/Formamidase	0.725083575	0.025043707	-0.06706181	1	
05	Solyc02g092540.2.1	Unknown	0	1	0	1	
06	Solyc02g092550.3.1	LOB	-5.130349897	4.98E-10	-1.052818397	0.04607281	
07	Solyc02g092560.3.1	BTB/POZ	0.013487801	1	-1.109558488	0.347541207	
08	Solyc02g092580.3.1	Peroxidase	-0.785689512	0.287749565	0.411338026	0.673968537	
09	Solyc02g092590.3.1	glycosyltransferase-like	-0.2522887	0.707720526	0.563797164	0.312837781	
10	Solyc02g092600.3.1	BTB/POZ	0.373068819	0.641588505	1.288780178	0.077989128	

gene_info_tomato_data_table

Ready Average: 35.16430986 Count: 22115 Sum: 374499.9 130%

(**d**)

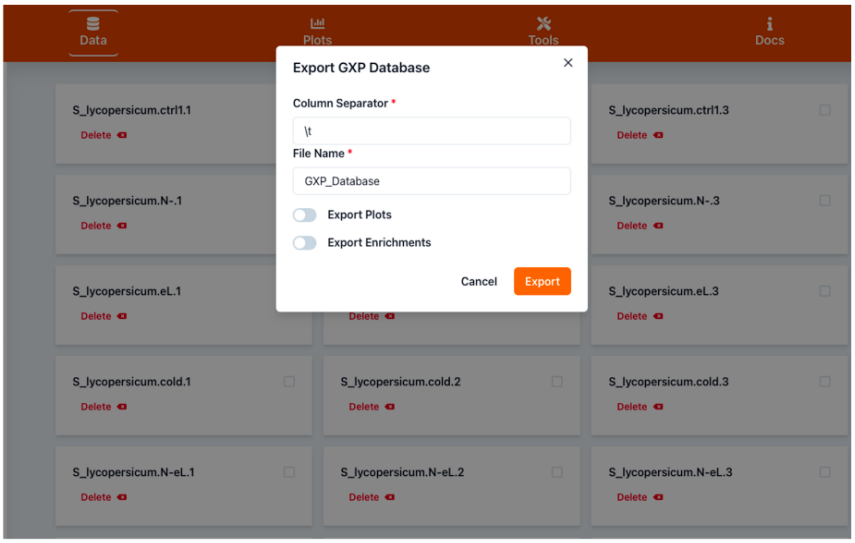


Solyc05g053550.3.1

GROUP	SAMPLE	REPLICATE	COUNT (RAW)
S_lycopersicum	ctrl1	1	8.3087991409749 4
S_lycopersicum	ctrl1	2	13.6608805337 53
S_lycopersicum	ctrl1	3	15.0446155412782
S_lycopersicum	N-	1	467.094506 13703
S_lycopersicum	N-	2	374.792981932547
S_lycopersicum	N-	3	319.253011047553
S_lycopersicum	eL	1	19.7703006518955
S_lycopersicum	eL	2	9.538781799307 94
S_lycopersicum	eL	3	15.1955824836498
S_lycopersicum	cold	1	51.7336529443882
S_lycopersicum	cold	2	34.763201807561
S_lycopersicum	cold	3	39.0486692460595
S_lycopersicum	N-eL	1	357.5119657755 16
S_lycopersicum	N-eL	2	316.1989053971 35
S_lycopersicum	N-eL	3	630.1079100372 92
S_lycopersicum	N-cold	1	419.3143051581658
S_lycopersicum	N-cold	2	504.690415964786
S_lycopersicum	N-cold	3	402.8056588262 77
S_lycopersicum	N-eLcold	3	750.1295584013 04

(**e**)

Figure 1. *Cont.*



(**f**)

Figure 1. Screenshot images of Gene Expression Plotter (GXP) showing interfaces for data input. (**a**) shows the quantitative data table file import form, triggered by the first document button in the upper left panel. This form is used to load a quantification table such as the one shown in (**b**). (**c**) shows the transcript information table file import form, triggered by the second document icon in the left panel. This form is used to load an optional information table (see Section 1) like the one shown in (**d**). After successful import, the user can search for genes or their annotations in the "gene browser" (helix icon on the lower left panel). In the shown example, the user searched for "chalcone synthase", a polyketide synthase involved in flavonoid biosynthesis. In (**e**) the user now inspects this gene's expression quantifications (highlighted foreground) and additional information such as the logarithmic fold changes of gene expression assessed for various comparisons of control and stress treatments (lightly faded out background). Furthermore, as shown in (**f**), by using the GXP export function triggered by the fifth, upward arrow on the box icon in the left panel, GXP enables the user to save the current state, i.e., all imported data, generated plots, and analysis results for later continuation or to share it with other researchers.

Gene Expression Plotter is freely available on GitHub for direct use (https://usadellab.github.io/GeneExpressionPlots; accession date 11 January 2022). It is provided with example RNAseq data from a study on stress response contrasting wild with domesticated tomato species [38]. In collaboration with one of the original authors, this data has been used to directly compare and benchmark the results produced by GXP with the already published findings. This example data can be loaded by clicking on "Load example data" in the "data" menu.

2.1.1. Browsing and Searching Gene Information

After having loaded the input data, the user can browse gene information in a searchable interface and make this information available to all who have access (Figure 1e). The presented information includes the respective quantitative data extracted from the "quantifications table" (see Section 2.1) and any further information about the respective transcripts or metabolites extracted from the "information table" (see Section 2.1). A user can, e.g., search for a gene of interest "chalcone synthase" and inspect all transcripts associated with this molecular function, their respective expression, and additional information such as conserved protein domains.

Figure 2. Screenshots of plots visualizing and comparing quantified gene expression between different treatments, experimental conditions, and genes: (**a**) bar plot, (**b**) individual lines plot and (**c**) stacked lines plot, are different modes of how Gene Expression Plotter (GXP) visualizes the expression profile of the example gene Solyc05g053550.3.1 (*CHALCONE SYNTHASE*). The three plots highlight how the expression of the example *CHALCONE SYNTHASE* responds to the experimental conditions. This *CHALCONE SYNTHASE's* expression is up-regulated in *S. lycopersicum* but conversely not in *S. pennellii* following stress treatments of nitrogen deficiency (N-) and in combination with chilling temperatures (cold) and elevated light intensity (eL). Plot (**d**) compares the genetic response of this *CHALCONE SYNTHASE* with another gene of interest Solyc08g075570.4.1 (*UREA PROTON SYMPORTER*). In contrast to the expression of *CHALCONE SYNTHASE*, gene expression of the *UREA PROTON SYMPORTER* is relatively low in both *S. pennellii* and *S. lycopersicum*.

2.1.2. Saving Work and Exporting Data

At any stage of using GXP, the user can export and save all imported data, plots, and analyses by using the dedicated "Export GXP Database" functionality (Figure 1f). The generated database can be used later to resume previous work or share plots and results. The exported GXP database contains a configuration file ("GXP_settings.json", see manual for more details) that can be used, e.g., to change the order of "x-axis factors", the unit of the expression values, and the various field separators used to load the tables into the application. At any stage all data is strictly kept on the user's computer and at no point is user data sent through the web.

2.2. Visualizing Quantitative Data

The user can generate plots showing the expression of individually selected genes. Available visualizations are bar- (Figure 2a) and line-plots (Figure 2b). In these plots, "x-axis factors" (see Section 2.1) define the position of points on the x-axis. If the user has provided "group factor" information (see Section 2.1) that further groups biological replicates, e.g., species, genotype, or different treatments. This information will be visualized by the color or position of plotted values (Figure 2; domesticated tomato *S. lycopersicum* in red on the left side and wild *S. pennellii* in green on the right side). Quantifications differing between group factors, e.g., domesticated versus wild tomato species, can either be plotted in two graphs side-by-side (Figure 2a,b) or in a single graph as differentially colored stacked curves (Figure 2c). The user can also visualize the expression of multiple genes/metabolites in the same plot using all mentioned types: bar, single, or stacked curves (Figure 2d for an example with bars). In this case, as mentioned, the color distinguishes genes while the group factors are shown in their separate graphs side-by-side. All plots are interactive in that upon hovering over data points with the mouse, the user is presented with the respective values in a little overlay window. Hovering with the mouse over a specific point will display the underlying plotted data corresponding to that point, or bar, respectively. All plots a user generates can be saved and downloaded in high quality scalable vector graphics and used for publication purposes. Furthermore, exporting the GXP data and state optionally saves generated plots and analyses as well (see Section 2.1).

2.3. Assessing Similarity of Biological Replicates Based on Either Gene Expression or Quantified Metabolites

Typically, during the analysis and interpretation or RNAseq or metabolomics experiment results, one wants to distinguish within, i.e., biological background noise, from between-group differences to enable drawing significant conclusions in terms of an organism's response to contrasting experimental conditions. Only if the in-between-group variation is not silenced by the background noise, can the data be used to elucidate the original biological questions motivating the study. To assess background noise and in-between-group variation, typically several (at least three in RNAseq experiments) biological replicates sharing the same experimental condition are quantified. Subsequently, similarity of the quantified replicates should be recognizable over the background noise to indicate that the quantifications can be used for the investigation of the original biological question motivating the study. This similarity can be assessed and visualized using correlation or Euclidean distance based hierarchical clustering and principal component analysis.

2.3.1. Hierarchical Cluster Analysis

Results of a hierarchical cluster analysis, executed on optionally z-transformed values, are visualized in a heatmap whose axis is accompanied with a tree dendrogram representing the hierarchical clusters that the respective biological replicates have been grouped into (Figure 3a). A transposed cluster analysis can also be carried out, in which all or a selected subset of transcripts/metabolites are grouped by similarity of their respective quantitative data. GXP computes and visualizes hierarchical cluster analysis on demand in the user's browser. The user can either select correlation or Euclidean distance between replicates'

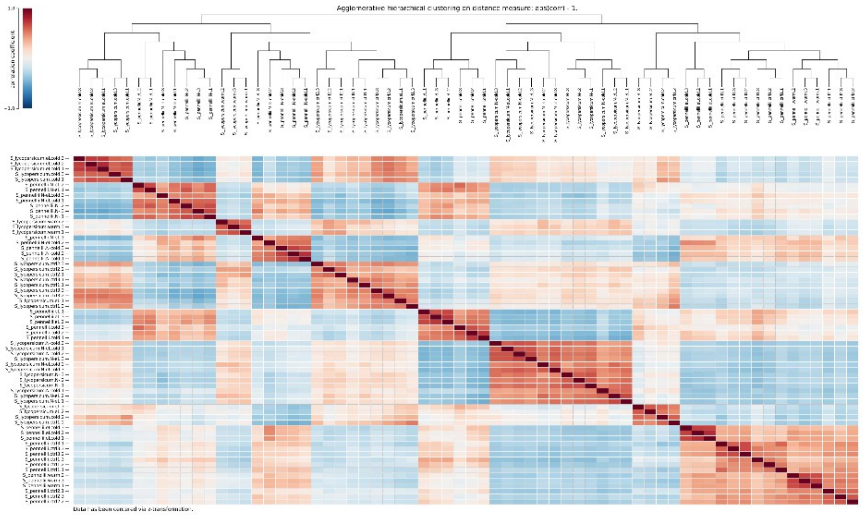
gene expression vectors as a basis for the clustering analysis (see Section 4.3 for more details). The potentially demanding analysis is carried out in the background, using so called "web-workers", and therefore does not block the user interface. While an analysis runs, a plot showing a loading icon is immediately created, indicating ongoing calculation. Once complete, the plot color indicates either the correlation values or euclidean distances between respective replicates, depending on the user's original choice. A color scale is provided, and the plot is interactive, in that hovering with the mouse over a heatmap cell will display the calculated likeness value assessed for the respective pair the cell's row and column corresponds to, respectively.

2.3.2. Principal Component Analysis

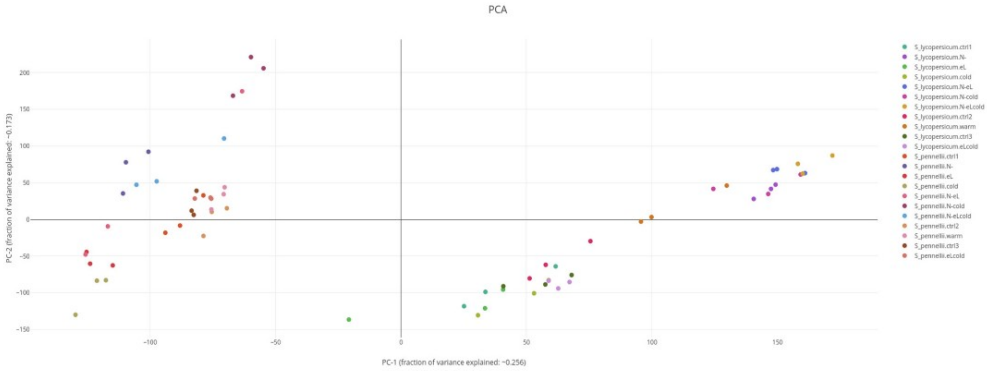
Results of principal component analysis (PCA) are typically visualized as a subclass of scatter-plots where the two axes represent the two principal components that explain most of the observed variance between samples. GXP carries out a PCA on user demand and runs the respective calculation in the user's browser in the background, thus not blocking the user interface. Once the calculation is finished, the loading icon indicating ongoing processing disappears and the respective scatter plot becomes visible (Figure 3b). Biological replicates are color coded so that replicates that have identical x-axis factors (Section 2.1), e.g., all replicates belonging to the wild tomato species *S. pennellii* that also have been exposed to the same cold stress conditions receive the same color. In the PCA scatter plot hovering with the mouse over a data point will display the name of the underlying biological replicate and the values of the two visualized principal components interactively.

2.4. Mapman Web Browser Plots

The Mapman frameworks (Mapman4 [36] and the older version Mapman v.3.6 [39]) comprise a manually curated vocabulary (ontology) to describe the function of land plant proteins. Mercator and Mercator4 [3,36] are efficient and accurate genome scale annotation pipelines that assign the descriptions of Mapman v.3.6 and of Mapman4, respectively, to query proteins or transcripts. The desktop application MapMan [1,3,36,39]) has been developed to visualize annotations of the Mapman frameworks in the context of gene expression data. Based on a proof-of-principle code [1,40] we also developed a simple MapMan web browser application. The same as in the MapMan desktop application, a user can choose one of several basic metabolic cellular sketches, e.g., "Metabolism overview", "Photosynthesis", or "Secondary metabolism" (Figure 4). In these sketches, small squares represent proteins or transcripts with functional annotations semantically corresponding to that region in the diagram. For example, all proteins with functions related to the Calvin cycle would appear as small boxes in the upper left area of the "Photosynthesis" sketch (Figure 4b).



(**a**)



(**b**)

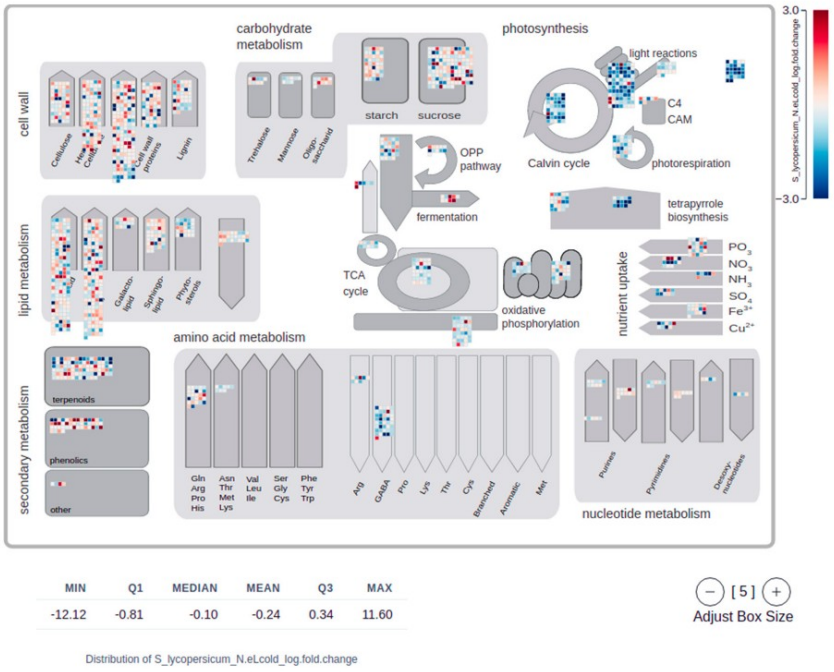
Figure 3. Screenshots of plots investigating likeliness of gene expression assessed in different biological replicates. Plot (**a**) shows the result of correlation-based hierarchical clustering in the form of a dendrogram and a correlation heatmap. In the top left corner, a scale color-codes the calculated correlation coefficients. In the top middle, the dendrogram represents the result of hierarchical clustering of all loaded biological replicates. In this example, the plot informs the user of their choice to z-transform the data before the calculation of correlation (lower left corner). Upon hovering with the mouse over single cells of the correlation matrix, the user is presented with the respective correlation value between the two biological replicates represented by the cell's row and column, respectively. This example shows how Gene Expression Plotter (GXP) helps the user to assess how well the applied experimental conditions and treatments are reflected in terms of quantified gene expression. Here, serving as a quality check, the statistical factors "species" and "stress treatments" mostly imply the grouping of biological replicates, highlighting that the experimental setup and bioinformatics analysis yielded data fit to carry out the original biological question of the study, namely to elucidate the genetic responses to the applied stress treatments and subsequently compare

these genetic responses between the two studied species of tomato. A plot highlighting similar patterns is shown in (**b**). Here, a principal component analysis (PCA) has been carried out on z-transformed data. The resulting scatter plot of the two most important principal components (PC) confirms that the color-coded biological replicates (legend in the top right corner) mostly group by the factors "species" and "stress treatment", i.e., are found in close proximity within the scatter plot. When hovering with the mouse over single data points, the user is presented with the exact PC values and the name of the respective biological replicate represented by the data point. Using the axes labels, the user is informed about how much of the observed variation is explained by the two respective principal components PC1 (here: approx 25.6 %) and PC2 (here: approx. 17.3 %). As in (**a**), the PCA and resulting scatter plot indicate that biological replicates group well together, implying that within this example study, the influence of treatment and genotype on gene expression is well distinguishable from the biological background noise.

The user chooses either a group of expression values for biological replicates belonging to the same experimental condition (x-axis factor; see Section 2.1), or any arbitrary numerical information provided in the optional information table. This can be the logarithmic expression fold change as typically shown in pathway diagrams comparing two experimental conditions, e.g., control versus cold stress. Instead of using logarithmic fold change values as a measure of the intensity of a transcriptional response under different experimental conditions, the user can also choose adjusted *p*-values produced by differential gene expression analyses. Finally, the user can choose how the color gradient is dispersed over the selected numerical values. The choice is between a divergent scale from a fixed negative value to the positive counterpart (as in the MapMan desktop application which focuses on log fold change data), or a continuous scale ranging from zero, or the first quartile, to the third quartile. The MapMan web browser plots are interactive, hovering with the mouse over a specific box displays the gene identifier, the Mapman4 protein description, and the numerical information assigned to the gene.

2.5. Overrepresentation (Enrichment) Analysis

To qualitatively describe a biological response, often, annotations about biological processes and molecular functions are analyzed. Those annotations found to be significantly overrepresented among selected genes or metabolites of interest characterize that group of genes. The selection criteria can, e.g., be significant up- or down-regulation of gene expression contrasting two experimental conditions, e.g., control versus cold stress treatment. Typically, Fisher's exact test is used to determine whether the number of annotations within the selected genes significantly deviates from the number of annotations found within the background, i.e., the whole genome or metabolome. GXP offers the user an easy way to carry out such overrepresentation analysis (ORAs). The user specifies a criterion shared by all selected quantified entities, i.e., either transcripts or metabolites, and additionally selects annotation terms for which overrepresentation should be tested. Alternatively, the user can select the transcripts or metabolites of interest manually by entering their respective identifiers. It is noteworthy that GXP is agnostic to the underlying data structure, consequently the user can use any information originally loaded with the "information table" (Section 2.1). In principle, ORAs can be carried out for metabolite data, even though these analyses are less common for targeted metabolomics studies. Consequently, a great variety of enrichment analyses can be carried out. In this, each single calculation of Fisher's exact test produces a corresponding single *p*-value, i.e., one *p*-value for each tested annotation. Currently, these *p*-values are corrected for multiple hypothesis testing using Bonferroni's or the Benjamini-Hochberg method. All calculations are carried out in the background, so that user experience is not interrupted. The final result is a table in which for each tested annotation the corresponding adjusted *p*-value is shown (Figure 5). These results can be exported along with the data and plots by clicking the "Export GXP Database" button in the "Data" submenu (see Section 2.1).

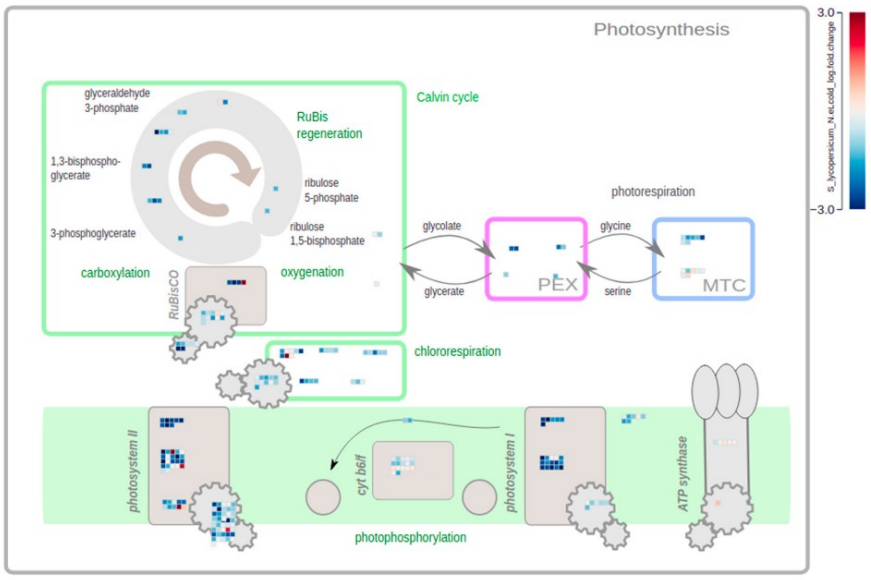


MIN	Q1	MEDIAN	MEAN	Q3	MAX
-12.12	-0.81	-0.10	-0.24	0.34	11.60

[5] Adjust Box Size

Distribution of S_lycopersicum_N.eLcold_log.fold.change

(**a**)



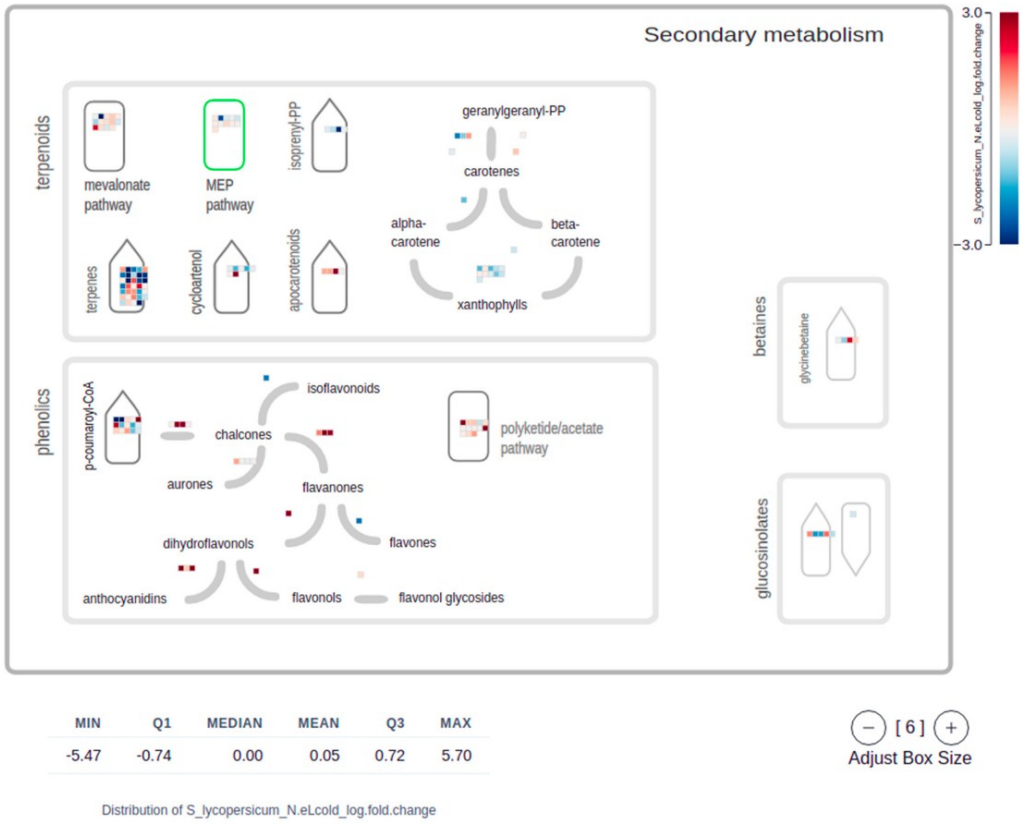
MIN	Q1	MEDIAN	MEAN	Q3	MAX
-8.22	-1.87	-1.13	-1.13	-0.29	6.70

[6] Adjust Box Size

Distribution of S_lycopersicum_N.eLcold_log.fold.change

(**b**)

Figure 4. *Cont.*



(**c**)

Figure 4. Screenshots of Mapman plots [1,2] used to visualize the genetic response to experimental stimuli in the form of metabolic sketches. Genes are mapped to areas in the sketches according to their molecular function. This gene function is directly extracted from the respective Mapman Bins [36] the genes are assigned to [3]. Each gene is represented by a single-colored box, where the color represents a numeric value, in this example the logarithmic fold change of gene expression (log-FC) between control and stress treatment. A legend in the top-right corner informs about the color-scale used to represent these numeric values. At the bottom of each Mapman plot, a summary statistic informs the user about the distribution of the respective numerical values, here the log-FC, shown in the plot. An interactive control in the bottom-right corner allows the user to adjust the sizes of the boxes, each representing one gene. Plot (**a**) shows a metabolic overview sketch and highlights how in the example data the expression of genes associated with photosynthesis is down regulated in *S. lycopersicum* following stress treatments (blue boxes in the respective top-right corner matrices). This down-regulation particularly affects genes of the light reaction, calvin cycle, and photorespiration pathways. Plot (**b**) sheds more light on this genetic response and zooms into the effect of stress treatments on the expression of genes associated with Photosystem I and II. Another detailed representation of the observed genetic response to stress treatment is shown in plot (**c**), elucidating how the expression of genes involved in terpene and carotene synthesis is down-regulated in *S. lycopersicum*.

2.6. Usage of GXP to Publish Data along with Plots and Analysis Results

As previously mentioned, a user can save a GXP work-session by exporting all data, plots, and analysis results into a downloadable "GXP Database". Such a GXP Database file can be made publicly available, e.g., by uploading it to a web-server such as GitHub or by

providing it in the form of a supplement to a scientific article. Readers can thus download the published GXP Database, load it into GXP and explore the data, plots, and analysis results restored from the original work session.

Another, more luxurious, mode of publishing data, plots, and analysis results along with GXP is included in the GXP manual. A user can deploy a copy of GXP together with an exported database to a dedicated webserver. GitHub-pages offers this option free of charge. This creates a link that can be cited in upcoming publications, does not require any maintenance work, and will be functioning as long as GitHub is maintained. A comprehensive (seven steps) step-by-step guide to set up such a tailored copy of GXP with specific user data, plots, and analysis results has been included in the online manual.

Thus, by using either of the above two methods, not only the raw expression counts and differential expression analysis result data could be provided, but also pregenerated supplemental plots highlighting the scientific results could be discussed in publications. This makes the data free to be explored by third parties in their own context of interest, possibly reaching beyond the scope of the publication.

Figure 5. Screenshot of the result of an enrichment analysis (EA) carried out on the example data. This analysis is available in the "Tools" menu (screwdriver and wrench icon in the top panel). In the lightly faded-out background, the carried-out enrichment analysis can be seen. If more such analyses were done by clicking on the round plus icon in the bottom-right corner, they would also appear in this list. Clicking on the respective analysis opens the table shown in the highlighted foreground overlay. In it, the user is presented with significant results, while the button in the bottom-right corner "Show all Entries" enables the inspection of all, not only the annotations significantly tested for overrepresentation. In the shown example, the EA identified molecular gene functions overrepresented among genes whose expression is down-regulated in *S. lycopersicum* in response to the applied stress treatments, a combination of nitrogen deficiency (N-), chilling temperatures (cold), and elevated light intensity (eL). In this case, the results support the observation made for the example data earlier in Figure 4a,b, i.e. the response to stress treatments in the form of down-regulation of genes associated with (i) photosynthesis and (ii) terpene and carotene biosynthesis. Among the down-regulated genes, the molecular functions (i) "Chlorophyll a-b binding protein" in the "LHC-II complex" (Mapman Bin 1.1.1.1.1) and (ii) "UmamiT solute transporter" (Mapman Bin 24.2.1.5), a "sesquiterpene synthase", and "diterpene synthase" (Mapman Bin 9.1.4.2) are significantly overrepresented (all adjusted p-values < 0.006). Thus the Mapman plots and enrichment analyses truly help to elucidate the genetic response in *S. lycopersicum* to the stress treatments applied in the example study.

3. Discussion

The availability, efficiency, and relatively low cost of next-generation sequencing and metabolomics technologies allows their application in a wide variety of plant science research projects. Quantification of gene expression or metabolites and contrasting these quantifications between different experimental conditions is implemented in many standard pipelines. However, the need for simple visualization, summarization, and further selected analyses revealing similarity between biological replicates and overrepresented molecular functions in sets of selected transcripts or metabolites is key for biological interpretation of these datasets. We revised platforms and software solutions that have been developed to provide the user with tools to quantify RNAseq raw data and contrast this quantified data in differential expression analysis. The revision includes tools that generate scientific plots, carry out clustering, principal component, and overrepresentation analysis [1,3,6–28,31,32,35,41]. However, the presented tools require either some programming expertise, or manual installation of software, or send potentially confidential data via the web to dedicated servers. Spreadsheet applications are often used to partially fill this gap, but the resulting plots are not interactive, and spreadsheet programs do not easily allow clustering, principal component, or overexpression analyses. In this context, we introduced Gene Expression Plotter (GXP) that provides the user with the means to load, analyze, and visualize quantitative and qualitative Omics data in the browser without the need for programming expertise or software installation. Additionally, GXP does all calculations locally in the browser without any need to submit data to servers. We incorporated into GXP the first mature and no-installation-required version of the popular Mapman tool [1,2]. This enables the user to summarize, in high quality plots, the gene functions, particularly up- and down-regulation, in a genetic response to experimental treatments, e.g., contrasting control and cold stress treatments. Hence, GXP offers simple solutions to explore and analyze Omics data and to generate publication grade plots. We furthermore explained how GXP can be used to publish Omics data along with plots and analysis results either by simply providing the community with a ready to use GXP database file, e.g., in the form of an article supplement, or by setting up an online copy of GXP already including the mentioned data, plots, and results. In this, the latter method can be done directly on GitHub free of charge by following seven simple steps that only require a web browser and a GitHub account.

To help the reader and user explore the value of using GXP and to benchmark GXP's functions, we included real research data from a published study on stress response in two tomato species [38]. We verified with the aid of one of the authors of this original study that (i) GXP reproduces and visualizes the already published findings, (ii) aids in the exploration of Omics data and promotes the formation of scientific hypotheses (Figures 1–5), and (iii) thus helps to elucidate, e.g., the genetic response to experimental stimuli, i.e., the original biological question motivating the study.

GXP is open-source software, runs entirely in the web browser and all code is automatically unit-tested, and thus is ensured to carry out correct calculations. Additionally, all code has been written adhering to current cutting-edge coding standards. Importantly, GXP is versatile in terms of its input. GXP consumes data about quantified entities, typically transcripts or chemical compounds (metabolites). Optionally, the user can supply further arbitrary either free text, categorical, and/or numerical information about the quantified entities and use, especially the latter two types of input data in GXP's plots and analyses. This generic approach to quantitative data visualization, exploration, and analysis, along with the option to easily setup a copy of GXP with specific user data, plots, and analysis results, specifically qualifies the GXP code base for reuse and extension in the typical open-source community approach. We indeed believe that GXP can become a platform for which over time, more and more functions, plots, and analyzes can be provided by third party developers.

4. Materials and Methods

Gene-Expression-Plotter (GXP) was implemented in TypeScript (version 4.5.2; https://www.typescriptlang.org/; access date 11 January 2022) as a standalone application (single page application; SPA) executed in the web-browser. This form of implementation implies that even though the SPA is obtained from a webserver, after opening it in the browser no data is ever sent to any webserver for analysis. All calculations, plots, and analysis are carried out right on the user's computer in the browser itself; thus, data confidentiality is guaranteed. For implementation, the ReactJS (version 17.0.1; https://reactjs.org/; access date 11 January 2022) and Chakra UI (version 1.7.2; https://chakra-ui.com/; access date 11 January 2022) libraries were used to build the user interface. The ViteJS library (version 1.3.6; https://vitejs.dev/; access date 11 January 2022) was used for tooling. GXP can be accessed on GitHub pages (https://usadellab.github.io/GeneExpressionPlots; access date 11 January 2022); every time a new version is pushed to GitHub, the new code is compiled and automatically deployed to GitHub pages using GitHub actions. GXP's source code is freely available on GitHub (https://github.com/usadellab/GeneExpressionPlots; access date 11 January 2022) under a GPL-3 license.

4.1. Input and Output Data

Expression or metabolite data, and additional information about transcripts or metabolites, can be loaded into GXP (see Section 2.1). Alternatively, a previous work-session can be restored using the "Import GXP Database" function in the "Data" menu (see Section 2.1). All data is stored in memory, no data is ever sent via the internet to any backend server. Memory state management has been implemented using the MobX library (version 6.3.7; https://mobx.js.org; access date 11 January 2022).

4.2. Gene Expression Plots

All introduced plots (see Section 2.2) were implemented with the plotly.js Javascript library (version 2.6.3; https://plotly.com/javascript/; access date 11 January 2022). Plot data and definitions are stored in memory and thus can be exported to and restored from GXP Databases (see Section 2.1.2).

4.3. Hierarchical Cluster Analysis

Similarity between biological replicates is either assessed using correlation or Euclidean distance between the respective gene expression vectors (see Section 2.3.1). In this, correlation values $c_{i,k}$ are transformed to distance values $d_{i,k}$ as follows:

$$d_{i,k} = 1 - \text{abs}(c_{i,k}),$$

with "abs" returning the absolute value of its real number argument.

Thus, complete anticorrelation as well as complete correlation are interpreted as maximum likeliness of numeric quantification vectors.

Euclidean distance measures are computed with the ml-distance Javascript library (version 3.0.0; https://github.com/mljs; access date 11 January 2022). Hierarchical clusters are identified using the ml-hclust library (version 3.1.0; https://github.com/mljs; access date 11 January 2022). The heatmap and the respective dendrogram visualizing the results of hierarchical clustering are plotted with the visx (version 2.4.0; https://github.com/airbnb/visx; access date 11 January 2022) library that incorporates the popular and well proven D3 library (version 7.1.1; https://d3js.org/; access date 11 January 2022into React.js.

4.4. Principal Component Analysis (PCA)

In GXP, a PCA can be carried out to identify and visualize likeliness between gene expression or metabolite concentrations of biological replicates (see Section 2.3.2). The principal component analysis is computed with the help of the ml-pca library (version 4.0.2; https://github.com/mljs; access date 11 January 2022). In this calculation, all data points are

considered. The respective plot visualizing the first two principal components contributing most to the observed differences is created with plotly.js (version 2.6.3; https://plotly.com/javascript/; access date 11 January 2022).

4.5. MapMan Visualizations

GXP offers to summarize genetic expression or responses to contrasting experimental conditions in the form of Mapman plots [1] (see Section 2.4). All available canvas sketches (version X4.3) upon which to draw boxes to represent transcripts' quantification values were downloaded from the respective "MapMan Store" online repository (version X4.3; https://mapman.gabipd.org/mapmanstore; access date 11 January 2022) and included in the GXP package. Based on the proof-of-concept implementation [40], the visualization was programmed with the D3 library (version 7.1.1; https://github.com/d3/d3; access date 11 January 2022).

4.6. Overrepresentation Analysis

Gene Expression Plotter offers the user the ability to define sets of transcripts (genes) or sets of metabolites of interest, either by stating the respective identifiers one-by-one or by defining a selection criterion (see Section 2.5). Subsequently, annotations assigned to the selected genes are tested for being over-presented in comparison to the background, which is the whole information table, i.e., the genome or the metabolome (see Section 2.1). Each of these tests is carried out as Fisher's exact test resulting in a single p-value indicating how likely the observed annotation numbers can be explained by the null hypothesis, i.e., variations of the background annotations. In Fisher's exact test contingency tables are created and p-values calculated using the hypergeometric probability distribution (HGD) [42]. The calculation of specific HGD p-values is carried out with the GNU scientific library (version 2.6; git://git.savannah.gnu.org/gsl.git; accessed on 11 January 2022) [43] which was compiled to web-assembly (version 1.0; https://webassembly.org/; access date 11 January 2022) for usage in the web-browser with Javascript. This compilation was done with emscripten (version 2.0.25; https://emscripten.org/; access date 11 January 2022). To calculate the likelihood of the alternative hypothesis that the observed numbers of annotations are not just as is in the contingency table but potentially greater, i.e., more extreme, more contingency tables of more extreme distributions are created and tested, respectively. Resulting p-values are summed up until no more extreme contingency tables can be generated, i.e., the respective cells contain zero. This procedure has been implemented in Javascript and correctness of the calculations is confirmed by dedicated automatic unit software tests.

4.7. Example Dataset

To demonstrate GXP's qualities, published data sets from two tomato species were used [38]. In brief, two tomato species (*S. lycopersicum* and *S. pennellii*) were grown in rockwool blocks and watered with water for 16 days. Afterwards, the seedlings were fertilized with half-strength Hoagland solution for 14 days, followed by full-strength Hoagland solution (5 mM KNO3, 5 mM Ca(NO3)2, 2 mM MgSO4, 1 mM KH2PO4, 90 µM FeEDTA, plus micronutrients) for a further 11 days. A total of 6 weeks after germination, plants were stressed by nitrogen deficiency (N-), chilling temperatures (cold), warmer temperature regime (warm), or elevated light intensity (eL) and combinations thereof (Ncold, N-eL, eLcold and N-eLcold). After 1 week of stress treatment, leaflets of the fourth leaf (counted from the tip) were sampled, immediately frozen in liquid nitrogen and stored at −80 °C. Total RNA was extracted and treated with DNase followed by mRNA enrichment, and subsequently analyzed using an Illumina-platform (HiSeq) sequencing 2×75 bp paired-end reads.

Raw reads were trimmed using Trimmomatic [44]. An artificial transcriptome was built using default settings of StringTie [45,46] and back mapped to the genome of *S. lycopersicum*

(version ITAG 4). Data analysis was performed using R (version 3.5.2) [29]. Read abundances were analyzed using R-packages limma [47], edgeR [48], and tximport [49].

4.8. Automated Software Tests Ensure Correctness of Implemented Analyses

The code written to carry out the z-transformation, correlation, clustering, principal component, Fisher's exact test, overrepresentation, and *p*-value adjustment calculations provided within GXP are all verified for correctness using automated software, so called unit tests. These tests use data obtained from real life science projects and ensure that the respective functions behave correctly even in "edge cases" where data is unexpectedly abnorm, e.g., empty. These tests can be run automatically and thus ensure correctness of calculations even if future extensions are programmed. All tests are located in the cypress/tests/integration directory in GXP's GitHub code repository.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/plants11060745/s1, Text S1: Review of comparable tools to visualize and analyze RNAseq or Metabolomics quantitative data, Text S2: Table of used software packages and frameworks.

Author Contributions: A.H. and B.U. conceptualized the project. A.H. headed the software design of the Gene Expression Plotter (GXP) browser-based application. J.A., C.E., D.V. and D.W. programmed the application software and carried out software tests. C.E. and D.V. designed the graphical interface. B.U. provided scientific guidance, especially of methods to be applied, and feedback particularly about the user interface. B.U. and R.S. implemented a proof-of-concept Javascript software implementation of the MapMan visualizations. A.S. provided an extensive review in the form of testing and editing the GXP user manual, and delivered detailed user feedback. V.W. iteratively used and tested GXP in a RNAseq project, contributed to feature design, and provided extensive user feedback. J.J.R. provided preprocessed biological data. R.P., S.F. and U.S. provided project administration, integration and application of GXP into ongoing scientific research projects and user feedback. A.H., J.A., C.E. and B.U. wrote the manuscript with the help of all authors. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the German Ministry of Education and Research BMBF, grant agreement No BreedPath 031B0890B and the European Commission for the project EPPN2020 under grant agreement No. 731013, the project EOSC-Life under the grant agreement No. 824087, and the project EMPHASIS-PREP under the grant agreement No. 739514.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All code, example and test data is available at https://github.com/usadellab/GeneExpressionPlots (accession date 11 January 2022).

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

References

1. Bolger, M.; Schwacke, R.; Usadel, B. MapMan Visualization of RNA-Seq Data Using Mercator4 Functional Annotations. *Methods Mol. Biol.* **2021**, *2354*, 195–212. [PubMed]
2. Usadel, B.; Poree, F.; Nagel, A.; Lohse, M.; Czedik-Eysenberg, A.; Stitt, M. A guide to using MapMan to visualize and compare Omics data in plants: A case study in the crop species, Maize. *Plant Cell Environ.* **2009**, *32*, 1211–1229. [CrossRef] [PubMed]
3. Lohse, M.; Nagel, A.; Herter, T.; May, P.; Schroda, M.; Zrenner, R.; Tohge, T.; Fernie, A.R.; Stitt, M.; Usadel, B. Mercator: A fast and simple web server for genome scale functional annotation of plant sequence data. *Plant Cell Environ.* **2014**, *37*, 1250–1258. [CrossRef] [PubMed]
4. The InterPro Consortium; Mulder, N.J.; Apweiler, R.; Attwood, T.; Bairoch, A.; Bateman, A.; Binns, D.; Biswas, M.; Bradley, P.; Bork, P.; et al. InterPro: An integrated documentation resource for protein families, domains and functional sites. *Brief. Bioinform.* **2002**, *3*, 225–235. [CrossRef]
5. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* **2000**, *25*, 25–29. [CrossRef]
6. Pimentel, H.; Bray, N.L.; Puente, S.; Melsted, P.; Pachter, L. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat. Methods* **2017**, *14*, 687–690. [CrossRef]

7. Su, W.; Sun, J.; Shimizu, K.; Kadota, K. TCC-GUI: A Shiny-based application for differential expression analysis of RNA-Seq count data. *BMC Res. Notes* **2019**, *12*, 133. [CrossRef]
8. Choi, K.; Ratner, N. iGEAK: An interactive gene expression analysis kit for seamless workflow using the R/shiny platform. *BMC Genom.* **2019**, *20*, 177. [CrossRef]
9. Sundararajan, Z.; Knoll, R.; Hombach, P.; Becker, M.; Schultze, J.L.; Ulas, T. Shiny-Seq: Advanced guided transcriptome analysis. *BMC Res. Notes* **2019**, *12*, 432. [CrossRef]
10. Marini, F.; Binder, H. pcaExplorer: An R/Bioconductor package for interacting with RNA-seq principal components. *BMC Bioinform.* **2019**, *20*, 331. [CrossRef]
11. Wang, S.; Zhang, Y.; Hu, C.; Zhang, N.; Gribskov, M.; Yang, H. Shiny-DEG: A Web Application to Analyze and Visualize Differentially Expressed Genes in RNA-seq. *Interdiscip Sci.* **2020**, *12*, 349–354. [CrossRef]
12. Reyes, A.L.P.; Silva, T.C.; Coetzee, S.G.; Plummer, J.T.; Davis, B.D.; Chen, S.; Hazelett, D.J.; Lawrenson, K.; Berman, B.P.; Gayther, S.A.; et al. GENAVi: A shiny web application for gene expression normalization, analysis and visualization. *BMC Genom.* **2019**, *20*, 745. [CrossRef]
13. Haering, M.; Habermann, B.H. RNfuzzyApp: An R shiny RNA-seq data analysis app for visualisation, differential expression analysis, time-series clustering and enrichment analysis. *F1000Research* **2021**, *10*, 654. [CrossRef]
14. Kim, S.C.; Yu, D.; Cho, S.B. COEX-Seq: Convert a Variety of Measurements of Gene Expression in RNA-Seq. *Genom. Inform.* **2018**, *16*, e36. [CrossRef]
15. Zhang, C.; Fan, C.; Gan, J.; Zhu, P.; Kong, L.; Li, C. iSeq: Web-Based RNA-seq Data Analysis and Visualization. *Methods Mol. Biol.* **2018**, *1754*, 167–181.
16. Li, R.; Hu, K.; Liu, H.; Green, M.R.; Zhu, L.J. OneStopRNAseq: A Web Application for Comprehensive and Efficient Analyses of RNA-Seq Data. *Genes* **2020**, *11*, 1165. [CrossRef]
17. Hoek, A.; Maibach, K.; Özmen, E.; Vazquez-Armendariz, A.I.; Mengel, J.P.; Hain, T.; Herold, S.; Goesmann, A. WASP: A versatile, web-accessible single cell RNA-Seq processing platform. *BMC Genom.* **2021**, *22*, 195. [CrossRef]
18. Harshbarger, J.; Kratz, A.; Carninci, P. DEIVA: A web application for interactive visual analysis of differential gene expression profiles. *BMC Genom.* **2017**, *18*, 47. [CrossRef]
19. Nelson, J.W.; Sklenar, J.; Barnes, A.P.; Minnier, J. The START App: A web-based RNAseq analysis and visualization resource. *Bioinformatics* **2017**, *33*, 447–449. [CrossRef]
20. Li, Y.; Andrade, J. DEApp: An interactive web interface for differential expression analysis of next generation sequence data. *Source Code Biol. Med.* **2017**, *12*, 2. [CrossRef]
21. Russo, F.; Angelini, C. RNASeqGUI: A GUI for analysing RNA-Seq data. *Bioinformatics* **2014**, *30*, 2514–2516. [CrossRef]
22. Bray, N.L.; Pimentel, H.; Melsted, P.; Pachter, L. Near-Optimal probabilistic RNA-seq quantification. *Nat Biotechnol.* **2016**, *34*, 525–527. [CrossRef]
23. Langmead, B.; Trapnell, C.; Pop, M.; Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **2009**, *10*, R25. [CrossRef]
24. Kim, D.; Langmead, B.; Salzberg, S.L. HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* **2015**, *12*, 357–360. [CrossRef]
25. Patro, R.; Duggal, G.; Love, M.I.; Irizarry, R.A.; Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **2017**, *14*, 417–419. [CrossRef]
26. Howe, E.; Holton, K.; Nair, S.; Schlauch, D.; Sinha, R.; Quackenbush, J. MeV: MultiExperiment Viewer. In *Biomed. Informatics for Cancer Research*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 267–277. [CrossRef]
27. Howe, E.; Holton, K.; Nair, S.; Schlauch, D.; Sinha, R.; Quackenbush, J. WebMeV: MultiExperiment Viewer. Available online: https://webmev.tm4.org/#/about (accessed on 11 January 2022).
28. Su, S.; Law, C.W.; Ah-Cann, C.; Asselin-Labat, M.-L.; Blewitt, M.E.; Ritchie, M.E. Glimma: Interactive graphics for gene expression analysis. *Bioinformatics* **2017**, *33*, 2050–2052. [CrossRef]
29. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. 2021. Available online: https://www.R-project.org/ (accessed on 11 January 2022).
30. R Studio Inc. Easy Web Applications in R. 2013. Available online: https://www.rstudio.com/shiny/ (accessed on 11 January 2022).
31. Hadizadeh Esfahani, A.; Maß, J.; Hallab, A.; Schuldt, B.M.; Nevarez, D.; Usadel, B.; Ott, M.C.; Buer, B.; Schuppert, A. Plant PhysioSpace: A robust tool to compare stress response across plant species. *Plant Physiol.* **2021**, *187*, 1795–1811. [CrossRef]
32. Hernández-De-Diego, R.; Tarazona, S.; Martínez-Mira, C.; Balzano-Nogueira, L.; Furió-Tarí, P.; Pappas, G.J., Jr.; Conesa, A. PaintOmics 3: A web resource for the pathway analysis and visualization of multi-omics data. *Nucleic Acids Res.* **2018**, *46*, W503–W509. [CrossRef] [PubMed]
33. Naithani, S.; Gupta, P.; Preece, J.; D'Eustachio, P.; Elser, J.L.; Garg, P.; Dikeman, D.A.; Kiff, J.; Cook, J.; Olson, A.; et al. Plant Reactome: A knowledgebase and resource for comparative pathway analysis. *Nucleic Acids Res.* **2020**, *48*, D1093–D1103. [CrossRef] [PubMed]
34. Waese, J.; Fan, J.; Pasha, A.; Yu, H.; Fucile, G.; Shi, R.; Cumming, M.; Kelley, L.A.; Sternberg, M.J.; Krishnakumar, V.; et al. ePlant: Visualizing and Exploring Multiple Levels of Data for Hypothesis Generation in Plant Biology. *Plant Cell.* **2017**, *29*, 1806–1821. [CrossRef] [PubMed]

35. Julkowska, M.M.; Saade, S.; Agarwal, G.; Gao, G.; Pailles, Y.; Morton, M.; Awlia, M.; Tester, M. MVApp—Multivariate Analysis Application for Streamlined Data Analysis and Curation. *Plant Physiol.* **2019**, *180*, 1261–1276. [CrossRef]
36. Schwacke, R.; Soto, G.Y.P.; Krause, K.; Bolger, A.M.; Arsova, B.; Hallab, A.; Gruden, K.; Stitt, M.; Bolger, M.; Usadel, B. MapMan4: A Refined Protein Classification and Annotation Framework Applicable to Multi-Omics Data Analysis. *Mol. Plant* **2019**, *12*, 879–892. [CrossRef]
37. Goto, M.K.A. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30.
38. Reimer, J.J.; Thiele, B.; Biermann, R.T.; Junker-Frohn, L.V.; Wiese-Klinkenberg, A.; Usadel, B.; Wormit, A. Tomato leaves under stress: A comparison of stress response to mild abiotic stress between a cultivated and a wild tomato species. *Plant Mol. Biol.* **2021**, *107*, 177–206. [CrossRef]
39. Thimm, O.; Bläsing, O.; Gibon, Y.; Nagel, A.; Meyer, S.; Krüger, P.; Selbig, J.; Müller, L.A.; Rhee, S.Y.; Stitt, M. MAPMAN: A user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* **2004**, *37*, 914–939. [CrossRef]
40. Usadel, B. MapManJS—Pure Web Implementations of MapMan. 2018. Available online: https://github.com/usadellab/MapManJS (accessed on 11 January 2022).
41. Lohse, M.; Bolger, A.M.; Nagel, A.; Fernie, A.R.; Lunn, J.E.; Stitt, M.; Usadel, B. RobiNA: A user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Res.* **2012**, *40*, W622–W627. [CrossRef]
42. Fisher, R.A. On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P. *J. R. Stat. Soc.* **1922**, *85*, 87. [CrossRef]
43. The Gnu Scientific Library Team. *Gnu Scientific Library 2.0*; Samurai Media Limited: Surrey, UK, 2015.
44. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef]
45. Pertea, M.; Pertea, G.M.; Antonescu, C.M.; Chang, T.-C.; Mendell, J.T.; Salzberg, S.L. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **2015**, *33*, 290–295. [CrossRef]
46. Pertea, M.; Kim, D.; Pertea, G.M.; Leek, J.T.; Salzberg, S.L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **2016**, *11*, 1650–1667. [CrossRef]
47. Ritchie, M.E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C.W.; Shi, W.; Smyth, G.K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **2015**, *43*, e47. [CrossRef]
48. Robinson, M.D.; McCarthy, D.J.; Smyth, G.K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **2010**, *26*, 139–140. [CrossRef]
49. Soneson, C.; Love, M.I.; Robinson, M.D. Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences. *F1000Research* **2015**, *4*, 1521. [CrossRef]

Supplementary Materials

Supplemental Text S1: Review Of Comparable Tools To Visualize And Analyze Rnaseq Or Metabolomics Quantitative Data

In this document we review the tools presented in the introduction section in order to explain the motivation that led to the development of Gene Expression Plotter (GXP) in the context of comparison with the mentioned tools. We briefly summarize their features, requirements, and technical traits. In this comparison, we focus on features that are key in the context of this article. Each tool has its section headed by the tool's name.

Abbreviations

ANOVA:	Analysis of variance
DE:	Differential Expression
DEG:	Differentially Expressed Gene
DEGES:	DEG Elimination Strategy
FC:	Fold-Changes
FDR:	False Discovery Rate
GSEA:	Gene Set Enrichment Analysis
GUI:	Graphical User Interface
KEGG:	Kyoto Encyclopedia of Genes and Genomes
MDS:	Multi-dimensional Scaling
ORA:	Over-Representation Analysis
PCs:	Principal Components
PCA:	Principal Component Analysis
PPI:	Protein-Protein Interaction
scRNA-Seq:	Single cell RNA sequencing
RLE:	Relative log expression
TCC:	Tag Count Comparison
TMM:	Trimmed mean of M values

PaintOmics3

- Hernández-de-Diego R, Tarazona S, Martínez-Mira C, Balzano-Nogueira L, Furió-Tarí P, Pappas GJ Jr, et al. PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data. Nucleic Acids Res. 2018;46: W503–W509.

General

Requires installation?	No
Sends data through the web?	Yes
Can it be used to publish data?	An option to share results exists
Coding required?	No

Specific

- Specialized in visualization of omic data onto KEGG pathways.
- “Paintomics takes complete transcriptomics and metabolomics datasets, together with lists of the significant gene or metabolite changes, and paints this information on KEGG pathway maps.”

PlantReactome

- Naithani S, Gupta P, Preece J, D’Eustachio P, Elser JL, Garg P, et al. Plant Reactome: a knowledgebase and resource for comparative pathway analysis. Nucleic Acids Res. 2020;48: D1093–D1103

General

Requires installation?	No
Sends data through the web?	Yes
Can it be used to publish data?	An option to share results exists
Coding required?	No

Specific

- Uses knowledge of gene-orthology to project omics data of any plant species onto rice known pathways.
- Takes in protein-protein interaction data as input.

ePlant

- Waese J, Fan J, Pasha A, Yu H, Fucile G, Shi R, et al. ePlant: Visualizing and Exploring Multiple Levels of Data for Hypothesis Generation in Plant Biology. Plant Cell. 2017;29: 1806–1821.

General

Requires installation?	No
Sends data through the web?	No - data cannot be uploaded
Can it be used to publish data?	An URL can be generated to work done
Coding required?	No

Specific

- Integrates interactome, transcriptome, and 3D molecular structure data published for *Arabidopsis thaliana.*
- Allows exploration of already integrated data through multiple ways.
- Allows for hypothesis formation and testing.
- Can be deployed on any server.

SLEUTH

- Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq incorporating quantification uncertainty. Nat Methods. 2017;14: 687–690.

General

Requires installation?	Yes
Sends data through the web?	No
Can it be used to publish data?	No
Coding required?	Yes

Specific

- R/Shiny application.
- Implements an alternative method for the identification of differentially expressed genes.
- Implements a ready to use RNA-Seq pipeline that generates expression counts and differential expression results.

MVapp

- Julkowska MM, Saade S, Agarwal G, Gao G, Pailles Y, Morton M, et al. MVApp—Multivariate Analysis Application for Streamlined Data Analysis and Curation. Plant Physiology. 2019. pp. 1261–1276. doi:10.1104/pp.19.00235

General

Requires installation?	No (optional for self-hosted)
Sends data through the web?	Yes, but does not store data on the server
Can it be used to publish data?	Yes
Coding required?	No

Specific

- Carries out multivariate analyses to reveal genotype-to-phenotype relationships.
- Can be used for interactive data curation, clustering, and quantile regression.
- Phenotypic trait modelling by curve fitting.
- Highlights potential outlier samples.
- Tests for significantly varying phenotypes (ANOVA, Kruskal-Walis)
- Plots: Heatmap/hierarchical clustering, PCA.
- Performs estimation of heritability.

GLIMMA

- Su S, Law CW, Ah-Cann C, Asselin-Labat M-L, Blewitt ME, Ritchie ME. Glimma: interactive graphics for gene expression analysis. Bioinformatics. 2017;33: 2050–2052.

General

Requires installation?	Yes
Sends data through the web?	No
Can it be used to publish data?	Yes
Coding required?	Yes

Specific

- R package extending popular plots generated for results obtained from differential gene expression identification.
- Plots: multidimensional scaling plots and mean-difference plots.
- Individual data points can be explored interactively by hovering over them in the generated plots. The expression of a single gene is then shown in a new plot side by side.

TCC-GUI

- Su W, Sun J, Shimizu K, Kadota K. TCC-GUI: a Shiny-based application for differential expression analysis of RNA-Seq count data. BMC Res Notes. 2019;12: 133.

General

Requires installation?	Yes
Sends data through the web?	No
Can it be used to publish data?	Maybe
Coding required?	No

Specific

- A GUI frontend to the TCC Bioconductor package for identification of differential expression.
- Test server available online.

- Performs DEGES (DEG elimination strategy) to iteratively normalize expression counts after removing top-ranked (differential expression) genes.
- Implements analysis, interactive plots: MDS, PCA, clustering.
- Generates interactive linked plots, e.g. volcano plot, hovering over a data point opens expression level barplot for that gene.
- Differences in expression between groups can be displayed using MA plots.

iGEAK

- Choi K, Ratner N. iGEAK: an interactive gene expression analysis kit for seamless workflow using the R/shiny platform. BMC Genomics. 2019;20: 177.

General

Requires installation?	Yes
Sends data through the web?	No
Can it be used to publish data?	Maybe
Coding required?	No

Specific

- R/Shiny desktop application.
- Carries out the identification of differential expression genes.
- Implements: PCA, correlation-based heatmap/clustering, gene expression boxplots (one box per sample/group), PPI based on pre-existing ENSEMBL data, ORA based on Reactome pathways.
- Plots: volcano plot.
- Specializes in usage for animals, as it uses Gene-Symbols and the Reactome pathways for published data-sets on human and mouse experiments.

Shiny-Seq

- Sundararajan Z, Knoll R, Hombach P, Becker M, Schultze JL, Ulas T. Shiny-Seq: advanced guided transcriptome analysis. BMC Res Notes. 2019;12: 432.

General

Requires installation?	No
Sends data through the web?	Yes
Can it be used to publish data?	Maybe
Coding required?	No

Specific

- R/Shiny GUI to pipeline analysis.
- Takes in Kallisto output or a count table.
- Performs normalization with DESeq2.
- Implements batch effect analysis and PCA plot is used to show the effect of surrogate variables.
- Implements differential gene expression analysis and co-expression network analysis.
- Gene set enrichment analysis (GSEA) for clusters can be done in the context of humans, mice, and yeast.
- Transcription factor binding side overrepresentation analysis in the context of human, mouse data.
- Generated plots include MA, PCA, Fold-Change x FC, volcano, expression plot with baseline.

pcaExplorer

- Marini F, Binder H. pcaExplorer: an R/Bioconductor package for interacting with RNA-seq principal components. BMC Bioinformatics. 2019;20: 331.

General

Requires installation?	Yes
Sends data through the web?	No
Can it be used to publish data?	Yes

Coding required?	No

Specific

- R/Shiny application.
- Implements PCA, normalization with DESeq2, heatmaps, box-plots, functional interpretation of principal components.
- Functional annotation based on Gene Ontology database can be done on the PCs.
- Box plots of data grouped based on experimental variables and selected genes of interest can be generated and explored.
- HTML reports can be generated at any state displaying code and output.

Shiny-DEG

- Wang S, Zhang Y, Hu C, Zhang N, Gribskov M, Yang H. Shiny-DEG: A Web Application to Analyze and Visualize Differentially Expressed Genes in RNA-seq. Interdiscip Sci. 2020;12: 349–354.

General

Requires installation?	Yes
Sends data through the web?	No
Can it be used to publish data?	Yes
Coding required?	No

Specific

- R/Shiny application.
- Supports multi-factor design models. DEG across all conditions can be identified.
- Plots: Boxplot of gene expression, Volcano plots showing the up and down-regulated genes, heatmap of cluster analysis (using Euclidean, Manhattan or Pearson correlation), PCA plot of DEG.

GENAVi

- Reyes ALP, Silva TC, Coetzee SG, Plummer JT, Davis BD, Chen S, et al. GENAVi: a shiny web application for gene expression normalization, analysis and visualization. BMC Genomics. 2019;20: 745.

General

Requires installation?	No (optional for self-hosted)
Sends data through the web?	Yes
Can it be used to publish data?	Yes
Coding required?	No

Specific

- R/Shiny application hosted on the Shiny server.
- Used for normalization and differential analysis of human or mouse feature count data.
- Supports a range of normalization strategies: t-score like, logCPM-edgeR, vst-DESeq2 and rlog-DESeq2.
- Pots: Clustered heatmap based on Pearson distance, PCA, volcano plot.
- DEA can be performed integratively using the DESeq2 workflow.
- Gene set enrichment analysis can be performed.

RNfuzzyApp

- Haering M, Habermann BH. RNfuzzyApp: an R shiny RNA-seq data analysis app for visualisation, differential expression analysis, time-series clustering and enrichment analysis. F1000Res. 2021;10: 654.

General

Requires installation?	Yes
Sends data through the web?	No

Can it be used to publish data?	Yes
Coding required?	No

Specific

- R/Shiny application.
- Supports normalization of data using DESeq2, TMM, RLE or edgeR's upper-quantile.
- Implements DEA using DESeq2, edgeR or bayseq.
- Pairwise comparisons of different conditions and time points can be made.
- The following Interactive plots can be generated: hierarchical cluster heatmaps, MA, Volcano plots and PCA.
- Performs enrichment analysis using gprofiler2.
- Implements RNA-seq Time series analysis.

COEX-Seq

- Kim SC, Yu D, Cho SB. COEX-Seq: Convert a Variety of Measurements of Gene Expression in RNA-Seq. Genomics Inform. 2018;16: e36.

General

Requires installation?	Yes
Sends data through the web?	No
Can it be used to publish data?	No
Coding required?	No

Specific

- R/Shiny application.
- Converts gene expression measurement data formats that are used in various bioinformatics tools for RNA-Seq analysis.
- Generates a summary report of the converted data in form of boxplots.

iSeq

- Zhang C, Fan C, Gan J, Zhu P, Kong L, Li C. iSeq: Web-Based RNA-seq Data Analysis and Visualization. Methods Mol Biol. 2018;1754: 167–181.

General

Requires installation?	No
Sends data through the web?	Yes
Can it be used to publish data?	Yes
Coding required?	No

Specific

- R/Shiny application.
- Supports data normalization by size factor (using DESeq) and quantile (using preprocessCore R package).
- DEA can be done using DESeq2.
- Functional enrichment is implemented using DAVID and GOseq R package.
- iSeq incorporates a link to DAVID official site where a list of previously obtained up or down-regulated genes can be used for functional enrichment.
- Plots: Box plot, PCA, Hierarchical Clustering, Volcano plots, Heatmap.

OnestopRNAseq

- Li R, Hu K, Liu H, Green MR, Zhu LJ. OneStopRNAseq: A Web Application for Comprehensive and Efficient Analyses of RNA-Seq Data. Genes. 2020;11. doi:10.3390/genes11101165

General

Requires installation?	No
Sends data through the web?	Yes

Can it be used to publish data?	Yes
Coding required?	No

Specific

- Web application hosted by Apache server.
- MySQL is used as the backend and PHP is used for the presentation layer.
- Implements analysis using a pipeline in Snakemake workflow.
- Performs RNA-seq analysis such as raw read quality check, differential analysis of gene expression, exon usage, alternative splicing, transposable element expression, allele-specific gene expression quantification, and gene set enrichment analysis
- RNA-seq data analyses based on human, mouse, yeast, fruit fly, zebrafish, and worm genomes as the tool provides their reference genomes and annotations for the user to select.
- The user can request for a genome of interest to be added if not present.
- Integrates DEBrowser and Shiny-seq for DEA and interactive investigation of DE data.

WASP

- Hoek A, Maibach K, Özmen E, Vazquez-Armendariz AI, Mengel JP, Hain T, et al. WASP: a versatile, web-accessible single-cell RNA-Seq processing platform. BMC Genomics. 2021;22: 195.

General

Requires installation?	Yes
Sends data through the web?	No
Can it be used to publish data?	Yes
Coding required?	No

Specific

- Snakemake automated preprocessing pipeline and R/Shiny application.

- Processes Drop-Seq-based scRNA-Seq data.
- Implements a pipeline in Snakemake workflow for data pre-processing. The results are fed to the R Shiny web application for gene expression matrix generations, post-processing and plot generation.
- The Shiny app component can also consume external data not generated by WASP.
- Designed for scRNA-seq data.

DEIVA

- Harshbarger J, Kratz A, Carninci P. DEIVA: a web application for interactive visual analysis of differential gene expression profiles. BMC Genomics. 2017;18: 47.

General

Requires installation?	Yes
Sends data through the web?	No
Can it be used to publish data?	Yes
Coding required?	No

Specific

- R/Shiny application.
- Single page web application.
- Consumes results of DEA statistical test results as input.
- Provide an interactive interface for the exploration of differential expression data.
- Summarizes statistical test results of DEA.
- Gene identification is implemented in an interactive approach.
- Generates MA plot and volcano plot.

RNASeqGUI

- Russo F, Angelini C. RNASeqGUI: a GUI for analysing RNA-Seq data. Bioinformatics. 2014;30: 2514–2516.

General

Requires installation?	Yes
Sends data through the web?	No
Can it be used to publish data?	Yes
Coding required?	No

Specific

- R-Tcl/Tk application.
- Can be used to explore bam files, counting mapped reads against an annotation file, identification of differential expression genes using DESeq, DESeq2, EdgeR, NOISeq, BaySeq.
- Plots: Fold change plots, false discovery rate (FDR) histograms, p-value histograms and volcano plot, heatmap and PCA plot.

START App

- Nelson JW, Sklenar J, Barnes AP, Minnier J. The START App: a web-based RNAseq analysis and visualization resource. Bioinformatics. 2017;33: 447–449.

General

Requires installation?	Yes
Sends data through the web?	No
Can it be used to publish data?	Yes
Coding required?	No

Specific

- R/Shiny application.
- Performs normalization and regression of raw gene expression count data.

- Plots: PCA plot, heatmap based on Euclidean distance, volcano plot, scatter, gene expression boxplot, analysis plot (using p-values and fold change).

DEApp

- Li Y, Andrade J. DEApp: an interactive web interface for differential expression analysis of next-generation sequence data. Source Code Biol Med. 2017;12: 2.

General

Requires installation?	No
Sends data through the web?	Yes
Can it be used to publish data?	Yes
Coding required?	No

Specific

- R/Shiny application.
- Takes in raw count table (output of transcript quantification tools) and meta-data table.
- Genetic features with low counts can be filtered out within the tool's data panel.
- Identifies differentially expressed genes using edgeR, DESeq-2 or limma-voom.
- Outputs of the different methods can be compared directly from within the interface.
- Plots: multidimensional scaling (MDS) plot, volcano plot, dispersion plot.

Plant Physiospace

- Hadizadeh Esfahani A, Maß J, Hallab A, Schuldt BM, Nevarez D, Usadel B, et al. Plant PhysioSpace: a robust tool to compare stress response across plant species. Plant Physiol. 2021;187: 1795–1811.

General

Requires installation?	Yes
Sends data through the web?	No

Can it be used to publish data?	No
Coding required?	depends

Specific

- R package provided along with a GUI written in R/Shiny
- A specialized algorithm compares changes in gene expression with curated signatures obtained from earlier published studies. Alike signatures are robustly identified and thus similar genetic responses to potentially very different stimuli or treatments can be identified. The tool enables the user also to compare two or more of her/his own differential expression profiles stemming from separate contrasted treatments.
- To get full access to the implemented functions a user needs to do some manual coding in R.
- The core functionality is exposed through a R/Shiny interface

MeV / WebMeV

- Howe E, Holton K, Nair S, Schlauch D, Sinha R, Quackenbush J. MeV: MultiExperiment Viewer. Biomedical Informatics for Cancer Research. 2010. pp. 267–277. doi:10.1007/978-1-4419-5714-6_15
- Howe E, Holton K, Nair S, Schlauch D, Sinha R, Quackenbush J. WebMeV: MultiExperiment Viewer. Available: https://webmev.tm4.org/#/about

General

Requires installation?	Yes (MeV), No (WebMeV)
Sends data through the web?	No (MeV), Yes (WebMeV)
Can it be used to publish data?	No (MeV), Yes (WebMeV)
Coding required?	No

Specific

- MeV, multi experiment viewer, is a software suite to analyze large genomic data, with a particular focus on RNAseq.
- For those, who do not want to install it locally, there is a web-frontend available (WebMeV) which runs the analyses in a dedicated cloud.
- Implemented functions enable the user to load quantified expression data, provide metadata to assign biological replicates to experimental conditions or treatments, carry out differential expression analysis, cluster the biological replicates by several algorithms, including principal component analysis, and identify overrepresented gene functions (Gene Ontology terms), carry out a metabolic pathway analysis and identify modules of co-expressed genes.
- Gene expression can be visualized with boxplots and clustering results in heatmap plots.

Supplemental Text S2: Overview of Software Packages and Frameworks

This supplement provides a comprehensive table of the software packages, modules, and frameworks utilized in the development of the Gene-Expression-Plotter (GXP) and its associated functionalities. This summary enables readers to quickly identify and assess the software dependencies essential for implementing GXP.

Table S2.1: Software packages, modules, and frameworks used in the implementation of Gene-Expression-Plotter (GXP). Entries are organized in the order in which they appear in the main text for easy reference

Package, module, or framework	Version	Link
typescript	4.5.2	https://www.typescriptlang.org/
ReactJS	17.0.1	https://reactjs.org/
Chakra UI	1.7.2	https://chakra-ui.com/
ViteJS	1.3.6	https://vitejs.dev/
MobX 6.3.7	6.3.7	https://mobx.js.org/
plotly.js	2.6.3	https://plotly.com/javascript/

ml-distance	3.0.0	https://github.com/mljs/
ml-hclust	3.1.0	https://github.com/mljs/
visx	2.4.0	https://github.com/airbnb/visx/
D3	7.1.1	https://d3js.org/
ml-pca	4.0.2	https://github.com/mljs/
MapMan Store	X4.3	https://mapman.gabipd.org/mapmanstore/
GNU Scientific Library	2.6	git://git.savannah.gnu.org/gsl.git/
web-assembly	1.0	https://webassembly.org/
emscripten	2.0.25	https://emscripten.org/

CHAPTER 4

General Discussion

This thesis explores the genetic diversity, adaptation mechanisms, and evolutionary insights offered by teosinte. Teosinte's complex ecological and genetic profile, spanning multiple taxa across distinct geographic and climatic regions, makes it a valuable genetic reservoir for understanding both the evolutionary history and the domestication of maize. Through a combination of genome-wide association studies (GWAS), population genetics analyses, and ecological profiling, this research provides insights into the genetic foundations underlying key adaptive traits. A significant focus of this work is the role of genetic variation within teosinte populations in shaping traits associated with environmental adaptation. By characterizing and comparing the adaptive features of different teosinte taxa across diverse environments, this study reveals patterns of local adaptation and selective pressures that have shaped the evolutionary trajectory of these populations (Chapter 2).

This thesis also introduces the Gene Expression Plotter (GXP), an integrative tool developed to facilitate the analysis and visualization of omics data. GXP enhances the accessibility and interpretability of high-dimensional omics data by providing user-friendly, customizable visualization options, making it easier to pinpoint gene expression patterns linked to specific adaptive traits. By leveraging GXP, this research aimed at effectively analyzing transcriptome data that was to be generated from the teosinte project (Chapter 3).

The combination of these studies underscores the importance of understanding genetic diversity and adaptation within plant species and the necessity for efficient data interpretation tools to advance research and conservation strategies.

Phenotypic Variation and Its Genetic Basis

The phenotypic traits investigated, including but not limited to plant surface area, flowering time, reproductive organ dimensions, and kernel weight, are key for understanding how teosinte populations cope with environmental constraints. Correlations between plant surface area and humidity and the impact of altitude on yield emphasize the relationship between morphology and environmental adaptation. Through GWAS, candidate genes

associated with these traits were identified, shedding light on the genetic basis of traits such as flowering synchronization in variable photoperiods. This phenotypic variation not only provides a glimpse into teosinte's capacity to adapt but also identifies traits that could be potential targets for crop improvement in maize. For example, flowering time genes from teosinte could be useful for maize breeding in regions where seasonal synchronization of flowering with rainfall is critical for yield stability.

Genetic Diversity and Population Structure: Implications for Evolution and Conservation

The results confirm that genetic diversity in teosinte is highly structured and shaped by both species-specific and environment-specific selection pressures. Our analyses reveal clear distinctions among teosinte taxa, where genetic variation corresponds closely with ecological and phenotypic diversification across climatic zones. The admixture and principal component analyses (PCA) indicate that teosinte's genetic population structure is aligned with its taxonomy and strongly reflects environmental variables such as altitude, temperature, and humidity.

These findings are critical to conservation planning, as each teosinte population represents a unique evolutionary history and local adaptation. If lost, this diversity cannot simply be replicated or restored from other populations due to the specificity of adaptive traits. This is consistent with the Genetic Biodiversity Framework (GBF) indicator for preserving population-level genetic diversity. Conservation strategies, therefore, should prioritize maintaining *in situ* populations to preserve this adaptive potential, especially in regions facing climate change.

Adaptive Mechanisms in Response to Climatic and Environmental Factors

Our research identified numerous candidate genes associated with adaptive traits in different teosinte populations, particularly those in *Zea mays ssp. parviglumis* and *Zea mays ssp. mexicana*. Notably, genes involved in cutin and suberin biosynthesis, glycerolipid synthesis, and photoperiod-sensitive pathways point to complex adaptations in *parviglumis*, enabling it to thrive in warmer, moderately wet regions. These adaptations reflect a finely tuned response to environmental pressures, with particular emphasis on mechanisms to mitigate heat and maintain reproductive synchrony with seasonal variations in light and temperature.

In contrast, *mexicana* exhibits genetic associations with chromatin remodeling, calcium-dependent signaling, and other pathways supporting rapid adaptation to humidity variability and pathogen resistance in colder, high-altitude environments. The discovery of altitude-specific alleles, linked to primary active transport and nutrient uptake, aligns with the unique challenges posed by nutrient-poor soils and high-altitude stresses. These adaptations highlight the independent evolutionary pathways of *parviglumis* and *mexicana*, even when both encounter overlapping environmental conditions. Together, these findings suggest that natural selection in these populations has crafted distinct genetic responses to similar ecological niches, adding depth to our understanding of adaptive evolution.

Genetic Links Between Maize Traits and Climatic Adaptations in Teosinte

The results of this study reveal a compelling link between the genetic architecture of agronomically important maize traits and the climatic conditions of teosinte habitats. This genetic-environmental interplay underscores the role of natural selection in shaping phenotypic adaptations in *Zea* species, providing critical insights into how genetic diversity within teosinte populations has influenced the evolution and optimization of maize traits under specific environmental conditions. Loci associated with key agronomic traits, such as yield, flowering time, and stress tolerance in maize, were also linked to teosinte's climatic conditions, including temperature, humidity, photoperiod, and evapotranspiration. This dual association suggests that these traits are not just a result of artificial selection during domestication but are deeply rooted in the adaptive strategies of teosinte to its diverse habitats. Consequently, the correlation between phenotypic traits and climatic variables highlights the genetic dependency of maize's agronomic performance on specific environmental conditions. This finding has significant implications for breeding strategies, as improving traits such as yield or stress tolerance may require concurrent optimization of climatic growing conditions to fully realize their genetic potential.

The clustering analysis, based on shared SNP markers, revealed that maize root traits are particularly influenced by environmental factors. This analysis linked maize root traits to variables such as temperature, longitude, and evapotranspiration (see Chapter 2, Figure 7, for the visual representation of the tree and more detail). This finding highlights the role of below-ground traits in plant adaptation, suggesting that root architecture and function are pivotal for maize's survival and productivity under changing environmental conditions. Similarly, the association of grain traits with winter evapotranspiration and minimum temperatures reflects the importance of environmental pressures on reproductive development and seed quality, critical determinants of yield.

In summary, our findings underline the importance of considering the genetic and environmental dependencies of agronomic traits in maize breeding. The genetic diversity present in teosinte populations offers a promising reservoir of alleles for improving maize adaptation and performance under diverse climatic conditions. However, this comes with the challenge of aligning breeding goals with environmental realities, as certain traits may inherently require specific climatic conditions to express their full potential. Future breeding strategies should, therefore, integrate genomic, phenotypic, and environmental data to create robust, climate-resilient maize varieties. Additionally, the conservation of teosinte genetic resources *in situ* and *ex situ* is paramount, ensuring the long-term availability of this invaluable genetic reservoir for sustainable agriculture.

Conservation Implications: *In Situ* and *Ex Situ* Approaches

A core insight from this thesis is the necessity of population-specific conservation for maintaining teosinte's genetic diversity. As shown, teosinte populations are not interchangeable; each harbors unique genetic attributes shaped by its environment. While *in situ* conservation is indispensable, *ex situ* methods offer supplementary protection, especially against catastrophic events that could irreversibly impact local populations.

Ex situ collections should aim to capture and maintain the genetic diversity observed in natural populations. However, reintroduction efforts should consider the unique genetic adaptations of each population, as relocating individuals without considering their adaptive profiles might disrupt local adaptation patterns. Thus, a combined approach leveraging both *in situ* and strategically curated *ex situ* conservation is crucial to preserving teosinte's evolutionary potential and protecting against loss of genetic diversity. Additionally, our findings support the Global Biodiversity Framework's Population Management (PM) indicator, suggesting that monitoring at the population level is necessary to preserve adaptive traits specific to each ecological niche.

The application of GWAS to crop wild relatives, such as teosinte, and the identification of genetic adaptations to habitat conditions is a new and valuable tool in ecology. This approach provides insights into how wild species adapt to their environments, which is crucial for developing effective conservation strategies. By understanding the genetic basis of climate adaptation in teosinte, we can make more informed decisions about how to protect and manage these populations in the face of environmental change.

The Role of Omics Data Visualization in Biological Research

In parallel with the GWAS analysis, the development of GXP represents a complementary approach to plant genetics research, addressing the challenges of handling

and visualizing omics data. As omics technologies like next-generation sequencing and metabolomics become increasingly integral to plant research, the need for accessible, intuitive data analysis tools grows. GXP offers an innovative solution by providing a no-installation, web-based platform that allows users to analyze and visualize data without advanced programming skills. By integrating functions such as differential expression analysis, clustering, and pathway overrepresentation, GXP aids in the exploration of gene expression responses to environmental conditions, thereby supporting hypothesis generation and experimental validation. GXP's versatility enables researchers to visualize both quantitative and categorical data, facilitating deeper insights into genetic and phenotypic responses to stressors.

Future Directions

While this study has contributed valuable insights into the genetic diversity, adaptation, and population structure of teosinte, there remain significant avenues for future research. One key limitation we encountered was the inability to conduct whole genome and transcriptome sequencing of the teosinte individuals in Mexico due to the denial of the Access and Benefit-Sharing (ABS) permit, as required under the Nagoya Protocol. Although this regulatory challenge affected our capacity to generate comprehensive genomic data, future efforts should prioritize overcoming such barriers to fully explore the genetic basis of adaptation in these populations.

Successfully obtaining the necessary permits for genome and transcriptome sequencing would unlock several important research opportunities. First, it would enable a deeper understanding of the genetic mechanisms underlying phenotypic traits and environmental adaptations in teosinte populations. High-resolution genomic data would allow for more precise identification of adaptive loci and gene expression changes that contribute to traits like drought tolerance, altitude adaptation, or heat resistance. Additionally, transcriptomic analysis could provide critical insights into how gene regulation varies in response to environmental conditions, offering a more dynamic view of adaptation.

Moreover, future research could validate the candidate genes identified as being associated with various climatic factors in this study. By integrating additional omics data, such as epigenomics and metabolomics, researchers could gain a more comprehensive picture of how genetic, environmental, and molecular factors interact to drive adaptation in teosinte. Epigenomic data would help explore gene regulation through DNA methylation and chromatin modification, revealing the impact of environmental stress on gene expression patterns. Metabolomic profiling, on the other hand, would offer insights into biochemical pathways

associated with stress tolerance, nutrient utilization, and energy metabolism, further uncovering the mechanisms that allow teosinte populations to thrive in diverse environments.

The integration of these multi-omics approaches could be paired with functional studies to experimentally validate gene-environment interactions and their influence on phenotypic traits. A promising direction would be the use of genome-wide association studies (GWAS) combined with expression quantitative trait loci (eQTL) analysis to link genetic variants with both gene expression changes and observable phenotypes. This would provide a more robust framework for understanding the genomic architecture of adaptation and identifying key regulatory networks involved in environmental resilience.

Expanding the study to include other teosinte subspecies and maize landraces not sequenced in this project would also be beneficial. Doing so would broaden the understanding of genetic diversity within Zea species and enable comparative genomic studies across subspecies, highlighting both convergent adaptations and unique evolutionary pathways of environmental adaptation.

To facilitate future research in this area, it will be essential to continue building collaborative relationships with local stakeholders and regulatory agencies to navigate the complex legal frameworks governing access to genetic resources. Strengthening partnerships with institutions in biodiversity-rich countries like Mexico, and maintaining transparent communication regarding the use and sharing of genetic data, will be crucial to ensuring compliance with international agreements like the Nagoya Protocol. This approach may help streamline permit processes and avoid delays similar to those encountered in this study.

Finally, future research should also contribute to advancing conservation strategies for teosinte populations. Given the increasing environmental pressures and habitat loss these species face, it is critical to incorporate genetic data into conservation planning. A combination of *in situ* conservation efforts protecting natural populations in their habitats and *ex situ* strategies, such as seed banking, would help preserve the genetic diversity of teosinte for future agricultural and ecological research, safeguarding these valuable resources in the face of ongoing environmental changes.

REFERENCES

Aguirre-Liguori, J. A., Gaut, B. S., Jaramillo-Correa, J. P., Tenaillon, M. I., Montes-Hernández, S., García-Oliva, F., Hearne, S. J., & Eguiarte, L. E. (2019). Divergence with gene flow is driven by local adaptation to temperature and soil phosphorus concentration in teosinte subspecies (Zea mays parviglumis and Zea mays mexicana). *Molecular Ecology*, *28*(11), 2814–2830.

Aguirre-Liguori, J. A., Tenaillon, M. I., Vázquez-Lobo, A., Gaut, B. S., Jaramillo-Correa, J. P., Montes-Hernandez, S., Souza, V., & Eguiarte, L. E. (2017). Connecting genomic patterns of local adaptation and niche suitability in teosintes. *Molecular Ecology*, *26*(16), 4226–4240.

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664.

Amusan, I. O., Rich, P. J., Menkir, A., Housley, T., & Ejeta, G. (2008). Resistance to Striga hermonthica in a maize inbred line derived from Zea diploperennis. *The New Phytologist*, *178*(1), 157–166.

Bioversity International, & University of Birmingham. (2017). *Crop wild relative checklist and inventory descriptors v.1*. Bioversity International.

Bolger, M., Schwacke, R., & Usadel, B. (2021). MapManMapMan Visualization of RNA-Seq Data Using Mercator4Mercator4 Functional Annotations. In D. Dobnik, K. Gruden, Ž. Ramšak, & A. Coll (Eds.), *Solanum tuberosum: Methods and Protocols* (pp. 195–212). Springer US.

Casas, A., Otero-Arnaiz, A., Pérez-Negrón, E., & Valiente-Banuet, A. (2007). In situ management and domestication of plants in Mesoamerica. *Annals of Botany*, *100*(5), 1101–1115.

Castañeda-Álvarez, N. P., Khoury, C. K., Achicanoy, H. A., Bernau, V., Dempewolf, H., Eastwood, R. J., Guarino, L., Harker, R. H., Jarvis, A., Maxted, N., Müller, J. V., Ramirez-Villegas, J., Sosa, C. C., Struik, P. C., Vincent, H., & Toll, J. (2016). Global conservation priorities for crop wild relatives. *Nature Plants*, *2*, 16022.

CBD. (2022). *Decision adopted by the Conference of the Parties to the Convention on Biological Diversity CBD/COP/DEC/15/5 Monitoring framework for the Kunming-Montreal Global Biodiversity Framework. CBD/COP/DEC/15/5*. https://www.cbd.int/doc/decisions/cop-15/cop-15-dec-05-en.pdf

Chen, L., Luo, J., Jin, M., Yang, N., Liu, X., Peng, Y., Li, W., Phillips, A., Cameron, B., Bernal, J. S., Rellán-Álvarez, R., Sawers, R. J. H., Liu, Q., Yin, Y., Ye, X., Yan, J., Zhang, Q., Zhang, X., Wu, S., … Yan, J. (2022). Genome sequencing reveals evidence of adaptive variation in the genus Zea. *Nature Genetics*, *54*(11), 1736–1745.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*, 13.

Corre, V. L., Siol, M., Vigouroux, Y., Tenaillon, M. I., & Délye, C. (2020). Adaptive introgression from maize has facilitated the establishment of teosinte as a noxious weed in Europe. *Proceedings of the National Academy of Sciences*, *117*(41), 25618–25627.

Curry, H. A. (2022). The history of seed banking and the hazards of backup. *Social Studies of Science*, *52*(5), 3063127221106728.

De La Torre S., J. F., González S., R., Cruz G., E. J., Pichardo G., J. M., Quintana C., M., Contreras T., A. R., & Cadena I., J. (2018). Crop Wild Relatives in Mexico: An Overview of Richness, Importance, and Conservation Status. In S. L. Greene, K. A. Williams, C. K. Khoury, M. B. Kantar, & L. F. Marek (Eds.), *North American Crop Wild Relatives, Volume 1: Conservation Strategies* (pp. 63–96). Springer International Publishing.

Drysdale, R., Cook, C. E., Petryszak, R., Baillie-Gerritsen, V., Barlow, M., Gasteiger, E., Gruhl, F., Haas, J., Lanfear, J., Lopez, R., Redaschi, N., Stockinger, H., Teixeira, D.,

Venkatesan, A., Elixir Core Data Resource Forum, Blomberg, N., Durinx, C., & McEntyre, J. (2020). The ELIXIR Core Data Resources: fundamental infrastructure for the life sciences. *Bioinformatics (Oxford, England)*, *36*(8), 2636–2642.

EFSA, Devos, Y., Aiassa, E., Muñoz-Guajardo, I., Messéan, A., & Mullins, E. (2022). Update of environmental risk assessment conclusions and risk management recommendations of EFSA (2016) on EU teosinte. *EFSA Journal*, *20*(4), e07228.

Eiteneuer, C., Velasco, D., Atemia, J., Wang, D., Schwacke, R., Wahl, V., Schrader, A., Reimer, J. J., Fahrner, S., Pieruschka, R., Schurr, U., Usadel, B., & Hallab, A. (2022). GXP: Analyze and plot plant omics data in web browsers. *Plants*, *11*(6), 745.

Ellstrand, N. C., Garner, L. C., Hegde, S., Guadagnuolo, R., & Blancas, L. (2007). Spontaneous hybridization between maize and teosinte. *The Journal of Heredity*, *98*(2), 183–187.

FAO. (2020). *FAO strategy on mainstreaming biodiversity across agricultural sectors*. FAO.

Flint-Garcia, S. (2017). Kernel evolution: from teosinte to maize. *Maize Kernel Development*, 1–15.

Flint-Garcia, S. A., Bodnar, A. L., & Scott, M. P. (2009). Wide variability in kernel composition, seed characteristics, and zein profiles among diverse maize inbreds, landraces, and teosinte. *Theoretical and Applied Genetics*, *119*(6), 1129–1142.

Funk, W. C., McKay, J. K., Hohenlohe, P. A., & Allendorf, F. W. (2012). Harnessing genomics for delineating conservation units. *Trends in Ecology & Evolution*, *27*(9), 489–496.

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., & Buckler, E. S. (2014). TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PloS One*, *9*(2), e90346.

Goettsch, B., Urquiza-Haas, T., Koleff, P., Acevedo Gasman, F., Aguilar-Meléndez, A., Alavez, V., Alejandre-Iturbide, G., Aragón Cuevas, F., Azurdia Pérez, C., Carr, J. A., Castellanos-Morales, G., Cerén, G., Contreras-Toledo, A. R., Correa-Cano, M. E., De la

Cruz Larios, L., Debouck, D. G., Delgado-Salinas, A., Gómez-Ruiz, E. P., González-Ledesma, M., … Jenkins, R. K. B. (2021). Extinction risk of Mesoamerican crop wild relatives. *Plants, People, Planet*, *3*(6), 775–795.

Goudet, J. (2020). *Jombart T. hierfstat: estimation and tests of hierarchical F-statistics. R package version 0.04-22. 2015*.

Guzzon, F., Arandia Rios, L. W., Caviedes Cepeda, G. M., Céspedes Polo, M., Chavez Cabrera, A., Muriel Figueroa, J., Medina Hoyos, A. E., Jara Calvo, T. W., Molnar, T. L., Narro León, L. A., Narro León, T. P., Mejía Kerguelén, S. L., Ospina Rojas, J. G., Vázquez, G., Preciado-Ortiz, R. E., Zambrano, J. L., Palacios Rojas, N., & Pixley, K. V. (2021). Conservation and use of Latin American maize diversity: Pillar of nutrition security and cultural heritage of humanity. *Agronomy (Basel, Switzerland)*, *11*(1), 172.

Hernández-de-Diego, R., Tarazona, S., Martínez-Mira, C., Balzano-Nogueira, L., Furió-Tarí, P., Pappas, G. J., Jr, & Conesa, A. (2018). PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data. *Nucleic Acids Research*, *46*(W1), W503–W509.

Hoban, S., Jessica, M., Hughes, A., Hunter, M. E., Stroil, K., Laikre, L., Mastretta-Yanes, A., Millette, K., Paz-Vinas, I., Bustos, L. R., Shaw, R. E., & Vernesi, C. (2024). Too simple, too complex, or just right? Advantages, challenges, and guidance for indicators of genetic diversity. *Bioscience*, *74*(4), 269–280.

Hufford, M. B., Lubinksy, P., Pyhäjärvi, T., Devengenzo, M. T., Ellstrand, N. C., & Ross-Ibarra, J. (2013). The genomic signature of crop-wild introgression in maize. *PLoS Genetics*, *9*(5), e1003477.

Hufford, M. B., Martínez-Meyer, E., Gaut, B. S., Eguiarte, L. E., & Tenaillon, M. I. (2012). Inferences from the historical distribution of wild and domesticated maize provide ecological and evolutionary insight. *PloS One*, *7*(11), e47659.

Jayakodi, M., Golicz, A. A., Kreplak, J., Fechete, L. I., Angra, D., Bednář, P., Bornhofen, E., Zhang, H., Boussageon, R., Kaur, S., Cheung, K., Čížková, J., Gundlach, H., Hallab, A.,

Imbert, B., Keeble-Gagnère, G., Koblížková, A., Kobrlová, L., Krejčí, P., … Andersen, S. U. (2023). The giant diploid faba genome unlocks variation in a global protein crop. *Nature*, *615*(7953), 652–659.

Jiao, Y., Zhao, H., Ren, L., Song, W., Zeng, B., Guo, J., Wang, B., Liu, Z., Chen, J., Li, W., Zhang, M., Xie, S., & Lai, J. (2012). Genome-wide genetic changes during modern breeding of maize. *Nature Genetics*, *44*(7), 812–815.

Julkowska, M. M., Saade, S., Agarwal, G., Gao, G., Pailles, Y., Morton, M., Awlia, M., & Tester, M. (2019). MVApp—Multivariate Analysis Application for Streamlined Data Analysis and Curation. In *Plant Physiology* (Vol. 180, Issue 3, pp. 1261–1276). https://doi.org/10.1104/pp.19.00235

Karn, A., Gillman, J. D., & Flint-Garcia, S. A. (2017). Genetic analysis of teosinte alleles for kernel composition traits in maize. *G3 (Bethesda, Md.)*, *7*(4), 1157–1164.

Kassambara, A. (2016). Factoextra: extract and visualize the results of multivariate data analyses. *R Package Version*, *1*. https://cir.nii.ac.jp/crid/1370004235968325765

Kohsaka, R. (2012). The negotiating history of the Nagoya Protocol on ABS: Perspective from japan. *Journal of Intellectual Property Association of Japan https://www.ipaj.org/english_journal/pdf/9-1_Kohsaka.pdf*, *9*, 56–66.

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*(1), 559.

Letunic, I., & Bork, P. (2024). Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Research*, *52*(W1), W78–W82.

Liu, X., Huang, M., Fan, B., Buckler, E. S., & Zhang, Z. (2016). Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. *PLoS Genetics*, *12*(2), e1005767.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550.

Mastretta-Yanes, A., da Silva, J. M., Grueber, C. E., Castillo-Reina, L., Köppä, V., Forester, B. R., Funk, W. C., Heuertz, M., Ishihama, F., Jordan, R., Mergeay, J., Paz-Vinas, I., Rincon-Parra, V. J., Rodriguez-Morales, M. A., Arredondo-Amezcua, L., Brahy, G., DeSaix, M., Durkee, L., Hamilton, A., … Hoban, S. (2024). Multinational evaluation of genetic diversity indicators for the Kunming-Montreal Global Biodiversity Framework. *Ecology Letters*, *27*(7), e14461.

Maxted, N. (2013). In Situ, Ex Situ Conservation. In *Encyclopedia of Biodiversity* (pp. 313–323). Elsevier.

Maxted, N., Brehm, J. M., & Kell, S. (2013). Resource book for the preparation of National Plans for Conservation of Crop Wild Relatives and Landraces. *FAO: Rome, Italy*, *463*.

Maxted, N., Kell, S., Ford-Lloyd, B., Dulloo, E., & Toledo, Á. (2012). Toward the systematic conservation of global crop wild relative diversity. *Crop Science*, *52*(2), 774–785.

Maxted, N., & Kell, S. P. (2009). Establishment of a global network for the in situ conservation of crop wild relatives: status and needs. *FAO Commission on Genetic Resources for Food and Agriculture*, *266*.

Maxted, N., Kell, S., Toledo, Á., Dulloo, E., Heywood, V., Hodgkin, T., Hunter, D., Guarino, L., Jarvis, A., & Ford-Lloyd, B. (2010). A global approach to crop wild relative conservation: securing the gene pool for food and agriculture. *Kew Bulletin / Royal Botanic Gardens*, *65*(4), 561–576.

Moreno-Letelier, A., Aguirre-Liguori, J. A., Piñero, D., Vázquez-Lobo, A., & Eguiarte, L. E. (2020). The relevance of gene flow with wild relatives in understanding the domestication process. *Royal Society Open Science*, *7*(4), 191545.

Parra-Quijano, M., Iriondo, J. M., & Torres, E. (2012). Ecogeographical land characterization maps as a tool for assessing plant adaptation and their implications in agrobiodiversity studies. *Genetic Resources and Crop Evolution*, *59*(2), 205–217.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–

419.

Peng, R. D. (2011). Reproducible research in computational science. *Science (New York, N.Y.)*, *334*(6060), 1226–1227.

Perrino, E. V., & Perrino, P. (2020). Crop wild relatives: know how past and present to improve future research, conservation and utilization strategies, especially in Italy: a review. *Genetic Resources and Crop Evolution*, *67*(5), 1067–1105.

Pyhäjärvi, T., Hufford, M. B., Mezmouk, S., & Ross-Ibarra, J. (2013). Complex patterns of local adaptation in teosinte. *Genome Biology and Evolution*, *5*(9), 1594–1609.

Rivera-Rodríguez, D. M., de Jesús Sánchez González, J., De la Cruz Larios, L., Santacruz-Ruvalcaba, F., & Ruiz Corral, J. A. (2019). Morphological and Climatic Variability of Teosinte (Zea spp.) and Relationships Among Taxa. *Systematic Botany*, *44*(1), 41–51.

Rivera-Rodríguez, D. M., Mastretta-Yanes, A., Wegier, A., De la Cruz Larios, L., Santacruz-Ruvalcaba, F., Ruiz Corral, J. A., Hernández, B., & Sánchez González, J. de J. (2023). Genomic diversity and population structure of teosinte (Zea spp.) and its conservation implications. *PloS One*, *18*(10), e0291944.

Rojas-Barrera, I. C., Wegier, A., Sánchez González, J. de J., Owens, G. L., Rieseberg, L. H., & Piñero, D. (2019). Contemporary evolution of maize landraces and their wild relatives influenced by gene flow with modern maize varieties. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(42), 21302–21311.

Sánchez González, J. de J., Ruiz Corral, J. A., García, G. M., Ojeda, G. R., Larios, L. D. la C., Holland, J. B., Medrano, R. M., & García Romero, G. E. (2018). Ecogeography of teosinte. *PloS One*, *13*(2), e0192676.

Satori, D., Tovar, C., Faruk, A., Hammond Hunt, E., Muller, G., Cockel, C., Kühn, N., Leitch, I. J., Lulekal, E., Pereira, L., Ryan, P., Willis, K. J., & Pironon, S. (2022). Prioritising crop wild relatives to enhance agricultural resilience in sub□Saharan Africa under climate change. *Plants, People, Planet*, *4*(3), 269–282.

Schwacke, R., Ponce-Soto, G. Y., Krause, K., Bolger, A. M., Arsova, B., Hallab, A., Gruden,

K., Stitt, M., Bolger, M. E., & Usadel, B. (2019). MapMan4: A Refined Protein Classification and Annotation Framework Applicable to Multi-Omics Data Analysis. *Molecular Plant*, *12*(6), 879–892.

Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q., & Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genetics*, *44*(7), 825–830.

Soneson, C., & Robinson, M. D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, *15*(4), 255–261.

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Reviews. Genetics*, *20*(11), 631–656.

Studer, A., Zhao, Q., Ross-Ibarra, J., & Doebley, J. (2011). Identification of a functional transposon insertion in the maize domestication gene tb1. *Nature Genetics*, *43*(11), 1160–1163.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545–15550.

Tian, D., Wang, P., Tang, B., Teng, X., Li, C., Liu, X., Zou, D., Song, S., & Zhang, Z. (2020). GWAS Atlas: a curated resource of genome-wide variant-trait associations in plants and animals. *Nucleic Acids Research*, *48*(D1), D927–D932.

Tobón-Niedfeldt, W., Mastretta-Yanes, A., Urquiza-Haas, T., Goettsch, B., Cuervo-Robayo, A. P., Urquiza-Haas, E., Orjuela-R, M. A., Acevedo Gasman, F., Oliveros-Galindo, O., Burgeff, C., Rivera-Rodríguez, D. M., Sánchez González, J. de J., Alarcón-Guerrero, J., Aguilar-Meléndez, A., Aragón Cuevas, F., Alavez, V., Alejandre-Iturbide, G., Avendaño-Arrazate, C.-H., Azurdia Pérez, C., … Koleff, P. (2022). Incorporating evolutionary and threat processes into crop wild relatives conservation. *Nature Communications*, *13*(1),

6254.

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., & Lappalainen, T. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1), 1–21.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*(11), 4414–4423.

Wang, H., Nussbaum-Wagler, T., Li, B., Zhao, Q., Vigouroux, Y., Faller, M., Bomblies, K., Lukens, L., & Doebley, J. F. (2005). The origin of the naked grains of maize. *Nature*, *436*(7051), 714–719.

Wang, J., & Zhang, Z. (2021). GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and Prediction. *Genomics, Proteomics & Bioinformatics*, *19*(4), 629–640.

Wang, L., Xu, C., Qu, M., & Zhang, J. (2008). Kernel amino acid composition and protein content of introgression lines from Zea mays ssp. mexicana into cultivated maize. *Journal of Cereal Science*, *48*(2), 387–393.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, *10*(1), 57–63.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J., da Silva Santos, L. O. B., Bourne, P., Bouwman, J., Brookes, A., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*. https://doi.org/10.1038/sdata.2016.18

Woodhouse, M. R., Cannon, E. K., Portwood, J. L., 2nd, Harper, L. C., Gardiner, J. M., Schaeffer, M. L., & Andorf, C. M. (2021). A pan-genomic approach to genome databases using maize as a model system. *BMC Plant Biology*, *21*(1), 385.

Wright, S. I., Bi, I. V., Schroeder, S. G., Yamasaki, M., Doebley, J. F., McMullen, M. D., & Gaut, B. S. (2005). The effects of artificial selection on the maize genome. *Science*, *308*(5726), 1310–1314.

Yang, N., Wang, Y., Liu, X., Jin, M., Vallebueno-Estrada, M., Calfee, E., Chen, L., Dilkes, B. P., Gui, S., Fan, X., Harper, T. K., Kennett, D. J., Li, W., Lu, Y., Ding, J., Chen, Z., Luo, J., Mambakkam, S., Menon, M., … Ross-Ibarra, J. (2023). Two teosintes made modern maize. *Science*, *382*(6674), eadg8940.

Zamora-Tavares, P., Vargas-Ponce, O., Sánchez-Martínez, J., & Cabrera-Toledo, D. (2015). Diversity and genetic structure of the husk tomato (Physalis philadelphica Lam.) in Western Mexico. *Genetic Resources and Crop Evolution*, *62*(1), 141–153.

Zavala-López, M., López-Tavera, E., Figueroa-Cárdenas, J. de D., Serna-Saldívar, S. O., & García-Lara, S. (2018). Screening of major phenolics and antioxidant activities in teosinte populations and modern maize types. *Journal of Cereal Science*, *79*, 276–285.

ACKNOWLEDGMENTS

First and foremost, I would like to express my heartfelt gratitude to my advisors, **Prof. Dr. Asis Hallab**, **Dr. Alicia Mastretta-Yanes**, and **Prof. Dr. Björn Usadel**, whose guidance, patience, and expertise were invaluable throughout my PhD journey. **Prof. Dr. Hallab**, as my daily supervisor, has been instrumental in shaping my research path, providing constant support and insightful advice. I am especially grateful to **Dr. Alicia Mastretta-Yanes** for her invaluable mentorship, insightful input, and genuine encouragement, which has profoundly impacted my development as a researcher. I am deeply grateful to my mentor, **Prof. Dr. Benjamin Stich**, for his availability and dedication to my growth as a researcher. His input and encouragement have been critical to the progress and completion of this work.

I am also thankful for the collaboration and expertise shared by my colleagues and mentors in Mexico. **Dr. José de Jesús Sánchez**, **Dr. Diana Rodríguez**, **Dr. Lino de la Cruz Larios**, **Dr. Ana Wegier**, and **Dr. Nancy Galvez**: thank you for your invaluable contributions and support that enriched my research experience and helped drive the findings of this thesis.

To my friends and colleagues in the **Usadel Lab**, thank you for the welcoming environment and constant support. I especially want to recognize **Dr. Elisa Senger**, **Dr. Marius Weisweiler,** and **Volkan Cevik**, whose kindness, encouragement, and advice made a significant difference, including in helping me adapt to life in Germany.

Lastly, I owe my deepest gratitude to my family, whose unwavering love, sacrifices, and belief in me have been the foundation of my success. This thesis would not have been possible without their enduring support.