

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Evolutionary adaptation of the tepary bean (*Phaseolus acutifolius* A. Gray): Exploring functional phenome variation in genebank collections

Diego Felipe Conejo Rodríguez

Universidad Nacional de Colombia
Facultad Ciencias Agropecuarias, Ciencias Agrícolas
Palmira, Colombia
2024

Evolutionary Adaptation of the Tepary Bean (*Phaseolus acutifolius* A. Gray): Exploring Functional Phenome Variation in Genebank Collection

Diego Felipe Conejo Rodriguez

Tesis o trabajo de investigación presentada(o) como requisito parcial para optar al título de:
Doctor en Ciencias Agrarias

Director (a):

Milan Urban, PhD

Codirector (a):

Jhon Ocampo, PhD

Asesor:

Joaquin Guillermo Ramirez Gil, PhD

Línea de Investigación:

Mejoramiento genético

Universidad Nacional de Colombia
Facultad Ciencias Agropecuarias, Ciencias Agrícolas
Palmira, Colombia
2024

Content

1. Chapter one: Introduction	18
1.1 Research question	24
1.2 Hypothesis	24
1.3 General aim	24
1.4 Specific aims	25
1.5 General conceptual aspects	26
1.6 Evolutionary adaptation in plants and crops	26
1.7 Evolution and botanical characteristics of Phaseolus	28
1.8 Tepary bean (<i>Phaseolus acutifolius</i> A. Gray)	28
1.9 Tepary bean domestication	30
1.10 Crop phenomics	31
1.11 High throughput plant phenotyping	32
1.12 Functional diversity	33
1.13 Representativeness genetic, ecogeographic and functional diversity in genebanks	34
1.14 References	37
2. Chapter 2: Digital descriptors sharpen classical descriptors, for improving genebank accession management: A case study on <i>Arachis</i> spp. and <i>Phaseolus</i> spp.	47
2.1 Introduction	49
2.2 Materials and Methods	53
2.2.1 Methodological approach	53
2.2.2 Plant material and cultivation	54
2.2.3 Classical descriptors	55
2.2.4 Image acquisition and pre-processing	55
2.2.5 Extraction of phenomics descriptors from images	56

2.2.6 Selection of classical, phenomic and combined descriptors	57
2.2.7 Functional diversity of phenomic, classical and their combination	58
2.3 Results	59
2.3.1 Selection of phenomic, classical, and combined descriptors and accession classification	59
2.3.2 Functional diversity indices are important for beans and forage peanut accessions' characterization	62
2.4 Discussion	65
2.4.1 Proposed methodology for genebank curators	70
2.4.2 Genebanks, characterization, phenomics and conceptualization	73
2.5 Conclusions	75
2.6 Acknowledgments	75
2.7 References	76
3. Chapter 3: Representativeness functional phenome and ecogeographic diversity in genebanks: A case study on the tepary bean collection	97
3.1 Introduction	99
3.2 Materials and Methods	102
3.2.1 Methodological approach	102
3.2.2 Ecogeographic variable extraction	102
3.2.3 Dimensionality reduction of ecogeographic variables	103
3.2.4 Ecogeographic clustering of tepary bean	104
3.2.5 Functional diversity in tepary bean collection	105
3.2.6 Merging ecogeographic groups, genetic structure and functional diversity	107
3.3 Results	107
3.4 Discussion	113
3.5 Conclusions	118

3.6 Acknowledgments	119
3.7 References	119
4. Chapter 4: Sun-induced Chl fluorescence as an index selection of genotypes to light use efficiency in physiological breeding: Tepary bean functional collection case study	142
4.1 Introduction	143
4.2 Materials and methods	145
4.2.1 Methodological approach	145
4.2.2 Tepary bean functional diversity collection	146
4.2.3 Experimental design and cultivation	146
4.2.4 Phenotyping of photosynthetic functional traits and yield components	147
4.2.5 Sun-induced Chl fluorescence (SIF)	148
4.2.6 Selection index for light use efficiency: Relationship between SIF, biomass, yield and pod harvest index	149
4.3 Results	151
4.4 Discussion	156
4.5 Conclusions	160
4.6 Acknowledgments	160
4.7 References	161
5. Chapter 5: Using phenomics to identify and integrate traits of interest for better-performing common beans: a validation study on an interspecific hybrid and its Acutifolii parents	175
5.1 Introduction	177
5.2 Materials and methods	180
5.2.1 Plant material and experimental design	180
5.2.2 Morphometric descriptors	182

5.2.3 Physiological descriptors	183
5.2.4 Yield components	184
5.2.5 Multivariable analysis and Phaseolus accessions classification	185
5.2.6 Phenomic proportions of the interspecific hybrid with respect to its parents	185
5.3 Results	187
5.3.1 Phenomic proportions of the hybrid respect to its parents	188
5.4 Discussion	190
5.5 Conclusions	197
5.6 Acknowledgments	198
5.7 References	199
6. Chapter 6: Synthesis of results	213
6.1 Future prospects	215

*A quien despierta mis metales,
más allá y más acá
de los posibles y agotables acontecimientos.*

*A mi bella y ponderada
compañera de caminos,
baúl de mis recuerdos y mis sombras.*

*La que me anima en el absurdo,
bálsamo del afecto,
que recorre sutil,
entre la máscara y el rostro.*

*A quienes me habitaron
desde el oficio y ejercicio de la palabra,
con quienes ejercí los encuentros
de mis primeros años,
al pueblo y sus ventanas,
al hambre de amor,
y a esta fe infinita de aún
seguir creyendo.*

Patricia Arregocés Delgado

Acknowledgments

I would like to express my gratitude to my parents, Elizabeth Rodriguez Murillo, Alvaro Conejo Fandiño, and my sister, Martha Janeth Conejo Rodriguez, for influencing my focus and strategy. I am extremely grateful for their ideas, guidance, and support in a multitude of topics that contributed to my todays and future scientific interest. Thank you very much. To my aunts Nelly Conejo and Aurora Conejo, who affected me from infancy to adult by projecting confidence in each of the sentences we discussed. I thank Marcela Santaella, Milan Urban, and Joaquin Guillermo Ramirez for providing a basis and incentive for me to improve my skills and cleverness during my research. I would like to express my gratitude to Dr. Daniel Debouck. Dr. Daniel Debouck will always be an inspiration to me; his study and thoughts have influenced my philosophy, life, and love of science. Thank you for introducing me to the field of science and for sharing your time and wisdom at the highest point of my professional development. Thank you!. I thank the universe for everyone who has walked my path, from environmental activists to academics and scientists. Finally, I am grateful for my perseverance in achieving my goals of generating knowledge and developing capabilities that drive scientific discoveries.

Abstract

Evolutionary Adaptation of the Tepary Bean (*Phaseolus acutifolius* A. Gray): Exploring Functional Phenome Variation in Genebank Collection

The tepary bean (*Phaseolus acutifolius* A. Gray) is important for the genetic improvement of common beans (*Phaseolus vulgaris* L). Its genetic resources conserved in genebanks have been the focus of several studies to understand the genetic, taxonomic, functional, and ecological variations driven by evolutionary processes. Due to tepary resistance to drought and high temperatures, this legume has food sustainability importance in current and upcoming climate change scenarios. Germplasm collections could help us to understand the diversity of the functional phenome since the environment, time, and humans have all influenced phenotypic variation of highly heritable functional traits within the accessions that currently are conserved in genebanks. In this thesis I have presented a methodological, analytical, and conceptual approach to discovering functional phenome variation utilizing digital tools, artificial intelligence, and data science to better understand evolutionary adaptation processes, specifically in the tepary bean. I showed that the use of digital descriptors facilitates germplasm characterization during seed regeneration, as well as capture comprehending variation in phenotype using functional diversity indices as functional specialization (F_{Spe}), functional originality (F_{Ori}) and functional identity (F_{Id}e). Data interpretation helps to understand ecological processes in wild accessions. Data also reflect human phenotypic selection during directed and unconscious domestication and adaptation of tepary bean in Arizona-Sonora and Central America. I have noticed this by analyzing spatial, phenotypic, and genetic data mining linking tepary bean phenome variation to evolutionary processes. In addition, I have suggested gathering functional traits for tepary beans accessions from the genebank of the International Center for Tropical Agriculture (CIAT) in Colombia. Following the collection of tepary bean functional traits, light use efficiency (LUE) is estimated by applying sun-induced fluorescence (SIF) to both wild and domesticated tepary bean accessions. The findings demonstrated that the collection of tepary bean traits exhibits differential LUE attributed to both natural and human selection processes. I also found that wild and domesticated accessions accumulate biomass

differently based on the pod harvest index (PHI), total dry matter, and yield by examining LUE variation in relation to genetic and ecological clusters.

With my colleagues (co-authors or the papers) we used digital image processing, artificial intelligence, and functional traits to determine the phenotypic proportions of interspecific hybridization. This should serve as a strategy for breeding selection procedures of elite genotypes that exhibit commercial features. We developed selection criteria for interspecific hybrids based on selected seed and pod morphometric traits, yield components, and photosynthetic performance.

Keywords: Crop phenomics, machine learning, crop evolution, phenotyping, characterization, digital image processing, digital descriptors, photophysics, sun induced fluorescence.

Resumen

Adaptación evolutiva del frijol tepari (*Phaseolus acutifolius* A. Gray): Explorando la variación del fenoma funcional en colecciones de bancos de germoplasma

El frijol tepari (*Phaseolus acutifolius* A. Gray) es de importancia para el mejoramiento genético de frijol común, y sus recursos genéticos conservados en bancos de germoplasma han sido foco en diversos estudios que permitan comprender los procesos de variación genética, taxonómica, funcional y ecológica que han sido dirigido por procesos evolutivos. La variación del fenoma funcional puede comprenderse a partir de las colecciones de germoplasma, partiendo de que el ambiente, el tiempo y el hombre determinaron la variación fenotípica de rasgos funcionales de alta heredabilidad de las accesiones conservadas actualmente en los bancos de germoplasma. En este trabajo de tesis hacemos un acercamiento metodológico, analítico y conceptual para el descubrimiento de la variación del fenoma funcional usando herramientas digitales, inteligencia artificial y ciencia de datos que apoyen a comprender los procesos de adaptación evolutiva, en este caso, del frijol tepari. Nosotros encontramos que el uso de descriptores digitales apoyan los procesos de caracterización de germoplasma durante la regeneración de semillas, además de comprender su variación fenotípica usando índices de diversidad funcional. Determinamos cómo a partir de minería de datos espaciales, fenotípicos y genéticos relacionamos la variación del fenoma del frijol tepari a los procesos de evolutivos, descubriendo cómo tanto los procesos ecológicos en accesiones silvestres y la selección fenotípica del hombre dirige la estructura genética durante la adaptación y domesticación dirigida e inconsciente del frijol tepari en Arizona - Sonora y Centro América. Igualmente proponemos la colección de rasgos funcionales de frijol tepari del banco de germoplasma *semillas del futuro* del Centro Internacional de Agricultura Tropical (CIAT). Posteriormente, a la colección de rasgos funcionales de frijol tepari se determina el uso eficiente de la luz (LUE) a partir de la fluorescencia inducida por el sol (SIF) a las accesiones de frijol tepari silvestres y domesticadas. Descubrimos que la colección de rasgos funcionales del frijol tepari presenta LUE diferencial dirigido por los procesos de selección natural y del hombre. Demostramos que las accesiones silvestres como domesticadas acumulan diferencialmente biomasa a partir del índice de cosecha a nivel de vaina (PHI), materia

seca total y el rendimiento, analizando la variación del LUE con su estructura genética y ecológica. Debido a los limitantes en el uso de parientes silvestres de cultivo (CWR), principalmente en los procesos de hibridación y selección de genotipos que presenten características comerciales y mantengan proporciones fenotípicas de los CWR. Desarrollamos una propuesta metodológica a partir de procesamiento digital de imágenes, inteligencia artificial y rasgos funcionales establecimos las proporciones fenómicas de las hibridaciones interespecíficas. Usamos criterios de selección de híbridos interespecíficos basados en rasgos morfométricos de semilla y vaina, componentes de rendimiento y desempeño fotosintético. Finalmente, proponemos un flujo de trabajo que puede ser usada por los bancos de germoplasma y curadores que contribuyan a la minería de rasgos adaptativos en las colecciones de germoplasma que asocie rasgos funcionales en la documentación de las colecciones que fomenten el uso y distribución de semillas que brinden soluciones.

Palabras clave: Fenomica de cultivos, aprendizaje automático, evolución de cultivos, fenotipado, caracterización, procesamiento digital de imágenes, descriptores digitales, fotofísica, fluorescencia inducida por el sol.

Figures

	Pág.
Chapter two: <i>Digital descriptors sharpen classical descriptors, for improving genebank accession management: A case study on Arachis spp. and Phaseolus spp</i>	
Figure 2.1. Methodological procedure for the phenome characterization of <i>Phaseolus</i> and <i>Arachis</i> accessions.	83
Figure 2.2 Functional phenomic descriptors that describe the phenotypic diversity of <i>Arachis</i> spp. accessions.	85
Figure 2.3 Functional phenomic descriptors that describe the phenotypic diversity of <i>Phaseolus</i> spp. accessions.	86
Figure 2.4. Selection of classical descriptors, phenomic descriptors and their combination used a minimal depth distribution on <i>Arachis</i> spp accessions and its classification.	88
Figure 2.5. Selection of classical descriptors, phenomic descriptors and their combination evaluated via a minimal depth distribution on <i>Phaseolus</i> spp accessions and its classification.	90
Figure 2.6. Vertices accessions of the functional space for each of the descriptor types used for bean and forage peanut accessions.	92
Figure 2.7. Functional diversity indices of the accession grouping by <i>Phaseolus</i> spp. for each type of descriptor.	93
Figure 2.8. Functional diversity indices of the accession grouping by <i>Arachis</i> species for each type of descriptor.	94
Figure 2.9. CIAT genebank workflow suggestion resulting from the proposed methodology	95
Chapter three: <i>Representativeness functional phenome and ecogeographic diversity in genebanks: A case study on the tepary bean collection</i>	
Figure 3.1 The ecogeographic representativeness and functional phenome of the tepary bean CIAT's germplasm collection were investigated using this methodological pipeline.	123

Figure 3.2 Selection of key ecogeographic variables for distinguishing ecogeographic groups of <i>P. acutifolius</i>	125
Figure 3.3 Ecogeographic groups of <i>P. acutifolius</i> accessions, as well as radar plots of important variables.	126
Figure 3.4 The vertices positions within the cultivated <i>P. acutifolius</i> accessions seen in all major components PC1 to PC5.	127
Figure 3.5 Wild <i>P. acutifolius</i> vertice accessions in the major components PC1 to PC5.	128
Figure 3.6 Functional diversity indices for cultivated (1C, 2C, 3C y 4C) and wild (1W and 2W) <i>P. acutifolius</i> accessions in each ecogeographic group.	129
Figure 3.7 Functional diversity indices of cultivated <i>P. acutifolius</i> accessions per ecogeographic group.	130
Figure 3.8 Functional diversity indices of wild <i>P. acutifolius</i> accessions per ecogeographic group.	131
Figure 3.9 Effect of ecogeographic variables and ecogeographic groups on phenotypic traits observed in our field experiment of <i>P. acutifolius</i> accessions.	132
Figure 3.10 Ecogeographic group and genetic clusters of the wild tepary bean collection.	133
Figure 3.11 Ecogeographic groups and genetic clusters of the cultivated tepary bean collection.	135
Figure 3.12 Relationship of centroid distances to genetic and eco-geographic groups of wild accessions of tepary bean.	136
Figure 3.13 Relationship of centroid distances to genetic and eco-geographic groups of cultivated tepary bean accessions.	138

Chapter four: Sun-induced Chl fluorescence as an index selection of genotypes to light use efficiency in physiological breeding: Tepary bean functional collection case study

Figure 4.1 Methodological process of physiological evaluation of the tepary bean functional diversity collection.	165
--	-----

Figure 4.2 Relationship of qL with SIF for a set of tepary bean accessions evaluated under drought and irrigation conditions.	167
Figure 4.3 Relationship of qL medium with SIF for tepary bean wild, weedy and cultivated accessions evaluated under drought and irrigation conditions.	168
Figure 4.4 Relationship of mean yield to median SIFmax for cultivated tepary bean accessions evaluated under drought and irrigated conditions from the ecological groups.	169
Figure 4.5 Relationship of median yield to median SIF for wild tepary bean accessions evaluated under drought and irrigated conditions from the ecological groups.	170
Figure 4.6 Relationship of median pod harvest index (PHI) to median SIF for cultivated tepary bean accessions evaluated under drought and irrigated conditions from the ecological groups.	171
Figure 4.7 Relationship of median pod harvest index (PHI) to median SIF for wild tepary bean accessions evaluated under drought and irrigated conditions from the ecological groups.	172
Figure 4.8 Relationship of median biomass to median SIF for wild tepary bean accessions evaluated under drought and irrigated conditions from the ecological groups.	173
Figure 4.9 Relationship of median biomass to median SIF for wild tepary bean accessions evaluated under drought and irrigated conditions from the ecological groups.	174
Figure 4.10 Correlations of biomass, SIFmax, leaf area, yield and pod harvest index under irrigation and drought conditions evaluated in the tepary bean functional collection.	175

Chapter five: Using phenomics to identify and integrate traits of interest for better-performing common beans: a validation study on an interspecific hybrid and its *Acutifolii* parents

Figure 5.1 The procedure consists of three steps: (A) capture and processing of phenomic descriptors, (B) classification of parents using random forest and PCA and (C) phenomic ratios based on the weighting of the descriptors with the contribution of the PCs and phenomic ratios as the prediction of the hybrid with respect to its parents in the confusion table.	206
---	-----

Figure 5.2 Procedure for the analysis of the phenomic proportions between interspecific hybrids and their parents.	207
Figure 5.3 Phenotypic characteristics of the Phaseolus lines and interspecific hybrid, from its initial stages until harvest.	208
Figure 5.4 Gini index and accuracy decrease for feature selection in random forest for each characterization component in the hybrid.	209
Figure 5.5 Phenomic proportions in the confusion matrices for classification of the interspecific hybrid in each of the characterization components using phenomics descriptors	210

Chapter six: *Synthesis of results*

Figure 6.1 Methodological proposal for adaptive trait mining during characterization and documentation in genebanks, integrating ecological, genetic, functional, and geographic groups.	214
---	-----

Tables

	Pág.
Chapter two	
Table 2.1. Indices of functional diversity for the accessions of <i>Phaseolus</i> and <i>Arachis</i> species evaluated with phenomic (Phe), classic (Cla) and their combined (Comb) descriptors.	96
Chapter three	
Table 3.1 Passport data from the vertices accessions of the functional spaces PC1 through PC5.	139
Chapter five	
Table 5.1. MANOVA of parents and interspecific hybrid in each of the characterization components.	211

Organization of the thesis

According to the regulations of the Universidad Nacional de Colombia, the format proposed by SINAB, and, in particular, the regulations created for the presentation of the thesis, the final document might be arranged into chapters, with the results presented in the format of scientific articles. Some articles may differ from others due to the different formats used by each journal. However, we attempted to present them in the following order: introduction, materials and techniques, findings, and discussion, followed by a list of references and figures at the end of the chapter.

This research work is organized in seven chapters, which have the following content: The first chapter is a conceptual introduction that guides the thesis by highlighting the important subjects from a broad to a specific viewpoint. We conduct a literature review to determine the subject areas and how they add to our conceptual knowledge of this work, as well as suggesting the need for more research into germplasm characterisation techniques in genebanks. The two chapter is headed "*Digital descriptors sharpen classical descriptors, in improving genebank accession management: A case study on Arachis spp. and Phaseolus spp.*" (Conejo - Rodríguez et al., 2024) published in the journal PLOS One vol. 19, no 5 in the section Plant Science. In this chapter, we offer a methodological approach that integrates the use of digital and classical descriptors, as well as the use of random forests to choose important descriptors. Furthermore, we provide germplasm management-related analysis methodologies based on functional diversity indices. The chapter three, "*Representativeness functional phenome and ecogeographic diversity in genebanks: A case study on the tepary bean collection,*" discusses how functional diversity and the use of phenotypic data from the collection enable us to investigate phenome variation in

germplasm collections. In this study, we look at the ecological, genetic, phenomic, and geographic components that help us understand phenotypic variation and its relevance to adaptive processes in tepary beans. We identify ecological groups by spatial analysis of climatic variables and the use of decision support artificial intelligences. In addition, we build the collection of tepary bean functional diversity, which serves as a basis for the following chapter. In Chapter fourth, "*Sun-induced Chl fluorescence as an index selection of genotypes to light use efficiency in physiological breeding: Tepary bean functional collection case study*," we develop a selection index based on sun-induced fluorescence (SIF) as a strategy for selecting accessions and genotypes to light use efficiency using pulses amplitude modulated (PAM). We show that under conditions of natural variation and drought, accessions differ in light use efficiency (LUE) depending on their genetic and ecological group, indicating how adaptive processes influenced the photosynthetic processes of tepary beans. In the five chapter, "*Using phenomics to identify and integrate traits of interest for better-performing common beans: a validation study on an interspecific hybrid and its Acutifolii parents*" (Conejo-Rodríguez et al. 2022, Frontiers plants, vol. 13), a methodological process is developed to support selection processes in interspecific hybridizations. This approach uses multivariate analysis strategies and random forests to incorporate digital descriptors and classification algorithms for parents and interspecific hybrids. Finally, in chapter seven, we develop general implications from the thesis work and present a methodological contribution that integrates all of the doctoral work. I hope it is useful and enjoyable to all of you. During different experiments that preceded the writing on my thesis I co-lead or conducted with CIAT Bean Physiology Team and with other students. It should be noted, that only my research outputs were used to write this thesis, unless it is explicitly

mentioned I used some conclusions from other students/researchers to support my ideas or to show a different view.

Chapter one: Introduction

Climate change has created significant challenges for sustainable agriculture (Magnan et al., 2021). Crop adaptation to changing environments will maintain food security in the future, as climate change models around the world predict that rising temperatures and droughts will likely have a significant impact on agricultural production (Challinor et al., 2014). High temperature stress and drought influences plant growth and development, altering physiological processes as well as crop yield (Driedonks et al., 2015). Temperature and drought have a deleterious impact on several important physiological processes in plants. A variety of physiological, biochemical, and molecular mechanisms regulate plant growth; additionally, photosynthesis is a major phenomena that contributes strongly to crop productivity (Mathur et al., 2014).

Common bean (*Phaseolus vulgaris* L.), is the most extensively cultivated and economically important species in the *Phaseolus* genus (Beebe et al., 2011). Reductions of 50% in common bean cultivation area are primarily due to expected climate variability in Latin America and Africa (Beebe et al., 2011; Paliwal et al., 2020). It is considered that temperatures above 30 °C reduced the quantity of seeds, reduced photosynthetic rates, and increased stomatal conductance (Traub et al., 2018). This generates a need to look for alternatives to reduce these sources of stress and adaptation to variability and climate change in this species. In this regard, diversification into the production of underutilized legume crops, such as tepary bean (*Phaseolus acutifolius* A Gray), provides alternatives to increasing food and nutritional security for

resource-poor farmers who are vulnerable to climatic variability (Mwale et al., 2020). Tepary bean has adapted to high temperatures and drought through diversification and domestication in Arid America (Nabhan, 1985; Pratt and Nabhan, 1988; Muñoz et al., 2006), resulting in distinctive morphological and physiological traits.

Tepary bean has been shown to have strong agronomic potential under stress conditions, as evidenced by biomass production, yield, assimilate translocation, osmolyte accumulation, increased photosynthetic efficiency, root depth, and stomatal resistance (Mwale et al., 2020), as well as nutritional content comparable to common bean (López-Ibarra et al., 2021). Tepary bean genetic resources are conserved in the USDA's genebank, which currently includes 204 accessions, 58 of which are *tenuifolius* and 148 of which are *acutifolius* (Bornowsky et al., 2023).

In Alliance Bioversity International and the International Center for Tropical Agriculture in Colombia (CIAT), the genetic resources program has 326 cultivated and wild accessions of tepary beans (Mhlaba et al., 2018). However, germplasm evaluation has been limited (Rainey and Griffiths, 2005; Suarez et al., 2021; Conejo-Rodríguez et al., 2022; Bornowsky et al., 2023). The enormous number of accessions in genebank collections poses one of the most significant problems to germplasm characterization and evaluation (Conejo - Rodriguez et al., 2024). The focused identification germplasm strategy (FIGS) offers an original approach to increasing the potential of collecting genebank accessions exhibiting specific adaptive traits (Street et al., 2016; Conejo - Rodríguez et al., 2024). The reasoning is that specific environments influence natural selection for or against adaptive evolution of crop traits in both wild and cultivated populations (Conejo - Rodríguez et al., 2024; Conejo - Rodríguez et al., See Chapter three). As a consequence, the spatial distribution of plant adaptive traits should be predicted using

ecogeographic profiles (Stenberg and Ortiz, 2020), allowing for the prioritization of actions that may show potential adaptations to extreme environments. This will reduce the number of accession candidates during evaluation; however, the variation of adaptive functional phenome has received little attention. Due to their ordinal and qualitative nature, the classic descriptors used in the characterization process have limited informative value. Allelic variation of various traits in genetic resources, including those for evaluation, such as high temperature and drought tolerance, needs efficient methods based on highly accurate phenotyping and quick modern data analysis techniques (Basavaraj and Rane, 2020; Nguyen and Norton, 2020).

Characterization in genebanks frequently employs lists of adapted classical morphological descriptors that have been used on many cultivated plants since the 1980s, based on highly heritable traits of the collected accessions (IBPGR, 1985). In the case of tepary bean, IBPGR (1985) created these descriptors, which are primarily based on metric, scalar, state, and presence-absence measurements, most of which are qualitative in character, restricting analytical capacity and integration with genomic data (Garcia-Fernandez et al. 2021). In recent years, the number of scientific papers that have worked on the development of analytical tools and techniques to detect traits of interest for plant breeding has hugely increased (Araus and Cairns 2014; Reynolds and Langridge 2016). Little tools have been conceptualized using biology knowledge as a discipline that truly is part of the omics sciences (Houle et al., 2010; Zavafer et al., 2023), or focusing research questions with attention on the evolution of adaptive traits and demonstrating their value for genebanks or breeders (Conejo - Rodriguez et al., 2024; Conejo - Rodríguez et al., See Chapter three).

Phenomics is the biological discipline that focuses on the study of phenomes (Zavafer et al., 2023). If the phenome is a set of phenotypes that originate from the relationship between

genotype (G), time (t) and environment (E) (Zavafer et al., 2023), in the context of plant genetic resources, the accessions collected today present the phenotypic traits that allowed adaptation to ecological conditions determined/conserved in time and by environment pressure; therefore, these phenotypic traits are mostly constitutive. Phenotypic variation of germplasm collections (characterization) would thus allow us to evaluate the genome expression of accessions that are part of the same species and to identify, evaluate and apply constitutive functional traits that relate diverse adaptation strategies from/for the environments in which they evolved (Conejo - Rodriguez et al., 2024; See Chapter 2 and 3).

The high throughput phenotyping employing proximal remote sensing has transformed germplasm evaluation, particularly in crop breeding collections (Nguye and Norton, 2020; Yang et al., 2020; Furbank and Tester, 2011). Recent developments in high-throughput phenotyping (HTTP) have facilitated crop improvement, particularly for some physiological traits (Araus and Cairns 2014; Reynolds and Langridge 2016). Several studies have been published using multispectral, hyperspectral and fluorometer sensors. These publications evidenced the phenotypic diversity of value of adaptive traits in resistance to high temperatures and drought. The positive conclusions were based on the prediction and accessions classification together with their agronomic potential using machine learning (ML - Machine Learning) techniques (Lopez-Cruz et al., 2020; Fernandez-Calleja et al., 2020; Lobos et al., 2019; Zhang et al., 2019; Hennessy et al., 2020; Sun et al., 2019; Kuhlgert et al., 2016), however exploration in genebanks has been limited. In addition, the more complex integration of ecogeographic, phenotypic and genomic structure data in genebanks is rarely described and the studies on beans have concentrated on the common bean, specifically on the adaptive traits associated with climate change (Beebe et al., 2011). Phenotypic diversity is closely related to species evolution and

adaptation to functional traits (Mammola et al., 2021). Functional diversity is a component of biological diversity and is related to the study of morpho-physiological variation of species associated with the ecosystem (Garnier et al., 2016). It is based on multivariate techniques and diversity indices that relate the variation of functional traits in one or more populations in order to relate ecogeographic assemblages with phenotypic variation.

Tepary bean is of particular interest in the genetic improvement of the common bean; nonetheless, tepary also has the potential to be a food source in different regions of the world where environmental circumstances have a direct impact on legume productivity (Bornowsky et al., 2023). There are knowledge gaps in some species, particularly regarding its adaptive evolution and phenotyping variation due to natural selection and domestication processes influenced by the environment, time and human activity (Conejo - Rodriguez et al., 2024).

In this study, we propose standardized operational techniques for documenting germplasm collections, as well as how to combine varied data and data mining research that contribute to responding to research questions during than gathering phenotypic data in germplasm collection characterization. We have standardized these techniques and offered comprehensive “work-flows.” Our research is organized into four chapters in which we standardize, create, and validate strategies for managing germplasm collections based on phenome variation. The first chapter, *“Digital descriptors sharpen classical descriptors, for improving genebank accession management: A case study on Arachis spp., and Phaseolus spp.”* (Conejo - Rodriguez et al., 2024), establishes a standard procedure for germplasm characterization that incorporates digital descriptors, artificial intelligence, and functional diversity in *Arachis* and *Phaseolus* accessions. In the second chapter, *“Representativeness functional phenome and ecogeographic diversity in genebanks: A case study on the tepary bean collection”* we explore the methodological

development integrating geographic distribution, ecology, genetic structure, and phenotypic variation with the aim of understanding and discovering phenome variation in wild and cultivated tepary bean accessions and their evolutionary adaptation process, and propose the tepary bean functional trait collection of the International Center for Tropical Agriculture (CIAT) Future Seeds genebank. In the third chapter, "*Sun-induced Chl fluorescence as an index selection of genotypes to light use efficiency in physiological breeding: Tepary bean functional trait collection case study*", we explore how sun-induced Chl fluorescence (SIF) can discover light use efficiency of in germplasm collections and in selection processes of genotypes with differential potential in photophysical efficiency, contributing to the development of selection indices for physiological breeding. Finally, we strongly feel that plant breeding requires the use of crop wild relatives (CWR). In chapter four, "*Using phenomics to identify and integrate traits of interest for better-performing common beans: a validation study on an interspecific hybrid and its Acutifolii parents*" (Conejo - Rodriguez et al., 2022), we develop a methodology to maintain desired traits of wild relatives in interspecific hybridizations based on phenomic proportions using multivariate techniques and artificial intelligence. This research presents a novel approach to understanding phenome variation in germplasm collections, as well as how doing so efficiently adds to germplasm documentation, use, and distribution. We are aware of the existing problems in ensuring characterisation or processes that go beyond conservation; nonetheless, we recognize the significance of the phenomics era, which, when combined with genomics and spatial data analysis, can contribute to *seed solutions* for improved crop adaptation. This is where the story begins.

1. 1 Research question

Are the functional traits of the conserved tepary bean germplasm collection affected by adaptive evolutionary processes driven by natural selection and human, and how novel tools (high throughput plant phenotyping, artificial intelligence, and functional diversity indices) support understanding the phenome variation in both wild and domesticated accessions?

1.2 Hypothesis

If the functional traits of the conserved tepary bean germplasm collection are conditioned by adaptive evolutionary processes driven by natural selection and human intervention, then phenomics, high throughput plant phenotyping, artificial intelligence, and functional diversity will reveal significant phenome variation between wild and domesticated accessions.

1.3 General aims

We will determine the extent to which adaptive evolutionary processes driven by natural selection and/or by human intervention affected the selected functional traits of the conserved tepary bean germplasm collection. We also want to present phenomics, high throughput plant phenotyping, artificial intelligence, and functional diversity to elucidate phenome variation in both wild and domesticated accessions.

1.4 Specific aims

1. We will propose a methodology for improving the characterization of bean (*Phaseolus*) and peanut (*Arachis*) accessions in the CIAT genebank by identifying phenomic descriptors comparable to classical ones. We will use AI models to pinpoint key descriptors for higher characterization precision and functional diversity quantification, and suggest ways of incorporating this methodology into the genebank workflow.
2. To assess the extent to which the functional phenome variation in the tepary bean (*Phaseolus acutifolius* A. Gray) collection is related to its geographic distribution, genetic structure, and ecological clusters.
3. From the collection of functional traits of tepary bean (*Phaseolus acutifolius* A. Gray) create a selection index for the light use efficiency in that contributes to select accessions with adaptive potential to drought conditions.
4. To develop a *Phaseolus*-targeted framework for detecting phenomic proportions of interspecific hybrids relative to their parental lines, we will use multivariate and machine learning methods to characterize and classify three parental line accessions as well as their interspecific hybrid accession.

1.5 General conceptual aspects

I elaborated the thesis from a conceptual standpoint. I will attempt to provide a conceptual overview of the previous work being done in the field of phenomics and how this can facilitate the characterization and evaluation of germplasm in genebanks.

1.6 Evolutionary adaptation in plants and crops

According to Anderson and Song (2020), plants are able to adjust to the direct and indirect effects of climate change on extinction risk, agricultural sustainability, and food security. Centers of agricultural diversity, which have been influenced over time by farmers, the environment, and evolutionary processes, are linked to genetic variation in crops (Brush, 2004). Major changes in adaptation capacity to a range of biotic and abiotic circumstances may be attributed to phenotypic plasticity, evolution, and gene flow, among other potential mechanisms (Mercer & Perales, 2010). When populations live in various environments, the spatial variation of abiotic and biotic circumstances may promote the evolution of local adaptation (Leimu and Fischer 2008). It is true that a variety of taxonomic groupings have demonstrated local adaptation (Wadgymar et al., 2022). Functional traits have a strong influence on this adaptation. According to Violle et al (2007), plant functional traits are those that affect growth, reproduction, and survival and make it possible for predicting how plants will react and adapt to changing environments at all levels of organization, including organs, species, and ecosystems. Indeed, functional traits can assist predicted species ranges, vital rates, and climate change responses (Pollock et al., 2012; He et al., 2020). These functional traits may differ within a species (Westerband et al., 2021). Abiotic and biotic factors promote phenotypic plasticity, which drives intraspecific variation in functional traits within and across individuals (Westerband et al., 2021).

At the population and community levels, intraspecific variation can influence overall plant performance. Over time, local genetic adaptation and epigenetic influences promote intraspecific variation, changing community dynamics and generating changes in species distributions (Westerband et al., 2021). Variation in functional traits among groups has revealed adaptive patterns linked to the climate of origin (Ahrens et al., 2020). Some traits, such as efficient use of water and light, will be heritable and show large variations over climatic gradients, whereas others, such as specific leaf area, will vary more within and between populations, originating in lower levels of heritability (Ahrens et al., 2020). Thus natural selection and domestication directed adaptive capability while accounting for intraspecific diversity in domesticated and wild relatives. The emergence of agriculture altered our species' relationship with the planet by creating new ecological niches of agricultural systems that became regions of coevolution (Fuller et al., 2023), in which domesticated plants and animals evolved alongside a series of commensals, such as pests and weeds that obtain their calories from the human-crop system. These new agricultural ecologies can be thought of as food web modifications that support human populations at increased population densities, allowing for the geographical extension of domesticated crops (Fuller et al., 2023). Domestication processes took millennia, possibly 3,000 or more generations of plants and up to 150 human generations (Fuller et al., 2023). The rates of domestication of various cereals are significant because they show that the period of coevolution occurred during significant climatic, cultural, and technological change, implying that the "origins" of domestication are not due to a single cause (Fuller et al., 2023). Understanding phenotypic variation in germplasm collections might help examine the phenome of conserved species (Conejo-Rodríguez et al., 2024) and address concerns about domestication and adaptation.

1.7 Evolution and botanical characteristics of Phaseolus

Beans (*Phaseolus* spp.), particularly the common bean *P. vulgaris* L., are the most significant legumes for direct human consumption worldwide, providing protein and micronutrients (Vaz Patto et al., 2015). The five domesticated *Phaseolus* species differ in several ways, including their life histories, levels of interbreeding, ecological adaptations, genetic diversity, and amount of domestication (Parker and Gepts, 2021). As *Phaseolus* beans spread beyond their native areas in the tropical and subtropical Americas, they underwent selection for traits that allow differentiation. Domesticated *Phaseolus* species differ in their sensitivity to light use efficiency (LUE), water use efficiency (WUE), and variations in temperature and precipitation regimes, which are directly related to cultural usage (Parker et al. 2021).

The five domesticated species are: *P. acutifolius* A. Gray (tepary bean), *P. coccineus* L. (scarlet bean), *P. dumosus* Macfady (year bean), *P. lunatus* L. (Lima bean), and *P. vulgaris* L. (common bean). The Lima bean (Motta-Aldana et al., 2010) and common bean (Bitocchi et al., 2013) were domesticated independently in Mesoamerica and the Andes of South America, while the tepari (Muñoz et al., 2006), scarlet runner (Guerra-García et al., 2017), and frijol año (Schmit and Debouck, 1991) were only domesticated once in Mesoamerica. The five domesticated species' original domestication events occurred roughly 8,000 years ago (Mamidi et al., 2011).

Following domestication, landraces spread throughout the Americas, and after 1493, *P. vulgaris*, *P. lunatus*, and, to a lesser extent, *P. coccineus* were transferred to various portions of the Old World (Kaplan and Kaplan, 1992). Existen alrededor de ochenta especies de Phaseolus que comparten diez rasgos diagnósticos (Debouck, 2019). The morphology of trifoliate leaves with stipules, primary bracts on the peduncle, and the rachis present until anthesis are among the distinctive traits, but the largest variability among wild and cultivated species of the genus is determined by geographic distribution (Debouck, 2019).

1.8 Tepary bean (*Phaseolus acutifolius* A. Gray)

Tepary bean (*P. acutifolius* A. Gray) is more heat and drought tolerant than common beans (Thomas et al., 1983; Mwale et al., 2019). It demonstrates that the cultivar contains traits that can be introduced into common bean to improve abiotic stress tolerance (Moghaddam et al., 2021; Souter et al., 2017). The common bean demonstrates low adaptation to high temperatures because of its restricted adaptation range (Beebe et al., 2008; Chacon et al., 2005). As a result, one ongoing issue for bean breeding programs is to increase the use of Tepary bean genes and/or strategically enhance breeding effort on this species (Debouck, 2019). Tepary bean genetic resources are conserved in the USDA genebanks, which currently have 204 accessions, and the Alliance of Bioversity International and International Center for Tropical Agriculture (CIAT) genetic resources program, which has 326 accessions (Mhlaba et al., 2018). CIAT collection include domesticated and wild accessions associated with the wild varieties of *P. acutifolius* var *acutifolius* and *P. acutifolius* var *tenuifolius*. Tepary bean (*P. acutifolius*) has lately received attention in modern crop breeding efforts as a source of genetic characteristics for biotic and abiotic stress tolerance in its sister species, common bean (*P. vulgaris*) (Debouck, 2019, Cruz et

al 2023). Tepary bean's adaptation to high ambient temperatures and dry, arid conditions stems from its natural habitat in the Sonoran Desert (Thomas, 1983), as an alternative source of protein in areas threatened by a warmer, drier future (Moghaddam et al., 2021), and as a genetic resource for abiotic stress tolerance to produce more resilient common and tepary beans in a changing climate (Buitrago-Bita et al., 2021). Currently, tepary bean is only produced on a small scale, mostly by dryland subsistence farmers in the years (Nabhan and Felger, 1978). However, it is improbable that the crop will be found planted in arid American regions due to its relocation. Tepary bean breeding efforts have recently focused on resistance to rust, common bacterial blight, yield, acclimation, seed size, and quality (Porch et al., 2013), demonstrating the bean's potential for improvement through breeding.

1.9 Tepary bean domestication

Tepary bean is a crop known for its abiotic stress tolerance, which is thought to be mostly owing to its origin and evolution in the Sonoran Desert's high temperatures and dry circumstances (Moghaddam et al., 2021). Assessments of genetic diversity based on structure and phylogenetic inference have confirmed that domesticated tepary beans were genetically most closely related; however, tepary bean is believed to have undergone multiple domestications, suggesting the possibility of one in Central America and one in Mexico/USA, implying that the geographical dispersion of domesticated tepary was the result of a single domestication event in central Mexico followed by migration (Gujaria-Verma et al., 2016, Bornowski et al., 2023). Bornowski et al. (2023) identified six genetic clustering of tepary bean, four for wild accessions: (I) Sonora-Sinaloa *acutifolius*, (II) Southeast Arizona *tenuifolius*, (III) Chihuahua *acutifolius*, and (IV) Durango. The cultivated accessions were divided into two groups: (I) Sonora-Sinaloa and

(II) Central America. Furthermore, it should be highlighted that they provided a number of ADMIXTURE - accessions that did not fit into any genetic cluster, particularly accessions that were characterized as transitional between wild and domesticated (Gujaria-Verma et al., 2016; Bornowski et al., 2023). There are significant gaps in knowledge in tepary domestication due to the absence of articulation of genetic diversity with functional traits and the differentiation processes between desert and subtropical tepary beans (Conejo - Rodriguez et al., Chapter 5).

1.10 Crop phenomics

The crop phenomics is more closely linked to the employment of digital technologies in the acquisition of phenotypic traits (Jiang and Zhu et al., 2024). Houle et al. (2010) propose the first approach to phenomics, which establishes that it is the systematic acquisition of phenotypic traits that relate to the interaction with the environment; however, conceptual shortcomings exist in the omics sciences (Zavafer et al., 2023), because the use of technologies does not imply the conceptualization of a science. Phenomics, as defined by its etymology, is a discipline that investigates the phenome. The phenome can be described as all of the phenotypes that a genotype or genome can exhibit as a result of the interaction between the environment and its development and growth (Zavafer et al., 2023). Zavafer et al. (2023) conceive of phenomics, its biological interpretation, and its foundation based on the link between time, space, and phenotype variation.

1.11 High throughput plant phenotyping

High-throughput plant phenotyping (HTPP) with proximal remote sensing is a non-destructive and non-invasive approach (White et al., 2012), relying primarily on information provided by visible/near-infrared radiation (VIS-NIR), image analysis, robotics, data mining, and artificial intelligence (Basavaraj and Rane, 2020; Araus et al., 2014). The ability to use phenomic tools and technologies to translate "knowledge of abiotic stress tolerance mechanisms" into "trait identification" will allow phenotyping to identify tolerant genotypes (Basavaraj and Rane, 2020) under controlled, natural environmental conditions.

Recent developments in HTPP, particularly those based on field data, have improved crop breeding using physiological parameters (Araus and Cairns 2014; Reynolds and Langridge 2016). Physiological trait breeding entails designing a plant type, crossing parents with complicated but complementary traits, and achieving the cumulative influence of genes on yield (Fernandez-Calleja et al., 2020). Several research are now underway that use proximal remote sensing to highlight the phenotypic variation in adaptive traits.

Meacham-Hensold et al. (2019) used hyperspectral reflectance to evaluate photosynthetic capability in tobacco (*Nicotiana tabacum* L.). The authors conducted the investigation using two (2) types of sensors: (I) Fieldspec4 with spectral resolutions of 3 nm in the visible and NIR ranges (350 - 1000 nm), and 8 mm in the near infrared (1000 - 2500 nm). Similarly, low-cost proximal remote sensors have been designed to assess photosynthetic capacity with high accuracy and distinguish between genotypes (Kuhlgert et al., 2016; Deva et al., 2020; Fernández-Calleja et al., 2020). Deva et al. (2020) used the MultispeQ proximal sensor to evaluate high temperature sensitive and tolerant genotypes of common bean (*Phaseolus vulgaris* L.), with leaf temperature differential as a response variable (LTD). The scientists demonstrated

that using this parameter can reveal tolerant genotypes while also simulating air temperature and relative humidity. Fernández-Calleja et al. (2020) used the MultispeQ sensor to analyze the efficient utilization of light in barley (*Hordeum vulgare* L.) hybrids and parents during drought. A number of photosynthetic variables were established by the authors, including the following: photosystem II yield (Φ_{II}), non-photochemical quenching (Φ_{NPQ}), energy loss by heat dissipation (Φ_{NO}), relative chlorophyll content (Chl), the fraction of "open" reaction centers [qL , $qL = (((Fm - Fs)/(Fm - F0)) * (F0 / Fs))$], steady-state fluorescence (Fs), maximum fluorescence (Fm), minimum fluorescence (F0), maximum variable fluorescence (Fv, $Fv = Fm - F0$), the coefficient of photochemical quenching (qP , $qP = (Fm - F0)$), and the fraction of "open" reaction centers (Fv, $Fv = Fm - F0$). Under situations of drought, they were able to clearly identify differences in the phenotypic responses of parents and hybrids, particularly with regard to the ΔNPQ trait.

1.12 Functional diversity

Functional diversity studies look at how morpho-physiological traits affect individual fitness both within and between populations (Bradshaw, 1987). In this sense, the term "function" can refer to both trophic levels and evolutionary processes, specifically the function of adaptations (Laureto et al., 2015). Functional diversity integrates and connects ecogeographic assemblages and phenotypic variation in order to better understand species that possess crucial traits that enable them to adapt to climate variability. Functional diversity research has been performed to answer two major questions: (a) how species influence ecosystem functioning and (b) how species respond to environmental change (Hooper et al., 2000). Functional diversity is measured using indices including richness, divergence, and regularity (Mammola et al., 2021). "Richness" refers

to metrics that reflect the total of differences among observations; 'divergence' refers to measurements that reflect average differences among observations; and 'regularity' refers to metrics that indicate the regularity of differences among observations. Despite its importance to crop biodiversity, functional diversity has received little attention in genebanks. These strategies will serve to identify accessions with adaptive potential related to ecogeographic groupings, which will aid in understanding phenotypic variation from an evolutionary perspective, allowing the detection of groups of accessions with adaptive traits.

Functional diversity enables the identification of species/accessions that exhibit highly discriminating functional phenotypic traits, increasing the likelihood of discovering functional traits with key adaptive potential for selecting useful accessions for genetic improvement and contributing to FIGS models (Sunitha et al., 2024). Studies have focused on animal populations in varied environments (Teixidó et al., 2018; Mouillot et al., 2014; Gardarin et al., 2021), but no studies have documented functional diversity applied to genetic resources for food and agriculture.

1.13 Representativeness of genetic, ecogeographic and functional diversity in genebanks

The identification of germplasm/genotypes tolerant to a given trait is a prerequisite for breeding; however, one of the major constraints in breeding is the identification of reliable and cost-effective phenotyping methods to facilitate this identification (Wahid et al., 2007; Basavaraj and Rane, 2020). Currently, methodical approaches are being established to rank possible accessions in genebank collections, working collections of genetic resources, and breeding programs. One of these is the Focused Identification Germplasm Strategy (FIGS), which offers a

novel approach to increasing the likelihood of acquiring genebank accessions with certain adaptive traits (Street et al., 2016). The reason is that particular conditions mediate natural selection for or against the adaptive evolution of crop traits in both wild and domesticated populations. As a result, ecogeographic profiles should predict the spatial distribution of plant adaptive traits (Stenberg and Ortiz, 2020), allowing us to prioritize accessions that are likely to exhibit adaptations to abiotic and biotic stresses and select a small number of accessions for germplasm evaluation.

Few studies have studied morpho-physiological traits in entire collections of tepary bean from both wild and domesticated accessions (Suarez et al. 2021), with a lesser degree employing phenotypic descriptors (Conejo-Rodríguez et al., 2022). Suarez et al. (2021) phenotypically examined the entire collection of *P. acutifolius* accessions from the Bioversity Alliance and CIAT genebanks. In this study, traits such as leaf area, specific leaf area, days to flowering, days to physiological maturity, pod weight, and seed weight were used to differentiate between domesticated and wild *P. acutifolius* var *tenuifolius* and var *acutifolius* in humid tropical conditions in the Colombian Amazon. One of the most difficult aspects of evaluating accessions in genebanks is identifying accessions with functional traits related to adaptability under abiotic and biotic stress conditions. One technique is to combine ecogeographic and phenotypic traits, which allows for the selection of highly dissimilar accessions and increases the likelihood of selecting potential parental accessions for abiotic stress breeding. The integration of phenotypic and eco-geographical data has centered on ecological studies, which have created various approaches and methods for interpreting functional diversity, a critical component of biological diversity (Mammola et al., 2021) that has received little attention in genebanks. FIGS (Focused Identification of Germplasm Strategy) is a mechanism for identifying and pre-selecting

accessions as candidate genotypes based on original/historical ecological data from each collecting location (Mackay et al., 2004). Candidate accessions for a specific stress factor are selected by comparing ecological data from the accession's original collection to the ecological distribution of that stress (Angessa and Li, 2016). This strategy takes advantage of genetic variety resulting from adaptation, in which environmental stressors are the driving force (Hill et al., 1998).

Numerous predictive characterisation studies have been conducted. For example, Garcia-Sánchez et al. (2019) developed a model of cyanogenesis expression in *Trifolium repens* L, identifying 470 accessions with a high probability of being acyanogenic among 160 a priori evaluated accessions and 3,022 non-evaluated accessions. Similarly, predictive characterisation was performed in *Aegilops* to detect the collection's susceptibility to drought and salinity (Garcia et al., 2017). Similarly, Maize for diverse abiotic adaptations (Tapia et al., 2015), *Vicia faba* L., *P. vulgaris* L., and *Triticum aestivum* L. for tolerance to water deficit (Khazaei et al., 2013; Cortés et al., 2013; Endresen et al., 2012), and only wheat for resistance to rust (*Puccinia striiformis*) (Bari et al., 2013; Endresen et al., 2012). FIGS can be used to execute pre-selection processes on tepary bean accessions that are possibly tolerant to abiotic stress; however, this needs the investigation of adaptive functional traits as well as research that examine tepary bean germplasm collections (Salazar et al., 2021). The exploration of functional traits from ecogeographic assemblages and phenotypic data will allow the selection of accessions that span the species' ecological ranges, allowing for the identification of adaptive functional traits that allow discrimination between these ecogeographic groups by identifying the response to increases in abiotic stress (Sunitha et al., 2024). These accessions will be the starting point as FIGS training populations that will allow predictive modeling utilizing artificial intelligence

from machine learning approaches, assisting in determining accessions with a high chance of differential morpho-physiological response to increased stress (Sunitha et al., 2024). Given the foregoing, the direct relationship between phenome/genome variation is influenced by ecological variation and human selection processes during crop domestication. The combination of proximal remote sensing and high-throughput phenotyping strategies will help to answer biologically based questions, allowing for germplasm characterization and evaluation processes that incorporate data science and artificial intelligence to support the development of decision support tools during breeding processes.

1.14 References

- Araus, J. L., & Cairns, J. E. (2014). Field high-throughput phenotyping: the new crop breeding frontier. *Trends in plant science*, 19(1), 52-61.
- Ashraf, M. H. P. J. C., & Harris, P. J. (2013). Photosynthesis under stressful environments: an overview. *Photosynthetica*, 51(2), 163-190.
- Barabás, G., Parent, C., Kraemer, A., Van de Perre, F., & De Laender, F. (2022). The evolution of trait variance creates a tension between species diversity and functional diversity. *Nature communications*, 13(1), 1-10.
- Basavaraj, P. S., & Rane, J. (2020). Avenues to realize potential of phenomics to accelerate crop breeding for heat tolerance. *Plant Physiology Reports*, 1-17.
- Beebe, S. E., Rao, I. M., Cajiao, C., & Grajales, M. (2008). Selection for drought resistance in common bean also improves yield in phosphorus limited and favorable environments. *Crop Science*, 48(2), 582-592.

Beebe, S., Ramirez, J., Jarvis, A., Rao, I. M., Mosquera, G., Bueno, J. M., & Blair, M. W. (2011). Genetic improvement of common beans and the challenges of climate change. *Crop adaptation to climate change*, 356-369.

Bhattacharya, A. (2019). Effect of high temperature on crop productivity and metabolism of macro molecules. Academic Press.

Bitocchi, E., Bellucci, E., Giardini, A., Rau, D., Rodriguez, M., Biagetti, E., ... & Papa, R. (2013). Molecular analysis of the parallel domestication of the common bean (*P. haseolus vulgaris*) in Mesoamerica and the Andes. *New Phytologist*, 197(1), 300-313.

Bornowski, N., Hart, J. P., Palacios, A. V., Ogg, B., Brick, M. A., Hamilton, J. P., ... & Porch, T. (2023). Genetic variation in a tepary bean (*Phaseolus acutifolius* A. Gray) diversity panel reveals loci associated with biotic stress resistance. *The Plant Genome*, 16(3), e20363.

Buitrago-Bitar, M. A., Cortés, A. J., López-Hernández, F., Londoño-Caicedo, J. M., Muñoz-Florez, J. E., Muñoz, L. C., & Blair, M. W. (2021). Allelic Diversity at Abiotic Stress Responsive Genes in Relationship to Ecological Drought Indices for Cultivated Tepary Bean, *Phaseolus acutifolius* A. Gray, and Its Wild Relatives. *Genes*, 12(4), 556.

Chacon, M.I, Pickersgill, B, Debouck, D. G. (2005). Domestication patterns in common bean (*Phaseolus vulgaris* L.) and the origin of the Mesoamerican and Andean cultivated races. *Theoretical and Applied Genetics*, 110(3), 432-444.

Challinor, A. J., Watson, J., Lobell, D. B., Howden, S. M., Smith, D. R., & Chhetri, N. (2014). A meta-analysis of crop yield under climate change and adaptation. *Nature Climate Change*, 4(4), 287-291.

Chaudhary, S., Devi, P., Bhardwaj, A., Jha, U. C., Sharma, K. D., Prasad, P. V., ... & Nayyar, H. (2020). Identification and characterization of contrasting genotypes/cultivars for developing heat tolerance in agricultural crops: Current status and prospects. *Frontiers in Plant Science*, 11, 1505.

Cruz, S., Lobatón, J., Urban, M. O., Ariza-Suarez, D., Raatz, B., Aparicio, J., ... & Beebe, S. (2023). Interspecific common bean population derived from *Phaseolus acutifolius* using a bridging genotype demonstrate useful adaptation to heat tolerance. *Frontiers in Plant Science*, 14, 1145858.

Debouck, D. G. (2019). Recursos genéticos del fríjol para su progreso.

Debouck, D. G. (2021). PHASEOLUS BEANS (LEGUMINOSAE: PHASEOLEAE). *Journal of the Botanical Research Institute of Texas*, 15(1), 73-111.

Deva, C. R., Urban, M. O., Challinor, A. J., Falloon, P., & Svitákova, L. (2020). Enhanced leaf cooling is a pathway to heat tolerance in common bean. *Frontiers in plant science*, 11, 19.

Driedonks, N., Xu, J., Peters, J. L., Park, S., & Rieu, I. (2015). Multi-level interactions between heat shock factors, heat shock proteins, and the redox system regulate acclimation to heat. *Frontiers in plant science*, 6, 999.

Fernández-Calleja, M., Monteagudo, A., Casas, A. M., Boutin, C., Pin, P. A., Morales, F., & Igartua, E. (2020). Rapid on-site phenotyping via field fluorimeter detects differences in photosynthetic performance in a hybrid—parent barley germplasm set. *Sensors*, 20(5), 1486.

Foster, N., & Cordell, L. S. (Eds.). (1992). Chilies to chocolate: food the Americas gave the world. University of Arizona Press.

García-Fernández C, Campa A, Garzón AS, Miklas P, Ferreira JJ. GWAS of pod morphological and color characters in common bean. *BMC plant biology*. 2021 Dec;21(1):1-3.

Gardarin, A., Pigot, J., & Valantin-Morison, M. (2021). The hump-shaped effect of plant functional diversity on the biological control of a multi-species pest community. *Scientific reports*, 11(1), 1-13.

Garnier, E., Navas, M., & Grigulis, K. (2016). Trait-based ecology: definitions, methods, and a conceptual framework. In *Plant Functional Diversity* (pp. 9-25). Oxford University Press.

Guerra-García, A., Suárez-Atilano, M., Mastretta-Yanes, A., Delgado-Salinas, A., & Piñero, D. (2017). Domestication genomics of the open-pollinated scarlet runner bean (*Phaseolus coccineus* L.). *Frontiers in plant science*, 8, 1891.

Gujaria-Verma, N., Ramsay, L., Sharpe, A. G., Sanderson, L. A., Debouck, D. G., Tar'an, B., & Bett, K. E. (2016). Gene-based SNP discovery in tepary bean (*Phaseolus acutifolius*) and common bean (*P. vulgaris*) for diversity analysis and comparative mapping. *BMC genomics*, 17, 1-16.

Hazekamp, T., Serwinski, J., & Alercia, A. (1997). Multicrop passport descriptors. Part A. Discussions and Recommendations, 40.

Hennessy, A., Clarke, K., & Lewis, M. (2020). Hyperspectral classification of plants: a review of waveband selection generalisability. *Remote Sensing*, 12(1), 113.

Hooper, D. U., Chapin III, F. S., Ewel, J. J., Hector, A., Inchausti, P., Lavorel, S., ... & Wardle, D. A. (2005). Effects of biodiversity on ecosystem functioning: a consensus of current knowledge. *Ecological monographs*, 75(1), 3-35.

Houle, D., Govindaraju, D. R., & Omholt, S. (2010). Phenomics: the next challenge. *Nature reviews genetics*, 11(12), 855-866.

IBPGR 1985. *Phaseolus acutifolius* descriptor list. International Board for Plant Genetic Resources, 1985.

Jiang, N., & Zhu, X. G. (2024). Modern phenomics to empower holistic crop science, agronomy, and breeding research. *Journal of Genetics and Genomics*.

Kuhlgert, S., Austic, G., Zegarac, R., Osei-Bonsu, I., Hoh, D., Chilvers, M. I., ... & Kramer, D. M. (2016). MultispeQ Beta: a tool for large-scale plant phenotyping connected to the open PhotosynQ network. Royal Society open science, 3(10), 160592.

Laureto, L. M. O., Cianciaruso, M. V., & Samia, D. S. M. (2015). Functional diversity: an overview of its history and applicability. Natureza & Conservação, 13(2), 112-116.

Li, B., Gao, K., Ren, H., & Tang, W. (2018). Molecular mechanisms governing plant responses to high temperatures. Journal of integrative plant biology, 60(9), 757-779.

Lobos, G. A., Escobar-Opazo, A., Estrada, F., Romero-Bravo, S., Garriga, M., Del Pozo, A., ... & Caligari, P. (2019). Spectral reflectance modeling by wavelength selection: studying the scope for blueberry physiological breeding under contrasting water supply and heat conditions. Remote Sensing, 11(3), 329.

Lopez-Cruz, M., Olson, E., Rovere, G., Crossa, J., Dreisigacker, S., Mondal, S., ... & de Los Campos, G. (2020). Regularized selection indices for breeding value prediction using hyper-spectral image data. Scientific reports, 10(1), 1-12.

López-Ibarra, C., de Jesús Ruiz-López, F., Bautista-Villarreal, M., Báez-González, J. G., Romero, B. A. R., González-Martínez, B. E., ... & Vázquez-Rodríguez, J. A. (2021). Protein Concentrates on Tepary Bean (*Phaseolus acutifolius* Gray) as a Functional Ingredient: In silico Docking of Tepary Bean Lectin to Peroxisome Proliferator-Activated Receptor Gamma. Frontiers in Nutrition, 8.

Magnan, A. K., Pörtner, H. O., Duvat, V. K., Garschagen, M., Guinder, V. A., Zommers, Z., ... & Gattuso, J. P. (2021). Estimating the global risk of anthropogenic climate change. Nature Climate Change, 11(10), 879-885.

Magneville, C., Loiseau, N., Albouy, C., Casajus, N., Claverie, T., Escalas, A., ... & Villéger, S. (2022). mFD: an R package to compute and illustrate the multiple facets of functional diversity. *Ecography*, 2022(1).

Mamidi, S., Rossi, M., Annam, D., Moghaddam, S., Lee, R., Papa, R., & McClean, P. (2011). Investigation of the domestication of common bean (*Phaseolus vulgaris*) using multilocus sequence data. *Functional Plant Biology*, 38(12), 953-967.

Mammola, S., Carmona, C. P., Guillerme, T., & Cardoso, P. (2021). Concepts and applications in functional diversity. *Functional Ecology*, 35(9), 1869-1885.

Mathur, S., Agrawal, D., & Jajoo, A. (2014). Photosynthesis: response to high temperature stress. *Journal of Photochemistry and Photobiology B: Biology*, 137, 116-126.

Mhlaba, Z. B., Mashilo, J., Shimelis, H., Assefa, A. B., & Modi, A. T. (2018). Progress in genetic analysis and breeding of tepary bean (*Phaseolus acutifolius* A. Gray): A review. *Scientia Horticulturae*, 237, 112-119.

Moghaddam, S. M., Oladzad, A., Koh, C., Ramsay, L., Hart, J. P., Mamidi, S., ... & McClean, P. E. (2021). The tepary bean genome provides insight into evolution and domestication under heat stress. *Nature communications*, 12(1), 2638.

Moghaddam, S. M., Oladzad, A., Koh, C., Ramsay, L., Hart, J. P., Mamidi, S., ... & McClean, P. E. (2021). The tepary bean genome provides insight into evolution and domestication under heat stress. *Nature communications*, 12(1), 1-14.

Motta-Aldana, J. R., Serrano-Serrano, M. L., Hernández-Torres, J., Castillo-Villamizar, G., Debouck, D. G., & Chacóns, M. I. (2010). Multiple origins of lima bean landraces in the Americas: evidence from chloroplast and nuclear DNA polymorphisms. *Crop Science*, 50(5), 1773-1787.

Mouillot, D., Villéger, S., Parravicini, V., Kulbicki, M., Arias-González, J. E., Bender, M., ... & Bellwood, D. R. (2014). Functional over-redundancy and high functional vulnerability in global fish faunas on tropical reefs. *Proceedings of the National Academy of Sciences*, 111(38), 13757-13762.

Muñoz, C., Blair, M. W., & Debouck, D. G. (2006). Hacia un mejor conocimiento del acervo genético de los Acutifolii.

Muñoz, L. C., Duque, M. C., Debouck, D. G., & Blair, M. W. (2006). Taxonomy of tepary bean and wild relatives as determined by amplified fragment length polymorphism (AFLP) markers. *Crop Science*, 46(4), 1744-1754.

Mwale, S. E., Shimelis, H., Mafongoya, P., & Mashilo, J. (2020). Breeding tepary bean (*Phaseolus acutifolius*) for drought adaptation: A review. *Plant Breeding*, 139(5), 821-833.

Nabhan, G. P. (1985). Native crop diversity in Aridoamerica: conservation of regional gene pools. *Economic Botany*, 39(4), 387-399.

Nabhan, G. P., & Felger, R. S. (1978). Teparies in southwestern North America. *Economic Botany*, 32(1), 3-19.

Nguyen, G. N., & Norton, S. L. (2020). Genebank phenomics: A strategic approach to enhance value and utilization of crop germplasm. *Plants*, 9(7), 817.

Paliwal, R., Abberton, M., Faloye, B., & Olaniyi, O. (2020). Developing the role of legumes in West Africa under climate change. *Current Opinion in Plant Biology*.

Parker, T. A., & Gepts, P. (2021). Population genomics of *Phaseolus* spp.: a domestication hotspot.

Parker, T. A., Lo, S., & Gepts, P. (2021). Pod shattering in grain legumes: emerging genetic and environment-related patterns. *The Plant Cell*, 33(2), 179-199.

Peñuelas, J., & Filella, I. (1998). Visible and near-infrared reflectance techniques for diagnosing plant physiological status. *Trends in plant science*, 3(4), 151-156.

Phillippo, D. M. (2022). multinma: Bayesian Network Meta-Analysis of Individual and Aggregate Data. R package version 0.5.0, doi: [10.5281/zenodo.3904454](https://doi.org/10.5281/zenodo.3904454).

Polanía Perdomo, J. A., Chaves, N., Cajiao V, C. H., Beebe, S. E., & Rao, I. M. (2014). Evaluación e identificación de líneas de frijol común (*Phaseolus Vulgaris L.*) tolerantes a altas temperaturas.

Porch, T. G., Beaver, J. S., Debouck, D. G., Jackson, S. A., Kelly, J. D., & Dempewolf, H. (2013). Use of wild relatives and closely related species to adapt common bean to climate change. *Agronomy*, 3(2), 433-461.

Pratt, R. C., & Nabhan, G. P. (1988). Evolution and Diversity of *Phaseolus acutifolius* Genetic Resources. In *Genetic resources of Phaseolus beans* (pp. 409-440). Springer, Dordrecht.

Rainey, K. M., & Griffiths, P. D. (2005). Evaluation of *Phaseolus acutifolius* A. Gray plant introductions under high temperatures in a controlled environment. *Genetic Resources and Crop Evolution*, 52(2), 117-120.

Reynolds, M., & Langridge, P. (2016). Physiological breeding. *Current Opinion in Plant Biology*, 31, 162-171.

Schmit, V., & Debouck, D. G. (1991). Observaciones sobre el origen del *Phaseolus polyanthus* Greenman. *Economic Botany*, 45, 345-364.

Smith, M. R., & Rao, I. M. (2021). Common bean. In *Crop Physiology Case Histories for Major Crops* (pp. 384-406). Academic Press.

Souter, J. R., Gurusamy, V., Porch, T. G., & Bett, K. E. (2017). Successful introgression of abiotic stress tolerance from wild tepary bean to common bean. *Crop Science*, 57(3), 1160-1171.

Sun, D., Cen, H., Weng, H., Wan, L., Abdalla, A., El-Manawy, A. I., ... & He, Y. (2019). Using hyperspectral analysis as a potential high throughput phenotyping tool in GWAS for protein content of rice quality. *Plant methods*, 15(1), 1-16.

Sunitha, N. C., Prathibha, M. D., Thribhuvan, R., Lokeshkumar, B. M., Basavaraj, P. S., Lohithaswa, H. C., & Anilkumar, C. J. G. R. (2024). Focused identification of germplasm strategy (FIGS): A strategic approach for trait-enhanced pre-breeding. *Genetic Resources and Crop Evolution*, 71(1), 1-16.

Teixidó, N., Gambi, M. C., Parravacini, V., Kroeker, K., Micheli, F., Villéger, S., & Ballesteros, E. (2018). Functional biodiversity loss along natural CO₂ gradients. *Nature communications*, 9(1), 1-9.

Thomas, C. V., Manshardt, R. M., & Waines, J. G. (1983). Teparies as a source of useful traits for improving common beans. *Desert plants*, 5(1), 43-48.

Thomas, C. V., Manshardt, R. M., & Waines, J. G. (1983). Teparies as a source of useful traits for improving common beans.

Traub, J., Porch, T., Naeem, M., Urrea, C. A., Austic, G., Kelly, J. D., & Loescher, W. (2018). Screening for heat tolerance in Phaseolus spp. using multiple methods. *Crop Science*, 58(6), 2459-2469.

Vacca, R. A., de Pinto, M. C., Valenti, D., Passarella, S., Marra, E., & De Gara, L. (2004). Production of reactive oxygen species, alteration of cytosolic ascorbate peroxidase, and impairment of mitochondrial metabolism are early events in heat shock-induced programmed cell death in tobacco Bright-Yellow 2 cells. *Plant physiology*, 134(3), 1100-1112.

Vaz Patto, M. C., Amarowicz, R., Aryee, A. N., Boye, J. I., Chung, H. J., Martín-Cabrejas, M. A., & Domoney, C. (2015). Achievements and challenges in improving the nutritional quality of food legumes. *Critical reviews in plant sciences*, 34(1-3), 105-143.

Wahid, A., Gelani, S., Ashraf, M., & Foolad, M. R. (2007). Heat tolerance in plants: an overview. *Environmental and experimental botany*, 61(3), 199-223.

Yang, W., Feng, H., Zhang, X., Zhang, J., Doonan, J. H., Batchelor, W. D., ... & Yan, J. (2020). Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives. *Molecular Plant*, 13(2), 187-214.

Zavafer, A., Bates, H., Mancilla, C., & Ralph, P. J. (2023). Phenomics: conceptualization and importance for plant physiology. *Trends in Plant Science*, 28(9), 1004-1013.

Zhang, C., Chen, W., & Sankaran, S. (2019). High-throughput field phenotyping of Ascochyta blight disease severity in chickpea. *Crop Protection*, 125, 104885.

2. Chapter two: Digital descriptors sharpen classical descriptors, for improving genebank accession management: A case study on *Arachis* spp. and *Phaseolus* spp.

Diego Felipe Conejo-Rodriguez^{1,2,3}, Juan José Gonzalez-Guzman¹, Joaquín Guillermo Ramirez-Gil⁴, Peter Wenzl¹ and Milan Oldřich Urban²

¹ Genetic Resources Program, International Center for Tropical Agriculture (CIAT), Palmira, Valle del Cauca, Colombia.

² Bean Physiology and Breeding Program, International Center for Tropical Agriculture (CIAT), Palmira, Valle del Cauca, Colombia

³ Universidad Nacional de Colombia sede Palmira, Facultad de Ciencias Agropecuarias, Palmira, Valle del Cauca, Colombia

⁴ Universidad Nacional de Colombia sede Bogotá, Facultad de Ciencias Agrarias, Departamento de Agronomía, Bogotá Colombia

Corresponding Author:

Diego Felipe Conejo Rodríguez
Milan Oldřich Urban

Email address: dfconejoro@unal.edu.co (DC); m.urban@cgiar.org (MOU)

This chapter has been published in the journal PLOS ONE (for more details please see:
<https://doi.org/10.1371/journal.pone.0302158>)

Abstract

High-throughput phenotyping brings new opportunities for detailed genebank accessions characterization based on image-processing techniques and data analysis using machine learning algorithms. Our work proposes to improve the characterization processes of bean and peanut accessions in the CIAT genebank through the identification of phenomic descriptors comparable to classical descriptors including methodology integration into the genebank workflow. To cope with these goals morphometrics and colorimetry traits of 14 bean and 16 forage peanut accessions were determined and compared to the classical International Board for Plant Genetic Resources (IBPGR) descriptors. Descriptors discriminating most accessions were identified using a random forest algorithm. The most-valuable classification descriptors for peanuts were 100-seed weight and days to flowering, and for beans, days to flowering and primary seed color. The combination of phenomic and classical descriptors increased the accuracy of the classification of *Phaseolus* and *Arachis* accessions. Functional diversity indices are recommended to genebank curators to evaluate phenotypic variability to identify accessions with unique traits or identify accessions that represent the greatest phenotypic variation of the species (functional agrobiodiversity collections). The artificial intelligence algorithms are capable of characterizing accessions which reduces costs generated by additional phenotyping. Even though deep analysis of data requires new skills, associating genetic, morphological and ecogeographic diversity is giving us an opportunity to establish unique functional agrobiodiversity collections with new potential traits.

Subjects: Plant Science, Agrobiodiversity, Genetic Resources, Data Mining and Machine Learning.

Keywords: Morphological characterization, Genebank, Image analysis, Digital colorimetric, Geometric morphometric, Random forest, Functional phenomics.

2.1 Introduction

Characterization of genebank accessions contributes to their use in agro-food systems. However, morphological characterization of large collections is expensive and requires substantial manual labor [1]. CIAT's genebank accessions collection, introduction and first duplications were made in the eighties and nineties. Currently during accession regeneration, the genebank systematically characterizes accessions, collecting basic phenological and morphological data [2,3]. The genebank attempt to characterize accessions during regeneration. This process has limited outputs mainly by using lower-priority descriptors and because it is still not recognized as important. This is despite the proclaimed need to collect valuable and reliable phenotypic data which could easily be associated with genomic data [4].

Since the late 1980s, plant genetic resources' morphological characterization has been based on the descriptor guidelines from the International Board for Plant Genetic Resources (IBPGR, now part of the Alliance of Bioversity International and CIAT) [5 – 9]. These classical descriptors have been used to evaluate the phenotypic diversity within the collections and to identify accessions that potentially have characteristics for different scenarios in pre-breeding processes [10]. However, classical descriptors have limitations in germplasm characterization due to the type of data used (qualitative nature which limits genetic association analysis) [11]. High-throughput phenotyping (HTP) methods using proximal sensors and digital images can improve the capture of morphometric and colorimetric descriptors, and monitor physiological traits, and expand the dimensionality of characterization data [12 – 14]. Phenomic descriptors

can complement classical descriptors [15, 16], but the sheer volume, variety and veracity of imaging and remote-sensing data still presents limits in data analysis [17].

Artificial intelligence (AI) promises to deliver real values for humanity including areas such as agriculture and food safety [18]. The high capacity of AI is driven by advances in mechanics, data management, computation, algorithms, and optimization of many mathematics and statistical methods [18, 19]. Machine learning (ML) stands out as a subset of AI as a useful tool for analyzing large datasets generated by non-destructive, field-based, high-throughput phenotyping in plants [17, 20, 21]. However, it has yet to be fully explored to optimize parameters and/or recognize descriptors able to discriminate genebank accessions via HTP analysis [22]. Generating rapid, comprehensive, and precise analysis of large datasets from HTP remains a challenge [17].

Characterization of large genebank collections has been limited due to the high demand of resources and qualified personnel needed even for conventional data capture. Cost-effective phenotyping may involve high initial investments in sensors, vehicles, and method development, but improves quality data capture along with low operational costs [23]. Developing operational, technical, and analytical procedures for genetic resources characterization will contribute to deeper and effective use of germplasm collections. Current characterization procedures do not integrate quantitative metrics for accessions' diversity, distinction, redundancy, and discovery of functional traits. These data are associated with the genome and contribute effectively to managing and documenting germplasm collections. In genebanks, one of the objectives is to conserve a wide range of crop-species diversity based on the components of biological diversity (genetic, taxonomic, ecological, and functional). Unfortunately, diversity-related traits used in genebanks so far do not consider functional diversity evaluated via phenomics as a component to

be investigated. Using functional diversity phenomics traits allows understanding the nature and interactions of traits allowing species to flourish within an ecosystem [24]. Therefore, some genebank accessions present functional traits that were directed by processes of natural selection and/or domestication which allowed them to adapt to diverse ecological conditions. In detail, functional diversity indices determine which species exhibit highly specialized or redundant functional traits and allow increase their resilience when subject to environmental or anthropogenic disturbances [25]. The indices of functional specialization (FSpe), functional identity (FIde) and functional originality (FOri) have been widely used in functional diversity studies in a wide range of ecosystems and organisms [26].

FSpe allows determining the degree to which species exhibit a distinctive set of functional traits compared to species with low FSpe values. These may also contain redundant functional traits. FOri measures the uniqueness of a species' functional traits within a community. A species with high FOri has a unique set of functional traits not found in other species within the same community. FIde measures the functional similarity or dissimilarity among species in a community. The FIde index quantifies to which level a species has similar functional traits with a community and how these traits are distributed among species [25]. These indices have been widely used in studies in diverse ecosystems, however, from the perspective of genebanks they have not been explored.

The genus *Phaseolus*, native to the Americas, is composed of more than eighty species, five of which were domesticated in pre-Columbian times [27]. Beans (*Phaseolus* spp.), and in particular the common bean *P. vulgaris*, represent the most important legume for direct human consumption worldwide [28]. In addition, the forage peanut (*Arachis* spp.) is a native genus of South America, tolerating waterlogging and prolonged drought, and grows well in sun or shade

[29]. Forage peanut is mainly used for livestock feeding due to its high protein content and can also be used to improve soil fertility through symbiotic nitrogen fixation [30]. Bean and forage peanut crop wild relatives (CWR) display agronomically interesting traits [31, 32], making them attractive species worldwide and generating increased interest in information associated with CIAT's germplasm collections.

The integration of highly informative digital descriptors, facilitated by HTP and data analysis using AI tools, represents a transformative approach to characterizing genebank accessions. This innovative set of methods should be used together with conventional techniques and augments the efficiency, precision, and informativeness of the outputs. The proposed integration methodology primary objective lies in complementarity with established methodologies and will enhance and optimize the utilization of plant resources by the stakeholders. The proposed synergy between advanced digital tools with traditional methods represents an accurate and insightful paradigm shift of modern plant genetic resources management.

This work propose a methodology to strengthen characterization in genebanks that allows: (1) improving the characterization processes of bean and peanut accessions in the CIAT genebank via identification of phenomic descriptors comparable with classical ones in their discriminatory power; (2) using artificial intelligence (AI) models to identify descriptors that contribute mostly to higher characterization precision and that will quantify functional diversity among accessions; and (3) to integrate the characterization methodology into workflow of genebanks.

2.2 Materials and Methods

2.2.1 Methodological approach

The methodological approach we used is divided into four main phases (for details on used technique, please see below): (1) Taking pictures of flowers, leaves, and seeds at the Palmira and Tenerife seed regeneration stations in the Valle del Cauca. A scanner was used to take leaf images. (2) Preprocessing of images: images are binarized for morphometric analysis and standardized using a color card before being isolated from the background for further colorimetric analysis. (3) Feature extraction of phenomic descriptors: The extraction of RGB color spaces by clustering and morphospaces from contour analysis is done using the R libraries colordistance and momocx. (4) Classical descriptors: the official list of *Phaseolus* and *Arachis* descriptors was used to determine morphological data, and the RHS color charts were utilized to obtain flower and seed color data. (5) Choosing key descriptors (phenomic, classical, and their combination) for the accessions classification: Random forests were used for the selection process, with the first 10 descriptors being prioritized, starting at the minimal average depth. (6) Functional diversity of accession: The convexhull areas of the 10 functional spaces' accession vertices (functional collection) are identified. (7) The species' functional diversity indicators are shown: The functional entities that join the related accessions will be called species (Figure 2.1).

2.2.2 Plant material and cultivation

Scans of leaf samples and flowers, seed and pod photos were taken from Arachis and Phaseolus accessions (S1Table - *Accession list and passport data*) during the regeneration in the first semester of 2020 at the CIAT Palmira station (03°32'22"N - 76°18'13"W; 965 m above sea level; 900 mm annual cumulative precipitation; 60 % average relative humidity; 23.8 °C average annual temperature) and at the genebank's research station at Tenerife (03°43'34"N - 76°04'26"W; 2,200 m above sea level; 1,800 mm annual cumulative precipitation; 70 % average relative humidity; 17 °C average annual temperature). During regeneration, 16 accessions of Arachis spp. and 14 Phaseolus spp. were selected. Each plot had 120 plants, and 10 healthy and adult plants per plot were chosen to take photos and data. The sample size follows the classical morphological characterization methods [5 - 7] and three important criteria were used to select accessions for this study: (1) The seed regeneration process: In Tenerife and Palmira, a mapping group of accessions were sown to produce seeds. (2) The present species diversity: In order to support the process of distinguishing and classifying species, we considered the diversity of Phaseolus and Arachis species as well as interspecific hybrids that were present during the particular seed regeneration process. (3) The phenotypic variation: Within a species, accessions were chosen based on morphological, colorimetric variation and seed availability. Since phenotypic characterisation of accessions was fully complementary during the routine seed regeneration process, our study could not change or take an experimental design into consideration.

2.2.3 Classical descriptors

The descriptors routinely used by genebank during regeneration for forage peanut and beans are based on the IBPGR guidelines for *Arachis* and *Phaseolus* species. For *Phaseolus coccineus*, *P. vulgaris*, *P. lunatus*, and *Arachis spp.*, we evaluated the following descriptors: days to flowering, days to harvest, weight of 100 seeds, primary seed color, secondary seed color, seed shape, seed color pattern, leaflet shape, standard flower color, flower wing color and pod peak for peanuts [5 - 7]. The classical descriptors were obtained from historical data from the CIAT regeneration program. The flower and seed color was obtained during photographic captures using Royal Horticultural Society Color card (RHS) which is not (but should be) routinely used in the genebank.

2.2.4 Image acquisition and pre-processing

Leaf, seed, pod, and flower images were captured for each bean and forage peanut accession. From ten individual plants the ten healthy leaves per accession were selected and images taken. Leaves were then separated into three (beans) or four (forage peanuts) leaflets and scanned using an EPSON 1000XL scanner to generate JPEG files. Flower images were taken in both experimental stations between 10:00 am and 1:00 pm on a sunny day and calibrated via color card. Images of harvested seed and pods were taken with a Canon SX60HS camera in ORTech Photo-e-Box Bio to standardize the light exposure [33]. To capture the highest image quality, we used RAW and JPEG formats. A 24ColorCard Camera Trax card was used to standardize the images if taken in somewhat different lighting environments

(https://imagejdocu.tudor.lu/plugin/color/chart_white_balance/start). In addition, image backgrounds were extracted for subsequent analyses of color features using the API (Application

Programming Interface) PhotoScissors API (<https://photoscissors.com/upload>). To proceed with the analysis of contours, images were converted into binary format using the ImageJ software with the SIOX plugin [34]. The images were stored and binarized and those used for colorimetric analysis separately.

2.2.5 Extraction of phenomics descriptors from images

After image binarization of pod, leaflet and seed, the outlines obtained were processed using the Momocs R library to perform a Fourier series analysis. From these we extracted the morphospaces values and the values of classical morphometric descriptors such as length, width, area, and perimeter [35]. The ten (10) main morphospace components from each accession were extracted. For the further analysis, binarized images on the white backgrounds were used to obtain each of the principal components or morphospaces. Each morphospace is considered as a phenomic descriptor that reports the phenotypic variation of Phaseolus spp. and Arachis spp. accessions. The colorimetric analysis was performed using colordistance library R that converts image pixels into three-dimensional coordinates into a defined "color space", producing a multidimensional color histogram for each picture [36]. Pairwise distances between histograms were calculated using the earth-movement distance [37], followed by colorimetric K-means clustering. Ten (10) colorimetric groups were determined and the ratio of each of them is calculated using the R colordistance library.

Each of the colorimetric groups is transformed into a colorimetric descriptor for both seeds and flowers. Each of the RGB values per colorimetric group were used as a phenomic descriptor showing the colorimetric variation within and between accessions (to clarify: R_F1=cluster Red number 1 of flowers etc.). The list of all used descriptors can be found in the supplementary

material (*S2 Table - Phenomic and classical descriptors*). Both, the descriptors obtained from the contour analysis and the colorimetric analysis are entered into databases for the selection of descriptors.

2.2.6 Selection of classical, phenomic and combined descriptors

The random forest (RF) ML algorithm was used to determine the weight of individual descriptors so that evaluated accessions can then be optimally grouped and separated [22, 38]. The RF algorithm was executed using the package caret [39] to create folds (data groups) that are a set of (usually consecutive) records of the dataset and fit the training data set. The classification model used 100 trees and the training was performed with 70% of the data and 30% was used for validation. The library random Forest [40] was used to run the RF model, and finally, *randomForestExplainer* [41] was used to visualize the RF models.

The relationship between the number of trees and the minimum depth of distribution will determine the most important descriptors able to discriminate accessions comparing the classical, phenomic (digital) and combined descriptors. Out-of-bag (OOB) data accuracy is determined for each descriptor class and its combination presenting values below 25%. This number indicates models of good to high precision in accession discrimination [42, 43]. The highest values of the minimum depth distribution and the lowest number of trees determined fifteen (15) the most important descriptors.

The confusion matrix is made by determining the error proportions of the predicted values for each of the models using classical, phenomic and combined descriptors. The confusion table was analyzed from two perspectives: (1) its high ability to discriminate between non-related

accessions/species, and (2) its low sensitivity to confusions between accessions of the same species.

2.2.7 Functional diversity of phenomic, classical and their combination

In addition to expensive genotype (DNA)-based diversity investigations, where the primary goal is to detect genetic duplicity or redundancy, genebank should also use cheaper phenotypic characterization techniques to confirm if accessions redundant genetically are also phenotypically redundant or not. For this purpose, a functional diversity analysis was carried out aiming to apply the theoretical foundation in the genebank contexts. Initially, each of the most important descriptors was clustered with the related species for both Phaseolus spp. and Arachis spp. and a first matrix "1. *Species - accessions*" was made. Here the presence or absence of each of the accessions grouped in each of the species was considered. Subsequently, a second matrix "2. *Type of descriptor*" was made where the descriptor and its class (number or characters) were associated. Finally, the descriptors selected in the phenomic and classical descriptor types and their combination constituted the third final matrix "3. *Functional descriptors*". Species were treated as functional entities (FE) and accessions as abundances per species. These matrices are required to determine the functional diversity indices. The area of the functional space was calculated for each forage peanut and bean accession descriptor types. Functional diversity indices were determined as: (1) the functional identity (FId), which has the centroid coordinates of the principal coordinate analysis performed in the functional diversity analysis; (2) the functional specialization (FSpe), which is the average distance from the accessions to the centroid of all species in the principal coordinates analysis; and (3) the functional originality

(FOri), which is the average distance between accessions of the same species [44, 45]. These metrics and visualizations were performed using the R mFD library [25]. The metrics were grouped by types of descriptors and functional indices.

Accessions of *P. coccineus* species to *P. lunatus* and *P. dumosus* species, as well as accessions of *A. pintoi* and *A. paraguariensis* species, can be used to visualize the importance of functional indices. We are able to observe the discrimination of every accession for every species for every assessed index thanks to this depiction. Figure 2.1 shows the methodological process in graphic form.

2.3 Results

2.3.1 Selection of phenomic, classical, and combined descriptors and accession classification

From image segmentation, processing and analyses, quantitative colorimetric and morphometric values were obtained for bean and forage peanut flowers, leaflets, seeds, and fruits (Figure 2.2 and 2.3). Morphometric and colorimetric quantification highlighted wide variation among the accessions within both bean and peanut. Based on the random forest algorithm and the minimum depth distribution (MDM) outputs, we selected the most discriminatory descriptors (within the classical descriptors, the phenomic descriptors and their combination) considering the lower the average minimum depth values the descriptor presents, the greater is its interaction with others, and shows greater capacity for accession discrimination. Classification of accessions using the confusion table for each of the descriptors and their combination provides clarity on the

descriptor sensitivity in discriminating between individual accessions and their grouping for both bean and forage peanut accessions (Figure 2.4 and 2.5).

Variation in pod and seed shape was observed in forage peanuts, determining differences especially in seed and pod length and width and in pod beak shape. In the colorimetry of the peanut flower, there was variation in yellow. In the case of seed we found differences in brown shades (Figure 2.2). In *Arachis*, the phenomic descriptors as leaf (PC1LM) and seed (PC1S) and green color of flower (G_8F) are the ones with the lowest MDM values and thus with high variability between accessions. Generally, the descriptors covering the first morphospaces of seed (PC1S and PC5S), leaf (PC1LM and PC3LM) and pod (PC2P) are the most important descriptors in the classification of forage peanut accessions, together with descriptors for seed color (R_10S, G_4S, G_1S, G_3S, B_1S, R_2S, B_6S), pod (R_10P) and flower (G_8F, B_5F) (Figure 2.4A). In classical forage peanut descriptors, the 100-seed weight (P100_weight), days to flowering (DTF), growth habit and days to harvest (DTH) presented the lowest MDM values (Fig 2.4C). Other important group of descriptors were those related to the primary and secondary seed color (RHS_sseedcolor; RHS_seedcolor). Similarly, the leaflet shape (Leaflet_shape) descriptor contributes to the discrimination of peanut accessions. The leaflet shape belonging to the classical descriptor showed the same values when taken as the phenomic descriptor (Figure 2.4C). Importantly, we observed that the *Arachis* spp. classical descriptors presented a better discriminative power than phenomic ones. For a descriptor combination (classical + phenomic) of *Arachis* accessions (Figure 2.4E), the classical descriptors of secondary seed color (RHS_sseedcolor), days to flowering (DTF), 100-seed weight, growth habit and days to harvest (DTH) showed the lowest MDM values and thus highest variability. The descriptors of the first morphospace in pod and leaf (PC1P, PC1L) are good discriminants of forage peanut accessions.

Also the RGB color space of seed and flower (G_3S) is very good for peanuts. Generally, compared to the classical descriptors, the phenomic descriptors showed a less number of descriptors suitable for discriminating among forage peanut accessions. In the confusion matrix of phenomic descriptors, *A. pintoi*'s wild accessions (18745, 18746, 1748, 22151, 22159, and 22176) had the highest level of similarity. This indicates that the phenomic descriptors accurately capture the phenotypic redundancy of those accessions where classical descriptors do not show differences. For example, the *A. paraguariensis* (22636) presents similarity with the *A. archeri* (22226) and *A. repens* (22162) indicating phenotypic closeness to these accessions(Figure 2.4B).

In the case of beans, leaf shape showed high variation among the accessions, allowing differentiation of leaf shape from lanceolate to globular, as well as discriminating seed shape from circular to kidney-shaped (Figure 2.3). The most important phenomic descriptors in Phaseolus spp. accessions are the first seed morphospace (PC1S) and leaf (PC1L). These two descriptors had the lowest values of MDM, thus high power to discriminate accessions. Seed color descriptors in RGB color spaces (B_8S, R_9S, R_8S, G_6S, B_6S, R_4S, B_10S, R_6S, G_8S and R_3S) in addition to the red color space (R) represent highly useful descriptors. This indicates that variability in seed color is the most important trait in discriminating the evaluated Phaseolus accessions (Figure 2.5A). The days to flowering (DTF), primary seed color (Colorps) and 100-seed weight (Weight100) were the most important classical descriptors in Phaseolus accession discrimination (Figure 2.5C). Similarly, descriptors associated with seed shape (Formasem), seed brightness and flower color (RHSwing and RHS_standard), leaf length (leaf_length), leaf width (leaf_width) and the ratio between them (ratio_leaf) were highlighted. For descriptor combinations (classical + phenomic), the 100-seed weight, seed shape, primary seed color and days to flowering presented lower MDM values (Figure 2.5E). Also first seed and

leaf morphospace (PC1S, PC1L) and the RGB seed color spaces (B_8S, B_6S, R_4S, R_8S and R_9S) showed better Phaseolus discrimination when combined descriptors were performed (Figure 2.5E). In the classification of Phaseolus accessions, there is a confusion between accessions of *P. lunatus* (G26467 and G26480), *P. dumosus* and *P. coccineus* (G35580, G35586, G35754A, G36211) (Figure 2.5A). Interestingly, the accessions that are of interspecific hybridization origin (G24764B and G40503) showed confusion with the species *P. tuerckheimii*, despite being phylogenetically distant (Figure 2.5B).

For the phenomic descriptors, it is evident that there is a confusion (uncertainty) in both Phaseolus and Arachis accessions. This demonstrates a greater accuracy/precision of phenomic descriptors in detecting redundancy among accessions. Importantly, ML selection of both species (Phaseolus spp. and Arachis spp.) included descriptors of different organs. This indicates a non-biased selection since it counterbalances classical traits used by taxonomists in the classification of plant species with the digital ones. The discrimination of Arachis spp. and Phaseolus spp. accessions using classical and combined descriptors showed no confusion between accessions (Figure 2.4A and F, and Figure 2.5D and F).

2.3.2 Functional diversity indices are important for beans and forage peanut accessions' characterization

We only considered descriptors prioritized by minimum mean depth (MDM) for each descriptor type. The functional diversity indices are determined to evaluate in quantitative and observable terms the capacity of each descriptor to show phenotypic diversity and redundancy among accessions of the same species. The functional spaces for each of the descriptor types provide evidence of the descriptors' ability to show phenotypic variation among accessions, as well as to

determine the descriptors' ability to unravel species. When the functional spaces present a greater area, it is related to a greater phenotypic variation among the accessions.

The diversity indices FSpe, FOri and FIde show the variation between descriptor types and species, allowing grouping accessions that are specialized, or redundant, or distinct (Table 1).

The phenomic descriptors in the bean accessions have a functional space area of 0.144 and 12 accession vertices. The vertices accessions of the first (PC1) and second (PC2) components are accessions G35271, G35998 and G35754A of the species *P. coccineus* and accessions G40503, which is an artificial hybrid between *P. lunatus* and *P. polystachyus* and accession G26467 *P. lunatus* (Figure 2.6A). In spite of the fact that the classical descriptors showed greater area than phenomic descriptors, they have the same number of vertices accessions (Figure 2.6C). The classical descriptors presented the vertices accessions G40503 (*P. lunatus* x *P. polystachyus*), G35998 (*P. coccineus*), G26736 and G24480 (*P. lunatus*) and the natural hybrid G24764B (*P. dumosus* x *P. vulgaris*) (Figure 2.6C).

When combined, the phenomic and classical descriptors revealed a larger area in the convexhull than when applied separately. This suggests that combined analysis distinguishes and captures the phenotypic variations of every accession assessed separately. Furthermore, the 14 accessions of Phaseolus (Figure 2.6E), are displayed as vertices, demonstrating the phenotypic identity of each accession. The functional specialization analysis (FSpe), with the phenomic descriptors, showed a clear discrimination of the species *P. dumosus* with *P. coccineus* (Figure 2.7A). Calculating FSpe, the classical descriptors do not clearly differentiate between *P. dumosus* and *P. coccineus* (Figure 2.7G), while when combined the accessions are more separated than using classical descriptors only (Figure 2.7M). The functional identity analysis (FIde) shows that the phenomic descriptors clearly separate *P. dumosus* and *P. coccineus* (Figure 2.7C). Importantly,

when using only classical descriptors, no clear difference between the two species in PC1 were found (Figure 2.7I). Finally, the combined descriptors allow a clear distinction between both accessions especially in PC2 (Figure 2.7O).

When comparing the *P. coccineus* with *P. lunatus*, a clear distinction between them is observed in all three types of descriptors (phenomic, classical and their combination) (Figure 2.7E, K, Q). In the combination of descriptors (Figure 2.7Q), the *P. lunatus* accessions are clearly separated. The separation of *P. lunatus* accessions in the combined descriptors clearly influences the discrimination of accessions. The combination of descriptors in FOrí shows a greater separation of *P. lunatus* accessions (Figure 2.7R) compared to the classical (Figure 2.7L) and phenomic (Figure 2.7F) descriptors.

FSp helps determine that the *P. dumosus* presents its lowest values for phenomic descriptors (0.32) and in descriptor combination (0.47). Meanwhile *P. coccineus* presents the lowest value of 0.45 in the classical descriptors. The *P. lunatus* accessions presented the highest FSp values in the phenomic descriptors (0.67) and the combination (0.84). The artificial hybridization *P. lunatus* x *P. polystachyus* showed the lowest value in the phenomic descriptors (0.32) but the highest value using the classical descriptors (1.00). The FIde index shows that the combination of descriptors generates intermediate values or discriminates to a greater extent (for example *P. lunatus* increasing the distance in the PC1 coordinate; -0.26). The descriptor combination determined that the accession of *P. tuerckheimii* increased discrimination (-0.21), with respect to the phenomic (-0.03) and classical (0.05) descriptors, determining that the accession is a vertices in the convexhull.

In forage peanuts, in two groups of accessions of *A. pintoi* and *A. paraguariensis* were observed the effect of different types of descriptors on functional diversity indices (Figure 2.8). The FSp

analysis showed the phenomic (Figure 2.8A) and classical descriptors (Figure 2.8D) discriminated *A. pintoi* against *A. paraguariensis*. Many *A. pintoi* accessions are clustered using the classical and phenomic descriptors, yet to a greater degree in the combination (Figure 2.8G). For FId, the three types of descriptors clearly and similarly separate accessions of *A. pintoi* and *A. paraguariensis* species (Figure 2.8B, E and H). In the FOri analysis, the phenomic descriptors allow determining a group of closely related *A. pintoi* accessions (Figure 2.8C). The accessions of *A. pintoi* are grouped on a cluster of closely related accessions when using the classical descriptors since FOri determines the functional similarity between the accessions (Figure 2.8F). Importantly, this is the confirmation that the descriptor combination method is more sensitive to determine similar accessions (Figure 2.8I). The FSpe for Arachis determines that the accessions *A. archeri*, *A. paraguariensis* and *A. repens* present greater distance with respect to the centroid. The highest value was observed for *A. paraguariensis* in comparison with the lowest values of accessions of *A. pintoi*. This again demonstrates the high functional redundancy that exists between these accessions.

2.4 Discussion

The results showed that the combination of classical and phenomic descriptors can discriminate between forage peanuts and between bean accessions more accurately than using them independently. For both bean and peanut forage accessions the first morphospaces of seed and leaf (Figure 2.3A) are important to discriminate accessions. In addition, the variation of flower and seed color spaces were identified as important. This means that the digital quantification of these common classical descriptors as leaf shape and seed color determine the existing morphological variation between the accessions and can speed up the process.

Historically, the variability in morphological and colorimetric traits in leaves, flowers, and seeds was a key factor in determining the genetic representativeness of the collected accessions during the germplasm explorations. List of standardized descriptors relating morphological and agronomic parameters served as a starting point for germplasm introductions into different agro-environments [46, 47].

The classical descriptors as days to flowering, days to harvest, 100-seed weight, and flower and seed colors (Royal Horticulture Society; RHS) allow discriminating both bean and forage peanut accessions (Figure 2.4C). For example, 100-seed weight is closely related to seed size. That emerged as a result of the domestication syndrome differentiating clearly between wild and domesticated genotypes in modern days. Day to flowering differentiate between different bean species accessions and it is influenced to a greater extent by ecogeographic regions of origin, as was published for both bean and peanut species [48, 49, 27].

Our study - not surprisingly - shows that digital phenomic descriptors have high concordance with classical descriptors. This means that the quantification of leaf shape from geometric morphometry and seed and flower color using both descriptor types will characterize genetic resources with the similar precision [50 - 53]. The phenomic descriptors such as PC1, PC1L and RGB color spaces for both seed and flower traits thus can be used as descriptors clearly differentiating the variation among accessions and via digital processing save time during and after evaluation.

When we combined both descriptor types, the machine learning analysis selected these descriptors as the most important: PC1L, PC1S, PC1LM, PC1P, days to flowering, PC1P, days to flowering, flower color (RHSwing and RHSstandard), 100 seed weight and classical seed color (RHS_seedcolor) and phenomic (B_6S, R_4S, R_8S, R_9S, G_5S and G_3S) for both bean and

peanut accessions evaluated. These descriptors are quantitative, except for the seed and flower color. The above mentioned will allow a better understanding of the accession phenotypic architecture and identify redundant accessions, as demonstrated by the confusion matrices (Figure 2.5) and the functional diversity analysis (Figure 2.6).

The phenomic descriptors logically reveal some uncertainty in accessions determination particularly in closely related accessions of forage peanuts and beans. This is not a surprise and will likely change with the newly generated knowledge. However, since all accessions were clearly differentiated, neither the classical descriptors nor the descriptor combination displayed any uncertainty at all. The understanding of the nature of the variables and its overall composition to a species characterization is a key to guarantee the quality of the discrimination, since classical descriptors such as seed brightness (Bllosem) and vigor (performance) are not technically clearly characterized, due to their "*ordinal*" nature. This makes the interpretation difficult in routine operations during seed regeneration processes in genebanks.

Although the classical descriptors presented a better classification than the digital phenomic descriptors, the classical descriptors, being more qualitative, limit its use for genetic studies [54, 55]. Phenomic descriptors aim to digitalize traits and thus quantify them into understandable, determinable, and measurable attributes convergent to genomics. Quantifying and dissecting the responses (physiological, morphological, and other traits) will help pre-breeders, curators, and crop physiologists to understand the complex spatio-temporal dynamics of individual traits orchestrating in different species and then use other techniques (QTL, GWAS, genomic selection etc.) to verify their heritability in different environments and select for them much effectively. This itself should be a valuable justification to add digital phenomics descriptors into any routine characterization of accessions by genebank curators worldwide.

In our study we demonstrate for the first time the use and importance of functional diversity indices in characterizing the genetic resources even possible during the routine regeneration. These indices can become metrics to assess the phenotypic diversity of accessions and more importantly identify redundancy within collections. One of our key results is the integration of phenomic and classical descriptors.

The descriptors combination allows us to determine a minimum number of accessions, but still represents the highest phenotypic variation in bean and forage peanut accessions (Figure 2.6) by identifying vertices accessions with differential and/or redundant phenotypic characteristics. Furthermore, the functional space (Figure 2.6) analysis will allow selection of key accessions in germplasm collections within species.

The convex hull can be evaluated in two ways: (I) Vertices accessions could be considered a core or functional collections; (II) Determining redundant accessions allows genomic evaluations to be performed on a smaller number of accessions and verify their similarity. For example, functional space associates *A. archeri* and *A. paraguariensis* that are closely related to section Erectoides. Also the accessions of *A. pintoi* and *A. repens* under the section Caulorhizae are well associated [56].

In the case of beans, there is a confirmation of the similarity (Figure 2.6) in functional space of the species *P. dumosus* and *P. coccineus*, as these species share phenotypic and genetic characteristics due to their co-evolution [57]. Importantly, *P. lunatus* is represented by a single group in a functional space showing the discrimination of the combined descriptors between these three species is in correspondence to the recognized gene pools [58]. Functional diversity indices such as functional specialization (FSp), functional originality (FOri), and functional identity (FIde) [59] have not been implemented in the genebanks so far despite their value. These

functional indices have been used to study species in diverse ecosystems and quantifying their level of adaptation to different anthropogenic or natural disturbances [60 - 62].

One of the main objectives of genebanks is to gather diversity in its different components (taxonomic, ecological, genetic, and functional) through the conservation of accessions that represent the variation in each of the components at both the accession and species levels. This obviously is a very important task of all genebanks, however not easily doable e.g. regarding the missing collections evaluations. There are currently no studies that integrate the characterization of germplasms with metrics associating functional diversity indices.

To understand the importance of our approach by evaluating phenotypic functionality in genebanks, we proposed the following points: (1) Accessions should come from diverse ecogeographic areas (passport data) with determined genetic and phenotypic variation [63], (2) The ability of germplasm accessions to adjust to particular environment is associated with their geographical spread, which is defined by the ecological habitat of origin [64]. (3) The variation of functional phenotypic traits (descriptors) could (in some cases) determine the level of resilience to different biotic and abiotic stress scenarios in the origin [65] but also in the new target population of environments too.

Considering critically the above-mentioned points, FSpe can determine the accessions of each species that represent extreme functional descriptors, as in the case of *P. lunatus* hybrid (*P. lunatus* x *P. polystachyus*; G40503) [66] (Figure 2.7) and *A. paraguariensis* (Figure 2.8). Furthermore, FOri can determine the phenotypic redundancy between accessions per species, as in the case of *A. pintoi*. Species which show the lowest FOri values among the other accessions allow generation of metrics with phenotypic redundancy and thus contribute significantly to the management of germplasm collections.

2.4.1 Proposed methodology for genebank curators

First, we recommend using the high-throughput phenotyping platforms (HTTP) as an alternative to capture classical descriptor values associated with color, shape, and size of the main plant organs. These can be likely/preferably captured on the plants during the very seed regeneration descriptors into processes in the genebank [50, 52, 53] The workflow of digital image captures must be based on a verified and standardized system by individual genebanks. This guarantees full traceability with the processes of seed regeneration and conservation, as well as required standard operating procedures (SOP) and training [67]. Some of the operational activities from our characterization methodology are depicted in Figure 2.9. These procedures could be incorporated into the CIAT genebank workflow primarily to facilitate population structure analysis, taxonomic curation, and purity verification. This means that HTTP data needs to be not only successfully captured in the right moment but also transferred in the recognized curator pathways to be added to the database without any confusion. It is not practical to create simultaneous datasets on the same accessions, but mainly to create a very new type of database able to process the classical and phenomic descriptors into meaningful documents. New inventions in HTTP have been giving plant breeders accessible and powerful tools to characterize many genotypes for important agronomic and morphological traits [68] and quantify them. Regarding the traits' names, we strongly recommend that the curated vocabulary is used to describe new digital phenotypic traits and possibly crop trait ontology needs to be applied accordingly.

Second, the image processing performed in different image software alters the results of the tests performed by the artificial intelligence affecting the efficiency of the clustering and discrimination processes [69]. Nowadays with convolutional networks and different trained neural networks it is possible to improve the processes of color correction, image segmentation and feature extraction in an automated way [70]. Descriptors identified by using digital images will allow high efficiency and precision characterization, as demonstrated by our results. The challenge in the image standardization process is to optimize operational processes and follow the technical and infrastructure requirements that could be limited to maintain stable controlled light conditions (quality, quantity, angles) and to have trichromatic sensors of more than 20 megapixels or multispectral sensors [71]. Genebanks also should strategically integrate the use of other genomics and phenomics techniques through the acquisition of high-performance technologies and computational infrastructures (sensors, algorithms, and processing/storage). That allows efficient and effective accession characterization even during routine genebank operational processes.

Third, analyses of phenotypic data in the genebanks should integrate various tools into one more universal one to reduce data dimensionality, facilitate selection of trait variables and optimize data outputs via machine learning results. The model used for data management will then depend on the objective of study. For example, it will differ for detailed classification of accessions required by curators, in comparison to more general data important for breeders or even deep understanding of accessions adaptability needed by crop physiologists. Our research demonstrates that digital descriptors extracted from organ images like flower and seed, as well as classical descriptors like days to flowering and days to harvest, can be selected to classify bean and peanut accessions. This also means that good curators should be brave and patient but

persistent in the process of transformation. Work carried out by others [72, 73], demonstrate the ability to classify species and genotypes of soybean, using machine learning algorithms such as super vector machine (SVM), random forest (RF), neural networks (NN), K-Nearest Neighbors (KNN) and linear discriminant analysis (LDA).

The use of machine learning algorithms is recommended to screen highly characterized training populations. This will contribute to the development of effective automated workflows that could reduce the costs and time currently required for germplasm evaluation and increase the precision of characterization. In addition, using higher spectral resolution proximal and/or remote sensors will certainly contribute to building species-specific data libraries of germplasms (species spectral signatures) to support future digital genebanks.

Fourth, when data capture and selection of functional descriptors is finished, functional diversity analyses can be performed, identifying, and selecting redundant accessions. The evaluation of accessions of interest including functional diversity analyses could be performed in two ways: 1) Evaluation of vertices accessions in pre-breeding processes, integrating selected morphological and physiological traits under stress conditions, especially in crop wild relatives, and 2) Verification of accessions with low FOri values by integrating phenomic and genomic data, identifying probably duplicated accessions in the collection etc. The results of functional diversity analyses will contribute significantly to crop breeding programs, improving and strengthening the parental lines selection and development of new lines that include genes from crop wild relatives that have not been explored for traits associated with tolerance to biotic and abiotic stresses and boost seed nutritional profile (biofortification process). In our previous study we showed that a similar approach can be used to even verify the successful transfer of traits between parents and its progeny [43].

To introduce a new material into a genebank, seed increase is necessary as well as a first characterization of the materials based on classical descriptors. Genetic (genomic) identification can be a significant player in the correct decision-making process too. The regeneration is carried out when the seed viability or number of conserved seeds has been reduced or for sanitary reasons to eliminate quarantine-type diseases, to conserve the materials in the long term etc. We propose that during these two procedures, trait capturing can be easily performed in the flowering phase (BBCH 65) and then in fruit formation stages (BBCH 75 and on) by capturing photos of leaves, flowers, and pods. If spectral signatures (spectroradiometer) will be taken they need to be stored as digital libraries for later use when calibration curves allow to add additional characteristics. Data needs to be further processed (pre-processing algorithms) and analyzed (machine learning and functional analysis); however, hopefully these processes will be fully automated in the future. In addition, during the seed quality verification (we strongly recommend this step as part of the routine protocol), a parallel path can be established where the capture of seed quality data are associated with seed HQ pictures using various types of sensors (RGB cameras, multispectral, hyper-radiometer) where phenomics descriptors of interest can be extracted and used by stakeholders.

2.4.2 Genebanks, characterization, phenomics and conceptualization

Genebank collections were collected with the objective of conserving the highest genetic representativeness of wild and cultivated relatives. Both wild and cultivated accessions were in an evolutionary process influenced by environmental conditions and selection pressure during domestication. From this postulate, it can be inferred that the conserved accessions present phenotypic characteristics driven by space (environment), time (evolution) and man (selection)

[74]. Based on the above, we could begin to delve deeper into the conceptual aspects that integrate the phenotypic characterization of plant genetic resources and their relationship with phenomics. Germplasm characterization is based on phenotypic traits that have high heritability and can be expressed in all environments [5 - 7], so that intra-accessional variation is low. Considering the above-mentioned, phenomics is the biological discipline that focuses on the study of phenomes. If the phenome is a set of phenotypes that originate from the relationship between genotype (G), time (t) and environment (E) [75], in the context of plant genetic resources, the accessions collected and conserved today present the phenotypic traits that allowed adaptation to ecological conditions determined by time and environment; therefore, these phenotypic traits are constitutive. The study of the phenotypic variation of germplasm collections (characterization) would thus allow us to evaluate the phenome expression of accessions that are part of the same species and to understand these constitutive functional traits that relate diverse adaptation strategies in the environments in which they evolved. Our work proposes a methodological route that allows the integration of high-throughput phenotyping, artificial intelligence and functional diversity indices that allow exploring the dimensionality of the phenome in genebanks. The study of the phenome in germplasm collections can contribute to the selection of accessions via functional traits with potential for crop improvement and to understand the evolutionary aspects that conditioned the phenotypic diversity of plant genetic resources.

2.5 Conclusions

In our study we demonstrate that via the integration of phenomic and classical descriptors it is possible to identify specific and/or redundant accessions and to characterize each accession (within one species) by using functional diversity indices and AI. Phenotyping via digital images is suitable for germplasm characterization. Image capture can be integrated into various operational processes required during the routine SOPs as seed regeneration, seed quality protocols and seed conservation. The identification of redundant accessions is one of the goals of genebank curators and currently requires the development of new procedures and metrics in phenotyping characterization. Our proposed methodology facilitates the identification of possible redundant accession groups more easily than the classical approach which is very laborious or almost impossible. Furthermore, we showed that functional diversity analyses using functional diversity indices could be a key in understanding the adaptive capacity of different species. Even though we know the deep analysis of data based on functional diversity indices is not easy and requires new skills, associating genetic, morphological and ecogeographic diversity will establish unique core collections with new potential that have not been explored in genebanks yet.

2.6 Acknowledgements

We express thanks to Maria Isabel Gomez for her support during image capture in the CIAT entomology laboratory. We would also like to thank the genetic resources program (especially Dr. Marcela Santaella) for the support in methodological development and an overall positive attitude. MOU is thankful to the Gesellschaft für Internationale Zusammenarbeit, Germany

(GIZ) for their systematic support. The authors would also like to acknowledge Vincent Johnson of the Bioversity-CIAT Alliance Science Writing Service for his editorial support.

2.7 References

1. FAO. The Second Report on the State of the World's Plant Genetic Resources for Food and Agriculture; Food and Agriculture Organization of the United Nations: Rome, Italy, (2010).
2. CIAT Centro Internacional de Agricultura Tropical, Genetic Resources Unit. 1982. Annual report. Cali. Colombia. p50. (1983)
3. CIAT Centro Internacional de Agricultura Tropical, Tropical Pastures germplasm. In Annual report - Genetic Resources Unit. Cali. Colombia. p 43-54. (1989).
4. Sackville Hamilton, R. "Scope and roles of the CGIAR genebanks: 2030 vision. Background paper No. 4 prepared for the 2020 System Level Review of Genebank Costs and Operations (GCO). CGIAR and the Global Crop Diversity Trust. Bonn: Global Crop Diversity Trust." (2020).
5. IBPGR Descriptors for Lima Bean (*Phaseolus lunatus*). International Plant Genetic Resources Institute, Rome, Italy, 1982.
6. IBPGR. Descriptors for groundnut. Rome, Italy and Patancheru, AP, India: Int. Board of Plant Genetic Resources and Int. Crops Res. Inst. For the Semi-Arid Tropics. 1992.
7. IPGRI. Descriptores para *Phaseolus vulgaris*. International Plant Genetic Resources Institute. Rome. 2001.
8. Kumar, Praveen, and Prashant Kaushik. "Evaluation of genetic diversity in cultivated and exotic germplasm sources of Faba Bean using important morphological traits." BioRxiv (2020). DOI:<https://doi.org/10.1101/2020.01.24.918284>
9. Rezaei, Mehdi, Ali Sarkhosh, and Ahmad Balandari. "Characterization of Valuable Indigenous Barberry (*Berberis* sp.) Germplasm by Using Multivariate Analysis." International Journal of Fruit Science 20, no. 1 (2020): 1-19. DOI: <http://dx.doi.org/10.1080/15538362.2018.1555508>
10. Nguyen, Giao N., and Sally L. Norton. "Genebank phenomics: A strategic approach to enhance value and utilization of crop germplasm." Plants 9, no. 7 (2020): 817. DOI: <https://doi.org/10.3390/plants9070817>
11. Ferreira, ME. "Molecular analysis of gene banks for sustainable conservation and increased use of crop genetic resources." *The role of biotechnology in exploring and protecting agricultural genetic resources* (2006): 121-127. ISBN: 9251054800.

12. Araus, José Luis, and Jill E. Cairns. "Field high-throughput phenotyping: the new crop breeding frontier." *Trends in plant science* 19, no. 1 (2014): 52-61. DOI: <https://doi.org/10.1016/j.tplants.2013.09.008>.
13. Van Belleghem, Steven M., Riccardo Papa, Humberto Ortiz-Zuazaga, Frederik Hendrickx, Chris D. Jiggins, W. Owen McMillan, and Brian A. Counterman. "patternize: An R package for quantifying colour pattern variation." *Methods in Ecology and Evolution* 9, no. 2 (2018): 390-398. DOI: <https://doi.org/10.1111/2041-210X.12853>
14. Weller, Hannah I., and Mark W. Westneat. "Quantitative color profiling of digital images with earth mover's distance using the R package colordistance." *PeerJ* 7 (2019): e6398. DOI 10.7717/peerj.6398
15. Nankar, Amol N., Ivanka Tringovska, Stanislava Grozeva, Daniela Ganeva, and Dimitrina Kostova. "Tomato phenotypic diversity determined by combined approaches of conventional and high-throughput tomato analyzer phenotyping." *Plants* 9, no. 2 (2020): 197. DOI: <https://doi.org/10.3390%2Fplants9020197>
16. Pereira-Dias, Leandro, Ana Fita, Santiago Vilanova, Elena Sánchez-López, and Adrián Rodríguez-Burrueto. "Phenomics of elite heirlooms of peppers (*Capsicum annuum* L.) from the Spanish centre of diversity: Conventional and high-throughput digital tools towards varietal typification." *Scientia Horticulturae* 265 (2020): 109245. DOI: <https://doi.org/10.1016/j.scienta.2020.109245>
17. Singh, Arti, Baskar Ganapathysubramanian, Asheesh Kumar Singh, and Soumik Sarkar. "Machine learning for high-throughput stress phenotyping in plants." *Trends in plant science* 21, no. 2 (2016): 110-124. DOI: <http://dx.doi.org/10.1016/j.tplants.2015.10.015>
18. Smith, Matthew J. "Getting value from artificial intelligence in agriculture." *Animal Production Science* 60, no. 1 (2018): 46-54. DOI: <https://doi.org/10.1071/AN18522>
19. Jha, Kirtan, Aalap Doshi, Poojan Patel, and Manan Shah. "A comprehensive review on automation in agriculture using artificial intelligence." *Artificial Intelligence in Agriculture* 2 (2019): 1-12. DOI: <https://doi.org/10.1016/j.aiia.2019.05.004>
20. Lee, Unseok, Sungyul Chang, Gian Anantrio Putra, Hyoungseok Kim, and Dong Hwan Kim. "An automated, high-throughput plant phenotyping system using machine learning-based plant segmentation and image analysis." *PloS one* 13, no. 4 (2018): e0196615. DOI: <https://doi.org/10.1371/journal.pone.0196615>
21. Koh, Joshua CO, German Spangenberg, and Surya Kant. "Automated machine learning for high-throughput image-based plant phenotyping." *Remote Sensing* 13, no. 5 (2021): 858. DOI: <https://doi.org/10.3390/rs13050858>
22. Henao-Rojas, Juan Camilo, María Gladis Rosero-Alpala, Carolina Ortiz-Muñoz, Carlos Enrique Velásquez-Arroyo, William Alfonso Leon-Rueda, and Joaquín Guillermo Ramírez-Gil. "Machine Learning Applications and Optimization of Clustering Methods Improve the Selection of Descriptors in Blackberry Germplasm Banks." *Plants* 10, no. 2 (2021): 247. DOI: [10.3390/plants10020247](https://doi.org/10.3390/plants10020247).
23. Reynolds, Daniel, Frederic Baret, Claude Welcker, Aaron Bostrom, Joshua Ball, Francesco Cellini, Argelia Lorence et al. "What is cost-efficient phenotyping? Optimizing

- costs for different scenarios." *Plant Science* 282 (2019): 14-22. DOI: <https://dx.doi.org/10.1016/j.plantsci.2018.06.015>
24. Petchey, O. L., & Gaston, K. J. (2006). Functional diversity: back to basics and looking forward. *Ecology letters*, 9(6), 741-758. DOI: <https://doi.org/10.1111/j.1461-0248.2006.00924.x>
 25. Magneville, Camille, Nicolas Loiseau, Camille Albouy, Nicolas Casajus, Thomas Claverie, Arthur Escalas, Fabien Leprieur, Eva Maire, David Mouillot, and Sébastien Villéger. "mFD: an R package to compute and illustrate the multiple facets of functional diversity." *Ecography* 2022, no. 1 (2022). DOI: <https://doi.org/10.1111/ecog.05904>
 26. Pimiento, C., F. Leprieur, D. Silvestro, J. S. Lefcheck, Camille Albouy, D. B. Rasher, M. Davis, J-C. Svenning, and J. N. Griffin. "Functional diversity of marine megafauna in the Anthropocene." *Science Advances* 6, no. 16 (2020): eaay7650. DOI: <https://doi.org/10.1126/sciadv.aay7650>
 27. Chacon-Sanchez, Maria Isabel, Jaime Martínez-Castillo, Jorge Duitama, and Daniel G. Debouck. "Gene flow in Phaseolus beans and its role as a plausible driver of ecological fitness and expansion of cultigens." *Frontiers in Ecology and Evolution* 9 (2021): 618709. DOI: <https://doi.org/10.3389/fevo.2021.618709>
 28. Bitocchi, Elena, Domenico Rau, Elisa Bellucci, Monica Rodriguez, Maria L. Murgia, Tania Gioia, Debora Santo, Laura Nanni, Giovanna Attene, and Roberto Papa. "Beans (*Phaseolus* spp.) as a model for understanding crop evolution." *Frontiers in plant science* (2017): 722. DOI: <https://doi.org/10.3389/fpls.2017.00722>.
 29. Kerridge PC, Hardy B, editors. *Biology and agronomy of forage Arachis*. International Center for Tropical Agriculture; 1994. ISBN 958-9183-96-4
 30. Cook, B. G., B. C. Pengelly, R. Schultze-Kraft, Matt Taylor, S. Burkart, J. A. Cardoso Arango, J. J. González Guzmán, K. Cox, C. Jones, and M. Peters. "Tropical Forages: An interactive selection tool. 2nd and Revised Edn." International Center for Tropical Agriculture (CIAT), Cali, Colombia and International Livestock Research Institute (ILRI), Nairobi, Kenya. www. tropicalforages. info (2020).
 31. Burridge, James D., Harini Rangarajan, and Jonathan P. Lynch. "Comparative phenomics of annual grain legume root architecture." *Crop science* 60, no. 5 (2020): 2574-2593.
 32. Carvalho, Paula Andreá Sampaio de Vasconcelos, Márcio de Carvalho Moretzsohn, Ana Cristina Miranda Brasileiro, Patrícia Messenberg Guimarães, Tânia da Silveira Agostini-Costa, Joseane Padilha da Silva, and Marcos A. Gimenes. "Presence of resveratrol in wild *Arachis* species adds new value to this overlooked genetic resource." *Scientific Reports* 10, no. 1 (2020): 1-9.
 33. Bohórquez-Chaux, Adriana, María Isabel Gómez-Jiménez, Luisa Fernanda Leiva-Sandoval, and Luis Augusto Becerra Lopez-Lavalle. "NYMPHSTAR: an accurate high-throughput quantitative method for whitefly (*Aleurotrachelus socialis* Bondar) resistance phenotyping in cassava." (2022).
 34. Wang, Fang. "SIOX plugin in ImageJ: area measurement made easy." *UV4Plants Bulletin* 2016, no. 2 (2017): 37-44. DOI: <https://doi.org/10.19232/uv4pb.2016.2.11>

35. Bonhomme, Vincent, Sandrine Picq, Cédric Gaucherel, and Julien Claude. "Momocs: outline analysis using R." *Journal of Statistical Software* 56, no. 13 (2014): 24-p. DOI: <https://doi.org/10.18637/jss.v056.i13>.
36. Weller, Hannah I.; Westneat, Mark W. Quantitative color profiling of digital images with earth mover's distance using the R package colordistance. *PeerJ*, 2019, vol. 7, p. e6398. DOI: [10.7717/peerj.6398](https://doi.org/10.7717/peerj.6398)
37. Rubner, Yossi, Carlo Tomasi, and Leonidas J. Guibas. "The earth mover's distance as a metric for image retrieval." *International journal of computer vision* 40, no. 2 (2000): 99-121. DOI: <https://doi.org/10.1023/A:1026543900054>
38. Pal, Mahesh. "Random forest classifier for remote sensing classification." *International journal of remote sensing* 26, no. 1 (2005): 217-222. DOI: <https://doi.org/10.1080/01431160412331269698>
39. Kuhn, M., J. Wing, S. Weston, A. Williams, C. Keefer, and A. Engelhardt. "Classification and Regression Training. R package version 6.0–81." (2018).
40. Brewer, S., and M. A. Liaw. "Package 'randomforest' University of California." Berkeley: Berkeley, CA, USA (2018).
41. Paluszynska, Aleksandra, Przemyslaw Biecek, Yue Jiang, and Maintainer Yue Jiang. "Package 'randomForestExplainer'." Explaining and visualizing random forests in terms of variable importance (2017).
42. Kennedy, Robert E., et al. "Attribution of disturbance change agent from Landsat time-series in support of habitat monitoring in the Puget Sound region, USA." *Remote Sensing of Environment* 166 (2015): 271-285. DOI: <https://doi.org/10.1016/j.rse.2015.05.005>
43. Conejo Rodriguez, Diego Felipe, Milan Oldřich Urban, Marcela Santaella, Javier Mauricio Gereda, Aquiles Darghan Contreras, and Peter Wenzl. "Using phenomics to identify and integrate traits of interest for better-performing common beans: A validation study on an interspecific hybrid and its *Acutifolii* parents." *Frontiers in Plant Science* (2022). DOI: <https://doi.org/10.3389/fpls.2022.1008666>.
44. Mouillot, David, Nicholas AJ Graham, Sébastien Villéger, Norman WH Mason, and David R. Bellwood. "A functional approach reveals community responses to disturbances." *Trends in ecology & evolution* 28, no. 3 (2013): 167-177. DOI: <https://doi.org/10.1016/j.tree.2012.10.004>
45. Freytag, George F.; Debouck, Daniel G. 2002. Taxonomy, distribution, and ecology of the genus *Phaseolus* (Leguminosae-Papilionoideae) in North America, Mexico and Central America. Botanical Research Institute of Texas (BRIT), Forth Worth, TX, USA. 298 p. (Sida, botanical miscellany no. 23)
46. Valls, José Francisco Montenegro. "Origem do germoplasma de *Arachis pintoi* disponível no Brasil. "Red Internacional de Evaluación de Pastos Tropicales RIEPT. Reunión Sabanas (1, 1992, Brasília, Brasil). Resultados presentados."(1992). <https://hdl.handle.net/10568/56399>

47. Debouck, Daniel G. "Phaseolus germplasm exploration." In Genetic resources of Phaseolus beans, pp. 3-29. Springer, Dordrecht, 1988. DOI: https://doi.org/10.1007/978-94-009-2786-5_1.
48. Simpson, C. E., A. Krapovickas, and J. F. M. Valls. "History of Arachis including evidence of *A. hypogaea* L. progenitors." Peanut Science 28, no. 2 (2001): 78-80. DOI: <https://doi.org/10.3146/i0095-3679-28-2-7>
49. Chacón-Sánchez, María Isabel. "The domestication syndrome in Phaseolus crop plants: A review of two key domestication traits." Origin and Evolution of Biodiversity (2018): 37-59. DOI: https://doi.org/10.1007/978-3-319-95954-2_3.
50. Plazas, Mariola, Isabel Andújar, Santiago Vilanova, Pietro Gramazio, F. Javier Herraiz, and Jaime Prohens. "Conventional and phenomics characterization provides insight into the diversity and relationships of hypervariable scarlet (*Solanum aethiopicum* L.) and gboma (*S. macrocarpon* L.) eggplant complexes." Frontiers in Plant Science 5 (2014): 318. DOI: <https://doi.org/10.3389/fpls.2014.00318>
51. Sankaran, Sindhuja, Meng Wang, and George J. Vandemark. "Image-based rapid phenotyping of chickpeas seed size." Engineering in agriculture, environment and food 9, no. 1 (2016): 50-55. ISSN : 1881-8366
52. Migicovsky, Zoë, Mao Li, Daniel H. Chitwood, and Sean Myles. "Morphometrics reveals complex and heritable apple leaf shapes." Frontiers in plant science 8 (2018): 2185. DOI: <https://doi.org/10.3389/fpls.2017.02185>
53. Rosero, Amparo, José-Luis Pérez, Deisy Rosero, William Burgos-Paz, Remberto Martínez, Julio Morelo, Iván Pastrana, Esteban Burbano, and Alfredo Morales. "Morphometric and colourimetric tools to dissect morphological diversity: An application in sweet potato [*Ipomoea batatas* (L.) Lam.]." Genetic resources and crop evolution 66, no. 6 (2019): 1257-1278. DOI: <https://doi.org/10.1007/s10722-019-00781-x>
54. García-Fernández, Carmen, Ana Campa, Alvaro Soler Garzón, Phil Miklas, and Juan Jose Ferreira. "GWAS of pod morphological and color characters in common bean." BMC plant biology 21, no. 1 (2021): 1-13. DOI: <https://doi.org/10.1186/s12870-021-02967-x>
55. Marsh, Jacob I., Haifei Hu, Mitchell Gill, Jacqueline Batley, and David Edwards. "Crop breeding for a changing climate: Integrating phenomics and genomics with bioinformatics." Theoretical and Applied Genetics 134, no. 6 (2021): 1677-1690. DOI: <https://doi.org/10.1007/s00122-021-03820-3>
56. Krapovickas, Antonio, Walton C. Gregory, David E. Williams, and Charles E. Simpson. "Taxonomy of the genus *Arachis* (Leguminosae)." Bonplandia 16 (2007): 7-205. <http://www.jstor.org/stable/41941433>.
57. Mina-Vargas, Angela M., Peter C. McKeown, Nicola S. Flanagan, Daniel G. Debouck, Andrzej Kilian, Trevor R. Hodgkinson, and Charles Spillane. "Origin of year-long bean (*Phaseolus dumosus* Macfady, Fabaceae) from reticulated hybridization events between multiple *Phaseolus* species." Annals of botany 118, no. 5 (2016): 957-969. DOI: <https://doi.org/10.1093/aob/mcw138>

58. Debouck, Daniel G. "Diversity in Phaseolus species in relation to the common bean." In Common bean improvement in the twenty-first century, pp. 25-52. Springer, Dordrecht, 1999.
59. Leclerc, Camille, Sébastien Villéger, Clara Marino, and Céline Bellard. "Global changes threaten functional and taxonomic diversity of insular species worldwide." *Diversity and Distributions* 26, no. 4 (2020): 402-414. DOI: <https://doi.org/10.1111/ddi.13024>
60. Mouillot, David, Sébastien Villéger, Valeriano Parravicini, Michel Kulbicki, Jesus Ernesto Arias-González, Mariana Bender, Pascale Chabanet et al. "Functional over-redundancy and high functional vulnerability in global fish faunas on tropical reefs." *Proceedings of the National Academy of Sciences* 111, no. 38 (2014): 13757-13762. DOI: <https://doi.org/10.1073/pnas.1317625111>
61. Teixidó, Nuria, Maria Cristina Gambi, Valeriano Parravacini, Kristy Kroeker, Fiorenza Micheli, Sébastien Villéger, and Enric Ballesteros. "Functional biodiversity loss along natural CO₂ gradients." *Nature communications* 9, no. 1 (2018): 1-9. DOI: <https://doi.org/10.1038/s41467-018-07592-1>
62. Cappelatti, Laura, Alizée RL Mauffrey, and John N. Griffin. "Functional diversity of habitat formers declines scale-dependently across an environmental stress gradient." *Oecologia* 194, no. 1 (2020): 135-149. DOI: <https://doi.org/10.1007/s00442-020-04746-1>.
63. Parra-Quijano, M., J. M. Iriondo, and E. Torres. "Improving representativeness of genebank collections through species distribution models, gap analysis and ecogeographical maps." *Biodiversity and Conservation* 21, no. 1 (2012): 79-96. DOI: <http://dx.doi.org/10.1007/s10531-011-0167-0>
64. Castañeda-Álvarez, N. P., Khoury, C. K., Achiganoy, H. A., Bernau, V., Dempewolf, H., Eastwood, R. J., ... & Toll, J. (2016). Global conservation priorities for crop wild relatives. *Nature plants*, 2(4), 1-6. DOI: <https://doi.org/10.1038/nplants.2016.22>
65. Stenberg, J. A., & Ortiz, R. Focused identification of germplasm strategy (FIGS): polishing a rough diamond. *Current Opinion in Insect Science*, (2021). 45, 1-6. DOI: <https://doi.org/10.1016/j.cois.2020.11.001>
66. Dohle, Sarah, Jorge Carlos Berny Mier y Teran, Ashley Egan, Theodore Kisha, and Colin K. Khoury. "Wild Beans (*Phaseolus* L.) of North America." In *North American Crop Wild Relatives*, Volume 2, pp. 99-127. Springer, Cham, 2019. DOI: https://doi.org/10.1007/978-3-319-97121-6_4.
67. Fahlgren, Noah, Malia A. Gehan, and Ivan Baxter. "Lights, camera, action: high-throughput plant phenotyping is ready for a close-up." *Current opinion in plant biology* 24 (2015): 93-99. DOI: <https://doi.org/10.1016/j.pbi.2015.02.006>.
68. Yoosefzadeh-Najafabadi, Mohsen, Hugh J. Earl, Dan Tulpan, John Sulik, and Milad Eskandari. "Application of machine learning algorithms in plant breeding: predicting yield from hyperspectral reflectance in soybean." *Frontiers in plant science* 11 (2021): 624273. DOI: <https://doi.org/10.3389/fpls.2020.624273>

69. Perez-Sanz, Fernando, Pedro J. Navarro, and Marcos Egea-Cortines. "Plant phenomics: An overview of image acquisition technologies and image data analysis algorithms." *GigaScience* 6, no. 11 (2017): gix092. DOI: <https://doi.org/10.1093/gigascience/gix092>
70. Ubbens, Jordan R., and Ian Stavness. "Deep plant phenomics: a deep learning platform for complex plant phenotyping tasks." *Frontiers in plant science* 8 (2017): 1190. DOI: <https://doi.org/10.3389/fpls.2017.01190>
71. Tardieu, François, Llorenc Cabrera-Bosquet, Tony Pridmore, and Malcolm Bennett. "Plant phenomics, from sensors to knowledge." *Current Biology* 27, no. 15 (2017): R770-R783. DOI: <http://dx.doi.org/10.1016/j.cub.2017.05.055>
72. Parmley, Kyle, Koushik Nagasubramanian, Soumik Sarkar, Baskar Ganapathysubramanian, and Asheesh K. Singh. "Development of optimized phenomic predictors for efficient plant breeding decisions using phenomic-assisted selection in soybean." *Plant Phenomics* 2019 (2019). DOI: <https://doi.org/10.34133/2019/5809404>
73. Parmley, Kyle A., Race H. Higgins, Baskar Ganapathysubramanian, Soumik Sarkar, and Asheesh K. Singh. "Machine learning approach for prescriptive plant breeding." *Scientific reports* 9, no. 1 (2019): 1-12. DOI: <https://doi.org/10.1038/s41598-019-53451-4>
74. Debouck, D. G. Genetic resources of Phaseolus beans: patterns in time, space, and people. Segundo seminario de la Judia. Actas de la asociación Española de leguminosas. 2000
75. Zavafer, A., Bates, H., Mancilla, C., & Ralph, P. J. (2023). Phenomics: conceptualization and importance for plant physiology. *Trends in Plant Science*. DOI: <https://doi.org/10.1016/j.tplants.2023.03.023>

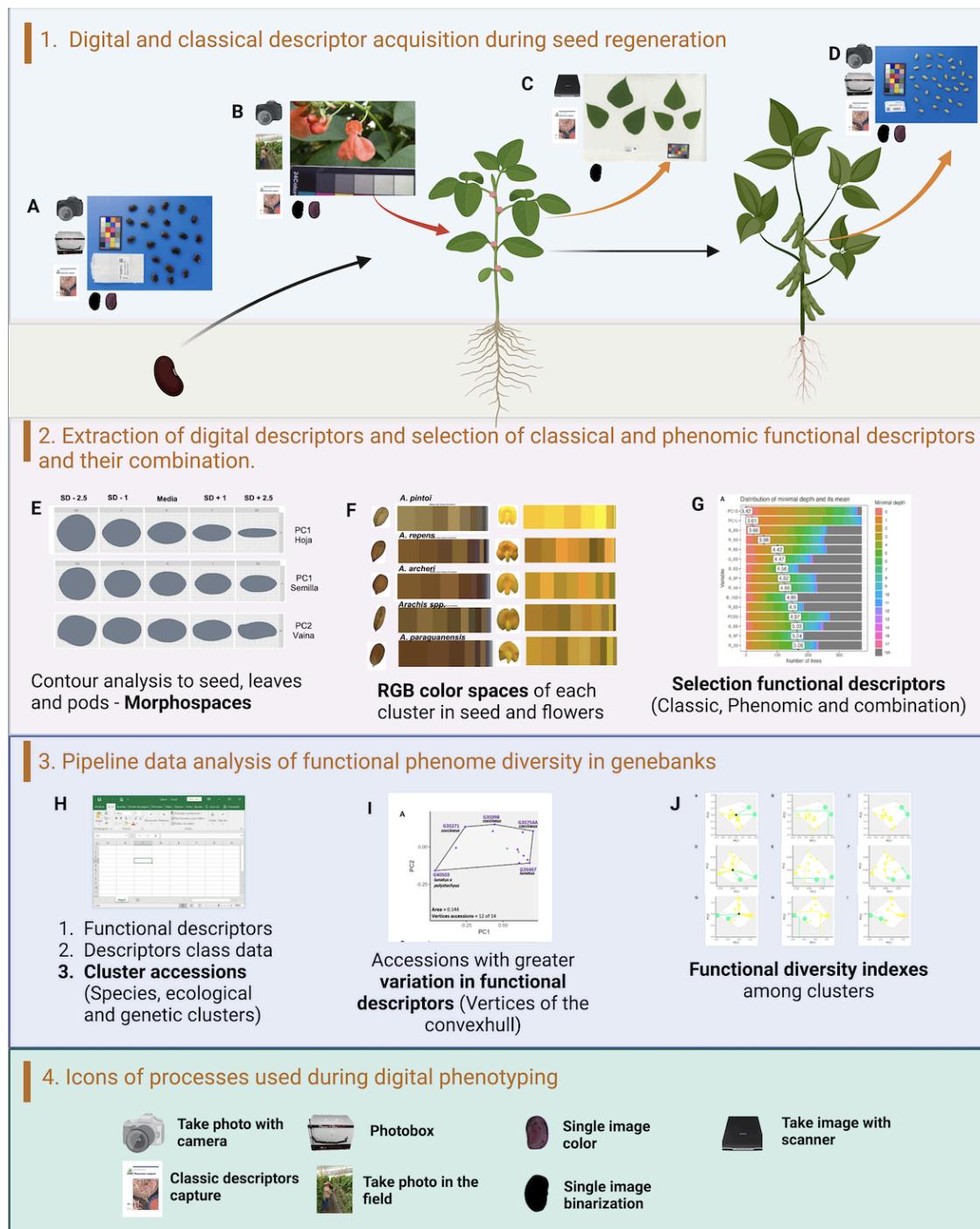


Figure 2.1 Methodological procedure for the phenotype characterization of *Phaseolus* and *Arachis* accessions. (1) Digital phenotyping during seed regeneration: A) Seed photographic capture using photobox and classic descriptors. Color and binarize image extraction; B) Field photographic capture of flowers and classic color descriptors. Color image extraction for digital colorimetry; C) Scanning of trifoliate leaves and classical descriptors. Binarized image extraction for geometric morphometrics analysis; D) Image capture of pods in the photo box and classic

descriptors. Binarized image is extracted for geometric morphometrics analysis. (2) Extraction of phenomic descriptors and selection: E) Contour analysis using geometric morphometry of seed, leaves and pods. The morphospaces (principal components -PC) are determined; F) Extraction of RGB color spaces from each colorimetric group using kmeans in seeds and flowers; G) Selection of descriptors using random forests from the mean minimum depth. (3) Data analysis of functional phenome: H) Three databases are generated that represent the selected functional descriptors, the class of the descriptors (Character, numerical and ordinal) and the associated groups that are the functional entities (Species, ecological group and genetic group); I) Vertices accessions of the functional spaces. These accessions represent the greatest phenotypic variation; J) Functional diversity indexes that relate accessions to functional entities. Functional specialization (FSpe), functional originality (FOri) and functional independence (FIde) are calculated. (4) Icons relating the stages of digital phenotyping. Image capture and image preprocessing are observed.

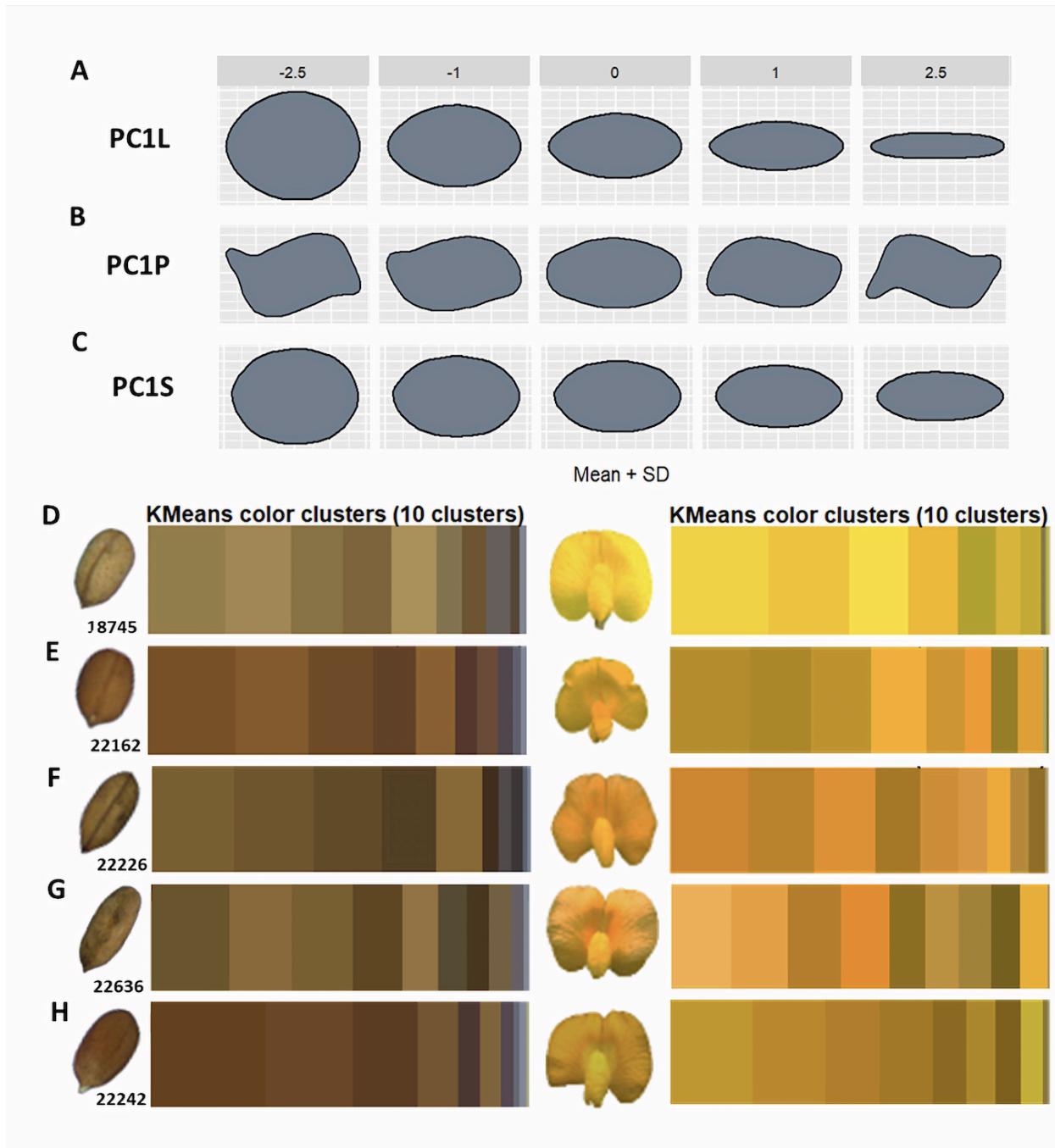


Figure 2.2 Functional phenomic descriptors that describe the phenotypic diversity of *Arachis* spp. accessions. (A) Principal component 1 of leaflet morphometrics. (B) Principal component 1 of pods morphometrics. (C) Principal component 1 of seeds morphometrics. (D) Flower and seed colorimetric clusters of accession 18745 *A. pintoi*. (E) Flower and seed colorimetric clusters of accession 22162 *A. repens*. (F) Flower and seed colorimetric clusters of accession 22226 *A. archeri*. (G) Flower and seed colorimetric clusters of accession 22636 *A. paraguariensis*. (H) Flower and seed colorimetric clusters of accession 22242 *Arachis* sp.

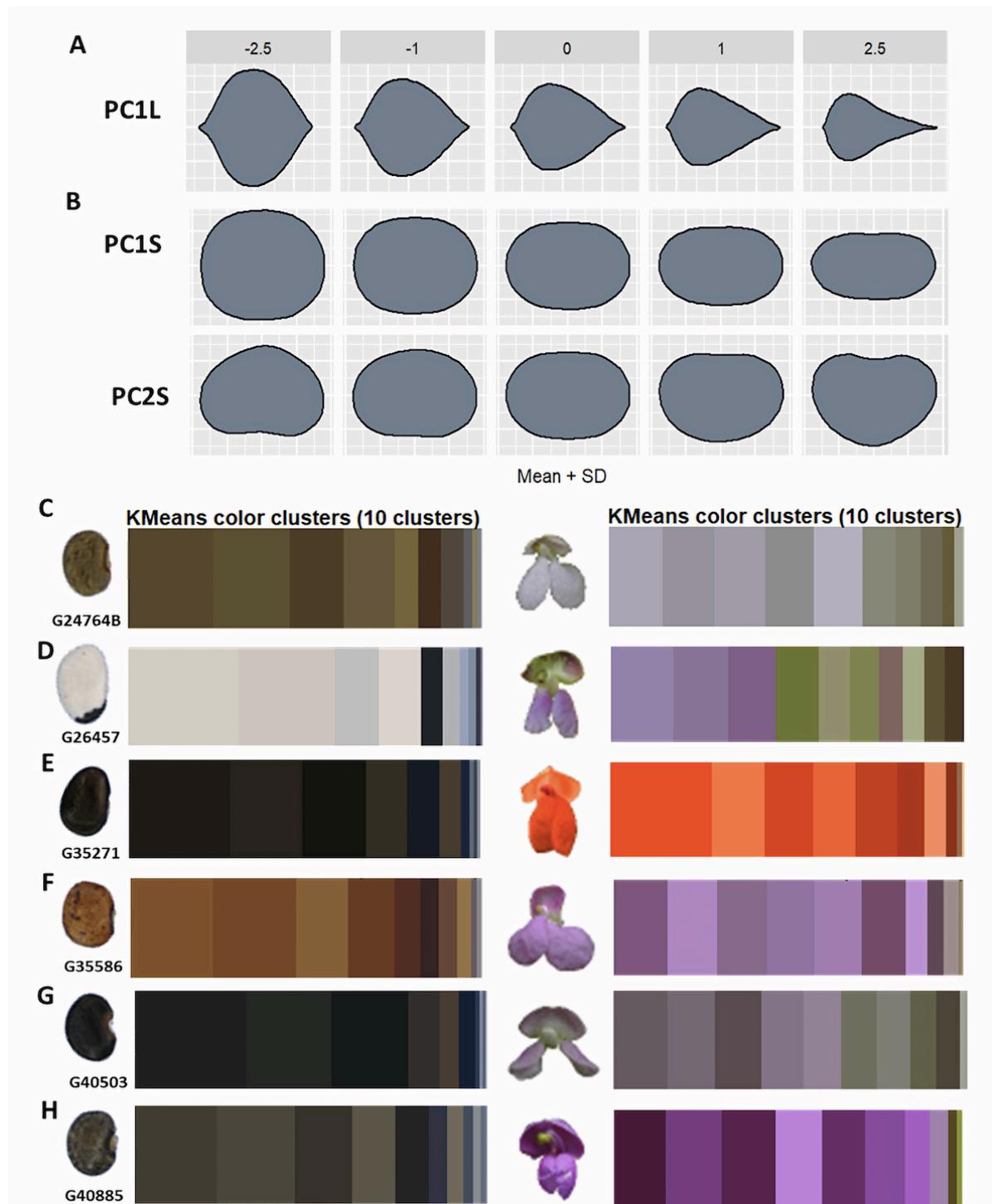


Figure 2.3 Functional phenomic descriptors that describe the phenotypic diversity of *Phaseolus* spp. accessions. (A) Principal component 1 of leaflet morphometrics. B) Principal components 1 and 2 of seeds morphometrics. (C) Flower and seed colorimetric clusters of accession G24764B *P. dumosus* × *P. vulgaris*. (D) Flower and seed colorimetric clusters of accession G26457 *P. lunatus*. (E) Flower and seed colorimetric clusters of accession G35271 *P.*

coccineus. (F) Flower and seed colorimetric clusters of accession G35586 *P. dumosus*. (G) Flower and seed colorimetric clusters of accession G40503 x (*P. lunatus* x *P. polystachyus*. (H) Flower and seed colorimetric clusters of accession G40885 *P. tuerckheimii*.

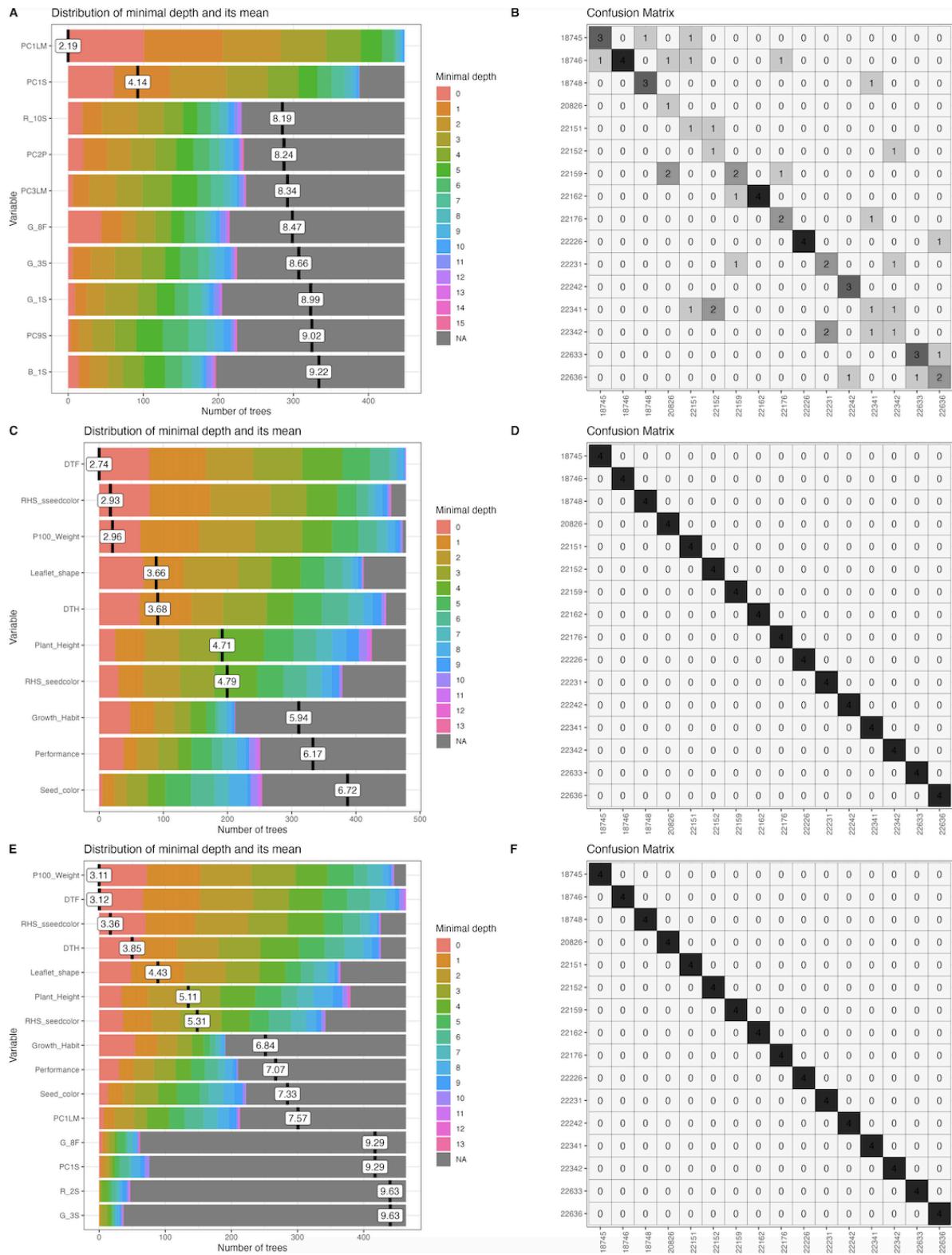


Figure 2.4 Selection of classical descriptors, phenomic descriptors and their combination used a minimal depth distribution on *Arachis* spp accessions and its classification. (A) Distribution of minimal depth of the phenomic descriptors (B) Classification of accessions using confusion matrix from phenomic descriptors. (C) Distribution of minimal depth of classical descriptors. (D) Classification of accessions using confusion matrix from classical descriptors. (E) Distribution of minimal depth of the combination of phenomic and classical. (F) Classification of species using confusion matrix from combining phenomic descriptors with classical descriptors.

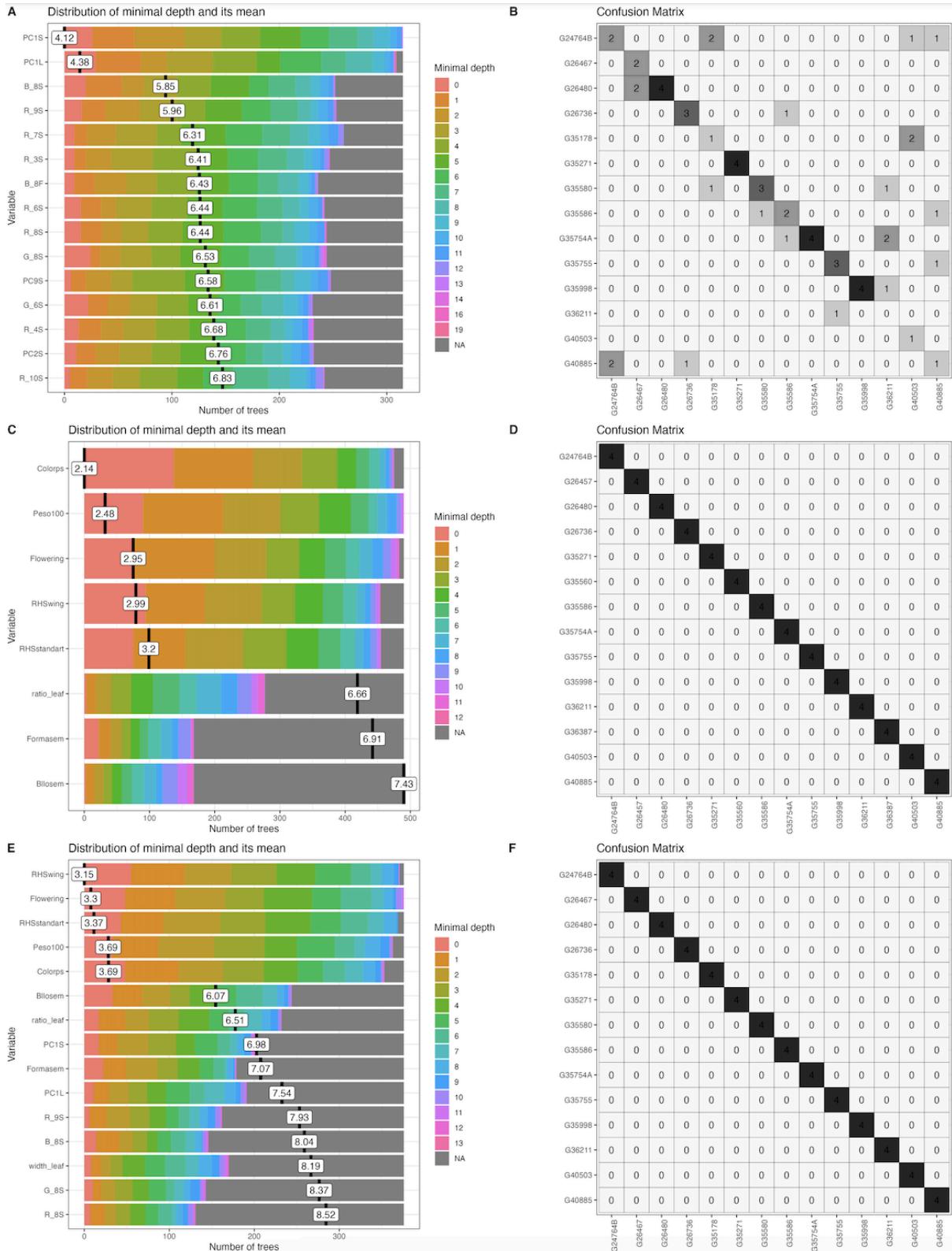


Figure 2.5 Selection of classical descriptors, phenomic descriptors and their combination evaluated via a minimal depth distribution on *Phaseolus* spp accessions and its classification. (A)

Distribution of minimal depth of the phenomic descriptors (B) Classification of accessions via confusion matrix from phenomic descriptors. (C) Distribution of minimal depth of classical descriptors. (D) Classification of accessions via confusion matrix from classical descriptors. (E) Distribution of minimal depth of the combination of phenomic and classical descriptors. (F) Classification of species via confusion matrix from combining phenomic descriptors with classical descriptors.

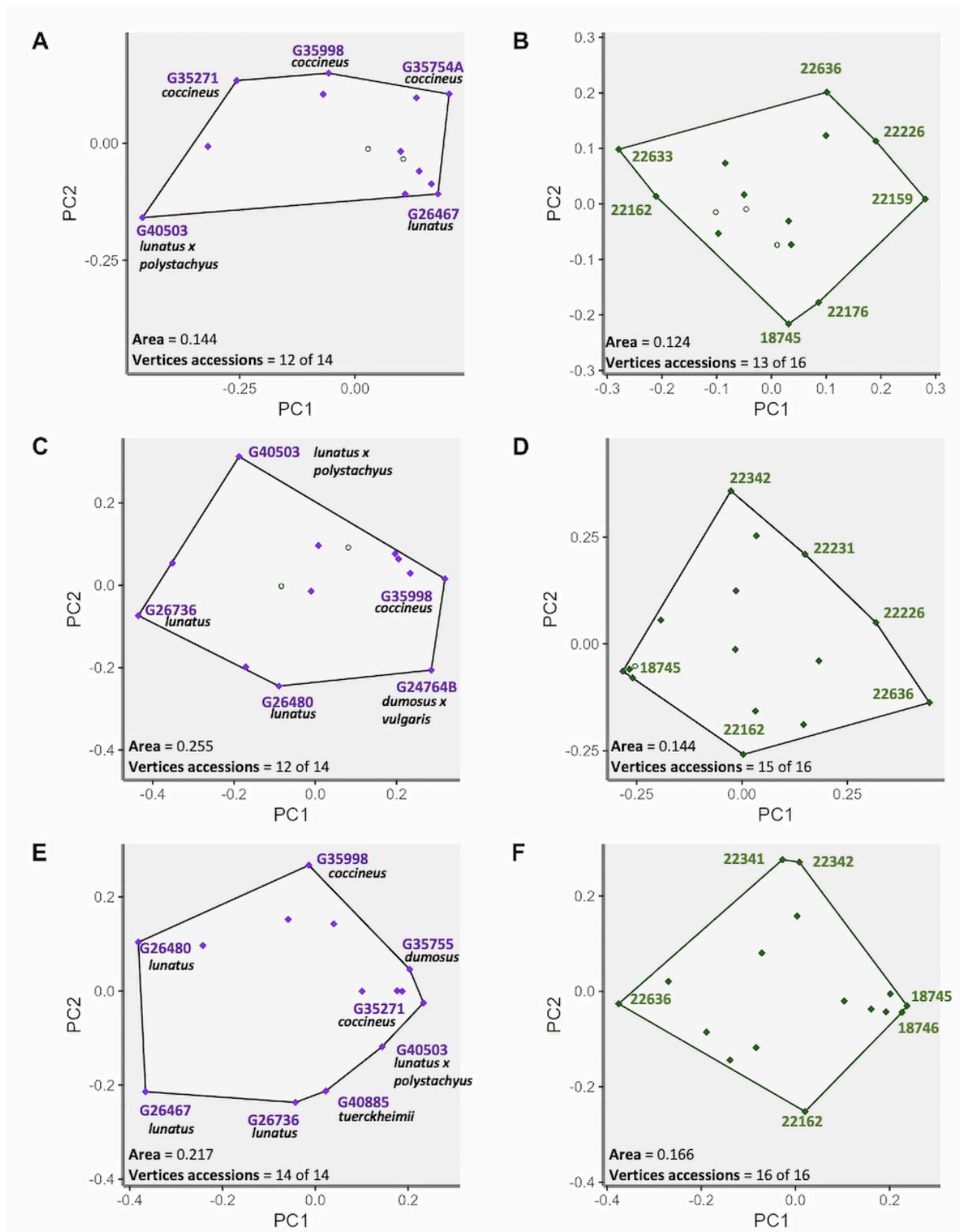


Figure 2.6 Vertices accessions of the functional space for each of the descriptor types used for bean and forage peanut accessions. (A-B) Bean and forage peanut vertices accessions with phenomic descriptors. (C-D) bean and forage peanut vertices accessions with classical

descriptors. (E - F) bean and forage peanut vertices accessions with the combination of descriptors. The value of the area of the convex hull and the number of vertex accessions for each of the descriptor types for both beans and forage peanuts.

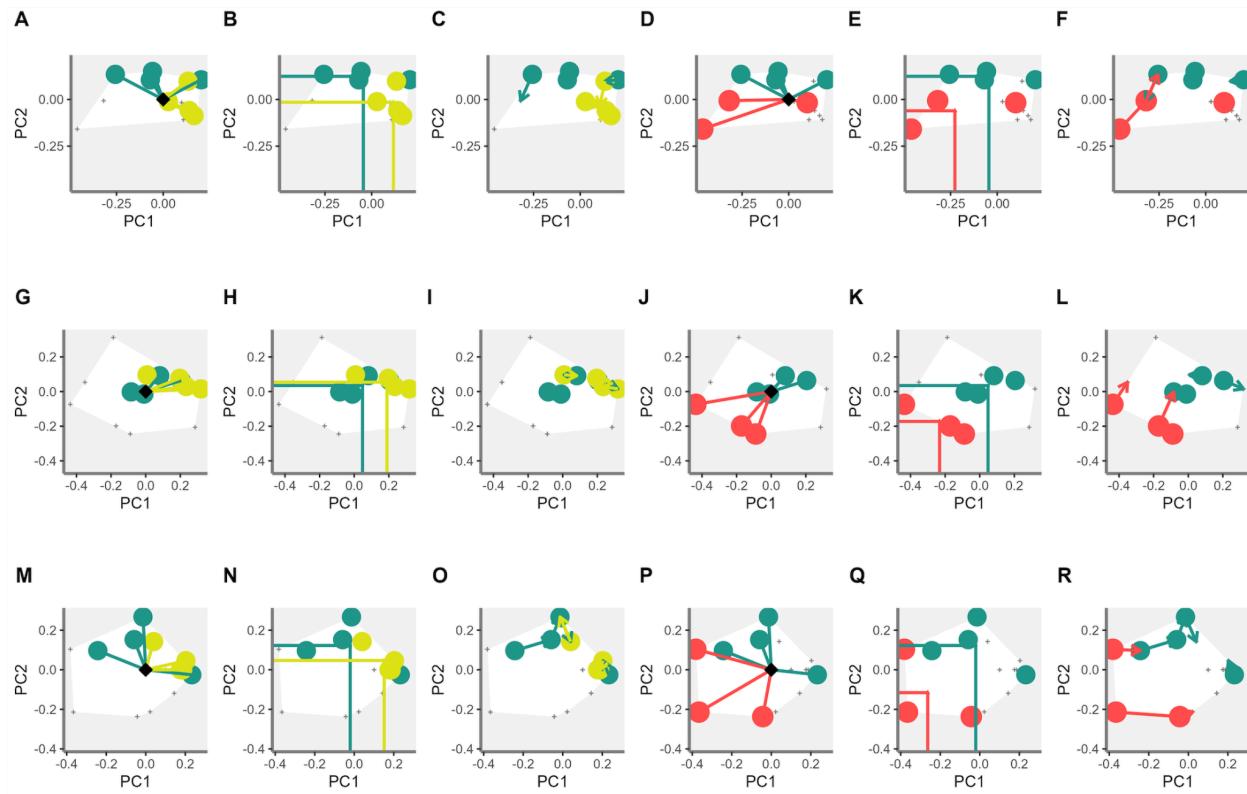


Figure 2.7 Functional diversity indices of the accession grouping by *Phaseolus* spp. for each type of descriptor. (A-F) Functional diversity indices for phenomic descriptors in the accession grouping of *P. dumosus*, *P. coccineus* and *P. lunatus*. (G-L) functional diversity indices for classical descriptors in the accession grouping of *P. dumosus*, *P. coccineus* and *P. lunatus*. (M-R) functional diversity indices for combined descriptors in the accession grouping of *P. dumosus*, *P. coccineus* and *P. lunatus*. For the indices of functional specialization (FSpe), functional identity (FId), and functional originality (FOri), the relationship between the first two primary coordinates (PC1, PC2) is observed.

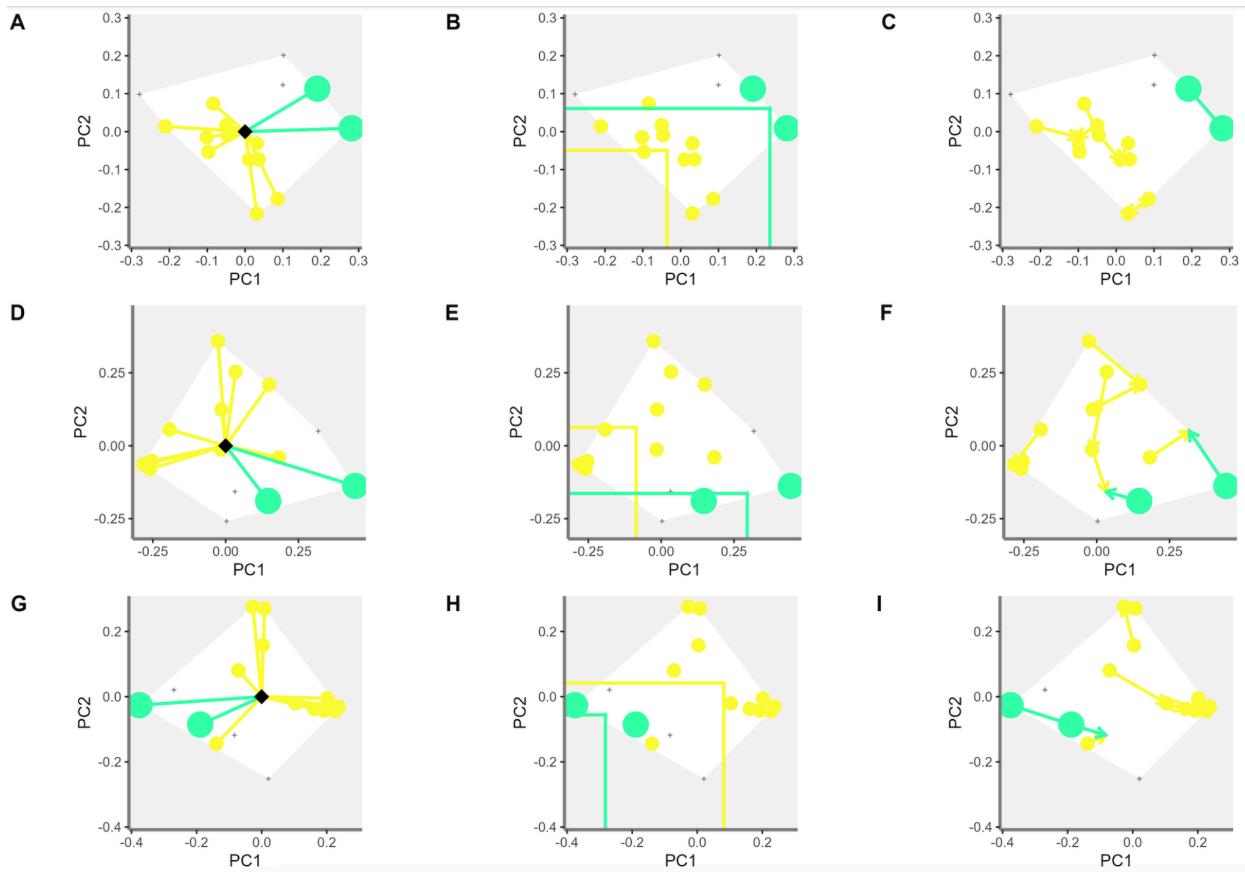


Figure 2.8 Functional diversity indices of the accession grouping by *Arachis* species for each type of descriptor. (A-C) Functional diversity indices for phenomic descriptors in the accession grouping of *A. pintoi* and *A. paraguariensis*. (D-F) functional diversity indices for classical descriptors in the accession grouping of *A. pintoi* and *A. paraguariensis*. G-I) functional diversity indices for combined descriptors in the accession grouping of *A. pintoi* and *A. paraguariensis*. For the indices of functional specialization (FSpe), functional identity (FIde), and functional originality (FOri), the relationship between the first two primary coordinates (PC1, PC2) is observed.

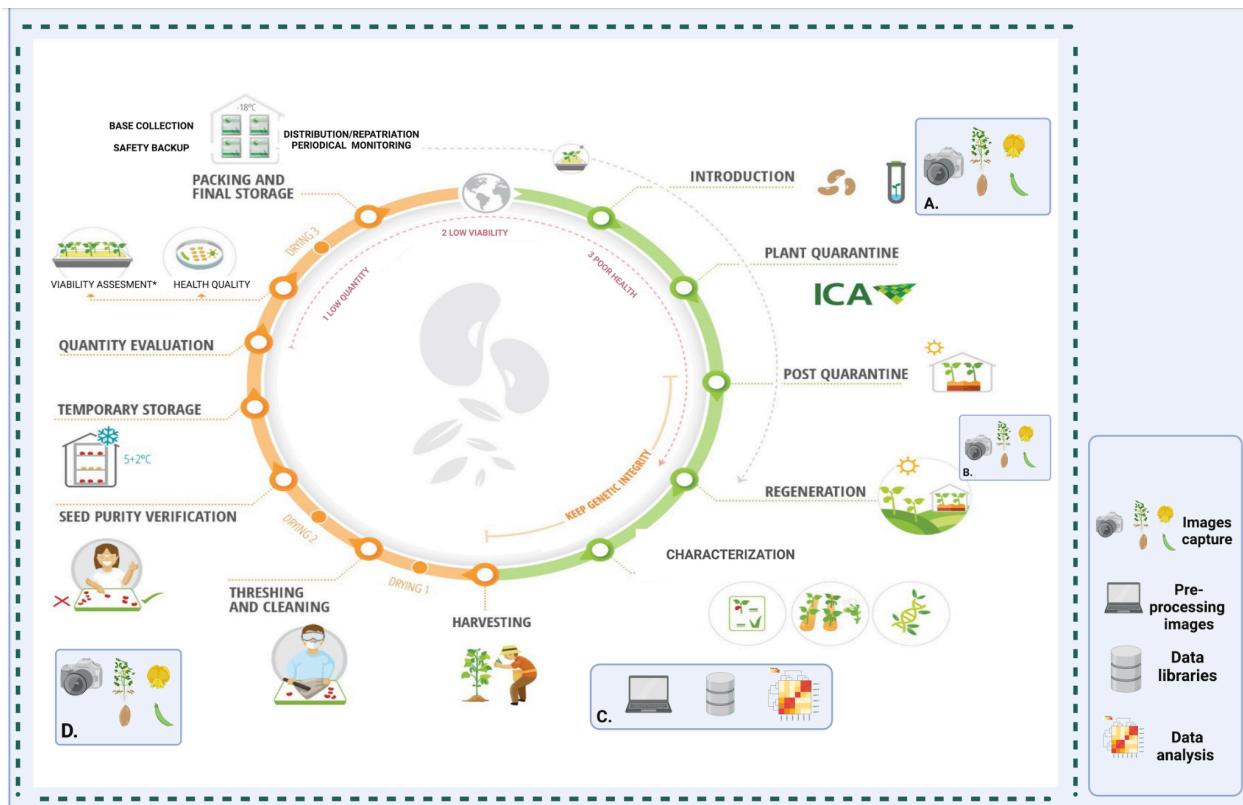


Figure 2.9. CIAT genebank workflow suggestion resulting from the proposed methodology.

(A) Capture of images of flower, pods, leaves, whole plant and herbarium sampling and digitization during the collection of new accessions. (B) Capture of images of flower, pods, leaves, whole plant and herbarium sampling and digitization during the seed multiplication and regeneration. (C) Image processing, morphometric and colorimetric data extraction, descriptor selection and data analysis with AI to classify accessions. (D) Capture images of pods and seeds, to feed the database and in the seed purity verification process.

Table 2.1 Indices of functional diversity for the accessions of Phaseolus and Arachis species evaluated with phenomic (Phe), classic (Cla) and their combined (Comb) descriptors. The indices of functional specialization (FSpe) and functional originality (FOri).

Species	Fspe			Fori		
	Phe	Cla	Com	Phe	Cla	Com
<i>P. coccineus</i> n = 4	0.45	0.40	0.62	0.51	0.33	0.60
<i>P. dumosus</i> n = 4	0.32	0.61	0.47	0.47	0.25	0.26
<i>P. lunatus</i> n = 3	0.67	0.82	0.84	0.78	0.62	0.73
<i>P. tuerckheimii</i> n = 1	0.23	0.88	0.55	0.37	0.52	0.25
<i>x(P.dumosus P.vulgaris)</i> n = 1	0.36	0.96	0.68	0.44	0.69	0.72
<i>x(P.lunatus P.polystachyus)</i> n = 1	0.46	1.00	0.47	0.32	1.00	0.44
<i>A. archeri</i> n = 1	0.78	0.82	0.86	0.76	0.60	0.85
<i>A. paraguariensis</i> n = 3	0.79	0.86	0.88	0.74	0.74	0.60
<i>A. pintoi</i> n = 11	0.42	0.64	0.67	0.36	0.45	0.51
<i>A. repens</i> n = 1	1.00	0.68	0.76	1.00	0.75	1.00
<i>Arachis</i> spp. n = 1	0.68	0.40	0.53	0.89	0.52	0.41

3. Chapter three: Representativeness functional phenome and ecogeographic diversity in genebanks: A case study on the tepary bean collection

Diego Felipe Conejo-Rodríguez^{1,2 *}, Joaquin Guillermo Ramírez-Gil³, Andrés J. Cortés^{4,5}, Juan Carlos Suárez^{6,7}, John A. Ocampo², Peter Wenzl¹ and Milan Oldřich Urban^{1*}

¹ Genetic Resources Program, International Center for Tropical Agriculture, Palmira, Valle del Cauca, Colombia. Bean physiology, International Center for Tropical Agriculture, Palmira, Valle del Cauca, Colombia

² Departamento de Ciencias Agrarias / Biológicas, Facultad de Ciencias Agropecuarias, Universidad Nacional de Colombia, Palmira, Valle del Cauca, Colombia

³Laboratorio de Agrocomputación y Análisis Epidemiológico, Center of Excellence in Scientific Computing, Departamento de Agronomía, Facultad de Ciencias Agrarias, Universidad Nacional de Colombia, Bogotá, Colombia

⁴ Corporación Colombiana de Investigación Agropecuaria - AGROSAVIA, Centro de Investigación La Selva, Km 7 vía Rionegro-Las Palmas, Rionegro, Colombia

⁵ Facultad de Ciencias Agrarias - Departamento de Ciencias Forestales, Universidad Nacional de Colombia - Sede Medellín, Medellín, Colombia

⁶ Programa de Ingeniería Agroecológica, Facultad de Ingeniería, Universidad de la Amazonia, Florencia, Colombia.

⁷ Centro de Investigaciones Amazónicas CIMAZ Macagual César Augusto Estrada González, Grupo de Investigaciones Agroecosistemas y Conservación en Bosques Amazónicos-GAIA. Universidad de la Amazonia.

Corresponding Author:

Diego Felipe Conejo Rodriguez - Milan Urban

Km 7 Recta Palmira – Cali, Colombia

Email address: dfconejoro@unal.edu.co; dehydrins@gmail.com

Abstract

The tepary bean (*Phaseolus acutifolius*) is important in breeding because of its ability to adapt to abiotic stress conditions. There are gaps in knowledge about how the tepary bean's genetic and phenotypic adaptations were influenced by natural selection and domestication. Our study explores the interaction between ecological, genetic, and functional factors in tepary bean. We first identify ecological groupings based on spatial data and the collection coordinates of the evaluated accessions. Dimensionality reduction techniques and machine learning classification algorithms are utilized to discover the most important environmental variables. Functional diversity indices, such as functional identity, functional specialization, and functional originality, were then determined and linked to the identified ecological groups. The genetic, functional, and ecological components' distances from the centroid were then calculated and related. Two ecological groupings were found in wild populations and four in cultivated populations. In the wild, the groupings are distinguished by soil roughness and rainfall. The *tenuifolius* variant loves sunny and rocky soils, whereas *acutifolius* grows in shaded and damp environments. We identified the wild and cultivated accessions with the highest functional phenotypic variation and proposed them as the tepary bean functional trait collection of the International Center for Tropical Agriculture Future Seeds genebank. Wild accessions exhibit phenotypic variability as a result of their ecological conditions, whereas cultivated accessions exhibit phenotypic variation as a result of the indigenous Bavavavavi O'odham communities of Arizona; however, due to phenotypic redundancies with the common bean, Mesoamerican accessions undergo unconscious selection. Our research looks at how integrating genetic, ecological, geographical, and phenotypic data might provide insight into evolutionary adaptation processes and help document genebank collections.

Keywords: Domestication, Crop wild relatives, Centroid distance, Spatial data, machine learning, Functional diversity.

3.1 Introduction

The tepary bean (*Phaseolus acutifolius* A. Gray) has lately attracted interest in modern crop breeding efforts as a source of genetic and physiological traits for abiotic stress resistance (Moghaddam et al., 2021). Tepary bean is a native plant from southwestern USA and Mexico deserts, but its use is limited due to poor seed quality, low yield, and problems associated with cooking properties (Porch and Beaver, 2012). Indigenous peoples of what is now northwestern Mexico and the southwestern United States domesticated the tepary bean first (Bornowski et al 2023). Tepary was grown by farmers as part of a cropping system that also included maize (*Zea mays* L), common beans (*Phaseolus vulgaris* L.), and squash (*Cucurbita* spp.) (Teiwes and Nabhan, 1983). In addition, tepary bean is also an important crop for subsistence of rural communities in southern Africa, primarily farmed by smallholder farmers (Nong et al., 2023).

The tepary bean has two botanical varieties: (I) *tenuifolius* and (II) *acutifolius* (c) (Freytag and Debouck., 2002). The fundamental difference between these two types is in leaf morphology, with the *tenuifolius* variety having lanceolate leaves and the *acutifolius* variety having heart-shaped leaves (Freytag and Debouck., 2002). Similarly, these two kinds can easily hybridize, resulting in a diverse genetic background among accessions (Bornowski et al., 2023).

Recently, the genetic variation of a diversity panel of 422 tepary bean accessions was determined, identifying six genetic groups: (GG1) Sonora-Sinaloa cultivated, (GG2) southeast AZ wild *tenuifolius*, (GG3) Central America cultivated, (GG4) Chihuahua wild *tenuifolius*, (GG5) Sonora-Sinaloa wild *acutifolius*, and (GG5) Durango wild (Bornowski et al., 2023).

However, the relationship of these groups with their phenotypic and environmental variation has not been studied in germplasm collections, indicating a gap in the knowledge of this species that allows better use. Characterization and evaluation of collected accessions is an essential basis for genebank knowledge, management, and use (Conejo - Rodriguez et al., 2024). Functional and geographic diversity analyses are vital tools for characterizing genebank entries (Calliope et al 2018; Dwivedi et al., 2017; Parra-Quijano et al 2012; Conejo - Rodriguez et al., 2024). Functional diversity analysis evaluates phenotypic and genetic variation across spatial and time, elucidating potential agricultural, ecological, and evolutionary benefits (Mammola et al., 2021; Conejo - Rodriguez et al., 2024). Meanwhile, geographic diversity analysis focuses on the distribution of genetic resources, aiming to identify geographically distinct gene pools (Parra-Quijano et al 2012). These two approaches enable comprehensive assessment of germplasm collections, enhancing their utility for crop improvement, adaptation, and even conservation strategies.

Consequently, functional and geographic diversity analyses bolster the essential role of genebanks in global food security and biodiversity conservation (Conejo-Rodriguez et al., 2024). Currently, studies of functional and ecogeographic diversity in plant genetic resources for food and agriculture have been limited (Conejo-Rodriguez et al., 2024). Because most germplasm collections were collected under varying environmental conditions that affected the genetic and phenotypic variation of the conserved accessions, germplasm collections provide an excellent opportunity to promote phenome variation studies (Conejo-Rodriguez et al., 2024). The phenome is defined as the interaction of genotype (G), environment (E), and time (t) (Zavafer et al., 2023). In agronomy we can add field management (M). The phenotypic variation of crop wild relatives (CWR) was determined by both t and E, and farmers have selected traits of interest during the

domestication process (Debouck, 2000). As a result, studying functional phenomes (manageable phenome; see Zavafer et al., 2023) in germplasm collections will help genebank collection management generate subsets of accessions that relate environmental influence and phenotypic traits. This will promote the use of accessions with pre-breeding potential by selecting traits closely associated with abiotic and biotic stresses, as well as understanding the role of the environment in phenotypic variation. The geographical and temporal distribution of a species within natural areas often reflects its adaptation to various environmental stresses (Hall et al., 1992). Understanding this distribution provides valuable information for genetic improvement programs aimed at improving resilience in crops (Parra-Quijano et al 2012). In regions prone to drought, certain genotypes may exhibit better tolerance or adaptive traits, making them ideal candidates for selection and breeding (Kamoshita et al.,2008; Mardeh et al., 2006). This information guides the selection process in breeding programs, where genotypes with differential responses to stress can be strategically selected to enhance the resilience of commercial crops (Mardeh et al., 2006). Furthermore, optimizing the search for germplasm by targeting specific geographical regions or ecosystems where favorable genotypes are naturally found accelerates the genetic improvement process (Poggi et al., 2023; Hall et al., 2012). Through this approach, genetic diversity and natural adaptation can be preserved and used more effectively to develop cultivars or varieties that can thrive in challenging environments such as drought and high temperatures. Therefore, our goal is to understand to what extent the functional variation of selected traits in tepary bean collection is influenced by its geographic distribution, genetic groups, and ecogeographic clusters.

3.2 Materials and Methods

3.2.1 Methodological approach

This study's approaches proceed is based on three phases that were developed independently for wild and cultivated accessions: (I) Determining the ecological groups using spatial data analysis, data mining, and machine learning; (II) Establishing the spaces of functional variation using phenotypic traits using functional diversity analysis and indices; and (III) Relating the distance from the centroid of the functional, genetic, ecological, and geographic components (Figure 3.1). This research will allow us to identify the tepary bean's evolutionary adaptation processes and develop an idea for its selection-driven domestication.

3.2.2 Ecogeographic variable extraction

The characterization of the ecogeographic variation in the *P. acutifolius* collection was carried out using spatial ecogeographic databases. A total of 61 ecogeographic variables were used including bioclimatic variables from WorldClim 2.1 (Fick and Hijmans, 2017), climatic indices from ENVIREM (Title and Bemmels, 2018), edaphic characteristics from SoilGrids (Hengl et al., 2017), geomorphological data (Amatulli et al., 2018), habitat information (Tuanmu and Jetz, 2015), cloud cover (Wilson and Jetz, 2016) and land use data (Tuanmu and Jetz, 2014) (Table 1S - Ecogeographic variables). We used a spatial resolution of the geodata set of 1 km x 1km. We extracted ecogeographic variables from the collection of 298 *P. acutifolius* CIAT accessions with coordinates and occurrences from the GBIF database that were associated with their geographic distribution (Freytag and Debouck, 2002). For each of the accessions with coordinates, the

ecogeographic variables were extracted using the extract function of the free R software using the raster package.

3.2.3 Dimensionality reduction of ecogeographic variables

Given the large number of ecogeographic variables, a dimensionality reduction approach was conducted using three methods: (I) Principal Component Analysis (PCA), (II) Uniform Manifold Approximation and Projection (UMAP), and (III) Autoencoder. These dimensionality reductions differ, particularly in terms of linear or nonlinear relationships (Tschanen et al. 2018). In the case of PCA it is based on the analyzed variables' linear correlations while retaining the greatest explained variance (Ringnér 2008). Meanwhile the UMAP method optimizes the layout of a low-dimensional data representation in order to reduce the cross-entropy between the two topological representations (McInnes et al. 2018). On the other hand, the autoencoder is a complex three-layer neural network designed to learn a relationship between high-dimensional observations and a lower-dimensional representation space, allowing the original observations to be reconstructed from the lower-dimensional representation (Tschanen et al. 2018). These dimensionality reduction were carried out with the use of the free R software using libraries FactorExtra, renv, kerasR, ruta, and umap. Each of the dimensionality reduction models produced latent variables. The PCA was used to select one of the five main components (PC1-PC5). The UMAP approach decreases the dimensionality of data using Laplacian maps and topological arrays (McInnes et al 2018). UMAP identifies two latent variables. Three latent variables were extracted from noise reduction analysis utilizing denoising and selu activation in Autoencoder. The Autoencoder reduces the dimensionality of data from convolutional neural networks. The ultimate aim was to find latent variables with minimal data loss during dimensionality reduction (Tschanen et al., 2018).

3.2.4 Ecogeographic clustering of tepary bean

The ecogeographic groups were determined via k-means clustering method using latent variables of the dimensionality reduction using PCA, UMAP, and Autoencoder methods described above. Running the FactoExtra library in R, we applied the k-means algorithm. The k-means algorithm was used to build groups based on the distance between the centroid of each identified group (Henao-Rojas et al., 2021). These clustering are formed based on prior knowledge of the species, including its geographic distribution and genetic structure (Bornowski et al. 2023). Four clusters are recommended for cultivated accessions, and two for wild accessions.

The groups were classified based on their biological status. Machine learning techniques were utilized in two ways to discover which environmental variables are important in ecological group discrimination: (I) to select the most accurate dimensionality reduction approach and (II) to select the relevant variables. To choose the dimensionality reduction approach, linear discriminant analysis (LDA), support vector linear and radial machines (SVML, SVMR), multinomial logistic regression (MLR), and naives bayes (NV) algorithms were utilized, and the average precision was calculated. Partial square regression (PLS) was used to determine important variables for the specified approach. The Abs coefficient of the three major components (PLS1, PLS2, and PLS3) was used to select variables of importance, avoiding false positives (Nawar et al., 2023). This analysis was carried out using the tidymodels library under free R software. *P. acutifolius* ecogeographic clusters were named after Mexican biogeographic provinces (Morrone, 2019), given that the majority of the gathered occurrences and accessions are found in this country (Bornowski et al., 2023). Following this step, using the free software R

with the radarchart package and the important ecogeographic variables, we created a radar plot for each ecogeographic group to display the ecogeographic variation.

3.2.5 Functional diversity in tepary bean collection

The functional diversity analysis was carried out based on the variation of phenotypic features of the collection related with the determined ecogeographic clusters. The variation of phenotypic traits were measured as part of the Bioversity - CIAT alliance characterization of the *P. acutifolius* collection in tropical environments at the Colombian Amazon (Suárez et al. 2021). Days to flowering (DF), days to physiological maturity (DPPM), root dry weight (RDW), pod dry weight (PDW), leaf dry weight (LDW), stem dry weight (SDW), leaf area (LA), specific leaf area (SLA), number of pods (PN), and 100-seed weight (W100) were the functional traits employed in this study (Suárez et al. 2021). These phenotypic traits have been reported in connection with domestication syndromes (Chacón-Sánchez, 2018). This fact allows wild accessions to be distinguished from cultivated ones in their intrinsic natural adaptation to local ecogeographical conditions. Each ecogeographic cluster and phenotypic trait were classified in terms of their functional diversity, and the number of accessions (*accession abundance*) was computed.

The class and type of descriptors were then determined (Table 4S - Class of descriptors), including determining whether the descriptor has numeric or character value. Finally, each cluster and accession was related to a specific phenotypic trait (Conejo - Rodriguez et al., 2024). First, we identified the vertices of the *P. acutifolius* accessions with the highest diversity in the functional phenome. Vertices are the boundary accessions that connect the functional space's convex area (Conejo - Rodriguez et al., 2024). Accessions around the convexhull represented the phenotypic variation within the *P. acutifolius* collection; we classified them as a functional

collection. The functional space was one of the ten primary components of the functional diversity analysis. Following that, the ecogeographic clusters were regarded as functional entities (FE), and the accessions were defined for their abundance in each ecogeographic cluster. Functional diversity indices such as functional richness (FRic), functional identity (FIde), functional dispersion (FDis), functional divergence (FDiv), functional evenness (FEve), functional originality (FOri), functional specialization (FSpe), functional mean pairwise distance (FMPD), and functional mean nearest neighbor distance (FNND) were calculated as mentioned in our previous study (Conejo-Rodriguez et al., 2024; Magneville et al., 2022). All analyses were partitioned between cultivated and wild genpools.

To generate and visualize the clusters the free R-library *mFD* was used (Mouillot et al., 2013; Magneville et al., 2022). These indicators demonstrated the level of representativeness and variations of the functional phenotype that was influenced by environmental conditions (Zavafer et al., 2023). The influence of ecogeographic clusters on phenotypic traits was further investigated. Their effect was estimated for the phenotypic parameters used here as "functional traits" (days to flowering (DF), days to physiological maturity (DPPM), pod number (PN), leaf area (LA), and 100-seed weight (W100)). The results were visualized in a *forestplot* using the free R software Forestplot package. Finally, the number of accessions in each ecogeographic group and the genetic group indicated by the genetic structure (Bornowski et al., 2023) were compared. Figure 3.1 depicts the methodological procedure (Figure 3.1 - Methodological approaches).

3.2.6 Merging ecogeographic groups, genetic structure and functional diversity

In order to see the relationship between eco-geographic, functional and genetic clusters, we compared the latent variables and the distance from the centroid for each relationship. Initially, Autoencoder's eco-geographic latent variables (EG: LV1 - LV2) were considered. Following that, the genetic clusters main components (GG: PC1GG - PC2GG), functional diversity components (FD: PC1FD - PC2FD), and geographic components (GGE: latitude and longitude) were interrelated. Finally, the distance from the centroid for each accession (EG, GG, FD, GGE) was determined in relation to one another and the free *R* software was used to visualize all plots.

3.3 Results

The average accuracy of the DR using an Autoencoder (AUT) was higher than the other two methods evaluated (PCA, and UMAP) (Table 1S – Accuracy). The number of ideal clusters is estimated using grouping indices based on the latent variables of the RD, resulting in four groups for cultivated accessions of tepary bean and two groups for wild accessions (Figure 1S). The most important environmental variables in cluster discrimination are reported in Figure 3.2 based on the PLS regression method. The precipitation of the driest month (PDM), precipitation of the driest quarter (PDRM), precipitation of the coldest quarter (PCOLQ), roughness, slope, and soil moisture content in mm (AWCH) were found to be the most important environmental variables in the discrimination of the four ecogeographic groups determined in the cultivated accessions of tepary bean (Figure 3.2A). However, the most important environmental variables in the discrimination of the two eco-geographical groups determined for wild tepary bean accessions were seasonal temperature (TS), soil bulk density (BLDFIE), annual temperature (TAR), soil

moisture content in mm (AWCH), silt content (SLTPPT), sand content (SNDPPT), and soil roughness (Figure 3.2B).

In general, cultivated accessions were more directly bound to the rainfall-related climatic factors, whereas wild accessions were more strongly related to temperature variation and edaphic characteristics. The ecogeographic groups for both wild and cultivated accessions were strongly related to the species' spatial distribution (Figure 3.3). Four groups were identified for cultivated accessions and two for wild accession. Cultivated group 1 (*cult_cluster_1*) was closely linked to the southern areas of Mexico bordering Central American countries such as Guatemala and El Salvador; however, there were accessions in the southwestern part of Mexico, too. Accessions from group 2 (*cult_cluster_2*) were related with the western slopes of the Sierra Madre Mexicana, whereas group 3 (*cult_cluster_3*) of cultivated accessions were found primarily in northern Mexico and southern USA in deserts and dry habitats. *Cult_cluster_4* was most common in Central America. The variation in TAR, roughness, slope, and soil acidity values distinguished the ecogeographic groups (Figure 3.3A, C, D, E). Furthermore, the SNDPPT and SLTPPT were related to soil humidity (AWCH) levels, which influenced the dispersion of cultivated tepary bean accessions. Accessions from *wild_cluster_1* were found on the western and eastern slopes of the Mexican Sierra, with some accessions also found in the desert zones of the Chihuahuan Desert, whereas the *wild_cluster_2* distribution was more associated with desert and arid conditions in the Arizona and Sonora deserts (Figure 3.3F and G). The roughness of the terrain, the PDM, and the slope of the terrain were the key differences between the two established groups of wild accessions, showing distinct survival strategies of the wild accessions in relation to the environmental conditions. The cultivated accessions of tepary bean were represented in the vertex accessions that reflect the most (highest) phenotypic variability for

accessions analyzed in the functional diversity analysis. The white seed types were more represented in the first five principal components (PC1 - PC5); however, this analysis also discriminated brown-orange, cream, speckled, and black seed types, which were the most representative for the cultivated tepary bean (Figure 3.4).

Similarly, the functional diversity analysis distinguished the wild accessions with the largest variation, displaying several seed types and colors ranging from speckled white, gray, brown, and black (Figure 3.5). We used both *acutifolius* and *tenuifolius* botanical varieties to identify vertice accessions. These particular accessions may be related to the functional collection of tepary bean because of their phenotypic variation presented in both groups of cultivated and wild accessions. These accessions likely have the highest phenotypic variation. The use of functional diversity analysis to choose vertices can also set the criteria for phenotypic redundancy within each of the identified ecogeographic groups. For example, the functional evenness index (FEve) relates that accessions of both wild and cultivated plants have values greater than 0.7. This indicates that there are no accessions with unique functional traits that are present within each group (Figure 3.6). The functional specialization (FSpe) and functional originality (FOri) indices show values less than 0.5, indicating that there were no specialized trait accessions in addition to similar functional traits within each eco-geographic group (Figure 3.6).

The functional traits of the various groups of beans from our field experiment differed (Figure 3.7). The functional indices of group 1 of cultivated beans differed from those of group 2 (Figure 3.7A, B, C, and D). Functional diversity (FDiv) (Figure 3.7A), FSpe (Figure 3.7B), functional identity (FId) (Figure 3.7C), and FOri (Figure 3.7D) were varied between groups, too. Group 1 of cultivars did not exhibit dissimilar functional features when compared to group 3 (Figure 3.7E, F, G, and H), and traits between accessions of each group overlapped for Fdiv (Figure

3.7E), Fspe (Figure 3.7F), Fide (Figure 3.7G), and FOri (Figure 3.7H). Group 4 exhibited discriminating functional traits when compared to group 1, particularly between PC2 (Figure 3.7I, J, K, and L). There were some changes in the functional attributes investigated between groups 1 and 2 of wild tepary bean accessions (Figure 3.8). In the FDiv (Figure 3.8A), it was apparent that both groups had varying numbers of accessions, in addition to contrasting accessions in the FSpe (Figure 3.8B). In the case of FId (Figure 3.8C), the separation of the groups could be seen regarding the centroid in the principal component one (PC1). In FOri, it was noted that each group provides accessions that are extremely distant from accessions of the same group (Figure 3.8D), yet there were accessions that present similar functional traits, indicating shared lineages between these accessions.

The effects on functional traits (DF, PPM, PN, LA, W100) were estimated from the ecogeographic groups (Figure 3.9). The only cultivar group with a positive influence on DF was group 4 (cult_cluster_4), which indicates lower DF than the other groups with negative effects (Figure 3.9A). PPM was positively influenced by both the eco-geographical groups of wild and cultivated accessions, as well as by elevation, but negatively by the environmental variables PDM and AWCH (Figure 3.9B). In PN, both wild and cultivated eco-geographical groups showed favorable effects; however, groups 1 and 2 of wild accessions had effects greater than 0.80 (Figure 3.9C). In terms of LA, both wild and cultivated groups have negative effects, with lower values for wild accessions, yet PDM had a beneficial effect on LA (Figure 3.9D). W100 shows that there was significant variation among the groups displaying favorable and negative impacts. Groups 3 and 4 had positive impacts compared to group 2, whereas PDM and AWCH had a negative effect on W100 (Figure 3.9E). In terms of LA, the wild and cultivated groups had a negative impact, the wild group had lower values, and PDM had a positive impact on LA

(Figure 3.9D). For W100, it was evident that there was variation among the groups, both of which exhibit positive and negative effects. The eco-geographical effects for the group of cultivated accessions was positive when compared to group 2's effects, while PDM and AWCH had negative effects on the W100 (Figure 3.9E).

In order to understand the diversity in functional traits in both wild and cultivated accessions, it is important to understand how the ecogeographic groups relate to functional, geographic and genetic components (Figures 3.10, 3.11). The first relationship between environmental conditions in wild accessions and ecogeographic and genetic groupings was between the latent variables LV1-LV2 (Figure 3.10A, B). The wild accessions of the botanical variety *tenuifolius* that were gathered in Arizona were associated with genetic group 2 (GG2), which comprises the wild ecogeographic group 2 (Figure 3.10A). Geographic echo group 1 of wild accessions was typically linked with GG4, GG5, and GG6 genetic clusters. This indicates that these genetic groups are connected to comparable environmental conditions (Figure 3.10B). Similar to how the genetic components (PC1.x - PC2.x) relate to the eco-geographic groups, it was evident how these two groups differ from one another. Nevertheless, some accessions belonged to eco-geographic group 1, which was connected to GG5 and was gathered in Sonora and Sinaloa (Figure 3.10C, D). It can be shown that eco-geographic group 2, which is a part of GG2, was associated with high latitudes in terms of the geographic linkage between eco-geographic and genetic groups.

Figure 3.11 shows the cultivated accessions relatedness given the eco-geographical and genetic clusters. The relationship between eco-geographical groups 2 and 4, and GG1 and GG3 was determined in relation to the variables LV1 and LV2, demonstrating the connection between eco-geography and genetic components (Figure 3.11A, B). It was also observed that GG1 and

GG3 separated similarly to the geographic eco-groups, indicating that traits are likely under selection by the environmental conditions that each group faces, taking into account that GG1 is related to the regions of Sonora and Sinaloa, and GG3 associates with accessions coming from Central America.

In order to detail and relate the ecological clusters with the genetic and functional components, the distances between ecogeographic groups, genetic clusters, and functional and geographic attributes were compared (Figures 3.12, 3.13). The ecological and genetic centroid distances were found to be closely related in the wild accessions. Some clusters in the genetic groups and ecological groups show greater variation, while the genetic distance was maintained (Figure 3.12A). Genetic groups distinguished ecological and functional distances (Figure 3.12). Regarding ecological and genetic distances from the centroid, GG2 exhibited no change in functional traits (Figure 3.12C). Although the ecological distance was dispersed across the evaluated range for the cultivated accessions (Figure 3.13A), and a very similar pattern was detected in the relationship between ecological and geographic distance (Figure 3.13B), ecological and genetic relationship clearly grouped the GG1 and GG3 discriminating groups. It was apparent that there were no clearly defined clusters regarding the relationship between functional and ecological distance, with a random distribution of all the genetic groups (Figure 3.13C). This was demonstrated by a gap between GG1 and GG3 in the link between functional and genetic distance, and F showed the same pattern in the association between ecological and genetic distance (Figure 3.13D). In terms of geographic and functional distance, it was observed that the genetic groups differed, particularly in terms of geographic distance, but that both genetic groups were spread throughout the functional distance range (Figure 3.13E). The genetic

component and the geographic distribution of the cultivated accessions of tepary bean, however, were clearly associated with the genetic and geographic distances (Figure 3.13F).

3.4 Discussion

Some tepary bean accessions contain functional features that allow the species to adapt to high temperatures and drought (Mwale et al., 2020; Muñoz et al., 2006), but the ecological, genetic, and functional trait linkages have not yet been investigated, until this study. The tepary bean accessions used in this study come from Arizona (USA) to Guanacaste (Costa Rica). However, the wild populations are primarily found in Arizona, Sonora, and Chihuahua associated with xerophytic environments (Nabhan, 1985; Pratt y Nabhan, 1988). Different phenotypic traits in cultivated accessions are closely related to their geographical distribution, particularly those associated with seed shape and color (Nabhan, 1985). In our study, we suggest two ecological clusters for wild populations and four for cultivated populations (Figure 3.3). Our findings reveal that each ecological group faces contrasting environmental conditions (Figure 3.3). Two ecological clusters of wild accessions are connected with comparable environmental conditions. Yet, the slope, terrain roughness, and precipitation during the driest month explain the ecological differences between each of these groups (Figure 3.2). In wild populations of *P. acutifolius*, there are two botanical varieties, *acutifolius* and *tenuifolius*, with distinct morphological characteristics, particularly in leaflet morphology (Nabhan, 1978). Interestingly, cluster two primarily contains *tenuifolius* accessions in contrast to cluster one with mostly *acutifolius* accessions (Figure 3.3). Both varieties have preferences for ecological conditions of their habitat distribution that likely can be attributed to morphological differences in leaflet shape (Nabhan 1978). The *tenuifolius* leaf is a xeromorphic trait (lateral leaflets narrowly lanceolate) that can

allow the plant to reduce undue transpiration. The wider leaves and longer internodes of *acutifolius* give it more tropical climber look, in addition to the fact that *acutifolius* is adapted to conditions with reduced light intensities growing under or climbing on plants and shrubs, colonizing especially floodplains and stream beds in somewhat drier places than *tenuifolius*. *Tenuifolius* variety is associated with places with more sun exposure, rougher rocky terrain, and more pronounced slopes (Nabhan, 1978; Freytag & Debouck 2002). This habitat preference is also supported by our research, as one of the most important ecological variables in the differentiation of ecological clusters was the degree of terrain roughness (Figure 3.3).

The cultivated accessions are divided into four ecogeographic clusters with differing ecological conditions (Figure 3.3). These clusters are distinguished by soil moisture conditions (AWCH), silt and clay particle size, and soil acidity. Cluster three and four, in particular, present contrasting ecological and geographic conditions (Figure 3.3). For example, the cultivated accessions of group three are related to desertic ecological conditions, particularly in Arizona. The group four contains accessions collected in Central America, which are associated with more acid soils and a higher clay / silt ratio, resulting in likely higher natural soil fertility (Figure 3.3). However, the tendency of man to control the environment during cultivation, especially in local hydrostatic and geomorphological modifications (irrigation), and the manipulation of the populations of cultivated plants in these modified environments, influence the process of domestication (Nabhan, 1978). Even though the soil acidity or alkalinity are difficult to modify (Schaetzl and Thompson, 2015), the tillage practices and soil modifications carried out by the indigenous communities, especially those of southern USA (Nabhan, 1978), determine the acclimatization of the variety of teparies cultivated in diverse ecological conditions.

The relationship between the ecological clusters and their phenotypic differences can be understood by evaluating the functional traits via functional diversity in both wild and domesticated accessions (Figure 3.5 and 3.6). The phenotypic variability of the diversity panel is clearly visible in cultivated accessions (Figure 3.4). The vertex accessions, which exhibit the greatest phenotypic variation among the cultivated accessions, indicated variation in seed coat color that are correlated with the evaluated functional traits. The functional diversity analysis demonstrates how ecological clusters are clearly distinguished, and how the seed coat color was modified by domestication. The prioritized vertices accessions (Figure 3.4) show that a greater percentage of accessions in the CIAT germplasm collection exhibit white seed color.

The geographic regions of Arizona, Durango and Sonora are related to the white seed coat accessions represented in ecological groups two and three (Table 3.1). These geographical locations are conditioned by xerophytic habitats, which most likely had an impact on the evolution of functional traits of these two ecogeographic clusters. These are connected to geographic areas ranging from Chiapas (Mexico) to Zacapa (Guatemala), in contrast to the cultivated accessions of cluster four (Table 3.1). The cluster four has seed coat colors varying from cream to cream mottled; these accessions have distinct functional traits profiles when compared to groups two and three (Figure 3.4). As discussed in this study, we were able to distinguish and identify the most contrasting accessions as well as the variation in seed coat color by using the functional traits of *P. acutifolius* in addition to the evaluation of functional diversity. This suggests the impact of unconscious selection processes during domestication (Darwin, 1876) by the indigenous communities of Mesoamerican (Kaplan, 1977 ;Nabhan, 1978; Bornowski et al. 2023). To *P. acutifolius* wild accessions are classified as two botanical varieties: tenuifolius and acutifolius (Freytag and Debouck, 2011). It has been reported that these botanical

varieties differ, particularly in their ecological niches (Nabhan, 1978), but their ecological clusters are not really contrasting, with only alterations in terrain roughness (Figure 3.3). This is confirmed by the functional diversity analysis (Figure 3.5). We can see similarities in the functional traits (Figure 3.5A), indicating sympatry between the two botanical varieties, a proof they prefer ecological niches with only subtle differences (Nabhan, 1978; Debouck et al., 1979; Debouck et al., 2023). Despite the ecological similarities between the identified ecogeographic clusters, the *tenuifolius* and *acutifolius* varieties exhibit distinct functional traits. For instance, the ecological groups are distinguished in the functional components PC1 clustering the *acutifolius* accessions in the positive scores and the *tenuifolius* accessions in the negative values. However, this difference is not seen in Figure 3.5B, C, and D, where the ecological overlap and the close genetic proximity of the populations led to intermediate genotypes between these two varieties (Nabhan, 1978; Debouck et al., 1979; Bornowski et al 2023). On the other hand, functional traits in wild accessions strongly differentiate between the two ecological clusters (Figure 3.8). The variation in terms of botanical traits clearly distinguishes the two clusters. These variations are geographically and sympatrically distributed under xerophytic environments (Figure 3.3). The wild ecological cluster one is represented to a greater extent by accessions collected in Mexico; however, the botanical variety *tenuifolius* exhibits functional traits completely different from the variety *acutifolius*, particularly in the variation of leaf shape, making it more related to the species *P. montanus*, which has a similar ecological niche (Freytag and Debouck, 2002; Muñoz et al., 2006). However, ecological clusters demonstrate a direct relationship in the phenotypic variability within each cluster, due to the evolutionary process that conditioned their functional identity (Figure 3.8C). It is worth noting that our data show a number of vertically-positioned accessions (Figure 3.5) that are separated from the centroid of

each ecological cluster (Figure 3.8B). This distancing indicates that within each botanical variety, there are accessions with high phenotypic variation compared to the cluster of accessions evaluated, despite having a restricted geographic distribution (Figure 3.7), exhibiting high functional uniformity (Figure 3.6; 1W), and being cleistogamous (Freytag and Debouck, 2002).

In addition, The centroid distance for each related group is derived after considering the latent variables and their relationship to the genetic and ecogeographic groups. The centroid distances generally relate the admixed genetic group's accessions. These accessions, particularly G40177, have been reported to exhibit intermediate functional traits between cultivated and wild populations (Chacon et al., 2022; Muñoz et al., 2006). *P. acutifolius* is a species that is cleistogamous in both cultivated and wild genepools, for this reason the probability of crossing is low. However this mechanism relates selection and migration processes to promote genetic exchange and new adaptations during the colonization of geographic areas with variable environmental conditions. In Phaseolus this type of regression has been widely reported in the different domesticated species (Chacon et al., 2022. (e.g., accessions derived from G40177, which Gary Nabhan collected in Pima County in 1974). Nabhan and Felger (1978) carried out research on the tepary bean in Arizona and found that the bean has both wild and domesticated genotypes, and they also recorded variations in the color of the seed's testa. These three scenarios were only present in Papago Indian communities, who have used the tepary bean — both wild and domesticated — in their diets and as a component of their cosmogony (Teiwes and Nabhan, 1983). It should be emphasized that these genetically admixed wild accessions are closely connected to the Sonora-Sinaloa and Durango groups, which are associated with the Sierra Madre Occidental (Muñoz et al., 2006); yet, they also exhibit genetic and functional variation in comparison to these groups.

Our findings indicate that the selective processes that influence domestication and subsequent spread to other regions of Mesoamerica and Central America occurred in this secluded area. The USA accessions collected at the Papago Indian Reserve, which exhibit distinct seed color characteristics ranging from yellow to brown and white, are the representatives of ecological group three (Table S1: Passport data). Previous research (Teiwes and Nabhan, 1983) has established that the Bavavavi O'odham, one of the indigenous communities of Arizona, valued the tepary bean because it was an indispensable component of their diet and because they associated it with spiritual and ancestral visions of the stars and the universe, particularly the white seed representing the color of the stars. Cultural differences in cosmogony and seed type preference (Debouck collection report, 1986) between the indigenous communities of Aridoamerica and Mesoamerica could have been determinants of the variation in functional traits between ecological groups three and four (Figure 3.7).

3.5 Conclusions

The tepary bean is a species of importance since millenary times being the staple food of indigenous communities of arid America and later its dispersion to Mesoamerica during the unconscious selection that directed the variation of traits among genetic groups as shown by our results. Our study provides an approximation of the tepary beans adaptive evolution processes, relating both the genetic structure, ecological groups, and the variation of functional features, as well as identifying the ecological variables that characterize the determined ecological groups. We identified the accessions that exhibit the largest variation in functional traits in both wild and domesticated forms. This methodological model gives genebank scientists a starting point for understanding phenotypic variation and emphasizes the importance of germplasm characterization in comprehending the variation of functional phenomes in germplasm

collections. We suggest how to use data from various data sources associated with germplasm collection to demonstrate the importance of documentation in genebanks, and how what was developed in this study can serve as a starting point for integrating data mining techniques to project the integration of documentation and germplasm characterization, which will contribute to germplasm utilization and distribution.

3.6 Acknowledgements

I would like to appreciate the Bioversity CIAT alliance's germplasm bank for their assistance during the research process and the associated data collecting. I am grateful to the Universidad de la Amazonia for their assistance in phenotyping the collection.

3.7 References

- Amatulli, G., Domisch, S., Tuanmu, M. N., Parmentier, B., Ranipeta, A., Malczyk, J., & Jetz, W. (2018). A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. *Scientific data*, 5(1), 1-15.
- Bornowski, N., Hart, J. P., Palacios, A. V., Ogg, B., Brick, M. A., Hamilton, J. P., ... & Porch, T. (2023). Genetic variation in a tepary bean (*Phaseolus acutifolius* A. Gray) diversity panel reveals loci associated with biotic stress resistance. *The Plant Genome*, e20363.
- Calliope, S. R., Lobo, M. O., & Sammán, N. C. (2018). Biodiversity of Andean potatoes: Morphological, nutritional and functional characterization. *Food chemistry*, 238, 42-50.
- Chacón-Sánchez, M. I. (2018). The domestication syndrome in *Phaseolus* crop plants: A review of two key domestication traits. *Origin and Evolution of Biodiversity*, 37-59.
- Debouck, D. G. 2000. Genetic resources of *Phaseolus* beans: patterns in time, space, and people. Segundo seminario de la Judia. Actas de la asociación Española de leguminosas.
- Dwivedi, S. L., Scheben, A., Edwards, D., Spillane, C., & Ortiz, R. (2017). Assessing and exploiting functional diversity in germplasm pools to enhance abiotic stress adaptation and yield in cereals and food legumes. *Frontiers in plant science*, 8, 1461.

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International journal of climatology*, 37(12), 4302-4315.

Freytag, George F.; Debouck, Daniel G. 2002. Taxonomy, distribution, and ecology of the genus phaseolus (Leguminosae-papilionoideae) in North America, Mexico and Central America. Botanical Research Institute of Texas (BRIT), Forth Worth, TX, USA. 298 p. (Sida, botanical miscellany no. 23)

Hall, A. E. (2012). Phenotyping cowpeas for adaptation to drought. *Frontiers in physiology*, 3, 155.

Hall, C. A., Stanford, J. A., & Hauer, F. R. (1992). The distribution and abundance of organisms as a consequence of energy balances along multiple environmental gradients. *Oikos*, 377-390.

Henao-Rojas, J. C., Rosero-Alpala, M. G., Ortiz-Muñoz, C., Velásquez-Arroyo, C. E., Leon-Rueda, W. A., & Ramírez-Gil, J. G. (2021). Machine learning applications and optimization of clustering methods improve the selection of descriptors in blackberry germplasm banks. *Plants*, 10(2), 247.

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., ... & Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLoS one*, 12(2), e0169748.

Kamoshita, A., Babu, R. C., Boopathi, N. M., & Fukai, S. (2008). Phenotypic and genotypic analysis of drought-resistance traits for development of rice cultivars adapted to rainfed environments. *Field crops research*, 109(1-3), 1-23.

Mammola, S., Carmona, C. P., Guillerme, T., & Cardoso, P. (2021). Concepts and applications in functional diversity. *Functional Ecology*, 35(9), 1869-1885.

Mardeh, A. S. S., Ahmadi, A., Poustini, K., & Mohammadi, V. (2006). Evaluation of drought resistance indices under various environmental conditions. *Field Crops Research*, 98(2-3), 222-229.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

Moghaddam, S. M., Oladzad, A., Koh, C., Ramsay, L., Hart, J. P., Mamidi, S., ... & McClean, P. E. (2021). The tepary bean genome provides insight into evolution and domestication under heat stress. *Nature communications*, 12(1), 2638.

Morrone, J. J. (2019). Regionalización biogeográfica y evolución biótica de México: encrucijada de la biodiversidad del Nuevo Mundo. *Revista mexicana de biodiversidad*, 90.

Nawar, S., Mohamed, E. S., Essam-Eldeen Sayed, S., Mohamed, W. S., Rebouh, N. Y., & Hammam, A. A. (2023). Estimation of key potentially toxic elements in arid agricultural soils using Vis-NIR spectroscopy with variable selection and PLSR algorithms. *Frontiers in Environmental Science*, 11, 1222871.

Nong, R. A., Gerrano, A. S., & Gwata, E. T. (2023). A preliminary evaluation of phenotypic traits of tepary bean (*Phaseolus acutifolius* A. Gray). *Journal of Underutilised Crops Research*, 2(1), 6.

Parra-Quijano, M., Iriondo, J. M., & Torres, E. (2012). Applications of ecogeography and geographic information systems in conservation and utilization of plant genetic resources. *Spanish journal of agricultural research*, 10(2), 419-429

Poggi, G. M., Corneti, S., Aloisi, I., & Ventura, F. (2023). Environment-oriented selection criteria to overcome controversies in breeding for drought resistance in wheat. *Journal of Plant Physiology*, 280, 153895.

Porch, T., & Beaver, J. (2012). Development of improved tepary bean (*Phaseolus acutifolius* A. Gray) germplasm. *Ann. Rep. Bean Improv. Coop*, 55, 123-124.

Ringnér, M. (2008). What is principal component analysis?. *Nature biotechnology*, 26(3), 303-304. <https://doi.org/10.1038/nbt0308-303>

Schaetzl, R. J., & Thompson, M. L. (2015). Soils. Cambridge university press.

Suárez, J. C., Contreras, A. T., Anzola, J. A., Vanegas, J. I., & Rao, I. M. (2021). Physiological Characteristics of Cultivated Tepary Bean (*Phaseolus acutifolius* A. Gray) and Its Wild Relatives Grown at High Temperature and Acid Soil Stress Conditions in the Amazon Region of Colombia. *Plants*, 11(1), 116.

Teiwes, H., & Nabhan, G. P. (1983). Tepary beans, O'odham farmers and desert fields.

Title, P. O., & Bemmels, J. B. (2018). ENVIREM: an expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling. *Ecography*, 41(2), 291-307.

Tschannen, M., Bachem, O., & Lucic, M. (2018). Recent advances in autoencoder-based representation learning. arXiv preprint arXiv:1812.05069.

Tschannen, M., Bachem, O., & Lucic, M. (2018). Recent advances in autoencoder-based representation learning. arXiv preprint arXiv:1812.05069.

Tuanmu, M. N., & Jetz, W. (2014). A global 1-km consensus land-cover product for biodiversity and ecosystem modelling. *Global Ecology and Biogeography*, 23(9), 1031-1045.

Tuanmu, M. N., & Jetz, W. (2015). A global, remote sensing-based characterization of terrestrial habitat heterogeneity for biodiversity and ecosystem modelling. *Global Ecology and Biogeography*, 24(11), 1329-1339.

Wilson, A. M., & Jetz, W. (2016). Remotely sensed high-resolution global cloud dynamics for predicting ecosystem and biodiversity distributions. *PLoS biology*, 14(3), e1002415.

Zavafer, A., Bates, H., Mancilla, C., & Ralph, P. J. (2023). Phenomics: conceptualization and importance for plant physiology. *Trends in Plant Science*.

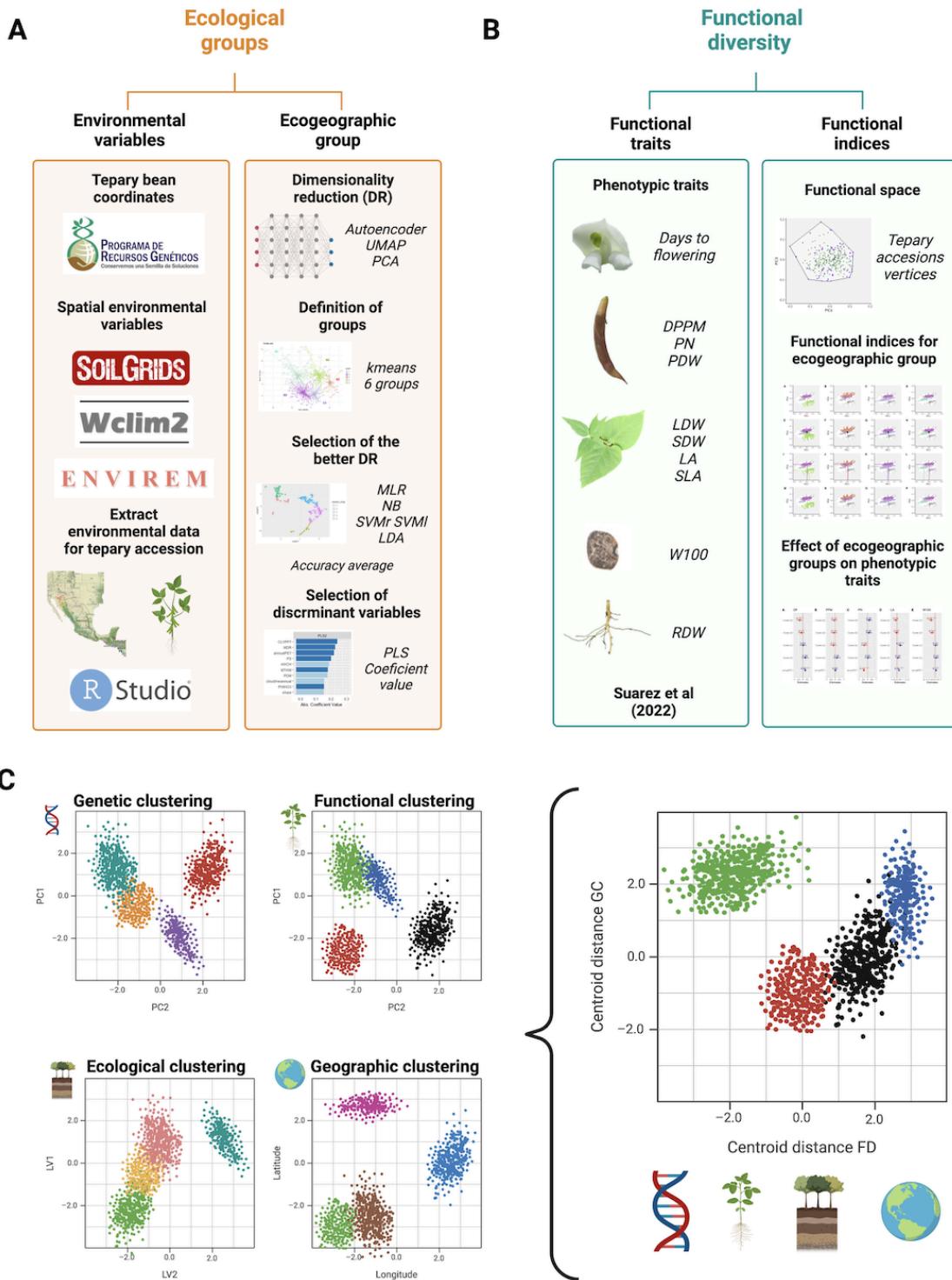
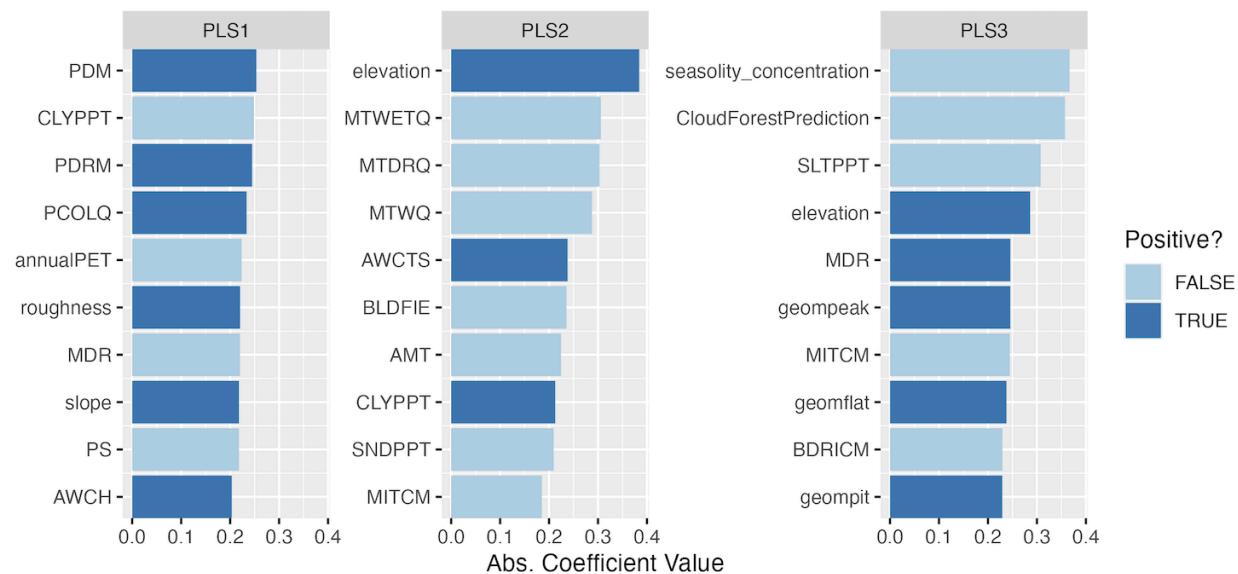


Figure 3.1 The ecogeographic representativeness and functional phenome of the tepary bean CIAT's germplasm collection were investigated using this methodological pipeline. Ecogeographic groups were determined by (A) extracting ecogeographic characteristics from *P.*

acutifolius accessions, and (B) employing ecogeographic variables and machine learning methods to determine ecogeographic groups. In terms of functional diversity, (III) phenotypic variables were utilized for the functional diversity analysis using (IV) functional diversity indices: Vertices accessions exhibiting the greatest variation in the functional phenome were determined and depicted at each ecogeographic cluster. (C) The distance from the centroid of genetic, functional, ecological, and geographic components.

A Partial Least Squares



B Partial Least Squares

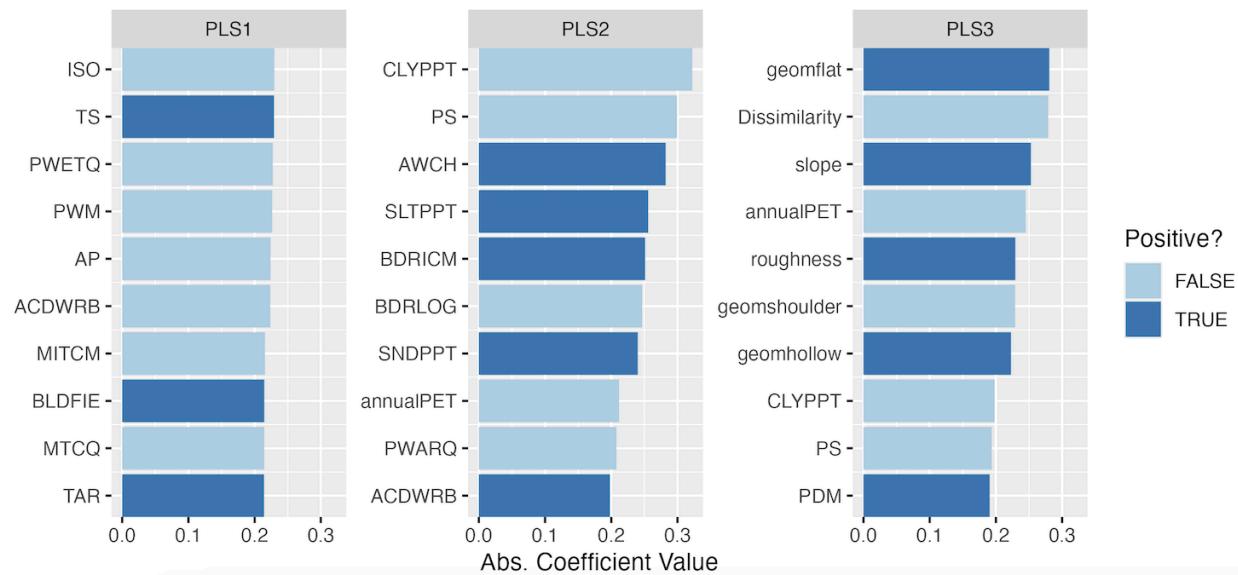


Figure 3.2. Selection of key ecogeographic variables for distinguishing ecogeographic groupings of *P. acutifolius* (A) cultivated and (B) wild genepools. The important variables in the discrimination of ecogeographic groups are represented in dark blue, whereas the false positives are shown in light blue.

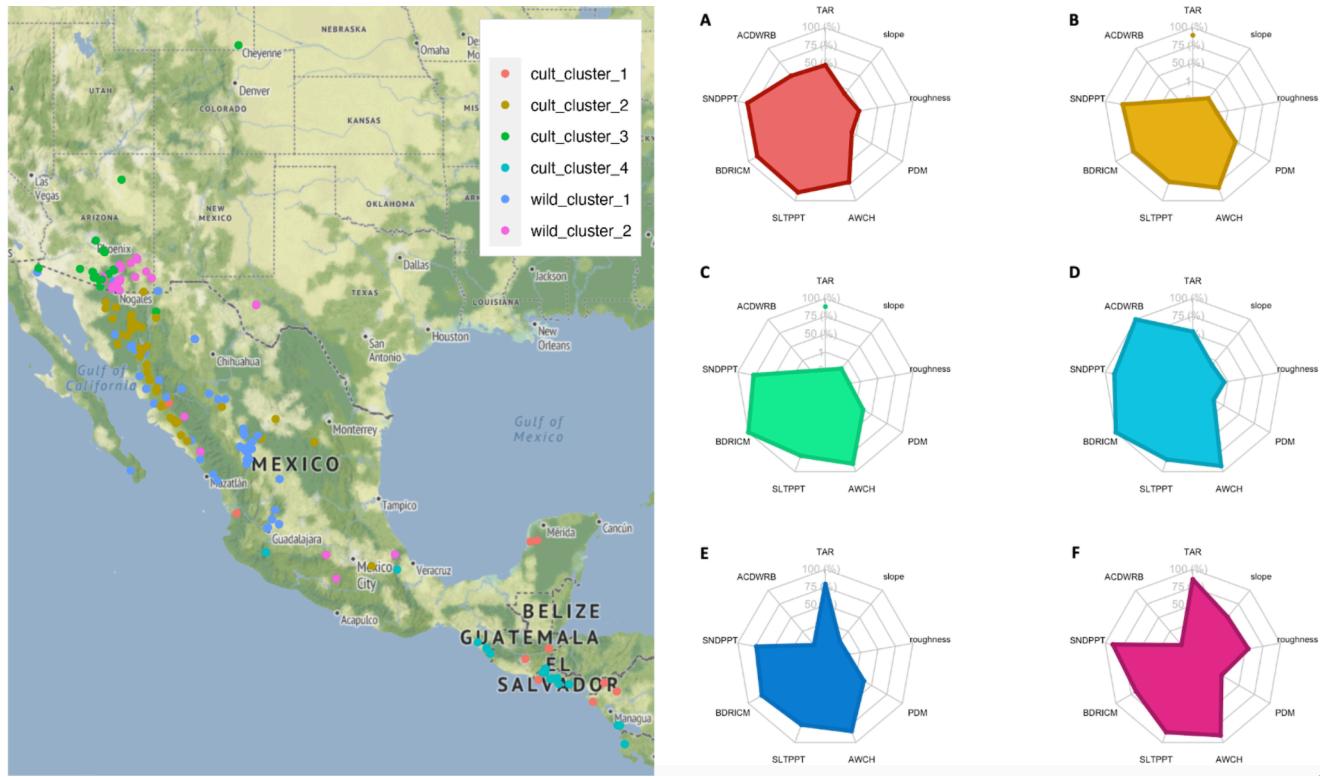


Figure 3.3 Ecogeographic groups of *P. acutifolius* accessions, as well as radar plots of important variables. The spatial distribution of each of the groups of cultivated and wild accessions is observed. The radar plots reflect the variation and importance of environmental conditions for each group: (A) cultivated group 1; (B) cultivated group 2; (C) cultivated group 3; (D) cultivated group 4; (E) wild group 1, and (F) wild group 2. The colored points are explained in the legend on the left-side within panel A.

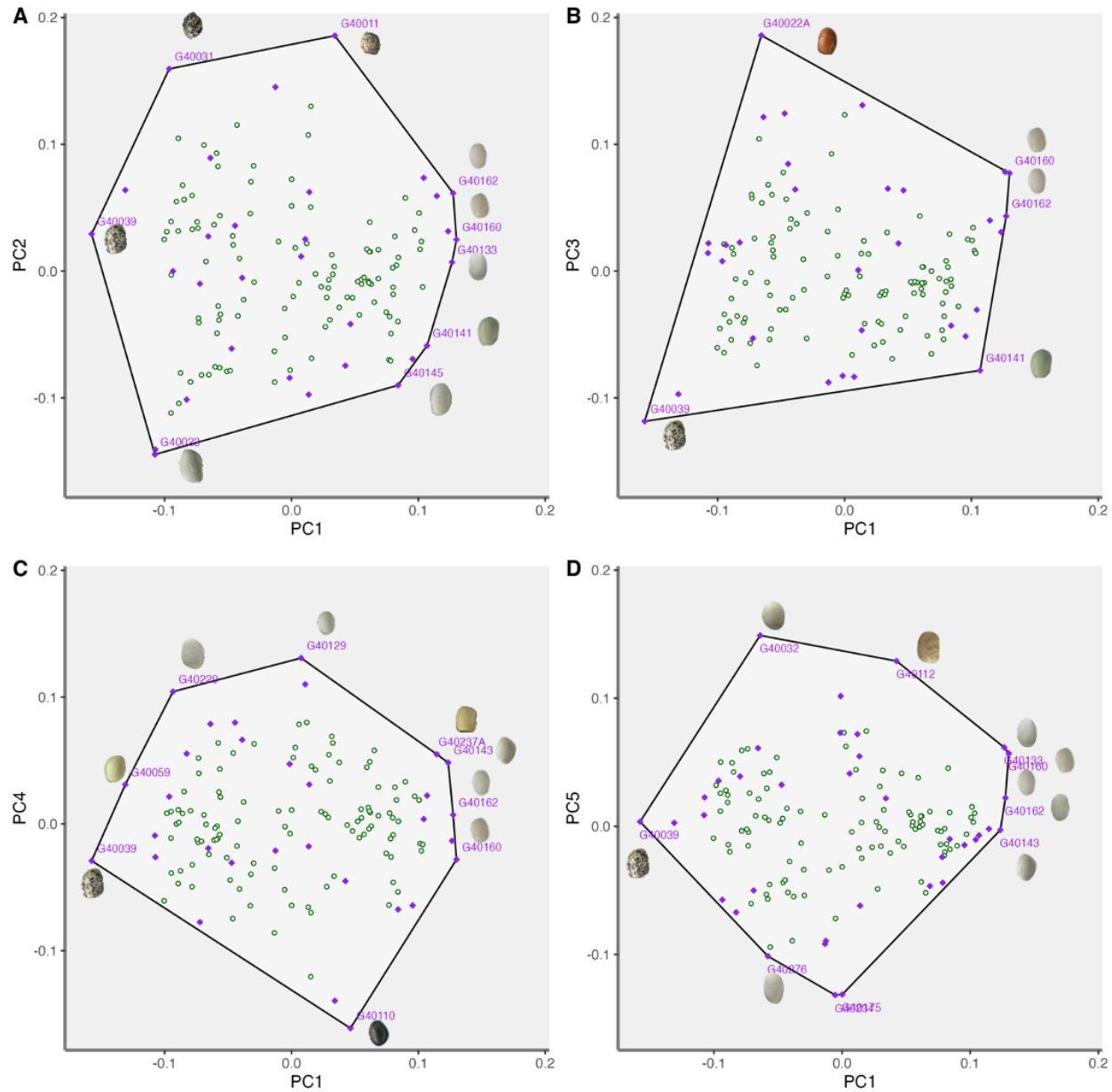


Figure 3.4. The vertices positions within the cultivated *P. acutifolius* accessions seen in all major components PC1 to PC5. Each purple dot represents a vertice accession in one of the major components. Green dots represent non-vertice accessions.

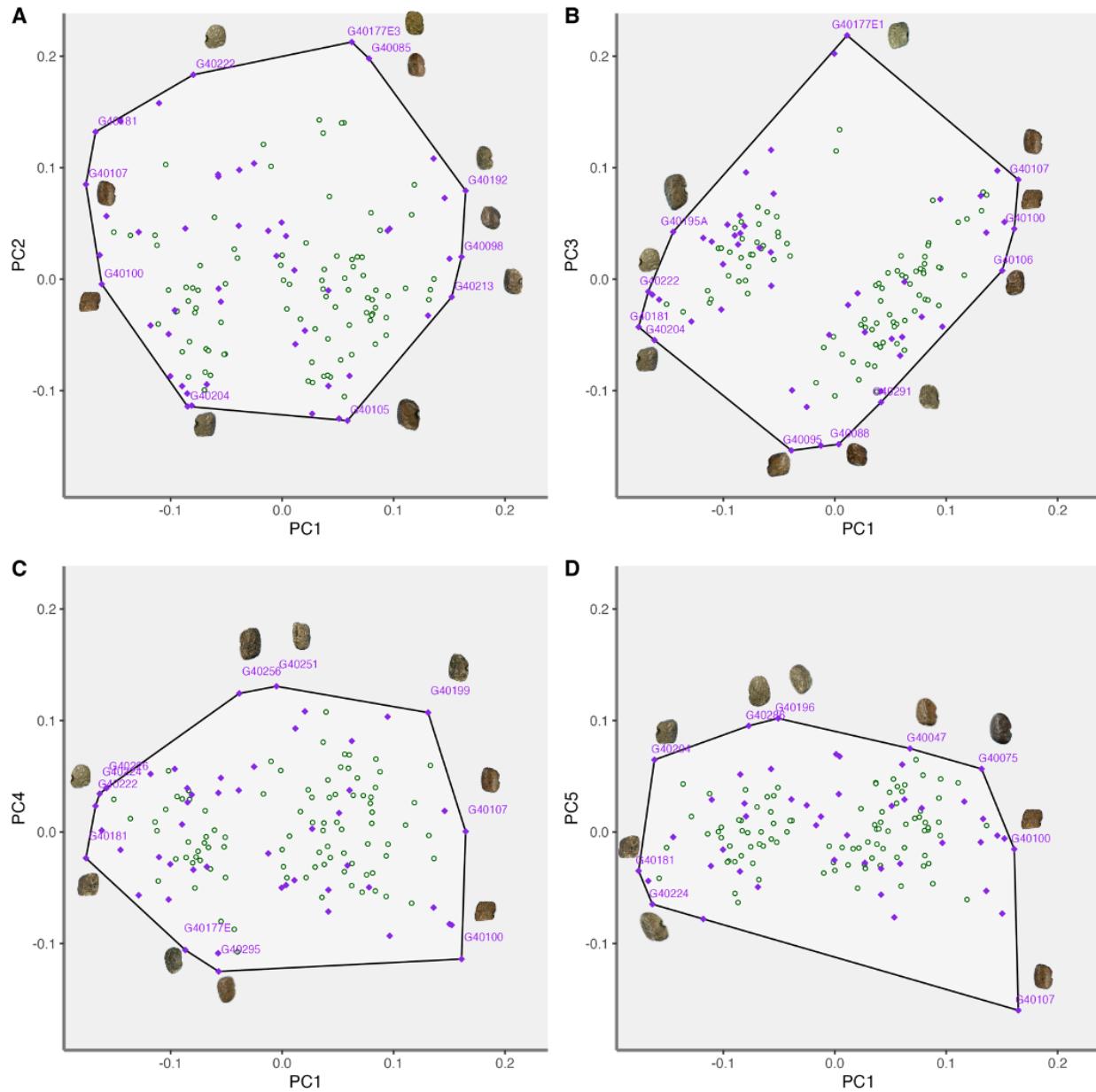


Figure 3.5. Wild *P. acutifolius* vertex accessions in the major components PC1 to PC5. Each purple dot represents a vertex accession in one of the major components. Green dots represent non-vertex accessions.

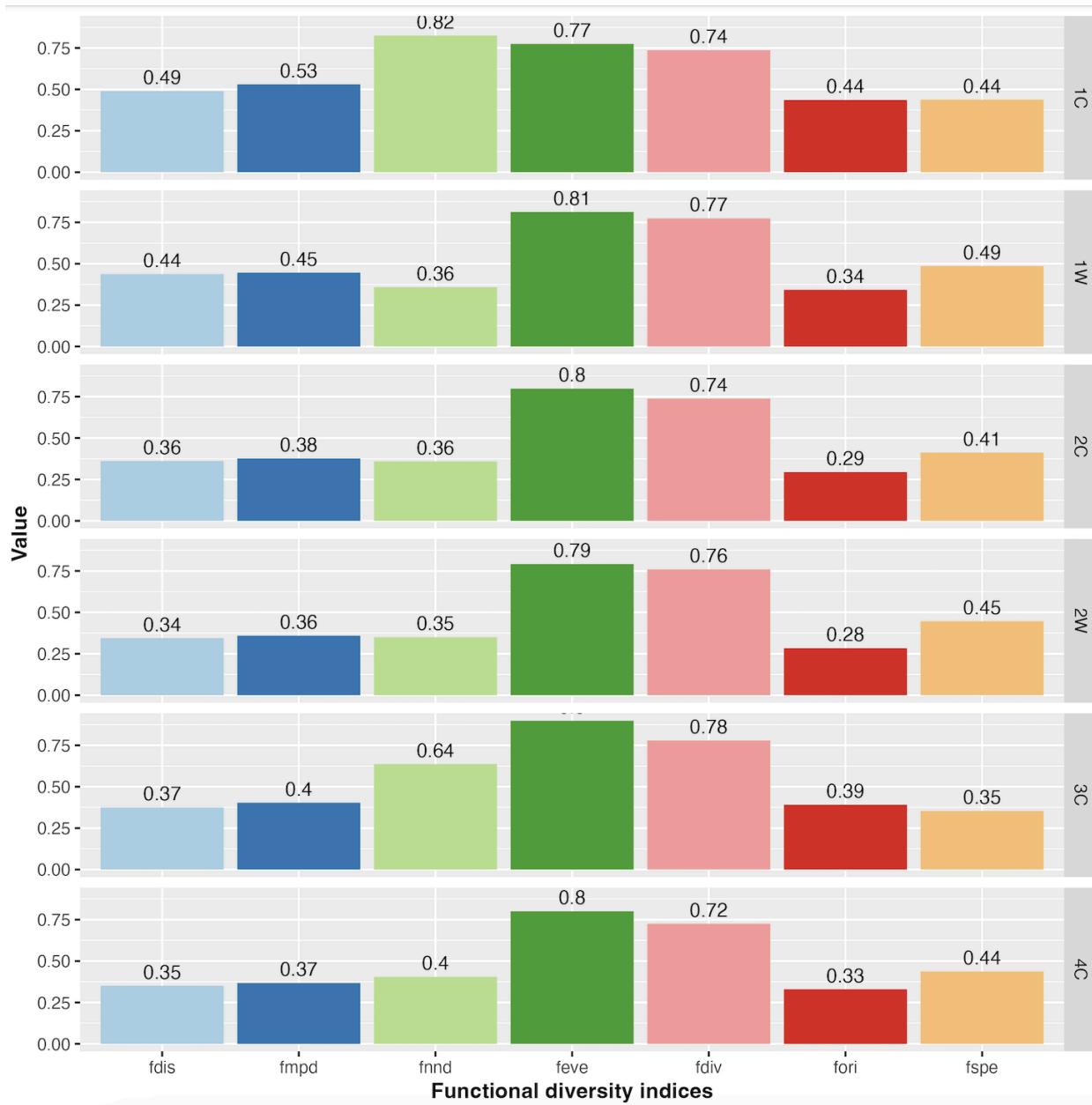


Figure 3.6. Functional diversity indices for cultivated (1C, 2C, 3C y 4C) and wild (1W and 2W) *P. acutifolius* accessions in each ecogeographic group. FDis, FMpd, FNnd, FEve, FDiv, FOri, and FSpe functional diversity indices are reported.

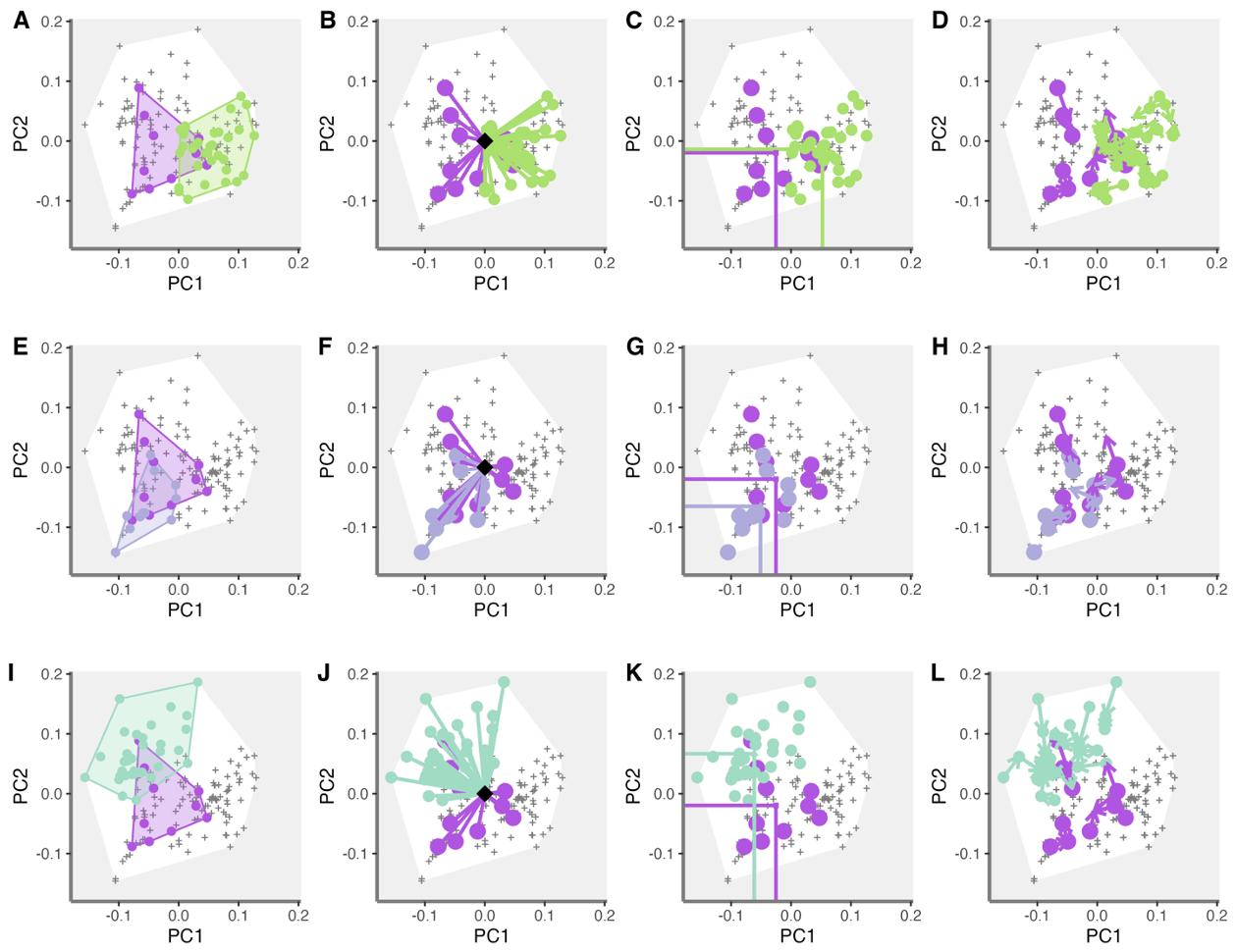


Figure 3.7. Functional diversity indices of cultivated *P. acutifolius* accessions per ecogeographic group. The functional richness index (FRic – A, E, I), functional specialization index (FSpec – B, F, J), functional identity index (Fide – C, G, K), and functional originality index (FOri – D, H, L) are depicted. Colors represent the geographical eco-groups.

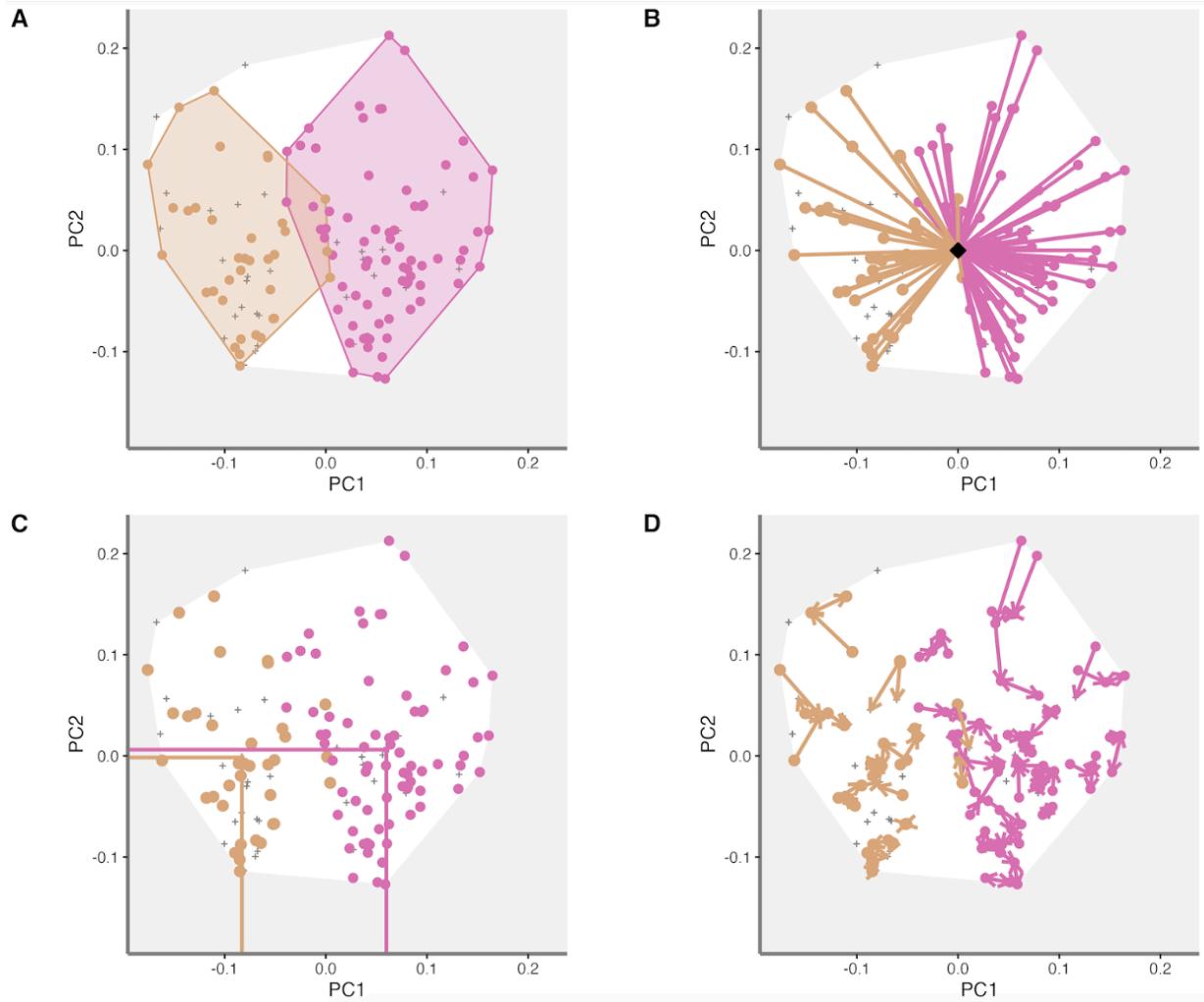


Figure 3.8. Functional diversity indices of wild *P. acutifolius* accessions per ecogeographic group. The functional richness index (A - FRic), functional specialization index (B-FSpec), functional identity index (C-FIde), and functional originality index (D-FOri) are shown. Colors represent the geographical eco-groups.

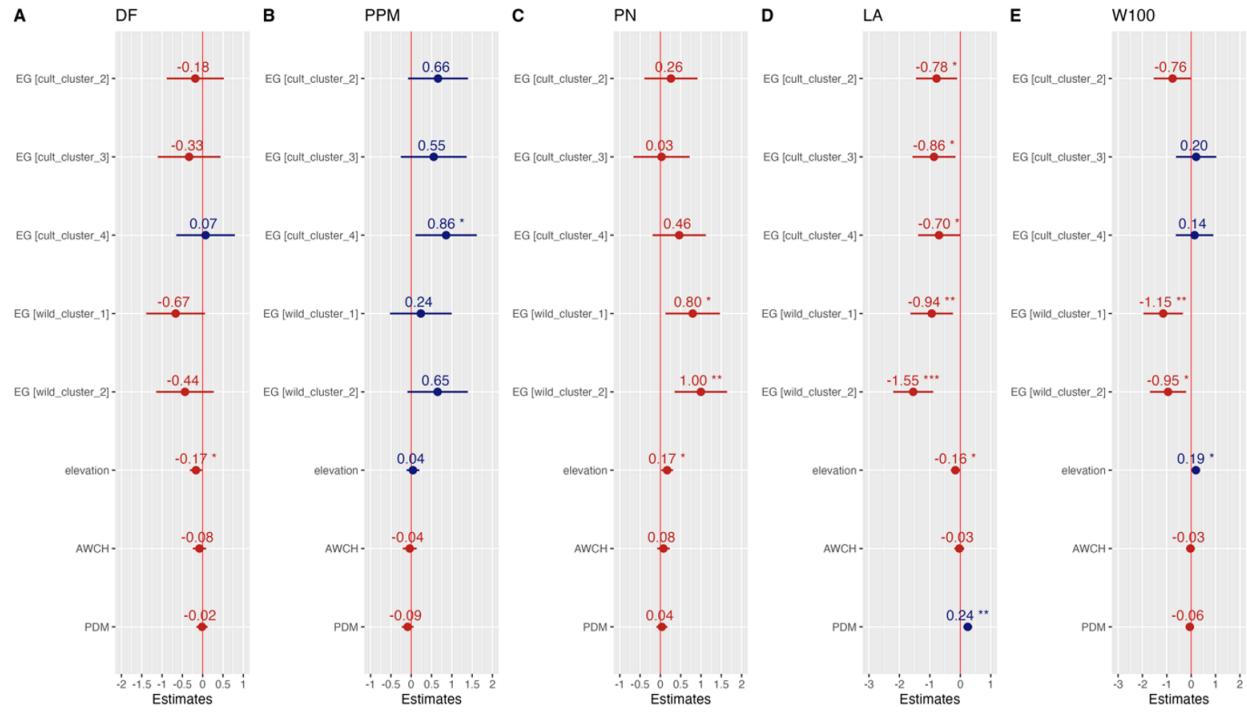


Figure 3.9. Effect of ecogeographic variables and ecogeographic groups on phenotypic traits observed in our field experiment of *P. acutifolius* accessions. A) The effect of days to flowering (DF), B) days to physiological maturity (PPM), C) number of pods (PN), D) leaf area (LA) and E) 100-seed weight (W100) is shown.

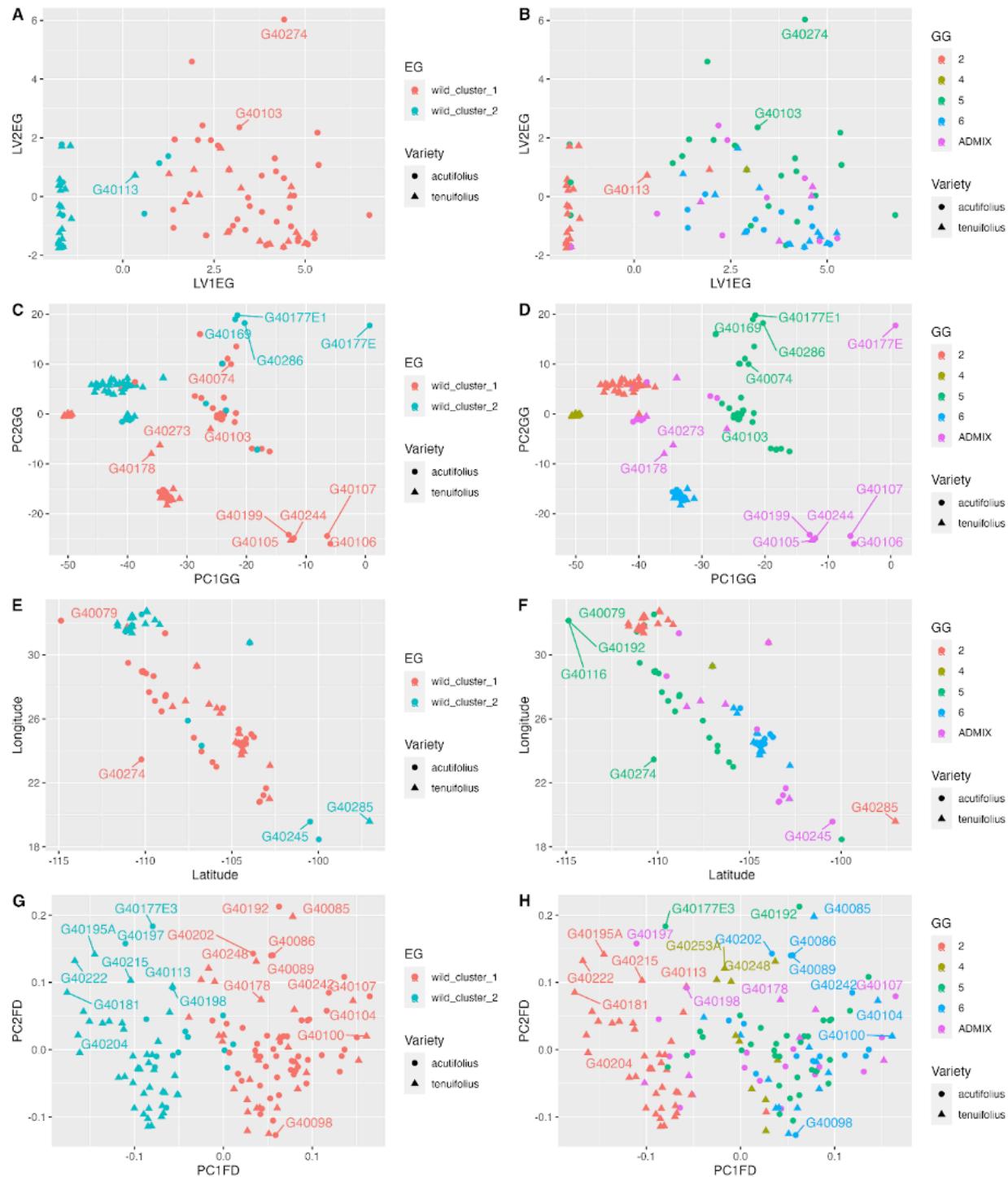


Figure 3.10. Ecogeographic group and genetic clusters of the wild tepary bean collection. (A - B) Latent variables of ecogeographic groups; (C - D) Principal components of genetic groups; (E - F) Relationship of latitude and longitude; (G - H) Principal components of functional diversity.

GG2-SAwt: Southeast Arizona wild var. *tenuifolius*; GG4-CHwt: Chihuahua wild var. *tenuifolius*; GG5-SSwa: Sonora-Sinaloa wild var. *acutifolius*; GG6-DUw: Durango wild and Admixed.

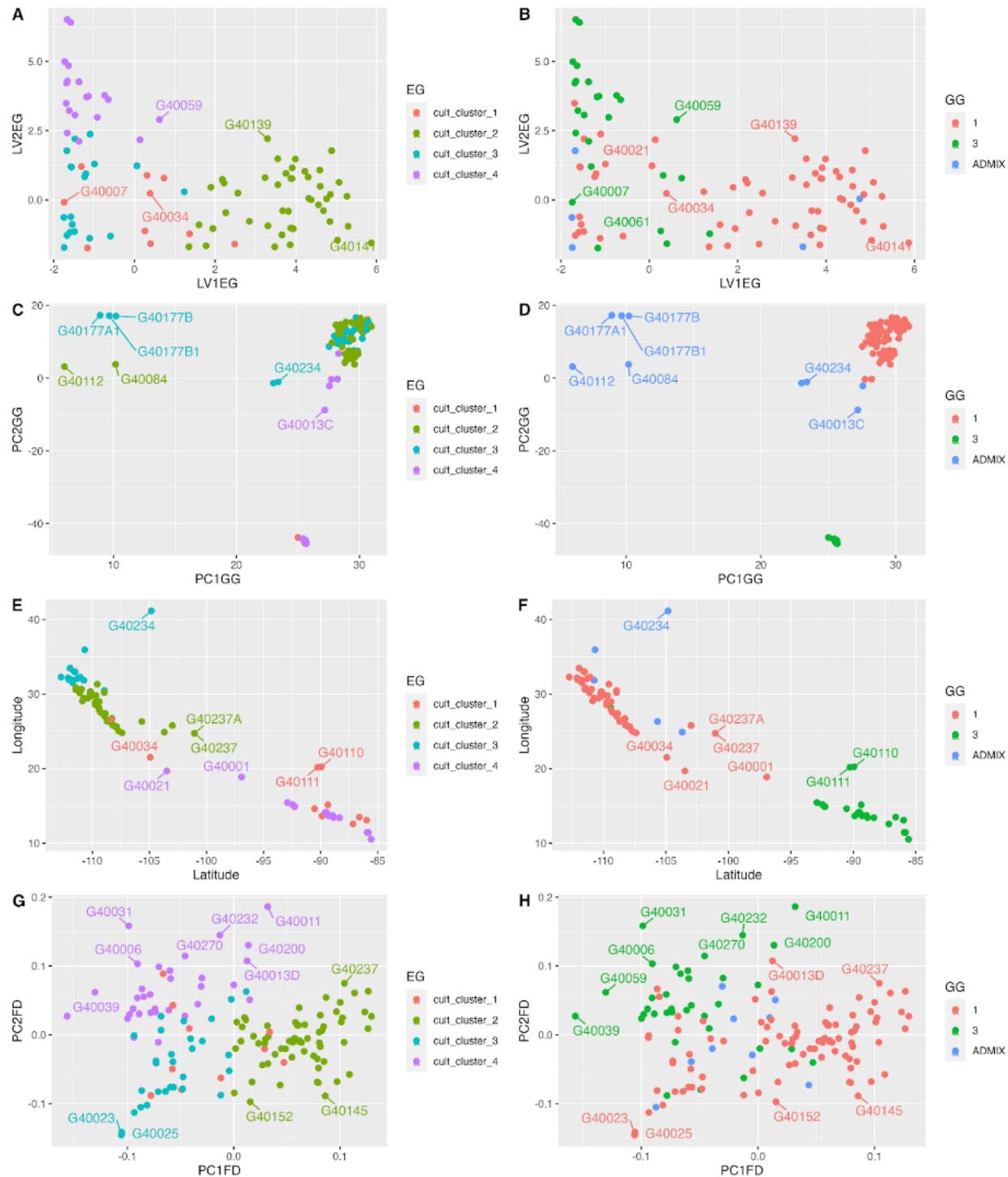


Figure 3.11. Ecogeographic groups and genetic clusters of the cultivated tepary bean collection.

(A - B) Latent variables of ecogeographic groups, (C - D) Principal components of genetic groups; (E - F) Relationship of latitude and longitude, and (G - H) Principal components of

functional diversity. Cultivated in GG1-SSc: Sonora-Sinaloa; GG3-CAc: Central America cultivated; GG4-CHwt and admixed.

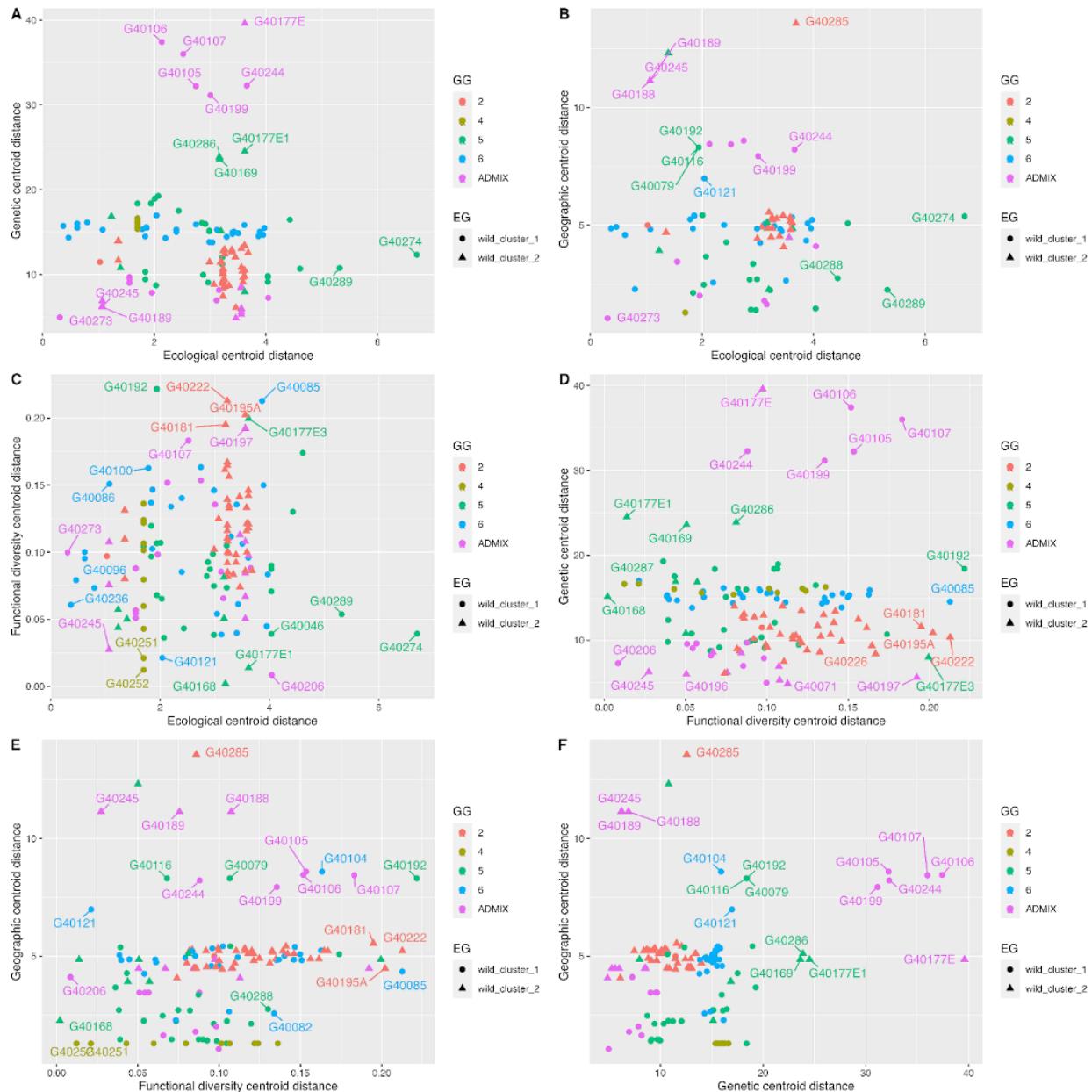


Figure 3.12. Relationship of centroid distances to genetic and eco-geographic groups of wild accessions of tepary bean. (A) ecological centroids and genetic centroids, (B) ecological centroids and geographic centroids, (C) ecological centroids and functional diversity centroids,

(D) functional diversity centroids and genetic centroids, (E) functional diversity centroids and geographic centroids, and (F) genetic and geographic centroids.

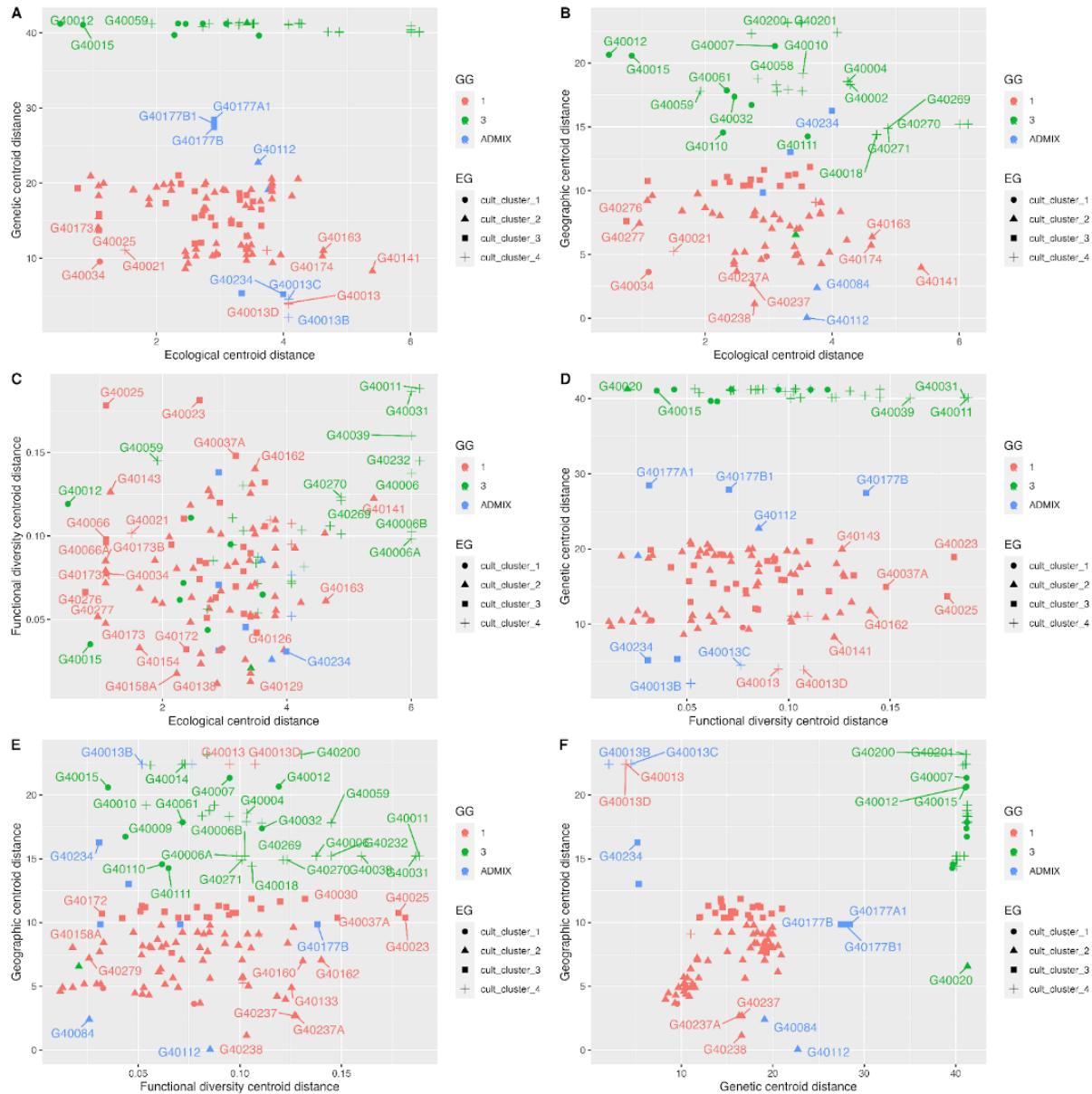


Figure 3.13. Relationship of centroid distances to genetic and eco-geographic groups of cultivated tepary bean accessions. (A) ecological centroids and genetic centroids. (B) ecological centroids and geographic centroids. (C) ecological centroids and functional diversity centroids. (D) functional diversity centroids and genetic centroids, (E) functional diversity centroids and geographic centroids. and (F) genetic and geographic centroids.

Table 3.1. Passport data from the vertices accessions of the functional spaces PC1 through PC5. The accession number, state and municipality of collection, biological status, botanical variety, doi, and weight of 100 seeds are all included.

Accession	Country	Estate	Municipalit y	Biologica l state	Variety	doi	W100 (g)
G40222	USA	Arizona	Pima County	Wild	<i>tenuifoliu s</i>	10.18730/JXJE V	2.5
G40204	USA	Arizona	Santa Cruz County	Wild	<i>tenuifoliu s</i>	10.18730/JXHM 1	2
G40219	USA	Arizona	Pima County	Wild	<i>tenuifoliu s</i>	10.18730/JXH W9	2.4
G40190	USA	New Mexico	Hidalgo County	Wild	<i>acutifoliu s</i>	10.18730/JXFY N	3.2
G40093	Mexico	Durango	Durango	Wild	<i>tenuifoliu s</i>	10.18730/JP865	3.8
G40160	Mexico	Sonora	Bacanora	Cultivate d	<i>acutifoliu s</i>	10.18730/JSHZ Q	16.4
G40011	Mexico	Chiapas	Tapachula	Cultivate d	<i>acutifoliu s</i>	10.18730/JGC0 E	16.8
G40032	Guatemala	Zacapa	Gualan	Cultivate d	<i>acutifoliu s</i>	10.18730/JFATS	15.2
G40177C	USA	Arizona	Pima County	Weedy	<i>acutifoliu s</i>	10.18730/S4ZR T	8
G40177E 1	USA	Arizona	Pima County	Wild	<i>acutifoliu s</i>	10.18730/S48Z0	6
G40055	Mexico	Sinaloa	Los Mochis	Wild	<i>acutifoliu s</i>	10.18730/JZ0M S	6.8
G40291	Mexico	Sinaloa	San Ignacio	Wild	<i>acutifoliu s</i>	10.18730/JYVT K	4.5
G40016	Mexico	Sonora	Alamos	Cultivate d	<i>acutifoliu s</i>	10.18730/K0XC 8	16.3
G40039	Mexico	Chiapas	Tapachula	Cultivate d	<i>acutifoliu s</i>	10.18730/K1GK Z	13.7

G40022A	USA	Arizona	Pima County	Cultivated	<i>acutifolius</i>	10.18730/S3XJ5	16.2
G40112	Mexico	Durango	Cuencame	Cultivated	<i>acutifolius</i>	10.18730/JP8ED	16.1
G40100	Mexico	Durango	Durango	Wild	<i>tenuifolius</i>	10.18730/JP8QP	4.8
G40098	Mexico	Durango	Francisco I. Madero	Wild	<i>acutifolius</i>	10.18730/JP8BA	4.2
G40093	Mexico	Durango	Durango	Wild	<i>tenuifolius</i>	10.18730/JP865	3.8
G40190	USA	New Mexico	Hidalgo County	Wild	<i>acutifolius</i>	10.18730/JXFYN	3.2
G40249	Mexico	Chihuahua	Santa Clara	Wild	<i>tenuifolius</i>	10.18730/JXVRX	2.1
G40251	Mexico	Chihuahua	Santa Clara	Wild	<i>tenuifolius</i>	10.18730/JXVTZ	2.1
G40256	Mexico	Chihuahua	Santa Clara	Wild	<i>tenuifolius</i>	10.18730/JXVDJ	2.3
G40192	Mexico	Sonora	Sonora	Wild	<i>acutifolius</i>	10.18730/JXG0Q	6.2
G40129	México	Sinaloa	Choix	Cultivated	<i>acutifolius</i>	10.18730/JSH2Z	13.4
G40048	Mexico	Sonora	San Bernardo	Wild	<i>acutifolius</i>	10.18730/JJQ09	3.2
G40083	Mexico	Durango	Villa Ocampo	Wild	<i>tenuifolius</i>	10.18730/JQ5BD	4.3
G40255	Mexico	Chihuahua	Santa Clara	Wild	<i>tenuifolius</i>	10.18730/JXVCH	2.3
G40055	Mexico	Sinaloa	Los Mochis	Wild	<i>acutifolius</i>	10.18730/JZ0MS	6.8
G40234	USA	Wyoming	Cheyenne	Cultivated	<i>acutifolius</i>	10.18730/JXJAQ	21.6

4. Chapter four: Sun-induced Chl fluorescence as an index selection of genotypes to light use efficiency in physiological breeding: Tepary bean functional collection case study

Diego Felipe Conejo-Rodríguez^{1,2,3}, Joaquín Guillermo Ramirez-Gil⁴, Jorge Enrique Aragon Medina³, Javier Mauricio Gereda^{1,3}, Peter Wenzl¹ and Milan Oldřich Urban²

¹ Genetic Resources Program, International Center for Tropical Agriculture (CIAT), Palmira, Valle del Cauca, Colombia.

² Bean Physiology and Breeding Program, International Center for Tropical Agriculture (CIAT), Palmira, Valle del Cauca, Colombia

³ Universidad Nacional de Colombia sede Palmira, Facultad de Ciencias Agropecuarias, Palmira, Valle del Cauca, Colombia

⁴ Universidad Nacional de Colombia sede Bogotá, Facultad de Ciencias Agrarias, Departamento de Agronomía, Bogotá Colombia

Corresponding Author:

Diego Felipe Conejo Rodríguez
Milan Oldřich Urban

Email address: dfconejoro@unal.edu.co (DC): dehydrins@gmail.com

Abstract

The light use efficiency (LUE) affects biomass production and crop yield significantly. The tepary bean has contributed to an important role in the genetic improvement of the common bean; nevertheless, the photosynthetic adaptation aptitude to abiotic stress conditions remain unknown. This study analyzed 56 accessions from the Bioversity-CIAT alliance's tepary bean, this collection was created from the functional trait variation evaluated from the entire tepary bean collection. created a Tepary functional diversity collection, with an emphasis on light use efficiency (LUE) and sun-induced Chl fluorescence (SIF) under drought and control conditions.

The selection index based on light and fluorescence is related to total biomass, pod harvest index, and yield by integrating the tepary bean genetic and ecological groups. Both cultivated and wild accessions from the Sonora-Sinaloa deserts produced better yields at higher fluorescence values, indicating a strong link between light use efficiency and sun-induced fluorescence. Under drought, tepary bean regulates electron transport via qL and rapidly releases energy via non-photochemical quenching , preventing photoinhibition. The weedy G40177F and G40177C accessions had different yield and sun-induced fluorescence responses than their wild relatives, indicating natural hybridizations that could improve photosynthetic efficiency. These findings highlight the need of examining germplasm collections in order to uncover adaptable physiological features important for genetic development.

Keywords: Crops phenomics, photophysics, genebanks, crop evolution, crop adaptation, biomass, functional traits.

4.1 Introduction

Photosynthesis is the conversion of light energy into chemical energy stored in the form of carbohydrates (Gu et al., 2023). Carbon dioxide (CO_2) from the atmosphere and water that the roots absorb are combined to create these carbohydrates, which are then carried to the plant's many organs by the xylem (Kamen, 1963). Photosynthesis is controlled by two photosystems: PSI and PSII (Shi et al., 2022). According to systems modeling, separating photosynthesis into three stages is conceptually important due to the temporal features of subsequent photosynthetic events (Kamen, 1963; Gu et al., 2023). Kamen (1963) proposed the three-step separation, which consists of (I) photophysics, (II) photochemistry, and (III) biochemistry. Photophysical responses include the storage and distribution of light between photosystems, as well as the transmission and distribution of excitation energy through various dissipation paths (Gu et al., 2023). With

respects to photochemical reactions, these include water splitting by the oxygen evolution complex, electron transport from PSII to the cytochrome b6f (Cyt) complex to PSI, the final NADP⁺ acceptor with plastoquinone, plastocyanin, and ferredoxin as electron carriers, as well as proton transport from the stroma to the lumen and ATP synthesis (Sun et al., 2023). Chlorophyll a fluorescence (ChlF) is the emission of red and far-red photons from excited chlorophyll molecules, competing with photochemical and non-photochemical energy uses (Gu et al., 2019). Sun-induced fluorescence Chl (SIF) occurs when the leaf absorbs irradiance, which stimulates chlorophyll molecules in PSII and lowers energy states via photochemical and non-photochemical quenching (NPQ) (Gu et al., 2019). In bean (*Phaseolus* spp.) physiological breeding, finding selection strategies for genotypes with improved light use efficiency (LUE) to increase genetic gains linked with drought and high temperature conditions is a top objective. In this regard Keller et al. (2023), confirmed that LUE can be used in a site-specific selection process of Andean climbing beans even with higher precision than genomic selection. Conserved genebanks collections can serve as the basis for understanding genome variation and relating genetic and ecological components to phenotypic variation in currently conserved collections (Conejo - Rodriguez et al., 2024), but few studies have examined photosynthetic trait variation in germplasm collections under natural variation or abiotic stress. While the tepary bean is known for its resistance to drought and high temperatures, little research has been conducted to enhance its photosynthetic capacity (Conejo-Rodríguez et al., 2022). Exploring the variation of photosynthetic features in the genetic resources of tepary bean will enable the selection of parents with the potential for genetic improvement of these physiological traits. Thus, to establish a LUE selection index of tepary bean accessions that contributes to the physiological improvement of tepary bean gene pool itself seems to be a tangible goal for the upcoming

breeding programs. The purpose of this study is to understand the variation of SIF in germplasm collections and how it relates to biomass, harvest index at the pod level, and yields that reflect the LUE of wild and domesticated accessions in terms of genetic structure and ecological distribution. This will enable the selection of genotypes with diverse LUE and the generation of diverse breeding schemes with an emphasis on the target prediction environment (TPE).

4.2 Materials and Methods

4.2.1 Methodological approach

The methodology of this study is built on the following stages: **(I) Plant material:** 56 tepary bean accessions from the functional trait collection (Conejo-Rodriguez et al., in review) were studied. These accessions exhibit the collection's greatest phenotypic variation and geographical distribution available at Alliance of Bioversity International and CIAT. **(II) Experimental design:** The functional collection was evaluated in two moisture level treatments (irrigation *versus* drought) using a spatial experimental design in which the spatial position of each accession is related. **(III) Phenotyping of functional photosynthetic traits:** Photosynthetic traits were measured during the reproductive cycle with the MultispeQ (PhotosynQ, USA) proximal fluorometer. **(IV) Sun-induced Chl fluorescence (SIF):** SIF was calculated using the methodology of Gu et al. (2019). **(V) Selection index:** Using the SIF relationship and photochemical quenching (qL), an index of selection for efficient light use is determined. The value of the selection index is related to the harvest index and yield (Figure 4.1).

4.2.2 Tepary bean functional diversity collection

Fifty-six accessions from the Alliance Bioversity - CIAT genebank's tepary bean functional trait collection (Conejo - Rodriguez et al in review) were analyzed. These accessions represent variation in phenotype within the species' conserved collection. These accessions are represented by both wild and cultivated accessions with corresponding geographic distributions (S1 - Passport data of evaluated accessions). The improved variety Tep23 was utilized as a control in this research (Porch et al., 2013) (Figure 4.1A).

4.2.3 Experimental design and cultivation

The functional collection of tepary bean under irrigation and drought conditions was evaluated using a spatial experimental design in which the spatial positions of each accession were associated. Four (4) replications in spatial design were carried out under two soil moisture conditions (I) irrigation and (II) drought. Drip irrigation with compensated drippers was employed to ensure that each of the treatments tested received the same amount of water. Each experimental unit consisted of four (4) plants per accession. Each treatment was assessed with 4 replications so with 16 plants. The experimental design was implemented in a spatial arrangement, with each experimental unit assigned a spatial position determined by row and column. Planting was initially done in trays to ensure a certain number of plants per experimental unit. A 6:3:1 mix of peat, sand, and compost was employed to aid in the germination and root growth of both wild and cultivated accessions. Seedlings transplanting was placed 10 days after sowing (DDS) for each treatment. The drought treatment was implemented 15 days after transplanting, with irrigation fully restricted until the end of the cycle. For the drought treatment a total of 1287 L was applied during the first 15 days after transplanting (DDT) while the

irrigation treatment was 1883 L until day 44 DDT. Soil, climate, irrigation curves, fertilizer and pesticide analyses used are given in Supplementary 1.

4.2.4 Phenotyping of photosynthetic functional traits and yield components

The MultispeQ hand-held fluorometer was used to measure photosynthetic traits in the central leaflet of the middle section of the plant from 10:30 a.m. to 1:00 p.m. (Kuhlgert et al., 2016; Conejo-Rodriguez et al., 2022). Each experimental unit has been evaluated with two biological replicates for a total of 8 measurements per accession per day of measurement. During the experiment under drought and irrigation environments, the Photosynthesis RIDES 2.0 protocol was used to score 2007 data points (Available at: <https://photosynq.org/projects/jorge2-tepary>, Project ID: 24330).

MultispeQ measures photosynthetic parameters, including I) photosystem II quantum yield (Φ_{II} - Phi2); II) non-photochemical quenching (Φ_{NPQ} - PhiNPQ); III) energy losses due to heat dissipation (Φ_{NO} - PhiNO); IV) relative chlorophyll (Chl); V) leaf temperature differential (LTD); VI) maximum variable fluorescence under steady-state conditions (F_m'); VII) minimum variable fluorescence during the dark phase after steady state (F_0'); VIII) variable fluorescence under steady state conditions (F_s'); IX) efficiency of open reaction centers in the light (F_v'/F_m') and; X) the fraction of PSII centers open when QA is oxidized (q_L), in addition to air temperature, relative humidity and photosynthetically active radiation (PAR) (Kuhlgert et al. , 2016). At phenological stages BBCH 61 and 81, the dry weight biomass and leaf area of each accession were determined (Aragon et al., under review). The pod harvest index (PHI) (Assefa et

al., 2013), number of pods, and yield were calculated for each accession studied. The PHI was computed using the total number of pods sampled for each accession.

4.2.5 Sun-induced Chl fluorescence (SIF)

Chl a fluorescence (ChlF) is the emission of red and far-red photons from excited Chl molecules in competition with photochemical and non-photochemical energy uses that are closely related to photosynthesis at the fundamental biochemical and biophysical processes (Porcar-Castell et al., 2014; Gu et al., 2019). The sun-induced Chl fluorescence (SIF) model (Gu et al., 2019) integrates photosynthesis dissipation mechanisms relation photochemical efficiency (Φ_{PSII} max), non-photochemical quenching (NPQ), photochemical quenching (qL), and photosynthetically active radiation (PAR) using the following equation:

$$SIF = \frac{1 - \phi_{PSII} \max}{(1 + kDF) * [(1 + NPQ) * (1 - \phi_{PSII} \max) + qL * (\phi_{PSII} \max)] * \alpha * \beta} \quad (1)$$

The SIF equation (1) relates PSII max, NPQ, qL, in addition it associates constants such as leaf absorbance ($\alpha = 0.84$, Björkman and Demmig, 1987; Schreiber, 2004); The fraction of light assigned to PSII ($\beta = 0.5$, von Caemmerer, 2000); and the ratio of the rate constants of thermal dissipation and fluorescence (KDF ($kF/kD = 10$, Pfundel, 1998) (kF = emission fluorescence; and kD = Constitutive thermal radiation, see Gu et al., 2019). KDF, as an intrinsic physical feature of chlorophyll (Chl) molecules, is considered constant.

4.2.6 Selection index for of light use efficiency: Relationship between SIF, biomass, yield and harvest index

Light use efficiency (LUE) models have been widely used to estimate terrestrial gross primary production (GPP) (Pei et al., 2022). The LUE models estimate GPP as the energy absorbed by plants multiplied by the actual LUE that converts energy to carbon fixed during the photosynthesis process (Monteith 1972). Gu et al. (2019) recently developed the mechanistic light reaction (MLR) model, which provides a mechanistic framework of the SIF emitted by PSII with the actual linear electron transport rate from PSII to PSI (J) and the rate of maximum photosynthesis (A_{max}), which is directly related to carbon fixation and is an improved version of the Farquhar, von Caemmerer, Berry (FvCB) biochemical model (Farquhar et al., 1980). MLR is a photophysical model that quantifies excitation energy dissipation pathways as a function of the fraction of open reaction centers in photosystem II (PSII) under the lake (q_L) model (Kramer et al., 2004), as well as NPQ, which estimates LUE from pulse amplitude modulated (PAM) (Kuhlgert et al., 2016.).

LUE is intimately related to q_L (Kramer et al., 2004; Gu et al., 2019). The parameter q_L denotes the fraction of reduction (opening) of the PSII reaction center, indicating the proportion of PSII reaction centers capable of transferring electrons downstream to the plastoquinone (QA) pool (Kramer et al., 2004). q_L is a mechanism that regulates photosynthetic efficiency in response to changing light conditions (Liu et al., 2024). q_L modulates the rate of electron transport in PSII (Gu et al., 2019). When the electron flow rate in photosynthesis accelerates and exceeds the energy usage capacity, q_L is activated to slow the flow and balance the energy demand (Ding et al., 2023).

Taking this into account, SIF and qL are related , with qL values that respond to differently photosynthetically active radiation (PAR). Tepary bean exhibits diverse traits as a result of natural selection processes and domestication (Conejo-Rodriguez et al., see Chapter 2); thus, LUE should differ depending on the ecological diversity (PAR, soils, temperature and rainfalls) to which it was exposed to these selection processes, resulting in variation in the SIF - qL relation. We propose a selection index based on quadrant-specific SIF and qL medians (Figure 4.1B) Assuming the foregoing, the median ratio between SIF and qL for each quadrat was determined under the following assumptions (Figure 4.1B):

$qL > \mu$ and $SIF > \mu$ in Q1 related to $> PAR > LUE$

$qL < \mu$ and $SIF > \mu$ in Q2 related to $< PAR > LUE$

$qL < \mu$ and $SIF < \mu$ in Q3 related to $< PAR < LUE$.

$qL > \mu$ and $SIF < \mu$ in Q4 related to $> PAR < LUE$

Taking into account the medians for each quadrant, the following selection index is determined to be used to evaluate the relationship with the harvest index, biomass, and yield using the equation:

$$SIFmax = \frac{Q1 + Q2 + Q3 + Q4}{4} \quad (2)$$

The equation for the selection index for light use efficiency (SIFmax) (2) is the sum of the medians of each quadrant divided by the total number of quadrants. This was estimated for each of the accessions examined. Finally, to validate the relationship between SIF and net primary productivity (GPP), we compared total biomass, yield, and harvest index at the pod level (PHI), as well as the relationship between ecological groups (Conejo - Rodriguez et al., see Chapter 3)

and the genetic structure of tepary bean (Bornowski et al., 2023). For the relationship of SIF and qL, general additive models (GAM) are performed using the R gam library. The tidyverse R library is used for data manipulation, analysis, and visualization.

4.3 Results

Our findings reveal that the photosynthetic performance of selected tepary bean accessions varies differently between drought and irrigation circumstances, and also between cultivated and wild accessions. The different accessions perform differently in terms of the relatedness between sun-induced Chl fluorescence (SIF) and the fraction of PSII reaction centers opened when plastoquinone is oxidized (QA) (qL) (Figure 4.2). The biological states show differences in photosynthetic performance among weedy, cultivated, and wild accessions (Figure 4.2); nonetheless, SIF drops are observable between 0.4 and 0.5 qL, particularly in weedy and cultivated accessions under both irrigation and drought conditions (Figure 4.2A, B).

The wild accessions perform similarly in both drought and irrigated conditions (Figure 2C), whereas the cultivated modified their SIF-qL ratio in drought conditions, displaying lower values of SIF - qL between 0.4 and 0.5 (Figure 4.2B). The accession G40177C, G40177D, and G40177F exhibit intermediate characteristics in between cultivated and wild biological states; however, G40177C in irrigated conditions exhibits higher and stable SIF values of 600 umol photons m⁻² s⁻¹ with low qL values of 0.2 to 0.4 (Figure 4.2A). When SIF is calculated from the distribution, a linear connection is shown for cultivated, wild and weedy accessions (Figure 4.3), revealing accessions with higher LUE under high and low PAR conditions.

The control irrigated conditions showed less data variance and higher linearity, which is expectable. The cultivated accession G40100 showed higher SIF values of 450 $\mu\text{mol photons m}^{-2} \text{ s}^{-1}$ under both irrigation and drought conditions, whereas the accession G40248A showed SIF values less than 350 $\mu\text{mol photons m}^{-2} \text{ s}^{-1}$ (Figure 4.3), indicating that some cultivated tepary bean accessions perform differently. Under drought and irrigation conditions, wild accessions demonstrated equal SIF activity, although values more than 300 $\mu\text{mol photons m}^{-2} \text{ s}^{-1}$ produced different qL values (Figure 4.3). The weedy accessions demonstrate significant variation under both drought and irrigation situations, with values greater than 400 $\mu\text{mol photons m}^{-2} \text{ s}^{-1}$. In terms of yield, cultivated accessions produced greater yields with higher SIF values, particularly under irrigation circumstances (Figure 4.3). Under drought conditions, cultivated accessions had yield below 1500 kg ha⁻¹, whereas under irrigated conditions, they approached 2000 kg ha⁻¹ (Figure 4.4).

When the ecological and genetic groups that contribute to the relationship of SIF with yield in cultivated accessions are compared, it is observed that the ecological group CC2 associated with the regions of Sinaloa, Nayarit, and Jalisco (Conejo - Rodriguez et al., see Chapter 3) presents yields greater than 1500 kg ha⁻¹, as well as the highest values of SIF greater than 400 $\mu\text{mol photons m}^{-2} \text{ s}^{-1}$ under irrigated conditions (Figure 4.4A). In contrast to CC4, which refers to accessions collected throughout the Pacific region, including Guatemala, El Salvador, and Costa Rica (Conejo - Rodriguez et al., See Chapter 3). These accessions had higher SIF values of 400 $\mu\text{mol photons m}^{-2} \text{ s}^{-1}$ but lower yields than

accessions from ecological group CC2 (Figure 4.4A). Under drought conditions, the ecological groups do not form clusters that highlight varied photosynthetic performances, indicating that the adaptation responses are similar (Figure 4.4A).

The genetic groups follow the same trend as the ecological groups, with accessions connected with the Sonora - Sinaloa (GG1_SSc) and Central America (GG3_CAc) genetic groups (Bornowski et al., 2023). It should be noted that accessions G40177F, G40177C, and G40177D1 belong to the ADMIX genetic group, which is related to accessions that were not classified in any related genetic group (Bornowski et al., 2023) and showed yields higher than 1000 kg ha⁻¹ in drought conditions and higher than 2500 kg ha⁻¹ with irrigation, with SIF values higher than 350 µmol photons m⁻² s⁻¹ in drought conditions and values between 400 µmol photons m⁻² s⁻¹ to 500 µmol photons m⁻² s⁻¹ in irrigation conditions (Figure 4.4B).

For the wild accessions, it is observed that the ecological group WC1 shows yields higher than 1000 kg ha⁻¹ under irrigated conditions, with accessions that can reach yields higher than 2000 kg ha⁻¹, with higher SIF values of 400 µmol photons m⁻² s⁻¹ (Figure 4.5A). Meanwhile the accessions associated with ecological group WC2 yield lower than 1500 kg ha⁻¹, but they show SIF values similar as accessions of ecological group WC1 (Figure 4.5A, B). It should be noted that the wild accessions did not show distinct yields between drought and irrigation (Figure 4.5, B). For example, accession G40107 yielded over 1000 kg ha⁻¹ in both soil moisture contents, despite having SIF of 300 µmol photons m⁻² s⁻¹ in drought and 400 µmol photons m⁻² s⁻¹ in irrigation (Figure 4.5A, B). The harvest index at

pod level (PHI) in control conditions for the cultivated accessions was higher than 75%; however, under drought conditions, accessions G40011 and G40039 had PHI values less than 70% (Figure 4.6A).

The accessions with the lowest PHI values are related to the Central American ecological group (CC4) (Figure 4.6A). This is also visible when comparing the genetic groups (Figure 4.5A), as the accessions with the lowest PHI values are related with the Central American genetic group (GG3_CAc) (Figure 4.6B), demonstrating the ecological and genetic influences on PHI trait variation. The PHI values for wild accessions were less than 76% under both drought and irrigation regimes (Figure 4.7). In our visualizations the PHI variation is mostly connected to SIF values between 300 and 600 $\mu\text{mol photons m}^{-2}\text{s}^{-1}$. The accession G40204 has the highest PHI in both drought and irrigation conditions, with SIF values ranging from 300 $\mu\text{mol photons m}^{-2}\text{s}^{-1}$ to 350 $\mu\text{mol photons m}^{-2}\text{s}^{-1}$ (Figure 4.7A). This accession belongs to the ecological group WC2, which includes accessions of the botanical variety *tenuifolius* collected in southeastern Arizona (Figure 4.7A, B).

The botanical variety *acutifolius* accessions exhibit SIF values ranging from 300 to 550 $\mu\text{mol photons m}^{-2}\text{s}^{-1}$, but have PHI values lower than 70% (Figure 4.7A, B). When the relationship is made with the total biomass in cultivated accessions, it is observed that accession G40177F presents the highest dry biomass near 40 g with SIF values greater than 500 $\mu\text{mol photons m}^{-2}\text{s}^{-1}$, whereas the majority of accessions present values less than 30 g in irrigated conditions (Figure 4.8A, B). In drought, the maximum dry biomass is 20 g, with accession G40177C having the highest dry biomass among all the groups, with

SIF values approaching 400 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$. It should be highlighted that accession G40112 also produced around 20 g of dry biomass in both drought and irrigation conditions. G40112 is a member of the ADMIX genetic group, which also includes above-mentioned accessions G40177F and G40177C (Figure 4.8). Regarding the genetic and ecological groups, it is observed that in both treatments the cultivated accessions from Central America (CC4 - GG3_CAc) showed lower dry biomass values than the ecological group CC2, which is related to the genetic cluster of Sinaloa - Sonora (GG1_SSc).

In the case of wild accessions, dry biomass values range from 5 g to 50 g in irrigated conditions and a maximum of 21 g in drought, with SIF values ranging from 150 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$ to 600 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$ (Figure 4.9A, B), though the highest values are associated with stocks belonging to the ADMIX genetic group (Figure 4.9B). Similarly, accessions from various genetic and ecological groups exhibit similar dry biomass under both drought and irrigation conditions (Figure 4.9), with accessions from the Sonora-Sinaloa region exhibiting greater dry biomass weight, in addition to being of the botanical variety *acutifolius* (Figure 4.9B). SIF's relationship to dry biomass, leaf area, and yield in cultivated and wild accessions under drought and irrigation conditions (Figure 4.10). In the accessions grown under irrigated conditions, SIF showed inversely proportional reactions in most of the evaluated variables, particularly yield ($r^2 = -0.35$), PHI ($r^2 = -0.33$), and yellow leaf area ($r^2 = -0.34$), with the exception of green leaf area ($r^2 = 0.51$) and qL ($r^2 = 0.61$) (Figure 4.10A).

Under drought conditions in cultivated accessions, SIF had directly proportional relationships with PHI ($r^2 = 0.11$), yield ($r^2 = 0.47$), biomass ($r^2 = 0.36$), yellow leaf area ($r^2 = 0.26$), and PHI had an inverse association with green leaf area ($r^2 = -0.53$) (Figure 4.10B). In wild accessions grown under irrigated conditions, SIF has directly proportional relationships with yield ($r^2 = 0.32$), green leaf area ($r^2 = 0.14$), and yellow leaf area ($r^2 = 0.55$), but inversely proportional interactions with PHI ($r^2 = -0.25$). Under drought conditions, SIF has no correlations with the other variables, but it has the strongest relationship with qL (0.77) when compared to the relationships of both cultivated accessions (Figure 4.10D).

5.4 Discussion

Tepary bean was of cosmogonic and social importance, particularly for people living in the dry parts of the Mexican Pacific and the arid regions and deserts of Chihuahua, Sonora, Sonaloa and Arizona (Rea, 1981; Nabhan and Teiwes, 1983). Several papers shown tepary bean as more resistant to drought and high temperatures than common bean (Beebe et al., 2008; Beebe et al., 2011); however, evaluations of tepary bean germplasm collections have been limited (Suarez et al., 2021), particularly with a focus on physiological breeding, as discussed in this paper. Our study explores 56 accessions from the Bioversity-CIAT alliance's tepary bean functional diversity collection (Conejo-Rodriguez et al., in review), allowing us to link ecological and genetic groups to light use efficiency (LUE) performance by using sun-induced Chl fluorescence (SIF) as a selection trait with potential for physiological/genetic improvement of beans.

Tepary bean has a geographic distribution that extends from Arizona to Costa Rica in cultivated accessions and from Arizona to Sinaloa to Durango in wild accessions, with cultivated accessions collected in both Mesoamerica (Tropical) and Arid America (Desert) regions (Conejo - Rodriguez et al., in review; Bornowski et al., 2023). These ecological variations generated photosynthetic efficiency adaptations in the evaluated tepary bean accessions. We demonstrated that LUE and biomass accumulation are influenced by crop biological status, ecological group, and genetic structure under both drought and irrigation conditions. The tepary bean exhibits adaptations to drought and high temperatures, owing primarily to ecological factors that influenced the evolutionary processes that established its genetic structure (Conejo - Rodriguez et al., in review), conditioning performance, LUE, and the relationship of SIF to net primary production (Biomass + Yield), which is directly related to qL (Gu et al., 2019).

In our study, we were able to identify the accessions with the highest values of SIF related to lowest/highest qL. This demonstrates what variations exist within the evaluated functional collection of tepary bean in both drought and irrigation conditions, primarily distinguishing between cultivated and wild accessions in drought conditions (Figure 4.2). The parameter qL denotes the fraction of reduction (opening) of the PSII reaction center, indicating the proportion of PSII reaction centers capable of transferring electrons downstream to the plastoquinone (PQ) pool (Kramer et al., 2004; Liu et al., 2024), it is also a photosynthetic trait that enables the link between stomatal conductance and carbon fixation (Han et al., 2022; Gu et al., 2019). Slow qL induction plays an important role in limiting photosynthesis under fluctuating light conditions (Han et al., 2022). Generally, tepary bean accessions evaluated both wild, cultivated and weedy show slow qL inductions ($qL < 0.5$, see Gu et al., 2019) under both drought and irrigated conditions (Figure 4.2). This is most likely due to the slow activation of the cytochrome b6f

complex (Cyt 6bf), which mediates electron transfer and proton translocation between the PSII and PSI reaction center complexes, positively contributing to linear electron flow and ensuring the products required for carbon fixation (Johnson and Berry, 2021).

Recent research has linked SIF to yield and physicochemical traits in wheat, rice, and avocado crops (Liu et al., 2024; Ding et al., 2023; Vitrack-Tamam et al., 2024). But none has evaluated germplasm collections and their genetic and ecological relationships under conditions of natural variation and drought, as well as a generating workflow for genotype selection using diverse LUE mechanisms. Tepary bean yield, PHI, biomass, and SIF vary according to drought conditions, as well as its genetic and ecological structure in wild and cultivated accessions, as demonstrated in this study.

We demonstrated that both cultivated and wild accessions collected in the Sonora-Sinaloa deserts (insert here group code) had higher yields at higher SIF values, demonstrating that there is a close relationship between LUE and SIF in both irrigation and drought conditions in tepary bean, in addition to being a component driven by evolutionary processes (selection and adaptation) (Alam and Purugganan, 2024). SIF is directly related to photosynthetic efficiency, particularly during photophysical reactions that include light harvesting and distribution across photosystems, as well as the transmission of excitation energy to various dissipation routes (Sun et al., 2023). Considering the foregoing, we discovered that under drought conditions, the tepary bean exhibits several adaptive mechanisms, primarily regulating electron transport through qL-directed relaxation and rapid dissipation of energy by NPQ heat (Gu et al., 2019; Kramer et al., 2004), preventing photoinhibition of photosystems, but necessitating energetic expenditures during photoacclimation (Bassi and Dall'Osto, 2021; Gu et al., 2019).

Recently, the use of crop wild relatives for genetic prebreeding or interspecific crossings of diverse crop species has been established and is increasing in importance (Sharwood et al., 2022; Croce et al., 2024). In this study, the weedy G40177F and G40177C accessions produce different yields and SIF responses under drought and irrigation conditions than their crop wild relatives (CWR). These accessions are probably natural hybridizations and were collected in early hybridization stages segregating diverse types of seed color (Conejo - Rodriguez et al., in review; Chacón-Sánchez et al., 2021), besides presenting mixed genetic structure (Bornowski et al., 2023) that can contribute to increase the transfer of photosynthetic traits that contribute to the improvement of photosynthetic efficiency (Gao et al., 2024; Croce et al., 2024) during interspecific hybridization.

The objectives of physiological breeding are to investigate variation in crop genetic resources, as well as to develop experiments under conditions of natural variation and abiotic stress that contribute to the expression of physiological traits, allowing the selection of probable progenitors through the use of High throughput phenotyping strategies (Reynolds and Langridge, 2016). The exploration of germplasm collections is of great interest for physiological improvement because of the potential for discovering photochemical traits associated with evolutionary processes directed by natural selection (Croce et al., 2024; Gao et al., 2024). We take an approach to explore the variation of photosynthetic traits in the functional collection of tepary bean (Conejo - Rodriguez et al., in review) demonstrating the importance of conducting studies in germplasm collections exploring adaptive phenotypic diversity.

4.5 Conclusions

In this study research, we investigate the photosynthetic traits in the tepary bean functional trait collection and their direct link to yield. We explore how germplasm screening can give light use efficiency (LUE) and how it is linked to evolutionary context. We present a methodology for selecting accessions with higher photosynthetic potential and explaining the variation under both natural and drought conditions, emphasizing the importance of investigating the genetic diversity retained in genebanks. The selection index based on sun-induced Chl fluorescence (SIF) provides for the discrimination of accessions with varying potential for LUE in both high and low light intensity conditions, allowing for the selection of parents based on their LUE.

4.6 Acknowledgments

I would like to thank the people who are part of the plant physiology laboratory for their support in data capture and the harmony in which each of the experiments was carried out. I thank the germplasm bank of the Bioversity - CIAT alliance for the support generated during the research. I would also like to thank the students of the phenomics, data science and artificial intelligence study group of the Universidad Nacional de Colombia, Palmira campus, for their support during the formulation and development of this experiment, especially Manuel Cortés for his support in the conceptual development.

4.7 References

- Alam, O., & Purugganan, M. D. (2024). Domestication and the evolution of crops: variable syndromes, complex genetic architectures, and ecological entanglements. *The Plant Cell*, koae013.
- Assefa, T., Beebe, S. E., Rao, I. M., Cuasquer, J. B., Duque, M. C., Rivera, M., ... & Lucchin, M. (2013). Pod harvest index as a selection criterion to improve drought resistance in white pea bean. *Field Crops Research*, 148, 24-33.
- Bassi, R., & Dall'Osto, L. (2021). Dissipation of light energy absorbed in excess: the molecular mechanisms. *Annual review of plant biology*, 72, 47-76.
- Beebe, S. E., Rao, I. M., Cajiao, C., & Grajales, M. (2008). Selection for drought resistance in common bean also improves yield in phosphorus limited and favorable environments. *Crop Science*, 48(2), 582-592.
- Beebe, S., Ramirez, J., Jarvis, A., Rao, I. M., Mosquera, G., Bueno, J. M., & Blair, M. W. (2011). Genetic improvement of common beans and the challenges of climate change. *Crop adaptation to climate change*, 356-369.
- Björkman, O., & Demmig, B. (1987). Photon yield of O₂ evolution and chlorophyll fluorescence characteristics at 77 K among vascular plants of diverse origins. *Planta*, 170, 489-504.
- Bornowski, N., Hart, J. P., Palacios, A. V., Ogg, B., Brick, M. A., Hamilton, J. P., ... & Porch, T. (2023). Genetic variation in a tepary bean (*Phaseolus acutifolius* A. Gray) diversity panel reveals loci associated with biotic stress resistance. *The Plant Genome*, 16(3), e20363.
- Bornowski, N., Hart, J. P., Palacios, A. V., Ogg, B., Brick, M. A., Hamilton, J. P., ... & Porch, T. (2023). Genetic variation in a tepary bean (*Phaseolus acutifolius* A. Gray) diversity panel reveals loci associated with biotic stress resistance. *The Plant Genome*, 16(3), e20363.
- Bornowski, N., Hart, J. P., Palacios, A. V., Ogg, B., Brick, M. A., Hamilton, J. P., ... & Porch, T. (2023). Genetic variation in a tepary bean (*Phaseolus acutifolius* A. Gray) diversity panel reveals loci associated with biotic stress resistance. *The Plant Genome*, 16(3), e20363.
- Chacón-Sánchez, M. I., Martínez-Castillo, J., Duitama, J., & Debouck, D. G. (2021). Gene flow in *Phaseolus* beans and its role as a plausible driver of ecological fitness and expansion of cultigens. *Frontiers in Ecology and Evolution*, 9, 618709.
- Conejo-Rodriguez, D. F. C., Urban, M. O., Santaella, M., Gereda, J. M., Contreras, A. D., & Wenzl, P. (2022). Using phenomics to identify and integrate traits of interest for better-performing common beans: A validation study on an interspecific hybrid and its *Acutifolii* parents. *Frontiers in Plant Science*, 13, 1008666.

Croce, R., Carmo-Silva, E., Cho, Y. B., Ermakova, M., Harbinson, J., Lawson, T., ... & Zhu, X. G. (2024). Perspectives on improving photosynthesis to increase crop yield. *The Plant Cell*, koae132.

Ding, H., Wang, Z., Zhang, Y., Li, J., Jia, L., Chen, Q., ... & Wang, S. (2023). A Mechanistic Model for Estimating Rice Photosynthetic Capacity and Stomatal Conductance from Sun-Induced Chlorophyll Fluorescence. *Plant Phenomics*, 5, 0047.

Ding, H., Wang, Z., Zhang, Y., Li, J., Jia, L., Chen, Q., ... & Wang, S. (2023). A Mechanistic Model for Estimating Rice Photosynthetic Capacity and Stomatal Conductance from Sun-Induced Chlorophyll Fluorescence. *Plant Phenomics*, 5, 0047.

Gao, Y., Stein, M., Oshana, L., Zhao, W., Matsubara, S., & Stich, B. (2024). Exploring natural genetic variation in photosynthesis-related traits of barley in the field. *Journal of Experimental Botany*, erae198.

Gao, Y., Stein, M., Oshana, L., Zhao, W., Matsubara, S., & Stich, B. (2024). Exploring natural genetic variation in photosynthesis-related traits of barley in the field. *Journal of Experimental Botany*, erae198.

Gu, L., Han, J., Wood, J. D., Chang, C. Y. Y., & Sun, Y. (2019). Sun-induced Chl fluorescence and its importance for biophysical modeling of photosynthesis based on light reactions. *New Phytologist*, 223(3), 1179-1191.

Han, J., Gu, L., Warren, J. M., Guha, A., McLennan, D. A., Zhang, W., & Zhang, Y. (2022). The roles of photochemical and non-photochemical quenching in regulating photosynthesis depend on the phases of fluctuating light conditions. *Tree Physiology*, 42(4), 848-861.

Johnson, J. E., & Berry, J. A. (2021). The role of cytochrome b6f in the control of steady-state photosynthesis: a conceptual and quantitative model. *Photosynthesis Research*, 148(3), 101-136.

Kramer, D. M., Johnson, G., Kiirats, O., & Edwards, G. E. (2004). New fluorescence parameters for the determination of QA redox state and excitation energy fluxes. *Photosynthesis research*, 79, 209-218.

Kuhlgert, S., Austic, G., Zegarac, R., Osei-Bonsu, I., Hoh, D., Chilvers, M. I., ... & Kramer, D. M. (2016). MultispeQ Beta: a tool for large-scale plant phenotyping connected to the open PhotosynQ network. *Royal Society open science*, 3(10), 160592.

Liu, Z., Guo, C., Yu, Q., Zhu, P., Peng, X., Dong, M., ... & Lu, X. (2024). A SIF-based approach for quantifying canopy photosynthesis by simulating the fraction of open PSII reaction centers (qL). *Remote Sensing of Environment*, 305, 114111.

Liu, Z., Guo, C., Yu, Q., Zhu, P., Peng, X., Dong, M., ... & Lu, X. (2024). A SIF-based approach for quantifying canopy photosynthesis by simulating the fraction of open PSII reaction centers (qL). *Remote Sensing of Environment*, 305, 114111.

- Nabhan, G. P., & Teiwes, H. (1983). Tepary beans, O'odham farmers, and desert fields [*Phaseolus acutifolius*, Southwestern USA desert population, planting and cultivation, harvesting, cooking; Western States (USA)]. *Desert Plants (USA)*, 5(1).
- Pei, Y., Dong, J., Zhang, Y., Yuan, W., Doughty, R., Yang, J., ... & Xiao, X. (2022). Evolution of light use efficiency models: Improvement, uncertainties, and implications. *Agricultural and Forest Meteorology*, 317, 108905.
- Pfundel, E. (1998). Estimating the contribution of photosystem I to total leaf chlorophyll fluorescence. *Photosynthesis research*, 56, 185-195.
- Porcar-Castell, A., Tyystjärvi, E., Atherton, J., Van der Tol, C., Flexas, J., Pfundel, E. E., ... & Berry, J. A. (2014). Linking chlorophyll a fluorescence to photosynthesis for remote sensing applications: mechanisms and challenges. *Journal of experimental botany*, 65(15), 4065-4095.
- Porch, T. G., Beaver, J. S., & Brick, M. A. (2013). Registration of tepary germplasm with multiple-stress tolerance, TARS-Tep 22 and TARS-Tep 32. *Journal of Plant Registrations*, 7(3), 358-364.
- Porch, T., & Beaver, J. (2012). Development of improved tepary bean (*Phaseolus acutifolius* A. Gray) germplasm. *Ann. Rep. Bean Improv. Coop*, 55, 123-124.
- Rao, I., Beebe, S., Polania, J., Ricaurte, J., Cajiao, C., Garcia, R., & Rivera, M. (2013). Can tepary bean be a model for improvement of drought resistance in common bean?. *African Crop Science Journal*, 21(4).
- Rea, A. M. (1981). Resource utilization and food taboos of Sonoran desert peoples. *Journal of Ethnobiology*, 1(1), 69-83.
- Reynolds, M., & Langridge, P. (2016). Physiological breeding. *Current opinion in plant biology*, 31, 162-171.
- Sharwood, R. E., Quick, W. P., Sargent, D., Estavillo, G. M., Silva-Perez, V., & Furbank, R. T. (2022). Mining for allelic gold: finding genetic variation in photosynthetic traits in crops and wild relatives. *Journal of Experimental Botany*, 73(10), 3085-3108.
- Suárez, J. C., Contreras, A. T., Anzola, J. A., Vanegas, J. I., & Rao, I. M. (2021). Physiological characteristics of cultivated tepary bean (*Phaseolus acutifolius* A. Gray) and its wild relatives grown at high temperature and acid soil stress conditions in the Amazon region of Colombia. *Plants*, 11(1), 116.
- Sun, Y., Gu, L., Wen, J., van Der Tol, C., Porcar-Castell, A., Joiner, J., ... & Luo, Z. (2023). From remotely sensed solar-induced chlorophyll fluorescence to ecosystem structure, function, and service: Part I—Harnessing theory. *Global change biology*, 29(11), 2926-2952.

Vitrack-Tamam, S., Yasour, H., Hamus-Cohen, D., Erel, R., Rubinovich, L., & Liran, O. (2023). Solar Induced Fluorescence Retrieved from Avocado (*Persea americana* Mill.) Canopies Along the Season Correlates with Sugar Levels in the Developing Fruit. *bioRxiv*, 2023-07.

von Caemmerer, S., Farquhar, G., & Berry, J. (2009). Biochemical model of C3 photosynthesis. In *Photosynthesis in silico: understanding complexity from molecules to ecosystems* (pp. 209-230). Dordrecht: Springer Netherlands.

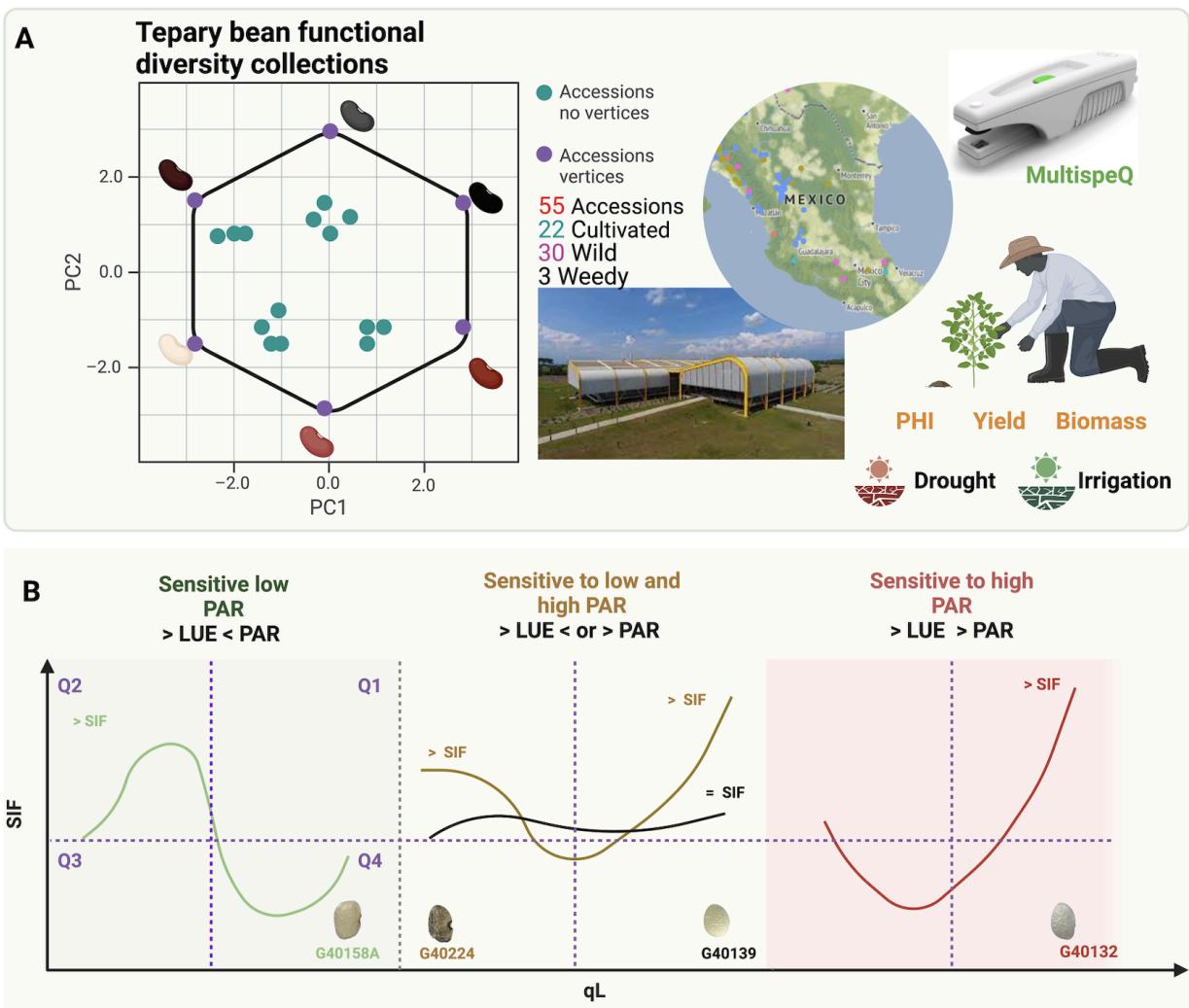


Figure 4.1. Methodological process of physiological evaluation of the tepary bean functional diversity collection. A) Accessions of the tepary bean functional diversity collection and phenotyping using the MultispeQ field fluorimeter under drought and irrigation conditions; B) Selection index to light use efficiency from quadrant analysis of the relationship between sun-induced Chl fluorescence (SIF) and the fraction of PSII reaction centers opened when plastoquinone is oxidized (QA) (qL).

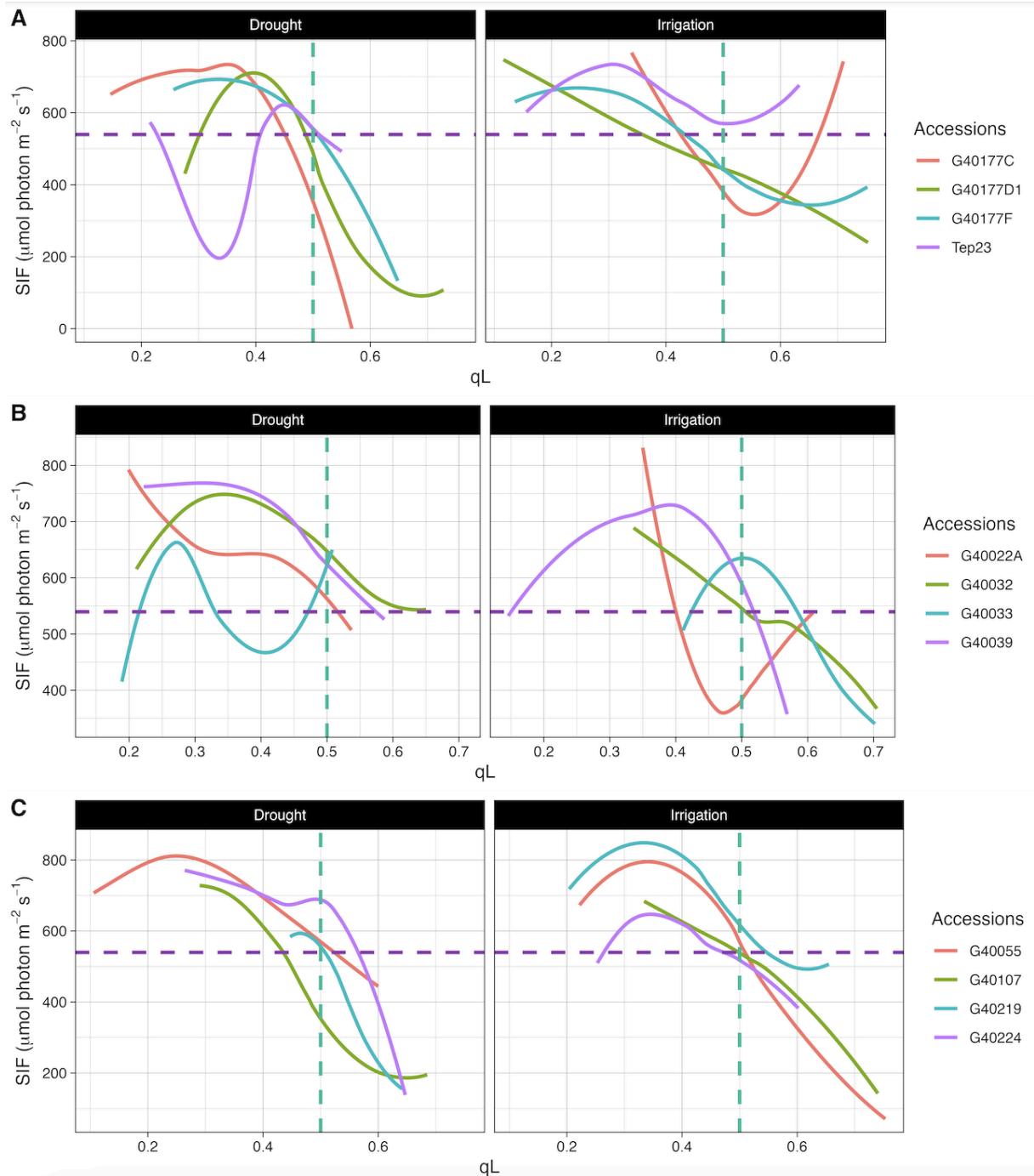


Figure 4.2. Relationship of qL with SIF for a set of tepeary bean accessions evaluated under drought and irrigation conditions. **A)** Regressions and control; **B)** Cultivated and **C)** Wild. The variation of performance for each of the accessions according to photosynthetically active radiation (PAR) is observed.

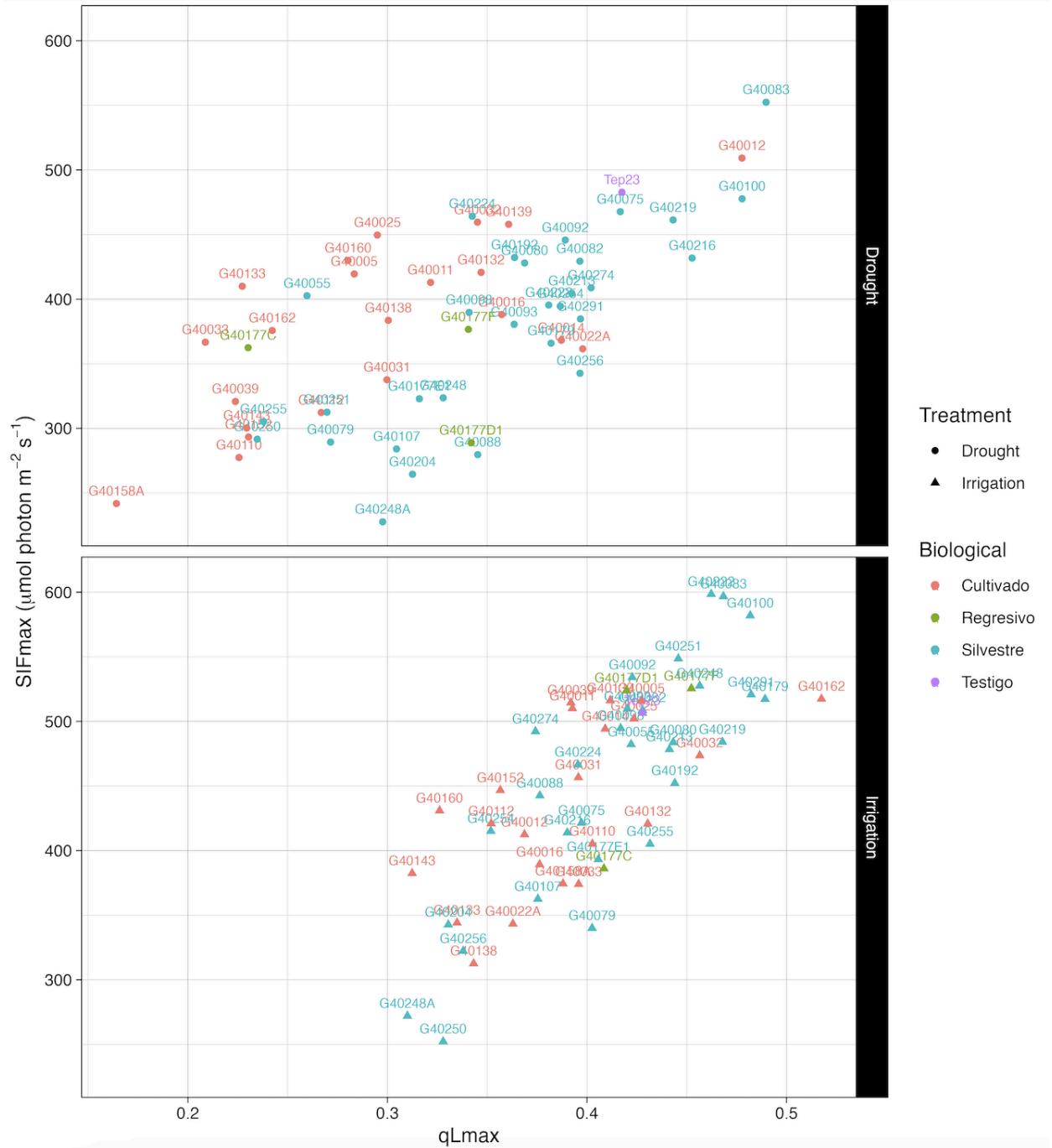


Figure 4.3. Relationship of qL medium with SIF for teary bean wild, weedy and cultivated accessions evaluated under drought and irrigation conditions. The variation of performance for each of the accessions according to photosynthetically active radiation (PAR).

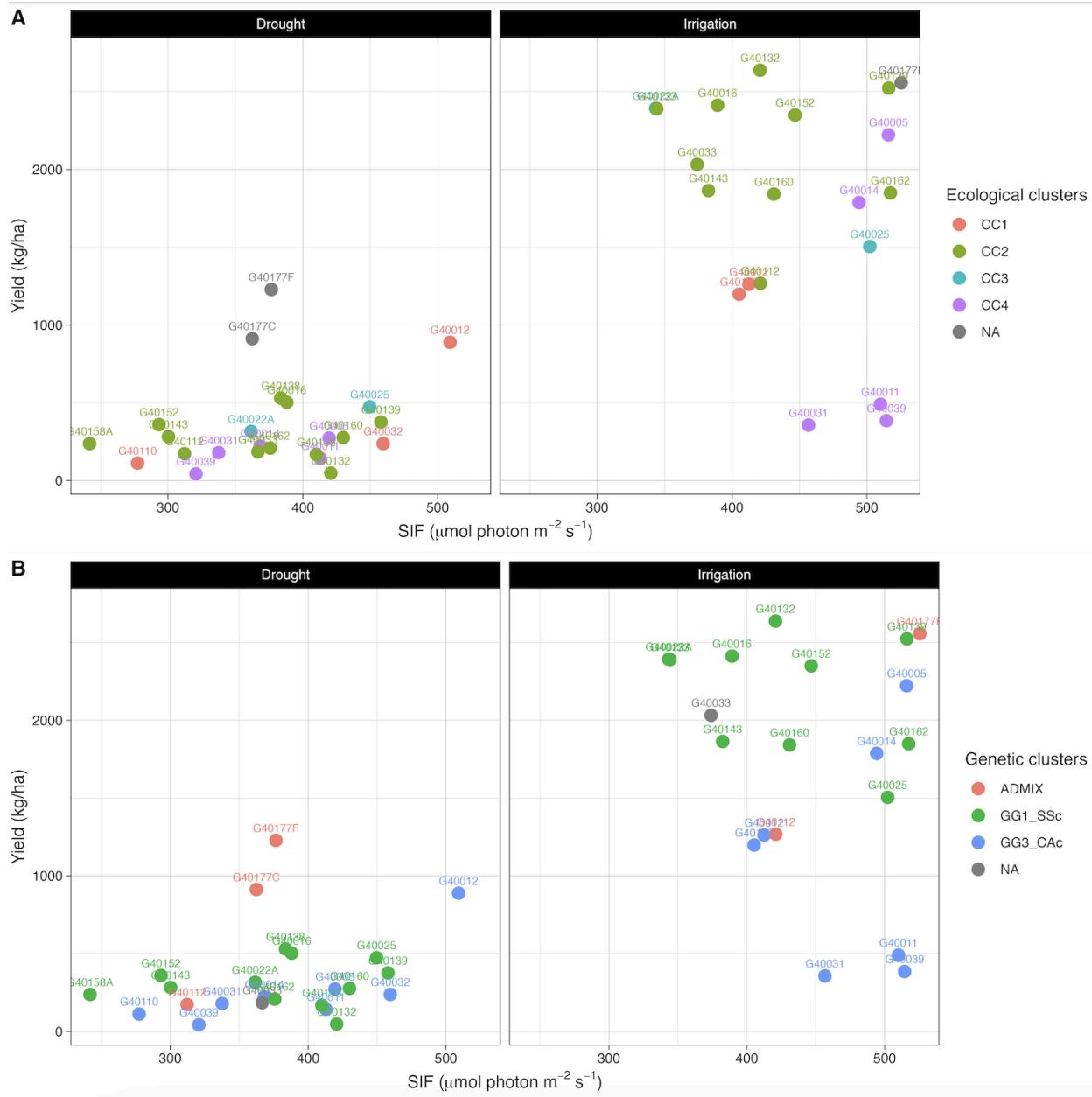


Figure 4.4 Relationship of mean yield to median SIFmax for cultivated tepary bean accessions evaluated under drought and irrigated conditions from the ecological groups. **A)** Relationship of yield and SIF with ecological groups and **B)** relationship of yield and SIF with genetic groups.

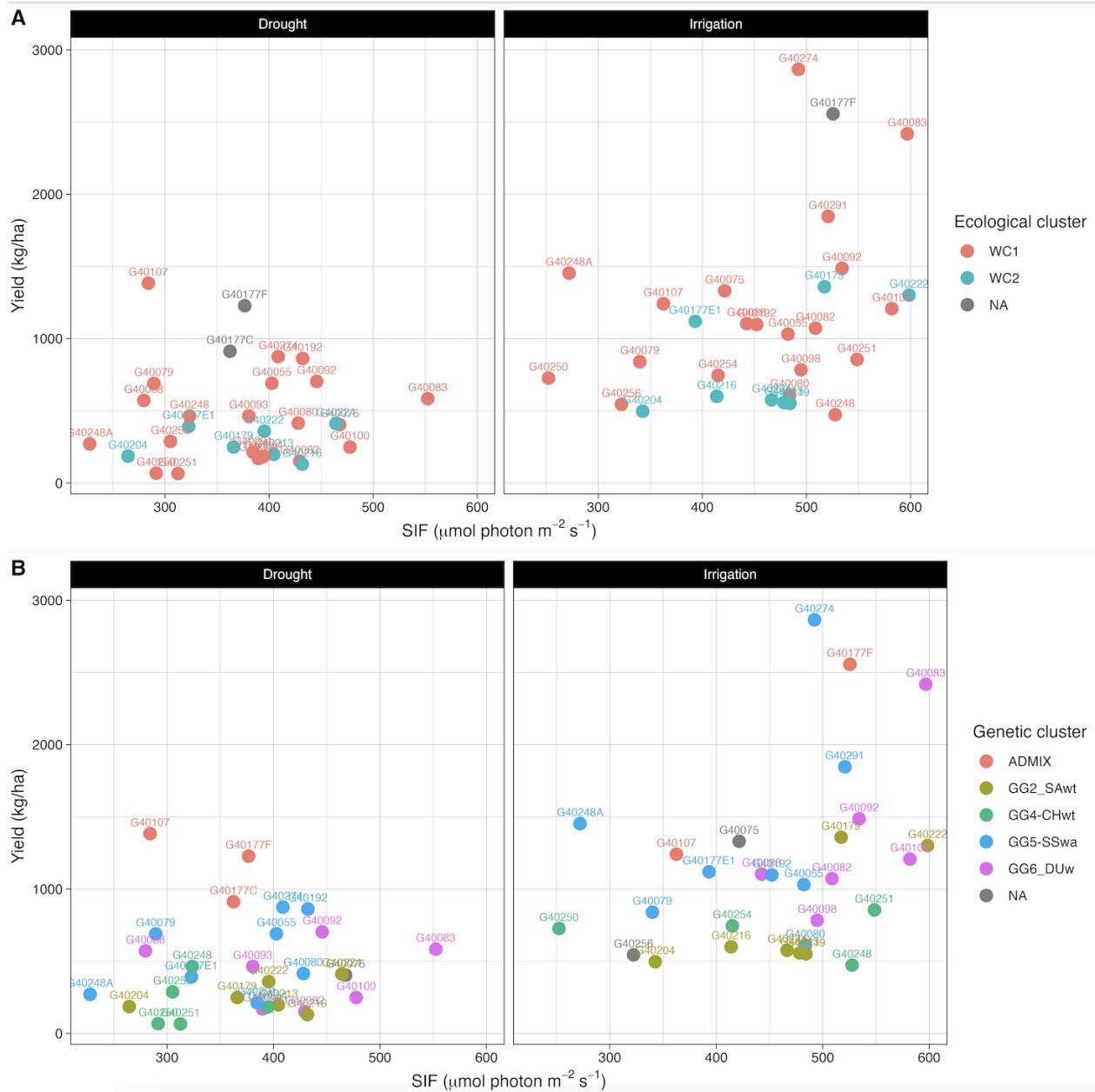


Figure 4.5. Relationship of median yield to median SIF for wild tepary bean accessions evaluated under drought and irrigated conditions from the ecological groups. **A)** Relationship of yield and SIF with ecological groups and **B)** relationship of yield and SIF with genetic groups .

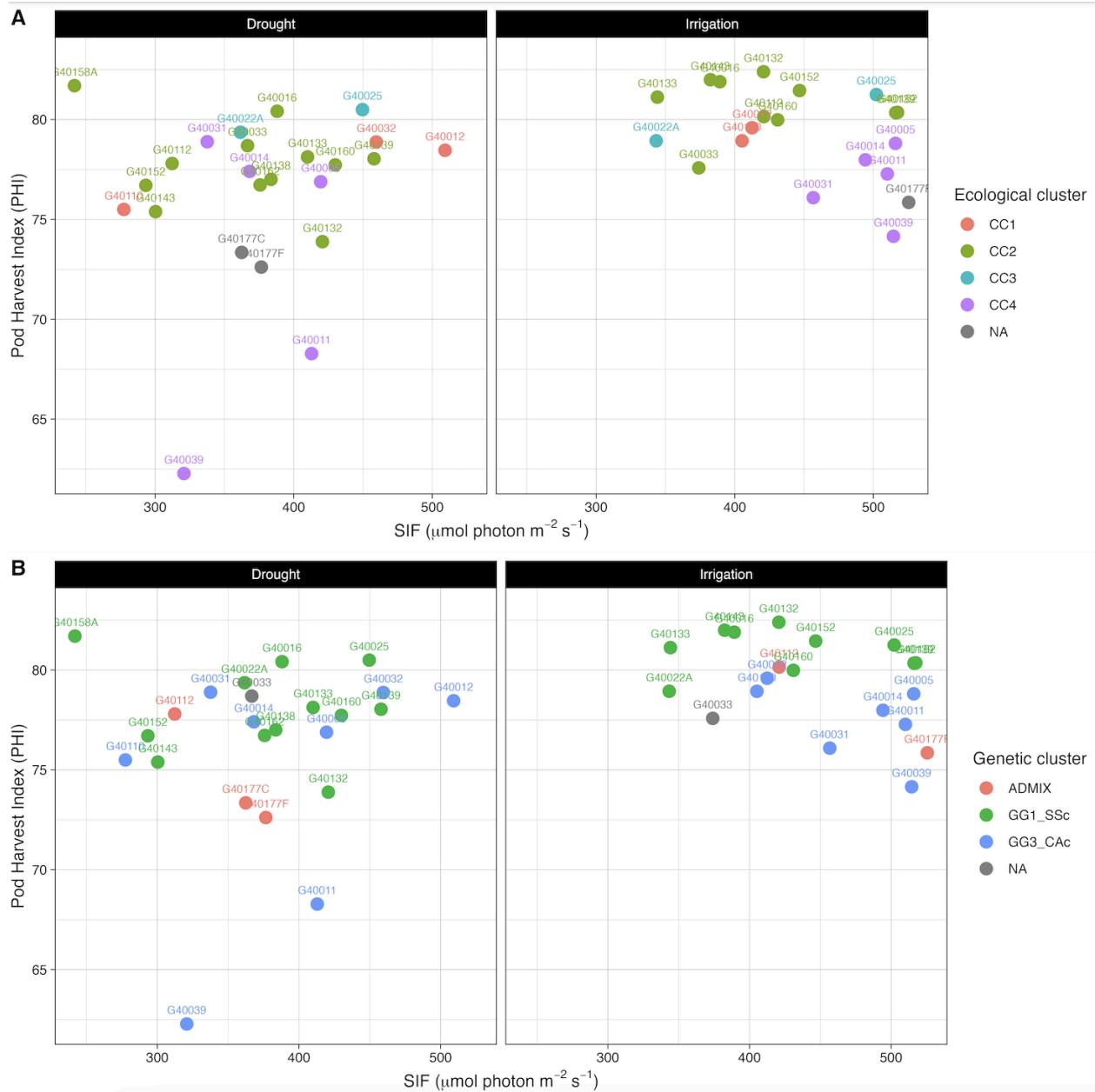


Figure 4.6. Relationship of median pod harvest index (PHI) to median SIF for cultivated tepary bean accessions evaluated under drought and irrigated conditions from the ecological groups. **A)** Relationship of PHI and SIF with ecological groups and **B)** relationship of PHI and SIF with genetic groups.

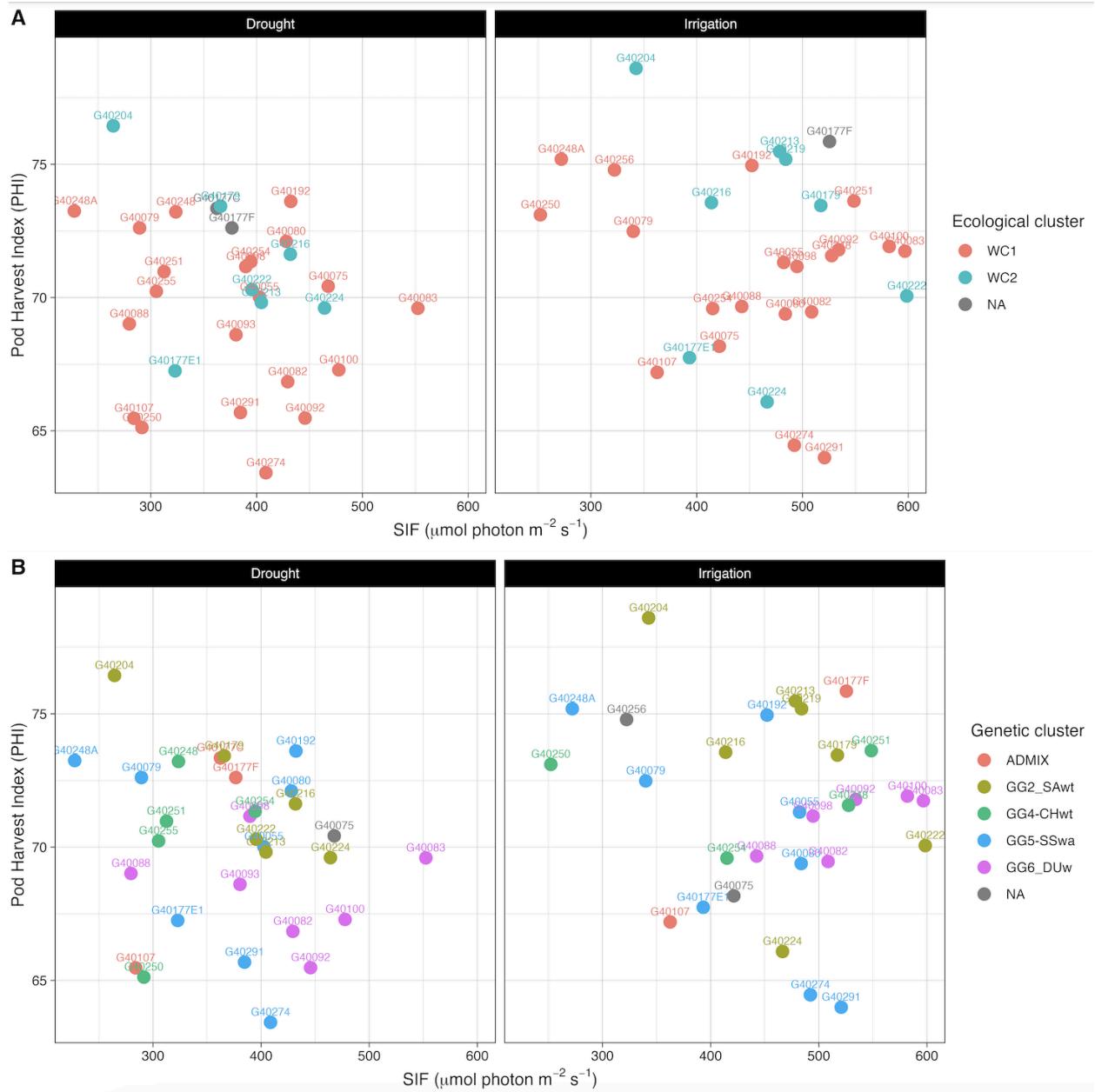


Figure 4.7. Relationship of median pod harvest index (PHI) to median SIF for wild tepary bean accessions evaluated under drought and irrigated conditions from the ecological groups. **A)** Relationship of PHI and SIF with ecological groups and **B)** relationship of PHI and SIF with genetic groups.

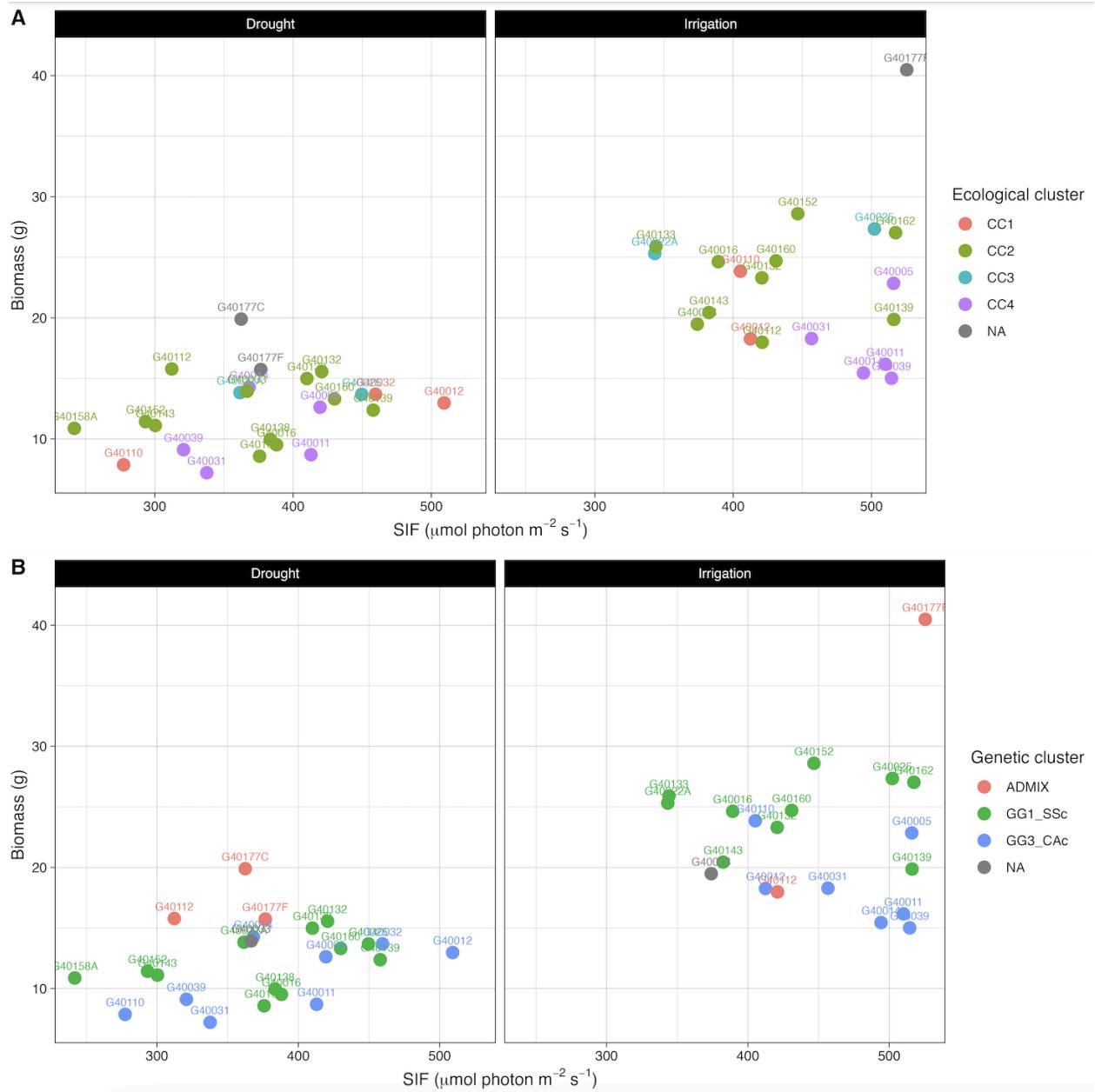


Figure 4.8. Relationship of median biomass to median SIF for wild tepary bean accessions evaluated under drought and irrigated conditions from the ecological groups. **A)** Relationship of biomass and SIF with ecological groups and **B)** relationship of biomass and SIF with genetic groups.

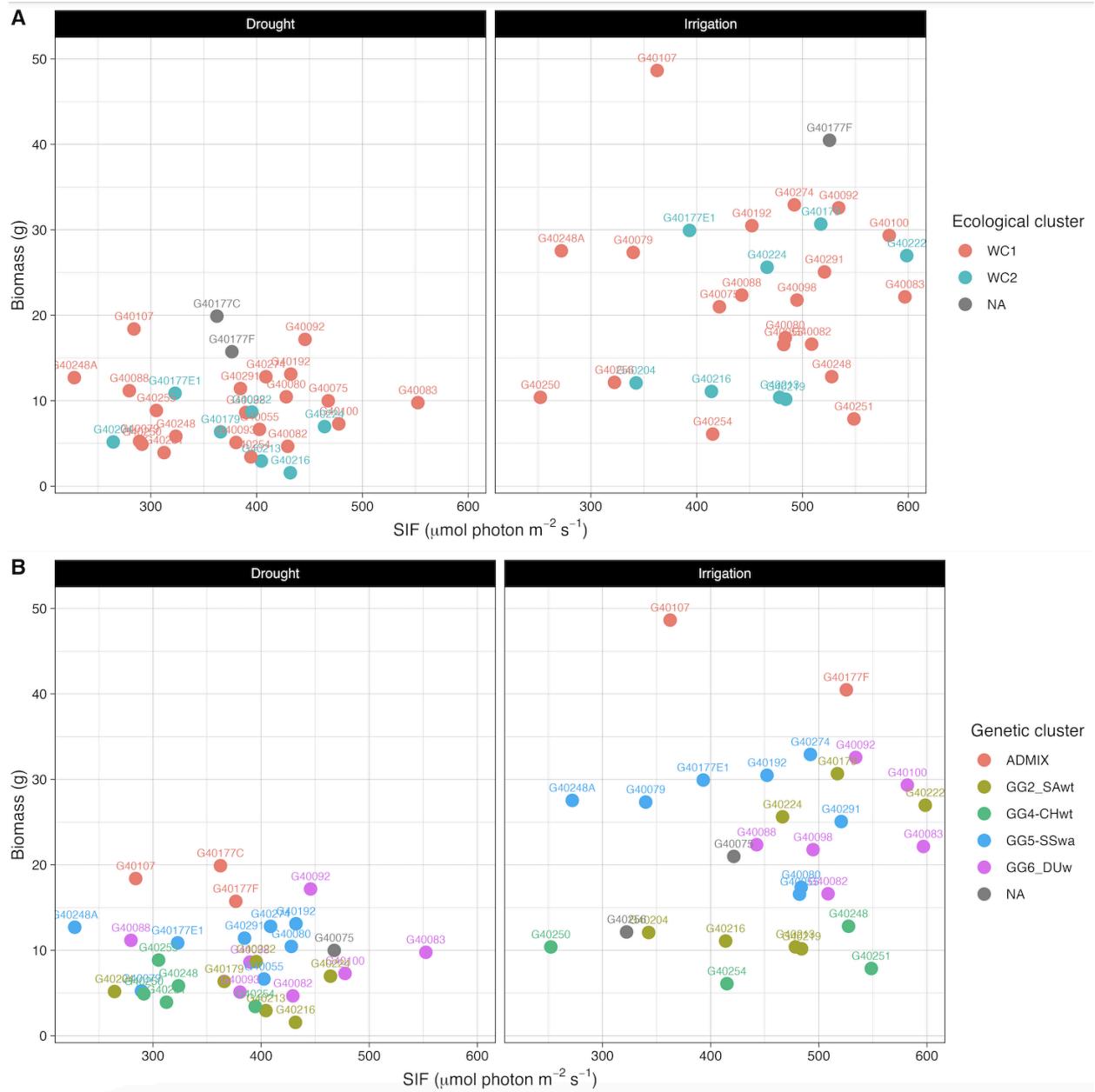


Figure 4.9. Relationship of median biomass to median SIF for wild tepary bean accessions evaluated under drought and irrigated conditions from the ecological groups. **A)** Relationship of biomass and SIF with ecological groups and **B)** relationship of biomass and SIF with genetic groups.

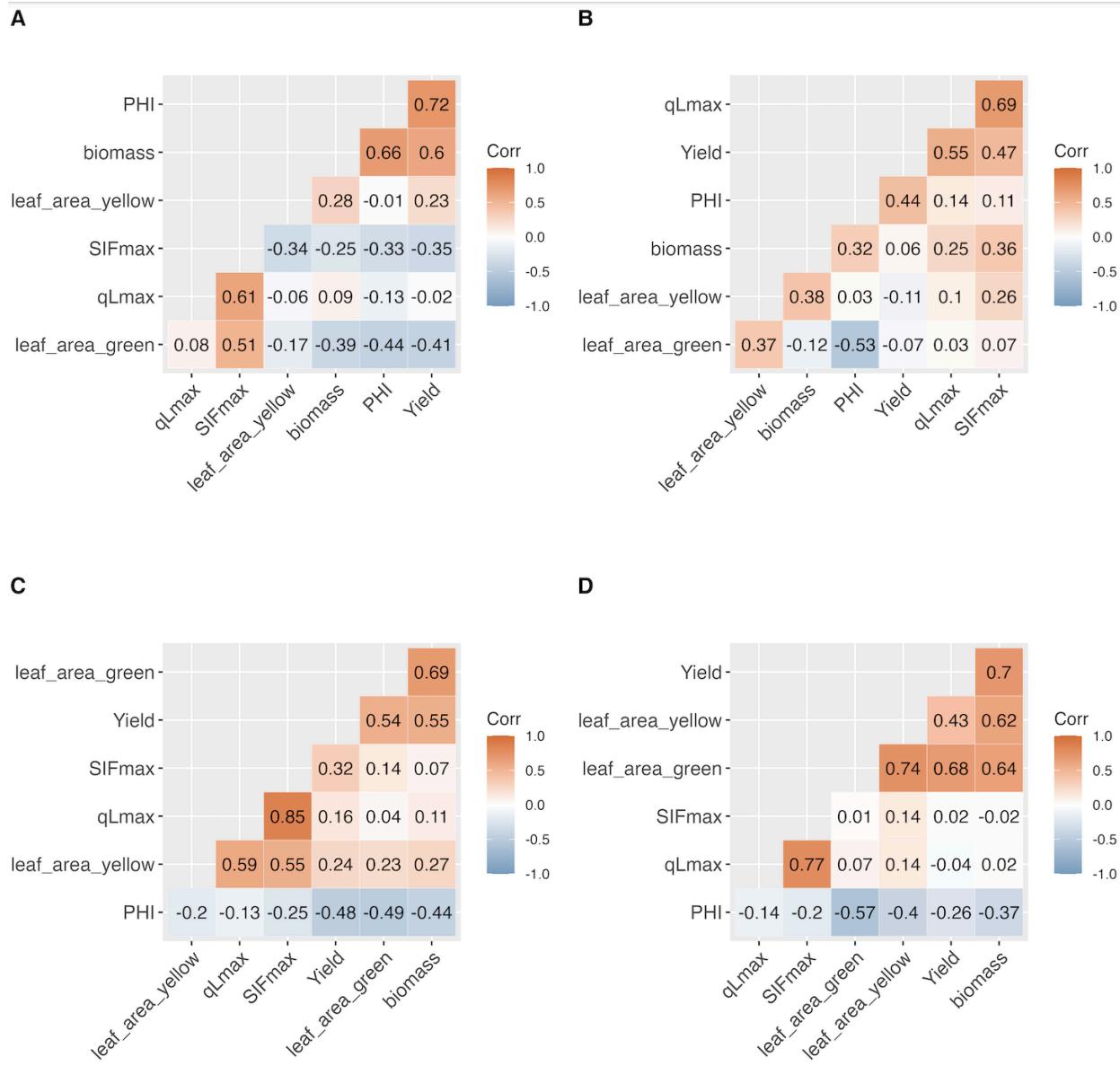


Figure 4.10. Correlations of biomass, SIFmax, leaf area, yield and pod harvest index under irrigation and drought conditions evaluated in the tepary bean functional collection. **A)** Irrigated cultivated accessions; **B)** Drought cultivated accessions; **C)** Irrigated wild accessions; **D)** Drought wild accessions.

5. Chapter five: Using phenomics to identify and integrate traits of interest for better-performing common beans: a validation study on an interspecific hybrid and its Acutifolii parents

Diego Felipe Conejo Rodriguez^{1,2*}, Milan Oldřich Urban^{2*}, Marcela Santaella¹, Javier Mauricio Gereda¹, Aquiles Darghan Contreras³ and Peter Wenzl¹

¹Genetic Resources Program, International Center for Tropical Agriculture (CIAT), Km 17, Recta Cali-Palmira, Valle del Cauca, Colombia

² Bean Physiology and Breeding Program, International Center for Tropical Agriculture, Km 17, Recta Cali-Palmira, Valle del Cauca, Colombia.

³ Universidad Nacional de Colombia, Bogotá, Department of Agronomy, Faculty of Agricultural Sciences, Bogotá Colombia.

* Correspondence:

Milan Oldřich Urban

m.urban@cgiar.org

Diego Felipe Conejo

d.conejo@cgiar.org

This manuscript is published in the journal Frontiers. Plant Sci...Sec. Plant Breeding (for details see <https://doi.org/10.3389/fpls.2022.1008666>).

Abstract

Common bean breeding aims to improve traits of abiotic and biotic resistance by using the genetic diversity of wild and domesticated relatives. Acutifolli relatives (*Phaseolus acutifolius*; *P. parvifolius* ~ *P. montanus*) are adapted to abiotic and biotic stresses and have been used in developing interspecific hybrids in common bean breeding programs. Evaluations of hybrids are limited, as classical genebank accession descriptors are semi-subjective, have qualitative traits and show complications when evaluating intermediate accessions. However, descriptors can be quantified using recognized phenomic traits. This digitalization can identify phenomic traits which correspond to the percentage of parental descriptors remaining expressed/visible/measurable in the particular interspecific hybrid. In this study, a line of *P. vulgaris*, *P. acutifolius* and *P. parvifolius* accessions and their crosses were sown in the mesh house according to CIAT seed regeneration procedures. Three accessions and one derived breeding line originating from their interspecific crosses were characterized and classified by selected phenomic descriptors using multivariate and machine learning techniques. The phenomic proportions of the interspecific hybrid (line INB 47) with respect to its three parent accessions were determined using a random forest and a respective confusion matrix. The seed and pod morphometric traits, physiological behavior and yield performance were evaluated. In the classification of the accession, the phenomic descriptors with highest prediction force were Fm', Fo', Fs', LTD, Chl, seed area, seed height, seed Major, seed MinFeret, seed Minor, pod AR, pod Feret, pod round, pod solidity, pod area, pod major, pod seed weight and pod weight. Physiological traits measured in the interspecific hybrid present 2.2% similarity with the *P. acutifolius* and 1% with the *P. parvifolius* accessions. In addition, in seed morphometric characteristics, the hybrid showed 4.5 % similarity with the *P. acutifolius* accession. Here we

were able to determine the phenomic proportions of individual parents in their interspecific hybrid accession. After some careful generalization the methodology can be used to: i) verify trait-of-interest transfer from *P. acutifolius* and *P. parvifolius* accessions into their hybrids; ii) confirm selected traits as “phenomic markers” which would allow conserving desired physiological traits of exotic parental accessions, without losing key seed characteristics from elite common bean accessions; and iii) propose a quantitative tool that helps genebank curators and breeders to make better-informed decisions based on quantitative analysis.

Keywords: phenomic proportions, image analysis, random forest, germplasm characterization.

5.1 Introduction

Genebank plant genetic resources comprise the representative diversity of genetic material contained in traditional varieties and modern cultivars, as well as in the crop wild relatives and other wild plant species that can be used now or in the future for food and agriculture (Wang and Zhang, 2011). Currently, there are about 1,750 genebanks worldwide that conserve 7.4 million accessions of agricultural genetic resources (Noriega *et al.*, 2019). Eleven CGIAR genebanks conserve about 730,000 accessions among crops, trees, and forages, of which the International Center for Tropical Agriculture (CIAT) conserves 37,987 bean accessions, 23,140 forage accessions and nearly 6,000 cassava accessions (Noriega *et al.*, 2019). Despite this great diversity, only approximately 10% of the accessions from the 1,750 genebanks is used in plant breeding, mainly because of poor phenotypic and genotypic characterization or lack of agronomic traits evaluation (De Carvalho *et al.*, 2013; Tadesse *et al.*, 2019; Nguyen and Norton, 2020; Kholova *et al.* 2020).

P. acutifolius (tepary bean) is an important species in common bean breeding, due to its adaptation to abiotic and biotic stress (Singh and Munoz, 1999; Porch *et al.*, 2013; Kusolwa *et al.*, 2016). The use of cultivated and wild relatives of *P. vulgaris* by the common bean breeding program at CIAT started in the 1980s, with the aim of generating lines with elevated levels of introgression from *P. acutifolius* and/or *P. parvifolius* ~ *P. montanus* (Debouck, 2021), using techniques such as congruity backcrossing (CBC) and recurrent backcrossing (RBC) (Haghghi and Ascher, 1988; Mejia-Jimenez *et al.*, 1994; Singh *et al.*, 1998) with the help of bridge genotypes.

Classification of hybrids based on phenotypic traits was done in the 1960s (Allendorf *et al.*, 2001), however, the detection of morphological traits usually assumes that hybrids are phenotypically intermediate to the parents. This is often not the case, because hybrids express a mosaic of parental phenotypes (Arnold, 1997) influenced also by environmental conditions. Furthermore, morphological characters do not allow determining whether an individual is a first-generation hybrid (F1), a backcross or late generation hybrid (Allendorf *et al.*, 2001).

Recently, there have been published studies that promote characterization processes using phenomic descriptors. When compared with conventional descriptors, these showed a better capacity for analysis of phenotypic variability (Rosero *et al.*, 2019; Nankar *et al.*, 2020). Phenomic descriptors has both qualitative and quantitative characters and deal with agronomic, morphological, physiological, and colorimetric traits of accessions which are captured by proximal sensors such as cameras, fluorometers, trichromatic, multispectral and hyperspectral sensors. Phenomic descriptors have a “high-throughput” character of data, which means, hundreds of accessions can possibly be characterized/screened in a reasonable time. However,

the sheer volume, variety and veracity of imagery and remote-sensing data still present limits in data analysis (Singh *et al.*, 2016).

To solve this problem, there are currently several machine learning models, such as partial least squares (PLS), random forest (RF), support vector machines (SVM), and neural networks (NN) used in phenomics data analytics (Araus *et al.*, 2022). Based on phenomic traits, classification of accessions and their hybrids has been attempted. This has been made possible with the development of phenomic and machine learning methods with thousands of data points from each individual, (Parmley *et al.*, 2019; Soltis *et al.*, 2020; Feldmann *et al.*, 2020, Henao-Rojas *et al.*, 2021).

In this work, we propose a Phaseolus-oriented methodology for the detection of phenomic proportions of interspecific hybrids with respect to their parents. We used multivariate and machine learning methods to characterize and classify three parental line accessions (a cultivated *P. vulgaris*, a domesticated *P. acutifolius*, and a wild *P. parvifolius* - *P. montanus*) with its interspecific hybrid accession.

The phenomic proportions correspond to the percentage of parental descriptors which remain expressed/visible/measurable in the particular interspecific hybrid. Correspondingly, in this study, phenomic proportions show the phenomic traits portions quantified in the three parental lines and verified in an interspecific hybrid. We hope this methodology will provide a first step to help genebank curators, breeders, physiologists and others to i) make detailed quantitative comparisons of selected phenomic traits between accessions of interest, and ii) better manage and understand their genetic resources. After some generalization using other parents/hybrid collections or random selection, this methodology could facilitate a deeper understanding about

i) crossings, heritability and breeding success; ii) functional trait diversity; iii) species domestication/evolution and genetic recombination; and iv) how to substantially increase genetic gains (in tandem with genomics).

5.2 Materials and methods

5.2.1 Plant material and experimental design

Materials used in this study were *P. acutifolius* (G40001 – CIAT genebank accession number), *P. parvifolius* - *P. montanus* (G40102), *P. vulgaris* (G5773) and their interspecific cross hybrid line INB 47 (G52443), all obtained from the bean collection at CIAT Genebank. Accessions G40001 (*P. acutifolius*) and G40102 (*P. parvifolius* - *P. montanus*) display a type III growth habit (indeterminate prostrate growth); while accessions G5773 and hybrid G52443 exhibit a type I growth habit (determinate bush growth; growth habits defined according to Fernández *et al.*, 1982). Accession G40001 showed heat and drought resistance (Mejía-Jiménez *et al.*, 1994) and G40102 was highly resistant to common bacterial blight (CBB) (Singh *et al.*, 1998).

The interspecific hybridization was a product of artificial crossings, carried out by the CIAT common bean breeding program, from crosses between a popular commercial bean variety Ica Pijao, *P. parvifolius* (G40102) and an interspecific line, with five cycles of congruity backcrossing (CBC₅) between Ica Pijao and *P. acutifolius* (G40001) (Mejía-Jiménez *et al.*, 1994). The pedigree of the interspecific hybrid line is as follows: INB 47 (G52443) = ICA PIJAO x (G40102 x (ICA PIJAO x (G40102 x (ICA PIJAO x (ICA PIJAO x (ICA PIJAO x G40001; CBC₅).

The INB 47 line was developed from ten (10) cycles of selection pressure based on commercial characteristics including growth habit, seed type and yield in the experimental station in Santander de Quilichao and in Palmira (both sites in Colombia), CIAT (Personal communication, Common bean breeding program, CIAT). The selection was mainly focused on conserving the commercial seed type similar to *P. vulgaris* – Ica Pijao, while introducing resistance to bacterial diseases (Mejia-Jimenez *et al.*, 1994). The phenotypic characteristics of the parental lines and the interspecific hybrid are shown in Figure 5.3. These characteristics were observed during the experiment at the regeneration station of CIAT's genetic resources program (GRP) in Palmira.

Our experiments were performed in two periods: from October 2018 to January 2019, and from January 2019 to April 2019, in a mesh house at CIAT, Palmira, Colombia ($3^{\circ}30'17''$ N, $76^{\circ}21'24''$ W, 950 masl). The cultivation protocol was used according to standard operating procedures used in CIAT genebanks when multiplying accessions. The experimental conditions inside the mesh house presented: i) an average daily temperature of 38°C , with a daily maximum temperature of 41°C at midday hours, and daily minimal temperature of 27°C ; ii) a minimum relative humidity of 31% and a maximum of 65%; and with iii) an average photosynthetically active radiation (PAR) of $1680 \mu\text{mol m}^{-2} \text{s}^{-1}$ during the sampling of physiological variables (please, explore variables at: <https://photosynq.org/projects/domestication-syndrome/explore>); iv) 12 h of natural light, and v) conventional agronomic management and fertigation using drip irrigation. The plants were sown in a substrate of coconut fiber substrate for hydroponic systems ($120 \times 40 \times 40$ cm), composed of 100 % peat, which is highly efficient in conserving water and guarantees health development in the early stages of cultivation (elaborated in Spain by the company Berger, <https://www.berger.ca/en/horticultural-products/>). Water irrigation and fertilization scenarios were the same for all plants (Supplementary Material Table 1 - Fertilization).

Each plant was tutored with agricultural mesh, starting at germination, to minimize human intervention in the span of plant development and guarantee adequate growth. The experimental design was a complete randomized block design, in which each accession acted as a treatment and each plant as an experimental unit. Five (5) plants per accession were planted, and three (3) independent technical repetitions measured in each period. To evaluate differences between accessions with special focus on the hybrid, we decided to compare morphometric, physiological, and agronomic traits.

5.2.2 Morphometric descriptors

We evaluated the morphometric aspects of pod shape and lateral seed shape. Seed morphological characterization was carried out using phenomic descriptors based on image analysis, including the following traits: seed Area (cm^2), Perimeter (cm), Width (cm), Height (cm), Major, Minor, MinorFeret, MajorFeret, Aspect Radio (AR), Circularity, Roundness, and Solidity according to Schlautman *et al* (2020) and Rosero *et al* (2019) (see details in Supplementary Material Table 2 – Morphometric descriptions).

The digital images of each of the accession's pod/seed were obtained using Canon SX60 HS camera with a digital resolution of 16.1 megapixels, an image area of 1080×1080 pixels in an automatic format. We captured images in a RAW format to ensure maximum image quality. Images were taken separately for seeds and pods. Each picture was processed using the DCRAW plugin of ImageJ 5.4 (<https://imagej.net/software/fiji/>).

After plant harvest, images were captured from pods and seeds of every accession, considering the level of luminosity and the contrast of the background. To capture color of the seeds, we used

a contrasting background and standardized color scale 24ColorCard Camera Trax Card-3x5 (Supplementary Material Figure 5.1 – Color scale and photobox used). Images of the pod shape and lateral shape of the seeds were processed using ImageJ software (Eliceiri *et al.*, 2012), following the protocol: (I) Settlement of the scale (pixels to cm), (II) Image binarization using the Max Entropy and Huang methods (Huang and Wang, 1995), (III) Definition of regions of interest (ROI) from pod and seed selections, and (IV) Extraction of morphometric measurements of each of the selected ROIs (Supplementary Material Figure 5.2 – imagen analysis process).

5.2.3 Physiological descriptors

For physiological measurements, we used the MultispeQ device (Kuhlgert *et al.*, 2016; PhotosynQ, USA). The device includes climatic and plant variables that facilitate characterizing the physiological performance of plants in their environment. All data were captured at midday between 11:00 am and 1:00 pm, with the aim of comparing the data collected under more stable temperature and lower air humidity (RH) under mesh house conditions. Measurements were taken three times a week in three technical (plant as experimental unit) repetitions per accession in all three replicates and both periods. In total, 1,022 MultispeQ observations were captured for the four accessions. Samplings were carried out every 15 days from December 22, 2018 to April 23, 2019. Samples were taken during phenological stages 22 to 85 according to the BBCN scale from plant branching until the harvest maturity (Feller *et al.*, 1995). The classical protocol was used: Leaf Photosynthesis MultispeQ V1.0 (the raw data are available at: <https://photosynq.org/projects/domestication-syndrome>; ID 5685).

Briefly, the MultispeQ proximal sensors measure photosynthetic parameters including: i) quantum yield of photosystem II (Φ_{II} – Phi2); ii) non-photochemical quenching (Φ_{NPQ} -

PhiNPQ); iii) energy losses for heat dissipation (Φ_{NO} - Phi $_{NO}$); iv) relative chlorophyll (Chl); v) linear electron flux (LEF); vi) leaf temperature differential (LTD); vii) maximum variable fluorescence at a steady-state conditions (Fm'); viii) minimum variable fluorescence during dark phase after a steady-state (Fo'); ix) variable fluorescence at a steady-state conditions (Fs'); x) efficiency of open reactions centers in the light (Fv'/ Fm'); xi) fraction of open PSII centers when QA is oxidized (qL); xii) photochemical quenching relating PSII maximum efficiency (qP); xiii) fluorescence decrease ratio (RFd), and xiv) leaf thickness (Kuhlgert *et al.*, 2016; Fernández-Calleja *et al.*, 2020; Deva *et al.*, 2020). In addition, the device also measures environmental conditions like light intensity (photosynthetically active radiation, PAR), air temperature and air humidity (see Supplementary Material – Figure 5.3 for temperature range, humidity and PAR). The data acquired can be visualized on the PhotosynQ platform (i.e. exploratory analysis of the data). Thus, at air temperatures above 32°C, photosynthetic activity is restricted in common bean of determinate growth type (Beebe *et al.*, 2011; Deva *et al.*, 2020).

5.2.4 Yield components

To calculate seed yield, ten pods were taken randomly from each of three similar plants per accession during the final harvest at BBCH 89. Seeds were pre-dried according to the Genetic Resources Program methodology (Salazar *et al.*, 2020). Seed weight was measured with high precision scales with seed average humidity of 14%. The dry weight of seed at harvest (PSW), weight of pod with seed at harvest (PW), dry weight of pod without seed (pod walls) (VW), number of seeds (SN) and the harvest index at the pod level (PHI = ((Dry weight of seeds at harvest)/(Dry weight of whole pod at harvest))x100) were determined.

5.2.5 Multivariable analysis and Phaseolus accessions classification

Initially, we performed an outlier detection test using the Dobin library of R software (Kandanaarachchi and Hyndman, 2021). Principal Component Analysis (PCA) was carried out first on the parents, and the Principal Components (PC) with the highest contribution to the explained variance were extracted in each characterization group for each variable (performed using the FactorExtra library of R; Kassambara and Mundt, 2017). Predictor analysis was performed by random forests (randomForest library of R; Liaw and Wirner, 2018). Classification was performed on 100 trees, using 70% of the data for tree training, and 30% of the data for validation. For evaluating the classification model prediction with “out of bag” accuracy (OOB accuracy). The OOB is an error estimation technique used to evaluate the accuracy of the random forest (Janitz and Hornung, 2018). The OOB estimates accuracy across all classes (values above 1 - 10% are estimated as high accuracy; Kennedy *et al.*, 2015).

The evaluation metric and confusion matrix were determined to observe the phenomic proportions of parent classification for each characterization group. The descriptors selected for each characterization component are determined by mean decrease in accuracy and the gini decrease index as parameters for feature selection. Analyses were run in R using the library “randomForestExplainer” and “caret” library (Paluszynska, 2017).

5.2.6 Phenomic proportions of the interspecific hybrid with respect to its parents

Initially, the contributions of the PCs from the classification of the parent lines were used for weighting the phenomic descriptors of the interspecific hybrid. Subsequently, the weighted

phenomic descriptor values of both parents and the interspecific hybrid are standardized to values between 0 - 1. The phenomic proportions are determined from classifying parents and an interspecific hybrid using random forests. Phenomic descriptors of importance in the classification will be determined for each characterization component using the gini index and mean decrease accuracy. The prediction of the confusion matrix in the interspecific hybrid will be considered as the phenomic proportions that it presents with respect to each of its parents. A confusion matrix is typically created representing the summary of the number of correct and incorrect prediction results broken down by each parental line .

Finally, a non-parametric multivariate analysis of variance (MANOVA) was performed to determine if there are significant differences between the parents and their hybrids in each of the characterization components with the already prioritized descriptors. In our study we used MANOVA developed by Friedrich and Pauli (2018) which allows flexibility of normality assumptions and incorporates general heteroscedastic designs and potentially singular covariance matrices. It also improves the performance of small samples through bootstrap techniques. The analysis was performed using 10,000 iterations, modified ANOVA-type statistics (MATS), and the p-resampling value was determined from the parametric bootstrap approach (paramBS). MANOVA was performed for each characterization component separately (physiology, pod morphometry, seed morphometry and yield). In order to observe the significant differences between parents and its hybrid, the *post hoc* Tukey multivariate test was performed. This was done using the MANOVA.RM library of the free software R. The summary of the data analysis procedure can be seen in Figure 5.1 and Figure 5.2.

5.3 Results

The three parental accessions used in this study show contrasting phenotypic differences in morphological characteristics of the leaf shape, growth habit, flower shape and color, and seed size (Figure 5.3). *P. acutifolius* (G40001) (Figure 5.3A) presents an oval-lanceolate leaf shape and acute angles (less than 90 degrees) in conditions of high light intensity at midday hours. The flowers and seeds of *P. acutifolius* are white, with a straight pod shape, and a predominantly round-oval seed shape. *P. parvifolius* (G40102) (Figure 5.3B) is a wild accession with a lobed leaf shape, a light-purple flower color, and dark-purple and curved pods. *P. parvifolius* seeds are small, black to mottled grayish black, with a flattened truncated shape. *P. vulgaris* - Ica Pijao (G52443) (Figure 5.3C) has an oval leaf shape, dark purple flowers, mottled purple pods with slightly curved pod shape, and black seeds with a flattened oval shape. The interspecific hybrid (G52443 - INB 47) (Figure 5.3D) has a lanceolate leaf shape and acute leaf angles under high light intensity. It has purple flowers, with some flowers showing malformations in the floral wings. The pods are slightly curved with mottled purple colors, the seeds are black, and the seed shape is round cuboid.

The classification of the parent accessions using random forests, determined as predictive descriptors for seed shape were: Major, Area, Minot, Height and MinFeret; while in pod morphometry the predictive descriptors were: Major, Feret, Round, Aspect Radio (AR), Solidity and Area, presenting corresponding values of gini index higher than 40 and accuracy decrease 0.06 (Supplementary Material Figure 5.4AS and 5.4CS). The physiological variables, the phenomic descriptors Fo', LTD, Fs' and Fm', and in the yield components VW (Valve weight), PW (Pod weight) SPW (Seed pod weight) and PHI (Pod harvest index), presented gini index

values higher than 35 and accuracy decrease higher than 0.075 (Supplementary Material Figure 5.4AS and 5.4CS). The table of the contributions of the PCs to each of the predictive descriptors is shown in supplementary material Table 3. The classification of the parents can be seen in the confusion matrix (Supplementary Material Figure 5.5 Matrix confusion parents' accessions).

5.3.1 Phenomic proportions of the hybrid respect to its parents

The low OOB value of 5.68% indicates very small classification mistakes and significant differences in the seed shape of our hybrid with respect to its parents in the evaluated phenomic descriptors. Interestingly, in the case of pod morphometric descriptors and physiological descriptors, the OOB values reached 22.44% and 36.76%, respectively; while 8.49% for the yield component (Supplementary Material Table 4 - OOB error).

The phenomic descriptors of importance in the classification of the parental accessions and the hybrid are presented in Figure 5.4. Generally, the descriptors of seed morphometry Area and Minor were the most important, while in pod morphometry the descriptors Major, Feret and AR presented accuracy higher than 0.250 (Figures 5.4A and 5.4C). In the physiological descriptors, Fm' and Chl are the most important, and in the yield components the descriptors PHI and PW are the ones that presented the highest values of accuracy decrease with values higher than 0.125 (Figures 5.4B and 4D).

Using the predictions in the confusion table, the phenomic proportions of each of the characterization components were determined (Figure 5.5). The relationships of the predictions of the hybrid with its parents are most closely related to the common bean parent *P. vulgaris* - Ica

Pijao (G52443). The values of proportions are as follows: 5.2% for physiological traits, 9.8% for pod morphometric traits, and 4.1% for yield components.

Despite the hybrid's high relatedness to its *P. vulgaris* parent accession (the effect of multiple back-crosses), the hybrid presents phenomic proportions of 2.2% also with *P. acutifolius* (G40001) and 1% with *P. parvifolius* (G40102) accessions in physiological descriptors (Figure 5.5B), while 4.5% with *P. acutifolius* in seed morphometrics (Figure 5.5A), 0.6% for yield and 4.9% for pod morphometrics indicating successfully inherited traits from these parents as well (Figure 5.5C).

Phenomic proportions of the physiological components showed trait discrimination (difference) between the parents in traits like the Chl, Fm', Fo', LTD and Fs', respectively. Furthermore (Figure 5.4B), it was observed that the parental *P. parvifolius* shows contrasting physiological behavior when compared with the other two parents. This can be explained by its wild origin and different morphometric characteristics. The interspecific hybrid is closely related to the *P. vulgaris* accession and also the *P. acutifolius* accession, indicating that it conserves physiological characteristics mainly from these two parents. However, the hybrid presents higher phenomic proportions of the *P. vulgaris* accession and not of the *P. acutifolius* accession (Figure 5.5B). For seed morphometry, the confusion matrix clearly separates the interspecific hybrid and the parents with high precision, showing the hybrid has unique characteristics in seed morphometry, despite sharing 4.5% of phenomic proportion with *P. acutifolius* (Figure 5.5A). In pod morphometry, it is clear that there is a difference of our hybrid regarding its *P. acutifolius* and *P. parvifolius* parental accessions (Figure 5.5C). This supports the fact that there is no clear trait separation between the *P. vulgaris* parental and the interspecific hybrid and again is likely an influence of multiple back-crossing and/or environment-based limitations of *P. acutifolius*-related traits.

The data after MANOVA fitting, supported the rejection of the statistical hypothesis associated with shared characteristics between the hybrid and its parents (Table 5.1). In the physiological characterization (MultispeQ data), there were no significant differences with the parentals *P. acutifolius* and *P. vulgaris* accession; while with *P. parvifolius* accession differences were highly significant (< 0.001) (Table 1). In pod morphometry, the interspecific hybrid showed no differences with *P. acutifolius* and *P. vulgaris* accessions. In seed morphometry, the hybrid showed no differences with the parental *P. vulgaris*, while in the yield components it showed no differences with *P. parvifolius* and *P. vulgaris* (Table 1) accessions. The MANOVA supports the statistical differences of the hybrid and its parents in each characterization component, contrasting with those obtained in the random forest analysis.

In addition, it is observed that using seed morphometric descriptors, the interspecific hybrid shows differences from the three parents, indicating characteristics of the interspecific hybrid determined probably by the hybrid vigor.

5.4 Discussion

High-throughput phenotyping methods can facilitate the use of genetic resources by estimating phenotypic traits of importance and identifying accessions of interest for pre-breeding and breeding programs (Nguyen and Norton, 2020). In this work, we explored phenomic descriptors that help discriminate selected *Phaseolus* accessions with their interspecific cross, using components based on physiological descriptors, seed and pod morphometrics and yield components.

Despite being crop relatives of *P. vulgaris*, Acutifolii species (*P. acutifolius* and *P. parvifolius*) have contrasting leaf, seed, and pod phenotypic characteristics (Figure 5.3). In addition, there are natural differences between domesticated and wild accessions (Mwale *et al.*, 2020). The domestication syndrome of the Phaseolus genus is characterized by a reduction in pod shattering and increase in seed size, being these the most important traits in the adaptation of domesticated populations (Chacón-Sánchez, 2018). This explains the differentiation in our data for seed and pod morphometrics and reveals why they can serve as the most significant traits in quantification of phenomic proportions between the studied hybrid and its parents and under some generalization and verification can be used for a wider spectrum of hybrid evaluations.

Each of the accession's classifications contained several phenomic descriptors that contributed to defining/ identifying its uniqueness. In pod and seed morphometry, it is observed that the descriptors with highest contributions (Figure 5.4A and 4C), such as seed/pod Area, pod Feret, seed Height, seed MinFeret, seed Minor and seed/pod Major, are descriptors directly related to the organ (seeds or pod) size; while Solidity shows that the pod appearance is influenced by its curvature and also by the shape of the seed. *P. parvifolius*, being a wild accession, does not present domestication syndromes (Chacón-Sánchez, 2018). This is evidenced by the pod and seed small size, being the primary discriminating descriptors in classifying the parvifolius accession. Both studied *P. acutifolius* and *P. vulgaris* have larger pod and seed sizes, likely due to the selection pressure of preferable domestication syndromes (Chacón-Sánchez, 2018). The domestication process directly influences pod and seed weights and can increase pod harvest index (PHI) (Rao *et al.*, 2013). Interestingly, *P. acutifolius* generally has smaller seed size (Freytag and Debouck, 2002) but higher PHI compared to *P. vulgaris* (Rao *et al.*, 2013). Higher PHI has strong heritability, is easily measured, and is related to drought resistance and low soil

fertility tolerance. The *P. parvifolius* accession also lacks pod shape curvature, allowing simple visual differentiation and classification between accessions.

The physiological descriptors are extremely useful for accession classifications (Figure 5.5B), although less comparable to morphometric aspects. However, the lower weight of physiological traits is understandable when considering all physiological descriptors (in our case we measured photosynthetic/fluorescence traits and leaf-based data by MultispeQ) as greatly influenced by the environment, with possibly limited heritability and biologically relevant biochemical acclimation thresholds (resistance). The physiological traits inherently hold considerable genetic complexity, considering their crucial role in plant development and survival (Reynolds and Langridge, 2016). Alternatively, it is possible that *P. acutifolius* shows similarly reduced physiological behavior as *P. parvifolius* in the conditions where experiments were done. Although *P. acutifolius* and *P. parvifolius* are two different species (Buhrow, 1983; Schinkel and Gepts, 1989), *P. acutifolius* var. *latifolius* has been reported as an intermediate species between the domesticated *P. acutifolius* and wild *P. parvifolius*. This suggests that both studied accessions may share similar genetic background (Freytag and Debouck, 2002; Muñoz *et al.*, 2006; Blair *et al.*, 2012) and thus some physiological performance as mentioned above.

Regarding the physiological descriptors, it is clear that leaf components as Chl, Fo', Fs', Fm', and LTD (supplementary material Figure 5.4B) present the highest contributions in the differentiation and classification between lines. Both *P. acutifolius* and its wild relative *P. parvifolius*, have similar physiological responses most likely due to the similar ecogeographic distribution of both species, associated with the arid areas of southern USA and northern Mexico (Freytag and Debouck, 2002). Similarly, the above-mentioned descriptors are closely related to photosynthetic efficiency and are recognized - in some scenarios - as indicators of abiotic stress

resistance of individual accessions (Sánchez-Reinoso *et al.*, 2019; Guidi *et al.*, 2019). *P. vulgaris* usually exhibits a limited response (higher sensitivity) to drought stress compared to more resistant *P. acutifolius* (Rao *et al.*, 2013; Polania *et al.*, 2016). This can be closely related to the two independent domestication processes of *P. vulgaris* (Chacón *et al.*, 2005), mainly influenced by differences in air/soil humidity and contrasting temperatures between the Mesoamerican and Andean races, presenting differences also in their photosynthetic adaptations (Lynch *et al.*, 1992; González *et al.*, 1995).

The studied interspecific hybrid line INB 47 is a product of interspecific crossings carried out by the CIAT common bean-breeding program. The selection process focused on obtaining adequate seed type, growth habit and yield characteristics from the parent *P. vulgaris*. No surprise then, that the studied interspecific hybrid presented low phenomic proportions with the *P. parvifolius* and *P. acutifolius* parent accessions. This is because agronomically-valued traits likely do not coincide with those two parental accessions. This is probably mainly because physiological traits were selected indirectly (in contrast to agronomically-important descriptors), with no apparent interest/knowledge in/of physiological traits at the time of selection by the breeders (Mejia-Jimenez *et al.*, 1994).

Mejia-Jimenez *et al.* (1994) developed a group of CBC₅ interspecific hybrids with *P. acutifolius*. These authors generated populations with high genetic frequencies of *P. acutifolius*, showing average introgressions of 8% in CBC₅ using amplified fragment length polymorphisms (AFLP). Considering that the interspecific hybridization used in our study employed CBC₅ crossed twice with the parents *P. vulgaris* and *P. parvifolius*, and that it was selected during ten selfing cycles, it is likely that introgression of the *P. acutifolius* has decreased. Nevertheless, the 2.2% of the phenomic proportions of *P. acutifolius* - predicted from the physiological characterization -

evidence the successful introgressions from this parental line. In addition, the studied interspecific hybrid also preserves morphological traits similar to *P. acutifolius*, such as the lanceolate leaf shape and acute leaf angle at high light intensities (Figure 5.3). It could be interesting to evaluate the effect of these morphological traits on the abiotic stress resistance of this or other hybrids (after the methodology generalization).

Moreover, the studied interspecific hybrid keeps some characteristics that can influence the acclimatization process during abiotic stresses. This argues strongly in favor of conserving and characterizing accessions with intermediate phenomic proportions and could allow better understanding and quantifying (based on their GxE base) of the inherited traits and their proportions. It also would support accelerating genetic advances during more effective selection processes based on more newly available data types (semi-automatic remote sensing collection of data of highest interest).

Additionally, in hybrids, phenomic descriptors with the highest distinction powers (discrimination) could allow conserving desired physiological traits of *P. acutifolius* or *P. parvifolius* accessions, without losing key seed characteristics (e.g. size, color or taste) from their *P. vulgaris* parental line. Targeted accession evaluations can be performed by continuous monitoring even during the selection process (starting in already in the F1 generation). This would be based on the suggested machine learning techniques and selected traits of special interest for validating the functional introgressions of desired traits from crop wild relatives.

Our results demonstrate that the use of phenomic descriptors and machine learning analyses offer a very useful alternative for classifying hybrids, by using useful phenotypic and morphometric traits (with some degree of generalization and verification). In reality, breeders focusing on

interspecific crossings should consider physiological and morphological traits identified in this study as part of an effective screening strategy. This would be especially true where some of these traits were to prove functional in certain environments (willow leaves, leaf angle, growth habit, PHI) or be connected to farmers preferences (seed color and size, pod shape etc.). Breeders would then be able to use other selection criteria apart from the laborious final yield components and seed type characteristics, and thus quickly estimate the introgression efficiency of functional traits in the progenies.

Currently, the CIAT genebank conserves 18 interspecific hybrid lines of *P. vulgaris* x *P. acutifolius* x *P. parvifolius* and 6 interspecific hybrid lines of *P. vulgaris* x *P. acutifolius*, which were selected based on the phenotypic traits conserving characteristics associated with its crop relatives. In addition, the CIAT genebank stores 326 accessions of *P. acutifolius*, including cultivated lines, landraces and wild accessions. However, only a fraction of the whole collection has been studied and characterized for key agronomic and physiological traits, heavily limiting their utilization in pre-breeding or breeding programs (Mwale *et al.*, 2020). This suggests the urgent need to conduct experiments to explore phenomic traits of the *P. acutifolius* collection, including the genetically diverse wild tepary bean accessions as these offer a unique opportunity to find desirable genes with potential for introducing them into the genetic background of the domesticated tepary bean (Mhlaba *et al.*, 2018; Mwale *et al.*, 2020).

Breeders may then be encouraged to start working within the *acutifolius* group. We believe that selected phenomic descriptors can also help identify suitable “bridge” genotypes for crossings between secondary and tertiary genepools and common beans. The development of phenomic markers (new phenomic proportions recognized as important descriptors) will contribute to germplasm management in genebanks as well (Nguyen *et al.*, 2020). Selected and recognized

phenomic descriptors will facilitate the detection of accessions with similar phenomic proportions, determining accessions with high phenomic redundancy, and likely helping germplasm curators even to effectively find duplicate accessions.

Our study demonstrates that selected phenomic descriptors' data processed by a machine learning approach have the potential to discriminate between parental accessions or our studied hybrid. After some generalization (trait verification on different hybrid systems), this methodological approach may help breeders quantify any trait-of-interest introgression directly from different genepools or wild relatives increasing the chances of identifying important consumer target traits in elite common bean lines. After generalization, this methodology also will be able to identify hybrids with hybrid vigor due to the performance of unique phenomic traits. In addition, genome-associated phenomic markers could further contribute to the detection of genes of deep agronomic interest under abiotic and biotic stress conditions (Al-Tamimi *et al.*, 2021; Pasala *et al.*, 2020; Dwivedi *et al.*, 2020).

Detailed characterization of CIAT genebank conserved interspecific hybrids or new early breeding materials will likely show new traits with physiological or agronomic potential. In our study, the most contrasting characterization components with the highest precision and stability of the selected *Phaseolus* taxonomy classification are the seed and pod morphometry data.

In our study we were able to evaluate selected phenomic traits and their ability to become “phenomic markers” and then establish digital descriptors. We also identified machine learning techniques, which allow us to differentiate between studied *Phaseolus* accessions and determine the similarities or differences of an interspecific hybrid with respect to its parental lines. In our experiment we performed the analysis with random forests, however the strategy can use various

machine learning algorithms (Parmley *et al.*, 2019), since the purpose is to determine the most important descriptors that discriminate generally between the parents and its hybrid. There are several algorithms that can have greater accuracy in the classification according to the needs of the researcher and the dimensionality of the phenomic data.

5.5 Conclusions

In our work we demonstrate the use of phenomics and machine learning approach as analytical tools in understanding the phenotypic variability of selected *Phaseolus* lines and quantifying the crossing effectivity in its related hybrid. This study was never intended as the end of classic crop descriptors used by genebanks curators. In reality, classic descriptors will always offer their unique potential. However, some of them can still be rather subjective, are often only qualitative, and require laborious effort to apply them. We have tried to build on the understanding and precision of such classic mostly qualitative descriptors by digitization of some of the crop responses, so as to use a quantitative approach to make some descriptors available to modern breeders and with potential selection power (QTL, GWAS, genomic selection etc.).

In our study, we quantified the physiological, morphological and yield proportional relatedness of parental lines with its hybrid, finding differences between all groups. Results indicate that the interspecific hybrid preserve intermediate yield characteristics from *P. vulgaris* and *P. acutifolius* parents; although, it has closer phenotypic proportions with *P. vulgaris* (6 %). The phenomic proportions method can be a useful tool for the analysis of the closeness of lines/hybrids to their parents even by using traits with clear agronomic potential. However, also physiological data

(MultispeQ) showed high potential for lines discrimination, especially towards the studied line of *P. parvifolius*. This complex of traits needs to be further studied and amplified in a wide range of genotypes to verify its value across species, genepools and environments. Using a machine learning approach and phenotyping tools will automate the process of data evaluation in addition to the integration of classification algorithms. This will select the algorithm that has the highest discrimination accuracy and thus confer power to differentiate between accessions in any environment of interest.

5.6 Acknowledgments

This work was financed under the GCRF-BBR: Developing a hybrid-bean collection to advance climate-ready bean breeding project, led by the National Institute of Agricultural Botany (NIAB) (Huntingdon Road, Cambridge, CB3 0LE, UK). We thank PhD Monica Carvajal and Gustavo Cardona for their help and the space they provided during the collection of morphological and agronomic data. We would also like to thank PhD Daniel Debouck for his deep support and invaluable thoughts. Finally, to the genetic resources program and especially to the regeneration area for the help provided during the agronomic management of the crop. MOU is thankful to GIZ and PIAF Germany for their support. Finally we would like to thank Vincent Johnson of the Alliance Science Writing Service for his editorial review.

5.7 References

- Allendorf, F.W., Leary, R.F., Spruell, P. and Wenburg, J.K., 2001. The problems with hybrids: setting conservation guidelines. *Trends in ecology & evolution*, 16(11), pp.613-622.
- Al-Tamimi, N., Oakey, H., Tester, M. and Negrão, S., 2021. Assessing Rice Salinity Tolerance: From Phenomics to Association Mapping. In *Rice Genome Engineering and Gene Editing* (pp. 339-375). Humana, New York, NY.
- Araus, J.L., Kefauver, S.C., Vergara-Díaz, O., Gracia-Romero, A., Rezzouk, F.Z., Segarra, J., Buchaillot, M.L., Chang-Espino, M., Vatter, T., Sanchez-Bragado, R. and Fernandez-Gallego, J.A., 2022. Crop phenotyping in a context of Global Change: What to measure and how to do it. *Journal of Integrative Plant Biology*, 64(2), pp.592-618.
- Arnold, M.L., 1997. *Natural hybridization and evolution*. Oxford University Press on Demand.
- Beebe, S., Ramirez, J., Jarvis, A., Rao, I.M., Mosquera, G., Bueno, J.M. and Blair, M.W., 2011. Genetic improvement of common beans and the challenges of climate change. *Crop adaptation to climate change*, pp.356-369.
- Blair, M.W., Pantoja, W. and Carmenza Muñoz, L., 2012. First use of microsatellite markers in a large collection of cultivated and wild accessions of tepary bean (*Phaseolus acutifolius* A. Gray). *Theoretical and Applied Genetics*, 125(6), pp.1137-1147.
- Buhrow, R., 1983. The wild beans of southwestern North America.
- Chacón S, M.I., Pickersgill, B. and Debouck, D.G., 2005. Domestication patterns in common bean (*Phaseolus vulgaris* L.) and the origin of the Mesoamerican and Andean cultivated races. *Theoretical and Applied Genetics*, 110(3), pp.432-444.
- Chacón-Sánchez, M.I., 2018. The domestication syndrome in *Phaseolus* crop plants: A review of two key domestication traits. *Origin and Evolution of Biodiversity*, pp.37-59.

de Carvalho, M.A., Bebeli, P.J., Bettencourt, E., Costa, G., Dias, S., Dos Santos, T.M. and Slaski, J.J., 2013. Cereal landraces genetic resources in worldwide GeneBanks. A review. *Agronomy for sustainable development*, 33(1), pp.177-203.

Debouck, D.G., 2021. Phaseolus beans (Leguminosae, Phaseoleae): a checklist and notes on their taxonomy and ecology. *Journal of the Botanical Research Institute of Texas*.

Deva, C.R., Urban, M.O., Challinor, A.J., Falloon, P. and Svitáková, L., 2020. Enhanced leaf cooling is a pathway to heat tolerance in common bean. *Frontiers in plant science*, 11, p.19.

Dwivedi, S.L., Goldman, I., Ceccarelli, S. and Ortiz, R., 2020. Advanced analytics, phenomics and biotechnology approaches to enhance genetic gains in plant breeding. *Advances in Agronomy*, 162, pp.89-142.

Eliceiri, K.W., Berthold, M.R. and Goldberg, I.G., 2012. Ibáñez L. Manjunath BS, Martone ME, Murphy RF, Peng H, Plant AL, Roysam B, Stuurman N, Stuurmann N, Swedlow JR, Tomancak P, Carpenter AE, pp.697-710.

Feller, C., Bleiholder, H., Buhr, L., Hack, H., Hess, M., Klose, R., Meier, U., Stauss, R., Boom, T.V.D. and Weber, E., 1995. Phanologische Entwicklungsstadien von Gemusepflanzen II. Fruchtgemüse und Hülsenfruchte. *Nachrichtenblatt des Deutschen Pflanzenschutzdienstes*, 47(9), pp.217-232.

Feldmann, M.J., Hardigan, M.A., Famula, R.A., Lopez, C.M., Tabb, A., Cole, G.S. and Knapp, S.J., 2020. Multi-dimensional machine learning approaches for fruit shape phenotyping in strawberry. *GigaScience*, 9(5), p.giaa030.

Fernández-Calleja, M., Monteagudo, A., Casas, A.M., Boutin, C., Pin, P.A., Morales, F. and Igartua, E., 2020. Rapid on-site phenotyping via field fluorimeter detects differences in photosynthetic performance in a hybrid—parent barley germplasm set. *Sensors*, 20(5), p.1486.

Freytag, G.F. and Debouck, D.G., 2002. Taxonomy, distribution, and ecology of the genus *Phaseolus* (Leguminosae-Papilionoideae) in North America, Mexico and central America. *BRIT*.

Friedrich, S. and Pauly, M., 2018. MATS: Inference for potentially singular and heteroscedastic MANOVA. *Journal of Multivariate Analysis*, 165, pp.166-179.

González, A., Lynch, J., Tohme, J.M., Beebe, S.E. and Macchiavelli, R.E., 1995. Characters related to leaf photosynthesis in wild populations and landraces of common bean. *Crop Science*, 35(5), pp.1468-1476.

Guidi, L., Lo Piccolo, E. and Landi, M., 2019. Chlorophyll fluorescence, photoinhibition and abiotic stress: does it make any difference the fact to be a C3 or C4 species? *Frontiers in plant science*, 10, p.174.

Haghghi, K.R. and Ascher, P.D., 1988. Fertile, intermediate hybrids between *Phaseolus vulgaris* and *P. acutifolius* from congruity backcrossing. *Sexual Plant Reproduction*, 1(1), pp.51-58.

Henao-Rojas, J.C., Rosero-Alpala, M.G., Ortiz-Muñoz, C., Velásquez-Arroyo, C.E., Leon-Rueda, W.A. and Ramírez-Gil, J.G., 2021. Machine Learning Applications and Optimization of Clustering Methods Improve the Selection of Descriptors in Blackberry Germplasm Banks. *Plants*, 10(2), p.247.

Huang, L.K. and Wang, M.J.J., 1995. Image thresholding by minimizing the measures of fuzziness. *Pattern recognition*, 28(1), pp.41-51.

Janitza, S. and Hornung, R., 2018. On the overestimation of random forest's out-of-bag error. *PloS one*, 13(8), p.e0201904.

Kandanaarachchi, S. and Hyndman, R.J., 2021. Dimension reduction for outlier detection using DOBIN. *Journal of Computational and Graphical Statistics*, 30(1), pp.204-219.

Kassambara, A. and Mundt, F., 2017. Package 'factoextra'. Extract and visualize the results of multivariate data analyses, 76(2).

Kennedy, R.E., Yang, Z., Braaten, J., Copass, C., Antonova, N., Jordan, C. and Nelson, P., 2015. Attribution of disturbance change agent from Landsat time-series in support of habitat monitoring in the Puget Sound region, USA. *Remote Sensing of Environment*, 166, pp.271-285.

Kholová, J., Urban, M.O., Cock, J., Arcos, J., Arnaud, E., Aytekin, D., Azevedo, V., Barnes, A.P., Ceccarelli, S., Chavarriaga, P. and Cobb, J.N., 2021. In pursuit of a better world: crop improvement and the CGIAR. *Journal of Experimental Botany*, 72(14), pp.5158-5179.

Kuhlgert, S., Austic, G., Zegarac, R., Osei-Bonsu, I., Hoh, D., Chilvers, M.I., Roth, M.G., Bi, K., TerAvest, D., Weebadde, P. and Kramer, D.M., 2016. MultispeQ Beta: a tool for large-scale plant phenotyping connected to the open PhotosynQ network. *Royal Society open science*, 3(10), p.160592.

Kusolwa, P.M., Myers, J.R., Porch, T.G., Trukhina, Y., González-Vélez, A. and Beaver, J.S., 2016. Registration of AO-1012-29-3-3A red kidney bean germplasm line with bean weevil, BCMV, and BCMNV resistance. *Journal of Plant Registrations*, 10(2), pp.149-153.

Liaw, M. A & Wirner, M, 2018. Package ‘randomforest’. University of California, Berkeley: Berkeley, CA, USA.

Lynch, J., González, A., Tohme, J.M. and García, J.A., 1992. Variation in characters related to leaf photosynthesis in wild bean populations. *Crop Science*, 32(3), pp.633-640.

Mejía-Jiménez, A., Muñoz, C., Jacobsen, H.J., Roca, W.M. and Singh, S.P., 1994. Interspecific hybridization between common and tepary beans: increased hybrid embryo growth, fertility, and efficiency of hybridization through recurrent and congruity backcrossing. *Theoretical and Applied Genetics*, 88(3), pp.324-331.

Muñoz, L.C., Duque, M.C., Debouck, D.G. and Blair, M.W., 2006. Taxonomy of tepary bean and wild relatives as determined by amplified fragment length polymorphism (AFLP) markers. *Crop Science*, 46(4), pp.1744-1754.

Mhlabi, Z.B., Mashilo, J., Shimelis, H., Assefa, A.B. and Modi, A.T., 2018. Progress in genetic analysis and breeding of tepary bean (*Phaseolus acutifolius* A. Gray): A review. *Scientia horticulturae*, 237, pp.112-119.

Mwale, S.E., Shimelis, H., Mafongoya, P. and Mashilo, J., 2020. Breeding tepary bean (*Phaseolus acutifolius*) for drought adaptation: A review. *Plant Breeding*, 139(5), pp.821-833.

Nankar, A.N., Tringovska, I., Grozeva, S., Ganeva, D. and Kostova, D., 2020. Tomato phenotypic diversity determined by combined approaches of conventional and high-throughput tomato analyzer phenotyping. *Plants*, 9(2), p.197.

Nguyen, G.N. and Norton, S.L., 2020. Genebank phenomics: A strategic approach to enhance value and utilization of crop germplasm. *Plants*, 9(7), p.817.

Noriega, I.L., Halewood, M., Abberton, M., Amri, A., Angarawai, I.I., Anglin, N., Blümmel, M., Bouman, B., Campos, H., Costich, D. and Ellis, D., 2019. CGIAR operations under the plant treaty framework. *Crop Science*, 59(3), pp.819-832.

Pasala, R. and Pandey, B.B., 2020. Plant phenomics: High-throughput technology for accelerating genomics. *Journal of biosciences*, 45(1), pp.1-6.

Parmley, K.A., Higgins, R.H., Ganapathysubramanian, B., Sarkar, S. and Singh, A.K., 2019. Machine learning approach for prescriptive plant breeding. *Scientific reports*, 9(1), pp.1-12.

Polania, J., Rao, I.M., Cajiao, C., Rivera, M., Raatz, B. and Beebe, S., 2016. Physiological traits associated with drought resistance in Andean and Mesoamerican genotypes of common bean (*Phaseolus vulgaris* L.). *Euphytica*, 210(1), pp.17-29.

Porch, T.G., Beaver, J.S. and Brick, M.A., 2013. Registration of tepary germplasm with multiple-stress tolerance, TARS-Tep 22 and TARS-Tep 32. *Journal of Plant Registrations*, 7(3), pp.358-364.

Paluszynska, A. (2017). Structure mining and knowledge extraction from random forest with applications to the cancer genome atlas project. University of Warsaw Master Thesis.

Rao, I., Beebe, S., Polania, J., Ricaurte, J., Cajiao, C., Garcia, R. and Rivera, M., 2013. Can tepary bean be a model for improvement of drought resistance in common bean? *African Crop Science Journal*, 21(4).

Rassizadeh, L., Cervero, R., Flors, V. and Gamir, J., 2021. Extracellular DNA as an elicitor of broad-spectrum resistance in *Arabidopsis thaliana*. *Plant Science*, 312, p.111036.

Rosero, A., Pérez, J.L., Rosero, D., Burgos-Paz, W., Martínez, R., Morelo, J., Pastrana, I., Burbano, E. and Morales, A., 2019. Morphometric and colorimetric tools to dissect morphological diversity: An application in sweet potato [Ipomoea batatas (L.) Lam.]. *Genetic resources and crop evolution*, 66(6), pp.1257-1278.

Salazar, D.E., Santos, L.G., Wenzl, P. and Hay, F.R., 2020. Effect of dry heat on seed germination of Desmodium and Stylosanthes species. *Seed Science and Technology*, 48(3), pp.419-437.

Sánchez-Reinoso, A.D., Ligarreto-Moreno, G.A. and Restrepo-Díaz, H., 2019. Chlorophyll α fluorescence parameters as an indicator to identify drought susceptibility in common bush bean. *Agronomy*, 9(9), p.526.

Schlautman, B., Diaz-Garcia, L. and Barriball, S., 2020. Morphometric approaches to promote the use of exotic germplasm for improved food security and resilience to climate change: a kura clover example. *Plant science*, 290, p.110319.

Singh, A., Ganapathysubramanian, B., Singh, A.K. and Sarkar, S., 2016. Machine learning for high-throughput stress phenotyping in plants. *Trends in plant science*, 21(2), pp.110-124.

Schinkel, C. and Gepts, P., 1989. Allozyme variability in the tepary bean, *Phaseolus acutifolius* A. Gray. *Plant Breeding*, 102(3), pp.182-195.

Singh, S.P., Debouck, D.G. and Roca, W., 1998. Interspecific hybridization between *Phaseolus vulgaris* L. and *P. parvifolius* Freytag.

Singh, S.P. and Munoz, C.G., 1999. Resistance to common bacterial blight among *Phaseolus* species and common bean improvement. *Crop Science*, 39(1), pp.80-89.

Soltis, P.S., Nelson, G., Zare, A. and Meineke, E.K., 2020. Plants meet machines: Prospects in machine learning for plant biology. *Applications in Plant Sciences*, 8(6).

Tadesse, W., Sanchez-Garcia, M., Assefa, S.G., Amri, A., Bishaw, Z., Ogbonnaya, F.C. and Baum, M., 2019. Genetic gains in wheat breeding and its role in feeding the world. *Crop Breed. Genet. Genom*, 1, p.e190005.

Wang, S.M. and Zhang, Z.W., 2011. The State of the World's Plant Genetic Resources for Food and Agriculture. *Journal of Plant Genetic Resources*, 12(3), pp.325-338.

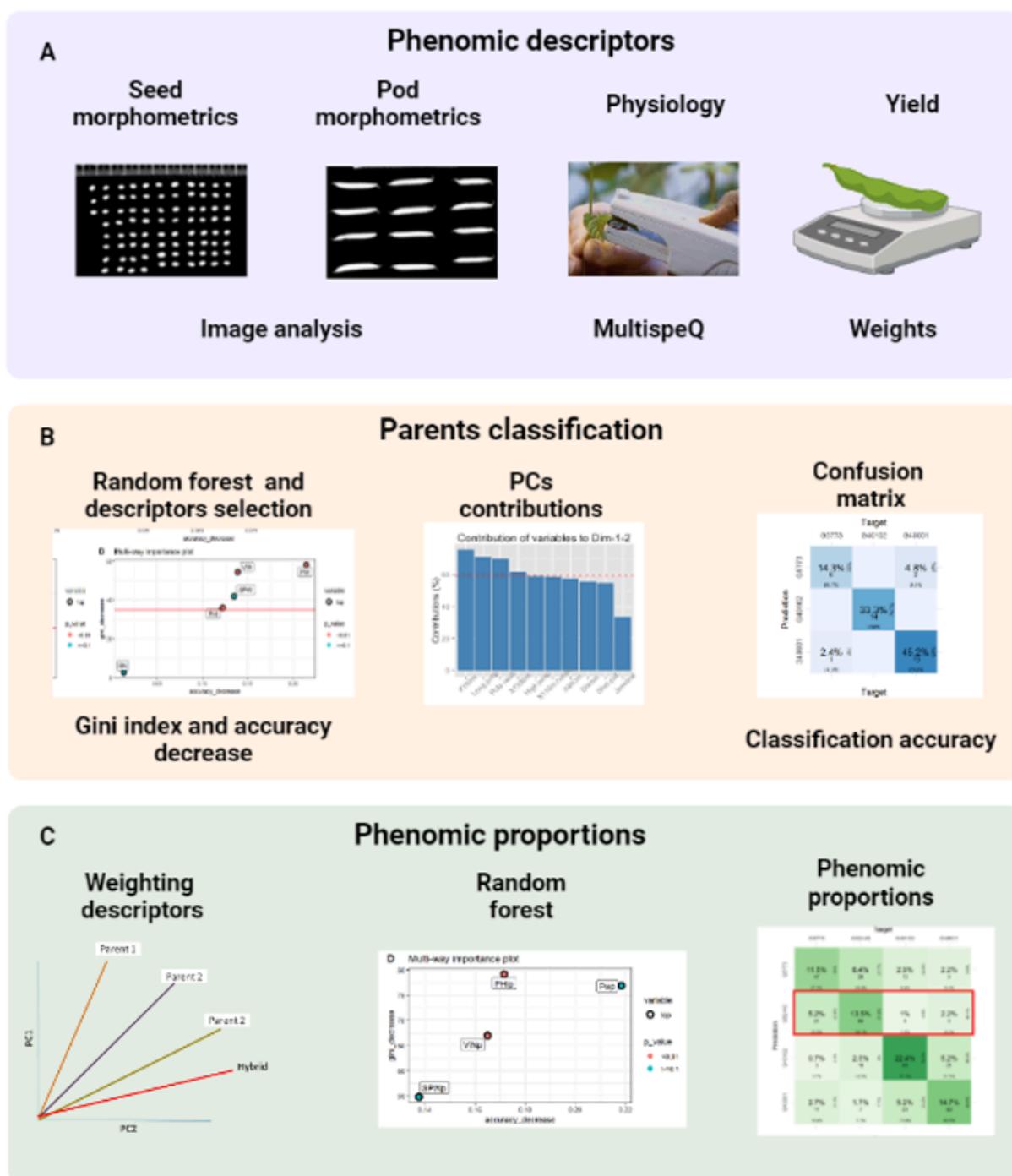


Figure 5.1. The procedure consists of three steps: (A) capture and processing of phenomic descriptors, (B) classification of parents using random forest and PCA and (C) phenomic ratios based on the weighting of the descriptors with the contribution of the PCs and phenomic ratios as the prediction of the hybrid with respect to its parents in the confusion table.

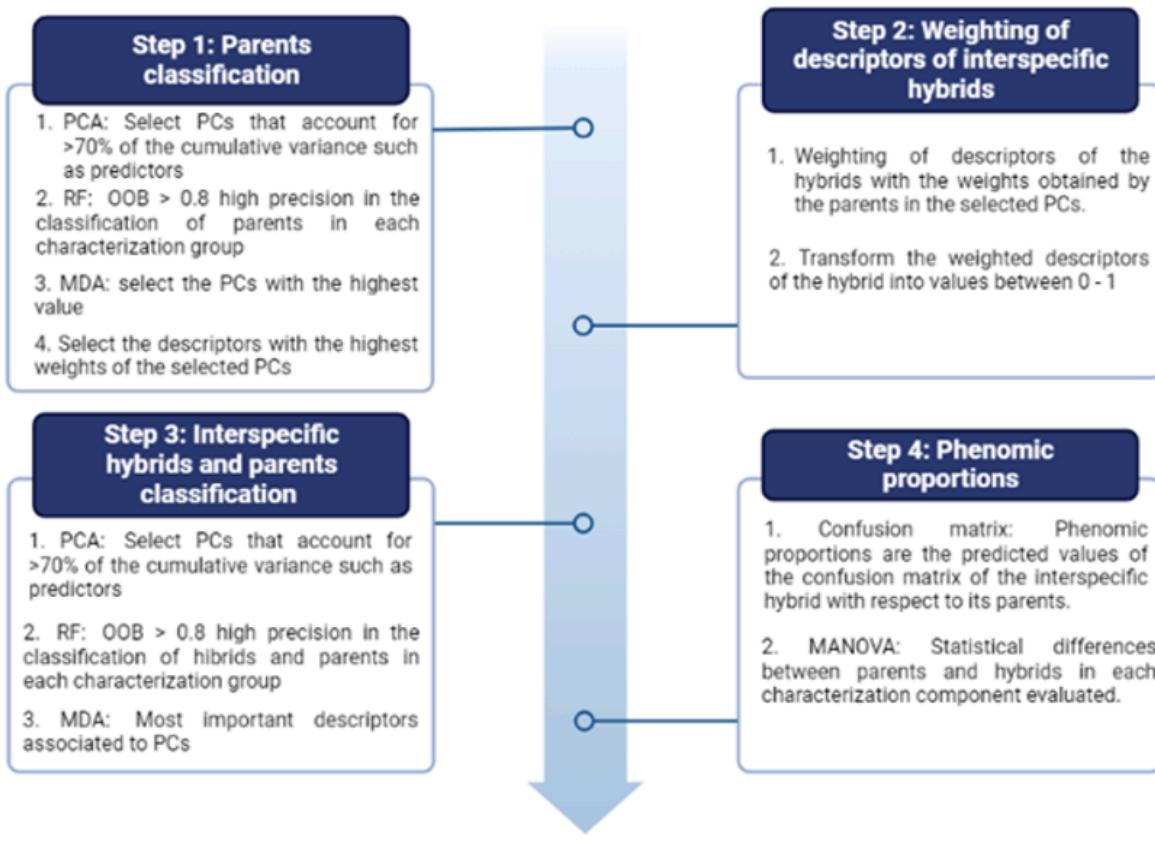


Figure 5.2. Procedure for the analysis of the phenomic proportions between interspecific hybrids and their parents. Four stages are observed that contemplate multivariate (PCA) and random forest (RF) analyses. **OOB:** Out of bag accuracy and **MDA:** Mean decrease accuracy.



Figure 5.3. Phenotypic characteristics of the *Phaseolus* lines and interspecific hybrid, from its initial stages until harvest. (A) *P. acutifolius* (G40001), (B) *P. parvifolius* (G40102), (C) *P. vulgaris* - ICA Pijao (G5773) and (D) Interspecific hybrid (G52443 - INB 47).

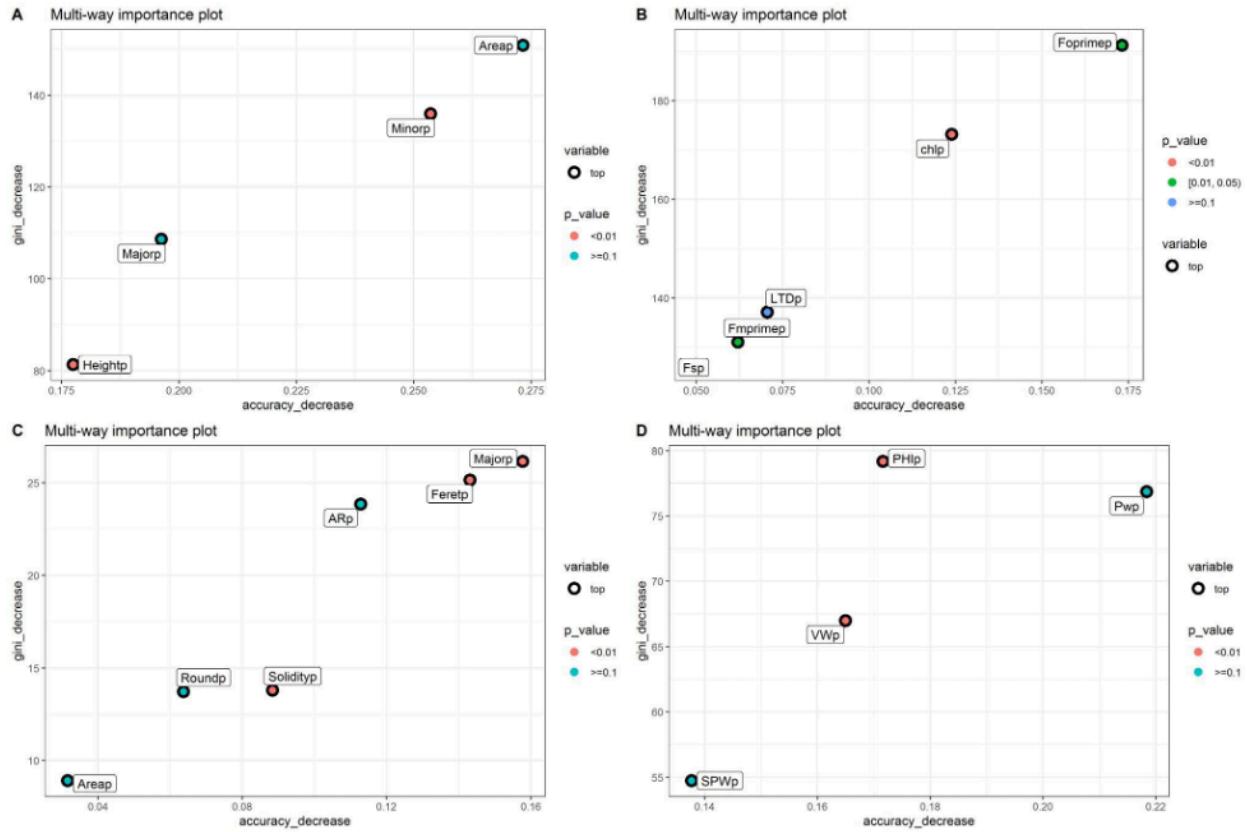


Figure 5.4. Gini index and accuracy decrease for feature selection in random forest for each characterization component in the hybrid. (A) Seed morphometric component, (B) Physiological component, (C) Pod morphometric component and (D) Yield component. The weighted phenomic descriptors are observed.

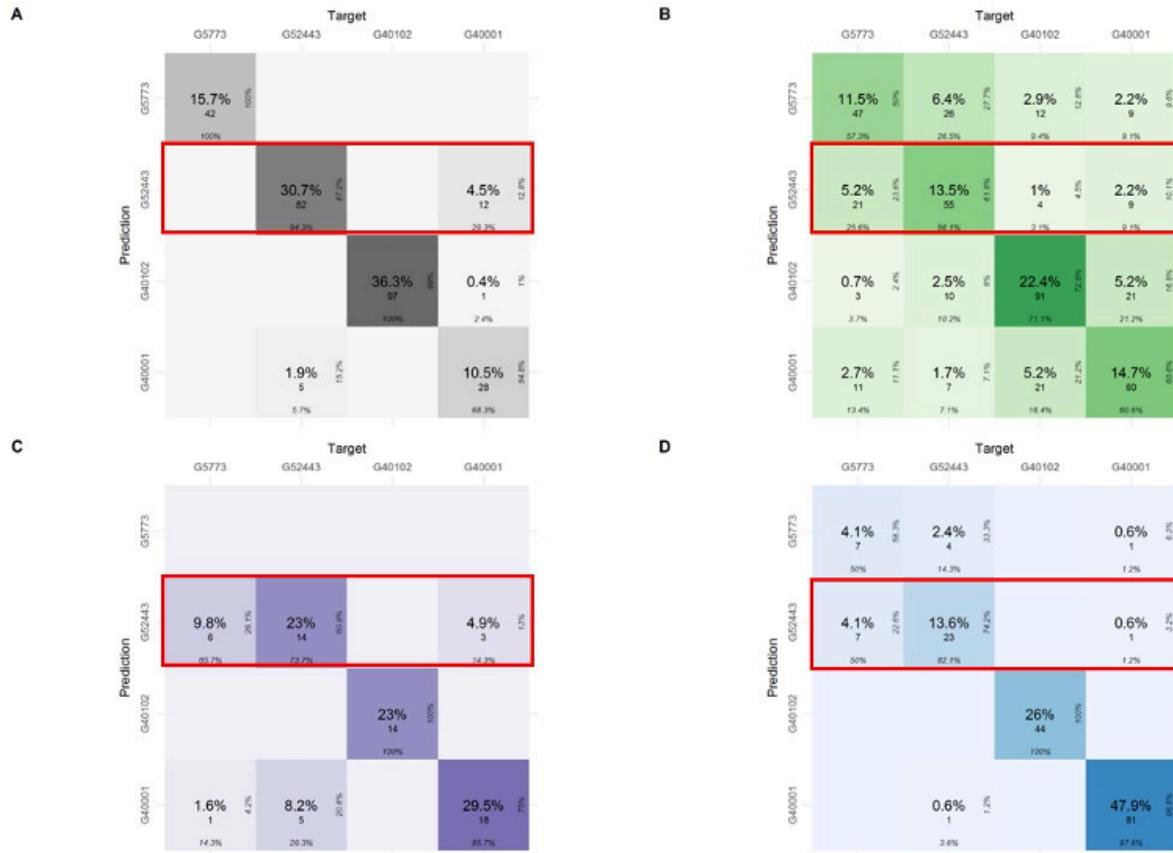


Figure 5.5. Phenomic proportions in the confusion matrices for classification of the interspecific hybrid in each of the characterization components using phenomics descriptors. (A) Seed morphometric component, (B) Physiological component, (C) Pod morphometric component and (D) Yield component.

Table 5.1. MANOVA of parents and interspecific hybrid in each of the characterization components. The p-value for each parent and its relationship to the hybrid is observed.

Characterization component	Accession	Contrast to:	p value	MATS	p - value Resampling
Physiological	G40001 <i>(P. acutifolius)</i>	parvifolius	< 0.001	1016.229	< 0.001
		vulgaris	0.0073		
		hybrid	0.1053		
	G40102 <i>(P. parvifolius)</i>	acutifolius	< 0.001		
		vulgaris	< 0.001		
		hybrid	< 0.001		
	G5773 <i>(P. vulgaris)</i>	acutifolius	0.0073		
		parvifolius	< 0.001		
		hybrid	0.7523		
Pod morphometrics	G40001 <i>(P. acutifolius)</i>	parvifolius	< 0.001	2538.441	< 0.001
		vulgaris	0.6099		
		hybrid	0.8856		
	G40102 <i>(P. parvifolius)</i>	acutifolius	< 0.001		
		vulgaris	0.0004		
		hybrid	< 0.001		
	G5773 <i>(P. vulgaris)</i>	acutifolius	0.6099		
		parvifolius	0.0004		

		hybrid	0.8301		
Seed morphometrics	G40001 <i>(P. acutifolius)</i>	parvifolius	< 0.001	26348.12	< 0.001
		vulgaris	< 0.001		
		hybrid	< 0.001		
	G40102 <i>(P. parvifolius)</i>	acutifolius	< 0.001		
		vulgaris	< 0.001		
		hybrid	< 0.001		
	G5773 <i>(P. vulgaris)</i>	acutifolius	< 0.001		
		parvifolius	< 0.001		
		hybrid	0.6546		
Yield	G40001 <i>(P. acutifolius)</i>	parvifolius	< 0.001	9697.0	< 0.001
		vulgaris	< 0.001		
		hybrid	< 0.001		
	G40102 <i>(P. parvifolius)</i>	acutifolius	< 0.001		
		vulgaris	0.0302		
		hybrid	0.015		
	G5773 <i>(P. vulgaris)</i>	acutifolius	< 0.001		
		parvifolius	0.0302		
		hybrid	0.9986		

6. Chapter six: Synthesis of results

Tepary bean adaptation has evolved in contrasting environmental conditions compared to its close relatives. The environment led to morphological, physiological, and agronomic traits that have positively influenced the relationship between the environment and man, as demonstrated by this study. The study's most important findings are that it demonstrates the conceptual, technological, and procedural capacity for implementing low-cost germplasm characterisation techniques that contribute significantly to the quality of the passport data and genebank documentation.

The classic documentation of germplasm collections typically includes data that limits the selection of parents and the usage of the accessions even if the senior advisor is present. One of the reasons is that systematic exploration requires expert(s) in many disciplines from computational statistics, and good agronomic practices to ethnobiology. The usage of collections should be promoted by documenting accessions and selecting different types of accessions based on the target environmental prediction (target population of environments; TPE). For example, the tepary bean collection naturally includes various morphological and photosynthetic adjustments that alter in light use efficiency (LUE).

These traits likely offer distinct accessions different sensitivity to high, low, or limited light conditions. As a result, bean breeders have germplasms providing new traits and their combinations according to the current environments and specific needs determined by experts that are part of the program. We propose a methodological flow (Figure 6.1) to carry out characterization and evaluation processes with digital descriptors that incorporate the usage of proximal sensors that help differentiate accessions in germplasm collections. We hope our

research raised awareness about the need of working with crop wild relatives that have agronomic potential for direct use by resource-poor populations. CIAT's Future Seeds alone conserves over 37,000 Phaseolus accessions, 85% of which have never been investigated but have all been successfully regenerated and are now conserved in Svalbard.

It is easy to imagine that these seeds are direct solutions to farmers with scarce economic resources, nutritional inadequacies, and demand for adaptations to abiotic-biotic stress. "*EVERYTHING WHAT WE HAVE*" neither nature, time, nor our ancestors are providing it; now we must know how to use it sustainably and targeted.

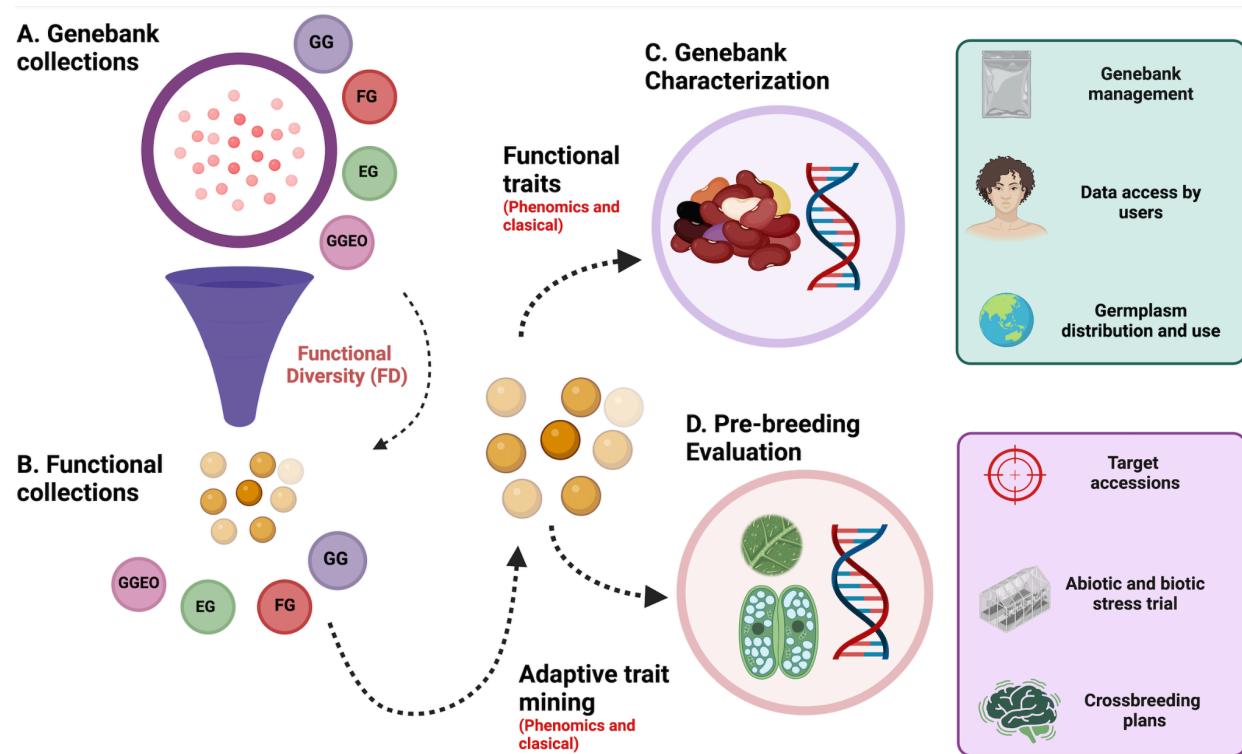


Figure 6.1. Methodological proposal for adaptive trait mining during characterization and documentation in genebanks, integrating ecological, genetic, functional, and geographic groups.

Germplasm exploration employing high-throughput phenotyping strategies identify and select accessions with genetic improvement potential by utilizing various proximal remote sensing techniques to enhance characterisation procedures during seed regeneration. These functional traits acquired during seed regeneration, in addition to being supported by genetic and ecological structure studies, will be documented in collections that will allow them to be employed later, either directly owing to specific properties or in breeding programs. The collections of functional traits will aid in exploring the knowledge of the accessions from the evolutionary processes that directed the diversity of the genome, as well as relating complementary traits such as LUE, water use efficiency (WUE), and nutrient use efficiency (NUE).

6.1 Future prospects

Genebanks are the most valuable asset because they conserve genetic resources, which can help food systems adapt to the effects of climate change. Several diversity studies are currently being created based on phenotype traits and more recently on the research of genomic structure variation, as sequencing technology becomes more accessible. The study of genome variation can help to articulate characterization processes in genebanks by directly relating digital descriptors that can give genetic association processes that contribute to the use of genetic resources, particularly crop wild relatives. Crop phenomics will be the next revolution in the study of adaptive traits in genetic resources for agriculture and food, especially to the possibility for trait mining utilizing analytical approaches that aid in germplasm selection.

The analysis of images and signals employing proximate remote sensors will allow for the acquisition of valuable data, a reduction in assessment times associated with seed regeneration operations, and a decrease in operating expenses for both specialized workers and independent characterization processes. Today, we have the genetic resources required for sustainable production that relates the use of potential genetic resources. Associating it with technologies that involve the efficient use of resources such as water, light, and nutrients will support better decisions during the selection process of germplasm that can be used as food potential as well as in breeding programs.