Universidad de Oviedo

Departamento de Ingeniería Química y Tecnología del Medio Ambiente

Programa de Doctorado en Ingeniería Química, Ambiental y Bioalimentaria

USO Y DESARROLLO DE HERRAMIENTAS GENÉTICAS Y GENÓMICAS PARA LA GESTIÓN DE LA DIVERSIDAD Y ACELERAR PROGRAMAS DE MEJORA EN JUDÍA

USE AND DEVELOPMENT OF GENETIC AND GENOMIC TOOLS TO DIVERSITY MANAGEMENT AND ACCELERATE BREEDING PROGRAMS IN COMMON BEAN

Doctorando:

María Jurado Cañas

Oviedo, Septiembre 2024

Universidad de Oviedo

Departamento de Ingeniería Química y Tecnología del Medio Ambiente

Programa de Doctorado en Ingeniería Química, Ambiental y Bioalimentaria

USO Y DESARROLLO DE HERRAMIENTAS GENÉTICAS Y GENÓMICAS PARA LA GESTIÓN DE LA DIVERSIDAD Y ACELERAR PROGRAMAS DE MEJORA EN JUDÍA

USE AND DEVELOPMENT OF GENETIC AND GENOMIC TOOLS TO DIVERSITY MANAGEMENT AND ACCELERATE BREEDING PROGRAMS IN COMMON BEAN

Doctorando:

María Jurado Cañas

Dirigida por: Dr. Juan José Ferreira Fernández

Dra. Ana Campa Negrillo

Oviedo, Septiembre 2024

Universidad de Oviedo

RESUMEN DEL CONTENIDO DE TESIS DOCTORAL

1.- Título de la Tesis	
Español/Otro Idioma: Uso y desarrollo de herramientas genéticas y genómicas para la gestión de la diversidad y acelerar programas de mejora en judía	Inglés: Use and development of genetic and genomic tools to diversity management and accelerate breeding programs in common bean

2.- Autor	
Nombre: María Jurado Cañas	
Programa de Doctorado: Ingeniería Química, Ambiental y Bioalimentaria	
Órgano responsable: Centro Internacional de Postgrado	

RESUMEN (en español)

Las legumbres juegan un papel muy importante en la seguridad alimentaria mundial, ya que tienen un alto valor nutricional, su cultivo tiene un impacto ambiental reducido y existe una gran diversidad intra e interespecífica. Entre ellas, destaca la judía, *Phaseolus vulgaris* L., como la leguminosa más importante para el consumo directo en el mundo, siendo la principal fuente de proteínas de los países en vías de desarrollo. En la región Norte de España el cultivo de judía representa un recurso socioeconómico importante, centrado principalmente en el cultivo del tipo varietal Fabada, que se caracteriza por su elevada calidad culinaria y una semilla bien diferenciada por forma y dimensiones.

Uno de los principales retos del cultivo de judía son los efectos del cambio climático. El desarrollo de variedades vía mejora genética es una de las herramientas más valiosas para hacer frente a estos desafíos. La mejora genética se basa, en gran medida, en los conocimientos sobre el control genético de los caracteres. En los últimos años, ha habido una gran expansión de nuevas técnicas y métodos de fenotipado y genotipado de alto rendimiento que aceleran la conexión fenotipo-genotipo. El objetivo principal de esta Tesis es hacer uso de las últimas técnicas desarrolladas para incrementar los recursos y el conocimiento de la diversidad genética de judía y del control genético de caracteres importantes en la mejora para contribuir a un uso y conservación eficiente de la diversidad y para acelerar los programas de mejora genética.

En el capítulo 1, se estudió la diversidad dentro del tipo varietal Fabada utilizando genotipado de alto rendimiento. Para ello se constituyó un panel de diversidad de 179 líneas homocigotas, incluyendo materiales antiguos conservados en colecciones de germoplasma y materiales cultivados actualmente. Se identificaron duplicaciones y sinonimias dentro del panel lo que permitió reducir su tamaño, facilitando así tanto su conservación como su uso en programas de mejora. Se constató que durante los últimos 30 años se ha producido una importante erosión genética en la región dentro de esta clase comercial.

En el capítulo 2 se aplicó RNA-seq para estudiar la expresión diferencial de genes en respuesta a la infección causada por la raza 38 de *Colletotrichum lindemuthianum* (antracnosis) en una línea mejorada de Fabada. Se identificaron 7 genes candidatos dentro del cluster de resistencia *Co-2*, localizado en el cromosoma Pv11, y se desarrollaron marcadores funcionales de estos genes. Estos marcadores serán de gran utilidad para aplicar en los programas de mejora genética.

En el capítulo 3, se estudió la arquitectura genética de caracteres complejos como son morfología y calidad de la semilla mediante mapeo por asociación y RNA-seq. Se fenotipó un panel de diversidad formado por 311 líneas y que se consideran representativas de la diversidad española para esta especie. Se identificaron 23 regiones genómicas implicadas en el control de estos caracteres, 5 de las cuáles fueron consistentes con regiones previamente

descritas. Mediante RNA-seq se estudió la expresión diferencial de genes durante el desarrollo de la semilla de la variedad 'Xana' y se identificaron 22 genes candidatos a controlar el tamaño de la semilla.

En el capítulo 4 se aborda la secuenciación *de novo*, el ensamblado y la anotación del genoma de la línea A25, derivada de la variedad 'Andecha'. Este recurso genético supone un importante material de trabajo para los programas de mejora de esta clase comercial. Además, se pondrá a disposición de la comunidad científica para su uso en estudios comparativos y evolutivos, así como su posible uso en el pan-genoma de *P. vulgaris*.

Esta Tesis proporciona recursos para el estudio de la diversidad genética, utiliza métodos de genotipado para contribuir a una mejor conservación de la diversidad y etiqueta regiones y genes candidatos a ser responsables del control genético de caracteres importantes en la mejora genética de la especie.

RESUMEN (en Inglés)

Legumes play a very important role in world food security since they have a high nutritional value, their cultivation has a reduced environmental impact and there is a great intra- and interspecific diversity. Among them, the common bean, *Phaseolus vulgaris* L., stands out as the most important legume for direct consumption in the world, being the main source of protein in developing countries. In the northern region of Spain, common bean cultivation represents an important socioeconomic resource, mainly focused on cultivating the Fabada varietal type, characterized by its high culinary quality and seed well differentiated by shape and dimensions.

One of the main challenges for common bean cultivation is the effects of climate change. Variety development through breeding is one of the most valuable tools to meet these challenges. Breeding mainly relies on knowledge about the genetic control of traits. In recent years, there has been a great expansion of new high-throughput phenotyping and genotyping techniques and methods that accelerate the phenotype-genotype connection. The main objective of this Thesis is to increase the resources and knowledge of the genetic diversity of *P. vulgaris* and the genetic control of important traits in breeding to contribute to the efficient use and conservation of diversity, as well as to accelerate breeding programs, making use of the latest techniques developed.

In Chapter 1, diversity within the Fabada varietal type was studied using high-throughput genotyping. For this purpose, a diversity panel of 179 homozygous lines within the market class Fabada was constituted, including old materials conserved in germplasm collections and currently cultivated materials. Duplications and synonymies were identified within the panel, which allowed reducing its size, thus facilitating both its conservation in the germplasm bank and its use in breeding programs. It was found that during the last 30 years, there has been a significant genetic erosion in materials grown by local farmers within this market class.

In Chapter 2, RNA-seq was applied to study differential gene expression in response to infection caused by the *Colletotrichum lindemuthianum* race 38 (anthracnose) in an improved Fabada line. Seven candidate genes were identified within the *Co-2* resistance cluster, located on chromosome Pv11, and functional markers for these genes were developed. These markers will be very useful to apply in breeding programs

In Chapter 3, the genetic architecture of seed morphology and quality were studied by association mapping and RNA-seq. A diversity panel (Spanish Diversity Panel) consisting of 311 lines which considered representative of the Spanish diversity for this species were phenotype for seven seed traits. Twenty-three genomic regions involved in the control of these traits were identified, 5 of which were consistent with previously described regions. Differential gene expression during seed development of the Fabada market class 'Xana' was studied by RNA-seq and 22 candidate genes controlling seed size were identified underline those QTL.

Chapter 4 deals with the de novo sequencing, assembly, and annotation of the genome of line A25, derived from the Fabada market class 'Andecha'. This genetic resource represents an important working material for the breeding programs of this commercial class. In addition, it will

be made available to the scientific community for use in comparative and evolutionary studies, as well as its possible use in the pan-genome of *P. vulgaris*.

In summary, this Thesis provides resources for the study of genetic diversity, uses genotyping methods to contribute to a better conservation of diversity and labels regions and candidate genes to be responsible for the genetic control of important traits in the breeding of the species.

SR. PRESIDENTE DE LA COMISIÓN ACADÉMICA DEL PROGRAMA DE DOCTORADO EN _____

This work has been developed in the Regional Service for Agrofood Research and Development (SERIDA), in collaboration with the University of Oviedo and thanks to the aid of the projects: **BRESOV** *Breeding for Resilient, Efficient and Sustainable Organic Vegetable Production* - EU; **INCREASE** *Intelligent Collections of Food Legumes Genetic Resources for European Agroffod Systems* - EU; **SUSTCROP** *Sustainable Crop* – Principality of Asturias; **AGL207-87050_R** *Análisis genéticos para la actualización de conocimientos y Desarrollo de herramientas útiles en la mejora genética de judía de grano y verde (Phaseolus vulgaris L.)* – AEI; **PID2021-123919OB-100** *Análisis genéticos para incrementar los conocimientos y herramientas para la mejora genética y sostenibilidad del cultivo de judía común* – AEI.

During this PhD, María Jurado Cañas has been supported by the Grant **PRE2019-091249** funded by MICIU/AEI/10.13039/501100011033 and by 'ESF Investing in your future', which also financed an international stay in INRAE – Université Paris Saclay.

En primer lugar, me gustaría mostrar mi agradecimiento a mis directores de Tesis. Al Dr. Juan José Ferreira por su confianza en mí desde el primer día, por su paciencia y en especial, por todo lo que me ha permitido aprender de él. A la Dra. Ana Campa, por su calma y su apoyo en todo momento, por tus enseñanzas pero también por el placer que ha sido estudiar juntas cosas nuevas. A los dos, gracias por estos 4 años de crecimiento personal y académico en los que me habeis acompañado y guiado de una forma excepcional. Gracias por hacer posible esta Tesis.

A esas personas que yo cariñosamente llamo "mis postdocs", esas que son tan imprescindibles para cualquier estudiante y que desde mi TFG he tenido la suerte de que me acompañen, las que te brindan siempre el consejo que necesitas: José Die, Patricia, Juan Camilo, Álvaro, Marta. Pero durante estos 4 años, mi agradecimiento más sincero es para Carmen García, la persona que estuvo conmigo desde mi primer día en el SERIDA y de la que espero siempre una silla junto a ella para cuando la conversación la requiera. Has superado todas las expectativas de lo que son esas personitas "postdoc", no he podido estar mejor acompañada durante mi tesis, ha sido un auténtico placer trabajar contigo como lo ha sido conocerte.

Gracias a todo el equipo de Genética Vegetal del SERIDA, a Marcos y Fernando por estar siempre dispuestos a ayudarme y a todo el personal que ha mantenido mis fabas con vida durante este tiempo sin poner en evidencia las dotes de agrónoma que aún me faltan. Gracias a Valerie Geffroy y a todo el grupo de GDYNPATH que me acogieron y enseñaron durante mis 3 meses en Paris y de los que guardo un precioso recuerdo.

A todos mis compañeros de despacho por hacer de éste un lugar tan amigable, en especial a mis "fruitis": Belén, Javi y Manu por llenar de risas, posits e ideas absurdas tanto el despacho como muchos otros lugares de Asturias y por la tranquilidad de saber que lo vamos a seguir haciendo. A mis amigas de siempre, que siguen ahí a pesar de todas mis mudanzas y a Marta y Fátima de las que siempre recibo visitas por muy lejos que me lleve la ciencia.

A tí Agus, por todas las veces que me sientas en el sofá y me pones los pies en la tierra, por la paciencia y por ayudarme tantísimo a crecer. Porque has hecho del Norte un hogar para mí y por estos maravillosos años. Pero sobre todo gracias por hacerme feliz.

Quiero agradecer a mi famila que siempre ha estado ahí con su apoyo incondicional a pesar de que no hayan sido los mejores años para estar lejos de casa, manchados por confinamientos y hospitales. Gracias por la ayuda que siempre he recibido en mis estudios y por manteneros enteros entre todas las lágrimas en la estación que tanto nos duelen a los cuatro.

TABLE OF CONTENTS

ABBREVIATIONS AND UNITS 5

GENERAL INTRODUCTION 11

1. Legumes crops 13

2. Common bean 14

2.1. Botanical description 15

2.2. Origins and evolution 17

2.3. Common bean genetic diversity 19

2.4. Fabada market class 21

3. Tools for the study of genetic diversity 23

3.1. High-throughput phenotyping 23

3.2. Genotyping-by-sequencing 24

3.3. Whole-genome sequencing 25

3.4. Whole-transcriptome analysis 27

4. Tools to study the genetic basis of characters 28

4.1. Biparental populations 28

4.2. Diversity Panels 29

5. Common bean breeding 31

OBJECTIVES 35

CHAPTER 1 39

1. Introduction 43

2. Materials and methods 45

2.1. Plant material 45

2.2. DNA isolation and genotyping 46

2.3. Panel filtering 47

2.4. Population diversity and clustering analyses 47

3. Results 48

3.1. Genotyping of the Faba Panel 48
3.2. Filtered from genotyping 49
3.3. Filtered from phenotyping 54
3.4. Genetic diversity in the Fabada market class 54
4. Discussion 56
5. Conclusion 58
CHAPTER 2 59
1. Introduction 63
2. Material and methods 66
2.1. Plant material 66
2.2. Inoculation with Colletotrichum lindemuthianum race 38 66
2.3. Total RNA isolation, cDNA library construction, and sequencing 67
2.4. Differentially expressed genes 67
2.5. Gene ontology (GO) analysis of DEGs 68
2.6. Genotyping by sequencing 68
2.7. RNA-Seq reads assembly and visualization 69
3. Results 69
3.1. Transcriptome sequencing of resistant and susceptible genotypes 69
3.2. Differentially expressed genes 70
3.3. Functional classification of DEGs 73
3.4. The physical position of the Co-2 cluster the line A4804 74
3.5. DEGs in the genomic Co-2 region *74*
3.6. Specific markers for the Co-2 cluster 77
4. Discussion 78
5. Conclusion 82
CHAPTER 3 83
1. Introduction 87

2. Materials and methods 90
2.1. Plant material 90
2.2. Genotyping 91
2.3. Phenotyping 91
2.4. Statistical analysis seed traits 92
2.5. GWAS and haplotype block detection 92
2.6. RNA-Seq 93
2.7. Gene ontology enrichment analysis 94
2.8. Approach to candidate genes 94
3. Results 95
3.1. Genotyping of the SDP 95
3.2. Phenotypic variation 95
3.3. Genome-wide association study 96
3.4. *Differentially expressed genes* 104
3.5. GO enrichment analysis of DEGs 105
3.6. Putative candidate genes for seed traits 105
4. Discussion 106
5. Conclusion 109
CHAPTER 4 111
1. Introduction 115
2. Material and methods *116*
2.1. Genomic DNA Extraction and Sequencing 116
2.2. Genome Assembly 116
2.3. Structural and functional annotation 117
2.4. Genomes comparison 117
3. Results 119
3.1. Raw sequencing data 119

3.2. Genome Assembly 120
3.3. Genome annotation 121
3.4. Genomes characteristics comparison 121
4. Discussion 123
5. Conclusion 124
GENERAL DISCUSSION 125
1. Genetic diversity and germplasm management 127
2. Genetic architecture of important traits 128
3. Increase knowledge of common bean genome 130
4. Essential issues 132
CONCLUSIONS/CONCLUSIONES 133
REFERENCES 137
ANNEX I 159

ABBREVIATIONS AND UNITS

A	*Andean gene pool*
AFLP	*Amplified Fragment Length Polymorphism*
bp	*base pair*
BP	*Biological Process*
CC	*Cellular Components*
cDNA	*Complementary DNA*
CIAT	*International Center for Tropical Agriculture*
Cl	*Colletotrichum lindemuthianum*
Cm	*centimeter*
cm^2	*square centimeter*
CNV	*Copy Number Variations*
Co	*Conserved population*
CP	*Coat Proportion*
CRF-INIA-CSIC	*Centro Nacional de Recursos Fitogenéticos*
Cu	*Cultivated population*
cv	*Cultivar*
DEG	*Differentially expressed genes*
DNA	*Deoxyribonucleic Acid*
FAO	*Food and Agriculture Organization of the United Nations*
FAOSTAT	*Global Food and Agriculture Statistics of FAO*
FASTmrEMMA	*fast multi-locus random-SNP-effect EMMA*
FM	*Functional Marker*
FP	*Faba Panel*

FPKM	*Fragments Per Kilobase per Million mapper fragments*
g	*grams*
GBS	*Genotyping by Sequencing*
GO	*Gene-ontology*
GWAS	*Genome-Wide Association Study*
H'	*Shannon-Wiener diversity index*
HCA	*Hierarchical clustering analysis*
HCPC	*Hierarchical Clustering on Principal Components*
Hpi	*hours post-inoculation*
HT3P	*High-Throughput Plant Phenotyping Platform*
HTG	*High-Throughput Genotyping*
HTP	*High-Throughput Phenotyping*
IGV	*Integrative Genomics Viewer*
InDel	*Insertion/Deletion*
InDel	*Insertions/Deletions*
KASP	*Kompetitive Allele Specific PCR*
LD	*Linkage disequilibrium*
LIS	*Legume Information System*
LOD	*Logarithm of the odds*
LRR	*Leucine-rich repeats*
LTR	*Long terminal repeat*
LWR	*Seed Length/Seed Width ratio*
MA	*Mesoamerican gene pool*

MABC	*Marker-assisted backcrossing*
MAF	*Minor Allele Frequency*
MAS	*Marker-Assisted Selection*
Max	*Maximum*
Mbp	*Million base pairs*
MF	*Molecular Function*
Min	*Minimum*
MLM	*Mixed Linear Model*
NCBI	*National Center for Biotechnology Information*
NGS	*Next-Generation Sequencing*
NIL	*Near-Isogenic Line*
ORA	*Over-representation analysis*
PAV	*Presence/Absence Variations*
PCA	*Principal Components Analysis*
PCoA	*Principal Coordinates Analysis*
PCR	*Polymerase Chain Reaction*
PDO	*Protected Denomination Origin*
PGI	*Protected Geographic Indication*
PTI	*Patterned-triggered immunity*
QTL	*Quantitative Trait Loci*
QTN	*Quantitative trait nucleotide*
R	*Plant disease-resistant genes*
RAD	*Restriction site associated DNA*

Re	*Reference population*
REML	*Restricted maximum likelihood*
RGB	*Red Green Blue*
RIL	*Recombinant Inbred Lines*
RNA	*Ribonucleic Acid*
RNA-Seq	*RNA sequencing*
SA	*Seed Area*
SD	*Standard Deviation*
SDP	*Spanish Diversity Panel*
SERIDA	*Regional Service for Agrofood Research and Development*
SL	*Seed Length*
SNP	*Single-nucleotide Polymorphism*
SV	*Structural Variations*
SW	*Seed weight*
SWI	*Seed Width*
TF	*Transcription Factor*
TMM	*Trimmed Mean of M-values*
WA	*Water Absorption*
WGS	*Whole-Genome Sequencing*

GENERAL INTRODUCTION

1. Legumes crops

Legumes belong to the family *Fabaceae* or *Leguminosae*. It includes approximately 770 genera and nearly 20,000 distributed species worldwide and represents the third-largest family of flowering plants (Lewis *et al*. 2005). Legumes are plants that produce their fruit as pods and are widely used in agriculture, both for animal feeding and human food, including major crops such as soybean (*Glycine max* (L.)Merr), common bean (*Phaseolus vulgaris* L.), chickpea (*Cicer arietinum* L.), lentil (*Lens culinaris* Medik), pea (*Pisum sativum* L.) or faba beans (*Vicia faba* L.).

Legume cultivation provides significant environmental benefits (Stagnari *et al*. 2017; Uebersax *et al*. 2022; Yanni *et al*. 2023). Legumes can perform a symbiotic relationship with nitrogen-fixing bacteria on their roots, which converts atmospheric nitrogen into a form that plants can use. This process improves soil fertility and reduces the need for synthetic nitrogen fertilizers, thus reducing the environmental impact of agriculture, as greenhouse gas emissions from agriculture come in large part from fertilizer application. Legumes have the ability to mobilize phosphorus and other essential nutrients and micronutrients into the soil because of their deeper root system, longer growth, improved soil structure, and enhanced carbon sequestration. The high organic content that legumes produce while growing feeds soil microorganisms and increases soil diversity benefiting soil health and leaving extra nutrients for the next crops to be grown on the same soils, making legume-based crop rotation or intercropping systems very environmentally suitable (Chamkhi *et al*. 2022).

In 2016, the Food and Agriculture Organization (FAO) introduced the term ‘pulses’ to refer to the legume crops harvested solely for dry grain, excluding legume crops harvested for green food which are classified as vegetable crops, as well as legume crops used mainly for oil extraction or used exclusively for sowing purposes. Beans, lentils, chickpeas, and peas are the most popular and common pulse types for human consumption. Pulses are characterized by high nutritional value compared to the seeds of other plants. They have a high protein content ranging from 20% to 30%, a low glycemic index, and high fiber content; are cholesterol-free; are an important source of minerals such as iron and vitamins, such as folate; and have been shown to have interesting bioactive properties, such as phenolic compounds (Geraldo *et al*. 2022; Grdeń and

Jakubczyk 2023). The results of several clinical studies suggest that pulse intake may have a protective effect against several diseases, such as cardiovascular diseases, type 2 diabetes mellitus, and specific types of cancer (Hayat *et al*. 2014; Messina 2014; Martín-Cabrejas 2019). The fact that pulses are an affordable source of protein and minerals compared to animal protein, that they have a long shelf life without losing their nutritional value and avoiding wastage makes them a major contributor to food security (FAO 2024).

In the society discourse on biodiversity loss, sustainable agriculture, and climate change, the revitalization of pulses as a vegetal protein source has emerged as a sensible and promising approach to shaping the future of our planet (Vasconcelos *et al*. 2020). European producers do not seem to favor these crops, with the percentage of arable land used for pulses being 1.5% in Europe, compared to 14.5% worldwide. Increasing the cultivation of pulses in Europe could play an important role in the agricultural and dietary sustainability objectives of the European Union's Farm-to-Fork strategy (Costa *et al*. 2021).

2. Common bean

The common bean (*Phaseolus vulgaris* L.) is the legume most cultivated worldwide for human consumption. The genus *Phaseolus* includes over 30 *Phaseolus* species native to the Americas, of which only five are cultivated (López *et al*. 1985; Miklas and Singh 2007): *P. vulgaris* L. (common bean), *P. coccineus* L. (runner bean), *P. lunatus* L. (lima bean), *P. acutifolius* A.Gray (tepary bean) and *P. dumosus* Macfad (year bean) (Freytag and Debouck 2002).

According to their human consumption, there are two main types of common beans: snap and dry beans. In dry beans, the seeds are cooked after rehydration. However, in some places, the seeds are consumed immediately before drying (green seeds). Bean seeds are interesting because of their nutritional composition, high protein content, carbohydrates, vitamins, and minerals, and the presence of functional compounds, such as dietary fiber and phenolic compounds, which exert protective effects against various diseases (Hayat *et al*. 2014; Rodríguez Madrera *et al*. 2024). Snap beans (syn. French

beans and green beans) are a group of common bean cultivars whose fresh pods are harvested at a physiologically immature stage and are consumed as green vegetables.

Bean-harvested areas and production have increased in recent years (Figure 1). According to FAO (2023)), the dry bean global harvested area was 36.79 million ha and production was 28.35 million tons in 2022 (Figure 1a); for snap beans, the global harvested area was 1.6 million ha and production was 23.34 million tons (Figure 1b). Asia is the major producer, accounting for 50% of the global production, followed by Africa, America, and Europe, where production represents 1% of the total (Nadeem *et al*. 2021).

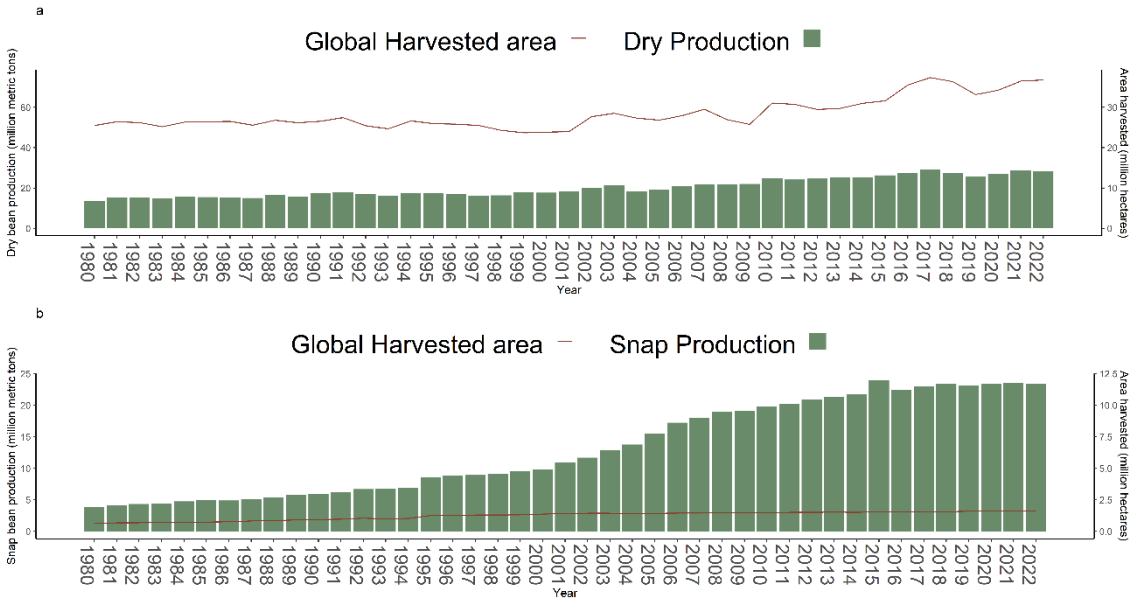

Figure 1. Evolution of the harvested area (red line) and production (green bar) worldwide of common bean between 1980 and 2022 (source: FAOSTAT 2023). a) Dry common bean data. b) Snap bean data.

2.1. Botanical description

The common bean is diploid (2n = 22) and autogamous species with an annual cycle. This species shows epigeal germination, where the cotyledons emerge with the hypocotyl. The root system tends to be fasciculate where the primary root is generally distinguished, with a low percentage of pivot types within the species. The stem is herbaceous, with cylindrical or angular sections, hairy, and tends to be vertical. The stem is formed by nodes or internodes and can end in an inflorescence (determinate growth habit; Figure 2a) that stops growth, or in a meristem that allows the growth to continue (indeterminate growth habit; Figure 2b). The pilosities and colors of stems are highly variable.

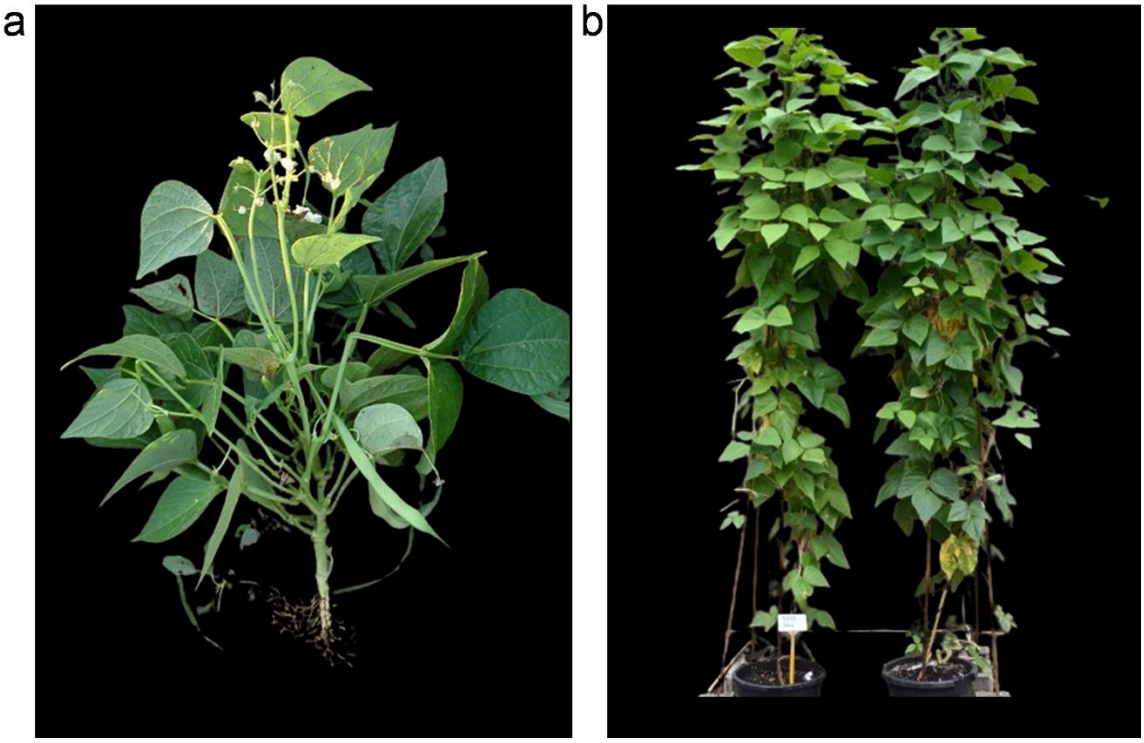

Figure 2. Growth habits of common bean. a) Determinate growth habit. b) Indeterminate growth habit.

There are two types of common bean leaves: simple and compound. The simple leaves are the primary leaves that appear on the second node of the stem and are formed on the seed during embryogenesis; they are opposite, cordate, and unifoliate. The second type is trifoliate, which is typical of bean leaves. The flower is papilionaceous, with five stamens, a superior ovary and anthers, and stigma at the same level, favoring self-fertilization. The fruit is a pod composed of two valves joined by dorsal and ventral sutures. The seed has no albumen, concentrating the nutritional reserves in the cotyledons, can be of various shapes, and has a covering or testa of maternal origin (Debouck and Hidalgo 1985a).

The development cycle of the plant usually lasts between 3-6 months, and is divided into two phases (Figure 3): i) the vegetative phase, from seed germination to the appearance of the first flower buds or first flower clusters; ii) the reproductive phase, from the end of the vegetative phase to harvest maturity. When the leaves begin to senesce, the mature pods are harvested (Debouck and Hidalgo 1985b).

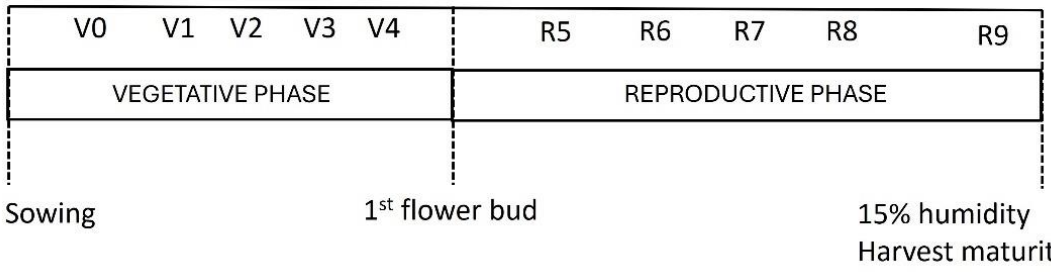

Figure 3. Stages in the development of a common bean plant. V0: Germination; V1: Emergence; V2: Primary leaves; V3: First trifoliate leaf; V4: Third trifoliate leaf; R5: Pre-flowering; R6: Flowering; R7: Pod formation; R8: Pod filling; R9: Ripening. (Own created; Debouck and Hidalgo (1985b)).

Considering the termination of the main stem, the number of nodes, and the ability to climb, were described 4 different growth habits in common beans: i) Type I. Determinate bush; ii) Type II. Indeterminate erect; and iii) Type III. Indeterminate prostrate; iv) Type IV. Indeterminate climbing (Debouck and Hidalgo 1985a).

2.2. Origins and evolution

Wild populations of *P. vulgaris* are located in two major ecogeographical regions in America: from northern Mexico to Colombia and from southern Peru to northwestern Argentina (Zizumbo-Villarreal *et al*. 2005). So, it considered that *P. vulgaris* has an American origin with two eco-geographical gene pools, the Mesoamerican (MA) and the Andean (A) which are differentiable at morphology (Salinas *et al*. 1988), seed protein patterns (Gepts and Bliss 1986) and allozymes (Koenig and Gepts 1989). Based on Amplified Fragment Length Polymorphism (AFLP) markers and sequence data evidence a larger amount of diversity in the MA wild germplasm than in the A. The most probable hypothesis is an MA origin of the common bean, most likely in Mexico, from where different migration events into South America occurred, resulting in a bottleneck during the formation of the wild A gene pool (Rossi *et al*. 2009; Bitocchi *et al*. 2012). Two independent domestication events occurred (Figure 4): one in MA and the other in A (Gepts *et al*. 1986; Chacón S *et al*. 2005; Kwak and Gepts 2009; Mamidi *et al*. 2011; Bitocchi *et al*. 2013). Domestication of common beans has led to numerous morphological and physiological alterations. These changes encompassed variations in growth patterns; the presence of seed dormancy; sensitivity to photoperiod; alterations in

the morphology of harvested parts, including shape, color, and size; and modifications in mechanisms for dissemination (Bellucci *et al*. 2014).

Figure 4. Geographic distribution of common bean following domestication (Created by biorender.com; Castro-Guerrero *et al.* (2016)).

Both gene pools were brought into Europe, probably through the Iberian Peninsula, after the expedition of Francisco Pizarro to northern Peru in 1529 (Gepts and Bliss 1988; Bellucci *et al*. 2023), and then evolved separately from their sources. Santalla *et al*. (2002) described local bean germplasm from the Iberian Peninsula based on allozymes and proposed southwestern Europe as a secondary center of genetic diversity of the common bean, giving rise to numerous distinct landraces (Puerta Romero 1961; Pérez-Vega *et al*. 2009; Angioi *et al*. 2010; Campa *et al*. 2018). Its dispersion throughout Europe followed routes that led to multiple exchanges between different European countries (Papa *et al*. 2006). The European germplasm is mostly of A origin, with most being 67% of the total, and the MA seeds found in Europe are the largest within this group owing to the introduction of varieties with larger seeds and the preference of the farmers for them or to the introgression of Andean type varieties (Logozzo *et al*. 2007; Angioi *et al*. 2010).

2.3. Common bean genetic diversity

Genetic diversity can be defined as the heritable variation in phenotypes and genotypes within a species. Genetic diversity is essential to crop sustainability because provides characters, genes, and gene combinations to breeding. Common bean exhibits wide phenotypic and genotypic diversity as a result of its particular evolution, domestication, adaptation, and breeding (Debouck and Hidalgo 1985a; López *et al*. 1985; Voysest 2000) (Figure 5). There is also variation in the nutritional components of the seed such as fiber, mineral, or phenolic content (Steckling *et al*. 2017; Moghaddam *et al*. 2018; Rodríguez Madrera *et al*. 2020).

Figure 5. Examples of flower, pod, and seed phenotypic variation in common bean.

Throughout history, farmers have used and selected all these morphological variations, giving rise to different local varieties or landraces. The landrace concept is useful for naming or distinguishing cultivated varieties through simple traits that are locally adapted to traditional farming systems (Zeven 1998b). The conservation of genetic diversity has been approached with the establishment of *ex-situ* germplasm

collections that are maintained in genebanks under controlled conditions (Offord 2017). The largest and most diverse *Phaseolus* germplasm collection in the world is located at the International Center for Tropical Agriculture (CIAT; Cali, Colombia) with 37,938 accessions, 86% of them of *P. vulgaris*. The second largest collection is located at the United States Department of Agriculture-Agricultural Research Service (USDA-ARS), with nearly 18,000 samples. The main collection of beans in Spain is preserved in the Centro Nacional de Recursos Fitogenéticos (CRF-INIA-CSIC), with ~3,000 *P. vulgaris* accessions. A *Phaseolus* collection of ~450 accessions is maintained at the Regional Service for Agrofood Research and Development (SERIDA; Asturias, Spain), which is representative of the diversity cultivated in northern Spain for this species (Ferreira *et al*. 2005). This collection was established in 1991 with the original goal of conserving the local diversity of the specific market class Fabada; for this reason, approximately 25% of this collection are accessions of this market class.

One of the main challenges of gene banks is the management of large number of conserved accessions, as well as the large amount of information generated per accession. Core collections have been established to facilitate collection management (Frankel 1984; Brown 1989; van Hintum *et al*. 2000). A core collection, or nuclear collection, is a smaller collection formed by a limited set of accessions, with minimum repetitiveness, representative of the genetic diversity maintained in the original collection (van Hintum *et al*. 2000; Gu *et al*. 2023). For example, a core collection was proposed to represent and manage the diversity of the Spanish bean collection maintained at the CRF-INIA-CSIC. This Spanish core collection consisted of 202 accessions that were selected based on morphological seed traits and passport data (De la Rosa *et al*. 2000; Pérez-Vega *et al*. 2009; Rivera *et al*. 2018). A duplicate of this core collection is maintained also at the SERIDA and was included as a part of the Spanish Diversity Panel (SDP; https://zenodo.org/records/10263706), which was established as a sample representative of the main Spanish diversity for this species (Campa *et al*. 2018). The establishment of core collections is of great help in the management of diversity, but it is still difficult to conserve the material and how best to handle this diversity for efficient use in plant breeding programs.

This wide phenotypic variation has been grouped into market classes within snaps and dry beans. For example, according to the phenotype of the pods, different snap bean market classes had been defined, such as 'Yellow wax, ' 'String snap bean, ' 'Romano type, ' 'Blue Lake type, ' 'Filet type, ' Fine' or 'Garrafal type' (García-Fernández *et al*. 2022). Also, there are numerous market classes for dry beans according to the seed phenotype, like 'Navy', 'Carioca'. 'White Kidney', 'Fabada', 'Canellini', Yellow', or 'Pinto' (Voysest Voysest 2000). In addition, some varieties may have differences due to their geographical origin and be under the protection of differentiated brands of the European Union (EU), such as protected designation of origin (PDO) or protected geographical indication (PGI). Some examples are PDO 'Cannellino di atina', PDO 'Bianchi di Rotonda, PDO 'Fesols de Santa Pau', PDO 'Mongeta del Ganxet', PGI 'Faba Asturiana', PGI 'Alubia de la Bañeza', PGI Judía del Barco de Ávila' or PGI 'Faba de Lorenzá' (https://www.mapa.gob.es/; https://www.politicheagricole.it/).

2.4. Fabada market class

The Fabada market class (syn. Favada, Faba Granja, and Faba de manteca) is characterized by a well-differentiated seed phenotype, featuring very large white seeds (~ 100 g/100 seeds), and an oblong shape with a length/width ratio greater than 2.2 (Figure 6). Fabada was already described in northern Spain by the mid-20th century by Puerta Romero (1961). Fabada landraces show indeterminate growth habits, whereas modern cultivars can have both determinate and indeterminate growth habits (Ferreira *et al*. 2017).

Figure 6. Flower, pods, plant, and seed phenotypic in Fabada market class (cv. 'Andecha').

The fabada bean crop has an importance in Spanish gastronomy and agricultural practices, especially in the north of Spain, the region of Asturias, and east of Galicia. Since 1996, had a PGI recognized by the EU (P.G.I Faba Asturiana), due to the particular quality, reputation, and characteristics attributable to the geographical origin. From 2013, the trend in production under this origin denomination has been increasing, in the 2021/2022 season 213 hectares were cultivated and 275 tons were produced in Asturias (https://faba-asturiana.org/). Fabada is a traditional legume that holds a special place in Asturian agriculture and cuisine, and its crop and preservation are essential for maintaining agricultural diversity, promoting sustainable farming practices, and

preserving cultural heritage. A large collection of landraces representing the Spanish diversity of the Fabada market class is maintained at the CRF-INIA-CSIC (Madrid; https://bancocrf.inia.es/es/). Most of these accessions were collected before 1991, and a duplicate collection has been maintained in SERIDA collection for 30 years.

3. Tools for the study of genetic diversity

Classic diversity studies are based on the characterization of the main morphological traits, such as growth habits, colors, and seed and pod shapes, or molecular markers such as seed proteins, allozymes, or DNA markers (Debouck and Hidalgo 1985a; Gepts *et al*. 1986; Koenig and Gepts 1989; Voysest Voysest 2000). In the last few years, progress in new technologies has led to the development of new phenomic techniques, tools, and platforms that rapidly generate comprehensive data about specific traits and characteristics of large populations of plants. In addition, new Next-Generation Sequencing (NGS) platforms provide many genotypic data, such as Single nucleotide polymorphism (SNP) and sequences, that have changed the way in harnessing the potential of genomic resources in the genetic improvement of crop plants. These new tools are essential for linking phenotypes and genotypes and for identifying genetic architectures that control important traits.

3.1. High-throughput phenotyping

Plant phenotyping is a comprehensive assessment of visible plant traits, such as growth, development, tolerance, resistance, architecture, physiology, ecology, yield, and the basic measurement of individual quantitative parameters that form the basis for more complex traits (Li *et al*. 2014). Conventional phenotyping methods may be expensive, time-consuming, invasive, and have the potential to diminish the accuracy and reliability of outcomes. Classical phenotyping involves manual measurement of plant traits, such as growth, morphology, phenology, architecture, stress response, and yield.

High-throughput phenotyping (HTP) is a nondestructive and rapid approach for monitoring and measuring multiple phenotypic traits related to growth, yield, and adaptation to biotic or abiotic stresses (Pabuayon *et al*. 2019). HTP includes techniques, such as multispectral imaging systems, scanning devices, chlorophyll fluorescence

sensors, control terminals, and image-based technologies. HTP is an important tool for breeding programs that generally aim to phenotype large populations for numerous traits in multiple environments or with replicated assays, as it allows for fast discovery and generates complex datasets for analysis.

High-throughput precision phenotyping platforms (HT3P) are novel and powerful tools that employ advanced sensors and data collection systems that allow the monitoring and quantification of phenotypic traits in a rapid, nondestructive, and high-throughput manner to achieve genomics-assisted breeding (Li *et al*. 2020). These platforms may include mobile phenotyping tools, drones, robotic platforms, and image analysis software. There are numerous image analysis software specialized in plant breeding, in the website Quantitative Plant (http://www.quantitative-plant.org/, accessed 04.03.2024) we can find 181 image software tools to measure different plant organs.

In common bean, the application of HTP, in combination with genetic variation, extends to addressing challenges such as low water availability or drought response through different platforms, such as sensors for visible (RGB), hyperspectral remote sensors, chlorophyll fluorescence, and multispectral imaging (Padilla-Chacón *et al*. 2019; Wong *et al*. 2023; Javornik *et al*. 2023), response to nutrient deficit by measuring morphological traits (Lazarević *et al*. 2022), and root architecture (Burridge *et al*. 2016; Jochua *et al*. 2020).

3.2. Genotyping-by-sequencing

Classical genotyping refers to traditional methods (typically PCR-based techniques) for determining the genotype of an organism and provides a low number of marker loci per analysis. In contrast, high-throughput genotyping (HTG) provides a high number of marker loci per analysis and is based on NGS methods. Among them, the Genotyping-by-sequencing (GBS) method is very common. GBS is a genotyping method based on reducing genome complexity by using restriction enzymes before sequencing (Elshire *et al*. 2011). This method assumes that an advance simplifies the computational alignment problems in plant species with complex genomes, high levels of genetic diversity, or frequent structural variations such as gene copy numbers or transposon rearrangements that produce extensive presence/absence variations. This technique is an evolution of

restriction site-associated DNA (RAD) sequencing, the earliest reduced-representation sequencing method that used restriction enzymes to divide the genome into DNA fragments (Miller *et al*. 2007). The original protocol for GBS includes the following steps: i) DNA digestion with *ApeKI*, ii) library preparation, iii) sequencing on NGS platforms, and iv) identification of SNPs. This protocol was extended to a two-enzyme version that combines a rare and common cutting restriction enzyme to generate uniform sequencing libraries, which allowed an increase in genotyping density (Poland *et al*. 2012).

GBS can be used in species without a reference genome and has gained popularity in crop research and plant breeding because of its high throughput and low cost. The use of SNPs as DNA markers for plant genotyping has increased the potential to score variation in specific DNA targets, being the most abundant and stable genetic variation marker, and is an attractive approach to saturate mapping and breeding populations with a high density of SNP markers in numerous crop species. GBS has been applied for different purposes, such as genomic selection, gene mapping, or genome-wide association analysis (GWAS) (He *et al*. 2014b; Wickland *et al*. 2017).

3.3. Whole-genome sequencing

Whole-genome sequencing (WGS) involves determining the complete DNA sequence of an organism's genome. The WGS workflow starts with lab procedures that involve DNA fragmentation, preparation of the library with fragment size selection and amplification, sequencing, and DNA sequence analysis, which includes read trimming, filtering, and assembly of the sequenced fragments in scaffolds and then in chromosomes (Ekblom and Wolf 2014). In plants, the large size and highly repetitive genomes, possible abundance of retrotransposons, and presence of duplicated genes make assembly a difficult task (Claros *et al*. 2012).

Genome assembly refers to the process of placing nucleotide sequences in the correct order, and assembly is required because sequence read lengths are much shorter than most genomes or even most genes. The availability of public databases, such as the National Center for Biotechnology Information (NCBI), Phytozome or Legume System Information (LIS), and reference genomes, facilitates this task; however, because the

position of genes is not always constant between genomes and copies of genes may appear, assembly may require even more time than obtaining sequencing data (Foxman 2012). It is important to distinguish between *de novo* assembly, which reconstructs genomes of organisms for which there is nothing similar, and re-sequencing or comparative approaches, which use a nearby organism genome during assembly (Pop 2009). Long-read sequencing technologies (e.g. Pacbio; Oxford Nanopore) allow genome assembly to be relatively easy, covering repetitive genomic regions, with read sizes ranging from 250 bp to 2.3 Mb, depending on the platform (Amarasinghe *et al*. 2020; Driguez *et al*. 2021).

Genome annotation is the process of identifying any functional element along the DNA sequence of a genome by identifying the location and function of genes and regulatory regions (Abril and Castellano 2019). Annotations in plant genomes can be divided into structural and functional annotations. Structural annotation refers to finding gene structures in the DNA sequence, which is usually done using gene prediction software, such as AUGUSTUS (Stanke and Morgenstern 2005). Functional annotation defines the function of a sequence, and it is possible to do so through protein alignment against a protein database (Ouyang *et al*. 2009).

More than a decade ago, the first genome of a legume, *Lotus japonicus* (Sato *et al*. 2008), was sequenced, and the genomes of 40 different legume species were sequenced and data stored in the LIS database (https://www.legumeinfo.org/, accessed on 02/06/2024). The first genome of *P. vulgaris,* organized into 11 chromosomes, was published in 2014 (Schmutz *et al*. 2014) for the genotype G19833, belonging to the A gene pool, with a relatively small size of 521.1 Mb, 28,134 annotated genes and 2,668 complete long terminal repeat (LTR) retrotransposons. The development of PacBio technology allowed the sequencing of version 2 of the G19833 genome in 2018 (*Phaseolus vulgaris* v2.1, DOE-JGI and USDA-NIFA, http://phytozome.jgi.doe.gov/), the size and number of annotated genes varied from version 1, being 537.2 Mb and 27,433 genes. To date, ten different bean genomes have been reported in public databases. The publication of genomes for different genotypes in cultivated species revealed the presence of variation in the type presence/absence (PAV), number of copies of some genes (CNV), and large-scale structural variation (SV).

WGS and the comparison between them provide the opportunity for researchers to identify genetic variations throughout the entire genome, such as SNPs, insertions, deletions, and structural variations, and use these variations to conduct more targeted and specific studies aligned with global demands (de Souza *et al*. 2023). The way to describe this genomic variation within a species is with a "Pan-genome" and could divide the genome into the core genome which contains the common genes between all the individuals, and the dispensable or variable genome consisting of partially shared DNA sequence elements. It has been reported that dispensable regions in crop plants are enriched with genes associated with agronomic traits (Tirnaz *et al*. 2020); therefore, their identification is a useful tool in breeding programs. Some legume species have pan-genomes comprising many accessions such as soybean with 2,898 deeply sequenced accessions (Liu *et al*. 2020) or chickpea with 3,366 genomes (Varshney *et al*. 2021b).

3.4. Whole-transcriptome analysis

NGS can acquire an unprecedented amount of data in a short time, and deep sequencing has rapidly transformed RNA research as well. The transcriptome is the entire set of RNA transcripts in a given cell. Microarray technologies have been used for high-throughput large-scale RNA-level studies, with a limit in the capacity of catalog and quantify RNA molecules expressed under various conditions (Yang and Kim 2015). RNA-Seq is currently the method of choice for studying gene expression and identifying novel RNA species. RNA-Seq uses recently developed deep-sequencing technologies for transcriptome sequencing using cDNA sequencing (Mortazavi *et al*. 2008; Wang *et al*. 2009). Compared with other methods, RNA-Seq offers less background noise and a greater dynamic range for detection, and it can directly reveal sequence identity, which is crucial for the analysis of unknown genes and transcript isoforms (Hrdlickova *et al*. 2017).

The RNA-Seq workflow involves several steps, including RNA isolation and fragmentation, cDNA generation, library amplification, sequencing, and mapping to a reference genome. Many software programs are available for transcript quantification and functional profiling (Love *et al*. 2016). RNA-Seq can be used for different purposes, such as differential expression analysis, transcriptome assembly, construction of expression

atlas, network analysis, and structural alterations. The RNA-Seq has been used in legumes studies to recognize development, stress response, composition or nitrogen fixation (Afzal *et al*. 2020). It is divided into two technologies: short-read (50–200 bp), which is dominant in plant sciences and provides the majority of the public datasets, and long-read (1,000 – 50,000 bp), which is particularly suited for *de novo* transcriptome assembly and identification of novel transcripts and isoforms (Tu *et al*. 2022).

The reference genomes of many organisms remain incomplete, and numerous novel genes and transcripts are yet to be discovered. RNA-Seq is also a very useful tool for identifying new genes in reference genomes, which may be incomplete, or for annotating new genomes by mapping RNA reads to the genome (Chen *et al*. 2017). The first reference genome of *P. vulgaris* was annotated with the help of Illumina RNA-seq data (Schmutz *et al*. 2014), and a Gene Expression Atlas was published in 2014 to facilitate functional genomic studies in common bean (O'Rourke *et al*. 2014).

4. Tools to study the genetic basis of characters

Knowledge of the inheritance of traits is essential to predict the phenotypes of progenies and, consequently, both for crop management and planning of breeding programs. The classical approach to the study of inheritance (forward genetics) is based on the development of mapping populations, phenotyped, genotyped, and genetic analysis of segregations. Following this approach, numerous major genes and Quantitative Trait Loci (QTL) have been described (e.g. see LIST OF GENES - *Phaseolus vulgaris* L.; http://Bean_Genes_List_2017.pdf). Currently, the availability of genomes allows us to go one step further and identify candidate genes that control specific traits. The inheritance studies were conducted on controlled populations, such as biparental (e.g., recombinant inbred populations, F2 population, etc.) or multi-parental populations (e.g., multi-parent advanced generation intercross population or nested association mapping populations).

4.1. Biparental populations

Biparental populations are genotypes derived from a cross between genotypes in which different traits are segreganted. The first step in developing linkage maps and the identification of quantitative trait loci (QTL and genes linked to traits of interest is the

development of genetic mapping populations by crossing genetically divergent parents. In linkage mapping, biparental populations, such as recombinant inbred lines (RIL) or F_2 population, are commonly used (Rani *et al*. 2023). A RIL population is a set of homozygous lines developed from a cross between two parental lines that differ in the traits of interest. RIL populations are created by self-pollination of the F2 population for several self-pollinated generations, typically six or more. Many RIL populations have been reported in common beans and used to generate genetic maps on which QTLs for many traits were located. One of the first was the XC population obtained from the cross 'Xana'/Cornell, in which many qualitative traits and QTL were mapped (Pérez-Vega *et al*. 2010; Campa *et al*. 2014; Murube *et al*. 2020). Near isogeneic lines (NIL) are a particular type of genotypes derived from biparental crosses, usually backcrossing. NILs are identical genotypes except for a specific genomic region or loci. The use of NIL to accumulate disease resistance genes in the same breeding line is very useful and also combining it with GBS allows delimiting the regions that control the resistances (Ferreira *et al*. 2017).

HTG has greatly boosted the capability of QTL mapping, and some software has been developed to detect QTLs with phenotyping and genotyping data of a mapping population (Rani *et al*. 2023). QTL mapping relies on detecting correlations between genetic markers and phenotypic traits in segregating populations. The use of RIL populations in QTL mapping has great advantages, such as finer mapping, which is the result of the high number of recombination events due to the multiple selfing processes, and the substantial contribution to QTL mapping of the species for various phenotypes when the RIL is established with all the genotypes fixed as homozygotes (Takuno *et al*. 2012). The efficiency of the NILs in identifying closely linked markers largely depends on the level of variation in the locus under study between the donor and recurrent parents (Rafalski *et al*. 1996).

4.2. Diversity Panels

Diversity panels are selected sets of genotypes that capture a large proportion of the genetic diversity within a species and a wide range of phenotypes. Different panels have been reported and used in association studies in common beans, such as the Andean

Diversity Panel (ADP; (Cichy *et al*. 2015a), Middle American Diversity Panel (MDP; (Moghaddam *et al*. 2016), Snap Bean Panel (SBP; (Campa *et al*. 2024), or the Spanish Diversity Panel (SDP) established in SERIDA from the local Spanish germplasm, as well as old and elite cultivars mainly used for snap consumption (Campa *et al*. 2018; see https://zenodo.org/records/10263706). SDP was used to explore the phenolic content variation in the panel (Rodríguez Madrera *et al*. 2021), and also to identify and validate genomic regions associated with resistance to white mold (Campa *et al*. 2020), and pod morphological traits (García-Fernández *et al*. 2021).

Linkage analysis for QTL mapping was the precursor of genome-wide association study (GWAS), which is another way to relate phenotypes to genotypes and genomes. To carry out a GWAS, phenotype and genotype data needs to be collected from a large sample of individuals such as a diversity panel. Genotype data are usually single nucleotide polymorphisms (SNPs). The genetic markers most strongly associated with the phenotype of interest were identified using different statistical approaches (Tibbs-Cortes *et al*. 2021). Several statistical models are available to identify the associations between marker loci and phenotypes, which are increasingly complex. Using covariates as population structure and kinship matrix in the model, as does the Linear Mixed Model (MLM) (Yu *et al*. 2006), can control for these confounding factors that lead to false negatives. In contrast to MLM, which is a single-locus approach, other multilocus models have been recommended (Wang *et al*. 2016; Zhang *et al*. 2020) for complex traits that are controlled by numerous loci simultaneously, and consider the information of all loci simultaneously (Kaler *et al*. 2019). Another factor to be considered in association studies is linkage disequilibrium (LD), which refers to the non-random association of alleles at different loci, which could determine the number of markers to cover the entire genome, with fewer markers needed in self-pollinated crops because of the typically larger LD decay compared to cross-pollinated crops. GWAS allows us to unravel the genetic contributors to complex traits where an individual's genes may have a minor influence, leaving much of the heritability unaccounted for (Sahito *et al*. 2024). Unlike family-bases populations, association mapping populations, such as diversity panels, allow a higher resolution mapping that can be attribute to the historical recombination events and the greater allele numbers incorporated. In these populations, historical recombinations that accumulated over

generations persist among the representative accessions improving the resolution of mapping through the rapid decay of LD (Alqudah *et al*. 2020).

The identification of false positives is a problem presented in association studies. This may be due to multiple factors, such as accuracy in phenotyping, population size, trait heritability, population composition, statistic method used, or variations with the reference genome. Therefore, validation of the results of these studies is recommended. There are different possibilities to validate the results of a GWAS like an analysis in biparental populations, comparison of studies, comparison of genome sequences, analysis of differential expression, or gene edition.

5. Common bean breeding

It is estimated that by 2050, agriculture will have to feed nine billion people, and the average income in the developing world is rising. Climate change, urbanization, land degradation, and the limited availability of inputs put pressure on food supply, and meeting future food demands will be a challenge. To address these food security challenges adequately, new crop cultivars must be developed, which must be performed faster (Lenaerts *et al*. 2019). Most cultivars are derived from plant breeding programs. Plant breeding aims to obtain new genotypes (syn. cultivars), with characteristics that present advantages over previously existing cultivars, providing an increase in production (kg), yield (kg/ha), or other qualities. To conduct breeding programs, breeders must define the traits to be improved, have inherited variations, and have tools for identifying the best individuals.

Breeding objectives depend on the needs of the production area or the requirements of each market and genotype. Many breeding programs for common beans focus on enhancing their resistance to biotic and abiotic stresses, such as disease resistance and drought tolerance (Singh and Schwartz 2010; Assefa *et al*. 2019; Mukankusi *et al*. 2019). Diseases are responsible for large losses of bean crops (Flood, 2010; Carvajal-Yepes *et al*. 2019; Ristaino *et al*. 2021; Singh *et al*. 2023), and European regulations that limit the use of pesticides make the availability of resistant crops necessary. Many efforts have also focused on morphological and quality characteristics such as plant architecture and seed and pod phenotypes. Finally, another important group of traits considered in bean

breeding programs is related to nutritional and sensory qualities. These efforts are crucial to meet consumer preferences across diverse markets (Assefa *et al*. 2019).

Breeders require available and heritable variation. In the last century, many studies have investigated the inheritance of important traits in common bean breeding and many major genes and QTL have been mapped using genetic linkage maps. However, with the publication of the bean genome (Schmutz *et al*. 2014), boosting the connection between the phenotype, genotype, and genome is necessary. This indicated the identification of annotated genes that control specific traits. These data provide relevant information for understanding the complex network of genes controlling specific phenotypes and allow the development of functional markers (FM) and DNA markers derived from functionally characterized sequence motifs within genes directly linked to phenotypic traits. FMs play a crucial role in plant breeding by aiding the fixation of beneficial alleles in breeding populations, the selection of complex traits, and MAS for crop improvement (breeding precision).

Breeders have always sought to accelerate improvement programs and obtain new cultivars efficiently, in terms of time and resources. For legume crops, developing new cultivars takes several years, because of the need to cross selected parent lines and perform 4-6 generations of inbreeding to have genetically stable lines for evaluation. Currently, some methods are available to reduce this time-consuming work. It is possible, for example, to increase the number of generations per year with the help of photoperiod-controlled conditions and temperature regulation to manipulate the growing environment, "Speed breeding" (Watson *et al*. 2018; Bhatta *et al*. 2021). Another way to identify the best genotypes in breeding programs is through cutting-edge phenotyping methods, such as digital image-based plant phenotyping (Abebe *et al*. 2023), phenotyping automatic centers (e.g., https://phenospex.com/), and aerial phenotyping with drones (Parker *et al*. 2020). Concerning identifying desirable genotypes, Marker-assisted selection (MAS) or Marker-assisted backcrossing (MABC) is widely used in modern plant breeding, as an indirect selection. MAS and MABC are based on using molecular markers to identify genes that control desirable traits (Hasan *et al*. 2015). An extension of MAS is genomic selection (GS), a method to predict the breeding values of progenies by association (Goddard and Hayes 2007; Budhlakoti *et al*. 2022).

Finally, plant breeders have shown an interest in genome editing approaches for genetic manipulation that use clustered regularly interspaced short palindromic repeats (CRISPR/Cas9), which implies a deep knowledge of the genetic control of traits. This product is now more feasible because of its looser regulations and widespread acceptance. This is an easy, quick, and efficient method for directly altering cellular genetic sequences (Ahmad 2023). Markers have been widely used in common bean breeding, and even with the recalcitrant nature of the transformation of this species, genetic manipulation using CRISPR has already been possible in root tissue (de Koning *et al*. 2023). All of these technologies require deep knowledge of the locus that controls specific characters of interest.

OBJECTIVES

The **main objective** of this Thesis was to increase the resources and knowledge of bean genetic diversity and the genetic control of important traits in bean breeding to contribute to the efficient handling and use of genetic diversity and to accelerate breeding programs. Three specific objectives have been addressed in this study, corresponding to the four chapters of this Thesis:

- **Objective 1.** Genotypic characterization of a diversity panel within the Fabada market class for efficient conservation, use of bean genetic diversity, and erosion study.

  CHAPTER 1. Genetic erosion within the Fabada dry bean market class revealed by high-throughput genotyping.

- **Objective 2.** Identification of Quantitative trait loci and candidate genes for important breeding characters to accelerate the development of new cultivars through precision breeding.

  CHAPTER 2. Differentially expressed genes against Colletotrichum lindemuthianum in a bean genotype carrying the Co-2 gene revealed by RNA-sequencing analysis.

  CHAPTER 3. Identification of consistent QTL and candidate genes associated with seed traits in common beans by combining GWAS and RNA-Seq.

- **Objective 3.** Development of new plant genomic resources. Whole-genome sequencing and annotation of a common fabada market-class bean.

  CHAPTER 4. A new bean genomic resource: de novo assembly and annotation of a Fabada *cultivar.*

CHAPTER 1

Genetic erosion within the Fabada dry bean market class revealed by high-throughput genotyping

A version of this Chapter has been published as:

Jurado M, García-Fernández C, Campa A, Ferreira JJ (2023) Genetic erosion within the Fabada dry bean market class revealed by high-throughput genotyping. *The Plant Genome*. 2023;16:e20379. https://doi.org/10.1002/tpg2.20379

Genetic erosion within the Fabada dry bean market class revealed by high-throughput genotyping

The Fabada market class within the dry beans has a well-differentiated seed phenotype with very large white seeds. This work investigated the genetic diversity maintained in the seed collections within this market class and possible genetic erosion over the last 30 years. A panel with 100 Fabada accessions was maintained in seed collections for 30 years, 57 accessions collected from farmers in 2021, six cultivars developed in SERIDA, and 16 reference cultivars of different market class were gathered and genotyped with 108,585 SNPs using the genotyping-by-sequencing method. Filtering based on genotypic and phenotypic data was carried out in a staggered way to investigate the genetic diversity among populations. The dendrogram generated from genotyping revealed 90 lines forming 16 groups with identical SNP profiles (redundant lines) from 159 lines classified as market-class Fabada according to their passport data. Seed phenotyping indicated that 19 lines were mistakenly classified as Fabada (homonymies), which was confirmed in the dendrogram built without redundant lines. Moreover, this study provides evidence of genetic erosion between the population preserved for 30 years and the currently cultivated population. The conserved population contains 54.6% segregation sites and 41 different SNP profiles, whereas the cultivated population has 19.6% segregation sites and 26 SNP profiles. The results allow for the more efficient preservation of plant genetic resources in genebanks, minimizing redundant accessions and incorporating new variations based on genotypic and phenotypic data.

1. Introduction

The common bean (*Phaseolus vulgaris* L.) is the second most important cultivated legume species in the world (33 million ha; FAO (2020)) and the most important legume for human consumption. The common bean is a diploid ($2n = 2x = 22$) and self-pollinated species domesticated by MA and A cultures. Two gene pools corresponding to those geographical regions have been reported for both wild and cultivated common beans based on morphological traits, seed proteins, and molecular marker variation (Singh *et al*. 1991; Kwak and Gepts 2009; Rodriguez *et al*. 2016). From these regions, the species was progressively dispersed worldwide (Gepts *et al*. 1986; Zhang *et al*. 2008; Asfaw *et al*. 2009; Angioi *et al*. 2010). In Europe, local germplasm from both gene pools are present, as well as intermediate genotypes probably derived from recombination between the two gene pools (Santalla *et al*. 2002; Angioi *et al*. 2010; Campa *et al*. 2018; Bellucci *et al*. 2023).

Bean seeds exhibit wide phenotypic diversity, including size, shape, and coat-color variations (Singh 1989; Voysest 2000). Seed coat colors vary widely (white, cream, yellow, brown, red, purple, and black) and have different intensities, as well as patterns that combine different colors (e.g., bicolor, mottled, and spotted). Cultivars with 20–100 g/100 seed-weights have been observed (Voysest 2000). Dry beans have been grouped into differentiated seed-based phenotypic groups or market classes, such as Navy, White Kidney, Great Northern, Canellini, Fabada, Yellow, Carioca, Small Red, Red Mexican, Red Kidney, Cranberry, Pinto, and Black. Among them, market-class Fabada (syn. Faba granja and favada) has a well-differentiated seed phenotype in the species with very large white seeds (~100 g/100 seeds) having an oblong shape with a length/width ratio greater than 2.2. Fabada had already been described in the north of Spain by the mid-20th century (Puerta Romero 1961). In the last 30 years, the Fabada crop has undergone notable changes: transition from bean-maize intercropping to monoculture, expansion of cultivation to new areas, diversification of uses, emergence of new cultivars from breeding programs, and modernization of farming methods. Landraces have climbing indeterminate growth habits, whereas modern cultivars can have both determinate and climbing indeterminate growth habits (Ferreira *et al*. 2017). Previous analyses indicated that Fabada landraces were not a homogeneous group. Santalla *et al*. (2002) reported

variations in this market class based on allozyme and phaseolin polymorphisms. Three genotypes of this market class (cv. 'Andecha', 'Maruxina', and 'Xana') were included in the SDP, and they were not grouped, with one of the two main gene pools showing an intermediate position (Campa *et al*. 2018).

Measures of genetic diversity in cultivated plants and their wild relatives are needed to make decisions, monitor changes, and warn of emerging problems in agricultural production (Brown and Hodgkin 2015). Most diversity studies in common beans have focused on a wide range of phenotypes from many places or maintained in wide seed collections (Angioi *et al*. 2010; Campa *et al*. 2018). The classification of bean-seed phenotypic diversity based on market classes is useful to describe the diversity or variation as a first approach, but within the same market class, there may be different levels of phenotypic and genotypic variation. Moreover, there are seed phenotypes that hardly fit the described market classes. A detailed characterization of variation in the market classes will help to efficiently preserve genetic diversity and to better differentiate and identify prominent genotypes. For diversity preservation, it is relevant to know the amount of variation to efficiently maintain and use diversity. Additionally, the changes in the diversity cultivated over time and possible genetic erosion provide interesting information on the diversity preserved in the *ex-situ* collections. The literature reports cases of lost genetic diversity in cultivated species and landraces over the past century and continuing into the present (Khoury *et al*. 2022). Crop erosion can occur at the level of crop species, variety, or allele. Genetic erosion, understood as a reduction in allelic evenness and richness, is directly related to the breeding capabilities, vulnerability, evolutionary potential, and resilience of crops (van de Wouw *et al*. 2010; Brown and Hodgkin 2015; Fu 2015). Crop genetic diversity has traditionally been analyzed using morphological traits. However, genetic diversity changes within a varietal type are more difficult to document owing to the limited number of phenotypic markers. At present, HTG methods (e.g., GBS; (Elshire *et al*. 2011)) identify many markers per genotype in comparison to a reference genome and provide a picture of the genetic diversity landscape.

A large collection of landraces representing the Spanish diversity of the Fabada market class is maintained in the CRF-INIA-CSIC (Madrid; https://bancocrf.inia.es/es/).

Most of these accessions were collected before 1991, and a duplicate collection has been maintained in the SERIDA collection for 30 years. On the other hand, new cultivars of the Fabada market class have been produced by different plant breeding programs (Ferreira *et al*. 2012) and disseminated over the last 20 years, such as cv. 'Andecha', 'Maruxina', 'Maximina', and 'Xana'. Thus, there is an opportunity to investigate the genetic erosion within this market class, which could be extrapolated to global conditions in which the rapid modernization of crops has occurred. This work investigated the genetic diversity of landraces grouped in the market class Fabada and the possible loss of genetic diversity in the currently cultivated material. The results are relevant for the efficient preservation of species diversity as well as the suitable use of genetic diversity in this market class.

2. Materials and methods

2.1. Plant material

A total of 179 *P. vulgaris* accessions were gathered into a panel (FabaPanel). Table 1S.1 presents the list of materials included in this study as well as the respective passport data. The FabaPanel contains 100 accessions conserved in the SERIDA seed collection and recorded as Fabada market class in respective passport data (conserved population; code FP032 to FP456). Most of these Fabada accessions were collected in Northern Spain before the 1990s. Moreover, 57 Fabada accessions with indeterminate growth habits were collected in Northern Spain from local farmers in 2021, and they were also included in the FabaPanel (cultivated population: FP500 to FP559). Six breeding lines developed in SERIDA having the Fabada seed phenotype were added as a control (cultivar population): A25 (cv. 'Andecha', an old cultivar marketed since the early 2000s; Figure 1.1), and the lines B8, 'Xana', A2806 (cv. 'Maximina', distributed since the early 2010s), X4562, and A4804. In addition, 16 well-known genotypes were added to this panel as a reference diversity population: AB136, BAT93, Cornell49242, 'Cannellini', DOR364, ''Musica'', G19833, 'Garrafal Oro', 'La Victorie', IVT7214, MDRK, 'Planeta', SanilacBc6Are, Tendergreen, TU, and Midas (reference population). Those reference genotypes are also included in the SDP (Campa *et al*. 2018).

Figure 1. 1. Seed phenotype of the Fabada market class. The bars represent 1 cm.

A homozygous line per accession was obtained by self-pollinating individual plants derived from each accession. The crop was developed in a greenhouse, and it was used to collect the tissues for genotyping and phenotyping of the harvested seeds. Seed phenotyping was based on their visual characteristics, including color, size, and shape.

2.2. DNA isolation and genotyping

Genomic DNA was isolated from young leaf samples using the SILEX method (Vilanova *et al*. 2020), and DNA quality was checked in agarose gels. The GBS method was carried out following Elshire *et al*. (2011) and optimized by Schröder *et al*. (2016) using the TaqαI and MseI restriction enzymes. Library construction was performed following the protocol of Poland *et al*. (2012) with modifications in the adaptors for ligation. In total, 20 barcoded samples were pooled for PCR amplification. Sequencing was performed in the Illumina platform by Macrogen Inc.

SNP calling was carried out by AllGenetics&Biology SL (www.allgenetics.eu) using the reference genome of *P. vulgaris* (Pvulgaris_442_v2.0) obtained from the JGI Data Portal (https://phytozome-next.jgi.doe.gov/info/Pvulgaris_v2_1).

2.3. Panel filtering

A stepped filtering method was developed to identify redundant or off-type lines. Genotyping data of the FabaPanel were filtered with the help of the software Tassel v5 (Bradbury *et al*. 2007). First, the homozygous genotypes with more than 50% missing data and SNPs located out of the 11 bean pseudo-chromosomes were removed from the FabaPanel. Then, the constituted subsets of the FabaPanel were filtered using the following criteria: the proportion of missing data (<10%) and minor allele frequency (MAF > 0.05 when the reference genotypes were included; MAF > 0.01 when the reference genotypes were not included). Finally, the lines classified as Fabada according to their passport data but having seed phenotypes that did not correspond to this market class were removed.

2.4. Population diversity and clustering analyses

The R package SambaR_v1.08 (de Jong *et al*. 2021) was used to import raw data files into R and perform the diversity analysis. Principal coordinate analyses (PCoA) based on Hamming's distance were conducted using the function ape_pcoa(). Population differentiation measures for all pair-wise population comparisons (Fst and Nei's distance) were performed with the function calcdistance().

Dendrograms were built from Euclidean distance using the unweighted pair group method with arithmetic mean method for clustering analysis with the help of the packages "ggplot2" (Wickham 2016), "FactoMinerR" (Lê *et al*. 2008), "factoextra" (Kassambara and Mundt 2020), "cluster" (Maechler *et al*. 2022), and "ape" (Paradis and Schliep 2019) in R software (R Core Team 2023).

The numbers of segregating sites (SNP) and SNP profiles per population were used to estimate the genetic diversity and putative genetic erosion. An SNP profile is defined by the same genotype for all SNPs. The number of segregating sites per population was obtained with the help of the software Tassel v5. The number of SNP profiles was obtained from the dendrogram constructed after filtering the populations. Finally, the diversity per population was estimated using the Shannon Diversity Index [H' = −Σpi * ln(pi)], where pi represents the proportion of the SNP profile i. The Shannon equitability index (which measures the evenness of profiles in a population) was also estimated as EH

= H′/ln(S), where H′ and S represent the Shannon index and the number of SNP profiles, respectively.

3. Results

3.1. Genotyping of the Faba Panel

Sequencing of the GBS libraries yielded approximately 90.7 million reads per line (an average of 816.4 million reads per library for 20 lines), and the Q20 value of each library was greater than 97.24%. The GBS analysis generated 108,585 SNPs. The genotyping data supporting this study are available at the Dryad repository (https://doi.org/10.5061/dryad.djh9w0w5d) Based on this genotyping, a staggered filtering process was carried out to specifically study the diversity of the Fabada market class (Figure 1.2). First, four lines showed missing data for more than 50% of the obtained SNPs (FP173, FP174, FP508, and F541) and were therefore removed (FilterFabaPanel1). The genotyping of the remaining 175 lines was filtered considering homozygous sites, location in one of the 11 bean chromosomes, missing values, and MAF, resulting in 22,259 SNPs. The number of SNPs per chromosome ranged from 1,214 for chromosome Pv06 to 3,088 for chromosome Pv11 (Figure 1S.1). After filtering and thinning the data, the mean proportion of missing data per individual was 3%, the GC content was 0.5, and the transition versus transversion ratio was 2.28. The most frequent transition and transversion events were A/G and C/T, respectively.

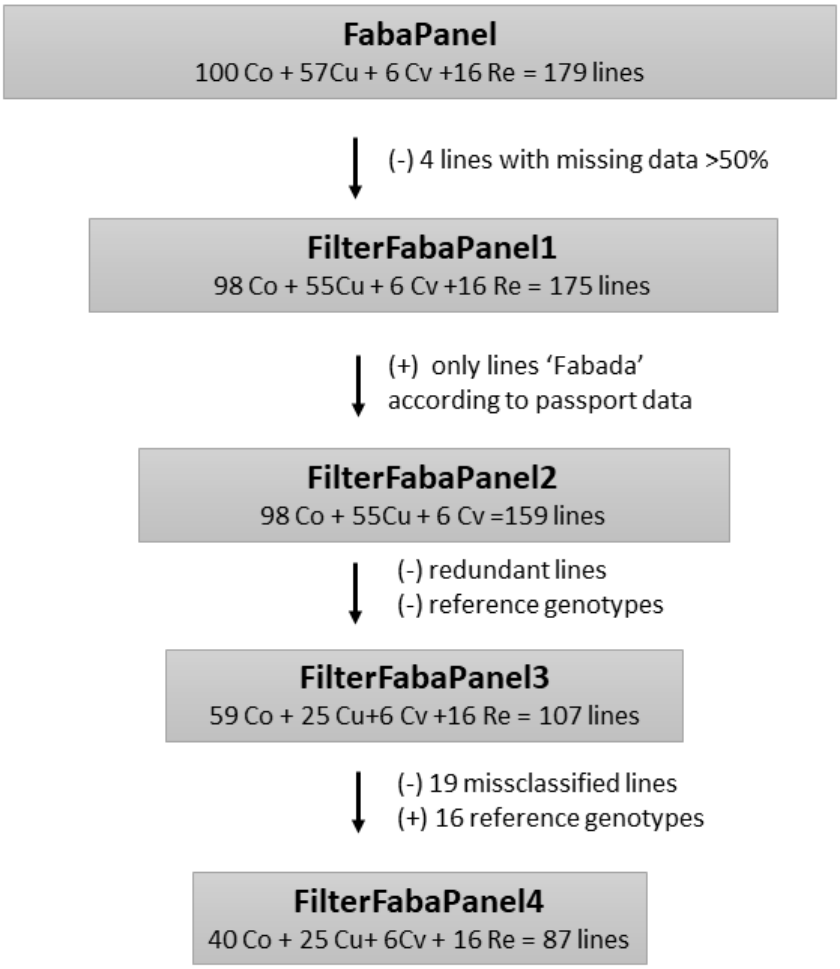

Figure 1. 2. Established subsets from FabaPanel to investigate changes in diversity in the Fabada market class. The composition of each subset and the filtering criteria are also indicated. The lines included in each subset are described in Table 1S.1. Co: conserved population; Cu: cultivated population; Cv: cultivars; Re: reference population.

3.2. Filtered from genotyping

Table 1S.1 presents the lines included in the subsets constituted to develop this study. A subset, named FilterFabaPanel2, was specifically designed from FilterFabaPanel1 to investigate the redundancies in the market class "Fabada" (see Figure 1.2). This FilterFabaPanel2 had 159 lines, all of them recorded as "Fabada" in the respective passport data (or Faba granja), and was genotyped with 21,837 SNP after filtering: 98 lines from the SERIDA collection (conserved population), 55 lines collected in 2021 (cultivated population), and six breeding lines (cultivar population; see Table 1S.1). A

PCoA based on Hamming's distance showed two main coordinates that explained a total of 93.9% of the variance (69.2% and 24.7%, respectively). The scatter plot built with these two main components (Figure 1S.2) exhibited a higher dispersion in the lines included in the conserved population than in the other two populations, the cultivar and cultivated populations. The plot also revealed many overlapping lines. In parallel, a dendrogram was constructed (Figure 1S.3), showing that 90 lines were grouped in 16 groups with more than one line with identical SNP profiles (41% of redundant lines or duplicate accessions). For instance, it was the case of the lines maintained in the SERIDA collection: FP102, FP103, FP106, FP108, FP115, FP119, FP133, FP147, FP148, FP154, and FP182. Additionally, 69 lines had unique SNP profiles, so 85 SNP profiles were detected for FilterFabaPanel2 (69 unique + 16 non-unique). However, we also observed closely related lines that differ by a few SNP; for example, the breeding line A2806 and the group with the lines FP501, FP524, and FP528 differ by two SNP, and the line A25 and the group with the lines FP502, FP540, and FP559 also differ by two SNP. Finally, the dendrogram also indicated that most of the lines collected in 2021 (cultivated population) were closely related and grouped together the line A25.

Redundant lines in FilterFabaPanel2 from the first generated dendrogram were removed, maintaining one genotype per SNP profile and population (conserved, cultivated, and cultivars). The resulting set contained 91 lines (Table 1S.1) that were used to build a new subs*et al*ong with the 16 lines from the reference population (FilterFabaPanel3, 107 lines). The PCoA of FilterFabaPanel3 revealed two main coordinates explaining 93.1% of the variance, and the generated plot showed wide dispersions for the conserved and reference populations (Figure 1.3). In contrast, the lines of the cultivar and the cultivated population were less dispersed.

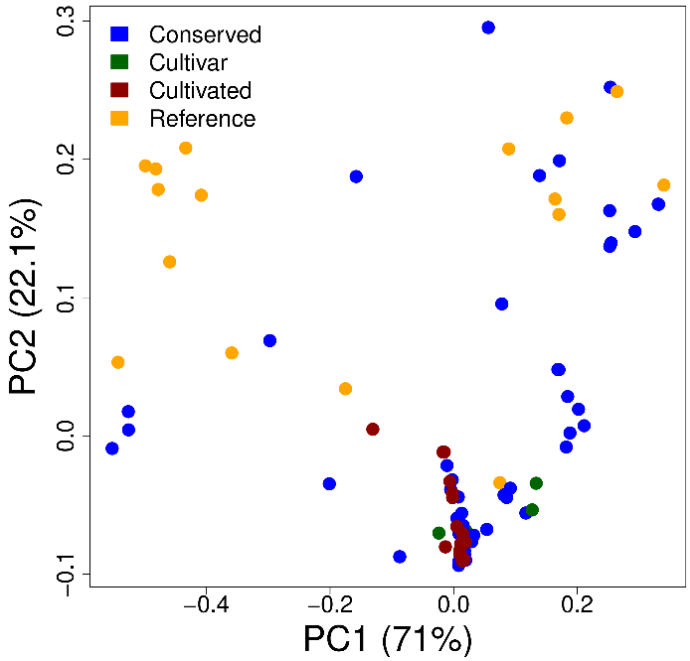

Figure 1. 3. Scatter plot obtained with the two mains coordinates revealed by the principal coordinate analysis (PCoA) based on Hamming's distance of the 107 lines genotyped with 21,618 SNPs.

As shown in Figure 1.4, the dendrogram has two main branches. The "A" group contained 13 lines, including typical Mesoamerican cultivars such as Sanilac, Cornell49242, and AB136. Group "A" included five lines from the conserved population (FP192, FP175, FP156, FP061, FP093), which were all recorded as Fabada market classes in their respective passport data. The "B" group contained 91 lines, including typical Andean cultivars such as MDRK, Tendergreen, and G19833. Most of the lines classified as the market class Fabada were classified as Group B, and the following four subgroups were established:

- **Group B1**, is formed by 16 lines, including FP139, FP125, FP437, FP453, FP110, MDRK, 'Garrafal Oro', FP354, and G19833, one of the bean genomes available (Schmutz *et al*. 2014).
- **Group B2**, is formed by three snap bean cultivars: Tender-green, Midas, and 'La Victorie'.
- **Group B3**, consisted of 71 lines that included the six cultivars of the Fabada market class, 39 lines from the conserved population, and 25 lines from the

collected population. The reference cultivar Canellini was included in this group and was close to the cultivars 'Xana' and X4562, which both have determinate growth habits.

- **Group B4**, is formed by the remaining two lines, FP037 and FP555.

Finally, the lines TU and FP293 were located far from the four B subgroups described above.

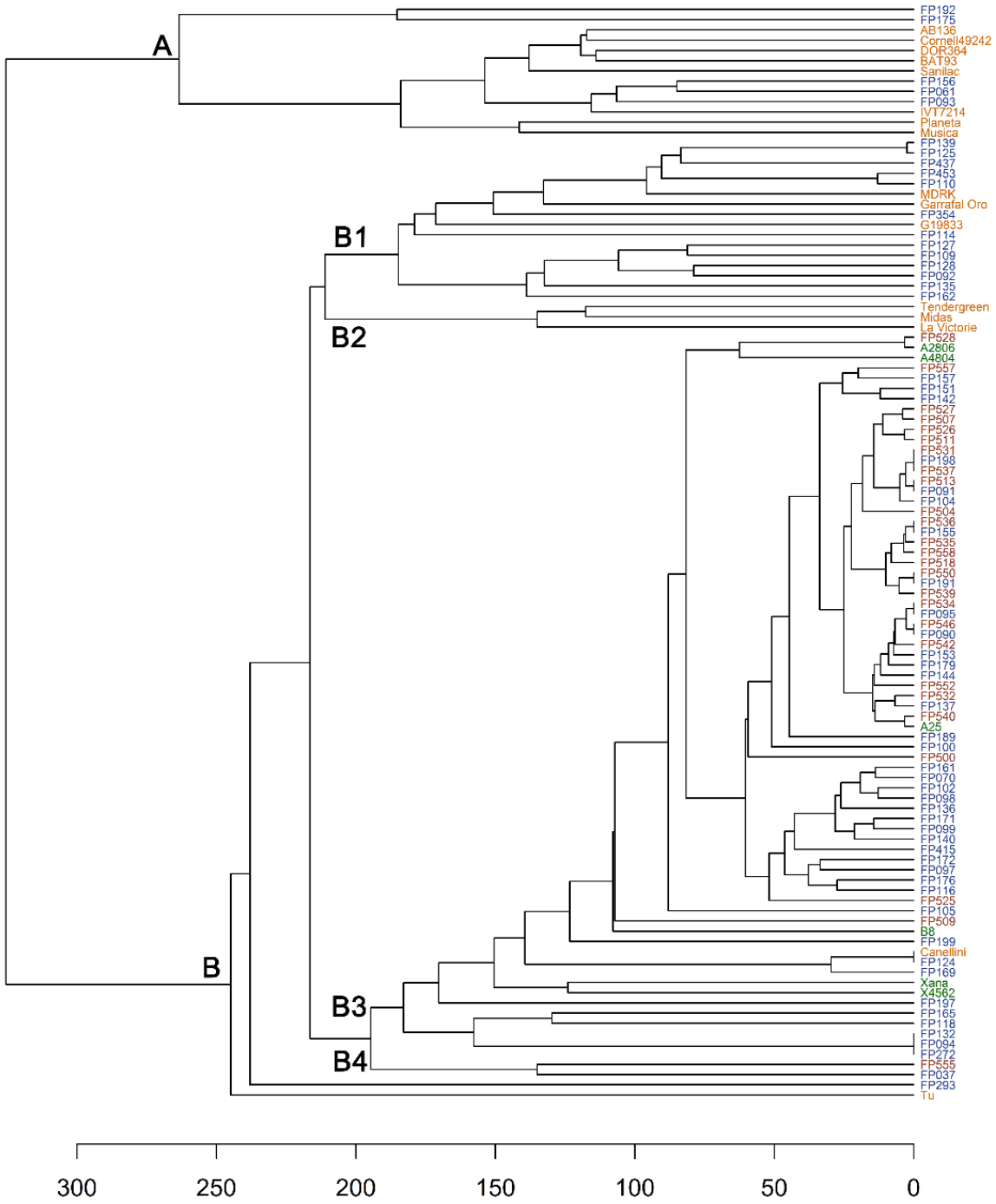

Figure 1. 4. Dendrogram generated from 107 lines genotyped with 21,618 SNPs using the Euclidean distance and the unweighted pair group method with arithmetic mean clustering method (FilterFabaPanel3).

3.3. Filtered from phenotyping

The harvested seeds from self-pollinated plants in the greenhouse were phenotyped by considering color, size, and shape. The phenotyping indicated that 19 lines registered as Fabada market class in their respective passport data did not have the full characteristics of the market class Fabada (homonymy; the same local name but different seed phenotype): FP037, FP061, FP093, FP110, FP114, FP118, FP124, FP125, FP128, FP135, FP139, FP156, FP165, FP175, FP192, FP293, FP354, FP437, and FP453 (see images at https://zenodo.org/records/7015279; Table 1S.1). These lines had white and oblong seeds, but they were of a smaller size than those of the Fabada market class. These 19 lines were located in the dendrogram outside the main group containing cultivar A25 (Group B3). The lines FP092, FP094, and FP272 were difficult to classify because their seed phenotypes were similar to the Fabada market class, with seed size being intermediate between the large Canellini and the Fabada market class.

3.4. Genetic diversity in the Fabada market class

To investigate the specific genetic diversity in the Fabada market class, a subset (FilterFabaPanel4) of the FabaPanel was created after removing redundant genotypes and mis- classifications in the conserved population, which had been revealed by the previous analysis. This subset of FabaPanel had 88 lines genotyped with 21,618 SNPs after filtering, and it contained the six cultivars: 26 lines from the cultivated population, 40 lines from the conserved population, and 16 lines from the reference population. The distribution of the segregating sites per population is shown in Figure 1S.5, and the percentages of segregating sites per chromosome in the four populations are shown in Table 1.1. The highest percentage was observed in the reference population (mean 99.68%), followed by the conserved population (54.56%). The cultivated population exhibited lower percentages of segregating sites than the conserved population for all the chromosomes except chromosome Pv11. The cultivar population had a higher percentage of segregating sites on chromosome Pv01 than the cultivated and conserved populations. Very high polymorphism levels (>90%) were found in the conserved population for chromosomes Pv04, Pv05, and Pv08. The estimations of the diversity indices using the SNP profiles showed a higher Shannon diversity index (H') for the conserved population (Table 1.1), whereas the lowest value was observed in the cultivar population. This

population contained only six breeding lines, all with different SNP profiles. Among the four populations, the fixation index (Fst) ranged from 0.018 (cultivar and conserved populations) to 0.329 (reference and cultivated populations), whereas the Neit's genetic distance varied between 0.028 (cultivated and cultivar populations) and 0.31 (reference and cultivated populations; Table 1S.2).

Table 1. 1. Diversity assessments in the FilterFabaPanel4.

		Segregating sites (%) per population			
Chr	**No. SNP**	**Cultivars**	**Cultivated**	**Conserved**	**Reference**
Pv01	2265	40.75	3.44	9.62	99.91
Pv02	1473	3.46	19.42	69.31	99.32
Pv03	2237	29.50	5.19	72.69	99.82
Pv04	2512	62.90	25.48	91.76	99.84
Pv05	2058	48.88	42.66	94.46	99.90
Pv06	1171	65.16	4.18	79.68	99.91
Pv07	1465	11.54	2.87	51.88	99.25
Pv08	2630	5.63	26.46	92.55	99.96
Pv09	1410	1.56	1.56	9.08	99.79
Pv10	1515	6.01	0.46	3.96	99.80
Pv11	2882	20.92	84.25	25.16	98.99
Mean		26.94	19.63	54.56	99.68
S		6	25	40	16
H'		0.17	1.79	2.64	0.47
E_H		0.09	0.38	0.49	0.17

Note: The lines included in this subset are described in Table 1S.1. Percentage of segregating sites (single-nucleotide polymorphism [SNP]) per chromosome in the four populations considered: conserved, cultivated, cultivar, and reference. Abbreviations: S, richness of genotypes (SNP profiles); H', Shannon diversity index; E_H, Shannon equitability index.

A map with the segregating sites along the bean genome showed relevant differences among the Fabada lines compared with the A25 line (Figure 1S.5). For example, the lines FP555 and FP509 show SNPs located on chromosome Pv11. The determinate cultivars 'Xana' and X4562 present many SNPs located at the end of Pv01, where the gene *fin* (introgressed in these lines) is located. The lines FP094, FP132, and FP272 were noted for localized variations along chromosome some Pv08, whereas lines FP092, FP109, FP137, and FP162 were noted for localized variations along chromosome Pv04. Lines FP555 and FP197 showed specific variations in Pv05, and lines A2806, A4804, X4504, and FP528 exhibited very specific variations at the end of chromosome Pv11, in the area of the Co-2 anthracnose resistance cluster.

4. Discussion

In this study, the variation within the Fabada dry bean group was investigated for efficient preservation and use of the genetic diversity in this market class. This market class has a well-differentiated seed phenotype and high-quality seeds. A panel with 179 accessions was gathered, and the lines were classified into four groups based on their origin: conserved, cultivated, cultivar, and reference populations. The genotyping of the panel yielded a high number of SNPs (108,585) that homogeneously covered the 11 bean chromosomes (Figure 1S.1). Filtering based on genotyping and phenotyping data was carried out to fine-tune the estimation of diversity in the Fabada market class. Genotypic variation was first detected in redundant lines; 16 groups consisted of more than one line that did not contain differences in the SNP genotypes. These 16 groups included 90 lines (52 from conserved and 38 from cultivated populations), resulting in 41.8% redundancy in the conserved population (41 of 98) and 50.9% in the cultivated population (28 of 55). Seven of those 16 groups included lines of both the conserved and cultivated populations; consequently, most of the genetic diversity in the cultivated population was already present in the SERIDA collection. In contrast, the lines FP555, FP509, FP525, FP500, FP552, and FP542, which have different SNP-related genotypes from the lines of the conserved populations, represent a source of variation not maintained in the SERIDA collection. These findings can be used to optimize preserving the collection by reducing the accessions conserved in lines having identical SNP profiles and incorporating those lines that were different from the cultivated population. Likewise, the variation observed within this market class agreed with the hypothesis that it is a landrace (Zeven 1998a) because if it was produced by a breeding program, a lower variation would be expected. The clustering also showed six lines that were very closely related (different for only two to four SNPs) to the old cultivar A25 (F559, FP502, and FP540) and the modern cultivars A2806 (FP528, FP501, and FP524), indicating that they probably are derived from these cultivars because it is very common for local farmers to use harvested seed for planting. The filtering carried out with genotyping data in the conserved and cultivated populations reduced the FabaPanel to 107 lines, removing redundant lines and maintaining a line per population (SNP profile). The PCoA plot and the generated cluster indicated the wide diversity of lines classified as Fabada market class. However, the phenotypic characterization of seeds indicated that 19 lines had seeds with characteristics that

differed from those of the Fabada market class (homonymy). These lines were located far from the A25 line in the generated dendrogram. Seed size is a quantitative inheritance trait with moderate heritability, and there was an environmental effect on the expression of the characteristic (Murube *et al*. 2020). In addition, there is an overlap in seed size between the large Canellini or White Kidney (seed length 17–20 mm) and Fabada (seed length 20–24 mm) market classes that can induce errors in the classifications. In fact, a 'Cannellini' line in the reference population shows an SNP profile that is identical to that of line FP124 and is similar to that of FP169. The Fabada cultivars 'Xana' and X4562, both with determinate growth habits derived from V203, are very close to Canellini accessions maintained in the SERIDA collection (Ferreira *et al*. 2017).

To study the putative genetic erosion into the Fabada market class in the past 30 years, those redundant lines and the misclassified lines were removed from FabaPanel. The diversity valued as the percentage of segregating sites indicated an important genetic erosion, a loss of genetic diversity between the cultivated and the conserved populations (19.62%–54.26%; Table 1.1). The changes are also reflected in the loss of SNP profiles and the Shannon diversity index; for example, the conserved population had 40 different profiles of SNPs (40 of 84, 47.6%), while the cultivated population had 26 profiles (30.9%), many of them common to the conserved population (6). Crop genetic erosion has been attributed to anthropogenic and environmental factors (van de Wouw *et al*. 2010; Khoury *et al*. 2022), such as the replacement of landraces by modern cultivars. Two Fabada commercial cultivars distributed more than 10 years ago were included in this study: A25 and A2806. However, the genetic erosion cannot be attributed to the diffusion of new cultivars because only six of the 57 lines of the cultivated population can be considered to be derived from commercial cultivars (10.2%). Rather, this genetic erosion may be due to the reduction in the number of farmers in the last 30 years, the selection of sowing seeds, and the increase of monocropping against the traditional maize-bean intercropping. Local farmers usually use their save seed from the previous season for replanting. This custom explains the limited dissemination of modern varieties and is a driver of selection of seed phenotypes.

The genomic exploration of segregation sites along the 11 bean chromosomes showed chromosomal regions with high and no variation (no SNPs; Figures 1S.4 and 1S.5). The

six cultivars included in this study exhibited higher diversity levels than the members of the cultivated population owing to the introgression of new genomic regions in chromosomes Pv01, Pv02, Pv04, Pv06, and Pv11. The lines 'Xana', X4562, A2806, and A4804 were originally derived from line A25 (Ferreira *et al*. 2017), and they carry new genes on chromosomes Pv01 (gene *fin*), Pv02 (gene I), Pv04 (gene *Pm1*), and Pv11 (gene *Co-2*). Major QTLs associated with seed weight have been reported in the 11 bean chromosomes (Arriagada *et al*. 2022). However, most of the Pv01, Pv09, and Pv10 regions did not show variation within the Fabada marker class (Figure 1S.5), whereas high concentrations of SNPs were observed in regions of the chromosomes Pv04, Pv05, Pv06, Pv07, Pv08, and Pv11. Many QTL for seed weight were located on chromosomes Pv01, Pv09, and Pv10 (Blair and Izquierdo 2012; González *et al*. 2016; Sandhu *et al*. 2018). These findings suggest that chromosomal regions without SNPs have relevant roles in controlling the Fabada seed phenotype. In contrast, SNP-rich regions do not have relevant roles in the genetic control of specific seed traits for this market class. In fact, in the chromosomes Pv04, Pv05, and Pv11, no major QTL for seed weight have been mapped in biparental populations having a Fabada parent (Murube *et al*. 2020). Nevertheless, the role of this type of variation should be confirmed in future studies using HTP of quantitative seed traits.

5. Conclusion

The genotyping of materials classified as Fabada showed a wide genotypic variation in this market class, suggesting that it is a landrace group rather than varieties derived from breeding programs. Genotyping also revealed redundant lines and homonymies among the accessions preserved in the SERIDA collection. Genetic erosion was detected when comparing diversity levels between the accessions preserved in the SERIDA collection and those collected during 2021. Most of the genetic diversity was maintained in the SERIDA collection, although some genotypes were not present. The observed erosion cannot be attributed to the diffusion of modern cultivars. In fact, these modern lines have higher genetic diversity levels than the currently cultivated lines. Therefore, their use could increase the variability within this market class.

CHAPTER 2

Differentially expressed genes against *Colletotrichum lindemuthianum* in a bean genotype carrying the *Co-2* gene revealed by RNA-sequencing analysis

A version of this Chapter has been published as:

Jurado M, Campa A, Ferreira JJ (2022) Differentially expressed genes against *Colletotrichum lindemuthianum* in a bean genotype carrying the *Co-2* gene revealed by RNA-sequencing analysis. *Frontiers in Plant Science*. 13:981517. https://doi.org/10.3389/fpls.2022.981517

Differentially expressed genes against *Colletotrichum lindemuthianum* in a bean genotype carrying the *Co-2* gene revealed by RNA-sequencing analysis

Anthracnose is responsible for large yield losses in common bean crops. RNA-sequencing was used to investigate the differentially expressed genes (DEGs) in response to race 38 of *Colletotrichum lindemuthianum* in two NIL (A25 and A4804) that differ in the presence of a resistance gene located in the cluster Co-2. Their responses were analyzed at different hours after inoculation (0, 24, and 48) and within and between genotypes. In all, 2,850 DEGs were detected, with 2,373 assigned to at least one functional GO term. Enriched GO terms in the resistant genotype were mainly related to functions as a response to stimulus, hormone signaling, cellular component organization, phosphorylation activities, and transcriptional regulation. The region containing the Co-2 cluster was delimited at the end of chromosome Pv11 (46.65–48.65 Mb) through a comparison with the SNP genotypes, obtained using GBS among seven resistant lines harboring the *Co-2* gene and the susceptible line A25. The delimited region contained 23 DEGs, including 8 typical R genes, that showed higher expression levels in the resistant genotype and non-changes in the susceptible genotype after inoculation. Six R genes encoding protein kinases and a Leucine-rich repeat (LRR) domain formed a cluster in a core region between 46.98 and 47.04 Mb. The alignment of the raw transcriptome reads in the core region revealed structural changes that were used to design four potential breeder-friendly DNA markers, and it revealed some alignments with the intergenic regions, suggesting the presence of genes in addition to those annotated in the reference genome.

1. Introduction

The common bean (*Phaseolus vulgaris* L.) is an edible legume crop worldwide. Bean crops can be affected by many diseases (Schwartz *et al*. 2005), such as anthracnose, caused by the fungus *Colletotrichum lindemuthianum* (Sacc. and Magnus) Lamb-Scrib. In the presence of the fungus and favorable conditions, the yield losses may be significant, reaching 100%. Typical symptoms are deep and well-delimited lesions on hypocotyls, stems, leaf veins, pods, and seeds that usually have salmon-colored spores. Disease progression is favored by humid environments with moderate temperatures, and the disease can lead to plant death. Conidia germinate on the host surface and form a specialized structure, an appressorium, to penetrate the host. Upon entering, the hyphal thread enlarges and penetrates the cells (O'Connell *et al*. 1985). This process continues for several hours without killing the cells (biotrophic phase). Then, the fungus switches to the necrotrophic phase by producing secondary hyphae, resulting in cell death, which gives rise to cavities that contain acervuli with conidial masses. The conidia can be easily dispersed by splashing raindrops, and the cycle can repeat several times in a growing season. In addition, the conidia can survive in soil, seeds, or plant debris for several years (Tu 1988, 1992); consequently, an efficient method of managing bean anthracnosis is the use of resistant bean genotypes.

C. lindemuthianum exhibits a high level of pathogenic variability. At least 182 races have been reported worldwide from about 1,590 isolates using a standardized set of 12 differential common bean cultivars (Padder *et al*. 2017). Resistance to anthracnose in common bean essentially follows the gene-for-gene model (Flor 1971), in which a specific resistance gene protects against specific isolates or races of the pathogen. On the basis of allelism tests and linkage analyses, many anthracnose resistance genes (named Co-) have been reported (Ferreira *et al*. 2013; Vaz Bisneta and Gonçalves-Vidigal 2020). These anthracnose resistance genes communally show complete dominance, although a few genes with a complementary mode of action have also been identified (Campa *et al*. 2014, 2017). Anthracnose-resistant loci have been located in specific genetic regions on the bean chromosomes Pv01, Pv02, Pv03, Pv04, Pv07, Pv08, and Pv11. Moreover, genetic mapping of genes conferring resistance to specific isolates revealed that the *Co-* genes were organized in clusters with very close race-specific resistance genes (e.g.,

cluster Co-3 on Pv04, cluster Co-5 on Pv07, and cluster Co-2 on Pv11) (Ferreira *et al*. 2013; Campa *et al*. 2014, 2017).

Plants can detect and trigger resistance reactions through the identification of conserved microbial elicitors using pattern recognition receptors, which gives rise to patterned-triggered immunity (PTI). Plants also have intracellular receptors that identify specific pathogen-virulence molecules and result in effector-triggered immunity (ETI; (Dodds and Rathjen 2010; Meng and Zhang 2013; Saijo and Loo 2020). Plant disease resistance genes (R) can detect a pathogen attack and facilitate a counter-attack against them. These genes encode for one or several typical protein domains, such as LRR, nucleotide-binding sites (NB), Toll/Interleukin-1-receptors, coiled-coil (CC), transmembrane domain, protein degradation domain, and protein kinase (Afzal *et al*. 2008; Gururani *et al*. 2012). In common bean, the positions of the reported Co-clusters co-locate with clusters of R genes encoding proteins with kinase or NB-LRR domains (Meziadi *et al*. 2016). For example, an important cluster of these R genes is located at the end of chromosome Pv11 (Meziadi *et al*. 2016; Vaz Bisneta and Gonçalves-Vidigal 2020). The gene *Co-2*, previously named *Are* and originally reported in the dry bean genotype Cornell 49-242 (Mastenbroek 1960; Adam-Blondon *et al*. 1994), has been mapped to this position. Then, using the RIL population 'Xana'/Cornell 49,242, a cluster of specific resistance genes to *C. lindemuthianum* races was mapped to the genetic position of gene *Co-2* (Campa *et al*. 2014). The anthracnose resistance located in this Co-2 cluster has been widely used in common bean breeding. It was introgressed in the navy bean cultivar Sanilac from Cornell 49-242 (Aylesworth *et al*. 1983), and in the fabada market class using the resistance sources SanilacBc6Are and A252 (Ferreira *et al*. 2012). The physical positions of the introgressed genomic region carrying the Co-2 cluster derived from SanilacBc6Are were delimited in the chromosome interval 46.72–48.65 Mb from the genotyping of a set of near-isogenic lines (Ferreira *et al*. 2017). This region contained 162 annotated genes, of which 70 encoded proteins containing NB-LRR, kinase, or Toll/Interleukin-1-receptors-nucleotide-binding site domains (Ferreira *et al*. 2017). Thus, the identification of candidate gene(s) involved in the resistance response required further analysis.

To determine the gene controlling specific traits in the genome (candidate gene) based on forward genetic analysis requires the study of large segregating populations, as well as large amounts of genotyping and phenotyping. A comparative transcriptomic analysis of the pathogen-host interactions in resistant and susceptible bean genotypes can provide data on the gene networks involved in the responses, including those mapped on the regions delimited by the genetic analysis. RNA-seq allows for the investigation of changes in complete transcript sets and their quantification for a specific developmental stage or physiological conditions (Wang *et al*. 2009). RNA-seq analysis identified 3,250 DEGs in response to anthracnose race 73 in the isogenic line T-9576 [derived from the cross Jaguar (Co-1) × Puebla152 (co-1)] through the comparison of susceptible and resistant genotypes (Padder *et al*. 2016). The DEGs included typical R genes and numerous transcription factors (TFs), some of them in or near the region containing the Co-1 locus. A detailed analysis of this region showed a small cluster of four genes encoding CRINKLY4 kinase in the bean genotypes BAT93 and G19833 (*Phvul.001G243500/KTR1*, *Phvul.001G243600/KTR2*, *Phvul.001G243700/KTR3*, *Phvul.001G243800/KFL*), but an additional gene encoding a truncated and chimeric *CRINKLY4* kinase (KTR2/3) was located within this CRINKLY4 kinase cluster in the resistant genotype JaloEEP558 (Cox;(Richard *et al*. 2021)). Expression analysis revealed that KTR2/3 is 3-fold up-regulated in JaloEEP558 (Cox) after *C. lindemuthianum* infection compared with the mock control at 24 h post-inoculation, whereas the expression levels of KTR2, KTR3, and KFL were not modified after infection. Interestingly, the candidate genes *Phvul.001G243500* and *Phvul.001G243700* were also differentially expressed in response to race 73 in NILs T-9576 (Padder *et al*. 2016).

In this study, RNA-seq was used to investigate DEGs in response to race 38 of *C. lindemuthianum* with a particular focus on the delimited genomic regions in which the Co-2 cluster is located. The analyses provide data for the gene networks involved in the response to *C. lindemuthianum*, an approach to identifying candidate gene(s) against race 38 in the Co-2 cluster, and the development of markers to accelerate breeding programs.

2. Material and methods

2.1. Plant material

The lines A25 and A4804 were used for the transcriptomic analysis. Line A25 is a selection of the market class fabada (white and very large seeds) that is susceptible to *C. lindemuthianum* race 38 (isolate Cl18). The NIL A4804 is a resistant genotype to *C. lindemuthianum* race 38 obtained from the cross A2806 × X4562. The NILs A2806 and X4562 are derived from A25 (Figure 2S.1), both having the seed phenotype of the market class fabada (Ferreira *et al*. 2012) and resistance to *C. lindemuthianum* race 38 controlled by a gene located in the Co-2 cluster (Ferreira *et al*. 2017). The resistance gene *Co-2* is derived from Sanilac Bc6Are (Figure 2S.1), which was obtained from Cornell 49-242. Finally, sources of the *Co-2* genes, Cornell 49-242 and Sanilac Bc6Are, were also included in this work.

2.2. Inoculation with *Colletotrichum lindemuthianum* race 38

The Cl18 monosporic isolate of *C. lindemuthianum*, classified as race 38 (Ferreira *et al*. 2008), was used in this work. To obtain abundant sporulation, the isolates were grown at 22°C in darkness for 10 days in potato-dextrose agar (Becton Dickinson and Company). Spore suspensions were prepared by flooding the plates with 5 ml of 0.01% Tween 20 in sterile distilled water and scraping the surface of the culture with a spatula. Inoculations were performed by spraying 10-day-old seedlings with a spore suspension containing 2 × 106 spores mL−1. Before sowing, seeds were disinfected in four steps: rinsed in distilled water to remove dirt particles, 30 s in 95% EtOH, 30 s in 15% hydrogen peroxide, and rinsed thoroughly in distilled water. The seedlings were maintained in a climate chamber at 23 to 24°C with 90 to 95% humidity and a 12-h photoperiod.

The experimental design had three replicates (corresponding to three resistance tests), two genotypes (susceptible genotype A25 and resistant genotype A4804), and three treatment assessment times: just before inoculation (named as 0), 24, and 48 h post-inoculation (hpi). On agar media at 24°C, the conidia germinated 4–6 h after sowing, and soon after, the appressoria was observed. At 48 h after sowing, there was extensive hyphal growth on the Petri plates, and a week later, sporulation was observed. Thus, the fungal attack started before 24 hpi, and the plant cells could detect the pathogen and start the

cascade of reactions. In susceptible genotypes, after less than 24 hpi, the cytoplasm of infected cells gradually degenerates (O'Connell *et al*. 1985). Two seedlings per genotype were included in each replicate and treatment. The leaf tissues were harvested, flash-frozen in liquid nitrogen, and stored at −80°C before RNA extraction. In all, the study included 18 samples, named S (susceptible) or R (resistant), follow by the time when the leaf was collected (0, 24, or 48) and then the replicate number (1, 2 or 3). For example, S0.1 represents the susceptible genotype at 0 h/control from experiment 1.

2.3. Total RNA isolation, cDNA library construction, and sequencing

Total RNA was isolated from samples using DNeasy Plant Mini Kit following the manufacturer's instructions (Qiagen, Germany). RNA was quantified by fluorometric methods and quantity was investigated by using 2,100 Bioanalyzer Instrument (Agilent Technologies, United Kingdom). RNA libraries were prepared using the TruSeq Stranded mRNA Sample Preparation Kit (Illumina) and sequencing was carried out on the Illumina platform. The reads were mapped to the reference genome with HISAT2 splice-aware aligner (Kim *et al*. 2015) using the bean genome G19833 v1.0 (Schmutz *et al*. 2014) (https://www.ncbi.nlm.nih.gov/genome/380). Expression profiles are represented as read count and normalization values which were calculated based on transcript length and depth of coverage. The counts for mapped reads were normalized by calculating the FPKM (Fragments Per Kilobase of transcript per Million mapped reads). This analysis was performed in Macrogen Inc. (Seoul, Republic of Korea).

2.4. Differentially expressed genes

A principal component analysis (PCA) and hierarchical clustering analysis (HCA) were performed to detect the possible sources of noise in the results. The DEGs were identified through comparisons within resistant (A4804) and susceptible genotypes (A25) at 0 and 24 hpi and at 0 and 48 hpi (comparisons named as R24-R0, R48-R0, S24-S0, and S48-S0). In addition, the DEGs were investigated through comparisons between the two genotypes at 0, 24, and 48 hpi (named R0-S0, R24-S24, and R48-S48). The NOISeq package (2.38.0;(Tarazona *et al*. 2011)) and pheatmap 1.0.12 package in R project (R Core Team 2021) were used to explore the quality of the samples and detect DEGs. The DEGs were identified using the criterion $q > 0.80$. Specific and common DEGs between

genotypes and treatments were visualized using Venn diagrams constructed using the package ggVenn/ ggplot2 in R project.

2.5. Gene ontology (GO) analysis of DEGs

To investigate functional groups of DEGs in response to fungus, a GO analysis was performed using the Ensembl database (organism dataset: pvulgaris_eg_gene, version 2022-02-10) considering the three categories: biological process (BP), molecular function (MF), and cellular components (CC). The GO enrichment of significantly over-represented terms was carried out using the R package ViSEAGO (Brionne *et al*. 2019) and the DEG list. Enrichment tests were assessed using Fisher's exact test ($p \leq 0.01$) for the resistant and susceptible genotypes at both time level datasets [24 hpi (S0-S24, R0-R24) and 48 hpi (S0-S48 R0-R48)] and the three GO categories. The enriched GO terms were grouped based on Wang's semantic similarity into functional clusters using hierarchical clustering between GO terms with GO graph topology and Ward's criterion.

2.6. Genotyping by sequencing

The bean lines Cornell 49-242, Sanilac Bc6Are, A25, A2806, X4562, and A4804 were genotyped using the GBS method (Elshire *et al*. 2011) optimized in accordance with Schröder *et al*. (2016). DNAs were isolated from young leaves following the SILEX method (Vilanova *et al*. 2020), and DNA quality was checked in agarose gels. Genomic DNAs from the lines were digested individually with the Taqα1 and MseI restriction enzymes. Libraries were built based on the Poland *et al*. (2012) protocol with a different adaptor for ligation. Then, individual samples were checked by PCR, and the resulting products were visualized on 2% agarose gels. In total, 20 barcoded samples were pooled for PCR amplification. Sequencing was performed in the Illumina platform by Macrogen Inc. (Seoul, Republic of Korea). SNP calling was carried out by AllGenetics&Biology SL (www.allgenetics.eu) using the *P. vulgaris* reference genome (genotypes G19833 v1.0, https://www.ncbi.nlm.nih.gov/genome/380,(Schmutz *et al*. 2014)). The SNPs supplied by GBS were filtered and extracted using TASSEL 5.1 software (Bradbury *et al*. 2007). SNPs with the following characteristics were considered in the analysis: (i) missing data, less than 50%; (ii) minor allele frequency, greater than 5%; (iii) mapped to one of the 11 pseudo-chromosomes. The genomic regions were delimited based on the physical

position of the SNPs flanking the introgressed region (coordinates of the first and last introgressed SNPs). Finally, the GBS results reported by Murube *et al*. (2017) for the NILs A1258, A2806, and X2776 were used.

2.7. RNA-Seq reads assembly and visualization

To identify variations useful to develop specific markers for the Co-2 region, raw RNA-seq reads of each experiment were assembled with the chromosome Pv11 of *P. vulgaris* v1 using the RBowtie package (Langmead *et al*. 2009). The alignments were visualized in Integrative Genomics Viewer (IGV) software (Robinson *et al*. 2011), and the selected region was explored to detect polymorphism, InDels, between the reference genome and the studied genotypes. The observed polymorphisms were also assessed by BLASTN and ClustalW alignments with the five bean genomes available in Phytozome (*Phaseolus vulgaris* G19833 v2.1; *Phaseolus vulgaris* 5 -593 v1.1; *Phaseolus vulgaris* UI111 v1.1; *Phaseolus vulgaris* Labor Ovalle v1.1). Those regions that showed an InDel in one genome were considered.

The primers for the PCR amplification of regions containing InDels were designed using the PrimerBlast tool (Ye *et al*. 2012). PCR reactions (20 µl) contained 2 µl of PCR buffer (10×), 1.2 µl MgCl2 (25 mm), 2 µl dNTP mixture, 0.8 µl of each primer (10 µm), 8.05 µl distilled water, 0.15 µl of TaKaRa LA Taq DNA Polymerase (5 U/µl; TaKaRa), and 50 µl 10 ng/µl DNA. The PCR reaction protocol performed on a Verity Thermal Cycler (Life Technologies Carlsbard, CA, United States) was as follows: an initial 5 min at 95°C; 30 cycles of 30 s at 95°C, 45 se at 62°C, 60 s at 72°C; final extension step at 72°C for 10 min. PCR products were separated on a 2% agarose gel, stained with RedSafe™ Nucleic Acid Staining Solution (iNtRON, Seoul, Republic of Korea), and visualized under ultraviolet light.

3. Results

3.1. Transcriptome sequencing of resistant and susceptible genotypes

At 7 days post-inoculation with *C. lindemuthianum* race 38, the susceptible genotype (A25) was dead, whereas symptoms were not observed on the resistant genotype A4804 (Figure 2.1). A total of 18 cDNA libraries were generated, one per genotype, replicate,

and treatment, and, in total, they generated 876,756,658 clean reads using the NovaSeq platform (Table 2S.1). The reads of all the samples were used for transcriptome assembly, and an average of 91.1% of reads were mapped to the reference genome. Mapped reads were normalized by calculating the FPKM. A PCA of FPKMs estimated for each genotype, replicate, and treatment revealed two components that explained 79% of the variation. The biplot shows 3 samples, S0.1, S24.1, and R24.1, separated from the remaining 15 samples (Figure 2S.2a). Similarly, an HCA classified the samples into two different main clusters separated from the samples S0.1, S24.1, and R24.1 (Figure 2S.2b). These samples had low RNA Integrity Numbers (RIN), less than 5.7, in the RNA extraction and were discarded from the analysis.

Figure 2. 1. Reactions of the genotype A4804 (left: resistant) and A25 (right: susceptible) against monosporic isolates Cl18 (race 38) 7 days after inoculation.

3.2. Differentially expressed genes

The DEGs were identified from seven comparisons (R24-R0, R48-R0, S24-S0, S48-S0, S0-R0, S24-R24, and S48-R48). A total of 5,740 differential expressions, involving 2,850 unique genes, were identified (Table 2S.2). A higher number of DEGs were found in the susceptible genotype than in the resistant genotype (Figure 2.2A) when the expression levels at 24 and 48 hpi were compared with the control (0 hpi). At 24 hpi after the susceptible genotype A25 was inoculated with *C. lindemuthianum*, 2,455 DEGs (686

upregulated; 1,769 downregulated) were detected, and the number decreased to 1,518 (469 upregulated; 1,049 downregulated) at 48 hpi (Figure 2.2a). Among the DEGs in the susceptible genotype, 1,615 genes only appeared in this genotype (Figure 2S.2b). The resistant genotype A4804 showed 831 DEGs (272 upregulated; 559 downregulated) at 24 hpi and 719 DEGs (306 upregulated; 413 downregulated) at 48 hpi. In addition, 127 DEGs were only detected in the resistant genotype, and 16 of them were identified at both 24 and 48 hpi (Figure 2.2B).

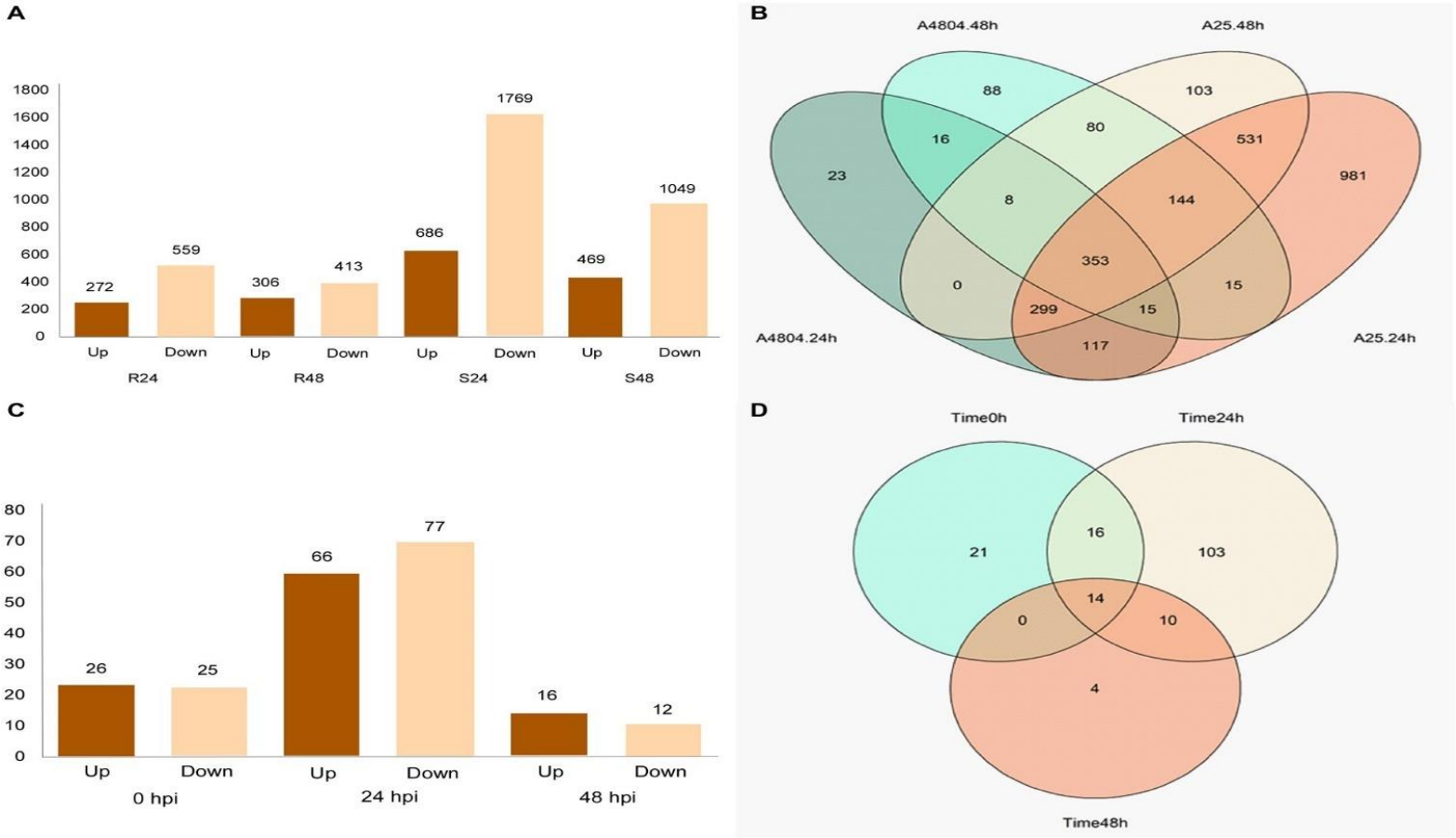

Figure 2. 2. Visualization of DEGs detected from comparison at different hpi and between genotypes in the same hours post-inoculation (hpi). (A) Histograms showing the number of DEGs, upregulated and downregulated, at different hpi in the two genotypes (B) Venn diagrams showing the numbers of specific and common DEGs at different hpi in the two genotypes. (C) Histograms showing the number of DEGs, upregulated and downregulated, at the same hpi. (D) Venn diagrams showing the numbers of specific and common DEGs at different hpi..

For the comparisons between resistant and susceptible genotypes at the same hpi, the highest number of DEGs, 143 (66 upregulated; 77 downregulated), were detected at 24 hpi. The number of DEGs decreased to 28 (16 upregulated; 12 downregulated) at 48 hpi (Figure 2.2C). In total, 21 genes were differentially expressed between both genotypes before inoculation and were discarded from the analysis. There were 14 DEGs between the resistant and susceptible genotypes that were common for the three hpi (Figure 2.2D). In contrast, there were 168 DEGs in both genotypes after inoculation, 10 of which maintained the differential expression at 24 and 48 hpi (Figure 2D): *PHAVU_003G011800g, PHAVU_004G005400g, PHAVU_004G046400g, PHAVU_004G094000g, PHAVU_ 007G216700g, PHAVU_008G103500g, PHAVU_010G012900g, PHAVU_011G044000g, PHAVU_011G201700g,* and *PHAVU_ 011G203000g*. Finally, four DEGs were only detected between both genotypes at 48 hpi: *PHAVU_001G083000g, PHAVU_ 003G011000g, PHAVU_003G0939001g*, and *PHAVU_007G276500g*.

3.3. Functional classification of DEGs

A GO analysis was performed using the DEGs in each genotype. Among the 2,850 DEGs, 2,373 were assigned to at least one functional GO term in the Ensembl database (analysed 02/23/2022, www.ensembl.org). To reveal the functional processes involved in the resistant and susceptible genotypes, the DEGs at 24 and 48 hpi were analyzed for enriched terms in the three GO categories (Figures 2S.3a–f, 2S.4). The differences were more evident at 48 than at 24 hpi, particularly in the BP category. The BP category contained 41 enriched GO terms at 24 hpi (Table 2S.4), 25 in the resistant genotype and 39 in the susceptible, with ‘translation’ and ‘biosynthetic process’ being the most significant pathways at both times. At 48 hpi, 111 GO terms were enriched (Table 2S.3), 89 in the resistant genotype, with ‘response to biotic stimulus,’ ‘cell wall organization,’ ‘hormone signalling’, ‘flavonoid metabolism’, ‘dephosphorylation’, and ‘phosphatase activity’ being highly featured, and 31 in the susceptible genotype, with ‘translation’ and ‘biosynthetic process’ being the more prevalent terms. For MF at 24 hpi (Table 2S.4), only nine GO terms were enriched, four in the resistant and eight in the susceptible genotypes. In both genotypes, the term ‘structural molecule activity’ was the most enriched. At 48 hpi, 34 GO terms were significant (Table 2S.3) in MF, 31 in the resistant

and 14 in the susceptible genotypes. Structural 'molecule activity' was the most enriched process in the susceptible genotype, whereas 'hydrolase' and 'signaling receptor activity' were the most enriched processes in the resistant genotype. For the CC category, 19 GO terms were enriched at 24 hpi (Table 2S.4), 13 of them in the resistant genotype and 14 in the susceptible genotype. The GO term 'organelle ribosome' was the most enriched in both cases, with 'non-membrane-bounded organelle' also being highly enriched in the resistant genotype. At 48 hpi 26 GO terms were enriched (Table 2S.3), 14 in the resistant genotype, with 'extracellular' and 'cell wall regions' being the most enriched term, and 15 in the susceptible, with 'ribosome' being the most enriched term.

3.4. The physical position of the Co-2 cluster the line A4804

Sequencing the GBS libraries yielded approximately 17.7 million reads in the six genotypes (Cornell 49-242, SanilacBc6Are, A25, A2806, X4562, and A2804), resulting in a total of 108,593 SNPs and 35,244 SNPs after filtering. The GBS revealed that 506 SNPs mapped on the end chromosome Pv11 (physical position>45 Mb). Genotypic comparisons of these 506 SNPs among the three resistant NILs harboring the *Co-2* gene (A2806, X4562, and A2804), the resistance sources Cornell 49-242 and Sanilac Bc6Are, and the susceptible line A25 revealed that the three resistant NILs exhibited an introgressed region at the end of chromosome Pv11. These regions are tagged by SNPs having the donor resistance genotype (Sanilac Bc6Are). On the basis of the physical position of these SNPs, line A2806 has a region located at 45.10–48.78 Mb tagged by 183 SNPs of the SanilacBc6Are genotype (except for 21 isolated SNPs). The lines X4562 and A4804 maintained an introgressed region at 46.65–50.32 Mb that was tagged by 132 SNPs of the SanilacBc6Are genotype (Figure 2S.4). There was a common introgressed region among the three NILs located between 46.65 and 48.65 Mb.

3.5. DEGs in the genomic Co-2 region

The introgressed region in line A4804 carrying the *Co*-2 gene (46.65–48.65) has 165 annotated genes. Among them, 23 were differentially expressed in response to *C. lindemuthianum*. The representation of the differential expressions of the 23 genes in a heatmap revealed a separation between the resistant and susceptible genotypes (Figure 2.3) and the following four main groups of genes:

- **Group I**, includes seven annotated genes that showed higher expression levels in the susceptible genotype and no changes in the resistant genotype, in response to *C. lindemuthianum*: **PHAVU_011G196600g, PHAVU_011G201800g, PHAVU_011G201700g, PHAVU_011G202000g, *PHAVU_011G193100g, PHAVU_011G203000g*, and *PHAVU_011G200600g*. The genes marked with an asterisk encode hypothetical proteins with LRR domains.
- **Group II**, contains five annotated genes that decreased in expression in response to *C. lindemuthianum* in both susceptible and resistant genotypes: *PHAVU_011G200200g, PHAVU_ 011G192700g, PHAVU_011G200400g, PHAVU_011G200000g*, and *PHAVU_011G189900g*.
- **Group III**, includes eight genes with higher expression in the resistant genotype and non-changes in the susceptible genotype in response to *C. lindemuthianum*: *PHAVU_011G193400g, PHAVU_011G193300g, PHAVU_011G194100g, PHAVU_ 011G193900g, PHAVU_011G193700g, *PHAVU_011G193800g, PHAVU_011G202200g*, and **PHAVU_011G202100g*. These genes have a serine/threonine-protein kinase function (6) or encode proteins with LRR domains (marked with an asterisk). They are located near two positions: a region with six genes (46.98–47.04 Mb, 'core region'; Figure 2.4A) and another region with two genes (48.02 Mb).
- **Group IV**, includes three annotated genes that tend to increase their expression in response to *C. lindemuthianum*, particularly in the susceptible genotype: *PHAVU_011G206500g, PHAVU_011G2015000g*, and *PHAVU_011G2015001g*. These genes have unknown functions in the bean reference genome.

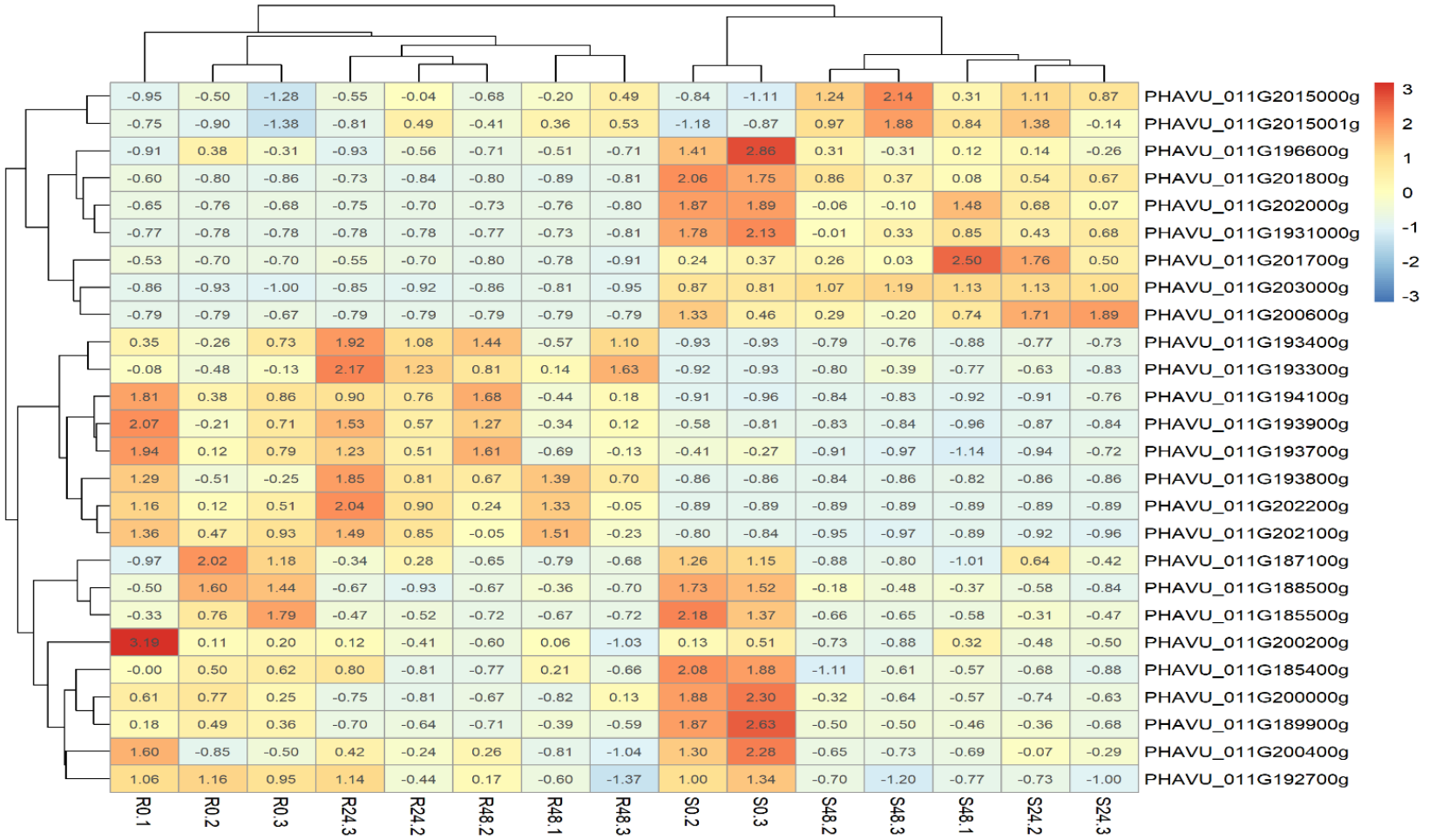

Figure 2. 3. Heatmap built with the package pheatmap showing the expression (FPKM) for the resistant and susceptible lines (A4804 & A25) per hour post-inoculation and sample in the 23 DGEs located in the delimited region introgressed with the resistance to C lindemuthinanum race 38 in the genotype A4804..

3.6. Specific markers for the Co-2 cluster

The assembly of the raw transcriptome reads in the 'core region' in the chromosome Pv11 revealed InDels when read sequences of the resistant genotype (A4804) and the reference genome were aligned. The occurrence of InDel was also checked by BLASTN with the four common bean genomes available through Phytozome database and five InDels were found in at least one of the genomes available. The susceptible genotype A25 did not show changes with the reference genome (G19833) in these four regions. The four polymorphic positions with InDel were:

- **M1 position**, located in the third exon of the gene *PHAVU_011G193000g* (see Figure 2.4A). The resistant genotype A4804 has an insertion of 18bp, also observed in the genomes 5–593 v1.1, UI111 v1.1, and, Labor Ovalle v1.1 (Figure 2S.5a).
- **M2 position**, located in the intergenic region between the DEG *PHAVU_011G193400g* (Figure 2.4A) and *PHAVU_011G193500g*. The genotypes A4804, 5–593 v1.1, and UI111 v1.1 have a deletion of 10 bp. The reads with these sequences aligned with the gene *Pv5-593.11G192900* annotated in the bean genome 5–593 (Figure 2S.5b).
- **M3 position**, located in the gene *PHAVU_011G193500g*. The genotypes A4804 and 5–593 show a deletion of 6 bp and closed a mutation of two bp (Figure 2S.5c).
- **M4 position**, located in the intergenic region between the genes *PHAVU_011G193600g* and *PHAVU_011G193700g* in the reference genome, but is annotated in the version 2 of G19833 in the gene *Phvul.011G193600.1*. The genotypes A4804, 5–593, and UI111 have a 12 bp deletion (Figure 2S.5d).

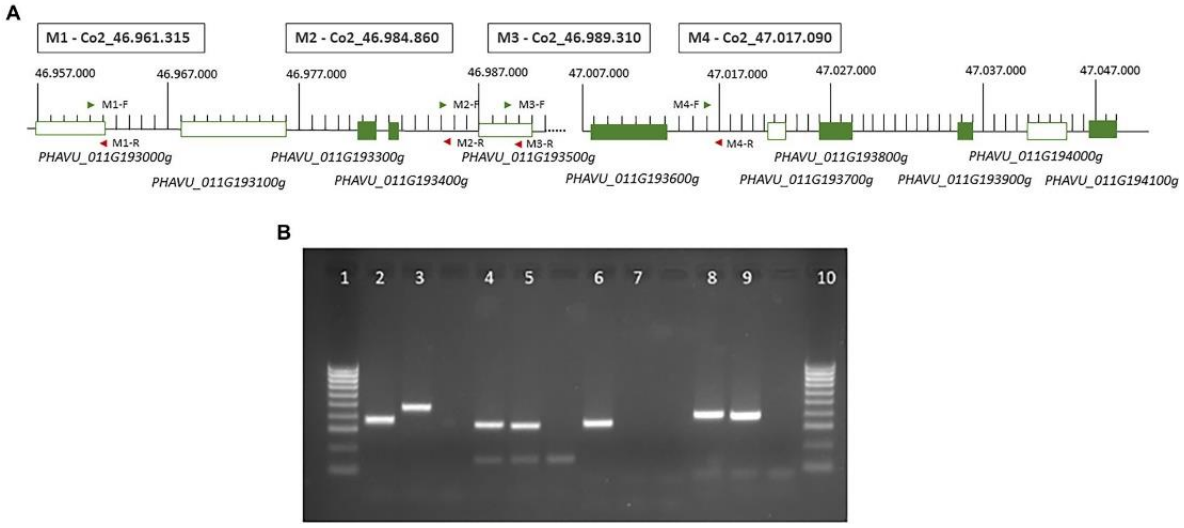

Figure 2. 4. A) Representation of the 'core region' with the 6 annotated genes in the reference genome and the position of the four markers developed in this study. Green boxes represent candidate genes revealed in transcriptomic analysis (Figure 3). B) Agarose gel (2%) showing the results of the PCR amplification for the four markers developed: lines 1 and 10, marker 100 bp; 2 and 3, results in n A25 and A4804 for the marker M1_ Co2_46.961.315; 4 and 5, results in A25 and A4804 for the marker M_ Co2_46.984.860; 6 and 7, results in A25 and A4804 for the marker M3_ Co2_46.989.310; 8 and 9, results in A25 and A4804 for the marker M4_ Co2_47.017.090.

4. Discussion

Anthracnose is an important disease of common bean that causes significant losses worldwide (Mohammed 2013). Resistance to anthracnose in common bean has been extensively studied through genetic analysis in segregating populations and many resistance genes have been described (*Co*-genes; (Ferreira *et al*. 2013)). However, information on the molecular responses to specific plant–fungus interactions is limited. Comparative transcriptome analyses have been used to study responses to disease in plants and to identify the specific genes involved (Kankanala *et al*. 2019), but they have not been extensively applied to investigate resistance to bean anthracnose (Oblessuc *et al*. 2012; Padder *et al*. 2017). In this study, the changes in the transcriptomic profile during the response to *C. lindemuthianum* race 38 were investigated in two NILs, the susceptible line A25 and the resistant line A4804 carrying a resistance gene located in the Co-2 cluster.

A total of 2,850 DEGs involved in the response to *C. lindemuthianum* were detected in this study. More DEGs were found in the susceptible genotype than in the resistant genotype, which corroborated the results of Padder *et al*. (2016) in response to race 73.

However, the resistant genotype showed a higher number of enriched GO terms than the susceptible genotype (Tables 2S.3, 2S.4), indicating the diversity of processes involved in the resistance response. At 48 hpi, the resistant genotype had significantly enriched GO terms for biological functions related to cellular events typically involved in response to stresses, such as cell wall biogenesis (e.g., GO:0044036, GO:0071555, GO:0042546; see Figure 2S.3A) and hormone network regulation (GO:0009737, GO:0009738, GO:0071215), as well as transcriptional, translational, and metabolic reprogramming. Notably, there was a great enrichment of phosphate inhibitor and regulation activities (GO:004864, GO:0019212) in the resistant genotype at 48 hpi, which represents a mechanism of response in plants reported in other species (Rakwal *et al*. 2001; Alvarado-Gutiérrez *et al*. 2008). In addition, DEGs involved in the regulation of protein phosphorylation and dephosphorylation (e.g., GO:0035305, GO:0016311, GO:0045936, GO:0010921) were identified as mitogen-activated protein kinases (GO:0004672). Protein kinases play crucial roles in plant resistance to pathogens because they are involved in signaling downstream of receptors/sensors that transduce extracellular stimuli into intracellular responses in eukaryotes (Meng and Zhang 2013).

Plant hormones, such as salicylic acid, jasmonic acid, abscisic acid (ABA), and ethylene, also play important roles in plant disease resistance (Mauch-Mani and Mauch 2005). For instance, cytokinin and ethylene responses were upregulated, whereas jasmonic acid, gibberellin, and abscisic acid responses were downregulated in response to *C. lindemuthianum* (race 73) in the genotype SEL1308 (Oblessuc *et al*. 2012). The terms GO:0009738 and GO:0009737 (abscisic acid-activated signalling pathway) were significantly enriched in this analysis and involved 14 genes (Table 2S.3A). The resistant genotype was enriched for GO terms involving salicylic acid (e.g., GO:2000031, *PHAVU_005G047200g*) and ABA (e.g., GO:0010427, *PHAVU_003G109000g*). Phytohormone networks are connected through crosstalk involving TFs or sequence-specific DNA-binding factor proteins that control the transcription rates of specific genes (Monson *et al*. 2022). The resistant genotype was enriched for terms, such as GO:0006355 and GO:1903506 (regulation of transcription), which involved a lot of genes. The role of TFs in the response to anthracnose was previously reported by Padder *et al*. (2016) and verified in this analysis. The following DEGs that code for TFs were identified in the resistant genotype (Tables 2S.2, 2S.3A): *PHAVU_002G056300g,*

PHAVU_002G260700g, and *PHAVU_002G265400g* (Table 2S.3; GO:0003700). DEGs *PHAVU_002G056300g* and *PHAVU_002G265400g* were also identified in response to race 73(Padder *et al*. 2016).

The genotyping of the NILs A2806, A4804, and X4562 through GBS allowed the introgressed region with the *Co-2* gene at the end of chromosome Pv11 to be delimited. The SNPs tagging this genomic region share the resistance donor's genotype (Sanilac Bc6 Are). This position co-located with the anthracnose resistance genes, forming the cluster Co-2. The resistance loci to races 6, 38, and 39 were mapped between makers IND11_46.8842 and Pv11_4600a at physical positions 46.8 and 47.07 Mb in the RIL 'Xana'/Cornell (unpublished data). The size and position of this region were similar to those reported by Murube *et al*. (2017) in the NILs A2806 and X2776. The lines X4562 (derived from X2776) and A4804 (from X2776 x A2806) maintained the introgressed region of the line X2776, between 46.65 and 50.10 Mb (see Figure 2S.4). The results showed a common introgressed region of ~2 Mb (46.65–48.65 Mb) among the resistant NILs A2806, A4804, X2776, and X4562. This common region also overlaps with the introgressed region in the NIL A1258 (46.65–48.65 Mb) obtained from a backcrossing program in which the line A252 was the donor parent and the line A25 the recurrent parent. The breeding line A252 has a resistance cluster mapped at the end of chromosome Pv11 that includes a resistance locus to race 38 (Rodríguez-Suárez *et al*. 2007). The NIL A1258 has a large introgressed region in chromosome Pv11 (Murube *et al*. 2017). The resistant NILs A1258, X2776, A2806, X4562, and A4804, all with resistance loci in the Co-2 cluster, have a common region between 46.65 and 47.07 Mb. A differential expression analysis revealed 23 DEGs in this region (Figure 2.3), including a group of 8 genes (cluster III) with higher expression levels in the resistant genotypes and non-changes in the susceptible genotype after *C. lindemuthianum* inoculation. Interestingly, six of these eight genes form a cluster at 46.98–47.04 Mb, at the border of the common regions among the six NILs (Figure 2S.4): *PHAVU_011G193300g, PHAVU_011G193400g, PHAVU_ 011G193700g, PHAVU_011G193800g, PHAVU_011G193900g*, and *PHAVU_011G194100g*. All of these genes are typical R genes with a serine/threonine protein kinase or LRR domain that may be associated with the initiation of plant defense response signals and with pathogen recognition (Dodds and Rathjen 2010; Meng and Zhang 2013; Monson *et al*. 2022). The genomic region

containing the six DEGs (core region) is the main candidate region for the resistance gene(s) to race 38.

RNA-seq technologies are powerful tools for studying gene expression, but they have limitations when using a unique genome as a reference because the annotated genes can vary between databases and genomes. Exploring RNA-seq data can be useful to improve the annotation of genetic variants (Chen *et al*. 2017). Da Silva *et al*. (2013) discovered 1,873 new genes in a local grapevine variety not annotated in the reference genome, and Tisserant *et al*. (2011) mapped 13% of the cDNA reads outside the predicted UTRs and gene models. Within the *P. vulgaris* species differences in the numbers of annotated protein-coding genes (e.g., 27,433 in G19833; 27,065 in 5–593) can be found; therefore, the appearance of RNA-seq reads in the 'core region' aligned to the intergenic regions (e.g., positions M2 and M4) of the reference genome can be treated and explored in later studies as putative novel genes/transcripts. Because genome-guided or *de novo* transcriptome reconstruction is needed to annotate these possible new genes, we only used the polymorphisms found in other *P. vulgaris* genomes to design markers and explore the genetic variation. Position M2 only corresponds with an annotated gene in the bean genome 5–593. This gene encodes a protein kinase (*Pv5-593.11G192900*), and an additional kinase gene in the Co-1 cluster, implicated in the defense against *C. lindemuthianum* race 100 in common bean, has been reported by Richard *et al*. (2021). Reads aligned in the M4 position revealed the problem of using only one database as a reference, because when the G19833 v1.1 genome in NCBI is used as a reference to record gene expression levels, this region is not a gene. However, it is the gene *Phvul.011G193600.1* in the G19833 v2.1 genome in the Phytozome database. As a result, the possible differences in expression are not explored in this gene, which encodes an LRR protein, included in the 'core region'. These alignments did, however, reveal polymorphisms that were used to develop four potential new breeder-friendly DNA markers. The most popular markers linked to the *Co-2* gene are SCH20 and SCAreoli (Geffroy *et al*. 1998), which are both Cleaved Amplified Polymorphic Sequences in which the polymorphism was revealed after a restriction enzyme cut. The polymorphisms in the four developed markers can be visualized in agarose gels, with the M1 marker showing size variations between the resistant and susceptible genotypes used in this work.

5. Conclusion

This study shows that by combining the physical locations and a comparative transcriptome analysis, a closer approximation of the region containing the candidate genes controlling resistance was possible. This approach allowed us to reduce an initial delimitated region of ~2 Mb to a 'core region' of ~60,000 bp. The 'core region' contains nine annotated genes in the G19833 genome, six of which were differentially expressed in response to *C. lindemuthianum*. They encode protein kinase or LRR domains, typical of R genes. However, additional resistance genes can also be present in the 'core region' of the resistant genotype as revealed by the alignments of obtained reads in intergenic regions of the bean reference genome. In addition, that alignments showed a major InDel that was used to design functional markers to help accelerate breeding programs and genetic analyses.

CHAPTER 3

Identification of consistent QTL and candidate genes associated with seed traits in common bean by combining GWAS and RNA-Seq

A version of this Chapter has been published as:

Jurado, M., García-Fernández, C., Campa, A. *et al*. Identification of consistent QTL and candidate genes associated with seed traits in common bean by combining GWAS and RNA-Seq. Theor Appl Genet 137, 143 (2024). https://doi.org/10.1007/s00122-024-04638-5

Identification of consistent QTL and candidate genes associated with seed traits in common bean by combining GWAS and RNA-Seq

Common beans show wide seed variations in shape, size, water uptake, and coat proportion. This study aimed to identify consistent genomic regions and candidate genes involved in the genetic control of seed traits by combining association and differential expression analyses. In total, 298 lines from the Spanish Diversity Panel were genotyped with 4,658 SNP and phenotyped for seven seed traits in three seasons. Thirty-eight significant SNP-trait associations were detected, which were grouped into 23 QTL genomic regions with 1,605 predicted genes. The positions of the five QTL regions associated with seed weight were consistent with previously reported QTL. Hierarchical Clustering on Principal Components (HCPC) analysis using the SNP that tagged these five QTL regions revealed three main clusters with significantly different seed weights. This analysis also separated groups that corresponded well with the two gene pools described: A and MA. Expression analysis was performed on the seeds of the cultivar ‘‘Xana’’ in three seed development stages, and 1,992 DEGs were detected, mainly when comparing the early and late seed development stages (1,934 DEGs). Overall, 91 DEGs related to cell growth, signaling pathways, and transcriptomic factors underlying these 23 QTL were identified. Twenty-two DEGs were located in the five QTL regions associated with seed weight, suggesting that they are the main set of candidate genes controlling this character. The results confirmed that seed weight is the sum of the effects of a complex network of loci, and contributed to the understanding of seed phenotype control.

1. Introduction

Common bean (*Phaseolus vulgaris* L.) is a diploid and self-pollinated species that is considered the most important legume crop for direct human consumption (FAO 2022). Bean crops are present worldwide, and depending on their genotype, they can be consumed as immature pods (snap beans or green beans) or seeds after rehydration (dry beans). Bean seeds are a valuable source of proteins, carbohydrates, dietary fiber, vitamins, minerals, and bioactive molecules as phenolic components (Hayat *et al*. 2014). In addition, bean crops provide benefits to the soil, have low carbon and water footprints, and integrate well into sustainable agricultural models (Uebersax *et al*. 2022).

In the common bean, two main gene pools were found in the analysis of variation in morpho-agronomic traits, seed size, isoenzymes, seed proteins such as phaseolin, and different types of DNA markers in both wild and cultivated populations (Gepts *et al*. 1986; Singh *et al*. 1991; Blair *et al*. 2009): A and MA. Each gene pool was domesticated independently in parallel domestication events. These two gene pools have also been observed in the European germplasm, although cultivars showing different levels of introgression between both gene pools have also been detected (Santalla *et al*. 2002; Campa *et al*. 2018).

Bean seeds exhibit extensive phenotypic variation (e.g., (Campa *et al*. 2018)), which can be described by considering a combination of seed shape, seed size, seed coat color, and color distribution. Seed shape is recorded as seed dimensions (length, width, thickness, and area) and the ratios among them, whereas seed size is usually recorded as 100 seed weight. The seed phenotype is an important trait in domestication and is related to consumer acceptability and its potential use as precooking and canned food. Seed size is a yield-related trait, along with the number of pods per plant and the number of seeds per pod (White and González 1990). Seed phenotype is also an important trait in snap bean varieties, which are preferred by white seed color, elongated seed shape, and smaller seed size (Silbernagel *et al*. 1991). Other relevant characteristics of bean seeds are water absorption and coat proportion because of their relationship with cocking response and consumer acceptability (Berry *et al*. 2020). Water uptake during soaking has been correlated with cooking time, and there is a relationship between seed size and speed of water absorption (Vidak *et al*. 2022). The seed coat plays a significant role in the hard-to-

cook process of bean hardening before and during storage (de León *et al*. 1989). The seed coat represents approximately 10% of the seed weight and shows high mineral content (for example, Fe, Ca, and Mg) and antioxidant capacity, as well as many anti-nutrients that affect mineral bioavailability (Blair *et al*. 2013). The proportion of seed coat is positively correlated with seed hardness; seeds with a higher percentage of coat tend to have hard shells (Escribano *et al*. 1997).

Some studies have addressed the inheritance of the traits involved in seed phenotypes. Seed size, shape, water uptake, and coat proportion exhibit quantitative inheritance with moderate-to-high heritability (Moghaddam *et al*. 2016; Berry *et al*. 2020). Many QTL involved in the genetic control of these seed traits have been described (González *et al*. 2016; Murube *et al*. 2020), although some QTL for seed size and shape have been collocated in different backgrounds and studies (Murube *et al*. 2020). These studies were conducted in different environments and used biparental populations, which revealed variation between the involved parents. GWAS, in which variation is captured among a defined population, have also reported genomic regions associated with seed size, shape, and quality traits (Blair *et al*. 2009; Schmutz *et al*. 2014; Cichy *et al*. 2015b; Moghaddam *et al*. 2016; Giordani *et al*. 2022; Amongi *et al*. 2023). All of these genetic studies have described major, minor, and epistatic QTL for seed traits across all 11 common bean chromosomes (Arriagada *et al*. 2022). However, most of the QTL involved in the inheritance of seed phenotype traits have not been validated in different backgrounds (genotypes and environments) or are not well delimited in the bean genome, an important feature before being used in plant breeding.

Concerning the candidate genes controlling seed morphological traits, Schmutz *et al*. (2014) suggested 15 candidate genes associated with seed weight in a Mesoamerican panel consisting of 280 genotypes, three of which were highlighted by (Moghaddam *et al*. 2016) for seed weight in the same population (*Phvul.006G069300, Phvul.008G013300*, and *Phvul.010G017600*). Other GWAS using multi-environment trials for 4 decades confirmed the involvement of genes located on chromosomes Pv02 and Pv10, and found two additional genes for seed weight, *Phvul.002G150600* and *Phvul.003G039900* (MacQueen *et al*. 2020). Additionally, 13 candidate genes for seed shape and size were proposed by Giordani *et al*. (2022) based on a GWAS conducted on

a Brazilian panel of 180 accessions. However, analysis of gene expression in the Negro Jamapa genotype during seed development showed that 10,453 genes modified their expression levels, with the majority (9,701) showing decreased expression (O'Rourke *et al*. 2014). Many of these genes are transcription factors, although the genes involved in starch biosynthesis (e.g., *Phvul.001G082500.1*) and sucrose synthesis are highly expressed in developing seeds. Also, the high expression of abscisic acid biosynthesis genes (e.g., *Phvul.002G018700.1* and *Phvul.005G031500.1*) was observed in developing seeds, with expression decreasing as the seeds matured (O'Rourke *et al*. 2014). In other species, multiple pathways, including G-protein signaling, ubiquitin–proteasome pathway, mitogen-activated protein kinase (MAPK) pathway, BR signaling, transcriptional regulatory factors, and auxin signaling, are involved in the regulation of seed development (Li *et al*. 2019). In the ubiquitin–proteasome pathway, the ubiquitin receptor DA1 and E3 ubiquitin ligases EOD1/BB and DA2 physically interact to control seed size in Arabidopsis by regulating cell proliferation in integuments (Xia *et al*. 2013; Li *et al*. 2019). MAPK pathway consists of the different combinations of MKKK, MKK, and MAPK proteins which the plants use to regulate distinct biological processes, like plant growth, development, and defense response (Xu *et al*. 2018; Jiang *et al*. 2022), as well as could be important to regulate the grain size in rice (Xu *et al*. 2018; Tian *et al*. 2021; Wu *et al*. 2022). In addition, a LECTIN RECEPTOR KINASE, LecRK-VIII.2, has been reported to coordinate silique number, seed size, and seed number to determine seed yield in Arabidopsis by acting upstream of the MAPK gene (Xiao *et al*. 2021). All evidence indicates that the bean seed phenotype is regulated by a complex network of genes.

The bean genome is available (e.g.,(Schmutz *et al*. 2014)). Therefore, it is necessary to strengthen the connection between phenotype, genotype, and genome to identify annotated genes related to the expression of particular characters. The main goal of this study was to identify the consistent genomic regions and candidate genes involved in the genetic control of common bean seed size, shape, and quality traits, by combining association and differential expression studies. These analyses contribute to the understanding of the complex network of genes involved in seed phenotype control.

2. Materials and methods

2.1. Plant material

The SDP, with wide variation in seed phenotypes, was used in this study (Campa *et al*. 2018; https://zenodo.org/records/10263706). The SDP has homozygous lines derived from 220 landraces, most of which are from the updated Spanish Core Collection. SDP included 51 snap bean cultivars, 37 lines derived from traditional old cultivars, and well-known breeding lines. In all, 298 homozygous lines of SDP were used in this study.

The genotype ‘‘Xana’’ was used for the analysis of DEG during seed development (Figure 3.1). The cultivar ‘‘Xana’’ was grouped in the A gene pool (Campa *et al*. 2018). This genotype has very large white seeds, determinate growth habits, and is classified as Fabada market class. ‘‘Xana’’ is included in the SDP as line SDP308.

Figure 3. 1. Pods of cv. ‘Xana’ showing the three different growth stages of seeds used in the differential expression analysis.

2.2. Genotyping

The GBS method, as described by Elshire *et al*. (2011), was performed at BGI-Tech (Copenhagen, Denmark) using *ApeKI* restriction enzyme (Campa *et al*. 2018; Table 3S.1). Sequencing reads from different genotypes were aligned using the bean genome G19833 v1.0, https://www.ncbi.nlm.nih.gov/ genome/380). Genotypic data were filtered using software Tassel v5 (Bradbury *et al*. 2007). Lines with more than 50% missing data were removed, and in the remaining genotypes, SNP were filtered using the following criteria: (i) proportion of missing data < 10% and (ii) minor allele frequency (MAF) > 0.05. SNPs were named according to their physical position in the bean genome G19833 v1 (chromosome followed by the physical position in base pairs).

2.3. Phenotyping

The SDP was grown in a greenhouse in Villaviciosa, Spain (43°29′01′ N, 5°26′11′ W; elevation 6.5 m) for three seasons (2016, 2017, and 2018). The experimental design was a randomized complete block design in which there was a replicate with a single 1 m row plot, including 8–10 plants per line. Standard agronomic practices for tillage, irrigation, fertilization, and pest control were followed to ensure adequate plant growth and development. Phenotyping was conducted for seven seed traits: seed shape (area, length, width, and length-to-width ratio), seed weight, and seed quality after rehydration (coat proportion and water absorption) (Table 3.1). Seed dimensions were digitally recorded and analyzed using SmartGrain software (Tanabata *et al*. 2012). The trait 25-seed weight was manually recorded. Coat proportions and water absorption traits were recorded manually according to Castellanos *et al*. (1995).

Table 3. 1. Description of the seven analyzed seed traits. Codes for each character are indicated in parentheses.

Trait	Unit	Method*	Description
Seed area (SA)	mm^2	2	Measure of 25 randomly chosen seeds
Seed length (SL)	mm	2	Measure in parallel to the hilum of 25 randomly chosen seeds
Seed width (SWI)	mm	2	Measure perpendicular to the length of 25 randomly chosen seeds
Seed SL/SWI ratio (LWR)		2	Ratio SL/SWI
25-seed weight (SW)	g	1	Measure of four sets of 25 dry seeds
Water absorption (WA)	%	1	Average of a set of 25 seeds per plot without considering non-rehydrated seeds (according to (Castellanos *et al*. 1995))
Coat proportion (CP)	%	1	Average of a set of 10 seeds per plot (according to (Castellanos *et al*. 1995)

*(1) manually measured; (2) measured from digital images with the help of the software SmartGrain

2.4. Statistical analysis seed traits

All statistical analyses were performed using R software version 4.3.0 (R Core Team 2023). Outliers were removed before mean estimation using the mode 'blup' in the phenotype package (Piepho *et al*. 2008). Phenotypic variation in individual traits was visualized using the probability density distribution generated by the ggplot2 package (Wickham 2016). Spearman's correlation coefficients between the seven traits were calculated using the R corrplot package (Wei and Simko 2021).

HCPC analysis was performed to identify the main clusters in which bean lines could be grouped based on the genotypic data. This analysis was performed using R software with the packages FactoMineR and FactoExtra (Lê *et al*. 2008). Putative significant differences in seed traits among the clusters established from the HCPC analysis were investigated using ANOVA, followed by a post hoc Tukey test.

2.5. GWAS and haplotype block detection

Association analysis was performed using the FASTmrEMMA model (Wen *et al*. 2018), implemented in the mrMLM package (Zhang *et al*. 2020) of the R project (R Core Team 2023). PCA and the kinship matrix obtained by the centered-IBS method were considered to account for multiple levels of relatedness within the lines included in the panel. A restricted maximum likelihood (REML) and critical LOD score of 3 were

considered critical thresholds of significance for the identification of significant Quantitative Trait Nucleotide (QTN).

Haplotype blocks around QTNs were investigated using the Haploview 4.2 software (Barrett *et al*. 2005) with default software parameters and algorithms. The identified blocks were named using the prefix 'Seed,' chromosome number, and start position in Mb. To complete the characterization of the delimited blocks revealed by GWAS, the genomic positions in the bean genome were compared with those of QTL previously reported in common bean (González *et al*. 2016; Murube *et al*. 2020; Berry *et al*. 2020; Bassett *et al*. 2021; Ugwuanyi *et al*. 2022; Arriagada *et al*. 2022). The list of annotated genes under the QTL was established based on those residing within the region delimited by the leftmost and rightmost flanking SNP in the defined haplotype blocks.

2.6. RNA-Seq

Plants of the genotype ''Xana'' were grown in pots of 7 liters during the summer of 2022 under greenhouse conditions. Seeds collected at three different growth stages were used to identify DEGs by RNA-seq (see Figure 3.1): D1, the beginning of seed development (seeds with 0.8–1 cm length and green color); D2, intermediate development stage (1.5–2 cm length and green color); and D3, final development stage (2–2.5 cm length and green–white appearance). The experimental design included two biological replicates corresponding to two seed samples from different plants and pods.

The seeds were extracted from the pods, flash-frozen in liquid nitrogen, and stored at − 80 °C before RNA extraction. Total RNA from two biological replicates per growth stage was isolated using the RNeasy Plant Mini Kit, following the manufacturer's instructions (Qiagen, Germany). RNA was quantified using fluorometric methods, and the quantity was determined using a 2100 Bioanalyzer Instrument (Agilent Technologies, UK). RNA libraries were prepared using the TruSeq Stranded mRNA Sample Preparation Kit (Illumina), and sequencing was performed on the Illumina platform (Macrogen, Korea).

The reads were mapped to the reference genome G19833 v1.0 (Schmutz *et al*. 2014) using HISAT2 splice-aware aligner (Kim *et al*. 2015). Expression profiles were represented as read counts and normalized by calculating the Trimmed Mean of M-values

(TMM). Genes with expression levels less than 33% were removed. A PCA was performed to detect the possible sources of noise in the results. The NOISeqBIO function of the NOISeq package in R (Tarazona *et al*. 2015) was used to identify DEG through comparisons at different growth stages: D3 versus D1, D2 versus D1, and D3 versus D2 (see Figure 3.1). DEGs were identified using $q > 0.99$. Specific and common DEGs among the three comparisons were detected and visualized using Venn diagrams constructed using the package ggVenn/ggplot2 (Wickham 2016).

2.7. Gene ontology enrichment analysis

Gene ontology (GO) annotation was done using the '*Phaseolus vulgaris*' organism database in AnnotationHub resource (Morgan and Shepherd 2023) considering the three categories: BP, MF, and CC to investigate the functional groups of the observed DEGs. Overrepresentation analysis (ORA) of candidate genes was performed using the R package clusterProfiler (Wu *et al*. 2021) based on the hypergeometric test (p value) and Benjamin–Hochberg method for controlling the false discovery rate (q value).

2.8. Approach to candidate genes

The identification of potential candidate genes was focused on considering some of the following criteria: (i) The QTL regions revealed by this study and colocalized with previously reported QTL for the same traits; (ii) DEGs located in the genomic regions delimited in this study; (iii) match with the candidate genes for domestication events previously proposed by Schmutz *et al*. (2014); (iv) match with the reported DEGs during seed development from the common bean expression atlas (O'Rourke *et al*. 2014).

Complete genomic sequences of the selected putative candidate genes were obtained from the Phytozome v12 database (MA gene pool genomes UI111 v1.1, Labor Ovalle1.1 and 5 593v1.1). The sequences obtained were analyzed by alignment (BLASTn) with sequences of the reference genome G19833 v1.1, using default parameters. The polymorphisms identified were nucleotide variation, insertions, deletions, and the number of predicted genes.

3. Results

3.1. Genotyping of the SDP

Genotyping of 298 lines included in the SDP was filtered considering homozygous sites, missing values (< 10%), and minor allele frequency (> 0.05), resulting in 4,658 SNPs distributed across the 11 bean chromosomes (Table 3S.1; Figure 3S.1). The number of SNPs per chromosome ranged between 298 (Pv10) and 619 (Pv02) SNPs.

3.2. Phenotypic variation

Table 3.2 shows the observed variation in the seven traits evaluated. Phenotypic evaluation of the SDP panel revealed a wide variation in all cases (see Figure 3S.2). For instance, SL and SW ranged from 8.68 mm (observed in SDP262) to 22.41 mm (SDP308), and from 3.9 g (SDP009) to 27.59 g (SDP308), respectively (see Table 3S.2). Similarly, WA showed wide variation in this panel, ranging from 39.3% (SDP083) to 63.3% (SDP106). All traits showed a good fit to a normal distribution (Kolmogorov–Smirnov test), except for LWR (Figure 3S.2). Significant correlations were detected in most cases, except for SW, LWR, WA, SA, SW, and CP (Figure 3.2). Correlations (r) ranged from − 0.07 to 0.94 and were positive in most cases. Overall, 11 of the 21 trait combinations showed moderately positive correlations (r > 0.4).

Table 3. 2. Observed variation in the seven seed traits analyzed in the SDP. Means, standard deviations (SD), and variation intervals (min–max) are indicated.

Seed trait	Mean	SD	Min-Max
Area (mm^2)	86.86	23.72	41.46 - 184.37
Length (mm)	13.61	2.41	8.68 - 22.41
Width (mm)	7.97	1.03	5.00 – 10.59
SL/SWI ratio	1.73	0.34	1.21 - 2.98
25Seed Weight (g)	12.53	4.14	3.90 - 27.59
Coat Proportion (%)	10.41	3.36	3.37 - 25.38
Water Absorption (%)	51.49	3.11	39.26 - 63.35

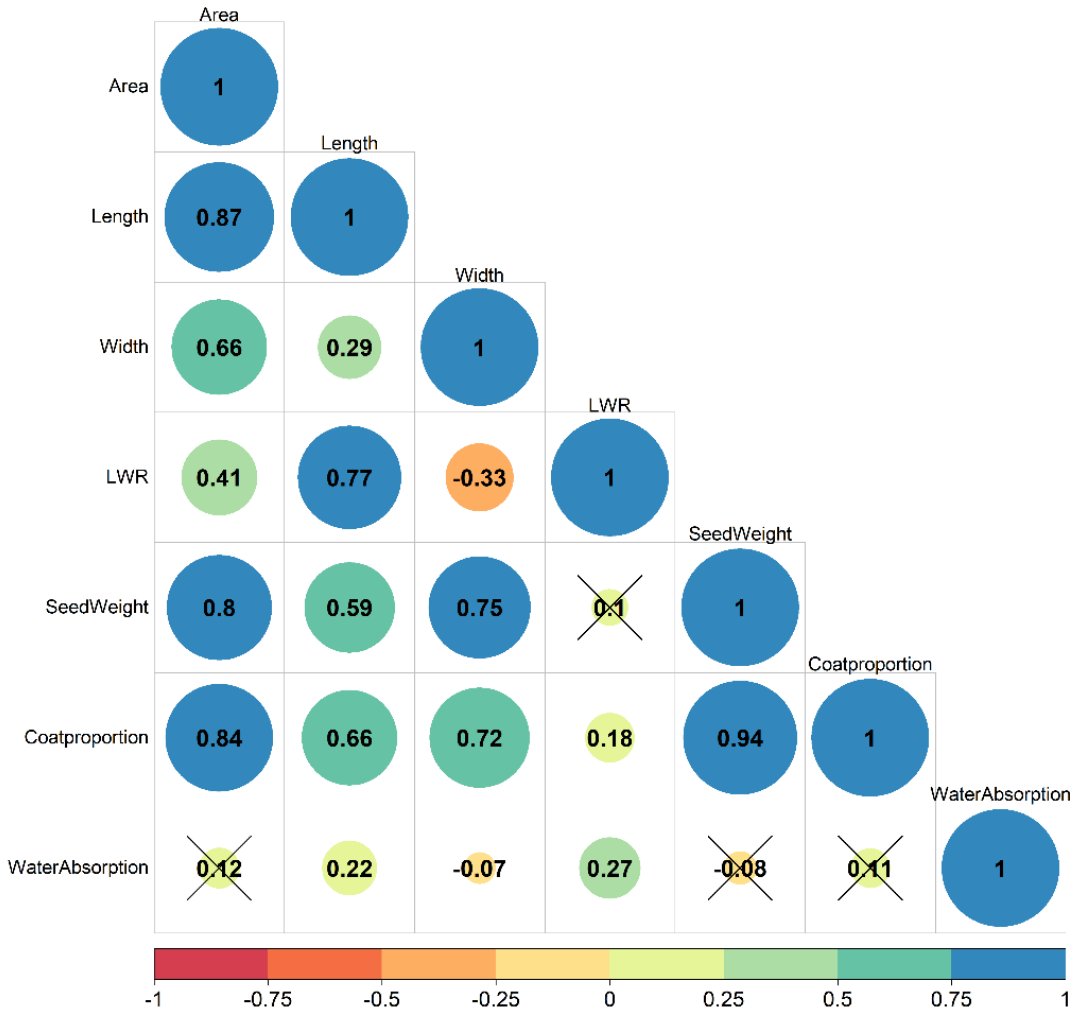

Figure 3. 2. Spearman's correlation coefficient between the evaluated seed traits. Black crosses represent not significantly associated (p>0.05).

3.3. Genome-wide association study

Association analysis revealed 38 QTN, although 11 SNPs were associated with more than one character. The 27 unique SNPs were located on 10 chromosomes (all except for chromosome Pv06). The distribution and characteristics of QTN are presented in Table 3.3 (see Figure 3S.3). For instance, the six QTN for SW were located on chromosomes Pv01, Pv03, Pv04, v07, Pv08, and Pv10, whereas those for SL were located on chromosomes Pv04, Pv07, Pv08 (two regions), Pv10, and Pv11.

Table 3. 3. Characteristics of the significant associations (LOD > 3) SNP-traits identified using the FASTmrEMMA method in the SDP for seven seed traits.

SNP	Seed trait	Chr	QTN effect	LOD score	$-\log_{10}(p)$	r^2 (%)	MAF
s1_50842559	CP	Pv01	1.78	6.09	6.92	4.46	0.21
s1_52137885	SW	Pv01	− 1.4	3.69	4.42	1.63	0.18
s2_28212493	LWR	Pv02	− 0.1	3.64	4.37	2.14	0.45
s2_37004268	SWI	Pv02	0.38	3.71	4.44	3.01	0.36
s2_39810114	LWR	Pv02	− 0.19	11.3	12.26	7.13	0.42
s3_46955356	SA	Pv03	16.12	6.09	6.93	9.1	0.29
s3_46955356	SWI	Pv03	− 0.87	11.34	12.3	9.8	0.17
s3_47996582	SW	Pv03	− 3.3	12.17	13.15	8.74	0.17
s3_47996582	CP	Pv03	− 2.54	10.37	11.32	7.88	0.17
s4_26787677	SA	Pv04	− 14.89	11.61	12.58	7.96	0.29
s4_26787677	SL	Pv04	− 1.03	4.96	5.76	3.66	0.29
s4_26787677	SW	Pv04	− 0.95	3.06	3.76	1.07	0.29
s4_43792143	LWR	Pv04	0.15	5.49	6.31	3.79	0.27
s5_3286816	WA	Pv05	2.11	4.92	5.72	10.82	0.41
s5_37260589	WA	Pv05	1.37	4.12	4.88	4.17	0.33
s5_37715374	LWR	Pv05	− 0.26	9.88	10.82	3.02	0.06
s7_633265	SL	Pv07	0.88	5.63	6.45	3.11	0.39
s7_663226	SW	Pv07	2.09	7.23	8.1	5.02	0.29
s7_3895030	LWR	Pv07	0.08	3.4	4.12	1.48	0.44
s7_557444	CP	Pv07	1.79	4.54	5.32	6	0.32
s8_55946412	SA	Pv08	11.05	6.02	6.85	3.46	0.2
s8_55946412	SL	Pv08	1.21	6.16	6.99	4	0.2
s8_55946412	LWR	Pv08	0.11	5.22	6.03	1.78	0.2
s8_55946412	SW	Pv08	2.08	7.05	7.91	4.02	0.2
s8_55946412	CP	Pv08	1.59	4.99	5.79	3.56	0.2
s9_12670503	SWI	Pv09	0.39	3.1	3.8	1.86	0.16
s9_31015943	LWR	Pv09	0.19	8.04	8.93	2.14	0.08
s10_32685348	WA	Pv10	1.43	5.08	5.87	5.01	0.44
s10_39185557	SA	Pv10	− 24.84	5.94	6.77	9.1	0.09
s10_39185557	SL	Pv10	− 1.77	4.72	5.5	4.43	0.09
s10_39193928	SW	Pv10	− 3.5	5.34	6.15	5.91	0.1
s10_39193928	CP	Pv10	− 2.7	4.68	5.46	5.35	0.1
s10_40537163	SA	Pv10	− 16.53	6.4	7.25	6.32	0.16
s10_40537163	SL	Pv10	− 1.55	5.2	6	5.32	0.16
s10_40979207	SWI	Pv10	− 0.63	5.6	6.42	4.8	0.15
s11_1587588	SL	Pv11	− 1.06	6.57	7.42	4.57	0.4
s11_1701280	SWI	Pv11	− 0.47	5.96	6.79	3.23	0.22
s11_2105912	SA	Pv11	− 12.36	7.05	7.92	5.02	0.25

Linkage disequilibrium (LD) analysis showed that the 27 SNP were organized in 23 genomic regions (blocks) ranging between 5 bp (Seed02_39.8) and 8.4 Mbp (Seed09_23.4) (Table 4). The QTL region Seed11_1.5 was tagged with an SNP. The positions of the genomic regions were compared to previously reported QTL regions associated with seed traits. Seven studies on biparental populations and diversity panels were considered, and 34 QTL associated with the genetic control of seed traits were found, revealing 13 overlapping regions on chromosomes Pv01, Pv02, Pv03, Pv04, Pv05, Pv08, Pv09, and Pv10 (Table 3.4). Interestingly, eight regions (most of which were associated with SW) were detected in more than one study: Seed01_50.7, Seed01_51.9, Seed02_28.1, Seed03_45.6, Seed08_55.3, Seed09_10.1, Seed09_23.4, Seed10_39.1, and Seed10_40.3 (see Figure 3.3).

Table 3. 4. The genomic regions carrying the identified SNPs were revealed using linkage disequilibrium analysis. Correspondence with reported QTL for similar traits is shown. The reference of the reported QTL (Ref.), the numbers of reported DEG (Rep. DEG), and observed DEG (Ide. DEG) in each genomic region are also indicated.

QTL name	Seed trait	Genomic región		QTL reported			Gene expression		
		Start	End	Traits	QTL name	Ref*	Num. Genes	Ide. DEG	Rep. DEG
Seed01_50.7	CP	50,711,659	51,126,515	CP; SW; SCP	SPE1.1; SW-1^{AM}; qProtein-a	1; 4; 7	73	9	2
Seed01_51.9	SW	51,943,288	52,159,049	SW; SPC	SW1.3; qProtein-a	4; 7	25	2	1
Seed02_28.1	LWR	28,125,848	29,131,665	SW; CP	SW-1MA; Yd_MQTL2.4; SW.2.2; SCP.2.2	1; 2; 5	62	4	4
Seed02_36.9	SWI	36,917,430	37,025,044				12	2	1
Seed02_39.8	LWR	39,810,109	39,810,114				0	0	
Seed03_45.6	SA SWI SW CP	45,689,287	48,502,614	SW	SW3.3; SW3.1	2; 4	234	15	16
Seed04_25.6	SA SL SW	25,641,883	27,416,740				24	0	1
Seed04-43.7	LWR	43,792,143	43,865,501	SW	SW4.1^{SA}	1	8	0	0
Seed05_3.2	WA	3,286,515	3,324,361	WA	WU.5.1	2	6	0	0
Seed05_37.1	WA	37,141,875	37,376,050				26	3	0
Seed05_37.6	LWR	37,685,717	37,988,349	SW	SW-5MA	1	35	2	3
Seed07_0.55	CP	557,444	620,707				9	1	1
Seed07_0.62	SL SW	627,376	1,090,742				76	1	5
Seed07_3.8	LWR	3,895,030	3,895,149				1	0	0

Seed08_55.3	SA SL LWR SW CP	55,393,031	56,063,905	SW; SWI; SA	SW8.3; SW8.1[AN,SA]; qSDia-b; S1_379992973	1; 4; 6; 7	66	3	4
Seed09_10.1	SWI	10,190,403	12,983,869	SW; CP; SP; SL	Yd_MQTL9.1; SCP.9.1; SP9[XB];SL9[XB]	3; 2; 5	263	14	23
Seed09_23.4	LWR	23,402,412	31,822,063	SW; SWI	qSDia-c; Yd_MQTL9.2	7; 5	559	27	43
Seed10_32.5	WA	32,549,846	32,685,348	WA; SW	WU10.1; SW10.1	4	4	0	1
Seed10_39.1	SA SL SW CP	39,180,865	39,358,330	SW; SWI; SL	SW10.1; qSDia-e; Yd_ MQTL10.2;SL10[XC]	3; 4; 7; 5	18	1	2
Seed10_40.3	SA SL SWI	40,365,242	40,979,207	SWI	SWI10[XC]	3	45	2	2
Seed11_1.5	SL	1,587,588	1,587,588				1	0	
Seed11_1.6	SWI	1,657,560	1,701,280				8	2	1
Seed11_1.9	SA	1,949,116	2,323,534				50	3	8

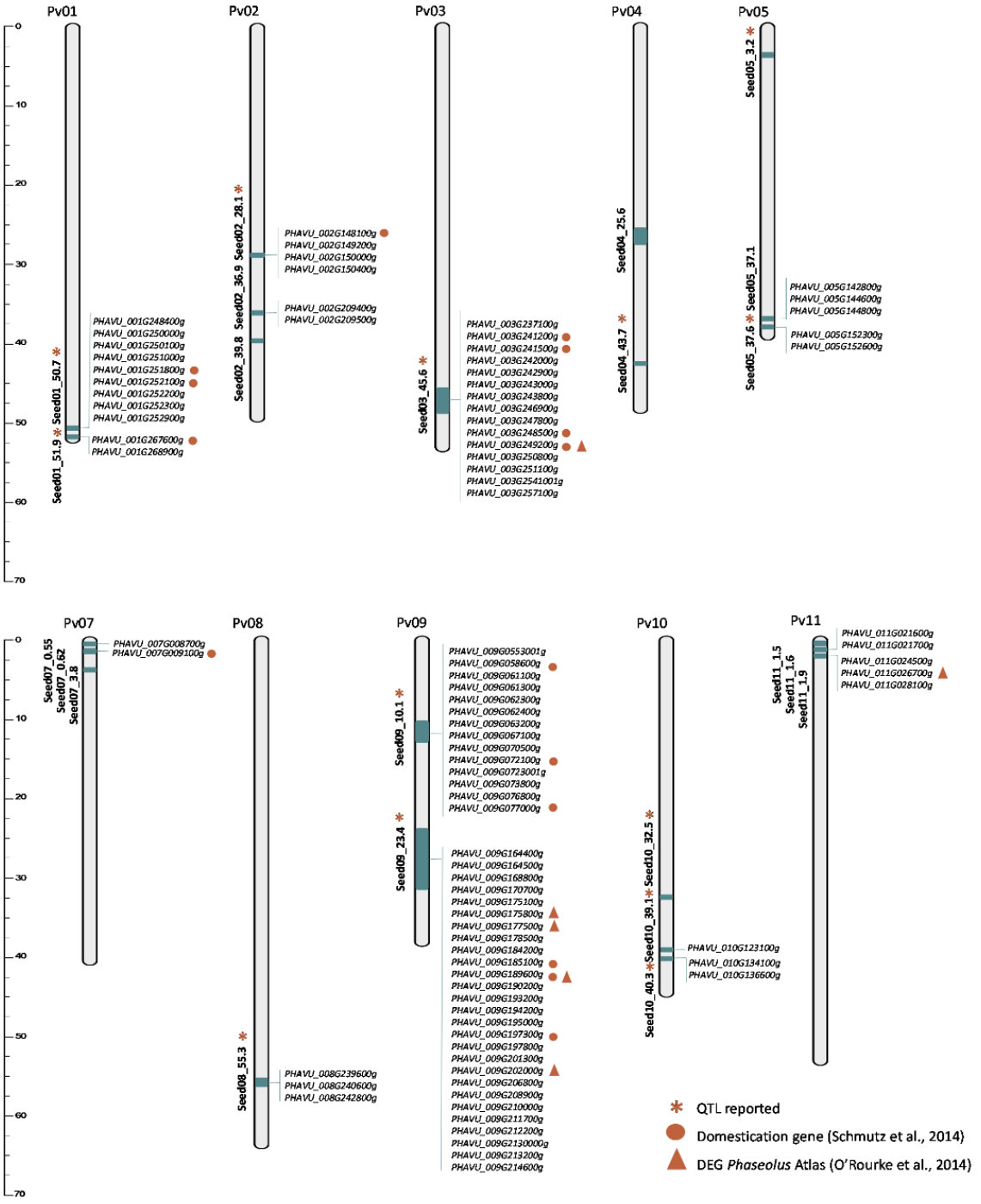

Figure 3. 3. Chromosomal positions of genomic regions associated with seven seed traits identified by GWAS (green boxes). * Regions colocated with previously reported QTL for seed traits. Genes differentially expressed during seed development underlying those genomic regions revealed by RNA-seq analysis are shown at the right of each chromosome.

Five QTL associated with SW have been consistently identified in other studies: Seed01_51.9, Seed03_45.6, Seed07_0.62, Seed08_55.3, and Seed10_39.1. A total of 102

SNPs that tagged these five QTL were selected, and HCPC analysis was performed with these SNPs revealing two main dimensions that explained 47.3% of the variance and led to the establishment of three main clusters with the SDP lines (Figure 3.4; Table 3S.2).

- **Cluster 1** is formed by 85 lines, including typical MA genotypes such as Sanilac (SDP290), Cornell49242 (SDP225), IVT7214 (SDP248), and AB136 (SDP005). The group had an average weight of 10.54 g per 25 seeds and contained 16 lines classified in the intermediate population and 1 classified in the A population (Campa *et al*. 2018).
- **Cluster 2** is formed by 43 lines, 38 of them classified as intermediate between both gene pool. The group had an average weight of 10.37 g per 25 seeds and did not differ significantly from cluster 1. Many of these lines are snap bean cultivars such as Fin de Bagnols (SDP232), Triomphe de Farcy (SDP293), Gloire De Saumur (SDP242), and Manteca de los Mercados (SDP247).
- **Cluster 3** consisted of 170 lines, of which 141 were classified in the A gene pool. This group included the typical A cultivars Tendergreen (SDP295), MDRK (SDP256), Perry Marrow (SDP276), and G19833 (SDP238), one of the bean genomes available (Schmutz *et al*. 2014). Cluster 3 included 28 lines classified as intermediate between both gene pools and a line grouped in the MA gene pool (SDP080). The mean of the 25 seeds weight was 14.06 g in this group, which differed significantly from that of clusters 1 and 2.

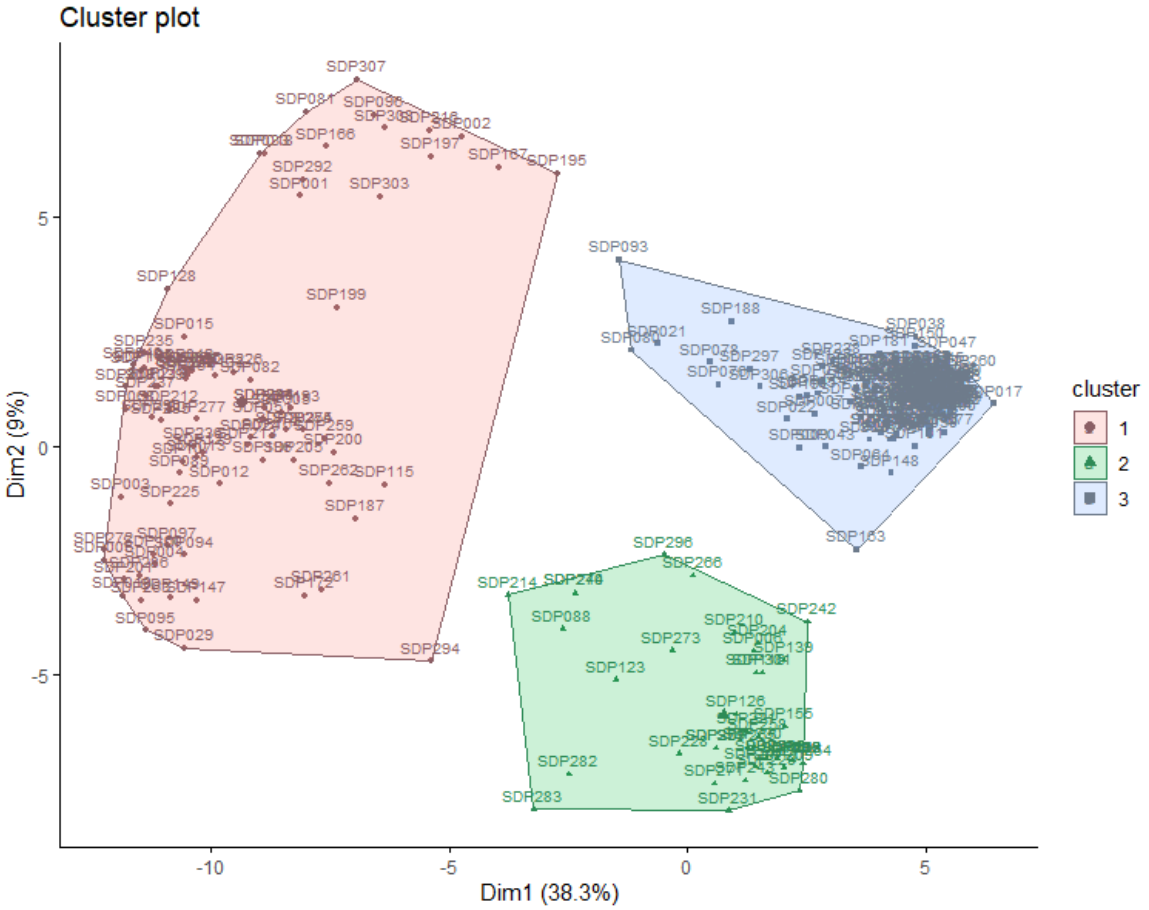

Figure 3. 4. Biplots showing the results of hierarchical clustering on principal components using 102 SNP tagging the five consistent QTL regions associated with SW.

An analysis was also carried out for each QTL region to evaluate the effect of each region on seed weight. HCPC analysis showed three groups for each QTL region (Figure 3S.4). The percentage of explained variance varied between 52.3% (Seed07_0.62) and 89% (Seed01_51.9). Significant differences in SW were detected between the groups of lines established for each QTL (Table 3.5). The QTL regions Seed01_51.9 and Seed10_39.1 showed the greatest differences between the two groups with extreme mean values (6.63 and 6.47 g, respectively) and significant differences in the means of the three groups.

Table 3. 5. Mean values for 25-seed weight in the three clusters obtained from the HCPC analysis using SNP tagging of the 5 consistent QTL regions associated with seed weight (see Figure 3S.4). Results of analysis of variance (ANOVA).

	5 QTL	Seed01_51.9	Seed03_45.6	Seed07_0.62	Seed08_55.3	Seed10_39.1
Num. SNP	102	7	42	20	21	12
Percentage explained variance (%)	47.30	89	67.30	52.30	73.40	68.90
Cluster 1						
Mean SW	10.54^{a}	13.02^{a}	13.26^{a}	13.34^{a}	9.27^{a}	9.24^{a}
Num. lines	85	118	211	215	45	89
Cluster 2						
Mean SW	10.37^{a}	16.80^{b}	8.22^{b}	14.09^{a}	10.34^{a}	15.71^{b}
Num. lines	43	55	39	9	68	29
Cluster 3						
Mean SW	14.06^{b}	10.17^{c}	12.78^{a}	9.94^{b}	14.11^{b}	13.63^{c}
Num. lines	170	125	48	74	185	180
ANOVA	$9.8e^{-14}$***	$< 2e^{-16}$***	$< 2.92e^{-12}$***	$1.4e^{-9}$***	$< 2e^{-16}$***	$< 2e^{-16}$***

3.4. Differentially expressed genes

The read counts for the expression levels at each seed development stage (D1, D2, and D3) and the replicates per locus are shown in Table 3S.3. The reads from all samples were used for transcriptome assembly, and an average of 91.2% of the reads were mapped to the reference genome. The mapped reads were normalized by calculating the TMM-normalized reads, which revealed two components that explained 67% of the variation (Figure 3S.5a). The obtained plot shows the grouped samples, except in the case of a sample derived from the D1 development stage (Figure 3S.5b). DEGs were identified by comparing three stages of seed development. (D2-D1; D3-D1; D3-D2; see Figure 3.1). In total, 2,085 differentially expressed genes involving 1,992 unique genes were identified in this analysis (Table 3S.4). The majority of DEGs were up-regulated (1,888) (Figure 3.5), and most of them were detected when the comparison was made between stages D3 and D1. Down-regulated DEGs were not detected in the comparison between the development stages, and up-regulated DEGs common among the three stages were also not detected.

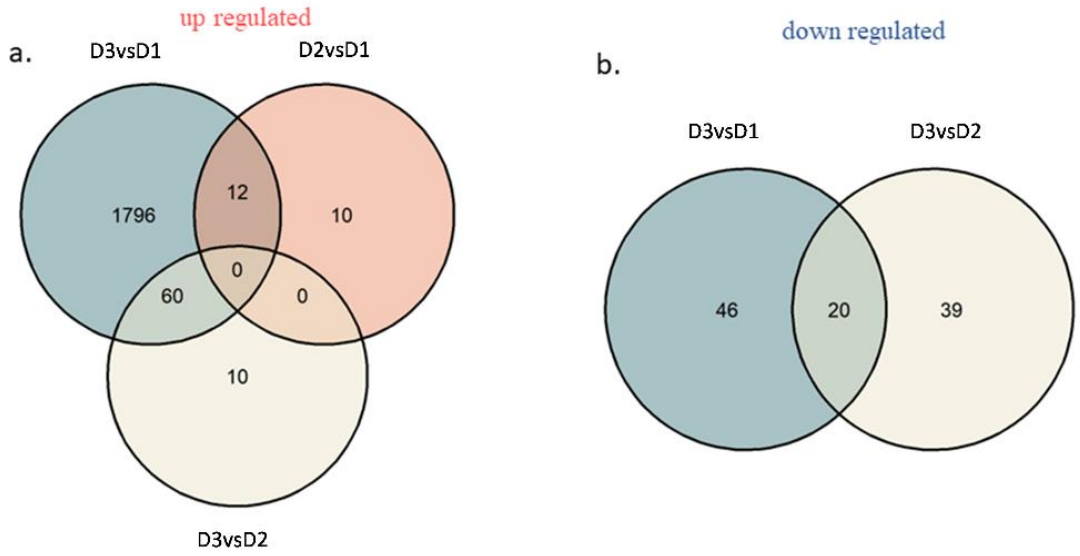

Figure 3. 5. Venn diagrams showing DEGs detected in each comparison (D2 vs. D1, D3 vs. D1, D3 vs. D2). A) Up-regulated genes. B) Down-regulated genes.

3.5. GO enrichment analysis of DEGs

GO enrichment analysis was performed with 1,934 genes, corresponding to the DEGs in D3 compared to D1. The BP and CC categories showed enrichment terms, whereas the MF category did not (Table 3S.5). For BP, 19 GO terms were enriched, with more important terms related to intracellular cell establishment and functions, such as protein transport, protein and macromolecule localization, and compound metabolic processes (Figure 3S.6a). For CC, 34 terms were enriched, most of which were implicated in functions related to endoplasmic reticulum or coated COPI-coated vesicles (Figure 3S.6b).

3.6. Putative candidate genes for seed traits

The 23 QTL contained 1,605 annotated genes in the G19833 genome, and 91 genes were differentially expressed in RNAseq analysis during seed development (Table 3S.6). These 91 differentially expressed genes were located in only 16 QTL regions (Table 3S.7). The number of DEGs per region ranged between one (Seed10_39.1) and 27 (Seed09_23.4). Six of these 91 genes were described in the gene expression atlas during

seed development: *PHAVU_003G249200g, PHAVU_009G175800g, PHAVU_009G177500g, PHAVU_009G189600g, PHAVU_009G202000g*, and *PHAVU_011G026700g* ((O'Rourke *et al*. 2014); Table 3S.6). In addition, 15 of these 91 genes have been reported as relevant genes in the domestication process by Schmutz *et al*. (2014). Functional annotation of these 91 DEG revealed molecular functions and biological processes already reported in seed development (see Table 3S.7).

The reference bean genome (G19833) contained 22 DEGs located in five consistent QTL regions associated with SW. The genomes of Labor Ovalle, UI111, and 5-593 had predicted homologous genes for each of these DEGs, except for *PHAVU_010G123100g*, which had two genes (Table 3S.8). The respective sequences of these genes were aligned, and different types of variation were observed, compared to the reference genome: mismatches, insertions, deletions, and duplications. The genome of 5-593 presented less variation than that of Labor Ovalle and UI111 compared to the genome of G19833. However, the levels of variation were not the same for the 22 DEGs. Thirteen of the 15 predicted genes located on chromosome 3 showed a very low variation with G19833. In contracts, eight of the 22 genes showed high variation in the three genomes in the respective alignments with G19883: *PHAVU_001G267600g, PHAVU_001G268900g, PHAVU_003G241500g, PHAVU_007G009100g, PHAVU_008G239600g, PHAVU_008G240600g, PHAVU_008G242800g*, and *PHAVU_010G123100g*. For example, the three homologous genes of *PHAVU_008G240600g* had an insertion of 46 bp.

4. Discussion

Seed phenotype is a relevant characteristic of both dry and snap beans. While many studies have focused on seed coat color inheritance, there is limited research on the genetic control of seed size, seed shape, and seed quality traits, as which is an amino acid positively correlated with protein concentration in soybean seeds (Pandurangan *et al*. 2012). The Seed10_39.1 region, which is also associated with CP, contains the gene *PHAVU_010G123100g*, which encodes a pectinesterase inhibitor involved in seed coat development in Arabidopsis. WA rate is also related to the emergence and germination of plants (Powell *et al*. 1986; Vidak *et al*. 2022), and some putative QTLs that control this

trait have been reported (Cichy *et al*. 2015b; Diaz *et al*. 2020; Berry *et al*. 2020; Bassett *et al*. 2021). The region Seed05_37.1, found in this work, associated with WA was reported by Berry *et al*. (2020), also, the region Seed10_32.5 was previously described as WA associated by Bassett *et al*. (2021).

Seed phenotype is the result of seed development. Two distinct phases during seed development have been described in legumes: The first phase involves cell division in the embryo, followed by a second phase, which regulates seed thickness via cell expansion and is highly influenced by the environment (Domoney *et al*. 2006). Changes in the transcriptomic profile of the ‘‘Xana’’ cultivar during seed development were studied and 1,992 DEGs were detected. Many DEGs were found in the comparison between development stages D3 and D1. GO enrichment analysis revealed GO enrichment only in the categories of biological processes and cellular components. The enriched GO terms have important functions in plant development processes, such as those related to the Golgi apparatus, endoplasmic reticulum, and coated vesicles, which are essential for plant growth (Ahn *et al*. 2015). In all, 91 of 1,992 DEGs were located under the 23 QTL regions and 15 of them (*PHAVU_001G251800g; PHAVU_001G252100g; PHAVU_001G267600g; PHAVU_009G072100g; PHAVU_003G241200g; PHAVU_003G241500g; PHAVU_003G248500g; PHAVU_003G249200g; PHAVU_007G009100g; PHAVU_009G058600g; PHAVU_009G072100g; PHAVU_009G07700g; PHAVU_009G185100g; PHAVU_009G189600g; PHAVU_009G197300g*) were considered associated to domestication events by Schmutz *et al*. (2014). On the other hand, six DEG identified in this study located underlying some QTL regions were also identified as DEG in seed development by O’Rourke *et al*. (2014): *PHAVU_003G249200g, PHAVU_009G175800g, PHAVU_009G177500g, PHAVU_009G189600g*, *PHAVU_009G202000g*, and *PHAVU_011G026700g*. Identification of the same genes in different studies consolidates their involvement in controlling seed phenotypes. The annotated function of these 91 DEG in the QTL region agreed with functions already reported in seed development. For example, functions related to ubiquitin activities are known to determine seed size in Arabidopsis and rice (Li and Li 2014); the gene *PHAVU_002G148100g* encodes a ubiquitin hydrolase; the gene *PHAVU_003G250800g* encodes a ubiquitin receptor DA1, which in *Arabidopsis thaliana* controls seed and organ growth by restricting cell proliferation (Li and Li 2014);

and the gene *PHAVU_005G142800g* encodes a ubiquitin ligase similar to the E3 ligase EOD1/BB identified as a negative regulator of seed size (Li *et al*. 2008). Across these 91 genes, we also found functions important in seed development in other species, such as AFP1-RELATED protein (*PHAVU_009G202000g*), expressed in embryos during the latest stages of seed maturation of Arabidopsis, and PPR protein (*PHAVU_009G175100g*), which play important roles in seed development in higher plants (Li *et al*. 2021). MAPKs control signaling cascades that play essential roles in plant growth, development, and defense response (Jiang *et al*. 2022), and are involved in regulating seed size in rice (Tian *et al*. 2021; Wu *et al*. 2022). The gene *PHAVU_009G062400g*, which encodes a MAP3K3/ MEKK3, could be part of this network. Acting upstream of the MAPK gene in Arabidopsis can be found LecRK-VIII.2 which coordinates silique number, seed size, and seed number to determine seed yield (Xiao *et al*. 2021). A homologous of this gene is a DEG located under the QTL Seed08_55.3, *PHAVU_008G239600g* which encodes a LECTIN RECEPTOR KINASE VIII.1.

Four bean genomes with predicted genes are available: one from the A gene pool (G19833) and three closely related to the MA gene pool (UI111, Labor Ovalle, and 5-593). Genotype G19833 has seeds larger than the other three genotypes. The 22 DEGs located in the consistent QTL associated with SW showed high variation when the respective sequences were aligned with the genes predicted in G19833, which may be a consequence of evolutionary differentiation and may contribute to phenotypic differentiation for SW. This variation was not homogeneous among the three MA genotypes, and 12 predicted genes in genotype 5-593 were very similar to those in G19833, suggesting that this observed variation should not be relevant to explicating the phenotypic variation between both gene pools for SW. In contrast, 10 genes were highly variable to those of G19833 in the three MA genotypes (see Table 3S.8), suggesting that they could be relevant for phenotypic variation between both gene pools for seed weight. Interestingly three of them (*PHAVU_001G267600g*, *PHAVU_001G268900g, PHAVU_010G123100g*) were located in the QTLs that provided the greatest differences in SW: Seed01_51.9 and Seed10_39.1. From the sequences of these genes, functional markers tagging both genes and QTL regions can be developed and used in MAS in breeding programs where SW is a trait considered. However, evidence suggests that SW

is not the result of variation in a few loci; rather, it is the consequence of variation in many loci and their interactions with the environment in which the plants develop. Thus, these markers can help to enrich segregating populations in certain SW phenotypes.

5. Conclusion

The combination of GWAS and RNA-seq analyses helped elucidate QTL regions and candidate genes that control seed size, shape, and quality traits. GWAS revealed 23 QTL regions that were significantly associated with the evaluated traits, 13 of which were consistent with the regions reported in previous studies. These QTL regions contained 1,605 annotated genes in the G19833 bean genome, of which 91 genes were differentially expressed during seed development in the cultivar ‘‘Xana’.’ DEGs were only found in 16 QTL, and 22 DEGs were located in five consistent QTL regions associated with SW. These regions and DEGs constitute a priority set for future genetic studies focused on SW control, their identification increases our knowledge of the genetic architecture of this trait, and a marker can be used as indirect selection tool, which is a relevant characteristic in many breeding programs.

CHAPTER 4

A new bean genomic resource: de novo assembly and annotation of a Fabada cultivar

This Chapter is a preliminary version that will be summited to the Journal *Data in Brief*. All data described in this chapter are planned for submission at the NCBI for public access.

A new bean genomic resource: de novo assembly and annotation of a Fabada cultivar

Fabada dry bean is a traditional high-quality common bean market class from Northern Spain characterized by a particular seed phenotype with white, oblong, and very large seeds (90-100 grams/100 seeds). Genotypic characterization of this market class reveals a genome mainly of A origin with approximately 30% introgression of M gene pool. In this study, *de novo* genome assembly and annotation of the Fabada bean line A25 maintained at the SERIDA seed collection was conducted. High-molecular-weight DNA was extracted from young trifoliate leaves and sequenced using the Illumina (genome coverage 128.75x) and PacBio Sequel II (genome coverage 73x) platforms. *De novo* genome assembly resulted in 470.324 Mbp organized in 1,368 scaffolds and the mitochondrial and chloroplast genomes. These scaffolds were assembled into the 11 chromosomes of the species based on the reference genome of G19833 (NCBI accession number GCF_000499845.1), and a total of 129 scaffolds were unplaced. RNA-seq reads of the seedling stage of the fabada bean line A25 were used for structural annotation (NCBI accession number: PRJNA851559). Annotation predicted a total of 59,426 protein-coding sequences. The mitochondrial genome was assembled and annotated based on that of *Phaseolus vulgaris* (NCBI accession number NC_045135), resulting in two FASTA files of 371,437 bp and 76 annotated genes, and 11,183 bp and 5 annotated genes. The chloroplast genome was assembled and annotated based on that of *Phaseolus vulgaris* (NCBI accession number NC_009259.1) resulting in a circular DNA molecule of 151,310 bp and 394 annotated genes. This is the second European genome available and, as far as is known, is the first genome available with recombination between both gene pools. It will be useful for the development of the future pangenome of the species and will be particularly interesting for the study and validation of genes involved in seed phenotype and seed quality.

1. Introduction

Common bean (*Phaseolus vulgaris*) is one of the most important food legumes for human consumption globally. Common bean is a highly diverse species, consisting of two main gene pools: A and MA. The fact that the domestication of the species took place in two independent events resulted in the differentiation of genes related to domestication. A total of 1,835 candidate genes of MA origin and 748 of A origin have been identified, among which only 59 are common, which indicates a null gene flow between gene pools (Schmutz *et al*. 2014). The MA gene pool is more structured and genetically diverse than the A gene pool. The first common bean reference genome was published in 2014 (Schmutz *et al*. 2014). The landrace line G19833, derived from the A pool (Race Peru), was sequenced and annotated resulting in 11 chromosomes of 472.5 Mb and 27,197 protein-coding loci. This reference genome has been of great use in the characterization of genetic diversity in terms of SNPs, InDels, and structural variation of genes. However, this genotype is photoperiod sensitive, and a single reference genome represents only a small fraction of all the genetic diversity of the species. In this context, the availability of other genomes, as well as the development of pan-genomes, is desirable to represent all the gene repertoires of a species (Bayer *et al*. 2020). As mentioned in the general introduction, eight *P. vulgaris* genotypes have been published to date, in total ten genomes, with only one from a cultivar of European origin (Carrère *et al*. 2023).

The fabada market class is a dry bean characterized by a distinct seed phenotype, featuring very large white seeds (≈ 100 g/100 seeds) and an oblong shape with a length/width ratio greater than 2.2 and described for the first time in the north of Spain by the mid-20th century (Puerta Romero 1961). This market class reveals a genome mainly of Andean origin, with about 30% introgression of Mesoamerican origin (Campa *et al*. 2018). In this chapter, a variety included in the market class Fabada is sequenced and annotated to expand the genetic knowledge of the species. This will be the second available genotype of European origin and will be especially useful for the study of seed size and seed quality genes.

2. Material and methods

2.1. Genomic DNA Extraction and Sequencing

Common bean line A25 (cv. 'Andecha') is an old cultivar within the market class Fabada since the early 2000s (white and very large seeds). High molecular weight genomic DNA from A25 leaf tissue was isolated following Schalamun and Schwessinger (2017) with minor modifications. The DNA was quantified using the Qubit High Sensitivity dsDNA Assay (Thermo Fisher Scientific). Sequencing was conducted using Illumina and PacBio technologies. The genomic DNA libraries were conducted by the company Allgenetics following Carøe and Bohmann (2020) with minor modifications for Illumina and the SMRTbell Express Template Prep Kit 2.0 for PacBio. The fragment size distribution and concentration of the libraries were checked in the Agilent 2100 Bioanalyzer. The Illumina library was sequenced in a NovaSeq PE150 flow cell, while the PacBio library was sequenced in a Sequel II platform with an SMRT (Single Molecule Real-Time) Cell 8M and using the Circular-Consensus Sequencing (CCS) mode. The resulting CCS raw reads were converted to HiFi reads. The quality of the reads obtained in FASTQ files was assessed using FastQC (Andrews 2010).

2.2. Genome Assembly

Short and long genomic sequencing reads were assembled *de novo* using the software MaSuRCA v3.4.2 (Zimin *et al*. 2017) and subsequently polished using POLCA (Zimin and Salzberg 2020). Genome scaffolding was performed with AGOUTI v0.3.3 (Zhang *et al*. 2016) using paired-end RNA reads to guide (NCBI accession number: PRJNA851559). To validate the assembly, the quality and completeness of the genome assembly were evaluated using BUSCO V5.beta.1 (Seppey *et al*. 2019). Chromosome assembly of the scaffolds was performed using the function scaffold in ragtag (Alonge *et al*. 2022) with the bean genome G19833 v1.0 as a reference (Schmutz *et al*. 2014). The quality of the chromosome assembly level was checked using QUAST (Gurevich *et al*. 2013).

Mitochondrial and chloroplastic reads were detected with the alignment of the initial quality-filtered reads to the complete chloroplastic and mitochondrial genomes of *P. vulgaris* (NCBI Reference Sequence: NC_009259.1 and NC_045135) using the BWA-MEM (0.7.15-r1140) algorithm (Li 2013). The mapped reads were extracted using

Samtools (Li *et al*. 2009) and Sabamba (Tarasov *et al*. 2015) and then used for *de novo* assembly using NOVOPlasty v4.2 (Dierckxsens *et al*. 2017).

2.3. Structural and functional annotation

The structural annotation was performed with the help of the software AGOUTI v0.3.3 (Zhang *et al*. 2016) using the RNA reads of the NCBI accession number PRJNA851559 (Jurado *et al*. 2022) completed with the gene prediction of AUGUSTUS (3.4.0) (Stanke and Morgenstern 2005) and *Arabidopsis thaliana* as a model species. TranDecoder v5.5.0 was used to identify the candidate coding regions in the predicted genes. The predicted protein-coding sequences were used to carry on the functional annotation using InterProScan (Jones *et al*. 2014) and Sma3s v2 (Muñoz-Mérida *et al*. 2014). RNA fasta sequences were used to do a spliced alignment with the whole genome using Minimap2 (Li 2018) to generate a GFF3 annotation file. The annotation of the organelles was performed using the GeSeq web server (Tillich *et al*. 2017) setting the option "BLAT Reference Sequences = *Phaseolus vulgaris"*.

2.4. Genomes comparison

Attending to the assembly level, the size of the genome and number of annotated genes ten common bean public genomes (Table 4.1) were explored in the following databases: NCBI (https://www.ncbi.nlm.nih.gov/genome/cgv), Phytozome (https://phytozome-next.jgi.doe.gov/) and LIS (https://www.legumeinfo.org/).

Table 4. 1. Common bean genomes information consulted (accessed 03 May 2024).

Genotype	Origin	Publication date	Database	Database ID	Sequencer type	Link
G19833 v1	A	2014	NCBI	ANNZ01	Illumina	ncbi/GCF_000499845.1
G19833 v2	A	2018/2024	Phytozome/NCBI	442/ ANNZ02	PacBio	phytozome/Pvulgaris_v2_1/ncbiGCA_000499845.2
Flavert	A	2023	NCBI	JARGYP01	PacBio	ncbi/GCA_029448765.1
OAC Rex	MA	2020	NCBI	JADFUL01	Illumina; PacBio	ncbi/GCA_015708805.1
JaloEEP558	A	2020	NCBI	JAAIFH01	Illumina	ncbi/GCA_016509735.1
BAT93	MA	2016	NCBI	LPQZ01	454 SOLiD;Sanger	ncbi/GCA_001517995.1
BAT93	MA	2020	NCBI	JAAIFG01	Illumina	ncbi/GCA_016509755.1
Labor Ovalle	MA	2021	Phytozome	670	PacBio	phytozome/LaborOvalle_v1_1
5-593	MA	2021	Phytozome	696	PacBio	phytozome/5_593_v1_1
UI111	MA	2019	Phytozome	534	PacBio	phytozome/UI111_v1_1

3. Results

3.1. Raw sequencing data

Table 4.2 shows the standard metrics parameters obtained for Illumina and PacBio sequence data.

Table 4. 2. Illumina and PacBio sequence data standard metrics parameters.

	Nº reads	Nº bases (bp)	Mean read length (bp)	N50
CCS (PacBio)	4,459,685	81,154,094,113	18197	20657
HiFi (PacBio)	1,831,809	34,429,383,781	18795	18917
Illumina (R1 + R2)	403,708,348	60,556,252,200	150	n/a

The k-mer sequence distribution was evaluated, with k-mer set to 21 according to the KMC v3.1.1 program (Kokot *et al*. 2017). Visualization of the k-mer profile fit to a model of the expected fractions allowed an initial estimate of the genome size, confirmation of the ploidy level, and estimation of a low sequencing error rate. This analysis was performed on the data after quality filtering for Illumina (Figure 4.1a) and PacBio (Figure 4.1b) reads.

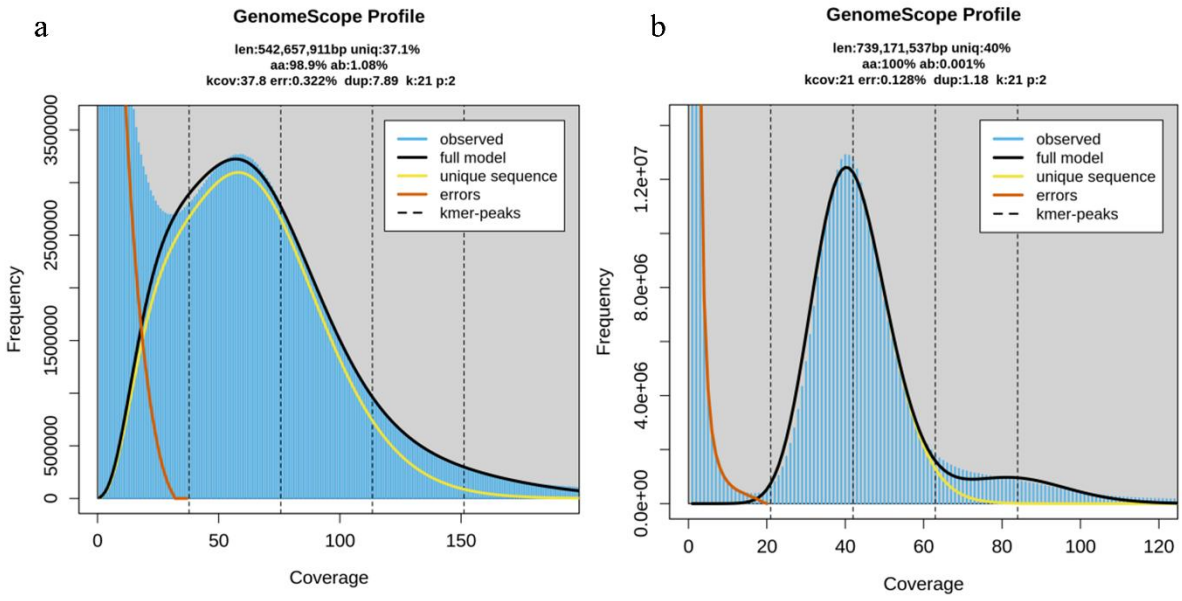

Figure 4. 1. GenomeScope k-mer profile and model fit plot for a) the Illumina data and b) the PacBio data based on a k-mer size of 21. The observed k-mer frequency distribution, depicted in blue, represents the number of times a given k-mer is observed in the sequencing data (coverage) and the total number of k-mers with a given coverage (frequency).

The genome coverage was calculated by dividing the number of base pairs sequenced by each platform (Table 4.2) by the resulting genome size, giving a coverage of 128.75x for Illumina reads, 172.5x for the CCs PacBio, and 73.2x after transforming to HiFi PacBio reads.

3.2. Genome Assembly

De novo assembly of the genome originated 470,324,335 bp with a GC content of 38.68% and 98.5% of completeness according to BUSCO (Seppey *et al*. 2019) and QUAST (Gurevich *et al*. 2013), organized in 1,368 scaffolds plus the mitochondrial and chloroplastic genomes (Table 4.3). The 1,368 scaffolds were organized in chromosomes according to the reference genome of the species (NCBI accession: GCF_000499845.1) resulting in 11 chromosomes (467,167,878 bp) and 129 scaffolds not aligned with the reference genome (2,622,527 bp). The mitochondrial genome was organized in two scaffolds of 371,437 bp and 11,183 bp. The chloroplastic genome results in a circular molecule of 151,310 bp.

The chromosomes were named with the prefix PvulA25 followed by Pv and the number corresponding to each one taking as reference the chromosomes of the G19833 v1 genome (Schmutz *et al*. 2014). Organelles were named with the prefix PvulA25 followed by Mt for mitochondrial and Cp for chloroplastic.

Table 4. 3. Data descriptor of the PvulA25 assembled genome.

Identifier	Molecule	Size (bp)	GC%	Transcripts
PvulA25_Pv01	Chromosome	46,263,949	34.65	5,917
PvulA25_Pv02	Chromosome	47,662,262	33.69	5,851
PvulA25_Pv03	Chromosome	46,158,017	33.09	5,265
PvulA25_Pv04	Chromosome	42,505,441	34.99	5,345
PvulA25_Pv05	Chromosome	35,400,105	35.27	4,698
PvulA25_Pv06	Chromosome	31,923,677	33.53	4,092
PvulA25_Pv07	Chromosome	43,922,236	34.43	5,523
PvulA25_Pv08	Chromosome	55,473,932	34.77	7,194
PvulA25_Pv09	Chromosome	37,738,617	32.21	4,053
PvulA25_Pv10	Chromosome	34,821,128	35.96	4,664
PvulA25_Pv11	Chromosome	45,298,514	35.63	5,919
Unplaced	Scaffolds	2,622,527	38.98	430
PvulA25_Mt1	Mitochondrial	371,437	45.05	76
PhvulA25_Mt2	Mitochondrial	11,183	45.06	5
PvulA25_Cp	Chloroplastic	151,310	34.95	394
PvulA25	Whole genome	470,324,335	38.68	59,426

3.3. Genome annotation

Annotation of the 1,368 *de novo* assembled scaffolds predicted 58,955 genomic protein-coding sequences, 58,951 of them were localized in the 11 chromosomes and unplaced scaffolds. Functional annotation of these protein-coding sequences can be consulted in Table 4S.1. The organelles were also annotated, comprising a total of 81 mitochondrial genes and 394 chloroplastic genes.

3.4. Genomes characteristics comparison

The A25 genome is the tenth common bean assembled genome available. From the first genome published in 2014, 2 genomes were sequenced with the Illumina platform, 5 with the PacBio technology, and one more combining the two platforms like the genome sequenced in this work (Table 4.1). Table 4.4 shows the characteristics of these different genomes and their significance variance in size and number of annotated genes. Genome size ranged from 449 Mb (JaloEEP558) to 615.2 Mb ('Flavert'), genome size varied also between the genomes of the same genotypes when a different technology was used in the sequencing. Also, the number of genes is different in all the genomes available, being

version 2 of G19833 the genotype with fewer genes, 27,433, and BAT93 the one with more genes, 30,491.

Table 4. 4. Characteristic of the available common bean sequenced genomes.

Genotype	Assembly level	Size	Number of genes/transcripts
G19833 v1	Chromosome	521.1 Mb	28,134
G19833 v2	Chromosome	537.2 Mb	27,433
Flavert	Chromosome	615.2 Mb	29,549
OAC Rex	Chromosome	423.7 Mb	Not annotated
JaloEEP558	Scaffold	449 Mb	Not annotated
BAT93	Chromosome	549.7 Mb	30,491
BAT93	Scaffold	452.7 Mb	Not annotated
Labor Ovalle	Chromosome	571.9 Mb	27,218
5-593	Chromosome	572.2 Mb	27,065
UI111	Chromosome	554.9 Mb	27,385
A25	Chromosome	470.3 Mb	59,426*

*Transcripts.

Table 4.5 shows a comparison of important genome features between A25 and G19833, which was used to guide the chromosome assembly level. All chromosomes except Pv09, were larger in the reference genome. GC content was stable between both genomes and the number of transcripts in A25 were in all the chromosomes much higher than the number of G19833 genes. Future transcripts filter works will be needed to accurate the A25 annotation.

Table 4. 5. Genome characteristics comparison between G19833 v1 and A25 genomes.

	G19833 v1			A25		
Chromosome	**Size**	**GC content**	**Genes**	**Size**	**GC content**	**Transcripts**
Pv01	52,205,531	35.0	2,779	46,263,949	34.7	5,917
Pv02	49,040,938	33.5	3,435	47,662,262	33.7	5,851
Pv03	52,284,309	34.0	3,058	46,158,017	33.1	5,265
Pv04	45,960,019	35.5	1,890	42,505,441	35.0	5,345
Pv05	40,819,286	35.5	1,928	35,400,105	35.3	4,698
Pv06	31,977,256	33.5	2,295	31,923,677	33.5	4,092
Pv07	51,758,522	35.5	2,895	43,922,236	34.4	5,523
Pv08	59,662,532	35.5	3,023	55,473,932	34.8	7,194
Pv09	37,469,608	32.0	2,719	37,738,617	32.2	4,053
Pv10	43,275,151	37.0	1,721	34,821,128	36.0	4,664
Pv11	50,367,376	36.5	2,253	45,298,514	35.6	5,919

4. Discussion

Common bean is a highly diverse species, with two main gene pools. In Europe, the common bean has traditionally been grown, and a wide diversity has also been reported which includes unique phenotypes in the species, such as market class fabada, ‘Garrafal Oro’, ‘Flavert’, ‘Cannellini’, ‘Borlotto’. Some authors have even suggested that Europe represents secondary diversity in species (Santalla *et al*. 2002). Except for the recently published ‘Flavert’ genome (Carrère *et al*. 2023), the available bean genomes were derived from American genotypes (Table 4.1). In addition, two of them are susceptible to photoperiod (G19833 and *Labor* Ovalle), which makes them difficult to use in Europe. All these reasons, led to *de novo* sequencing and annotation of line A25, a homozygous line into the market class fabada (cv. ‘Andecha’).

The variance in the sequence of the genome can determine differences in the organization of the genetic information, some genetic analyses could have different results depending on the reference genome used, for example, a GWAS or RNA-Seq (Castro *et al*. 2024). The observed genetics differences between the common bean genomes (Table 4.4), in size and number of genes, show the need to use a genome close to the working cultivar as well as to sequence a greater volume of genomes to expand knowledge of the consensus and variable regions within the species. Also, the sequences of five assembled genomes in *P. vulgaris* were compared. They showed high variation among them, finding 2.4 Mbp of extra sequence with 7000 annotated genes not present

in this first sequenced genome (Cortinovis *et al*. 2023). However, due to the high differences in the number of genes in the A25 genome compared to the other available genomes, the 59,426 predicted transcripts are currently in the process of verification and validation.

The plant community is committed to sequencing more individuals from the same species because intraspecific diversity in plants can be very high. Therefore, many efforts are focusing on the development of pan-genomes and super-pangenomes that represent the collection of the entire diversity of DNA sequences in a species/genus, and are expected to be more complete and accurate than the use of a single reference genome (Shi *et al*. 2023). For this reason, and taking advantage of new sequencing technologies, which can reveal differences even between genomes of the same genotype, the sequencing of new genomes should be encouraged to expand knowledge of the diversity of the species.

5. Conclusion

A new common bean genome will be published and will be a useful tool for future research in common bean breeding, especially for seed phenotype and evolutionary traits and also add new information about the species variation at a genomic level and contribute to the establishment of the *P. vulgaris* pangenome.

GENERAL DISCUSSION

The objective of this Thesis was to use and develop new tools taking advantage of sequencing and phenotyping techniques and developing new genomic resources. These tools were applied to study the common bean diversity for more efficient germplasm conservation and to increase the genetic knowledge of important traits in breeding.

1. Genetic diversity and germplasm management

Genetic diversity is essential for breeders because it provides genes or gene combinations to develop new cultivars that meet consumer and environmental demands. Part of this genetic diversity is conserved in large germplasm collections. Although there is an awareness of conserving genetic diversity, breeders may lack information to determine which accessions would be most beneficial for their breeding objectives (Pathirana and Carimi 2022). Germplasm characterization through phenotyping and genotyping helps to detect duplications in collections, simplify large collections, and identify the best genotypes to be used in breeding programs.

HTG was used in Chapter 1 to characterize the panel established for the Fabada market class (FP). The problem in this study was the characterization of diversity within a homogeneous phenotypic group. This FP had 100 accessions maintained in seed collections for 30 years, 57 accessions collected from farmers in 2021, six cultivars developed in SERIDA, and 16 reference cultivarsSurprisingly, genotyping reveals high diversity for a phenotypically homogeneous group. These results allowed us to identify redundant lines in the SERIDA collection (synonyms, duplications) and some homonyms (identical names assigned to different phenotypes in the passport data). The results lead to more efficient preservation of plant genetic resources in the SERIDA seed collection, prioritizing the conservation and use of certain lines that represent the diversity of this group. This genotyping also revealed genetic erosion when the diversity maintained in SERIDA collection was compared with the diversity collected in 2021. This result confirmed the importance of preserving germplasm in collections because we found important genetic erosion within this market class. The population preserved for 30 years in the SERIDA collections was more diverse than the currently cultivated population. In addition, since the phenotype is a bottleneck in diversity studies owing to the time-

consuming and required resources, a first approach based on HTG resulted in the simplification of the collection and streamlining of the breeding process.

Characterization can always be expanded or improved, currently, the characterization of FP is being completed. A subset of FP lines selected from genotyping results is being phenotyped in the field. Although phenotypic variation is limited, the phenotyping can lead to establishing differentiated lines into the market class fabada, and perform future association analysis to identify the genomic regions involved in this differentiation. The question is whether the observed genotypic variation corresponds to phenotypic variation.

2. Genetic architecture of important traits

The characteristics that can be enhanced are those that are genetically determined. The availability of annotated genomes allows to go further connecting phenotypes, genotypes, and genomes to identify candidate genes, explore different types of variation in nucleotide sequences, and develop specific markers for genes or genomic regions. When a close linkage between a marker and a trait of interest is established (linked marker), the marker can be used for MAS (Singh and Singh 2015). MAS allows an increase in the efficiency of plant breeding by accelerating the identification of the best phenotypes (Ibitoye and Akin-Idowu 2010). Similarly, a marker developed from the nucleotide sequence of a candidate gene controlling a specific trait (functional marker) can be used for indirect selection (Anders *et al*. 2021; Varshney *et al*. 2021a). For this reason, the expansion of the knowledge on major genes and QTL and the involved candidate genes offers an opportunity to improve/accelerate the breeding process. The establishment of the genetic base for the traits is useful for this and other species. In Chapters 2 and 3, different analyses were conducted to identify the genomic regions and candidate genes that control important traits. A combination of approaches was used in each study to verify and validate the identified genomic regions. The methodologies included, QTL mapping, differential expression by RNA-seq, GWAS, co-location with reported QTL, and genomic exploration of bean genomes.

Chapter 2 explores the region and candidate genes controlling resistance to an important disease, common bean anthracnose. The disease is caused by the fungus *C. lindemuthianum*, which causes large crop yield losses. The Co-2 cluster, located at the

end of chromosome Pv11 (46.01-47.77 Mb) in the NIL used, is responsible for resistance to race 38 of anthracnose in the genotype Cornell49242. RNA-Seq results revealed six potential candidate genes overexpressed in response to the pathogen in the Co-2 cluster, which is made up of more than 80 LRR domain genes.. The functions of these genes, LRR domain and serine/threonine protein kinase, were also reported before as responsible for disease resistance (Dodds and Rathjen 2010; Richard *et al*. 2021). The Co-2 cluster already has a linked molecular marker that requires cutting with restriction enzymes (Geffroy *et al*. 1998), but the mapping of the RNA raw reads and sequence comparison between four available genomes of *P. vulgaris* led to the design of new molecular markers linked to the proposed candidate genes. Unlike the previously described marker, the markers designed in this chapter are found within the cluster and most of them are codomain and functional markers. These markers have already been incorporated into the SERIDA plant breeding program, and their use in the development of new genotypes using MAS has saved time in the laboratory. Interestingly, transcriptomic analysis revealed several overexpressed read sequences located in the delimitated region (46.98-47.04 Mb in Pv11) that do not correspond to genes annotated in the G19833 genome. It is evidence of the variation in this telomeric region. Results also open the possibility to investigate other bean genotype – fungus genotype interactions in order to provide data in this specific interaction.

The seed phenotype is a relevant trait in common beans. Most of the cultivation area for this species is dry beans (FAO 2022). Bean seeds exhibit extensive phenotypic variation (see https://zenodo.org/records/10263706), which can be described by considering a combination of seed shape, seed size, seed coat color, and color distribution. Chapter 3 focused on the identification of regions associated with seed phenotypes, namely seed morphological traits such as seed shape and size through GWAS in SDP. Nine regions were associated with these characteristics and co-located with previous QTL studies (González *et al*. 2016; Murube *et al*. 2020; Bassett *et al*. 2021; Giordani *et al*. 2022; Ugwuanyi *et al*. 2022; Arriagada *et al*. 2022). These QTL had 70 differentially expressed genes during seed development, which were proposed as potential candidate genes in the control of traits. Some of them, such as *PHAVU_003G249200g*, were differentially expressed in the development of the cultivar 'Xana' and in the expression atlas of common bean (O'Rourke *et al*. 2014), indicating their importance in the development of the seed and in the final phenotype related to the size. The effect SNPs

tagging 5 consensuses QTL associated with seed weight was explored by HPCP analysis showing significant effects on seed weight in all cases and allowing the classification of SDP lines by origin gene pools as well, supporting the hypothesis that seed size plays a crucial role in the domestication of the species, for which 115 domestication genes (Schmutz *et al*. 2014) located under the total number of QTL identified in this chapter were labeled as candidate genes.

Concerning seed quality traits, Chapter 3 examines the genes/QTLs responsible for water absorption capacity and seed coat proportion. Water absorption and coat proportion, traits related to the protein content and seed cooking time. Shorter cooking times are more convenient for consumers, especially those with limited time for food preparation and energy consumption (Bassett *et al*. 2021). The QTL Seed01_50.7 was associated with coat proportion in the SDP and protein content in a diversity panel study by Ugwuanyi *et al*. (2022) and two candidate genes were identified. The chromosome Pv05 was interesting for the two regions associated with water absorption of the seed; the one located at the beginning of the chromosome was a consensus between our study and that reported by Berry *et al*. (2020). Chromosome Pv10 may be interesting as well for the water absorption of the seed; a region ranging from 32.54 to 32.68 Mbp, was detected in this study and through QTL mapping by Bassett *et al*. (2021).

All common bean chromosomes except Pv06 were associated with at least one of the traits studied, where the same character is associated with more than one region, highlighting the difficulty in the breeding of the studied characters. These QTL could be used in plant breeding, the most important for this purpose being those with the highest SNP effect and previously reported in the literature. Kompetitive Allele Specific PCR (KASP), is an allele-specific oligo extension-based PCR assay that uses fluorescence to detect genetic variations (He *et al*. 2014a; Dipta *et al*. 2024). KASP markers could be the solution to identify these major QTN in MAS to increase or reduce the value of the quantitative trait following the breeding requires.

3. Increase knowledge of common bean genome

Plant reference genomes are essential for advancing plant biological research and offer numerous benefits for plant breeding. Plant genomes help identify the location and

variations of specific DNA sequences, enabling the development of new marker-assisted breeding, understanding plant evolution, taxonomy, and gene regulation, and knowing the network of genes controlling important plant traits (Marks *et al*. 2021; Hong *et al*. 2023). In forward genetic analysis, the availability of genomes accelerated the identification of sequences controlling the genotypes for a specific trait. This led to the proposal of numerous candidate genes in the published plant genomes from linkage analysis, QTL mapping, and association studies (GWAS). Also, the availability of genomes offers a new perspective in transcriptomic analysis, allowing the identification of the network of differentially expressed genes throughout the genome (RNA-seq; (Wang *et al*. 2009)) in specific tissues or stresses. However, the publication of different genomes within a species revealed the variations in single nucleotide variants, and structural variants as duplications, InDels, CVN, inversions, and translocations. This variation among genomes represents a limitation in the genetic analysis when a single reference genome is considered. Ideally, the availability of the genome of the genotypes involved in our study would be the solution.

Chapters 2 and 3 confirm differences in sequence and number of annotated genes in delimited regions. Moreover, both studies showed variation in the sequence of candidate genes between five common bean genomes, as base pair change and InDels. This underlines the importance of having a genome close to your working materials, thus avoiding the possible errors derived from using any reference genome as a mirror in sequencing work, such as the loss of identification of cultivar-specific genes in an RNA-Seq.

Chapter 4 presents a genome draft for the fabada market class, and it will represent an advance in the breeding program of this market class, allowing a more precise forward genetic analysis and plant breeding in those studies in which this genome is involved. The availability of this new genome allows for the exploration of the differences between the Fabada market class and those already sequenced, especially traits associated with the morphology of the seed, which is responsible for the well-differentiated phenotype of this variety. Future works can be focused in the exploration of the regions associated with seed morphology reported in Chapter 3 in the genome of A25, and the comparison with the other genomes to detect variations involved in the seed size control. The assembly

and annotation of a new genome in this Thesis also offer the opportunity to extend the databases used in future studies of this type and move toward the development of the pangenome of the species.

4. Essential issues

The results obtained in this Thesis provided useful information for current common bean programs. On the one hand, the results help in the efficient management and use of the germplasm collections and show a way for it (Chapter 1). These results are already being implemented in the management of the SERIDA collection, and it is expected that they will give rise to a set of new lines within the Fabada market class.

On the other hand, the results of the Thesis expand the knowledge of genomic regions and candidate genes associated with the genetic control of important traits like resistance to anthracnose (Chapter 2) and seed phenotype (Chapter 3). These studies provides a new information on the gene network that controls these traits and allows the development of markers for assisted selection, several of which are being used in current breeding programs.

Finally, this Thesis provides an assembled and annotated new bean genome, which will be a great support for future genetic studies and will contribute to the construction of the pangenome of the species.

CONCLUSIONS/CONCLUSIONES

Conclusions

1. A detailed characterization of common bean accessions by high-throughput genotyping allows the identification of redundant lines (duplications), and misclassified accessions (homonyms) and increases the efficiency of preserving and using genetic diversity.

2. High-throughput genotyping of the fabada panel revealed both genotypic variation within this market class and genetic erosion over the last 30 years.

3. The genotypic variation within the market class Fabada supports the usefulness of *ex situ* germplasm collection.

4. The combination of physical mapping in the bean reference genome and differential expression analysis in NILs enabled the identification of candidate genes for the *Co-2* resistance gene, which controls the response to the race 38 of bean anthracnose by the fungus *Colletotrichum lindemuthianum.*

5. Genomic regions delimited by association analysis for seed size traits and differential expression analysis during seed development allows the identification of candidate genes involved in controlling seed size.

6. A combination of different approaches such as GWAS, QTL mapping, genomic mapping, differential expression analysis, and functional analysis of annotated genes, are ways to validate the phenotype-genotype-genome connection delimiting the genomic regions and approximating to annotated genes that control the studied traits.

7. Mapping of RNA transcripts in differential expression analysis and genomic comparisons among bean genomes revealed high variability within the *P vulgaris* L. genotypes and emphasized the importance of having a reference genome closely related to the field of study, as well as a pangenome in the species.

8. De novo sequencing of the A25 genotype, its assembly, and functional annotation led to the development of a new bean genome, which constitutes a tool for future studies in this species, and particularly in the market class fabada.

9. The availability of diverse and well-characterized plant material is essential for genetic analysis and plant breeding programs. This Thesis provides both well-characterized materials and knowledge for new genetic studies and breeding programs.

Conclusiones

1.Una caracterización detallada de las colecciones de judías mediante genotipado de alto rendimiento permite identificar las líneas redundantes y las accesiones mal clasificadas, además de aumentar la eficacia de la conservación y el uso de la diversidad genética.

2.El genotipado de alto rendimiento del panel de diversidad de fabada reveló tanto la variación genotípica dentro de esta clase de mercado como la erosión genética a lo largo de los últimos 30 años.

3.La variación genotípica observada en la clase comercial Fabada apoya apoyan la utilidad de la conservación de germoplasma *ex situ*.

4.La combinación de mapeo físico en genoma de judía y análisis de expresión diferencial en NILs permitió la identificación de genes candidatos para el cluster de resistencia Co-2, que controla la respuesta a la raza 38 de antracnosis en la línea A4804.

5.El análisis de asociación para rasgos de tamaño de semilla y el análisis de expresión diferencial durante el desarrollo de la semilla permiten la identificación de genes candidatos implicados en el control de los fenotipos de semilla.

6.Una combinación de diferentes enfoques como GWAS, mapeo QTL, mapeo genómico, análisis de expresión diferencial y análisis funcional de genes anotados, son formas de validar la conexión fenotipo-genotipo-genoma delimitando las regiones genómicas y aproximándose a genes anotados que controlan los caracteres estudiados.

7.El mapeo de transcritos de ARN en el análisis de expresión diferencial y las comparaciones genómicas entre genomas de frijol revelaron una alta variabilidad dentro de los genotipos de *P vulgaris* L., y enfatizaron la importancia de tener un genoma de referencia estrechamente relacionado con el campo de estudio, así como un pangenoma en la especie.

8.La secuenciación de novo del genotipo A25, su ensamblaje y anotación funcional condujeron al desarrollo de un nuevo genoma de la judía, que constituye una herramienta para futuros estudios en esta especie.

9.La disponibilidad de material vegetal diverso y bien caracterizado es esencial para el análisis genético y los programas de mejora. Esta Tesis proporcionar materiales bien caracterizados para nuevos estudios genéticos y programas de mejora vegetal.

REFERENCES

Abril JF, Castellano S (2019) Genome Annotation. In: Encyclopedia of bioinformatics and computational biology. Elsevier, pp 195–209

Adam-Blondon AF, Sévignac M, Bannerot H, Dron M (1994) SCAR, RAPD and RFLP markers linked to a dominant gene (Are) conferring resistance to anthracnose in common bean. Theor Appl Genet 88:865–870. https://doi.org/10.1007/BF01253998

Afzal AJ, Wood AJ, Lightfoot DA (2008) Plant receptor-like serine threonine kinases: roles in signaling and plant defense. Mol Plant Microbe Interact 21:507–517. https://doi.org/10.1094/MPMI-21-5-0507

Afzal M, Alghamdi SS, Migdadi HH, *et al* (2020) Legume genomics and transcriptomics: from classic breeding to modern technologies. Saudi J Biol Sci 27:543–555. https://doi.org/10.1016/j.sjbs.2019.11.018

Ahn H-K, Kang YW, Lim HM, *et al* (2015) Physiological functions of the COPI complex in higher plants. Mol Cells 38:866–875. https://doi.org/10.14348/molcells.2015.0115

Alonge M, Lebeigle L, Kirsche M, *et al* (2022) Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. Genome Biol 23:258. https://doi.org/10.1186/s13059-022-02823-7

Alqudah AM, Sallam A, Stephen Baenziger P, Börner A (2020) GWAS: Fast-forwarding gene identification and characterization in temperate Cereals: lessons from Barley - A review. J Advanc Res 22:119–135. https://doi.org/10.1016/j.jare.2019.10.013

Alvarado-Gutiérrez A, Del Real-Monroy M, Rodríguez-Guerra R, *et al* (2008) A *Phaseolus vulgaris* EF-hand calcium-binding domain is induced early in the defense response against *Colletotrichum lindemuthianum* and by abiotic stress: Sequences shared between interacting partners. Physiological and Molecular Plant Pathology 72:111–121. https://doi.org/10.1016/j.pmpp.2008.04.005

Amarasinghe SL, Su S, Dong X, *et al* (2020) Opportunities and challenges in long-read sequencing data analysis. Genome Biol 21:30. https://doi.org/10.1186/s13059-020-1935-5

Amongi W, Nkalubo ST, Ochwo-Ssemakula M, *et al* (2023) Phenotype based clustering, and diversity of common bean genotypes in seed iron concentration and cooking time. PLoS ONE 18:e0284976. https://doi.org/10.1371/journal.pone.0284976

Anders S, Cowling W, Pareek A, *et al* (2021) Gaining acceptance of novel plant breeding technologies. Trends Plant Sci 26:575–587. https://doi.org/10.1016/j.tplants.2021.03.004

Andrews S (2010) . FastQC: a quality control tool for high throughput sequence data. Babraham Institute

Angioi SA, Rau D, Attene G, *et al* (2010) Beans in Europe: origin and structure of the European landraces of *Phaseolus vulgaris* L. Theor Appl Genet 121:829–843. https://doi.org/10.1007/s00122-010-1353-2

Arriagada O, Arévalo B, Cabeza RA, *et al* (2022) Meta-QTL Analysis for Yield Components in Common Bean (*Phaseolus vulgaris* L.). Plants 12:. https://doi.org/10.3390/plants12010117

Asfaw A, Blair MW, Almekinders C (2009) Genetic diversity and population structure of common bean (*Phaseolus vulgaris* L.) landraces from the East African highlands. Theor Appl Genet 120:1–12. https://doi.org/10.1007/s00122-009-1154-7

Aylesworth JW, Tu JC, Buzzell RI (1983) Sanilac BC6-Are white bean breeding line. Hortscience 18:115:

Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263–265. https://doi.org/10.1093/bioinformatics/bth457

Bassett A, Katuuramu DN, Song Q, Cichy K (2021) QTL Mapping of Seed Quality Traits Including Cooking Time, Flavor, and Texture in a Yellow Dry Bean (*Phaseolus vulgaris* L.) Population. Front Plant Sci 12:670284. https://doi.org/10.3389/fpls.2021.670284

Bayer PE, Golicz AA, Scheben A, *et al* (2020) Plant pan-genomes are the new reference. Nat Plants 6:914–920. https://doi.org/10.1038/s41477-020-0733-0

Bellucci E, Benazzo A, Xu C, *et al* (2023) Selection and adaptive introgression guided the complex evolutionary history of the European common bean. Nat Commun 14:1908. https://doi.org/10.1038/s41467-023-37332-z

Bellucci E, Bitocchi E, Rau D, *et al* (2014) Genomics of Origin, Domestication and Evolution of *Phaseolus vulgaris*. In: Tuberosa R, Graner A, Frison E (eds) Genomics of plant genetic resources. Springer Netherlands, Dordrecht, pp 483–507

Berry M, Izquierdo P, Jeffery H, *et al* (2020) QTL analysis of cooking time and quality traits in dry bean (*Phaseolus vulgaris* L.). Theor Appl Genet 133:2291–2305. https://doi.org/10.1007/s00122-020-03598-w

Bitocchi E, Bellucci E, Giardini A, *et al* (2013) Molecular analysis of the parallel domestication of the common bean (*Phaseolus vulgaris*) in Mesoamerica and the Andes. New Phytol 197:300–313. https://doi.org/10.1111/j.1469-8137.2012.04377.x

Bitocchi E, Nanni L, Bellucci E, *et al* (2012) Mesoamerican origin of the common bean (*Phaseolus vulgaris* L.) is revealed by sequence data. Proc Natl Acad Sci USA 109:E788-96. https://doi.org/10.1073/pnas.1108973109

Blair MW, Díaz LM, Buendía HF, Duque MC (2009) Genetic diversity, seed size associations and population structure of a core collection of common beans (*Phaseolus vulgaris* L.). Theor Appl Genet 119:955–972. https://doi.org/10.1007/s00122-009-1064-8

Blair MW, Izquierdo P, Astudillo C, Grusak MA (2013) A legume biofortification quandary: variability and genetic control of seed coat micronutrient accumulation in common beans. Front Plant Sci 4:275. https://doi.org/10.3389/fpls.2013.00275

Blair MW, Izquierdo P (2012) Use of the advanced backcross-QTL method to transfer seed mineral accumulation nutrition traits from wild to Andean cultivated common beans. Theor Appl Genet 125:1015–1031. https://doi.org/10.1007/s00122-012-1891-x

Bradbury PJ, Zhang Z, Kroon DE, *et al* (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635. https://doi.org/10.1093/bioinformatics/btm308

Brionne A, Juanchich A, Hennequet-Antier C (2019) ViSEAGO: a Bioconductor package for clustering biological functions using Gene Ontology and semantic similarity. BioData Min 12:16. https://doi.org/10.1186/s13040-019-0204-1

Brown AHD, Hodgkin T (2015) Indicators of genetic diversity, genetic erosion, and genetic vulnerability for plant genetic resources. In: Ahuja MR, Jain SM (eds) Genetic diversity and erosion in plants. Springer International Publishing, Cham, pp 25–53

Brown AHD (1989) Core collections: a practical approach to genetic resources management. Genome 31:818–824. https://doi.org/10.1139/g89-144

Burridge J, Jochua CN, Bucksch A, Lynch JP (2016) Legume shovelomics: High—Throughput phenotyping of common bean (*Phaseolus vulgaris* L.) and cowpea (*Vigna unguiculata* subsp, *unguiculata*) root architecture in the field. Field Crops Res 192:21–32. https://doi.org/10.1016/j.fcr.2016.04.008

Campa A, Geffroy V, Bitocchi E, *et al* (2024) Screening for resistance to four fungal diseases and associated genomic regions in a snap bean diversity panel. Front Plant Sci 15:1386877. https://doi.org/10.3389/fpls.2024.1386877

Campa A, Murube E, Ferreira JJ (2018) Genetic Diversity, Population Structure, and Linkage Disequilibrium in a Spanish Common Bean Diversity Panel Revealed through Genotyping-by-Sequencing. Genes 9:. https://doi.org/10.3390/genes9110518

Campa A, Rodríguez-Suárez C, Giraldez R, Ferreira JJ (2014) Genetic analysis of the response to eleven *Colletotrichum lindemuthianum* races in a RIL population of common bean (*Phaseolus vulgaris* L.). BMC Plant Biol 14:115. https://doi.org/10.1186/1471-2229-14-115

Campa A, Trabanco N, Ferreira JJ (2017) Identification of Clusters that Condition Resistance to Anthracnose in the Common Bean Differential Cultivars AB136 and

MDRK. Phytopathology 107:1515–1521. https://doi.org/10.1094/PHYTO-01-17-0012-R

Carøe C, Bohmann K (2020) Tagsteady: A metabarcoding library preparation protocol to avoid false assignment of sequences to samples. Mol Ecol Resour 20:1620–1631. https://doi.org/10.1111/1755-0998.13227

Carrère S, Mayjonade B, Lalanne D, *et al* (2023) First whole genome assembly and annotation of a European common bean cultivar using PacBio HiFi and Iso-Seq data. Data Brief 48:109182. https://doi.org/10.1016/j.dib.2023.109182

Castellanos JZ, Guzmán-Maldonado H, Acosta-Gallegos JA, Kelly JD (1995) Effects of hardshell character on cooking time of common beans grown in the semiarid highlands of Mexico. J Sci Food Agric 69:437–443. https://doi.org/10.1002/jsfa.2740690406

Castro-Guerrero NA, Isidra-Arellano MC, Mendoza-Cozatl DG, Valdés-López O (2016) Common bean: A legume model on the rise for unraveling responses and adaptations to iron, zinc, and phosphate deficiencies. Front Plant Sci 7:600. https://doi.org/10.3389/fpls.2016.00600

Castro P, Carmona A, Perez-Rial A, *et al* (2024) Finding consensus on the reference genomes : A chickpea case study. Legume Science 6:. https://doi.org/10.1002/leg3.224

Chacón S MI, Pickersgill B, Debouck DG (2005) Domestication patterns in common bean (*Phaseolus vulgaris* L.) and the origin of the Mesoamerican and Andean cultivated races. Theor Appl Genet 110:432–444. https://doi.org/10.1007/s00122-004-1842-2

Chamkhi I, Cheto S, Geistlinger J, *et al* (2022) Legume-based intercropping systems promote beneficial rhizobacterial community and crop yield under stressing conditions. Industrial Crops and Products 183:114958. https://doi.org/10.1016/j.indcrop.2022.114958

Chen G, Shi T, Shi L (2017) Characterizing and annotating the genome using RNA-seq data. Sci China Life Sci 60:116–125. https://doi.org/10.1007/s11427-015-0349-4

Cichy KA, Porch TG, Beaver JS, *et al* (2015a) A diversity panel for andean bean improvement. Crop Sci 55:2149. https://doi.org/10.2135/cropsci2014.09.0653

Cichy KA, Wiesinger JA, Mendoza FA (2015b) Genetic diversity and genome-wide association analysis of cooking time in dry bean (*Phaseolus vulgaris* L.). Theor Appl Genet 128:1555–1567. https://doi.org/10.1007/s00122-015-2531-z

Claros MG, Bautista R, Guerrero-Fernández D, *et al* (2012) Why assembling plant genome sequences is so challenging. Biology (Basel) 1:439–459. https://doi.org/10.3390/biology1020439

Cortinovis G, Vincenzi L, Anderson R, *et al* (2023) Adaptive gene loss in the common bean pan-genome during range expansion and domestication. BioRxiv. https://doi.org/10.1101/2023.11.23.568464

Costa MP, Reckling M, Chadwick D, *et al* (2021) Legume-Modified Rotations Deliver Nutrition With Lower Environmental Impact. Front Sustain Food Syst 5:. https://doi.org/10.3389/fsufs.2021.656005

Da Silva C, Zamperin G, Ferrarini A, *et al* (2013) The high polyphenol content of grapevine cultivar tannat berries is conferred primarily by genes that are not shared with the reference genome. Plant Cell 25:4777–4788. https://doi.org/10.1105/tpc.113.118810

De la Rosa L, Lázaro A, Varela F (2000) Racionalización de la colección española de *Phaseolus vulgaris* L. Actas de Asoaciación Española de Leguminosas, Villaviciosa, Asturias, pp 55–62

de León LF, Bressani R, Elías LG (1989) Effect of the seed coat on the hardening of common beans (*Phaseolus vulgaris*). Arch Latinoam Nutr 39:405–418

Debouck DG, Hidalgo R (1985a) Morfología de la planta de frijol común. In: Centro Internacional de Agricultura Tropical (ed) Frijol: Investigación y producción. Cali, Colombia, pp 7–43

Debouck DG, Hidalgo R (1985b) Etapas de desarrollo de la planta de frijol. In: Centro Internacional de Agricultura Tropical (ed) Frijol: Investigación y producción. Cali, Colombia, pp 61–81

de Jong MJ, de Jong JF, Hoelzel AR, Janke A (2021) SambaR: An R package for fast, easy and reproducible population-genetic analyses of biallelic SNP data sets. Mol Ecol Resour 21:1369–1379. https://doi.org/10.1111/1755-0998.13339

de Souza IP, de Azevedo BR, Coelho ASG, *et al* (2023) Whole-genome resequencing of common bean elite breeding lines. Sci Rep 13:12721. https://doi.org/10.1038/s41598-023-39399-6

Diaz S, Ariza-Suarez D, Ramdeen R, *et al* (2020) Genetic Architecture and Genomic Prediction of Cooking Time in Common Bean (*Phaseolus vulgaris* L.). Front Plant Sci 11:622213. https://doi.org/10.3389/fpls.2020.622213

Dierckxsens N, Mardulyn P, Smits G (2017) NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res 45:e18. https://doi.org/10.1093/nar/gkw955

Dipta B, Sood S, Mangal V, *et al* (2024) KASP: a high-throughput genotyping system and its applications in major crop plants for biotic and abiotic stress tolerance. Mol Biol Rep 51:508. https://doi.org/10.1007/s11033-024-09455-z

Dodds PN, Rathjen JP (2010) Plant immunity: towards an integrated view of plant-pathogen interactions. Nat Rev Genet 11:539–548. https://doi.org/10.1038/nrg2812

Domoney C, Duc G, Ellis THN, *et al* (2006) Genetic and genomic analysis of legume flowers and seeds. Curr Opin Plant Biol 9:133–141. https://doi.org/10.1016/j.pbi.2006.01.014

Driguez P, Bougouffa S, Carty K, *et al* (2021) LeafGo: Leaf to Genome, a quick workflow to produce high-quality *de novo* plant genomes using long-read sequencing technology. Genome Biol 22:256. https://doi.org/10.1186/s13059-021-02475-z

Ekblom R, Wolf JBW (2014) A field guide to whole-genome sequencing, assembly and annotation. Evol Appl 7:1026–1042. https://doi.org/10.1111/eva.12178

Elshire RJ, Glaubitz JC, Sun Q, *et al* (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6:e19379. https://doi.org/10.1371/journal.pone.0019379

Escribano MR, Santalla M, de Ron AM (1997) Genetic diversity in pod and seed quality traits of common bean populations from northwestern Spain. Euphytica 93:71–81

FAO (2024) Pulses and soils: a dynamic duo | FAO Stories | Food and Agriculture Organization of the United Nations. https://www.fao.org/fao-stories/article/en/c/1676528/. Accessed 10 Apr 2024

FAO (2020) Statistical database. Food and agriculture organization of the United Nations. https://www.fao.org/faostat/en/#home. Accessed 13 Mar 2024

FAO (2022) Statistical Database. Food and Agriculture Organization of the United Nations. https://www.fao.org/faostat/en/#home. Accessed 2 Nov 2022

Ferreira JJ, Campa A, Kelly JD (2013) Organization of genes conferring resistance to anthracnose in common bean. In: Varshney RK, Tuberosa R (eds) Translational genomics for crop breeding: biotic stress. John Wiley & Sons Ltd, Chichester, UK, pp 151–181

Ferreira JJ, Campa A, Pérez-Vega E, Giraldez R (2008) Reaction of a Bean Germplasm Collection Against Five Races of *Colletotrichum lindemuthianum* Identified in Northern Spain and Implications for Breeding. Plant Dis 92:705–708. https://doi.org/10.1094/PDIS-92-5-0705

Ferreira JJ, Campa A, Pérez-Vega E (2005) Conservación y utilización de variedades tradicionales de faba en Asturias: colección activa de judías del Principado de Asturias

Ferreira JJ, Campa A, Pérez-Vega E, *et al* (2012) Introgression and pyramiding into common bean market class fabada of genes conferring resistance to anthracnose and potyvirus. Theor Appl Genet 124:777–788. https://doi.org/10.1007/s00122-011-1746-x

Ferreira JJ, Murube E, Campa A (2017) Introgressed Genomic Regions in a Set of Near-Isogenic Lines of Common Bean Revealed by Genotyping-by-Sequencing. Plant Genome 10:. https://doi.org/10.3835/plantgenome2016.08.0081

Flor HH (1971) Current status of the gene-for-gene concept. Annu Rev Phytopathol 9:275–296. https://doi.org/10.1146/annurev.py.09.090171.001423

Foxman B (2012) A primer of molecular biology. In: Molecular tools and infectious disease epidemiology. Elsevier, pp 53–78

Frankel OH (1984) Genetic Perspectives of Germplasm Conservation. In: Arber WK, Llimensee K, Peacock WJ, Stralinger P (eds) Genetic Manipulation: Impact on Man and Society. Cambridge University Press, Cambridge, pp 161–170

Freytag GF, Debouck DG (2002) Taxonomy, distribution, and ecology of the genus phaseolus (Leguminosae-papilionoideae) in North America, Mexico and Central America. Forth Worth, TX, USA

Fu Y-B (2015) Understanding crop genetic diversity under modern plant breeding. Theor Appl Genet 128:2131–2142. https://doi.org/10.1007/s00122-015-2585-y

García-Fernández C, Jurado M, Campa A, *et al* (2022) A core set of snap bean genotypes established by phenotyping a large panel collected in europe. Plants 11:. https://doi.org/10.3390/plants11050577

Geffroy V, Creusot F, Falquet J, *et al* (1998) A family of *LRR* sequences in the vicinity of the *Co-2* locus for anthracnose resistance in *Phaseolus vulgaris* and its potential use in marker-assisted selection. Theor Appl Genet 96:494–502. https://doi.org/10.1007/s001220050766

Gepts P, Bliss FA (1986) Phaseolin Variability among Wild and Cultivated Common Beans (*Phaseolus vulgaris*) from Colombia. Econ Bot 40:469–478

Gepts P, Bliss FA (1988) Dissemination pathways of common bean (*Phaseolus vulgaris*, Fabaceae) deduced from phaseolin electrophoretic variability. II. Europe and Africa. Econ Bot 42:86–104. https://doi.org/10.1007/BF02859038

Gepts P, Osborn TC, Rashka K, Bliss FA (1986) Phaseolin-protein Variability in Wild Forms and Landraces of the Common Bean (*Phaseolus vulgaris*): Evidence for Multiple Centers of Domestication. Econ Bot 40:451–468. https://doi.org/10.1007/BF02859659

Geraldo R, Santos CS, Pinto E, Vasconcelos MW (2022) Widening the Perspectives for Legume Consumption: The Case of Bioactive Non-nutrients. Front Plant Sci 13:772054. https://doi.org/10.3389/fpls.2022.772054

Giordani W, Gama HC, Chiorato AF, *et al* (2022) Genome-wide association studies dissect the genetic architecture of seed shape and size in common bean. G3 (Bethesda) 12:. https://doi.org/10.1093/g3journal/jkac048

González AM, Yuste-Lisbona FJ, Saburido S, *et al* (2016) Major contribution of flowering time and vegetative growth to plant production in common bean as deduced from a comparative genetic mapping. Front Plant Sci 7:1940. https://doi.org/10.3389/fpls.2016.01940

Grdeń P, Jakubczyk A (2023) Health benefits of legume seeds. J Sci Food Agric 103:5213–5220. https://doi.org/10.1002/jsfa.12585

Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Gururani MA, Venkatesh J, Upadhyaya CP, *et al* (2012) Plant disease resistance genes: Current status and future directions. Physiological and Molecular Plant Pathology 78:51–65. https://doi.org/10.1016/j.pmpp.2012.01.002

Gu R, Fan S, Wei S, *et al* (2023) Developments on core collections of plant genetic resources: do we know enough? Forests 14:926. https://doi.org/10.3390/f14050926

Hayat I, Ahmad A, Masud T, *et al* (2014) Nutritional and health perspectives of beans (*Phaseolus vulgaris* L.): an overview. Crit Rev Food Sci Nutr 54:580–592. https://doi.org/10.1080/10408398.2011.596639

He C, Holme J, Anthony J (2014a) SNP genotyping: the KASP assay. Methods Mol Biol 1145:75–86. https://doi.org/10.1007/978-1-4939-0446-4_7

He J, Zhao X, Laroche A, *et al* (2014b) Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Front Plant Sci 5:484. https://doi.org/10.3389/fpls.2014.00484

Hong K, Radian Y, Manda T, *et al* (2023) The development of plant genome sequencing technology and its conservation and application in endangered gymnosperms. Plants 12:. https://doi.org/10.3390/plants12234006

Hrdlickova R, Toloue M, Tian B (2017) RNA-Seq methods for transcriptome analysis. Wiley Interdiscip Rev RNA 8:e1364. https://doi.org/10.1002/wrna.1364

Ibitoye DO, Akin-Idowu PE (2010) Marker-assisted-selection (MAS): A fast track to increase genetic gain in horticultural crop breeding. African Journal of Biotechnology 9:8889–8895

Javornik T, Carović-Stanko K, Gunjača J, *et al* (2023) Monitoring drought stress in common bean using chlorophyll fluorescence and multispectral imaging. Plants 12:. https://doi.org/10.3390/plants12061386

Jiang M, Zhang Y, Li P, *et al* (2022) Mitogen-Activated Protein Kinase and Substrate Identification in Plant Growth and Development. Int J Mol Sci 23:. https://doi.org/10.3390/ijms23052744

Jochua CN, Strock CF, Lynch JP (2020) Root phenotypic diversity in common bean reveals contrasting strategies for soil resource acquisition among gene pools and races. Crop Sci 60:3261–3277. https://doi.org/10.1002/csc2.20312

Jones P, Binns D, Chang H-Y, *et al* (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Jurado M, Campa A, Ferreira JJ (2022) Differentially expressed genes against *Colletotrichum lindemuthiamum* in a bean genotype carrying the *Co-2* gene revealed by RNA-sequencing analysis. Front Plant Sci 13:981517. https://doi.org/10.3389/fpls.2022.981517

Kankanala P, Nandety RS, Mysore KS (2019) Genomics of plant disease resistance in legumes. Front Plant Sci 10:1345. https://doi.org/10.3389/fpls.2019.01345

Kassambara A, Mundt F (2020) Factoextra: Extract and Visualize the Results of Multivariate Data Analyses. Version Package Version 1.0.7

Khoury CK, Brush S, Costich DE, *et al* (2022) Crop genetic erosion: understanding and responding to loss of crop diversity. New Phytol 233:84–118. https://doi.org/10.1111/nph.17733

Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12:357–360. https://doi.org/10.1038/nmeth.3317

Koenig R, Gepts P (1989) Allozyme diversity in wild *Phaseolus vulgaris*: further evidence for two major centers of genetic diversity. Theor Appl Genet 78:809–817. https://doi.org/10.1007/BF00266663

Kokot M, Dlugosz M, Deorowicz S (2017) KMC 3: counting and manipulating k-mer statistics. Bioinformatics 33:2759–2761. https://doi.org/10.1093/bioinformatics/btx304

Kwak M, Gepts P (2009) Structure of genetic diversity in the two major gene pools of common bean (*Phaseolus vulgaris* L., Fabaceae). Theor Appl Genet 118:979–992. https://doi.org/10.1007/s00122-008-0955-4

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25. https://doi.org/10.1186/gb-2009-10-3-r25

Lazarević B, Carović-Stanko K, Živčak M, *et al* (2022) Classification of high-throughput phenotyping data for differentiation among nutrient deficiency in common bean. Front Plant Sci 13:931877. https://doi.org/10.3389/fpls.2022.931877

Lê S, Josse J, Husson F (2008) FactoMineR : An *R* package for multivariate analysis. J Stat Softw 25:. https://doi.org/10.18637/jss.v025.i01

Lewis G, Schrire B, Mackinder B, Lock M (2005) Legumes of the World. The Royal Botanic Gardens, Kew, United Kingdom

Liu Y, Du H, Li P, *et al* (2020) Pan-Genome of Wild and Cultivated Soybeans. Cell 182:162-176.e13. https://doi.org/10.1016/j.cell.2020.05.023

Li D, Quan C, Song Z, *et al* (2020) High-Throughput Plant Phenotyping Platform (HT3P) as a Novel Tool for Estimating Agronomic Traits From the Lab to the Field. Front Bioeng Biotechnol 8:623705. https://doi.org/10.3389/fbioe.2020.623705

Li H, Handsaker B, Wysoker A, *et al* (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. https://doi.org/10.48550/arxiv.1303.3997

Li L, Zhang Q, Huang D (2014) A review of imaging techniques for plant phenotyping. Sensors 14:20078–20111. https://doi.org/10.3390/s141120078

Li N, Li Y (2014) Ubiquitin-mediated control of seed size in plants. Front Plant Sci 5:332. https://doi.org/10.3389/fpls.2014.00332

Li N, Xu R, Li Y (2019) Molecular networks of seed size control in plants. Annu Rev Plant Biol 70:435–463. https://doi.org/10.1146/annurev-arplant-050718-095851

Li X, Sun M, Liu S, *et al* (2021) Functions of PPR proteins in plant growth and development. Int J Mol Sci 22:. https://doi.org/10.3390/ijms222011274

Li Y, Zheng L, Corke F, *et al* (2008) Control of final seed and organ size by the DA1 gene family in Arabidopsis thaliana. Genes Dev 22:1331–1336. https://doi.org/10.1101/gad.463608

Logozzo G, Donnoli R, Macaluso L, *et al* (2007) Analysis of the contribution of Mesoamerican and Andean gene pools to European common bean (*Phaseolus vulgaris* L.) germplasm and strategies to establish a core collection. Genet Resour Crop Evol 54:1763–1779. https://doi.org/10.1007/s10722-006-9185-2

López M, Fernández f, Schoonhoven A (1985) Frijol: Investigación y producción

Love MI, Anders S, Kim V, Huber W (2016) RNA-Seq workflow: gene-level exploratory analysis and differential expression [version 2; peer review: 2 approved]. F1000Res 4:1070. https://doi.org/10.12688/f1000research.7035.2

MacQueen AH, White JW, Lee R, *et al* (2020) Genetic associations in four decades of multienvironment trials reveal agronomic trait evolution in common bean. Genetics 215:267–284. https://doi.org/10.1534/genetics.120.303038

Maechler M, Rousseeuw P, Struyf A, *et al* (2022) cluster: Cluster Analysis Basics and Extensions. Version R package version 2.1.4

Mamidi S, Rossi M, Annam D, *et al* (2011) Investigation of the domestication of common bean (*Phaseolus vulgaris*) using multilocus sequence data. Functional Plant Biol 38:953. https://doi.org/10.1071/FP11124

Marks RA, Hotaling S, Frandsen PB, VanBuren R (2021) Representation and participation across 20 years of plant genome sequencing. Nat Plants 7:1571–1578. https://doi.org/10.1038/s41477-021-01031-8

Martín-Cabrejas MÁ (ed) (2019) Legumes: nutritional quality, processing and potential health benefits. Royal Society of Chemistry, Cambridge

Mastenbroek C (1960) A breeding programme for resistance to anthracnose in dry shell haricot beans, based on a new gene. Euphytica 9:177–184. https://doi.org/10.1007/BF00022219

Mauch-Mani B, Mauch F (2005) The role of abscisic acid in plant-pathogen interactions. Curr Opin Plant Biol 8:409–414. https://doi.org/10.1016/j.pbi.2005.05.015

Meng X, Zhang S (2013) MAPK cascades in plant disease resistance signaling. Annu Rev Phytopathol 51:245–266. https://doi.org/10.1146/annurev-phyto-082712-102314

Messina V (2014) Nutritional and health benefits of dried beans. Am J Clin Nutr 100 Suppl 1:437S–42S. https://doi.org/10.3945/ajcn.113.071472

Meziadi C, Richard MMS, Derquennes A, *et al* (2016) Development of molecular markers linked to disease resistance genes in common bean based on whole genome sequence. Plant Sci 242:351–357. https://doi.org/10.1016/j.plantsci.2015.09.006

Miklas PN, Singh SP (2007) Common Bean. In: Kole C (ed) Pulses, sugar and tuber crops. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 1–31

Miller MR, Dunham JP, Amores A, *et al* (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res 17:240–248. https://doi.org/10.1101/gr.5681207

Moghaddam SM, Brick MA, Echeverria D, *et al* (2018) Genetic architecture of dietary fiber and oligosaccharide content in a middle american panel of edible dry bean. Plant Genome 11:. https://doi.org/10.3835/plantgenome2017.08.0074

Moghaddam SM, Mamidi S, Osorno JM, *et al* (2016) Genome-Wide Association Study Identifies Candidate Loci Underlying Agronomic Traits in a Middle American Diversity Panel of Common Bean. Plant Genome 9:. https://doi.org/10.3835/plantgenome2016.02.0012

Mohammed A (2013) An overview of distribution, biology and the management of common bean anthracnose. J Plant Pathol Microbiol 04: https://doi.org/10.4172/2157-7471.1000193

Monson RK, Trowbridge AM, Lindroth RL, Lerdau MT (2022) Coordinated resource allocation to plant growth-defense tradeoffs. New Phytol 233:1051–1066. https://doi.org/10.1111/nph.17773

Morgan M, Shepherd L (2023) AnnotationHub: Client to access AnnotationHub resources. Version 3.8.0. R package

Mortazavi A, Williams BA, McCue K, *et al* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621–628. https://doi.org/10.1038/nmeth.1226

Muñoz-Mérida A, Viguera E, Claros MG, *et al* (2014) Sma3s: a three-step modular annotator for large sequence datasets. DNA Res 21:341–353. https://doi.org/10.1093/dnares/dsu001

Murube E, Campa A, Ferreira JJ (2017) Identification of new resistance sources to powdery mildew, and the genetic characterisation of resistance in three common bean genotypes. Crop Pasture Sci 68:1006. https://doi.org/10.1071/CP16460

Murube E, Campa A, Song Q, *et al* (2020) Toward validation of QTLs associated with pod and seed size in common bean using two nested recombinant inbred line populations. Mol Breeding 40:7. https://doi.org/10.1007/s11032-019-1085-1

Nadeem MA, Yeken MZ, Shahid MQ, *et al* (2021) Common bean as a potential crop for future food security: an overview of past, current and future contributions in genomics, transcriptomics, transgenics and proteomics. Biotechnology & Biotechnological Equipment 35:759–787. https://doi.org/10.1080/13102818.2021.1920462

O'Connell RJ, Bailey JA, Richmond DV (1985) Cytology and physiology of infection of *Phaseolus vulgaris* by *Colletotrichum lindemuthianum*. Physiological Plant Pathology 27:75–98. https://doi.org/10.1016/0048-4059(85)90058-X

O'Rourke JA, Iniguez LP, Fu F, *et al* (2014) An RNA-Seq based gene expression atlas of the common bean. BMC Genomics 15:866. https://doi.org/10.1186/1471-2164-15-866

Oblessuc PR, Borges A, Chowdhury B, *et al* (2012) Dissecting *Phaseolus vulgaris* innate immune system against *Colletotrichum lindemuthianum* infection. PLoS ONE 7:e43161. https://doi.org/10.1371/journal.pone.0043161

Offord CA (2017) Germplasm Conservation. In: Encyclopedia of applied plant sciences. Elsevier, pp 281–288

Ouyang S, Thibaud-Nissen F, Childs KL, *et al* (2009) Plant genome annotation methods. Methods Mol Biol 513:263–282. https://doi.org/10.1007/978-1-59745-427-8_14

Pabuayon ILB, Sun Y, Guo W, Ritchie GL (2019) High-throughput phenotyping in cotton: a review. J Cotton Res 2:18. https://doi.org/10.1186/s42397-019-0035-0

Padder BA, Kamfwa K, Awale HE, Kelly JD (2016) Transcriptome Profiling of the *Phaseolus vulgaris* - *Colletotrichum lindemuthianum* Pathosystem. PLoS ONE 11:e0165823. https://doi.org/10.1371/journal.pone.0165823

Padder BA, Sharma P, Awale HE, Kelly JD (2017) *Colletotrichum lindmuthianum*, the causal agent of bean anthracnose. Journal of Plant Pathology 99:317–330

Padilla-Chacón D, Peña Valdivia CB, García-Esteva A, *et al* (2019) Phenotypic variation and biomass partitioning during post-flowering in two common bean cultivars (*Phaseolus vulgaris* L.) under water restriction. South African Journal of Botany 121:98–104. https://doi.org/10.1016/j.sajb.2018.10.031

Pandurangan S, Pajak A, Molnar SJ, *et al* (2012) Relationship between asparagine metabolism and protein concentration in soybean seed. J Exp Bot 63:3173–3184. https://doi.org/10.1093/jxb/ers039

Papa R, Nanni L, Sicard D, *et al* (2006) The evolution of the genetic diversity in *Phaseolus vulgaris* L. In: Motley T (ed) Darwin's Harvest new approaches to the origins, evolution and conservation crops. Columbia University Press, pp 121–142

Paradis E, Schliep K (2019) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35:526–528. https://doi.org/10.1093/bioinformatics/bty633

Parker TA, Palkovic A, Gepts P (2020) Determining the Genetic Control of Common Bean Early-Growth Rate Using Unmanned Aerial Vehicles. Remote Sens (Basel) 12:1748. https://doi.org/10.3390/rs12111748

Pathirana R, Carimi F (2022) Management and utilization of plant genetic resources for a sustainable agriculture. Plants 11:. https://doi.org/10.3390/plants11152038

Pérez-Vega E, Campa A, De la Rosa L, *et al* (2009) Genetic Diversity in a Core Collection Established from the Main Bean Genebank in Spain. Crop Sci 49:1377–1386. https://doi.org/10.2135/cropsci2008.07.0409

Pérez-Vega E, Pañeda A, Rodríguez-Suárez C, *et al* (2010) Mapping of QTLs for morpho-agronomic and seed quality traits in a RIL population of common bean (*Phaseolus vulgaris* PL.). Theor Appl Genet 120:1367–1380. https://doi.org/10.1007/s00122-010-1261-5

Piepho HP, Möhring J, Melchinger AE, Büchse A (2008) BLUP for phenotypic selection in plant breeding and variety testing. Euphytica 161:209–228. https://doi.org/10.1007/s10681-007-9449-8

Poland JA, Brown PJ, Sorrells ME, Jannink J-L (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS ONE 7:e32253. https://doi.org/10.1371/journal.pone.0032253

Pop M (2009) Genome assembly reborn: recent computational challenges. Brief Bioinformatics 10:354–366. https://doi.org/10.1093/bib/bbp026

Powell AA, Oliveira MDA, Matthews S (1986) The Role of Imbibition Damage in Determining the Vigour of White and Coloured Seed Lots of Dwarf French Beans (*Phaseolus vulgaris*). J Exp Bot 37:716–722. https://doi.org/10.1093/jxb/37.5.716

Puerta Romero J (1961) Variedades de judias cultivadas en Espana: nueva clasificacion de la especie *Phaseolus vulgaris* (L. ex p.) Savi

Rafalski JA, Vogel JM, Morgante M, *et al* (1996) Generating and using DNA markers in plants. In: Nonmammalian Genomic Analysis. Elsevier, pp 75–134

Rakwal R, Shii K, Agrawal GK, Yonekura M (2001) Protein phosphatase inhibitors activate defense responses in rice (*Oryza sativa*) leaves. Physiol Plant 111:151–157. https://doi.org/10.1034/j.1399-3054.2001.1110204.x

Rani K, Kumar M, Razzaq A, *et al* (2023) Recent advances in molecular marker technology for QTL mapping in plants. In: QTL mapping in crop improvement. Elsevier, pp 1–15

Richard MMS, Gratias A, Alvarez Diaz JC, *et al* (2021) A common bean truncated CRINKLY4 kinase controls gene-for-gene resistance to the fungus *Colletotrichum lindemuthianum.* J Exp Bot 72:3569–3581. https://doi.org/10.1093/jxb/erab082

R Core Team (2023) R: A Language and Environment for Statistical Computing. Version 4.3.1. R Foundation for Statistical Computing

Rivera A, Plans M, Sabaté J, *et al* (2018) The Spanish Core Collection of Common Beans (*Phaseolus vulgaris* L.): An Important Source of Variability for Breeding Chemical Composition. Front Plant Sci 9:1642. https://doi.org/10.3389/fpls.2018.01642

Robinson JT, Thorvaldsdóttir H, Winckler W, *et al* (2011) Integrative genomics viewer. Nat Biotechnol 29:24–26. https://doi.org/10.1038/nbt.1754

Rodriguez M, Rau D, Bitocchi E, *et al* (2016) Landscape genetics, adaptive diversity and population structure in *Phaseolus vulgaris*. New Phytol 209:1781–1794. https://doi.org/10.1111/nph.13713

R Core Team (2021) R: A Language and Environment for Statistical Computing. Version 4.1.0. R Foundation for Statistical Computing, Vienna, Austria

Rodríguez-Suárez C, Méndez-Vigo B, Pañeda A, *et al* (2007) A genetic linkage map of *Phaseolus vulgaris* L. and localization of genes for specific resistance to six races of anthracnose (*Colletotrichum lindemuthianum*). Theor Appl Genet 114:713–722. https://doi.org/10.1007/s00122-006-0471-3

Rodríguez Madrera R, Campa Negrillo A, Ferreira Fernández JJ (2024) Modulation of the nutritional and functional values of common bean by farming system: organic vs. conventional. Front Sustain Food Syst 7:. https://doi.org/10.3389/fsufs.2023.1282427

Rodríguez Madrera R, Campa Negrillo A, Suárez Valles B, Ferreira Fernández JJ (2020) Characterization of extractable phenolic profile of common bean seeds (*Phaseolus vulgaris* L.) in a Spanish diversity panel. Food Res Int 138:109713. https://doi.org/10.1016/j.foodres.2020.109713

Rossi M, Bitocchi E, Bellucci E, *et al* (2009) Linkage disequilibrium and population structure in wild and domesticated populations of *Phaseolus vulgaris* L. Evol Appl 2:504–522. https://doi.org/10.1111/j.1752-4571.2009.00082.x

Saijo Y, Loo EP-I (2020) Plant immunity in signal integration between biotic and abiotic stress responses. New Phytol 225:87–104. https://doi.org/10.1111/nph.15989

Salinas AD, Bonet A, Gepts P (1988) The wild relative of *Phaseolus vulgaris* in middle america. In: Gepts P (ed) Genetic resources of phaseolus beans. Springer Netherlands, Dordrecht, pp 163–184

Sandhu KS, You FM, Conner RL, *et al* (2018) Genetic analysis and QTL mapping of the seed hardness trait in a black common bean (*Phaseolus vulgaris*) recombinant inbred line (RIL) population. Mol Breed 38:34. https://doi.org/10.1007/s11032-018-0789-y

Santalla M, Rodiño P, De Ron M (2002) Allozyme evidence supporting southwestern Europe as a secondary center of genetic diversity for the common bean. Theor Appl Genet 104:934–944. https://doi.org/10.1007/s00122-001-0844-6

Sato S, Nakamura Y, Kaneko T, *et al* (2008) Genome structure of the legume, Lotus japonicus. DNA Res 15:227–239. https://doi.org/10.1093/dnares/dsn008

Schalamun M, Schwessinger B (2017) DNA size selection (>1kb) and clean up using an optimized SPRI beads mixture v1. https://doi.org/10.17504/protocols.io.idmca46

Schmutz J, McClean PE, Mamidi S, *et al* (2014) A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet 46:707–713. https://doi.org/10.1038/ng.3008

Schröder S, Mamidi S, Lee R, *et al* (2016) Optimization of genotyping by sequencing (GBS) data in common bean (*Phaseolus vulgaris* L.). Mol Breeding 36:6. https://doi.org/10.1007/s11032-015-0431-1

Schwartz HF, Steadman JR, Hall R, Forster RL (2005) Compendium of Bean Diseases, 2nd Edition (Disease Compendium), 2nd edn. Amer Phytopathological Society, St. Paul, Minn

Seppey M, Manni M, Zdobnov EM (2019) BUSCO: assessing genome assembly and annotation completeness. Methods Mol Biol 1962:227–245. https://doi.org/10.1007/978-1-4939-9173-0_14

Shi J, Tian Z, Lai J, Huang X (2023) Plant pan-genomics and its applications. Mol Plant 16:168–186. https://doi.org/10.1016/j.molp.2022.12.009

Silbernagel MJ, Janssen W, Davis JHC, Montes de Oca G (1991) Snap bean production in the tropics: Implications for genetic improvement. In: Schoonhoven A van, Voysest O (eds) Common beans : Research for crop improvement. CAB International; Cali, CO : Centro Internacional de Agricultura Tropical (CIAT), Oxon, GB., pp 835–862

Singh BD, Singh AK (2015) Introduction to Marker-Assisted Crop Improvement. In: Marker-Assisted Plant Breeding: Principles and Practices. Springer India, New Delhi, pp 3–16

Singh SP, Gepts P, Debouck DG (1991) Races of common bean (*Phaseolus vulgaris*, Fabaceae). Econ Bot 45:379–396. https://doi.org/10.1007/BF02887079

Singh SP (1989) Patterns of variation in cultivated common bean (*Phaseolus vulgaris*, Fabaceae). Econ Bot 43:39–57. https://doi.org/10.1007/BF02859324

Stagnari F, Maggio A, Galieni A, Pisante M (2017) Multiple benefits of legumes for agriculture sustainability: an overview. Chem Biol Technol Agric 4:2. https://doi.org/10.1186/s40538-016-0085-1

Stanke M, Morgenstern B (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res 33:W465-7. https://doi.org/10.1093/nar/gki458

Steckling S de M, Ribeiro ND, Arns FD, *et al* (2017) Genetic diversity and selection of common bean lines based on technological quality and biofortification. Genet Mol Res 16:. https://doi.org/10.4238/gmr16019527

Takuno S, Terauchi R, Innan H (2012) The power of QTL mapping with RILs. PLoS ONE 7:e46545. https://doi.org/10.1371/journal.pone.0046545

Tanabata T, Shibaya T, Hori K, *et al* (2012) SmartGrain: high-throughput phenotyping software for measuring seed shape through image analysis. Plant Physiol 160:1871–1880. https://doi.org/10.1104/pp.112.205120

Tarasov A, Vilella AJ, Cuppen E, *et al* (2015) Sambamba: fast processing of NGS alignment formats. Bioinformatics 31:2032–2034. https://doi.org/10.1093/bioinformatics/btv098

Tarazona S, Furió-Tarí P, Turrà D, *et al* (2015) Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. Nucleic Acids Res 43:e140. https://doi.org/10.1093/nar/gkv711

Tarazona S, García-Alcalde F, Dopazo J, *et al* (2011) Differential expression in RNA-seq: a matter of depth. Genome Res 21:2213–2223. https://doi.org/10.1101/gr.124321.111

Tian X, He M, Mei E, *et al* (2021) WRKY53 integrates classic brassinosteroid signaling and the mitogen-activated protein kinase pathway to regulate rice architecture and seed size. Plant Cell 33:2753–2775. https://doi.org/10.1093/plcell/koab137

Tillich M, Lehwark P, Pellizzer T, *et al* (2017) GeSeq - versatile and accurate annotation of organelle genomes. Nucleic Acids Res 45:W6–W11. https://doi.org/10.1093/nar/gkx391

Tirnaz S, Edwards D, Batley J (2020) The importance of plant pan-genomes in breeding. In: Kang MS (ed) Quantitative genetics, genomics and plant breeding. CABI, UK, pp 27–32

Tisserant E, Da Silva C, Kohler A, *et al* (2011) Deep RNA sequencing improved the structural annotation of the *Tuber melanosporum* transcriptome. New Phytol 189:883–891. https://doi.org/10.1111/j.1469-8137.2010.03597.x

Tu JC (1988) Control of Bean Anthracnose Caused by the Delta and Lambda Races of *Colletotrichurn lindemuthianurn* in Canada. Plant Dis 72:5. https://doi.org/10.1094/PD-72-0005

Tu JC (1992) *Colletotrichum lindemuthianum* on Bean: Population Dynamics of the Pathogen and Breeding for Resistance. In: Colletotrichum: Biology, Pathology and Control (Cabi), First. CABI, Wallingford, Oxon, UK, pp 203–224

Tu M, Zeng J, Zhang J, *et al* (2022) Unleashing the power within short-read RNA-seq for plant research: Beyond differential expression analysis and toward regulomics. Front Plant Sci 13:1038109. https://doi.org/10.3389/fpls.2022.1038109

Uebersax MA, Cichy KA, Gomez FE, *et al* (2022) Dry beans (*Phaseolus vulgaris* L.) as a vital component of sustainable agriculture and food security—A review. Legume Science 5:. https://doi.org/10.1002/leg3.155

Ugwuanyi S, Udengwu OS, Snowdon RJ, Obermeier C (2022) Novel candidate loci for morpho-agronomic and seed quality traits detected by targeted genotyping-by-sequencing in common bean. Front Plant Sci 13:1014282. https://doi.org/10.3389/fpls.2022.1014282

van de Wouw M, Kik C, van Hintum T, *et al* (2010) Genetic erosion in crops: concept, research results and challenges. Plant Genet Res 8:1–15. https://doi.org/10.1017/S1479262109990062

van Hintum ThJL, Brown AHD, Spillane C, Hodgkin T (2000) Core collections of plant genetic resources. International Plant Genetic Resources Institute, Rome, Italy

Varshney RK, Bohra A, Yu J, *et al* (2021a) Designing Future Crops: Genomics-Assisted Breeding Comes of Age. Trends Plant Sci 26:631–649. https://doi.org/10.1016/j.tplants.2021.03.010

Varshney RK, Roorkiwal M, Sun S, *et al* (2021b) A chickpea genetic variation map based on the sequencing of 3,366 genomes. Nature 599:622–627. https://doi.org/10.1038/s41586-021-04066-1

Vasconcelos MW, Grusak MA, Pinto E, *et al* (2020) The biology of legumes and their agronomic, economic, and social impact. In: Hasanuzzaman M, Araújo S, Gill SS (eds) The plant family fabaceae: biology and physiological responses to environmental stresses. Springer Singapore, Singapore, pp 3–25

Vaz Bisneta M, Gonçalves-Vidigal MC (2020) Integration of anthracnose resistance loci and RLK and NBS-LRR-encoding genes in the *Phaseolus vulgaris* L. genome. Crop Sci 60:2901–2918. https://doi.org/10.1002/csc2.20288

Vidak M, Lazarević B, Javornik T, *et al* (2022) Seed water absorption, germination, emergence and seedling phenotypic characterization of the common bean landraces differing in seed size and color. Seeds 1:324–339. https://doi.org/10.3390/seeds1040027

Vilanova S, Alonso D, Gramazio P, *et al* (2020) SILEX: a fast and inexpensive high-quality DNA extraction method suitable for multiple sequencing platforms and recalcitrant plant species. Plant Methods 16:110. https://doi.org/10.1186/s13007-020-00652-y

Voysest O (2000) Mejoramiento genético del frijol (*Phaseolus vulgaris* L.). In: Legado de variedades de América Latina 1930-1999. Centro Internacional de Agricultura Tropical, Cali, Colombia

Voysest Voysest O (2000) Mejoramiento genético del frijol (*Phaseolus vulgaris* L.): Legado de variedades de América Latina 1930-1999. Centro Internacional de Agricultura Tropical (CIAT), Cali, CO

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63. https://doi.org/10.1038/nrg2484

Wei T, Simko V (2021) R package "corrplot": Visualization of a Correlation Matrix. Version 0.92. R core

Wen Y-J, Zhang H, Ni Y-L, *et al* (2018) Methodological implementation of mixed linear models in multi-locus genome-wide association studies. Brief Bioinformatics 19:700–712. https://doi.org/10.1093/bib/bbw145

White JW, González A (1990) Characterization of the negative association between seed yield and seed size among genotypes of common bean. Field Crops Res 23:159–175. https://doi.org/10.1016/0378-4290(90)90052-D

Wickham H (2016) ggplot2: Elegant Graphics for Data Analysis (Use R!), 2nd edn. Springer, Cham

Wickland DP, Battu G, Hudson KA, *et al* (2017) A comparison of genotyping-by-sequencing analysis methods on low-coverage crop datasets shows advantages of a new workflow, GB-eaSy. BMC Bioinformatics 18:586. https://doi.org/10.1186/s12859-017-2000-6

Wong CY, Gilbert ME, Pierce MA, *et al* (2023) Hyperspectral remote sensing for phenotyping the physiological drought response of common and tepary bean. Plant Phenomics 5:0021. https://doi.org/10.34133/plantphenomics.0021

Wu T, Hu E, Xu S, *et al* (2021) clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation (Camb) 2:100141. https://doi.org/10.1016/j.xinn.2021.100141

Wu X, Cai X, Zhang B, *et al* (2022) ERECTA regulates seed size independently of its intracellular domain via MAPK-DA1-UBP15 signaling. Plant Cell 34:3773–3789. https://doi.org/10.1093/plcell/koac194

Xiao W, Hu S, Zou X, *et al* (2021) Lectin receptor-like kinase LecRK-VIII.2 is a missing link in MAPK signaling-mediated yield control. Plant Physiol 187:303–320. https://doi.org/10.1093/plphys/kiab241

Xia T, Li N, Dumenil J, *et al* (2013) The ubiquitin receptor DA1 interacts with the E3 ubiquitin ligase DA2 to regulate seed and organ size in Arabidopsis. Plant Cell 25:3347–3359. https://doi.org/10.1105/tpc.113.115063

Xu R, Duan P, Yu H, *et al* (2018) Control of Grain Size and Weight by the OsMKKK10-OsMKK4-OsMAPK6 Signaling Pathway in Rice. Mol Plant 11:860–873. https://doi.org/10.1016/j.molp.2018.04.004

Yang IS, Kim S (2015) Analysis of whole transcriptome sequencing data: workflow and software. Genomics Inform 13:119–125. https://doi.org/10.5808/GI.2015.13.4.119

Yanni AE, Iakovidi S, Vasilikopoulou E, Karathanos VT (2023) Legumes: A vehicle for transition to sustainability. Nutrients 16:. https://doi.org/10.3390/nu16010098

Ye J, Coulouris G, Zaretskaya I, *et al* (2012) Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics 13:134. https://doi.org/10.1186/1471-2105-13-134

Zeven AC (1998a) Landraces: A review of definitions and classifications. Euphytica 104:127–139

Zeven A (1998b) Landraces: A review of definitions and classifications. Euphytica 104:127–139

Zhang SV, Zhuo L, Hahn MW (2016) AGOUTI: improving genome assembly and annotation using transcriptome data. Gigascience 5:31. https://doi.org/10.1186/s13742-016-0136-3

Zhang X, Blair MW, Wang S (2008) Genetic diversity of Chinese common bean (*Phaseolus vulgaris* L.) landraces assessed with simple sequence repeat markers. Theor Appl Genet 117:629–640. https://doi.org/10.1007/s00122-008-0807-2

Zhang Y-W, Tamba CL, Wen Y-J, *et al* (2020) mrMLM v4.0.2: An R Platform for Multi-locus Genome-wide Association Studies. Genomics Proteomics Bioinformatics 18:481–487. https://doi.org/10.1016/j.gpb.2020.06.006

Zimin AV, Puiu D, Luo M-C, *et al* (2017) Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Res 27:787–792. https://doi.org/10.1101/gr.213405.116

Zimin AV, Salzberg SL (2020) The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. PLoS Comput Biol 16:e1007981. https://doi.org/10.1371/journal.pcbi.1007981

Zizumbo-Villarreal D, Colunga-GarcíaMarín P, de la Cruz EP, *et al* (2005) Population structure and evolutionary dynamics of wild–weedy–domesticated complexes of common bean in a mesoamerican region. Crop Sci 45:1073. https://doi.org/10.2135/cropsci2004.0340

ANNEX I

Supplementary materials information

All supplementary material can be found in the USB card.

Supplementary materials information

CHAPTER 1. Genetic erosion within the Fabada dry bean market class revealed by high-throughput genotyping.

Table 1S.1. List of the common bean lines included in the FabaPanel with the passport data. The lines that conform to the different subsets to carry out the diversity analysis are indicated.

Table 1S.2. Estimation of Nei's genetic distances and Fst fixation index among the four populations (Conserved, Cultivar, Cultivated, Reference) using the package SambaR in the filtered FabaPanel population (107).

Figure 1S.1. Distribution along the 11 bean chromosomes of the 22,259 SNP with which the FilterFabPanel1 was genotyped after filtering.

Figure 1S.2. Scatter plot obtained with the two mains coordinates revealed by the Principal Coordinate Analysis (PCoA) based on Hamming's genetic distance of the 159 lines classified as Fabada market class and genotyped with 21,837 SNP.

Figure 1S.3. Dendrogram obtained from FilterFabaPanel2 (159 lines recorded as Fabada market class) genotyped with 21837 SNP using the Euclidea distance and the UPGMA method.

Figure 1S.4. Segregating site distributions per population along the 11 bean chromosomes. a) Conserved population. b) Cultivated population. c) Cultivar population.

Figure 1S.5. Distribution of segregating sites (SNP) per bean chromosome and genotypes of the FilterFabaPanel3. Black, major allele; yellow, minor allele.

CHAPTER 2. Differentially expressed genes against Colletotrichum lindemuthianum in a bean genotype carrying the Co-2 gene revealed by RNA-sequencing analysis

Table 2S.1. Quality of RNA sequencing data. Sample nomenclature: S, susceptible genotype A25; R, resistant genotype A4804; 0, 24, 48, hpi; 1, 2, 3, repetitions (different resistance tests).

Table 2S.2. List of differential genes expressed observed in the seven in the nine comparisons made (R24-R0, R48-R0, S24-S0, S48-S0, R0-S0, R24-S24 and R48- S48) using the package NOISeq. FPKM (Fragments Per Kilobase of transcript per Million Mapped reads); M (which is the log2-ratio of the two conditions) and D (the value of the difference between conditions). GO (Gene Ontology) assigned to each gene are also indicated.

Table 2S.3. List of GO terms significantly enriched at 48 hpi revealed by the package ViSEAGO. **a)** List of GO terms significantly enriched for Biological Functions (see Figure 2S.3a); **b)** List of GO terms significantly enriched for Molecular Functions (see Figure 2S.3b); **c)** List of GO terms significantly enriched for Cellular Component (see Figure 2S.3c).

Table 2S.4. List of GO terms significantly enriched at 24 hpi revealed by the package ViSEAGO. **a)** List of GO terms significantly enriched for Biological Functions (see Figure 2S.3d); **b)** List of GO terms significantly enriched for Molecular Functions (see Figure 2S.3e); **c)** List of GO terms significantly enriched for Cellular Component (see Figure 2S.4f).

Table 2S.5. Characteristics of the markers developed from the raw read sequences revealed by transcriptome analysis. Forward and reverse primer sequence, primer annealing temperature (AT) that produced the PCR products. Markers were named with Co2 following by the position in the chromosome 11 of the polymorphism to be amplified in *Phaseolus vulgaris* v1.

Figure 2S.1. Pedigree of lines analyzed in this work. Line A25 proceeds from a selection of the landrace 'Andecha', classified in the market class fabada. Line A4804 is an isogenic line originally derived from A25 carrying a resistance locus to anthracnose located in the cluster Co-2. Additionally, line A4804 carries resistance to BCMV (gene I) and powdery mildew (gene *Pm1*).

Figure 2S.2. RNA seq quality exploration. a) Scatterplots of the two principal components over FPKM normalized data contain all the samples. b) Heatmap and hierarchical clustering analysis (HCA) of FPKM values contains all samples.

Figure 2S.3. Visualization of ViSEAGO's functional analysis from 1,899 DEGs with assigned GO for the categories 'Biological process (BP)', 'Molecular Function (MF)', and 'Cellular Components (CC)'. Clustering heatmap plot that combines a dendrogram based on Wang's semantic similarity distance and ward.D2 aggregation criterion, a heatmap of -log10(pvalue) from functional enrichment tests and the information content (IC). **a)** Functional enrichment terms for BP at 48 hpi in the two genotypes (DEGs from comparations R0-R48 and S0-S48, respectively); **b)** Functional enrichment term for MF at 48 hpi; **c)** Functional enrichment terms for CC at 48 hpi; **d)** Functional enrichment term for BP at 24 hpi; **e)** Functional enrichment term for MF at 24 hpi; **f)** Functional enrichment term for CC at 24 hpi.

Figure 2S.4. The end of bean chromosome Pv11 with the introgression regions (SNPs with genotypes SanilacBc6 Are) in the NILs A1258, X2776, A2806, and A4804. * GBS results of Murube *et al* (2017) ** GBS results of the present study.

Figure 2S.5. Results of the alignments of four read obtained in the genotypes A25 (susceptible) and A4804 (resistant) with the bean genomes G19833v1, G19833v2.1, 5-592, UI111 and Labor Ovalle (https://phytozome-next.jgi.doe.gov/). Used polymorphisms to develop specific molecular markers are indicated in the box. Sequences and positions of the forwards

and reverses primers are shown as underlines. **a)** M1 - Co2_46.961.315; **b)** M2 - Co2_46.984.860; **c)** M3 - Co2_46.989.310; **d)** M4 - Co2_47.017.090.

CHAPTER 3. Identification of consistent QTL and candidate genes associated with seed traits in common bean by combining GWAS and RNA-Seq.

Table 3S.1. Genotypic data (SNP markers) of the 298 SDP lines used in this study.

Table 3S.2. Phenotype data and results of HCPC analysis. Clust, cluster assigned by HPCP analysis using the 5 QTL associated with seed weight. Seed01_51.9, Seed03_45.6, Seed07_0.62, Seed08_55.3 and Seed10_39.1, the FIVE QTL identified for 25-seed weight. Mean data for the 7 trait measures in the SDP lines used in this study. Gene pool were described by Campa *et al* (2018).

Table 3S.3. Results of RNA-seq analysis. D1, D2, and D3 represent developmental stages for seeds. R1 and R2 indicate biological replicates. TMM (trimmed mean of M values).

Table 3S.4. List of differentially expressed genes (DEGs) observed in the comparisons made (D2-D1; D3-D1 and D-D1) using the package NOISeq. TMM, trimmed mean of M values. 1, TMM mean of the first stage in the comparison. 2, TMM mean of the second stage of the comparison.

Table 3S.5. List of GO terms significantly enriched from the comparation between the stages D3 and D1 revealed by the the R package ClusterProfiler.

Table 3S.6. List of annotated genes underlaying the identified QTL in this study. X shows if the gene is a differentially expressed genes in this study or in the *Phaseolus* Atlas.

Table 3S.7. List of differentially expressed genes (DEGs) underlaying to QTL regions revelated in the association analysis.

Table 3S.8. Results of the alignment of the 22 DEGs in the reference genome G19833 with the respective genes predicted in the bean genotypes LaborOvalle, 5-593 and UI111.

Figure 3S.1. Distribution of the SNPs across the eleven bean chromosomes.

Figure 3S.2. Phenotypic frequency distribution of adjusted means of the seven seed traits evaluated in Spanish Diversity Panel.

Figure 3S.3. Manhattan and QQ plots obtained with FASTmrEMMA method for the morpho-agronomics traits. a) Area; b) Length; c) Width; d) LWR; e) Seed Weight; f) Coat proportion; g) Water absorption.

Figure 3S.4. SNP data HPCP of the 6 regions associated with SW. a) Seed01_51.9. b) Seed03_45.6. c) Seed07_0.62. d) Seed08_55.3. e) Seed10_39.1.

Figure 3S.5. RNA-Seq data quality exploration. a) Scatterplots of the two principal components over TMM normalized data contains all the samples. b) Boxplots with the TMM normalized data contains all the locus after normalization.

Figure 3S.6. GO Terms enrichment for the DEGs found in the comparison D3 vs D1. a) GO terms in Biological Process (BP) category. b) GO terms in Cellular Components (CC).

CHAPTER 4. A new bean genomic resource: de novo assembly and annotation of a Fabada cultivar

Table 4S.1. Functional annotation of the 58,955 transcripts of line A25 scaffolds assembly.