2021

# Deciphering seed development mechanisms in *Phaseolus vulgaris* L.

**José Ricardo Torrão Dias Parreira Salvado**

Dissertation presented to obtain the **Ph.D. degree in Plant Sciences**

**Plants for Life**

Oeiras, July, 2021

**INSTITUTO DE TECNOLOGIA QUÍMICA E BIOLÓGICA** ANTÓNIO XAVIER | UNIVERSIDADE NOVA DE LISBOA

# Deciphering seed development mechanisms in *Phaseolus vulgaris* L.

José Ricardo Torrão Dias Parreira Salvado

Dissertation presented to obtain the Ph.D. degree in Plant Sciences – Plants for Life

Instituto de Tecnologia Química e Biológica António Xavier | Universidade Nova de Lisboa

Oeiras, July 2021

*‘People are very fond of giving away what they need most themselves.*

*It is what I call the depth of generosity.’*

**Oscar Wilde, The Picture of Dorian Gray**

# Deciphering seed development mechanisms in *Phaseolus vulgaris* L.

José Ricardo Torrão Dias Parreira Salvado

Dissertation presented to obtain the Ph.D. degree in Plant Sciences – Plants for Life

Instituto de Tecnologia Química e Biológica António Xavier | Universidade Nova de Lisboa

Oeiras, July 2021

Research work performed within the Plants for Life PhD Program (PD/00035/2013), supported by Fundação para a Ciência e a Tecnologia (PD/BD/113474/2015).

**The research described in this thesis was performed under the supervision of:**

**Professor Pedro Fevereiro**
Head of Plant Cell Biotechnology Laboratory, Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa, Avenida da República, 2780-157 Oeiras, Portugal
CEO of InnovPlantProtect CoLab, Estrada de Gil Vaz, Apartado 72, 7351-901 Elvas, Portugal

**Doctor Susana Araújo**
Research Director - North Delegation at Association BLC3 -Technology and Innovation Campus (North Delegation), Rua Comendador Emílio Augusto Pires, 14, Edifício SIDE UP, 5340-257 Macedo de Cavaleiros, Portugal

**All work was performed at:**
Instituto de Tecnologia Química e Biológica António Xavier
Universidade Nova de Lisboa
Av. da República, 2780-157 Oeiras, Portugal

**Front Cover:** Close-up of SER 16 *Phaseolus vulgaris* seeds.

# Acknowledgements

I would like to express my sincere gratitude to everyone that contributed, directly or indirectly, to this thesis. Moreover, I would like to extend my thanks to the hosting institution ITQB NOVA for the excellent working conditions. To the financial support provided by Fundação para a Ciência e Tecnologia for the project PTDC/AGR-GPL/110224/2009, my Plants for Life PhD grant (PD/BD/113474/2015) within the scope of the PhD program Plants for Life (PD/00035/2013) in the GREEN-it: Bioresources for Sustainability (UID/Multi/04551/2020).

To my supervisors, Susana and Pedro, for introducing me to the wonders of plant biology, for the opportunity to develop this thesis, for the support and scientific discussions, and more importantly for always making try to see the “big picture”. Pedro, thank you for trying to make me more pragmatic and for teaching me the responsibility of being exact in my words. Also, for always questioning “have you checked if there aren’t any interesting results even if it didn’t change in abundance?”. Susana, thank you very much for believing in me since the beginning, for all the support and for the opportunities to grow and learn. I will cherish forever all the knowledge, the “show me your controls”, the “are you sure of that? Where is the reference?”, all the scientific meetings around a good coffee or amazing meals, but most of all for teaching me that being a supervisor is so much more than a “boss”. It is a colleague and a friend. Thank you. To my thesis committee, Dr. Jorge Paiva and Prof. Alma Balestrazzi, for the many insightful discussions, all the sharing of knowledge and help you always provided me.

To all of my former and present colleagues at the Plant Cell Biotechnology Laboratory and Genetics and Genomics of Plant Complex Traits Laboratory. Hélia Sales, Letice Gonçalves, Davide Martins, Priscila Pereira, Catarina Bicho, André Almeida, Diana Branco, Victor Carocha, Clara Graça, Sofia Duque, Rita Morgado, Ehsan Valiollahi, Francisco Fernandes, Joana Lérias, Matilde Cordeiro, Carlota Vaz Patto, Nuno Almeida, Susana Pera,

Carmen Santos, Susana Leitão, Mara Alves, Cláudia Lopes, Margarida Basaloco, Margarida Sampaio, Maria Assunção, Olivia Costa, Sara Costa, Sara Tedesco, Michela Cappuccio, Raquel Guedes, Carla Amaro and Carolina Gomes. Thank you very much for all the scientific discussions, critical suggestions and all the good moments. Much more than that, thank you for all the laughs and friendship.

Dear Avecs, we make a very peculiar group. Thank you so much for being a part of this journey, since before it began, until the finish line… All the support, suggestions, questions, discussions, and revisions were invaluable. But most of all, for the great friendship.

A todos os meus amigos, que sabem exatamente quem são. O meu sincero obrigado à Margarida Lima, Ana Pena, Catarina Santos, Inês e Ana Fragata, Carlos Almeida Carla Nunes, Inês Trindade, Jorge Faria, Nélia Penteado, André Morais, Ana Oliveira, Ana Isabel, Hélio, Hugo, Sofia, Tiago Gonçalves, Marisa, Maria João, Teresa Neves, André e Beatriz. Ao grande tradutor Ivan Figueiras etc. Obrigado Rute Silva, Joana Figueiras e Rita Caré por me mostrarem que mudar é necessário para crescermos e sermos felizes; Elsa Sousa, a tua resiliência e força inspiram-me a lutar mais; Ana Maria, por estares sempre presente e ensinares o valor de dizer “não”; Luís Fernandes, pelo teu apoio constante, independentemente das minhas decisões. Miguel Guerreiro, a tua força, apoio incondicional, e me obrigares a sair da minha zona de conforto, ajudam-me diariamente. Somos uma equipa. Sempre.

Obrigado a todos por me aturarem, por acreditarem em mim, por me ouvirem, mesmo quando sou chato e parvo. E quando o meu mau humor e respostas tortas vêm ao de cima. Que, convenhamos, é pelo menos metade do tempo. Por último, mas não menos importantes: à minha família, avó, tios e primas, por todo o apoio. Sempre. Em especial aos meus pais, que incondicionalmente me apoiaram, sacrificaram, ouviram, educaram, e pelos valores que me incutiram - e ainda incutem. Muito obrigado. **Não há muito mais a dizer do que sem todos vocês não chegaria a alcançar esta meta. Não tenho qualquer dúvida.**

# Table of Contents

# List of most used abbreviations

| | |
|---|---|
| **BC** | Mercator Bin Code |
| **Cn/µL** | Transcript copy number per microliter |
| **DAA** | Days after anthesis |
| **DAP** | Differentially abundant proteins |
| **DDR** | DNA Damage Response |
| **DNA** | Deoxyribonucleic acid |
| **dPCR** | Digital PCR |
| **EC** | Enzyme Commission Number |
| **GO** | Gene Ontology |
| **LAFL** | LEAFY COTYLEDON1 (LEC1), ABI3, FUSCA3 (FUS3) and LEAFY COTYLEDON2 (LEC2) |
| **LC-MS/MS** | Liquid Chromatography-tandem Mass Spectrometry |
| **MACE** | Massive Analysis of 3'-cDNA Ends |
| **MFEI** | Minimal folding energy index |
| **miRNA** | MicroRNAs |
| **mRNA** | Messenger RNA |
| **nt** | Nucleotide |
| **PCA** | Principal Component Analysis |
| **PCR** | Polymerase chain reaction |
| **psRNATarget** | Plant small RNA target server |
| **pvu** | *Phaseolus vulgaris* |
| **RNA** | Ribonucleic acid |
| **RNA-Seq** | RNA Sequencing |
| **RT-qPCR** | Reverse transcription quantitative PCR |
| **SD** | Seed Development |
| **sRNA-Seq** | Small RNA Sequencing |
| **TF** | Transcription Factor |

# Abstract

Seed development is one of the most relevant developmental processes in seed-producing crops. Seeds are essential to plants since they assure the next generation and the survival of the plant species. Nevertheless, they are important as food or feed for humans and animals, as a source of nutrients, playing important agronomic and socio-economic roles. Seed development is a complex developmental process that involves numerous molecular players and complex regulatory networks that is still poorly addressed in the literature, particularly for Legumes. This gap of knowledge was addressed within the scope of this thesis.

*Phaseolus vulgaris* L., the common bean, is a grain legume that plays an important agronomical role due to its capacity for fixing atmospheric nitrogen and, nutritionally, being a good source of proteins, carbohydrates, dietary fibres, vitamins, and minerals. Indeed, the common bean seeds are one of the most consumed staple foods worldwide, with a relevant role ensuring food security. In Portugal, common bean is the most consumed grain legume *per capita*. Several published reports have indicated common bean as one of the best agricultural model crops to study seed biology. However, preceding the onset of this PhD research, limited insights into the molecular mechanisms that shape *P. vulgaris* seed development were available. There was no comprehensive description of the proteins and genes accumulated at the different developmental stages, from late embryogenesis to seed desiccation, which is relevant to identify the acting molecular mechanisms and regulatory networks. The research conducted herein aimed at addressing several open research questions in this important species: i) how do phenotypic characteristics change during seed development and what is the timeframe of the main developmental stages? ii) what are the main molecular mechanisms and metabolic pathways at each

seed developmental stage? and iii) what regulatory networks and key regulators might impact seed traits of interest?

To address these questions, a comprehensive molecular analysis using Proteomics, and Coding and Non-Coding Transcriptomics, was used to describe the molecular and regulatory mechanisms underlying seed development in *P. vulgaris*. We aimed at: 1) the characterization of the main seed developmental stages to define the timeframe where molecular analysis would be conducted; 2) identify the main molecular mechanisms and metabolic pathways underlying seed development; 3) characterize the genome integrity maintenance mechanisms during this developmental process, and; 4) Understand the role of post-transcriptional regulatory mechanisms mediated by miRNAs in the developing seeds of *P. vulgaris*.

Limited knowledge about the molecular mechanisms controlling seed development in *P. vulgaris* was available prior to this thesis. In **Chapter II**, we aimed at providing a comprehensive description of proteome dynamics during this developmental process. A high-throughput gel-free proteomics approach (LC-MS/MS) was conducted on seeds at 10, 20, 30 and 40 days after anthesis (DAA), spanning from late embryogenesis until desiccation. Over 400 differentially accumulated proteins were observed. An accumulation of DAPs belonging to the MapMan functional categories of “protein”, “glycolysis”, “TCA”, “DNA”, “RNA”, “cell” and “stress” were found at early developmental stages, reflecting an extensive metabolic activity. Intermediate stages were characterized by the accumulation of storage, signalling, starch synthesis and cell wall-related proteins. In the later developmental stage, an increase in proteins related to redox and, to a less extent protein degradation/modification/folding and nucleic acid metabolisms reflect that seed desiccation-resistance mechanisms were activated. Overall, we described the main molecular mechanisms, metabolic pathways and stage specific proteomic signatures. Furthermore, clues about the putative

role of PTMs, such as neddylation, and new insights into the maintenance of genome integrity, were unveiled.

Contrary to what is described for germination, the knowledge on the mechanisms implicated in maintenance of genome integrity during seed development was still limited at this point, even with results from **Chapter II**. In **Chapter III** we aimed at the characterization of these mechanisms acting during this developmental process. For this, a qualitative and quantitative description of gene expression dynamics, using high-throughput transcriptomics methodology Massive Analysis of 3'-cDNA Ends (MACE) and digital PCR (dPCR) respectively, across developmental stages was presented. Our results highlight new information regarding DNA integrity maintenance in developing seeds, including DNA damage repair and chromatin remodelling mechanisms, which ensure genome integrity. At early seed development, a high expression of genes related to DNA damage sensing and repair suggests there is a tight control of DNA integrity. At the end of filling and the onset of seed desiccation, the upregulation of genes implicated in sensing of DNA double-strand breaks suggests that genome integrity is challenged. The expression of chromatin remodelers seems to imply a concomitant action of chromatin remodelling with DNA repair machinery, maintaining genome stability. The expression of genes related to nucleotide excision repair and chromatin structure is evidenced during the desiccation stage. The results obtained support **Chapter II** findings, while adding new knowledge and molecular resources on relevant mechanisms acting during seed development.

The previous chapters characterized the proteomic and transcriptomic landscapes during seed development, highlighting a time-frame of molecular and metabolic events. Still, the molecular mechanisms controlling the observed temporal changes in gene expression remained to be understood. **Chapter IV** aimed at addressing this gap of knowledge and elucidate the role of post-transcriptional regulation mediated by miRNAs on

this process. For this, a high throughput non-coding transcriptomics methodology, small RNA Sequencing (sRNA-Seq), coupled with a degradome analysis and a target prediction algorithm were used to describe miRNAs and identify or predict, respectively, their targets. A total of 72 known miRNAs, from 25 miRNA families, and 39 new miRNAs were identified. These constitute the first comprehensive overview of miRNAs expression dynamics during seed development in this species. Most miRNAs were more abundant at 10 and 40 DAA, suggesting that late embryogenesis/early filling and desiccation were seed developmental stages in which miRNA action is more pronounced. Target analyses identified targets for 77 expressed miRNAs. While several known miRNAs were predicted to target *HD-ZIP*, *ARF*, *SPL* and *NF-Y* transcription factors families, most new miRNAs' targets were predicted to encode for functional proteins. MiRNAs more accumulated at early stages were implicated on regulating the end of embryogenesis, postponing the seed maturation program, storage compound synthesis and allocation. MiRNAs more accumulated later in development seem to have a role on seed quiescence, desiccation tolerance and longevity.

In conclusion, this PhD thesis produced an extensive number of novel knowledge and molecular resources that contributed to address the existing gap of knowledge on the molecular mechanisms underlying legume seed development, especially on this species. The herein identified proteins, transcripts and miRNAs were associated with, or implicated in, the regulation of several biological processes, such as nutrient allocation, which impact the final seeds' quality traits. Other identified molecules can have a potential role on seed viability and germination. Not only these novel molecular resources can be extended to other legume species and shed light in new lines of research, but they can also be used in new biotechnology-based approaches to improve legume seeds and meet the current and future food security challenges.

# Sumário

O desenvolvimento da semente é um dos processos mais relevantes em plantas, particularmente em culturas para produção de grão. As sementes são essenciais para as plantas, uma vez que asseguram o estabelecimento da próxima geração e a sobrevivência da espécie. Adicionalmente, são também importantes para o consumo humano e animal, pois representam uma importante fonte de nutrientes tendo, portanto, um papel relevante em termos agronómicos e socioeconómicos. O desenvolvimento da semente, apesar de ser um processo complexo onde várias moléculas e vias de regulação atuam concertadamente, tem sido pouco explorado na literatura, especialmente em Leguminosas. Esta é uma lacuna identificada no conhecimento científico e sobre a qual o trabalho descrito ao longo desta tese de doutoramento (Ph.D.) se debruçou.

*Phaseolus vulgaris* L., o feijão, é uma leguminosa com importância agronómica, dado a sua capacidade de fixar azoto atmosférico, e nutricionalmente relevante por ser uma boa fonte de proteínas, hidratos de carbono, fibras, vitaminas e minerais. O feijão é, efetivamente, um dos alimentos mais consumidos a nível mundial, tendo um papel fulcral na segurança alimentar. Em Portugal, o feijão é a leguminosa de grão mais consumida *per capita*. Vários estudos publicados sugerem o feijão como uma das melhores plantas modelo para estudar biologia de sementes. No entanto, previamente ao início dos estudos desta tese, o conhecimento sobre os mecanismos moleculares subjacentes ao desenvolvimento da semente de *P. vulgaris* era ainda limitado. Por exemplo, não existia uma descrição detalhada dos genes e das proteínas acumulados nas distintas fases de desenvolvimento da semente, desde o final da embriogénese até à dessecação. Estes dados são relevantes para a identificação dos mecanismos moleculares e das redes de regulação a atuar neste processo. Os estudos descritos nesta tese de doutoramento focaram-se na resposta a

várias questões científicas ainda por esclarecer nesta espécie: i) como mudam os parâmetros fenotípicos durante o desenvolvimento da semente e em que momentos temporais ocorrem as diferentes fases de desenvolvimento?; ii) quais são os principais mecanismos moleculares e vias metabólicas que caracterizam cada fase de desenvolvimento?; e iii) quais são os principais reguladores e vias de regulação com um potencial impacto nas características finais da semente?

De forma a dar resposta a estas questões, aplicou-se uma abordagem molecular recorrendo a tecnologias de alto-débito, que incluiu Proteómica e Transcritómica (codificante e não codificante), para descrever os mecanismos moleculares e regulação subjacentes ao desenvolvimento da semente em *P. vulgaris*. Assim, os objetivos destes estudos foram: 1) caracterizar as principais fases de desenvolvimento da semente de forma a definir os pontos temporais onde as análises moleculares serão realizadas; 2) identificar os principais mecanismos moleculares e vias metabólicas subjacentes ao desenvolvimento da semente; 3) caracterizar os mecanismos de manutenção da integridade do genoma durante este processo; e 4) compreender o papel dos mecanismos pós-transcricionais, mediados por miRNAs, que regulam o desenvolvimento da semente de *P. vulgaris*.

O objetivo do **Capítulo II** é fornecer uma descrição ampla dos perfis de abundância das proteínas durante o processo de desenvolvimento da semente. Para atingir este objetivo, uma metodologia de alto débito proteómica (LC-MS/MS) foi utilizada em sementes aos 10, 20, 30 e 40 dias após a antese (DAA), incluindo estádios desde o final da embriogénese até à dessecação da semente. Foram identificadas 418 proteínas cuja abundância foi alterada durante o processo de desenvolvimento. Uma acumulação de proteínas pertencentes às categorias funcionais MapMan “protein”, “glycolysis”, “TCA”, “DNA”, “RNA”, “cell” e “stress” foi observada, o que reflete uma grande atividade metabólica nas fases iniciais de

desenvolvimento. Durante as fases intermédias observou-se uma acumulação de proteínas relacionadas com o armazenamento, sinalização e síntese de amido, bem como de proteínas relacionadas com a parede celular. Na fase final de desenvolvimento observou-se um aumento de proteínas relacionadas com *redox* e, em menor extensão, degradação e modificação de proteínas e metabolismo dos ácidos nucleicos, refletindo uma ativação dos mecanismos de resistência à dessecação. Globalmente, foi descrito os principais mecanismos moleculares, as vias metabólicas e os perfis proteicos característicos e associados cada fase de desenvolvimento da semente. Além disso, foram obtidas pistas sobre o papel das modificações pós-tradução, tais como a nedilação, e nova informação sobre a manutenção da integridade do genoma.

Contrariamente ao observado na germinação, o conhecimento sobre os mecanismos implicados na manutenção da integridade do genoma durante o desenvolvimento da semente era limitado, mesmo incluindo os resultados obtidos no **Capítulo II**. O objetivo principal do **Capítulo III** é a caracterização dos mecanismos de manutenção de integridade do genoma durante o desenvolvimento da semente. Para isto, utilizámos uma metodologia transcritómica de alto-débito (*Massive Analysis of 3'-cDNA Ends* -MACE), juntamente com PCR digital (*Digital* PCR - dPCR), de forma a realizar uma descrição qualitativa e quantitativa, respetivamente, da expressão génica ao longo das fases de desenvolvimento da semente. Os principais resultados incluem nova informação sobre a manutenção da integridade do genoma nas sementes em desenvolvimento, incluindo mecanismos de reparação de dano e de remodelação da cromatina. Durante as fases iniciais, foi observada uma elevada expressão de genes relacionados com a deteção e reparação de dano no ADN o que sugere um controlo apertado da sua integridade. No final da fase de enchimento da semente e início da dessecação, foi observado um aumento na expressão de genes implicados na deteção de quebras da cadeia dupla do ADN, o que

sugere que a integridade do mesmo pode estar em causa. A expressão de genes que codificam remodeladores de cromatina sugere uma ação concomitante entre a remodelação e os mecanismos de reparação do dano para manter a estabilidade da cromatina. A expressão de genes relacionados com os mecanismos de excisão de nucleótidos e com a estrutura da cromatina foi evidenciada nas fases finais do desenvolvimento. Os resultados obtidos estão de acordo com as conclusões do **Capítulo II** e adicionam novos recursos moleculares e conhecimento sobre estes mecanismos de elevada relevância para o desenvolvimento da semente.

Os capítulos anteriores caracterizaram os perfis proteómicos e transcritómicos do desenvolvimento da semente, destacando os pontos onde ocorrem os principais eventos metabólicos e moleculares. No entanto, os mecanismos moleculares que regulam as alterações de expressão génica observadas são ainda pouco conhecidos. O **Capítulo IV** aborda esta lacuna no conhecimento através do estudo da regulação pós-transcricional da expressão génica mediada por pequenos ARN (do inglês, microRNAs – miRNAs) neste processo. Assim, foi utilizada metodologia de alto-débito para a análise do transcritoma não codificante, sRNA-Seq (*Small RNA Sequencing*), para caracterizar as famílias de miRNAs, combinado com uma análise do degradoma e de um algoritmo de predição de genes alvo para identificar ou prever, respetivamente, os seus alvos. Setenta e dois miRNAs previamente descritos (conhecidos), de 25 famílias, e 39 novos miRNAs foram identificados, constituindo a primeira caraterização detalhada realizada durante o desenvolvimento da semente de *P. vulgaris*. Foi observada uma maior abundância de miRNAs aos 10 e aos 40 DAA, sugerindo o final da embriogénese e a dessecação da semente como fases onde a ação dos miRNAs é mais pronunciada. Transcritos alvo para 77 miRNAs expressos foram identificados. A análise de predição bioinformática de alvos identificou genes codificando para fatores de transcrição de várias famílias, tais como HD-ZIP, ARF, SPL e NF-Y como possíveis alvos dos miRNAs conhecidos. Por outro lado, a mesma análise de predição indicou

como alvos dos novos miRNAs identificados no estudo, transcritos de genes que codificam para proteínas funcionais. Os miRNAs mais acumulados nas fases iniciais de desenvolvimento parecem estar implicados na regulação do fim da embriogénese, prorrogação do programa de maturação da semente e na síntese e alocação de compostos de reserva. Os miRNAs mais abundantes nas fases finais de desenvolvimento têm um potencial papel na quiescência, tolerância à dessecação e longevidade das sementes.

Em suma, esta tese de doutoramento produziu uma vasta quantidade de novos conhecimentos e recursos moleculares que contribuíram para preencher a lacuna no conhecimento existente sobre os mecanismos moleculares subjacentes ao desenvolvimento de sementes de Leguminosas, nomeadamente em sementes de *P. vulgaris*. As proteínas, os transcritos e os miRNAs identificados foram associados ou implicados na regulação de processos biológicos relevantes, tais como a alocação de nutrientes, aspetos com impacto nas características finais ou qualidade das sementes. Outras moléculas identificadas poderão ter um papel ao nível da germinação e da viabilidade das sementes. Estes são recursos fulcrais que podem ser ainda transferidos e aplicados a outras espécies de Leguminosas e providenciam informação relevante para diferentes linhas de investigação futuras. O conhecimento e recursos moleculares obtidos podem ainda ser utilizados para o melhoramento de características da semente desta e outras leguminosas, recorrendo a abordagens baseadas em biotecnologia, visando responder aos atuais e futuros desafios de segurança alimentar.




# General Introduction

**Author's contributions to the chapter:**

José Ricardo T. D. Parreira Salvado wrote this chapter based on the referred bibliography.

# Seed development in *Phaseolus vulgaris* L.: a key process reviewed



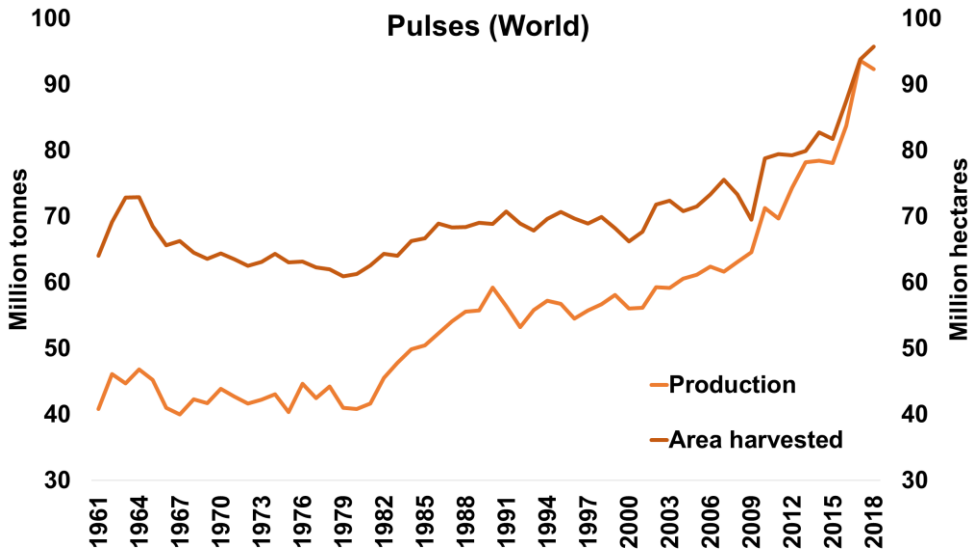
## 1. The Fabaceae: a family of agricultural importance

The Fabaceae is the third-largest vascular plant family, with almost 20,000 species [1]. This is an important family due to its high nutritional value and the capacity to perform atmospheric nitrogen fixation, in root nodules, in symbiosis with soil bacteria such as from *Rhizobium* and *Bradyrhizobium* families [2]. Plants from Fabaceae, commonly known as the legume family, can also be used in the production of important compounds for industrial and pharmaceutical purposes. Included in this family are several species with agricultural importance, such as *Arachis hypogaea* L. (peanuts), *Cicer arietinum* L. (chickpeas), *Glycine max* (L.) Merrill (soybean), *Pisum sativum* L. (peas), and *Phaseolus vulgaris* L. (common bean). Other species have high relevance for scientific research, including *G. max*, *Medicago sativa* L. (alfalfa), *Medicago truncatula* L. (barrel medic), and *Lotus japonicus* (Regel) K. Larsen.

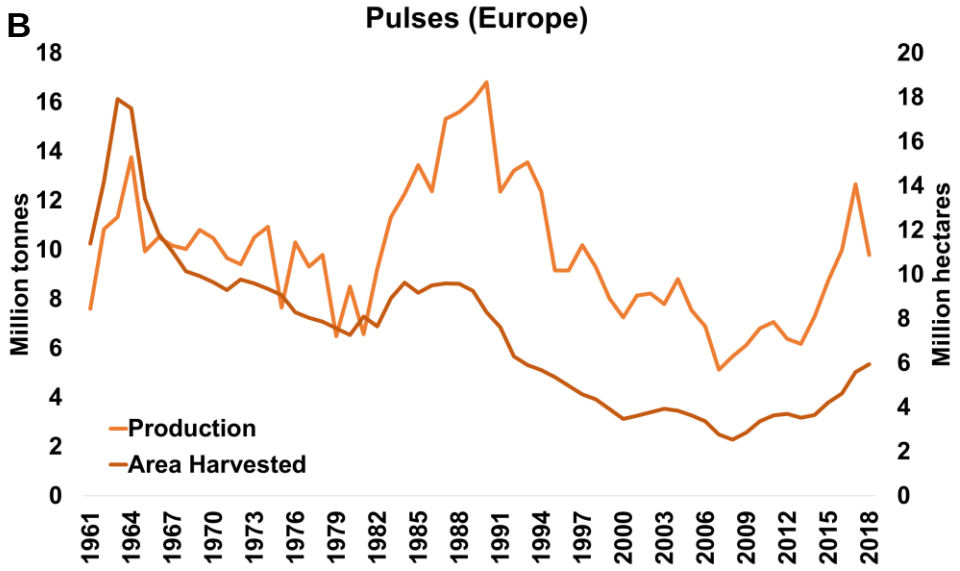
Grain legumes, also called pulses, are annual crops from the Fabaceae family which are normally harvested for their grain to be used for food and/or feed [3]. Leguminous crops play important roles within the agricultural systems such as their ability to fix nitrogen symbiotically and, by having a diversifying effect, reducing the requirement for pesticide use [4,5]. Furthermore, pulses represent an important source of proteins, complex carbohydrates, dietary fibre, vitamins, and minerals for human nutrition [6]. Indeed, some species, including *Vicia faba* L. (faba beans), *Lens culinaris* Medik (lentils), and chickpeas, are deeply rooted within the agronomic and culinary culture of Asian and European populations [7].

While worldwide pulses production and harvested area has seen an increase since 1961 (**Figure 1A**), the same pattern was not observed at the European (**Figure 1B**) or national (Portugal, **Figure 1C**) levels.

**A**




**B**




**C**

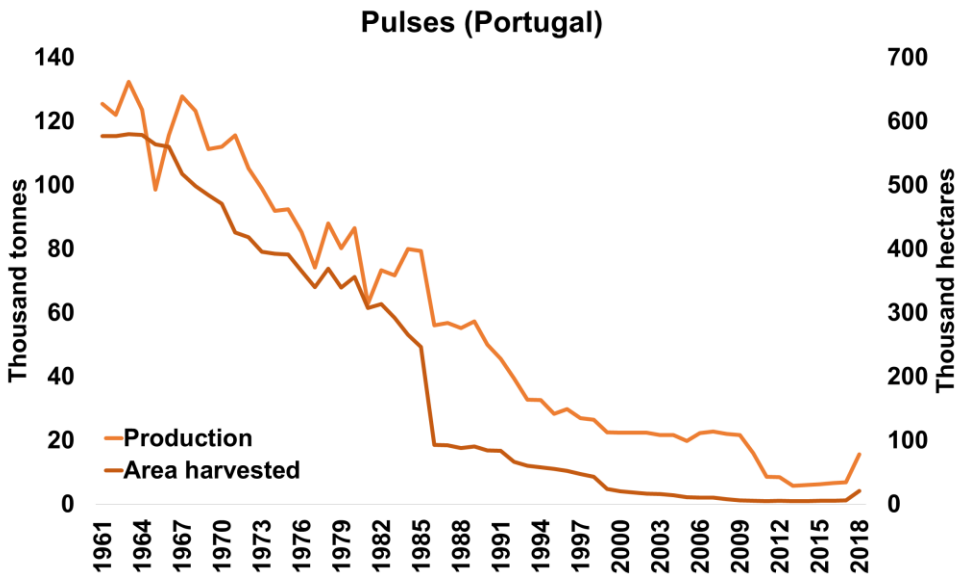
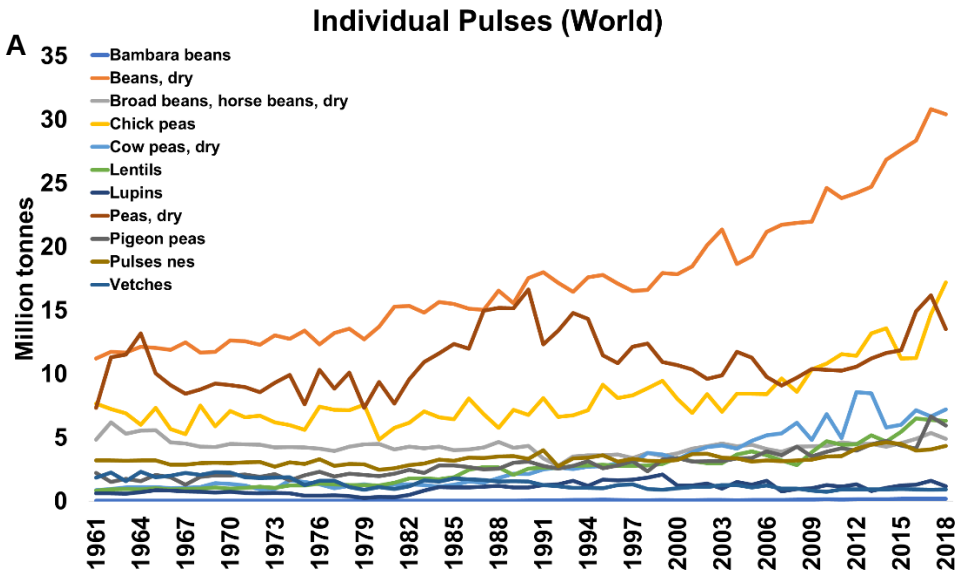
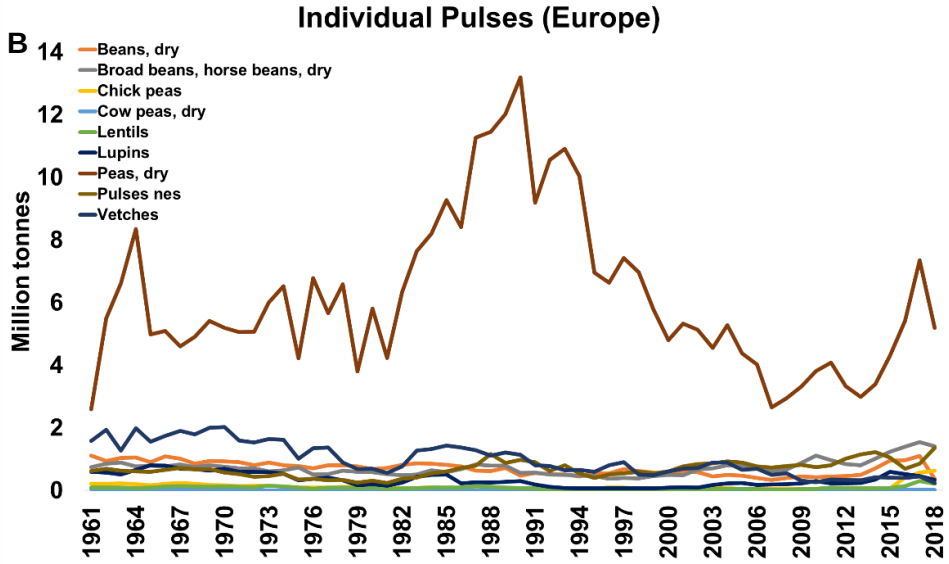



**Figure 1** Chronological evolution of production and harvested area of pulses in the World **(A)**, Europe **(B)**, and Portugal **(C)** from 1961 to 2018. Values were retrieved from FAOSTAT, accessed on the 1st December of 2020 (A) and 7th March 2021 (B and C).

According to the Food and Agriculture Organization of the United Nations (FAO), the *Phaseolus* spp. has the highest worldwide production within pulses, estimated at 30.43 million tonnes in 2018 (**Figure 2A**). Chickpeas are the second-highest produced pulse in the world with roughly half of the production (17.19 million tonnes), followed closely by peas (13.53 million tonnes). Over the recent years, *Phaseolus* spp. production has increased worldwide (**Figure 2A**). Contrastingly, when looking at the European level, a prevalence in peas production can be observed (**Figure 2B**), while in Portugal a decrease in pulses production can be observed from 1961 (**Figure 2C**).


A
Individual Pulses (World)
Million tonnes
Bambara beans
Beans, dry
Broad beans, horse beans, dry
Chick peas
Cow peas, dry
Lentils
Lupins
Peas, dry
Pigeon peas
Pulses nes
Vetches
0
5
10
15
20
25
30
35
1961
1964
1967
1970
1973
1976
1979
1982
1985
1988
1991
1994
1997
2000
2003
2006
2009
2012
2015
2018


B
Individual Pulses (Europe)
Million tonnes
Beans, dry
Broad beans, horse beans, dry
Chick peas
Cow peas, dry
Lentils
Lupins
Peas, dry
Pulses nes
Vetches
0
2
4
6
8
10
12
14
1961
1964
1967
1970
1973
1976
1979
1982
1985
1988
1991
1994
1997
2000
2003
2006
2009
2012
2015
2018

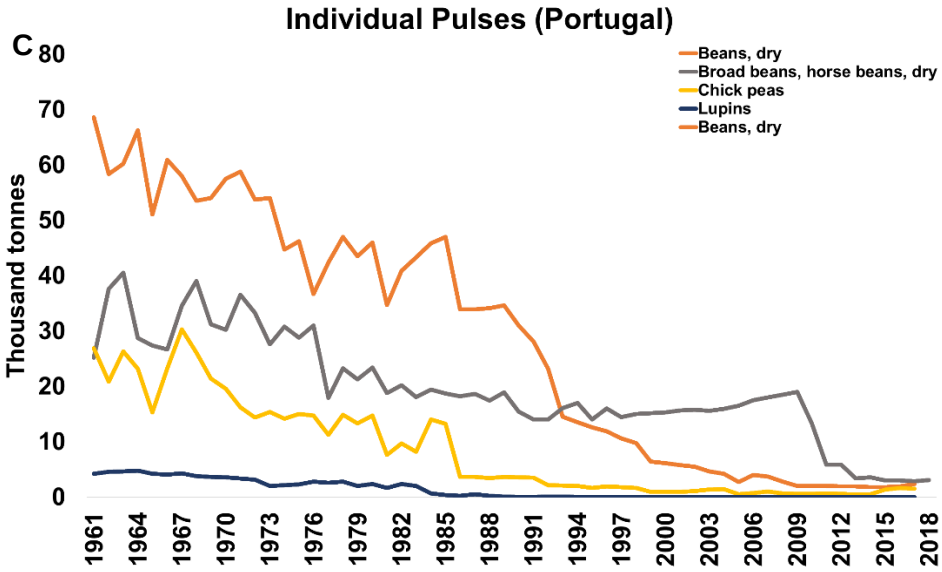



**Figure 2** Chronological evolution of production of pulses in the World **(A),** Europe **(B)** and Portugal **(C)** from 1961 to 2018. Values were retrieved from FAOSTAT, accessed on the 1st December of 2020 (A) and 7th March 2021 (B).

## 1.1 *Phaseolus vulgaris* L., the common bean

*Phaseolus vulgaris* L. (2n = 22), the common bean, is a mainly self-pollinated herbaceous annual plant, grown worldwide for its edible green pods and dry seeds. Within the Fabaceae family, this pulse has also an important agronomical role, due to its capacity for fixing atmospheric N and, nutritionally, being a good source of proteins, carbohydrates, dietary fibres, vitamins, and minerals [8]. Even though the common bean is one of the most known pulses within the agronomic and culinary culture and is one of the most consumed grain legumes in Europe, it was only recently introduced following the arrival of the first Europeans in America [3].

The origin of the common bean is the Mesoamerican region and, with the expansion to South America, seems to have diverged in two distinct geographic gene pools, Mesoamerican and the Andean, where it was independently domesticated [9]. Both gene pools hold a variety of landraces differing in terms of seed and leaf sizes, growth habits, seed coat colour, and

patterns. For instance, the Mesoamerican germplasm has small/medium seed sizes and S or B phaseolin patterns (the major seed storage proteins of the common bean) [10]. In contrast, seeds from the Andean germplasm are predominantly large and have T, C, H, or A phaseolin patterns [10]. In Portugal, a study of 175 accessions from Portuguese mainland and islands traditional bean-growing regions was performed [11]. The authors used molecular markers (microsatellites and DNA marker for phaseolin-type diversity analysis), and performed seed and plant morphological characterization [11]. The performed genetic structure analysis showed three main clusters, two more related to the Andean gene pool while the third was more related to the Mesoamerican gene pool. While most of the Portuguese germplasm grouped with the Andean region race representatives and wild relatives, one-third had an admixed origin and might represent putative hybrids between both original gene pools [11].



An interesting *P. vulgaris* line is the small red (SER) 16 (**Figure 3**), developed at the International Center for Tropical Agriculture (CIAT) in Cali (Colombia). From the Mesoamerican gene pool, this line is described as drought-resistant, short bush habit, and productive branching with excellent combining ability (readily transmitting its characteristics to the progeny) [12,13]. Indeed, SER 16 showed superior seed yield and pod harvest index under rainfed conditions [12] and to have the highest yield both in irrigated and terminal drought conditions, when compared with other drought-tolerant lines [14]. In this study, and under irrigated conditions, SER 16 showed a least square mean of 3364 kg/ha, decreasing to 1928 kg/ha under terminal drought, and a high partitioning and remobilization capacity [14]. All research performed and described in this thesis was performed with the SER 16 line.


**Figure 3** *Phaseolus vulgaris* SER 16 line in a growing chamber **(A)**, open flowers **(B)**, a pod at 10 days afters anthesis **(B)**, collected mature seeds **(C)** and a detail of a seed at late maturation **(C)**.

### 1.1.1 Production and nutritional value

As previously stated, a worldwide increase in *Phaseolus* spp. production can be observed, with the highest worldwide production within pulses, estimated at 30.43 million tonnes in 2018 (**Figure 2A**). The top producers in 2018 were India, Myanmar, and Brazil (FAOSTAT, accessed on the 1st December of 2020). Indeed, the common bean seeds are one of the most consumed staple foods worldwide, being crucial for nutrition, food security, and as an income source, particularly for smallholder farmers [15–17]. Interestingly, it's the most consumed grain legume per capita in Portugal, between 2019 and 2020 (Instituto Nacional de Estatística, available at www.ine.pt), still the 3000 tonnes of common bean produced cover only 6.7% of the country's demand (Estatísticas Agrícolas 2018, available at www.ine.pt).

Nitrogen ($N_2$) fixation is a key source of N for farmers, especially those that use little or no fertilizer [18]. Some common bean cultivars were shown to accumulate almost 100 kg of atmospheric nitrogen per hectare

when inoculated with rhizobia [19]. This is a similar value to what commercial producers apply of N, thus revealing a potential to decrease farmer costs when using rhizobia inoculation for *P. vulgaris* production [19].

Regarding the common bean protein content, it differs between wild, weedy, and cultivated varieties, with some reports showing a range between 166 g/Kg up to 330 g/Kg [20]. Common beans supply approximately 8% to 10% of the daily recommended allowance of protein per 100 g serving of cooked beans [8]. Of notice, even though the common bean is considered a high source of protein, it is limited in terms of concentration of sulphur amino acids, methionine, and cysteine, which are important in human nutrition [21]. Common bean seeds also contain high amounts of carbohydrates (~54g per 100g) and dietary fibre (~20g per 100g), the latter important for, among others, gastrointestinal transit time and nutrient bioavailability [22]. Furthermore, minerals, vitamins, and a rich variety of bioactive and/or antinutritional molecules, such as phytate, oligosaccharides, phenolic compounds, nonprotein amino acids, lectins, and enzyme inhibitors are also found in the dry seeds [8,23–25]. Indeed, consumption of beans is suggested to be associated with several health benefits, including anti-oxidant, anti-diabetic, anti-obesity, anti-inflammatory, anti-mutagenic, and anti-carcinogenic properties [8,26,27].



### 1.1.2 *P. vulgaris*, a potential model species

*Phaseolus vulgaris* has been suggested as a model species within the legume species [25,28]. By browsing PubMed in the National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov/) a recent increase of published research articles in which *P. vulgaris* studies were conducted was observed (**Figure 4**). A total of 6538 research articles were retrieved from the database until 2020 (as of 1st of December 2020). While this value falls short of the 7247 research articles found for the model *G. max*, it is more than double of those found for other model species *M. sativa* (2740) and *M. truncatula* (2592). Several resources, including genetic and genomic,

were made available and are, in part, responsible for the increase observed in published research. Interestingly, the first record for common bean dates to 1916, while for *G. max* it was almost 50 years later.

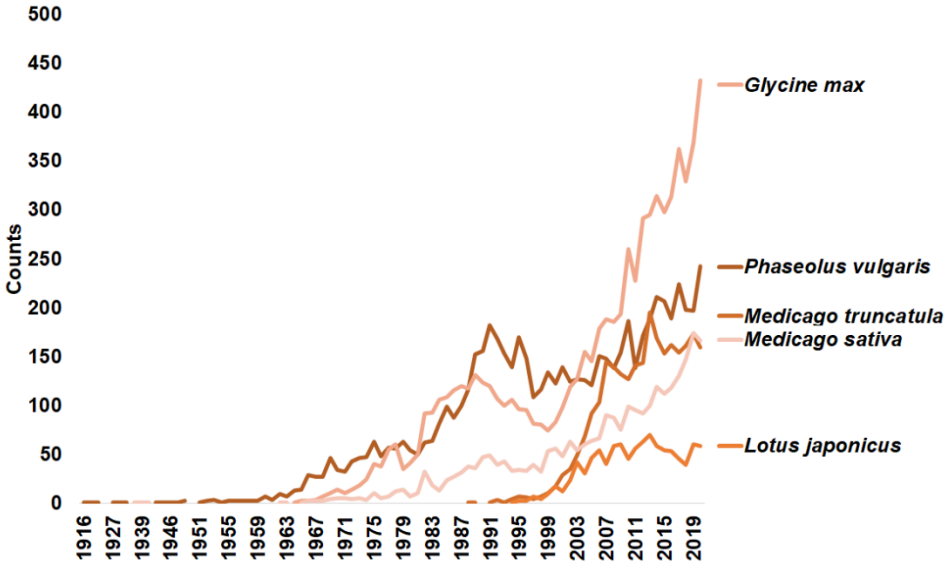



**Figure 4** Query of research scientific papers for *Phaseolus vulgaris*, *Medicago truncatula*, *Medicago sativa, Lotus japonicus,* and *Glycine max* retrieved from PubMed, published between 1916 and 2020. The search was performed individually for each species using the following settings in advanced search: [species name (Title/Abstract)] NOT review (Publication type). Review manuscripts were not included. Search conducted on the 1st of December 2020.

During the last decade, several important research advances were accomplished. In 2011, the first large-scale transcriptome study in common bean derived from 454 pyrosequencing from root, leaf, pod and flower samples was made available [29]. The authors aimed to provide valuable resources for the future annotation of the common bean genome and were able to increase by 150% the number of *P. vulgaris* expressed sequence tags (ESTs) available until that point.

Indeed, one of the hallmarks of scientific research in any species is the release of the genome sequence, and this was first accomplished for *P.*

*vulgaris* in 2014 [30]. Currently, the genome and annotation have reached version 2.1, available at Phytozome 13 (https://phytozome-next.jgi.doe.gov/). The new version combines a new PacBio-based assembly of the genome with a reannotation using the Gene Model Improvement pipeline and presents 27,433 total loci, containing 36,995 protein-coding transcripts. Since the beginning of this thesis other genome sequences were made available, such as the line BAT93 [31] (available at https://denovo.cnag.cat/bean), as well as the cultivar UI111 (Phytozome 13, available at https://phytozome-next.jgi.doe.gov/).



The release of the genome sequence and annotation allowed the development of other resources, such as the RNA-Seq based gene expression atlas of the common bean [32]. Expression atlases provide added value for future studies, allowing the query for genes or sets of genes of interest and exploration of their expression in different contexts, such as tissue, developmental stage, or other biological processes. In this study [32], the authors analysed samples from roots, nodules, stems, flowers, leaves, pods, and seeds at different stages of development. The expression profiles allowed the study of biological processes related to seed and pod development, nodulation, and responses to N availability. For instance, the authors were able to increase the knowledge about nitrogen metabolism in the species. They observed an increased expression of ARFs in the leaves of N deficient plants, suggesting increased auxin levels supporting its role as a N status mediator in the common bean [32]. The datasets obtained in this study were later used in the first *Phaseolus vulgaris* transcription factor database [33]. This database contains 2,370 putative TF gene models in 49 families and including sequence data, functional annotation, and gene ontology, among others.

Not only does the genome sequence and annotation allow for expressed protein-coding genes studies, but it also facilitates the analysis of gene expression regulatory mechanisms. In 2012, Peláez *et al.* published

what would represent the first massive-scale sequencing study to identify and characterize the miRNA population in *P. vulgaris* [34]. The authors generated small RNA libraries from roots, flowers, leaves, and seedlings using the Genome Analyzer II and Illumina Cluster Station. A total of 109 conserved miRNAs, from 29 families, were identified. Like for other legume species, members of the miR159, miR319, miR156, miR166, among others, were found as the most expressed. A total of 29 candidates for new miRNAs were identified, based on the small RNA reads and precursor prediction [34]. Later, a genome-wide identification of small RNAs in *P. vulgaris* using sRNA-Seq and degradome sequencing was published in 2015 [35]. The authors constructed and sequenced libraries from seedlings, roots, leaves, flowers, and nodules, using a Genome Analyzer IIx and identified 185 mature miRNAs, including novel miRNAs present only in common bean, as well as a total 181 targets from the degradome data. The authors built weighted correlation networks of miRNAs and provided, among others, new knowledge on post-transcriptional regulatory mechanisms controlling nodulation. On the same year, 2015, de Sousa Cardoso *et al.* published a report on the processing pathway genes and an *in silico* characterization of *P. vulgaris* miRNAs and their targets using the recently released genome [36]. Regarding the small RNA processing pathway, the authors identified 68 proteins putatively involved in the process, including 17 from the ARGONAUTE family and 6 DICER-LIKE proteins. A total of 221 mature miRNAs from 52 families were identified, including 129 new miRNAs. The target analysis, using the tool psRNATarget (available at http://plantgrn.noble.org/psRNATarget/home), predicted 483 targets, confirming some previously reported results, such as the ARF TFs being putatively targeted by MIR160 in *P. vulgaris* [36]

Even though plenty of genomic and genetic resources exist, research in *P. vulgaris* still faces a critical drawback: a reduced number of tools for functional gene validation. An efficient and reproducible transformation

system to obtain stable transgenic common bean plants is still lacking, even though substantial efforts have been made to develop methodologies for common bean genetic engineering [37–40]. The primary limitation for this seems to be *P. vulgaris* recalcitrance towards *in vitro* regeneration and rooting [41]. Still, limited success has been achieved since 2014. The induction of somatic embryogenesis from the apical meristem and cotyledonary nodes of zygotic embryos was accomplished [42]. The authors combined osmotic stress, a cytokinin (BAP), and an adenine-free base to induce somatic embryos from different common bean genotypes. The highest regeneration efficiency was observed for Negro Querétaro with only 25% of the embryogenic calli giving rise to plants. Similarly, to the observed in other species, this methodology is genotype-dependent. Recently a new publication on the regeneration of *P. vulgaris* from epicotyls and hypocotyls via direct organogenesis was published [43]. The authors used epicotyls and hypocotyls of six cultivars and four variants of two basal media (Murashige-Skoog and Gamborg). This work supported the hypothesis that common bean regeneration is cultivar-specific, with the observed values for the epicotyl response to be between 70% for Czerwona and 46.11% for Złota Saxa.

Regarding genetic transformation protocols for *P. vulgaris*, some have shown promise for the future of the species as a model plant. These include particle bombardment for cultivars Olathe and Carioca tolerance to the herbicide glufosinate ammonium [38] or RNAi-mediated resistance of cv. Olathe to bean golden mosaic virus [44]. Other protocols were developed, such as the use of indirect organogenesis to perform an *Agrobacterium*-mediated transformation of the var. CIAP7247F [37], or the transient transformation of leaf mesophyll protoplasts from cv. Negro Jamapa [39]. More recently, Song *et al.* (2020) [40] were able to produce stable transgenic common bean plants through *A. tumefaciens*-mediated transformation of cv. Olathe. While the authors observed a low transformation frequency between 0.5% and 1.1% when using embryo axes, when these were precultured on

regeneration media, the efficiency increased to 1.5 to 2.5%. This increase in efficiency might be due to an increase in competent cells prior to the infection. The authors also observed that a low concentration of *A. tumefaciens* increased the inoculated explant survival.

One interesting advancement in 2014, an efficient viral vector system was developed, using *Bean pod mottle virus* (BPMV), that allows expression of foreign genes, as well as virus-induced gene silencing (VIGS) in the common bean [45]. While efficient, the method has several limitations such as it requires virus-susceptible cultivars, for instance, the cv. Black Valentine, for a systemic infection to occur. Another limitation is the inability of the virus to penetrate the seed coat, impairing seed-specific gene expression studies.



While slow, these efforts seem to be paying off with an increase in stable transformation frequency. Nonetheless, tissue culture of *P. vulgaris* has repeatedly been found to be difficult, still hampering the application of genetic engineering studies.

## 2. Seed development

Seed development is one of the most relevant developmental processes in seed-producing crops. Seeds represent the next generation in plants, but they are also essential for animals, as a source of nutrients, being used as food/feed, and playing important agronomic and socio-economic roles. Seed development is a complex process that impacts seed storage composition and size, thus, the final quality traits. The knowledge of the molecular processes and regulatory networks occurring during this complex developmental process is particularly relevant to improve seed quality and yield.

Seed development is usually divided into three developmental stages in orthodox (desiccation-tolerant) seeds: embryogenesis, maturation/filling,

and desiccation [46]. A schematic representation of the seed developmental process is provided in **Figure 5.**

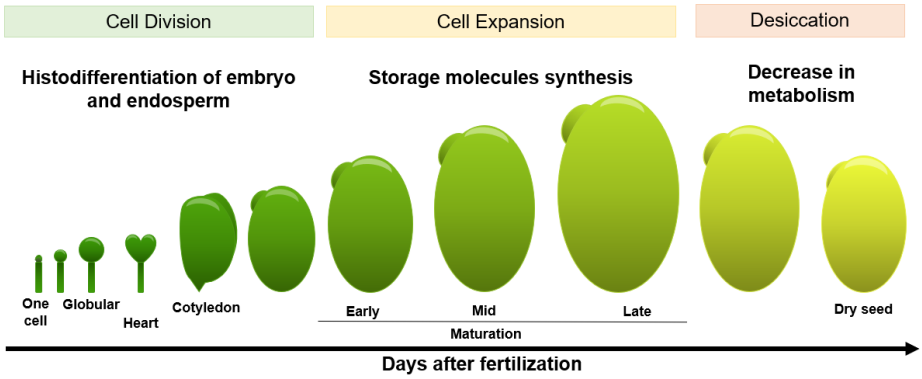



**Figure 5** Schematic representation of seed development in a legume species. Embryo morphologies and developmental events adapted from [46,47].

In angiosperms, it is during double fertilization that one sperm cell nucleus fuses with the egg cell nucleus to form the diploid zygote [48]. In parallel, the nucleus from the other sperm cell fuses with the two polar nuclei (or the two fused central diploid nucleus) to form a triploid nucleus, from which the endosperm develops [48]. After fertilization, successive cell divisions and differentiation processes allow the embryo to develop from the zygote, and the endosperm to form [48–50]. After the embryo is formed, maturation begins. Embryo cells, especially those of the cotyledons, start accumulating proteins, lipids, and carbohydrates that will provide energy and basic building blocks for germination and seedling growth, as well as desiccation tolerance [48,50]. During this developmental stage, a rapid increase in embryo mass and size can be seen. Lastly, desiccation occurs in orthodox seeds, leading to a metabolically quiescent state where protection mechanisms are activated and the accumulation of biomass ceases [48].

## 2.1 Molecular players and regulatory networks regulating seed development

The final seed traits depend on the timely accumulation of storage molecules during this developmental process which, in turn, is influenced by the genotype and adaptive changes to the environment [51]. Considering the potential improvement and modulation of seed traits, it is essential to identify the molecular players and regulatory networks acting during the developmental process.



Most published studies on these topics have focused on model species, such as *Arabidopsis thaliana* [52,53] but some examples can be found for agronomically important crops, such as *Oryza sativa* (rice) [54] and *Zea mays* (maize) [55]. Several studies in Fabaceae species were also published that focused on the molecular mechanisms and regulatory networks acting during seed development. A brief description of these reports will be provided on the next sections.

### 2.1.1 Proteomics and transcriptomics-based studies

Several studies, using transcriptomics and proteomics, made significant contributions to our knowledge about the molecular mechanisms implicated in seed development. In 2003, a gel-based proteomics approach, two-dimensional gel electrophoresis (2-DE) coupled with a MALDI-TOF, was used to profile and identify proteins during seed development in *Medicago truncatula* [56]. This study constitutes the first proteomics report that focused on protein profiles during seed filling in a legume species. To provide a framework for the proteomics study, the authors performed characterization of the seeds' fresh and dry weight, as well as water content, and linked to the different stages of seed development: from 8 to 12 days after pollination (DAP) corresponding to histodifferentiation, from 12 to 36 DAP to seed filling and from 36 to 44 DAP to desiccation [56]. Regarding the protein abundance profiles, the authors characterized the proteome of seeds harvested at five

timepoints during seed filling: seeds with low dry weight and unable to germinate (12 DAP) and seeds increasing in dry weight and developing the ability to germinate (14, 16, 18, and 20 DAP). The authors observed proteins involved in cell division, such as β-TUBULIN and ANNEXIN, to be more accumulated at the end of embryogenesis (12 DAP). Interestingly, a turnover was observed for these proteins' abundances reflecting a developmental switch from embryogenesis to seed filling, from 12 to 14 DAP. Storage proteins accumulated in specific timepoints during seed filling: vicilins (14 DAP), legumins (16 DAP), and convicilins (18 DAP). The most abundant proteins were 7S (VICILIN and CONVICILIN) and 11S (LEGUMIN) globulins. An accumulation of proteins associated with cell expansion and enzymes, such as SUCROSE SYNTHASE and STARCH SYNTHASE, were also observed during early seed filling, suggesting that these enzymes might be involved in the supply of carbon substrates for the synthesis of storage compounds [56].



Since then, more insights were provided regarding the molecular players behind this complex developmental process. An integrative transcriptomic (microarrays) and proteomic (2-DE) approach was published, focusing on *M. truncatula* seeds during seed filling (12 to 36 DAP), [57]. Around 77% of the protein spots and 50% of the transcripts changed their abundance more than 2-fold during the time course of the experiment. The authors suggest that, when compared with Arabidopsis, their results hinted at a more complex network of gene expression during the development of legume seeds. In addition, the authors analysed tissue-specific responses from the seed coat, endosperm, and embryo at the onset of seed filling (14 DAP), which constitutes the first tissue analysis reported for a legume species. This analysis revealed a complex compartmentalization of the amino acid metabolic pathways, such as methionine metabolism, which might impact the availability of sulphur-containing amino acids for embryo protein synthesis during seed filling [57].

In 2013, Verdier *et al.* studied the molecular events occurring from late seed filling to desiccation in order to understand the acquisition of desiccation tolerance (DT) and longevity mechanisms in *M. truncatula* seeds [58]. In this study, the authors used a combination of physiology and transcriptomic approaches to build a putative genetic regulatory network with different modules (subnetworks of gene interactions) related to the acquisition of these traits. The acquisition of DT module includes genes more expressed at 36 and 40 DAP, corresponding to the late maturation. Interestingly, in this module revealed most seed-specific genes, and was enriched in genes with Gene Ontology (GO) terms related to response to stresses (e.g. response to desiccation, heat, and abiotic stimulus), including late embryogenesis abundant (LEA), heat shock proteins, and abscisic acid-induced genes [58].



Within the crop species from the Fabaceae family, *Glycine max* has also been studied regarding the molecular mechanisms acting during seed development. Microarrays were used to disclose the expression dynamics of genes at 5 different seed development stages in soybean, ranging from histodifferentiation to the desiccated seed [59]. Cluster analysis of gene expression profiles revealed a timeframe for transcript accumulation during this developmental process. Indeed, several genes related to cell growth and energy processes were found more expressed at earliest developmental stage, such as *SUCROSE SYNTHASE*, *TUBULIN* and *ANNEXIN* [59]. The authors also observed an increase in expression, from the first stages studied to mid-maturation, of several genes encoding storage proteins, including glycinin genes, a major storage protein of soybean. Many genes related to protein degradation, such as ubiquitin-conjugating enzymes, proteases, and proteasome regulatory subunits, were found more expressed at the final stages of seed development. The authors proposed that the products of these genes may be useful for breaking down proteins no longer needed as the seed prepares for quiescence [59].

In another study, Jones *et al.* used RNA-Seq to analyse seven timepoints, ranging from embryogenesis to desiccation, during soybean seed development [60]. As before [59], a cluster analysis of gene expression profiles revealed a timeframe for gene expression during this developmental process. Genes encoding for storage proteins, like glycinin, showed the highest expression at the stages of largest fresh weight, validating the observation previously made [59]. Worth to highlight that soybean is a highly valuable crop not only because of its high protein contents in seeds, but it also accumulates large amounts of oil. Some genes, such as omega-6-fatty acid desaturase genes, involved in the conversion of oil to more highly desaturated forms, had higher expression at the stage of largest fresh weight. The authors found genes encoding for hydrophilic proteins, such as LEA proteins and dehydrins more expressed at the final dry whole seed, with a potential function on the preservation of the cellular structures during desiccation [60].



Overall, these reports provided an overview of the molecular players' accumulation patterns at specific time-points during seed development in important legume species.

### 2.1.2 Regulation of gene expression

Regulation of gene expression ensures that genes are expressed at the appropriate times, and the previously described research provided some clues. These clues included transcription factors (TFs) acting in *M. truncatula* and in *G. max*. In the following sections, regulation of gene expression mediated by transcription factors and miRNAs will be briefly described and key players/regulatory networks will be highlighted.

### 2.1.2.1 Transcriptional regulation of gene expression mediated by transcription factors

Regulation of gene expression, such as transcriptional regulation mediated by transcription factors (TF), plays an essential role during the seed developmental process [61]. From the previously described studies, in *M. truncatula* a seed-specific co-expression network showed that the transcriptional regulation of LEA genes was tightly connected to identified known and putative transcriptional regulators of DT, including *ABSCISIC ACID-INSENSITIVE3* (*ABI3*), *ABI4,* and *ABI5* [58]. In soybean, Jones and Vodkin [60] observed several transcription factor families to be highly expressed, reflecting the complex network of regulation of gene expression essential for this developmental process. For instance, members of the NAC family were found with a peak expression at the dry seed stage and are possibly involved in the desiccation response. Still, other studies were performed and will be briefly described.



The first real-time quantitative PCR expression profiling of 712 *M. truncatula* genes encoding putative TFs was published in 2008 [62]. The authors analysed seven time-points of seed development, ranging from late embryogenesis (10 DAP) to the onset of desiccation (36 DAP) and obtained 169 differentially expressed putative TFs. A cluster analysis allowed the association of TFs with developmental stages. For instance, TFs upregulated just before storage protein synthesis included candidate orthologues of *LEAFY COTYLEDON1-LIKE* (*L1L*) and *CUP-SHAPED COTYLEDON LIKE*. The authors combined the data with a microarray-derived transcriptome dataset to identify possible TF targets. A model for the transcriptional regulation of vicilin genes was constructed. In this model, possible *M. truncatula* orthologues of members of the LAFL (LEAFY COTYLEDON1 (LEC1), ABI3, FUSCA3 (FUS3) and LEAFY COTYLEDON2 (LEC2)) regulatory network were identified and a timeframe of their expression disclosed.

Indeed, the members of the LAFL regulatory network include master transcriptional regulators that interact, forming distinct complexes, to promote embryogenesis and seed maturation, regulate storage compound synthesis, and prevent precocious germination [63–65]. These transcriptional regulators include LEC1, a member of the NF-YB protein family (NF-YB9) that has a closely related homolog L1L (NF-YB6), and also ABI3, FUS3, and LEC2, that belong to the plant-specific family of B3 domain TFs and are named “AFL-B3” regulators [65]. Several functions have been described for this regulatory network and its members and will be briefly described.

LEC1 is considered a central regulator of seed development due to its multifunctionality during this process (for a review see Jo *et al*., (2019) [66]). This TF is involved in the acquisition of desiccation tolerance, storage macromolecule accumulation, seed dormancy, and suppression of vegetative development [66]. Still, other roles have been proposed for this regulator during seed development, such as maintenance of the embryonic suspensor identity, specification of cotyledon identity, and regulation of gene expression involved in embryo morphogenesis, including *PHAVOLUTA, PHABULOSA,* and *SCARECROW* [66]. As LEC1 role seems to extend beyond the control of the maturation phase, its expression seems to be tightly controlled. In that regard some regulators have been described, such as PICKLE (PKL) and VIVIPAROUS ABI3-LIKE (VAL) proteins that act at the epigenetic level [66].

The “AFL-B3” regulators' roles are also essential during seed development and are involved in the regulation of other transcription factors, as well as regulation of seed storage proteins or oil and fatty acid synthesis (for a review see [67]). Interestingly, some of these TFs have partial overlapping and synergistic functions. For instance, *FUS3* and *ABI3* expression is controlled by partial redundant regulations involving LEC1 and LEC2 and, also by positive regulatory feedback loops by FUS3 and ABI3

(**Figure 6**) [67]. Other specific functions have also been described, such as the involvement of ABI3 in chlorophyll degradation in late-maturing seeds, or the involvement of LEC1, LEC2, and FUS3, but not ABI3, in the control of embryo identity, and a loss of their function induced the formation of true leaves in embryos [67].



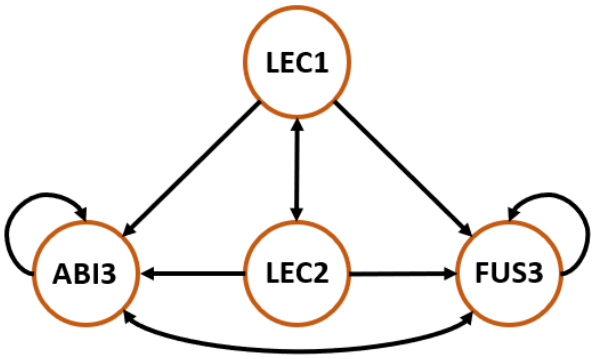



**Figure 6** Simplified schematic representation of the LAFL (LEAFY COTYLEDON1 (LEC1), ABSCISIC ACID-INSENSITIVE3 (ABI3), FUSCA3 (FUS3) and LEAFY COTYLEDON2 (LEC2)) regulatory network based on [67]. Arrows indicate activation of expression.

Due to the complex mutual interactions among the LAFL regulatory network members, the network is neither strictly hierarchical nor linear [68]. A review of this complex regulatory network, which is tightly controlled on multiple levels such as epigenetic, transcriptional, post-transcriptional, and post-translational, is available [65]. For instance, post-transcriptional regulation of *LEC2* and *FUS3* gene expression is performed, although not directly, via miRNA156 and miRNA166 [67].

### 2.1.2.2 Post-transcriptional regulation of gene expression mediated by microRNAs

The combination of miRNAs and TFs seems to allow a tight regulation over gene expression at both temporal and spatial scales, including in the seeds [69,70]. Indeed, small non-coding RNAs, including miRNAs, play a

pivotal role in this process (for a review see [70]). Thus, regulation of gene expression at the transcriptional and post-transcriptional level is hypothesized to be crucial for a successful seed development.

MiRNAs are endogenous small non-coding RNAs (~21-22 nucleotides) that act as negative post-transcriptional regulators of gene expression, either by transcript cleavage or translation repression [72]. **Figure 7** depicts a schematic representation of miRNAs biogenesis in plants.



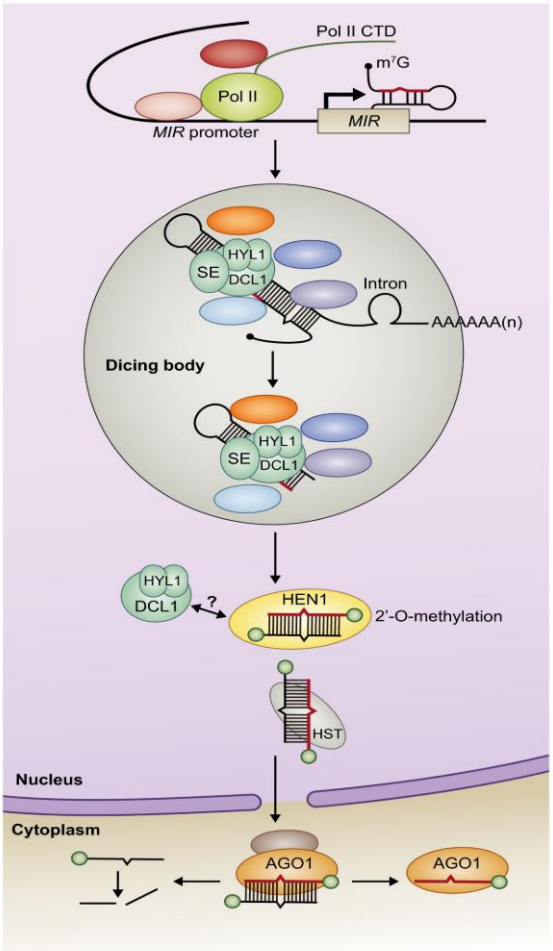



**Figure 7** Schematic representation of miRNAs biogenesis. RNA polymerase II (Pol II)-mediated miRNA gene (*MIR*) transcription is regulated via multiple TFs. Pol II activity itself is also subjected to phospho-regulation at its C-terminal domain (CTD). miRNA precursors are processed at the dicing bodies by the dicing complex, which is mainly composed of DICER-

LIKE 1 (DCL1), HYPONASTIC LEAVES 1 (HYL1) and SERRATE (SE). Probably the dicing complex interacts with HUA ENHANCER 1 (HEN1) and contributes to miRNA/miRNA* duplex export and RNA-induced silencing complex (RISC) assembly. During RISC loading, one strand of the small RNA duplex is selected as the guide strand (red) and incorporated into ARGONAUTE 1 (AGO1) to form a functional RISC, whereas the other strand (the passenger strand) is removed and degraded. m7G,7-methylguanylate cap at the 5′ end of primary miRNAs; HST, HASTY. (From [72]).



Although the tissues of developing seeds are technically challenging to sample and study, it is important to focus on miRNAs (and other small non-coding RNAs) regulatory processes, as the modulation of miRNAs abundance may be used to improve agronomic traits, such as crop yield and tolerance to stress agents [73]. However, modulation of miRNAs depends on their identification and understanding of their role in major developmental processes or in environmental responses.

Most of the efforts regarding post-transcriptional regulation mediated by miRNAs during seed development were performed in maize, rice, barley, and other model species [70]. For instance, in Arabidopsis, miRNAs regulate embryo development starting at the eight-cell stage [74]. The authors found that the lack of *DICER-LIKE1* (*DCL1*), which is required for miRNA biogenesis, leads to the abnormal division of the hypophysis and subprotoderm cells. Furthermore, overexpression of several miRNA targets was observed in this condition, including TFs such as *PHV, PHB,* and *SQUAMOSA PROMOTER BINDING PROTEIN-LIKE* (*SPL*) *10* and *SPL11*. The authors conclude that miRNAs might prevent the precocious expression of differentiation-promoting transcription factors, enabling a proper embryo patterning. Other impactful roles for miRNAs have been proposed during seed development. During morphogenesis in Arabidopsis, miR160 targets *AUXIN RESPONSE FACTOR* (*ARF*) *10*, *ARF16,* and *ARF17* impacting cell division, patterning, and differentiation [75] or in rice the miR397 impacts grain size via its target a laccase-like protein Os*LAC* [76].

Still, some studies have focused on the identification and understanding of miRNAs functions during legume seed development. For instance, Song *et al*., (2011) [77] constructed small RNA and degradome libraries from soybean seeds at 15 days after flowering. Using Solexa sequencing by synthesis (SBS) technology, the authors were able to identify 55 previously annotated miRNAs, as well as 26 new miRNAs, and confirmed their expression using stem-loop RT-qPCR. The authors also observed that, of the identified targets, 82% were transcription factors including ARF, NAC, and HD-ZIP family members, suggesting a role in plant growth or responses to the environment. Interestingly, the authors found that the miRNAs only found in soybean might regulate fewer targets than conserved miRNAs do. Additionally, and unlike the conserved miRNAs identified in the study, the new soybean miRNAs were not enriched in transcription factors, suggesting a new feature of miRNA regulation in soybeans [77].

To further understand the miRNA-mediated post-transcriptional regulation during seed development, Shamimuzzaman and Vodkin (2012) constructed five separate degradome libraries derived from seed coats and cotyledons of different soybean developmental stages, from early to mid-maturation. The authors were able to obtain 183 different targets for 53 known soybean miRNAs. While these results provided support for previous computational predictions, the authors used RNA ligase mediated 5’ rapid amplification of cDNA ends (5’RLM-RACE) to further confirm the degradome data. For instance, miR160 was found to have five identified targets in the degradome library. Using 5’RLM-RACE, the authors were able to validate the cleavage of four of the five ARF TFs targets of this miRNA in the cotyledon, suggesting a role in auxin signalling and indicating that degradome sequencing is an efficient strategy to identify miRNA targets in plants [78]. The comparison of these results with data from Song *et al.* [77] revealed similarities between miRNA targets, such as members of the ARF, NAC, and HD-ZIP TFs families.

More recently, new reports have been published regarding the role of miRNAs in legume seeds. For instance, in 2019, Yu *et al*. [79] sampled soybean seeds (from different lines with distinct oil and protein content) at early maturation, mid maturation, and at the dry seed stage for a sRNA-Seq analysis. The authors focused on target genes that might be involved in the development of soybean seeds and related to proteins, fatty acids, or sugars. For instance, two genes encoding *ALCOHOL DEHYDROGENASE 1* were identified as targets of miR2119 and are associated with the accumulation of oil in soybean seeds.

In another legume species, DeBoer *et al*., (2019) [80] analysed five seed developmental timepoints of *Lupinus angustifolius*, ranging from early maturation to the onset of seed desiccation (between 4 to 44 days after anthesis). The authors were able to profile the abundances of miRNA families members using an Illumina MiSeq, and confirmed expression of miR156, miR166, miR164, miR1507, and miR396 using quantitative RT-qPCR. Additionally, thirteen new lupin-specific miRNAs were also described, and, using 5'RLM-RACE the authors were able to validate the targets of miR399 and miR159. These targets are a putative ubiquitin-conjugating E2 enzyme, involved in the maintenance of phosphate homeostasis in other plant species, as well as a putative MYB transcription factor, respectively [80].

Pongamia (*Millettia pinnata* syn. *Pongamia pinnata)* is a legume tree species, with a high yield of non-edible seed oils that are highly suitable for biodiesel preparation [81]. Jin *et al*. [82] studied its seed development via the analysis of three distinct seed developmental stages, from embryogenesis to desiccation. Using an Illumina HiSeq 2500 platform and bioinformatic analysis, the authors were able to identify 82 conserved and 33 new miRNAs whose expression changed significantly between developmental stages. An enrichment analysis of all predicted target genes for the differentially expressed miRNAs revealed DNA repair as one of the most enriched GO

terms and, also, non-homologous end-joining and base excision repair as significantly enriched KEGG pathways. These results suggested that DNA damage repair is crucial for the maintenance of genome integrity during Pongamia seed development [82].

Indeed, maintenance of genome integrity is particularly important to the seed phase of the plant lifecycle [83]. Chromatin integrity is constantly being challenged by environmental factors like drought and ionizing radiation or free radicals and alkylating agents generated by endogenous processes [84,85]. These agents cause a variety of DNA damage seriously threatening the integrity of the plant genome [84]. To maintain genome stability, all organisms have evolved DNA Damage Response (DDR) mechanisms that activate cell cycle checkpoints and DNA repair pathways, and programmed cell death [86].



While genome integrity mechanisms have been extensively characterized during seed germination, the knowledge concerning the link between DDR and chromatin remodelling during seed development was scarce at the beginning of this thesis. Embryogenesis, with a high rate of cell division, as well as desiccation, in which orthodox seeds lose water while maintaining viability, are developmental stages prone to DNA damage. Consequently, disturbances during seed development may impact subsequent seed physiology, as desiccation, dormancy, longevity, and germination [87]. As an example, an essential role for the apurinic/apyrimidinic endonucleases APE1L and APE2, involved in DNA repair during embryo development has been described in *Arabidopsis* [88]. Also, a differential accumulation of DNA repair-related proteins, such as RAD23-like protein, has been described during seed development of *Brassica campestris* L. [89]. Nevertheless, more studies are needed to provide a comprehensive picture of the different DDR mechanisms activated to ensure the maintenance of genome integrity in seeds.

Taken together, all these reports highlight the complexity and the still limited knowledge available on seed development, reinforcing the need to use holistic approaches to deepen our knowledge of this process. Indeed, the combination of new high-throughput methodologies for transcriptome and proteome profiling allows the identification of a larger number of molecules, with higher specificity and accuracy, providing a deeper overview of the molecular intricacies behind complex processes [90,91].



## 2.2 Seed development in *P. vulgaris*: insights over a complex process

This species has been emerging as one of the best agricultural model crops to study seed biology and cumulative studies have been addressing this topic. Numerous reports have focused on this process, revealing molecular players and regulatory networks behind seed development in *P. vulgaris*. These reports will be discussed in the following sections.

### 2.2.1 Phenological characterization

In 1986, CIAT published a study guide with a description of the 10 different stages of *P. vulgaris* development [92], including both vegetative and reproductive stages (**Figure 8**). Within the latter, Preflowering (R5) starts from the first bud appearance until the flower fully develops and opens. From this moment starts Flowering (R6) that lasts until pod appearance, with a detached or still hanging corolla. Then Pod formation (R7) begins, which is characterized by an increase in length of the pod with a little increase of the seeds. When the valves reach their full-size Pod filling (R8) begins. During this stage, a change in pod colour can occur, and seeds are actively growing. Afterward comes Maturation (R9), which is the last stage as maturity is reached. There is a discoloration of the pods and seeds lose water, acquiring their typical coloration. Many factors seem to affect the duration of the developmental stages, including the genotype and the climate [92].

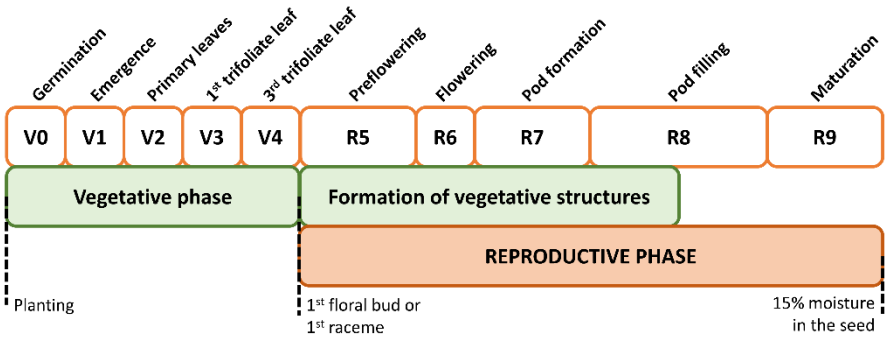





**Figure 8** Stages of development of a common bean plant. Adapted from [92]**.**

One of the early works describing seed development in common bean is dated of 1971, performed in variety Taylor's Horticultural [93]. Seed length and dry weight was measured at successive timepoints during the 5 weeks of embryo development. While the first 2 weeks are characterized by a slow embryo growth, the following 2 weeks are characterized by the rapid growth and differentiation of the mature embryo. Afterwards, the seeds start to dehydrate and lose up to 90% of their water content. The authors focused on RNA metabolism and found that the synthesis of RNA occurs quickly between days 14 to 24 days after anthesis (DAA), with a maximum at 16-20 DAA. An interesting hypothesis raised by the authors is that the RNA synthesized by day 25 seems to be maintained during maturation, dormancy, and persisting during the first hours of germination for protein synthesis [93].

Seed development in the cultivar Black Valentine was characterized by Geiger *et al.* (1989) [94]. The authors found that, while seeds volume increased as pods grew in length, dry weight accumulation in the seeds increased quickly after the pods stopped elongating, at about 12 DAA. Additionally, the authors proposed the timing of developmental events in pods and seeds, reflecting the different stages of embryo development. In this work, the authors state a need to know the time course for dry weight accumulation in individual fruits and seeds, as it helps to schedule sampling for research on specific stages.

More recently, Michelangeli *et al.* [95] focused on the phenological markers to model reproductive development in the common bean. Interestingly the authors used one Andean (Calima) and one Mesoamerican (Jamapa) genotype, as well as 3 recombinant inbred lines, to identify the phenological markers of reproductive development. Seed growth followed a 3-phase sigmoidal pattern, where initially there is a lag phase (until 15-18 DAA), followed by an increase in dry weight and rapid growth (until 36-42 DAA), and a constant dry weight phase afterwards. Interestingly, the authors observed that regardless of the genotype, the end of pod elongation seems to mark the end of the lag phase and the beginning of seed growth, while the change in pod hue from green to yellow/brown seems to mark the end of the rapid growth phase [95].



At the beginning of this thesis, no reports were found regarding the characterization of seed development in SER 16, thus there was still a lack of knowledge about the timing of developmental events that reflected the different stages of embryo development.

### 2.2.2 Molecular players and regulatory networks implicated in seed development in common bean

The knowledge over species-specific molecular mechanisms and regulatory networks operating during the span of this developmental process in *P. vulgaris* was limited by the beginning of this thesis. Nonetheless, some advances were achieved in this regard.

#### 2.2.2.1 Protein metabolism

Nitrogen is an essential macronutrient for plants due to its impact on growth and development. Allatoin and allantoic acid are the major transported forms of nitrogen in tropical and subtropical legumes, including *P. vulgaris* [96]. Besides the stem, allatoin seems to be more accumulated at the seed coat in this species [97]. The allantoin transporter *UPS1* was

found to be expressed in the seed coats and localized in the inner epidermis [97]. The authors relate the results with previous reports that allantoin is metabolized in the seed coat and most of the N is translocated in the form of amino acids for cotyledon nutrition [97].

Phaseolin is the most important storage protein in common bean seeds, accounting for around 50% of total protein stored in the cotyledons, and 35-46% of the total seed nitrogen [98]. Indeed, the expression of β-PHASEOLIN gene (*PHAS*) was described as confined to the cotyledons of developing embryos [99]. Phytohemagglutinin (PHA), another storage protein and the major seed lectin of the common bean, was found more accumulated at the cotyledons but is also present in the embryonic axis [100]. The embryonic axis contains about 15% of the PHA available in the cotyledons [101]. Its gene expression, synthesis, and localization were studied in common bean seeds [100,101]. As expected, PHA synthesis seems to be higher in the cotyledons than in the embryo axes, and this result is correlated with the amounts of mRNA found [100]. Another protein present in the axis and cotyledons is the α-AMYLASE-INHIBITOR, a plant defence protein that inhibits insect α-amylases [102]. The authors have shown that this lectin-type protein is synthesized during maturation and accumulates in the protein storage vacuoles.



Regulation of expression of phaseolin has been shown to be highly regulated and complex [103–107]. The ABI3-LIKE FACTOR (PvALF) was described as positively regulating phaseolin and phytohemagglutinin promoters [106]. *PvALF* expression is embryo-specific and temporally complex. Higher expression was observed preceding the onset of maturation and, thus, the expression of phaseolin and phytohemagglutinin genes, and is expressed again after these genes' expression begin to decrease [106]. While PvALF was initially described as regulating *PHAS* expression, other factors are also involved in this process. such as FUS3 and ABI5 [107].

Technological advances have allowed a research shift to focus on broader approaches regarding common bean storage proteins. In 2010, Marsolais *et al*. performed proteomic analysis of common bean mature seeds using a combination of label-free quantification by spectral counting and 2-DE [21]. The authors used two genetically related lines, SARC1 and SMARC1N-PN1. In SARC1 the lectin ARCELIN-1 is the dominant seed storage protein and a 3-fold decrease of phaseolin levels was observed. In contrast, SMARC1N-PN1 lacks PHASEOLIN and major lectins, through introgression from a *Phaseolus coccineus* accession and Great Northern 1140. Both lines share a similar level (ca. 85%) of the recurrent parental genotype from the commercial cv. Sanilac. The authors observed in the SMARC1N-PN1 line, an increase in 11S globulin legumin and α-amylase inhibitors among others, concomitantly with a decrease in phaseolin, phytohemagglutinin, and arcelin [21]. Curiously, while no change was observed in starch and amylose content, an increased abundance of starch and raffinose metabolic enzymes was observed [21]. These results have an impact on the nutritional and quality improvement of common bean seeds, especially with the increased sulphur amino acid content observed, potentially leading to a more balanced amino acid composition of the seeds.

Indeed, to understand the regulation of sulphur metabolism in developing seeds further research was developed using the SARC1 and SMARC1N-PN1 lines [108]. Liao *et al.* analysed the differentially expressed genes between both lines using microarrays at four time points during seed development, ranging from the cotyledon stage to the mature seed. Indeed, transcripts of sulphur-rich proteins were elevated during seed development in SMARC1N-PN1 [108]. These include those identified previously by the proteomics approach [21], such as including *LEGUMIN*, *ALBUMIN-2*, and *BOWMAN–BIRK TYPE PROTEINASE INHIBITOR 2*. Also, a significant increase in the abundance of transcripts related to sulphur metabolism was observed at maturation and desiccation. As an example, transcript levels of

two sulphate transporters, *SULTR1;2* and *SULTR3;3*, were elevated in SMARC1N-PN1 at maturation, which coincides with the active phase of storage protein accumulation. This study contributed with knowledge about the regulation of sulphur metabolism in the developing seeds as an answer to the change in the composition of endogenous proteins in the SMARC1N-PN1 line [108].



Another relevant study was the 2-DE-based proteomic analysis of common bean seeds [109]. The authors tested three different protein extraction methods (TCA–acetone, phenol, and a commercial clean-up kit) and were able to identify several storage proteins, including PHASEOLIN, PHYTOHEMAGGLUTININ, and α-AMYLASE INHIBITOR. Furthermore, proteins with other roles were identified in this study, including those related to carbohydrate metabolism or with stress responses, such as the GRANULE-BOUND STARCH SYNTHASE I or a DEHYDRIN [109]. In 2013, a more comprehensive two-dimensional map of proteins from the common bean seeds was published [110]. The authors were able to identify 141 spots using MALDI-TOF/TOF, which gave more information about the proteins accumulated in mature seeds and providing clues about the molecular processes that might be acting in the seeds. Most of the identified proteins are storage proteins, including several spots reported as PHASEOLIN, as well as LEGUMIN and ALBUMIN. Still, the authors also identified proteins related to other functional categories, such as defence-related, including α-AMYLASE INHIBITOR, or stress-related, including LEA proteins and dehydrins [110].

In another study, an ethyl methanesulfate (EMS) *P. vulgaris* mutant was studied at the seed development level [111]. This mutant, with arrested development after the embryo early globular stage and, thus, defective at the heart and later stages, was compared with the BAT93 cultivar, a non-abortive control. A set of 250 genes previously described as required for normal Arabidopsis embryo development were used to perform an *in silico* analysis

to search for EST sequences available in *Phaseolus.* The authors obtained a total of 22 ESTs with significant homology and confirmed *L1L*, *MINUTE-LIKE 1* (*AML1*), *BIOTIN AUXOTROPH 2* (*BIO2*), and *PASTICCINO 2* (*PAS2*) as highly and specifically expressed in the seeds at 15 days after pollination. A disrupted expression of *AML1* and *L1L*, among others, was observed in the mutant line, suggesting an essential role for these genes during seed development in *P. vulgaris* [111]. Indeed, L1L is a homolog of LEC1, a member of the LAFL network that regulates storage compound synthesis [65].



Overall, several reports have tackled protein metabolism in common bean seeds. A complex regulatory network seems to control these processes, especially during maturation. Interestingly, these reports have also shown the potential for common bean protein quality adaptation, as well as other related metabolisms.

#### 2.2.2.2 Carbohydrate metabolism

Besides proteins, carbohydrates are also important storage compounds in *P. vulgaris* seeds. Sucrose is the major carbohydrate transported to sink tissues for storage and development [112]. Sucrose is also the predominant sugar in *P. vulgaris* mature seeds [113]. SUCROSE SYNTHASE (SUS) is a key enzyme in the sucrose metabolism of plants and has been considered a biochemical marker for sink strength [114]. SUS is responsible for the degradation of sucrose into uridine-diphosphoglucose (UDP-glucose) and fructose for various metabolic and signalling processes such as starch or cellulose biosynthesis [114].

During common bean seed development, SUS was shown to have two high activity peaks at 16 and 24 days after anthesis, coincident with the peaks of dry weight accumulation [115]. In 2012, the expression profiles of a *SUS* gene in *P. vulgaris* were reported [116]. The authors observed a higher expression of *SUS* in the seeds, when compared with other plant tissues,

reflecting a high requirement for carbon and energy for rapid cell growth. RT-qPCR was used to analyse gene expression at the initial seed development time-points, from early globular to cotyledon stage. The dynamic expression profile obtained for *SUS*, with a peak expression at 9 days after pollination seems to be related to the biosynthesis of storage compounds [116]. Furthermore, evidence of tissue-specific expression of *SUS* was obtained suggesting a role for *SUS* initially for embryo and endosperm development, and then later for storage compound biosynthesis. In 2014 a RNA-Seq based gene expression atlas of the common bean was released [32]. The authors also analysed gene expression dynamics in the seeds at heart stage seeds (3 to 4 mm across), seeds between 6 and 7 mm, and seeds between 8 and 10 mm. Around half of the predicted genes in the study were found expressed in seeds with 75% of these expressed at all timepoints studied. Interestingly, 9702 genes were found to be downregulated as the seed develops, contrasting with 753 genes that increased expression [32]. Unsurprisingly, genes involved in carbohydrate biosynthesis are upregulated during seed development, such as the *STARCH SYNTHASE* and *SUS genes* [32]. Two sucrose transporters putatively involved in sucrose efflux from maternal seed coats to the seed apoplasm were identified [117]. The PvSUT1 exhibited functional characteristics of a sucrose/$H^+$ symporter, while PvSUF1 supported the pH- and energy-independent transport of sucrose. While expression of *PvSUF1* is higher than *PvSUT1* in several plant tissues including leaves, stem, and flowers, at the seed coats and cotyledons the expression profiles are reversed [117].

The SnRK1 is part of a subfamily of serine/threonine kinases that act as metabolite sensors, adapting metabolism to the supply/demand for energy [118]. Also, SnRK1 is implicated in the control of carbohydrate metabolism [119]. In common bean seeds, SnRK1 activity reaches its highest point at around 20 days after pollination, coincidently with the beginning of protein and starch accumulation as well as a large decrease in hexose content [120]. In this report, the authors found that regulation of

SnRK1 activity seems to be complex and tissue-dependent, where different inhibitors distinctively impact its activity, either in the cotyledon or the embryo axis. In the cotyledon, SnRK1 is inhibited by trehalose 6-phosphate (T6P) and adenosine diphosphoglucose (ADPG), while in the embryo axis it is virtually insensitive to T6P, but exhibits some inhibition by ADPG [120]. Furthermore, SnRK1 seems to be part of a mechanism required for nutrient remobilisation under conditions of stress in common bean seeds [121]. In this report, 35% increased activity of SnRK1 was observed in the cotyledons and embryo axes of seeds under reduced nutrient supply, when pods are removed from the plant [121].



Besides these roles, carbohydrates have other important roles. Oligosaccharides contents and the activities of antioxidant enzymes were found related to the acquisition of desiccation tolerance in developing common bean seeds [122]. In particular, desiccation tolerance acquisition was coincident with raffinose and stachyose accumulation in the embryonic axis. Also, in tolerant desiccating seeds, high catalase and glutathione reductase activities were observed, comparatively with low superoxide dismutase and ascorbate peroxidase activities [122].

#### 2.2.2.3 Phosphate metabolism

Phytic acid (myo-inositol hexaphosphate; $InsP_6$) is a major phosphate reserve in seeds and, in its storage form, phytin or phytate, is a mixed salt with mineral elements such as $K^+$, $Mg^{2+}$, and $Ca^{2+}$ [48]. It is considered nutritionally undesirable because it reduces essential dietary mineral availability for absorption [48].

Coelho *et al.* studied the dynamics of inositol phosphate pools concerning the rate of phytate synthesis during common bean seed development [123]. Samples were collected from 12 until 44 DAA. The authors found that total phosphorus increased coincidentally with dry matter accumulation. Additionally, the authors found evidence of two possible

regulatory points for phytate synthesis during seed development, one upstream of $InsP_3$ and another involved in the conversion of $InsP_5$ to $InsP_6$. In another report, Abid *et al.*, [124] studied the *MYO-INOSITOL-1-PHOSPHATE SYNTHASE* (*MIPS*) enzyme expression, and found it to be mainly expressed in seeds but also in flowers, cotyledons and leaves. The authors observed a decrease in expression from 3 (early globular stage) to 6 DAP (globular stage), followed by an increased expression to 9 DAP (torpedo stage) decreasing slightly to 12 DAP (bent cotyledon stage). This expression profile is suggested to be correlated with the accumulation of phytic acid in the seeds [124].



### 2.2.2.4 Regulation of gene expression

On the same RNA-Seq based gene expression atlas described previously, the authors also observed that 1197 TFs decrease their expression, while only 71 increase their expression throughout the time-points studied [32]. Members of several TF families, including HB, MYB, and ARF, were found to have seed-specific expression. Importantly, the authors also analysed the expression of some LAFL network members, being *LEC1* and *ABI3* highly expressed at the last time-points, while *LEC2* showed an opposite expression profile. The authors suggest that the low expression of both *LEC2* and *WRI1* observed may relate to the lower oil composition of the common bean seeds [32]. Using the genome and the same datasets produced by O'Rourke *et al.* [32], Wang *et al.* [125] characterized the basic leucine zipper (bZIP) transcription factor gene family in *P. vulgaris.* The authors found 72 *P. vulgaris* bZIP genes encoding 92 distinct proteins. Analysis of bZIP members' expression revealed three distinct expression profiles, suggesting their expression to be developmentally dependent [125]. During common bean seed development, the bZIP10 and bZIP33, homologous of *ABI5*, might be involved in ABA signalling and mediating embryogenesis in late embryo development. Putatively involved in ABA-mediated seed development and maturation, the bZIP71, homologous to

AREB3/DPBF3, was also found expressed at the middle time point studied [125].

Regarding miRNAs, several studies focused on different organs, such as leaves, roots or nodules, and/or growth conditions [34,35,126–131]. While most of these reports do not focus on the seeds, a few studies have addressed this topic, focusing however in one or few members of miRNA families on specific time-points [126,128]. As an example, Arenas-Huertero *et al*. (2009) [126] used total RNA from immature and dry, mature *P. vulgaris* embryos (among other tissues) to analyse the expression of conserved and newly identified miRNAs, including miRNA397, miRNA408, and miRNA2118 [126]. On the other hand, Contrera-Cubas *et al.* (2012) reported a second small RNA, from the *P. vulgaris* miRNA159a precursor, denoted as miRNA159.2. This miRNA accumulation seems to be easily detectable in the species, including the developing seeds analysed by the authors.



## 3. Scope of the thesis

The use of omics approaches, such as next-generation sequencing, and implementation of associated bioinformatic tools allows researchers to turn species of interest into their model organism. Indeed, the combination of the new high-throughput methodologies for transcriptome and proteome profiling allows the identification of a larger number of molecules, with higher specificity and accuracy, providing a deeper overview of the molecular complexities of a variety of organisms and developmental processes [90,91]. As Zargar *et al.* stated, the use of proteomics, genomics, and transcriptomics would likely enhance the agronomic merit and quality traits of the common bean by unravelling regulatory pathways and then enable the manipulation of these regulatory pathways to attain an improved and more sustainable crop [132].

As evidenced throughout this chapter, and prior to the onset of my PhD studies, limited insights into the molecular mechanisms that shape *P.*

*vulgaris* seed development were available. There was no comprehensive description of the proteins and genes accumulated at the different developmental stages, or the acting molecular mechanisms and regulatory networks. As such, several research questions remained to be addressed for this important species, namely: i) how do measurable phenotypic characteristics change during seed development and what is the timeframe of the main developmental stages? ii) what are the main molecular mechanisms and metabolic pathways at each seed developmental stage? and iii) what regulatory networks and key regulators might impact seed traits of interest?



The broad goal of this PhD thesis was to describe the molecular and regulatory mechanisms underlying seed development in *P. vulgaris*. To address these gaps in knowledge, several specific objectives were defined:

1) To perform a phenotypic characterization of the seed development process in *P. vulgaris*. To address this objective, an experimental scheme was implemented to harvest seeds at different developmental stages, spanning from embryogenesis to desiccation. Such step was crucial to identify the timeframe and main seed developmental stages in which molecular analysis would be conducted.

2) To identify the main molecular mechanisms and metabolic pathways underlying seed development in *P. vulgaris*. To accomplish this goal, a comprehensive characterization of proteome dynamics using a high-throughput gel-free proteomics approach was performed at the different developmental stages identified in objective 1.

3) To characterize the genome integrity maintenance players and mechanisms underlying seed development in *P. vulgaris.* For this goal, a high-throughput transcriptomics profiling of genes related to DNA damage response and chromatin remodelling was executed at the same seed development stages.

4) To understand the role of post-transcriptional regulatory mechanisms mediated by miRNAs in the developing seeds of *P. vulgaris*. A non-coding transcriptomics approach, coupled with a degradome, a predictive target genes analysis and the transcriptomics datasets obtained in objective 3, was performed at the same timepoints.



To successfully fulfil these objectives, the research described herein made use of comprehensive molecular analysis using Proteomics, and Coding and Non-Coding Transcriptomics. In **Chapter 2**, the proteome dynamics were described for seeds at late embryogenesis until desiccation. Protein accumulation profiles revealed distinct signatures of each stage studied, including clues about the regulation of seed development mediated by post-translational modifications and maintenance of genome integrity. In **Chapter 3**, the transcriptomic profiling of the expression of genes related to maintenance of genome integrity was performed at the same developmental stages described in **Chapter 2**. This analysis suggested that mechanisms for DNA damage response and chromatin remodelling are more active during embryo morphogenesis, while also provided clues about protection mechanisms acting during seed desiccation. **Chapter 4** focused on the post-transcriptional regulation of gene expression mediated by miRNAs during seed development. The initial and late stages of seed development seem to be where miRNAs action is more pronounced. The identified miRNAs, and respective targets, seem to be implicated in developmental stages transition, nutrient allocation and storage molecule synthesis, and desiccation tolerance and longevity. Finally, in **Chapter 5**, the results obtained were further discussed and contextualized, integrating the outcomes with the latest reports on the seed development subject.

Overall, the outputs of this thesis addressed an existing gap of knowledge and contributed to expand the information available on Seed Development, especially on *P. vulgaris*. The main metabolic pathways and molecular players were disclosed, providing evidence of a coordinated

regulation of gene expression while identifying candidate genes/proteins/miRNAs with potential impact on seed traits. Taken together, the produced knowledge potentially enables the future modulation of these pathways to enhance seed quality traits and ultimately enhance *P. vulgaris* production and yield.



## 3. Bibliography


1. Azani, N.; Babineau, M.; Bailey, C.D.; Banks, H.; Barbosa, A.; Pinto, R.B.; Boatwright, J.; Borges, L.; Brown, G.; Bruneau, A.; et al. A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny – The Legume Phylogeny Working Group (LPWG). *Taxon* **2017**, *66*, 44–77.

2. Vasconcelos, M.W.; Grusak, M.A.; Pinto, E.; Gomes, A.; Ferreira, H.; Balázs, B.; Centofanti, T.; Ntatsi, G.; Savvas, D.; Karkanis, A.; et al. The Biology of Legumes and Their Agronomic, Economic, and Social Impact. In *The Plant Family Fabaceae*; Springer Singapore: Singapore, **2020**; pp. 3–25 ISBN 9789811547522.

3. Vaz Patto, M.C.; Araújo, S.S. Positioning Portugal into the context of world production and research in grain legumes. *Rev. Ciências Agrárias* **2016**, *39*, 471–489.

4. Rubiales, D.; Mikic, A. Introduction: Legumes in Sustainable Agriculture. *CRC. Crit. Rev. Plant Sci.* **2015**, *34*, 2–3.

5. Jensen, E.S.; Peoples, M.B.; Boddey, R.M.; Gresshoff, P.M.; Hauggaard-Nielsen, H.; J.R. Alves, B.; Morrison, M.J.; Henrik, H.N.; Alves, B.J.R.; Morrison, M.J. Legumes for mitigation of climate change and the provision of feedstock for biofuels and biorefineries. A review. *Agron. Sustain. Dev.* **2012**, *32*, 329–364.

6. Maphosa, Y.; Jideani, V.A. The Role of Legumes in Human Nutrition. In *Functional Food - Improve Health through Adequate Food*; InTech, **2017**; Vol. i, p. 13.

7. González-Bernal, M.J.; Rubiales, D. Las leguminosas grano en la agricultura española y europea. *Arbor* **2016**, *192*, a311.

8. Câmara, C.; Urrea, C.; Schlegel, V. Pinto Beans (*Phaseolus vulgaris* L.) as a Functional Food: Implications on Human Health. *Agriculture* **2013**, *3*, 90–111.

9. Bitocchi, E.; Bellucci, E.; Giardini, A.; Rau, D.; Rodriguez, M.; Biagetti, E.; Santilocchi, R.; Spagnoletti Zeuli, P.; Gioia, T.; Logozzo, G.; *et al*. Molecular analysis of the parallel domestication of the common bean (*Phaseolus vulgaris*) in Mesoamerica and the Andes. *New Phytol.* **2013**, *197*, 300–313.

10. Singh, S.P.; Gepts, P.; Debouck, D.G. Races of common bean (*Phaseolus vulgaris*, Fabaceae). *Econ. Bot.* **1991**, *45*, 379–396.

11. Leitão, S.T.; Dinis, M.; Veloso, M.M.; Šatović, Z.; Vaz Patto, M.C. Establishing the Bases for Introducing the Unexplored Portuguese Common Bean Germplasm into the Breeding World. *Front. Plant Sci.* **2017**, *8*, 1–18.

12. Beebe, S.; Rao, I.; Polanía, J.; Grajales, M.; Cajiao, C. Improved harvest index in drought resistant common beans and possible effects on combining ability. *Bean Improv. Coop. Annu. Rep. (USA).* **2008**, *51*, 8–9.

13. Polania, J.A.; Poschenrieder, C.; Beebe, S.; Rao, I.M. Effective Use of Water and Increased Dry Matter Partitioned to Grain Contribute to Yield of Common Bean Improved for Drought Resistance. *Front. Plant Sci.* **2016**, *7*, 1–10.

14. Klaedtke, S.M.; Cajiao, C.; Grajales, M.; Polanía, J.; Borrero, G.; Guerrero, A.; Rivera, M.; Rao, I.; Beebe, S.E.; Léon, J. Photosynthate remobilization capacity from drought-adapted common bean (*Phaseolus vulgaris* L.) lines can improve yield potential of interspecific populations within the secondary gene pool. *J. Plant Breed. Crop Sci.* **2012**, *4*, 49–61.

15. Araújo, S.S.; Beebe, S.; Crespi, M.; Delbreil, B.; González, E.M.; Gruber, V.; Lejeune-Henaut, I.; Link, W.; Monteros, M.J.; Prats, E.; *et al*. Abiotic Stress Responses in Legumes: Strategies Used to Cope with Environmental Challenges. *CRC. Crit. Rev. Plant Sci.* **2015**, *34*, 237–280.

16. Petry, N.; Boy, E.; Wirth, J.; Hurrell, R. Review: The Potential of the Common Bean (*Phaseolus vulgaris*) as a Vehicle for Iron Biofortification. *Nutrients* **2015**, *7*, 1144–1173.

17. Beebe, S. Common bean breeding in the tropics. In *Plant Breeding Reviews volume 36*; Janick, J., Ed.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, **2012**; Vol. 36, pp. 357–426 ISBN 9781118358566.

18. Chianu, J.N.; Nkonya, E.M.; Mairura, F.S.; Chianu, J.N.; Akinnifesi, F.K. Biological nitrogen fixation and socioeconomic factors for legume

production in sub-Saharan Africa: a review. *Agron. Sustain. Dev.* **2011**, *31*, 139–154.

19. Beshir, H.; Walley, F.; Bueckert, R.; Tar'an, B. Response of Snap Bean Cultivars to *Rhizobium* Inoculation under Dryland Agriculture in Ethiopia. *Agronomy* **2015**, *5*, 291–308.

20. Guzmán-Maldonado, S.H.; Acosta-Gallegos, J.; Paredes-López, O. Protein and mineral content of a novel collection of wild and weedy common bean (*Phaseolus vulgaris* L). *J. Sci. Food Agric.* **2000**, *80*, 1874–1881.

21. Marsolais, F.; Pajak, A.; Yin, F.; Taylor, M.; Gabriel, M.; Merino, D.M.; Ma, V.; Kameka, A.; Vijayan, P.; Pham, H.; *et al*. Proteomic analysis of common bean seed with storage protein deficiency reveals up-regulation of sulfur-rich proteins and starch and raffinose metabolic enzymes, and down-regulation of the secretory pathway. *J. Proteomics* **2010**, *73*, 1587–1600.

22. de Almeida Costa, G.E.; da Silva Queiroz-Monici, K.; Pissini Machado Reis, S.M.; de Oliveira, A.C. Chemical composition, dietary fibre and resistant starch contents of raw and cooked pea, common bean, chickpea and lentil legumes. *Food Chem.* **2006**, *94*, 327–330.

23. Gabriele, M.; Sparvoli, F.; Bollini, R.; Lubrano, V.; Longo, V.; Pucci, L. The Impact of Sourdough Fermentation on Non-Nutritive Compounds and Antioxidant Activities of Flours from Different *Phaseolus vulgaris* L. Genotypes. *J. Food Sci.* **2019**.

24. Ribeiro, N.D.; Maziero, S.M.; Prigol, M.; Nogueira, C.W.; Rosa, D.P.; Possobom, M.T.D.F. Mineral concentrations in the embryo and seed coat of common bean cultivars. *J. Food Compos. Anal.* **2012**, *26*, 89–95.

25. Broughton, W.J.; Hernández, G.; Blair, M.; Beebe, S.; Gepts, P.; Vanderleyden, J. Beans (*Phaseolus* spp.) – model food legumes. *Plant Soil* **2003**, *252*, 55–128.

26. Ganesan, K.; Xu, B. Polyphenol-Rich Dry Common Beans (*Phaseolus vulgaris* L.) and Their Health Benefits. *Int. J. Mol. Sci.* **2017**, *18*, 2331.

27. Chávez-Mendoza, C.; Hernández-Figueroa, K.; Sánchez, E. Antioxidant Capacity and Phytonutrient Content in the Seed Coat and Cotyledon of Common Beans (*Phaseolus vulgaris* L.) from Various Regions in Mexico. *Antioxidants* **2018**, *8*, 5.

28. Castro-Guerrero, N.A.; Isidra-Arellano, M.C.; Mendoza-Cozatl, D.G.; Valdés-López, O. Common Bean: A Legume Model on the Rise for Unraveling Responses and Adaptations to Iron, Zinc, and Phosphate

Deficiencies. *Front. Plant Sci.* **2016**, *7*, 1–7.

29. Kalavacharla, V.; Liu, Z.; Meyers, B.C.; Thimmapuram, J.; Melmaiee, K. Identification and analysis of common bean (*Phaseolus vulgaris* L.) transcriptomes by massively parallel pyrosequencing. *BMC Plant Biol.* **2011**, *11*, 135.

30. Schmutz, J.; McClean, P.E.; Mamidi, S.; Wu, G.A.; Cannon, S.B.; Grimwood, J.; Jenkins, J.; Shu, S.; Song, Q.; Chavarro, C.; *et al*. A reference genome for common bean and genome-wide analysis of dual domestications. *Nat. Genet.* **2014**, *46*, 707–713.

31. Vlasova, A.; Capella-Gutiérrez, S.; Rendón-anaya, M.; Hernández-Oñate, M.; Minoche, A.E.; Erb, I.; Câmara, F.; Prieto-barja, P.; Corvelo, A.; Sanseverino, W.; et al. Genome and transcriptome analysis of the Mesoamerican common bean and the role of gene duplications in establishing tissue and temporal specialization of genes. *Genome Biol.* **2016**, *17*, 32.

32. O'Rourke, J.A.; Iniguez, L.P.; Fu, F.; Bucciarelli, B.; Miller, S.S.; Jackson, S.A.; McClean, P.E.; Li, J.; Dai, X.; Zhao, P.X.; *et al*. An RNA-Seq based gene expression atlas of the common bean. *BMC Genomics* **2014**, *15*, 866.

33. Bhawna; Bonthala, V.S.; Gajula, M.P. PvTFDB: a Phaseolus vulgaris transcription factors database for expediting functional genomics in legumes. *Database* **2016**, *2016*, baw114.

34. Pelaez, P.; Trejo, M.S.; Iniguez, L.P.; Estrada-Navarrete, G.; Covarrubias, A. a; Reyes, J.L.; Sanchez, F. Identification and characterization of microRNAs in *Phaseolus vulgaris* by high-throughput sequencing. *BMC Genomics* **2012**, *13*, 83.

35. Formey, D.; Iñiguez, L.P.; Peláez, P.; Li, Y.-F.; Sunkar, R.; Sánchez, F.; Reyes, J.L.; Hernández, G. Genome-wide identification of the *Phaseolus vulgaris* sRNAome using small RNA and degradome sequencing. *BMC Genomics* **2015**, *16*, 423.

36. de Sousa Cardoso, T.C.; Portilho, L.G.; de Oliveira, C.L.; McKeown, P.C.; Maluf, W.R.; Gomes, L.A.A.; Teixeira, T.A.; do Amaral, L.R.; Spillane, C.; de Souza Gomes, M. Genome-wide identification and in silico characterisation of microRNAs, their targets and processing pathway genes in *Phaseolus vulgaris* L. *Plant Biol.* **2016**, *18*, 206–219.

37. Collado, R.; Bermúdez-Caraballoso, I.; García, L.R.; Veitía, N.; Torres, D.; Romero, C.; Martirena-Ramírez, A.; Angenon, G. Agrobacterium-mediated transformation of *Phaseolus vulgaris* L. using indirect organogenesis. *Sci. Hortic. (Amsterdam).* **2015**, *195*,

89–100.

38. Aragão, F.J.L.; Vianna, G.R.; Albino, M.M.C.; Rech, E.L. Transgenic Dry Bean Tolerant to the Herbicide Glufosinate Ammonium. *Crop Sci.* **2002**, *42*, 1298.

39. Nanjareddy, K.; Arthikala, M.-K.; Blanco, L.; Arellano, E.S.; Lara, M. Protoplast isolation, transient transformation of leaf mesophyll protoplasts and improved *Agrobacterium*-mediated leaf disc infiltration of *Phaseolus vulgaris*: tools for rapid gene expression analysis. *BMC Biotechnol.* **2016**, *16*, 53.

40. Song, G.; Han, X.; Wiersma, A.T.; Zong, X.; Awale, H.E.; Kelly, J.D. Induction of competent cells for *Agrobacterium tumefaciens*-mediated stable transformation of common bean (*Phaseolus vulgaris* L.). *PLoS One* **2020**, *15*, e0229909.

41. Hnatuszko-Konka, K.; Kowalczyk, T.; Gerszberg, A.; Wiktorek-Smagur, A.; Kononowicz, A.K. *Phaseolus vulgaris* - Recalcitrant potential. *Biotechnol. Adv.* **2014**, *32*, 1205–1215.

42. Cabrera-Ponce, J.L.; López, L.; León-Ramírez, C.G.; Jofre-Garfias, A.E.; Verver-y-Vargas, A. Stress induced acquisition of somatic embryogenesis in common bean *Phaseolus vulgaris* L. *Protoplasma* **2014**, *252*, 559–570.

43. Hnatuszko-Konka, K.; Kowalczyk, T.; Gerszberg, A.; Glińska, S.; Grzegorczyk-Karolak, I. Regeneration of *Phaseolus vulgaris* from epicotyls and hypocotyls via direct organogenesis. *Sci. Rep.* **2019**, *9*, 6248.

44. Bonfim, K.; Faria, J.C.; Nogueira, E.O.P.L.; Mendes, É.A.; Aragão, F.J.L. RNAi-Mediated Resistance to Bean golden mosaic virus in Genetically Engineered Common Bean (*Phaseolus vulgaris*). *Mol. Plant-Microbe Interact.* **2007**, *20*, 717–726.

45. Pflieger, S.; Blanchet, S.; Meziadi, C.; Richard, M.; Thareau, V.; Mary, F.; Mazoyer, C.; Geffroy, V. The “one-step” Bean pod mottle virus (BPMV)-derived vector is a functional genomics tool for efficient overexpression of heterologous protein, virus-induced gene silencing and genetic mapping of BPMV R-gene in common bean (*Phaseolus vulgaris* L.). *BMC Plant Biol.* **2014**, *14*, 232.

46. Kang, Y.; Li, M.; Sinharoy, S.; Verdier, J. A Snapshot of Functional Genetic Studies in *Medicago truncatula*. *Front. Plant Sci.* **2016**, *7*, 1175.

47. Le, B.H.; Wagmaister, J. a; Kawashima, T.; Bui, A.Q.; Harada, J.J.; Goldberg, R.B. Using genomics to study legume seed development.

*Plant Physiol.* **2007**, *144*, 562–74.

48. Bewley, J.D.; Bradford, K.J.; Hilhorst, H.W.M.; Nonogaki, H. *Seeds*; 3rd Edition; Springer New York: New York, NY, 2013; ISBN 978-1-4614-4692-7.

49. Coelho, C.; Benedito, V.; Maria, C.; Coelho, M.; Augusto, V.; Coelho, C.; Benedito, V.; Maria, C.; Coelho, M.; Augusto, V.; et al. Seed development and reserve compound accumulation in common bean (*Phaseolus vulgaris* L.). *Seed Sci. Biotechnol* **2008**, *2*, 42–52.



50. West, M.; Harada, J.J. Embryogenesis in Higher Plants: An Overview. *Plant Cell* **1993**, *5*, 1361–1369.

51. Mondoni, A.; Orsenigo, S.; Donà, M.; Balestrazzi, A.; Probert, R.J.; Hay, F.R.; Petraglia, A.; Abeli, T. Environmentally induced transgenerational changes in seed longevity: maternal and genetic influence. *Ann. Bot.* **2014**, *113*, 1257–1263.

52. Le, B.H.; Cheng, C.; Bui, A.Q.; Wagmaister, J.A.; Henry, K.F.; Pelletier, J.; Kwong, L.; Belmonte, M.; Kirkbride, R.; Horvath, S.; *et al*. Global analysis of gene activity during Arabidopsis seed development and identification of seed-specific transcription factors. *Proc. Natl. Acad. Sci.* **2010**, *107*, 8063–8070.

53. Fait, A.; Angelovici, R.; Less, H.; Ohad, I.; Urbanczyk-Wochniak, E.; Fernie, A.R.; Galili, G. Arabidopsis seed development and germination is associated with temporally distinct metabolic switches. *Plant Physiol.* **2006**.

54. Deng, Z.Y.; Gong, C.Y.; Wang, T. Use of proteomics to understand seed development in rice. *Proteomics* 2013, *13*, 1784–1800.

55. Gu, Y.; Liu, Y.; Zhang, J.; Liu, H.; Hu, Y.; Du, H.; Li, Y.; Chen, J.; Wei, B.; Huang, Y. Identification and characterization of microRNAs in the developing maize endosperm. *Genomics* **2013**, *102*, 472–478.

56. Gallardo, K.; Le Signor, C.; Vandekerckhove, J.; Thompson, R.D.; Burstin, J. Proteomics of *Medicago truncatula* seed development establishes the time frame of diverse metabolic processes related to reserve accumulation. *Plant Physiol.* **2003**, *133*, 664–682.

57. Gallardo, K.; Firnhaber, C.; Zuber, H.; Héricher, D.; Belghazi, M.; Henry, C.; Küster, H.; Thompson, R. A Combined proteome and transcriptome analysis of developing *Medicago truncatula* seeds. *Mol. Cell. Proteomics* **2007**, *6*, 2165–2179.

58. Verdier, J.; Lalanne, D.; Pelletier, S.; Torres-Jerez, I.; Righetti, K.;

Bandyopadhyay, K.; Leprince, O.; Chatelain, E.; Vu, B.L.; Gouzy, J.; et al. A regulatory network-based approach dissects late maturation processes related to the acquisition of desiccation tolerance and longevity of *Medicago truncatula* seeds. *Plant Physiol.* **2013**, *163*, 757–774.

59. Jones, S.I.; Gonzalez, D.O.; Vodkin, L.O. Flux of transcript patterns during soybean seed development. *BMC Genomics* **2010**, *11*, 136.

60. Jones, S.I.; Vodkin, L.O. Using RNA-Seq to profile soybean seed development from fertilization to maturity. *PLoS One* **2013**, *8*, e59270.



61. Agarwal, P.; Kapoor, S.; Tyagi, A.K. Transcription factors regulating the progression of monocot and dicot seed development. *BioEssays* **2011**, *33*, 189–202.

62. Verdier, J.; Kakar, K.; Gallardo, K.; Le Signor, C.; Aubert, G.; Schlereth, A.; Town, C.D.; Udvardi, M.K.; Thompson, R.D. Gene expression profiling of *M. truncatula* transcription factors identifies putative regulators of grain legume seed filling. *Plant Mol. Biol.* **2008**, *67*, 567–580.

63. Verdier, J.; Thompson, R.D. Transcriptional regulation of storage protein synthesis during dicotyledon seed filling. *Plant Cell Physiol.* **2008**, *49*, 1263–1271.

64. Sreenivasulu, N.; Wobus, U. Seed-development programs: a systems biology-based comparison between dicots and monocots. *Annu. Rev. Plant Biol.* **2013**, *64*, 189–217.

65. Lepiniec, L.; Devic, M.; Roscoe, T.J.; Bouyer, D.; Zhou, D.-X.X.; Boulard, C.; Baud, S.; Dubreucq, B. Molecular and epigenetic regulations and functions of the LAFL transcriptional regulators that control seed development. *Plant Reprod.* **2018**, *31*, 291–307.

66. Jo, L.; Pelletier, J.M.; Harada, J.J. Central role of the LEAFY COTYLEDON1 transcription factor in seed development. *J. Integr. Plant Biol.* **2019**, *61*, 564–580.

67. Fatihi, A.; Boulard, C.; Bouyer, D.; Baud, S.; Dubreucq, B.; Lepiniec, L. Deciphering and modifying LAFL transcriptional regulatory network in seed for improving yield and quality of storage compounds. *Plant Sci.* **2016**, *250*, 198–204.

68. Jia, H.; Suzuki, M.; Mccarty, D.R. Regulation of the seed to seedling developmental phase transition by the LAFL and VAL transcription factor networks. *Wiley Interdiscip. Rev. Dev. Biol.* **2014**, *3*, 135–145.

69. Hobert, O. Gene regulation by transcription factors and MicroRNAs.

*Science (80-. ).* **2008**, *319*, 1785–1786.

70. Rodrigues, A.S.; Miguel, C.M. The pivotal role of small non-coding RNAs in the regulation of seed development. *Plant Cell Rep.* **2017**, *36*, 653–667.

71. Martin, R.C.; Martínez-Andújar, C.; Nonogaki, H. Role of miRNAs in seed development. In *MicroRNAs in Plant Development and Stress Responses*; Sunkar, R., Ed.; Signaling and Communication in Plants; Springer Berlin Heidelberg: Berlin, Heidelberg, 2012; Vol. 15, pp. 109–121 ISBN 978-3-642-27383-4.



72. Yu, Y.; Jia, T.; Chen, X. The 'how' and 'where' of plant microRNAs. *New Phytol.* **2017**, *216*, 1002–1017.

73. Djami-Tchatchou, A.T.; Sanan-Mishra, N.; Ntushelo, K.; Dubery, I.A. Functional roles of microRNAs in agronomically important plants—potential as targets for crop improvement and protection. *Front. Plant Sci.* **2017**, *8*, 378.

74. Nodine, M.D.; Bartel, D.P. MicroRNAs prevent precocious gene expression and enable pattern formation during plant embryogenesis. *Genes Dev.* **2010**, *24*, 2678–2692.

75. Liu, X.; Huang, J.; Wang, Y.; Khanna, K.; Xie, Z.; Owen, H.A.; Zhao, D. The role of floral organs in carpels, an Arabidopsis loss-of-function mutation in MicroRNA160a, in organogenesis and the mechanism regulating its expression. *Plant J.* **2010**, *62*, 416–428.

76. Zhang, Y.C.; Yu, Y.; Wang, C.Y.; Li, Z.Y.; Liu, Q.; Xu, J.; Liao, J.Y.; Wang, X.J.; Qu, L.H.; Chen, F.; et al. Overexpression of microRNA OsmiR397 improves rice yield by increasing grain size and promoting panicle branching. *Nat. Biotechnol.* **2013**, *31*, 848–852.

77. Song, Q.-X.; Liu, Y.-F.; Hu, X.-Y.; Zhang, W.-K.; Ma, B.; Chen, S.-Y.; Zhang, J.-S. Identification of miRNAs and their target genes in developing soybean seeds by deep sequencing. *BMC Plant Biol.* **2011**, *11*, 5.

78. Shamimuzzaman, M.; Vodkin, L. Identification of soybean seed developmental stage-specific and tissue-specific miRNA targets by degradome sequencing. *BMC Genomics* **2012**, *13*, 310.

79. Yu, J.-Y.; Zhang, Z.-G.; Huang, S.-Y.; Han, X.; Wang, X.-Y.; Pan, W.-J.; Qin, H.-T.; Qi, H.-D.; Yin, Z.-G.; Qu, K.-X.; et al. Analysis of miRNAs targeted storage regulatory genes during soybean seed development based on transcriptome sequencing. *Genes (Basel).* **2019**, *10*, 408.

80. DeBoer, K.; Melser, S.; Sperschneider, J.; Kamphuis, L.G.; Garg, G.; Gao, L.; Frick, K.; Singh, K.B. Identification and profiling of narrow-leafed lupin (*Lupinus angustifolius*) microRNAs during seed development. *BMC Genomics* **2019**, *20*, 135.

81. Karmee, S.K.; Chadha, A. Preparation of biodiesel from crude oil of Pongamia pinnata. *Bioresour. Technol.* **2005**.

82. Jin, Y.; Liu, L.; Hao, X.; Harry, D.E.; Zheng, Y.; Huang, T.; Huang, J. Unravelling the microRNA-mediated gene regulation in developing pongamia seeds by high-throughput small RNA profiling. *Int. J. Mol. Sci.* **2019**, *20*, 3509.

83. Waterworth, W.M.; Footitt, S.; Bray, C.M.; Finch-Savage, W.E.; West, C.E. DNA damage checkpoint kinase ATM regulates germination and maintains genome stability in seeds. *Proc. Natl. Acad. Sci.* **2016**, *113*, 9647–9652.

84. Roy, S. Maintenance of genome stability in plants: repairing DNA double strand breaks and chromatin structure stability. *Front. Plant Sci.* **2014**, *5*, 487.

85. Manova, V.; Gruszka, D. DNA damage and repair in plants – from models to crops. *Front. Plant Sci.* **2015**, *6*, 885.

86. Spampinato, C.P. Protecting DNA from errors and damage: an overview of DNA repair mechanisms in plants compared to mammals. *Cell. Mol. Life Sci.* **2017**, *74*, 1693–1709.

87. Leprince, O.; Pellizzaro, A.; Berriri, S.; Buitink, J. Late seed maturation: drying without dying. *J. Exp. Bot.* **2016**, *68*, erw363.

88. Murphy, T.M.; Belmonte, M.; Shu, S.; Britt, A.B.; Hatteroth, J. Requirement for abasic endonuclease gene homologues in Arabidopsis seed development. *PLoS One* **2009**, *4*, e4297.

89. Li, W.; Gao, Y.; Xu, H.; Zhang, Y.; Wang, J. A Proteomic analysis of seed development in *Brassica campestri* L. *PLoS One* **2012**, *7*, e50290.

90. Rani, B.; Sharma, V.K. Transcriptome profiling: methods and applications- A review. *Agric. Rev.* **2017**, *38*.

91. Aslam, B.; Basit, M.; Nisar, M.A.; Khurshid, M.; Rasool, M.H. Proteomics: technologies and their applications. *J. Chromatogr. Sci.* **2017**, *55*, 182–196.

92. Fernández, F.; Paul, G.; Marceliano, L. Stages of development of the common bean plant 1986, 32.

93. Walbot, V. RNA metabolism during embryo development and germination of *Phaseolus vulgaris*. *Dev. Biol.* **1971**, *26*, 369–379.

94. Geiger, D.R.; Shieh, W.-J.; Saluke, R.M. Carbon partitioning among leaves, fruits, and seeds during development of *Phaseolus vulgaris* L. *Plant Physiol.* **1989**, *91*, 291–297.

95. Clavijo Michelangeli, J.; Bhakta, M.; Gezan, S.; Boote, K.; Vallejos, C.E. From flower to seed: identifying phenological markers and reliable growth functions to model reproductive development in the common bean (*Phaseolus vulgaris* L.). *Plant. Cell Environ.* **2013**, *36*, 2046–58.



96. Pélissier, H.C.; Frerich, A.; Desimone, M.; Schumacher, K.; Tegeder, M. PvUPS1, an allantoin transporter in nodulated roots of French bean. *Plant Physiol.* **2004**, *134*, 664–675.

97. Pélissier, H.C.; Tegeder, M. PvUPS1 plays a role in source - sink transport of allantoin in French bean (*Phaseolus vulgaris*). *Funct. Plant Biol.* **2007**, *34*, 282.

98. Emani, C.; Hall, T.C. Phaseolin: Structure and Evolution. *Open Evol. J.* **2008**, *1*, 66–74.

99. van der Geest, A.H.M.; Hall, T.C. A 68 bp element of the β-phaseolin promoter functions as a seed-specific enhancer. *Plant Mol. Biol.* **1996**, *32*, 579–588.

100. Chrispeels, M.J.; Vitale, A.; Staswick, P. Gene expression and synthesis of phytohemagglutinin in the embryonic axes of developing *Phaseolus vulgaris* seeds. *Plant Physiol.* **1984**, *76*, 791–796.

101. Greenwood, J.S.; Keller, G.A.; Chrispeels, M.J. Localization of phytohemagglutinin in the embryonic axis of *Phaseolus vulgaris* with ultra-thin cryosections embedded in plastic after indirect immunolabeling. *Planta* **1984**, *162*, 548–555.

102. Moreno, J.; Altabella, T.; Chrispeels, M.J. Characterization of α-amylase-inhibitor, a lectin-like protein in the seeds of *Phaseolus vulgaris*. *Plant Physiol.* **1990**, *92*, 703–709.

103. Chandrasekharan, M.B.; Bishop, K.J.; Hall, T.C. Module-specific regulation of the β-phaseolin promoter during embryogenesis. *Plant J.* **2003**, *33*, 853–866.

104. Ng, D.W.-K.; Chandrasekharan, M.B.; Hall, T.C. Ordered histone modifications are associated with transcriptional poising and activation of the phaseolin promoter. *Plant Cell* **2006**, *18*, 119–132.

105. Sundaram, S.; Kertbundit, S.; Shakirov, E. V.; Iyer, L.M.; Juricek, M.; Hall, T.C. Gene networks and chromatin and transcriptional regulation of the phaseolin promoter in Arabidopsis. *Plant Cell* **2013**, *25*, 2601–2617.

106. Bobb, A.J.; Eiben, H.G.; Bustos, M.M. PvAlf, an embryo-specific acidic transcriptional activator enhances gene expression from phaseolin and phytohemagglutinin promoters. *Plant J.* **1995**, *8*, 331–343.

107. W-K Ng, D.; Hall, T.C. PvALF and FUS3 activate expression from the phaseolin promoter by different mechanisms. *Plant Mol. Biol.* **2008**, *66*, 233–244.



108. Liao, D.; Pajak, A.; Karcz, S.R.; Chapman, B.P.; Sharpe, A.G.; Austin, R.S.; Datla, R.; Dhaubhadel, S.; Marsolais, F. Transcripts of sulphur metabolic genes are co-ordinately regulated in developing seeds of common bean lacking phaseolin and major lectins. *J. Exp. Bot.* **2012**, *63*, 6283–6295.

109. De La Fuente, M.; Borrajo, A.; Bermúdez, J.; Lores, M.; Alonso, J.; López, M.; Santalla, M.; De Ron, A.M.M.; Zapata, C.; Alvarez, G. 2-DE-based proteomic analysis of common bean (*Phaseolus vulgaris* L.) seeds. *J. Proteomics* **2011**, *74*, 262–267.

110. Natarajan, S.S.; Pastor-Corrales, M.A.; Khan, F.H.; Garrett, W.M. Proteomic analysis of common bean (*Phaseolus vulgaris* L.) by Two-Dimensional Gel Electrophoresis and Mass Spectrometry. *J. Basic Appl. Sci.* **2013**, 424–437.

111. Abid, G.; Muhovski, Y.; Jacquemin, J.-M.; Mingeot, D.; Sassi, K.; Toussaint, A.; Baudoin, J.-P. *In silico* identification and characterization of putative differentially expressed genes involved in common bean (*Phaseolus vulgaris* L.) seed development. *Plant Cell, Tissue Organ Cult.* **2011**, *107*, 341–353.

112. Avigad, G.; Dey, P.M. Carbohydrate metabolism: storage carbohydrates. In *Plant Biochemistry*; Elsevier, 1997; pp. 143–204 ISBN 0122146743.

113. Altamirano-Hernández, J.; López, M.G.; Acosta-Gallegos, J.A.; Farías-Rodríguez, R.; Peña-Cabriales, J.J. Influence of soluble sugars on seed quality in nodulated common bean (*Phaseolus vulgaris* L.): the case of trehalose. *Crop Sci.* **2007**, *47*, 1193–1205.

114. Braun, D.M.; Wang, L.; Ruan, Y.-L. Understanding and manipulating sucrose phloem loading, unloading, metabolism, and signalling to enhance crop yield and food security. *J. Exp. Bot.* **2014**, *65*, 1713–1735.

115. Sung, S.S.; Sheih, W.J.; Geiger, D.R.; Black, C.C. Growth, sucrose synthase, and invertase activities of developing *Phaseolus vulgaris* L. fruits. *Plant, Cell Environ.* **1994**, *17*, 419–426.

116. Abid, G.; Muhovski, Y.; Jacquemin, J.-M.; Mingeot, D.; Sassi, K.; Toussaint, A.; Baudoin, J.-P. Characterization and expression profile analysis of a sucrose synthase gene from common bean (*Phaseolus vulgaris* L.) during seed development. *Mol. Biol. Rep.* **2012**, *39*, 1133–1143.



117. Zhou, Y.; Qu, H.; Dibley, K.E.; Offler, C.E.; Patrick, J.W. A suite of sucrose transporters expressed in coats of developing legume seeds includes novel pH-independent facilitators. *Plant J.* **2007**, *49*, 750–764.

118. Jossier, M.; Bouly, J.-P.; Meimoun, P.; Arjmand, A.; Lessard, P.; Hawley, S.; Grahame Hardie, D.; Thomas, M. SnRK1 (SNF1-related kinase 1) has a central role in sugar and ABA signalling in *Arabidopsis thaliana*. *Plant J.* **2009**, *59*, 316–328.

119. Wang, F.; Ye, Y.; Chen, X.; Wang, J.; Chen, Z.; Zhou, Q. A sucrose non-fermenting-1-related protein kinase 1 gene from potato, StSnRK1, regulates carbohydrate metabolism in transgenic tobacco. *Physiol. Mol. Biol. Plants* **2017**, *23*, 933–943.

120. Coello, P.; Martínez-Barajas, E. SnRK1 is differentially regulated in the cotyledon and embryo axe of bean (*Phaseolus vulgaris* L) seeds. *Plant Physiol. Biochem.* **2014**, *80*, 153–9.

121. Coello, P.; Martínez-Barajas, E. The activity of SnRK1 is increased in *Phaseolus vulgaris* seeds in response to a reduced nutrient supply. *Front. Plant Sci.* **2014**, *5*, 196.

122. Bailly, C.; Audigier, C.; Ladonne, F.; Wagner, M.H.; Coste, F.; Corbineau, F.; Côme, D. Changes in oligosaccharide content and antioxidant enzyme activities in developing bean seeds as related to acquisition of drying tolerance and seed quality. *J. Exp. Bot.* **2001**, *52*, 701–708.

123. Coelho, C.M.M.; Tsai, S.M.; Vitorello, V.A. Dynamics of inositol phosphate pools (tris-, tetrakis- and pentakisphosphate) in relation to the rate of phytate synthesis during seed development in common bean (*Phaseolus vulgaris*). *J. Plant Physiol.* **2005**, *162*, 1–9.

124. Abid, G.; Sassi, K.; Muhovski, Y.; Jacquemin, J.-M.; Mingeot, D.; Tarchoun, N.; Baudoin, J.-P. Comparative expression and cellular localization of Myo-inositol Phosphate Synthase (MIPS) in the wild type and in an EMS mutant during common bean (*Phaseolus vulgaris*

L.) seed development. *Plant Mol. Biol. Report.* **2012**, *30*, 780–793.

125. Wang, Z.; Cheng, K.; Wan, L.; Yan, L.; Jiang, H.; Liu, S.; Lei, Y.; Liao, B. Genome-wide analysis of the basic leucine zipper (bZIP) transcription factor gene family in six legume genomes. *BMC Genomics* **2015**, *16*, 1–15.

126. Arenas-Huertero, C.; Pérez, B.; Rabanal, F.; Blanco-Melo, D.; De la Rosa, C.; Estrada-Navarrete, G.; Sanchez, F.; Covarrubias, A.A.; Reyes, J.L. Conserved and novel miRNAs in the legume *Phaseolus vulgaris* in response to stress. *Plant Mol. Biol.* **2009**, *70*, 385–401.

127. Valdés-López, O.; Yang, S.S.; Aparicio-Fabre, R.; Graham, P.H.; Reyes, J.L.; Vance, C.P.; Hernández, G. MicroRNA expression profile in common bean (*Phaseolus vulgaris*) under nutrient deficiency stresses and manganese toxicity. *New Phytol.* **2010**, *187*, 805–818.

128. Contreras-Cubas, C.; Rabanal, F. a; Arenas-Huertero, C.; Ortiz, M. a; Covarrubias, A. a; Reyes, J.L. The *Phaseolus vulgaris* miR159a precursor encodes a second differentially expressed microRNA. *Plant Mol. Biol.* **2012**, *80*, 103–115.

129. Han, J.; Xie, H.; Kong, M.L.; Sun, Q.P.; Li, R.Z.; Pan, J.B. Computational identification of miRNAs and their targets in *Phaseolus vulgaris*. *Genet. Mol. Res.* **2014**, *13*, 310–322.

130. Valdés-López, O.; Arenas-Huertero, C.; Ramírez, M.; Girard, L.; Sánchez, F.; Vance, C.P.; Luis Reyes, J.; Hernández, G. Essential role of MYB transcription factor: PvPHR1 and microRNA: PvmiR399 in phosphorus-deficiency signalling in common bean roots. *Plant. Cell Environ.* **2008**, *31*, 1834–1843.

131. Liu, J.-Q.; Allan, D.L.; Vance, C.P. Systemic signaling and local sensing of phosphate in common bean: cross-talk between photosynthate and MicroRNA399. *Mol. Plant* **2010**, *3*, 428–437.

132. Zargar, S.M.; Nazir, M.; Rai, V.; Hajduch, M.; Agrawal, G.K.; Rakwal, R. Towards a common bean proteome atlas: looking at the current state of research and the need for a comprehensive proteome. *Front. Plant Sci.* **2015**, *6*, 1–4.

# CHAPTER 2



# Differential proteomics reveals the hallmarks of seed development in common bean (*Phaseolus vulgaris* L.)

**Adapted from:**

Parreira JR, Bouraada J, Fitzpatrick MA, Silvestre S, Bernardes da Silva A, Marques da Silva J, Almeida AM, Fevereiro P, Altelaar AFM, Araújo SS. 2016. Differential proteomics reveals the hallmarks of seed development in common bean (*Phaseolus vulgaris* L.). Journal of Proteomics 143, 188–198.

**Author contributions to the chapter:**

José Ricardo Torrão Dias Parreira Salvado participated in the experimental setup and design, performed the experiments except for LC-MS/MS and corresponding analysis, analysed the data and wrote the chapter.

## Abstract

Common bean (*Phaseolus vulgaris* L.) is one of the most consumed staple foods worldwide. Little is known about the molecular mechanisms controlling seed development (SD). This study aims to comprehensively describe proteome dynamics during SD of common bean. A high-throughput gel-free proteomics approach (LC-MS/MS) was conducted on seeds 10, 20, 30 and 40 days after anthesis, spanning from late embryogenesis until desiccation. Of the 418 differentially accumulated proteins (DAPs) identified, 255 were characterized, most belonging to protein metabolism. An accumulation of DAPs belonging to the MapMan functional categories of “protein”, “glycolysis”, “TCA”, “DNA”, “RNA”, “cell” and “stress” were found at early SD stages, reflecting an extensive metabolic activity. In the intermediate SD stages, accumulation of storage, signalling, starch synthesis and cell wall-related proteins stood out. In the later SD stage, an increase in proteins related to redox, protein degradation/modification/folding and nucleic acid metabolisms reflect that seed desiccation-resistance mechanisms were activated. Our study unveils new clues to understand the regulation of SD mediated by PTMs and the maintenance of genome integrity. This knowledge enhances the understanding on the SD molecular mechanisms that may be used in the design and selection of common bean seeds with desired quality traits.

**Keywords:** Legume, Shotgun Proteomics, Seed Development, Common Bean

# Differential proteomics reveals the hallmarks of seed development in common bean (*Phaseolus vulgaris* L.)



## 1. Introduction

Common bean (*Phaseolus vulgaris* L.) seeds are one of the most consumed staple foods worldwide, with a global production estimated at around 12 million tons per year, mainly in Latin America and Africa [1]. Like other legumes, common bean seeds are rich in protein, carbohydrates, fibers, vitamins, minerals as well as other health-promoting phenolic compounds, being crucial for food security, nutrition and income source, particularly for local farmers [2–5]

Common bean is emerging as one of the best agricultural model crops to study seed biology. Its genome was recently released [6], several genomic and genetic resources are available [7], as well as, functional reverse genetics tools [8–10]. Moreover, this species has a high level of synteny with model legumes such as *Glycine max*, *Lotus japonicus* and *Medicago truncatula*, which is relevant for comparative genomic and genetic approaches [11,12]. Cumulative studies on common bean seed traits are emerging, addressing both seed phenotype markers for development, yield and quality, as well as environment interaction [3,13–18].

Seed quality traits depend on the accumulation of various storage molecules during the seed development (SD) process, which is influenced in turn by the genotype and adaptive changes to environment [19]. In the first SD stage (histodifferentiation/embryogenesis) differentiation of seed embryo, endosperm and seed coat occurs concomitantly with a slow

biomass accumulation rate [18]. During the second SD stage, known as maturation/filling phase, a fast continuous boost of biomass accumulation occurs and the embryo grows [18]. In the final stage, the seed starts to dehydrate, the accumulation of biomass ceases and mechanisms for embryo desiccation-tolerance are activated [20].

Seed development has been extensively studied in important crops such as rice [21,22], soybean [23–25] and maize [26,27] and model plants like Arabidopsis [28–30] and *M. truncatula* [31–36]. The *P. vulgaris* seed proteome is still poorly studied and there is a need for a comprehensive proteome resource for this species [37]. Changes in protein composition of *P. vulgaris* mature seeds lacking phaseolin were studied in 2010 by Marsolais and co-workers [38]. The authors described the compensatory molecular mechanisms for the absence of the major storage proteins, resulting in an increase in sulphur amino acid contents. A mature seed proteome map of *P. vulgaris,* usable as a reference for 2-DE assays was published in 2011 by De La Fuente [39]. In 2013, a more comprehensive proteome map of *P. vulgaris* mature seeds, identifying and functionally categorizing 141 protein spots, was published by Natarajan [40]. Nevertheless, the temporal changes occurring in the proteome during common bean seed development remain poorly addressed.



Herein, a high-throughput gel-free Liquid Chromatography-tandem Mass Spectrometry (LC-MS/MS) proteomics study was conducted to extend our knowledge on the molecular mechanisms underlying SD in the common bean. The broadness, comprehensiveness and high-throughput nature of LC-MS/MS (for a recent review on label-free proteomics methodology, refer to [41]) are advantages over 2-DE approaches. Four time-points were analysed, spanning from late embryogenesis until desiccation stage. Specific temporal protein accumulation patterns for each time point were described, providing a comprehensive seed proteome atlas with new information for

future development and selection of new common bean seeds with desired quality traits.

## 2. Materials and Methods

### 2.1 Plant material



In this study, *P. vulgaris* seeds from the SER 16 genotype were kindly supplied by Dr. Steve Beebe from the International Center for Tropical Agriculture (CGIAR, Cali, Colombia). Seeds were placed in water-soaked paper in Petri dishes at 27° C for 2 days, followed by 3 days at 23°C, always in the dark. Seedlings were transferred to watered vermiculite trays where they could grow under controlled conditions: 50% humidity, photoperiod of 16/8-hour day/night at 25/18°C, respectively, and light intensity of 400 µmol $m^{-2}$ $s^{-1}$ (Fitoclima 5.000 EH, ARALAB, Portugal). A week later, seedlings were individually transferred to 2.5L pots with a (3:1) mixture of commercial soil (Compo Sana S.A., Barcelona, Spain) and vermiculite, respectively. SER16 flowers were tagged and pods/seeds were harvested at 10, 20, 30 and 40 days after anthesis (DAA). Plants were kept in the environmental conditions described above during the assay and were watered 3 times per week. Harvested seeds were divided into two batches: one immediately frozen in liquid nitrogen and stored at -80ºC until further analysis and another one was used to measure seed length, fresh and dry weight to characterize the seed development process. Four biological replicates (individual plants) per seed developmental stage were used.

### 2.2 Protein Sample Extraction

One hundred milligrams of seed material were ground to a fine powder in liquid nitrogen using mortar and pestle. Ground material was solubilized in a buffer containing 8 M urea, 50 mM triethylammonium bicarbonate containing one tablet of cOmplete, Mini, EDTA-free Protease

Inhibitor Cocktail Tablets in EASYpacks (Roche Diagnostics GmbH, Mannheim, Germany). The samples were centrifuged (20,000g for 10 min at 4 ˚C) and the supernatants were reduced with 2 mM dithiothreitol (Fluka, Buch, Switzerland) for 30 min at 56 ˚C and alkylated with 4 mM iodoacetamide (Sigma-Aldrich, Steinheim, Germany) for 30 min at room temperature in the dark. The proteins were then digested with Lys-C (enzyme/substrate ratio 1:100) for 4h at 37 ˚C. Subsequently trypsin was added for an overnight digestion at 37 ˚C (enzyme/substrate ratio 1:100) (Promega, Madison, WI, USA). Peptides were separated and analysed using a LC-MS/MS system with a label-free approach.



## 2.3 LC-MS/MS and Data Analysis

Samples were analysed on a LTQ-Orbitrap Elite mass spectrometer (Thermo Scientific, Bremen, Germany) at the Netherlands Proteomic Center (Utrecht, Netherlands).

After being re-suspended in 10% formic acid, peptides were analysed using nanoflow reversed phase liquid chromatography on a Proxeon Easy-nLC 100 (Thermo Scientific, Bremen, Germany) connected to an LTQ-Orbitrap Elite mass spectrometer (Thermo Scientific, Bremen, DE). The peptides were first trapped with 100% buffer A (0.1% formic acid in water) on an in-house-made double-fritted trap column (Reprosil C18, 3 µm, 2 cm x 100 cm, Dr Maisch, GmbH Ammerbuch, Germany) before being separated on an in-house-made analytical column (2.7 µm, 50 µm x 50 cm, Agilent Zorbax SB-C18, Amstelveen, The Netherlands). The peptides were separated with a gradient of an aqueous (water, A) and an organic (acetonitrile, B) solvent, both with 0.1% formic acid. The 120 min gradient starts with 7% buffer B and goes up to 30% in 90 min and then to 100% B in 30 min at a flow rate of 150 nl/min. After the separation, peptides were introduced through a nanoelectrospray ion source by using an in-house-made fused silica needle (10 µm inner diameter tip) applied at a voltage of 1.7 kV and analysed using an Orbitrap Elite mass spectrometer. Full-scan

MS spectra from m/z 375 to 1600 were acquired in the Orbitrap at a resolution of 60 000. The 20 most intense ions were selected for collision-induced fragmentation in the linear ion trap at a normalized collision energy of 35%. The MaxQuant software (version 1.5.2.8) was used for the analysis of raw data [42,43].



The data was searched against the *P. vulgaris* L. UniProt FASTA database (version of June 3rd, 2015): (http://www.uniprot.org/uniprot/?query=*&fil=organism:"Phaseolus vulgaris (Kidney bean) (French bean) [3885]"). The following parameters were used for database search: initial tolerance of 20 ppm and a final mass tolerance of 7 ppm for precursor masses, 0.6 Da for CID ion trap fragment ions, and two missed cleavages allowed. The thresholds used were as follows: Minimum number of peptides: one, Orbitrap MS/MS match tolerance: 20 ppm, Orbitrap MS/MS de novo tolerance: 10 ppm, Orbitrap MS/MS de-isotoping tolerance: 7 ppm, Orbitrap top peaks per 100 Da: 12. Ion trap MS/MS match tolerance: 0.5 Da. Ion trap MS/MS de-isotoping tolerance: 0.15 Da, ion trap top peaks per 100 Da: eight, maximum peptide mass: 4600 Da, minimum peptide length for unspecific search: eight, maximum peptide length for unspecific search is 25. Carbamidomethylation was used as a fixed modification, and protein N-terminal acetylation and oxidation as a variable modification. The false discovery rate was set at 1% for peptide and proteins. The minimum peptide length specified was seven amino acids.

Data acquired from MaxQuant was analyzed using Perseus software (version 1.5.1.6). Intensities normalization was performed by subtracting the median of the log-transformed intensities for each nano-LC-MS/MS run. A two-sided two-sample t-test was performed with a permutation-based false discovery rate of 0.05. The t-test was performed using four biological replicates with comparisons between each subsequent time point (10 vs. 20, 20 vs. 30, 30 vs. 40). Results were plotted in volcano plots, showing P values

(-log10) depicted against difference. The positive constant s0 parameter was calculated as described [44] to ensure that variance of relative difference is independent of protein accumulation. We obtained values of 0.4 (10 vs 20), 0.6 (20 vs 30 and 30 vs 40) and 0.3 (early vs late). The s0-adjusted cut-off was used to identify significantly accumulated proteins (DAPs) in each test.

All raw mass spectrometry data files and MaxQuant output files were deposited in the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the dataset identifier PXD002254. All peptide spectral matches and their MS/MS spectra can be accessed in the freely available MaxQuant Viewer.



Additionally, the BLAST tool of UniProtKB (http://www.uniprot.org/, version of 10th July 2015) database was used to find homologues for uncharacterized proteins identified in *P.vulgaris*. Only query proteins that exhibited expectation values (E-values) = 0.0; query coverage ≥75%, max identity >90% and the same annotation at least in 3 plant species were accepted as protein homologues and its identity used.

## 2.4 Bioinformatic Analysis

Protein sequences retrieved from UniprotKB were used to create a mapping file for the Mercator pipeline from the MapMan webtools (http://mapman.gabipd.org/app/mercator) [45,46] to assign functional categorization to all DAPs.

Venn diagrams were obtained from the Bioinformatics & Evolutionary Genomics platform (http://bioinformatics.psb.ugent.be/webtools/Venn/) using the DAPs between each developmental stage transition.

To assess the relative distribution of all biological replicates of each experimental group, we used Principal Component Analysis (PCA) and the Statistica 6.0 software.

Protein–protein interaction (PPI) networks were retrieved using the publicly available program STRING v. 10 (http://string-db.org/). STRING is a database of known and predicted protein-protein interactions. We used all interaction sources available: the genomic context, high-throughput experiments, co-expression and previous published knowledge. All identified proteins were blasted against the *Glycine max* (soybean) protein database to obtain annotated protein entries for STRING predictions. Results with a confidence score higher than 0.900 and 0 putative interactions were selected. For the STRING prediction, the proteins were separated in six groups: DAPs down-accumulated in 10 DAA vs 20 DAA, DAPs up-accumulated in 10 DAA vs 20 DAA, DAPs down-accumulated in 20 DAA vs 30 DAA, DAPs up-accumulated in 20 DAA vs 30 DAA, DAPs down-accumulated in 30 DAA vs 40 DAA and DAPs up-accumulated in 30 DAA vs 40 DAA.



## 3. Results

### 3.1 Characterization of seed development: length, dry weight and fresh weight evolution

Fresh weight, dry weight and seed length were recorded throughout the seed development (**Figure 1**). Seed length and fresh weight increased considerably till 30 DAA, declining afterwards, as seed started to dehydrate. Similarly, dry weight also increased but stabilized between 30 DAA and 40 DAA. Seed coat color changed gradually during SD, from green (10 DAA) to dark red (40 DAA).

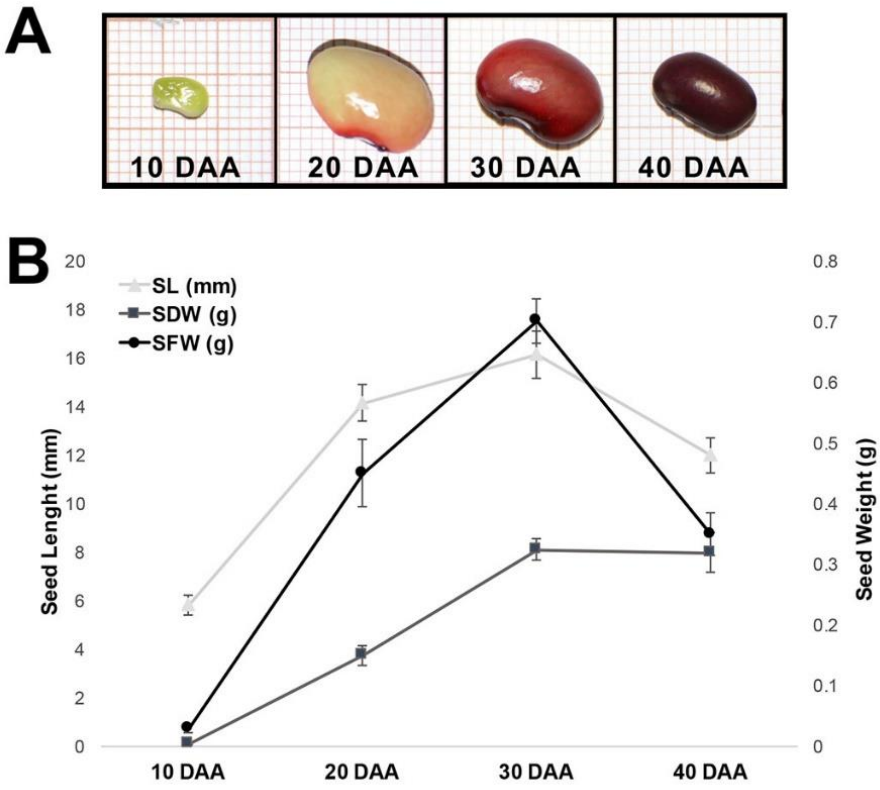


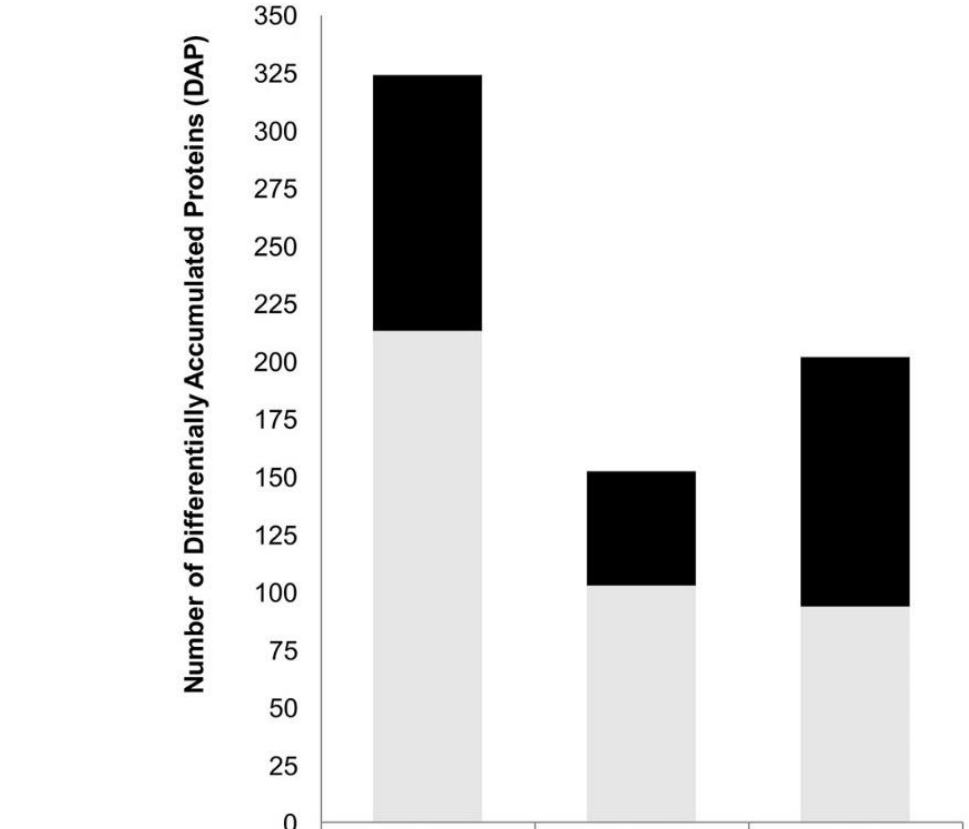

**Figure 1.** *Phaseolus vulgaris* seed development. DAA – Days after anthesis. A) Seed morphology; B) Seed fresh weight (SFW), seed dry weight (SDW), seed length (SL).

## 3.2 Proteins identified by LC-MS/MS and functional categorization

Four hundred and eighteen DAPs were retrieved from LC-MS/MS data analysis, in which pair wise comparisons were tested between samples (**Supplementary Figure 1**). Of these, 324 were differentially accumulated (DA) in the 10 DAA vs 20 DAA comparison (**Figure 2**), 152 in the 20 DAA vs 30 DAA comparison (**Figure 2**) and 202 in the 30 DAA vs 40 DAA (**Figure 2**).

Certain DAPs were restricted to specific sample comparison studied (**Figure 3**). One hundred and thirty-six proteins DA from the 10 DAA to the

20 DAA were restricted to this stage of seed development. Similarly, 26 and 48 proteins were DA in the comparison between 20 DAA and 30 DAA and between 30 DAA and 40 DAA, respectively.




| | 10 vs 20 DAA | 20 vs 30 DAA | 30 vs 40 DAA |
|---|---|---|---|
| DAP Up-accumulated | 111 | 49 | 108 |
| DAP Down-accumulated | 213 | 103 | 94 |

**Figure 2.** Number of differentially accumulated proteins (DAPs) in developing common bean seeds. The number of up -accumulated and down- accumulated DAPs is also displayed for the main SD comparison studied. DAA: days after anthesis.

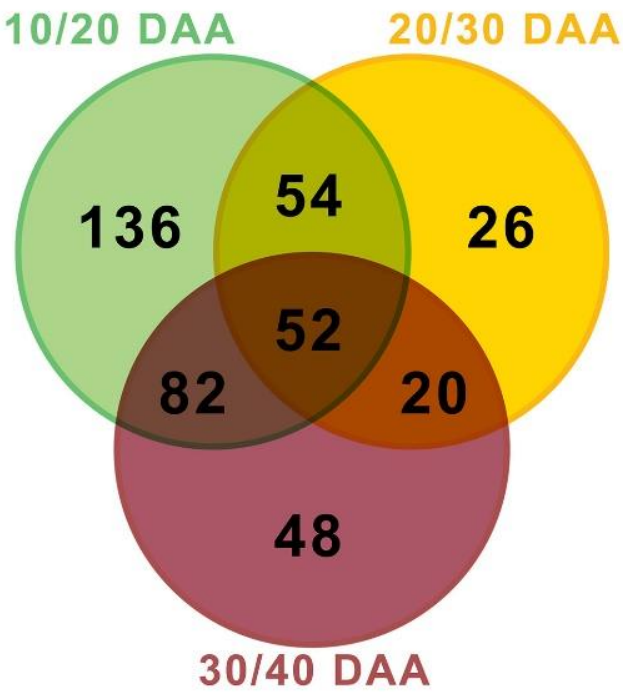





**Figure 3.** Venn diagram analysis of differentially accumulated proteins identified in developing common bean seeds. The overlapping regions denote common proteins among the main seed comparisons made. DAA: days after anthesis

MapMan functional categorization revealed that “protein” is the most representative category with 31.90% of DAPs, followed by “stress” (6.11%), “redox” (4.98%), “cell” (4.75%), “miscellaneous” (4.75%), “amino acid metabolism” (4.07%) and “development” (4.07%) (**Supplementary Table 1**). The percentage of DAPs belonging to each MapMan functional category varies depending on the seed sample comparison established (**Figure 4**).

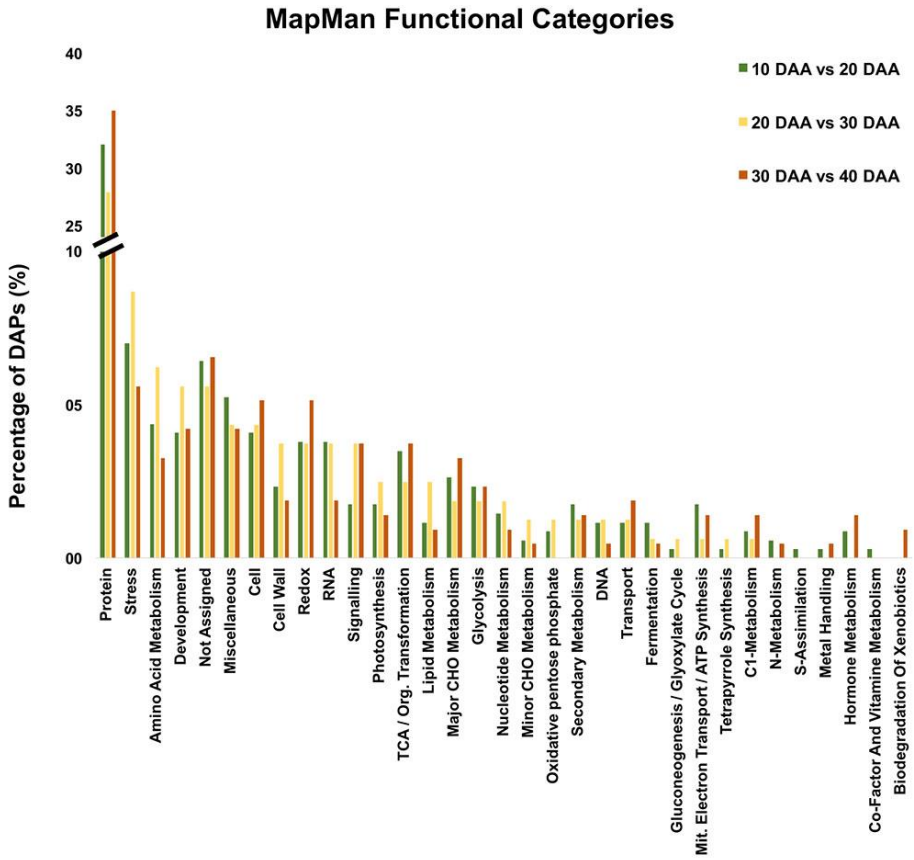



**Figure 4.** Functional categories of differentially accumulated proteins during common bean seed development. The percentage of proteins in each category were displayed between the main comparison studied (10 DAA vs 20 DAA; 20 DAA vs 30 DAA and 30 DAA vs 40 DAA). The percentage of proteins changed was calculated by comparison of the number of DAPs in each category in relation to the total of DAPs identified within each comparison.

From the 418 DAPs observed, 255 were annotated and considered characterized (**Supplementary Table 2**). PCA analysis conducted on annotated DAPs showed that replicates of each studied stage clustered together (**Figure 5**), discriminating well the 4 experimental groups. The protein-protein interactions, obtained using the String software (**Supplementary Figure 3**), showed the interaction of the metabolisms and proteins between each DAPs comparison studied.

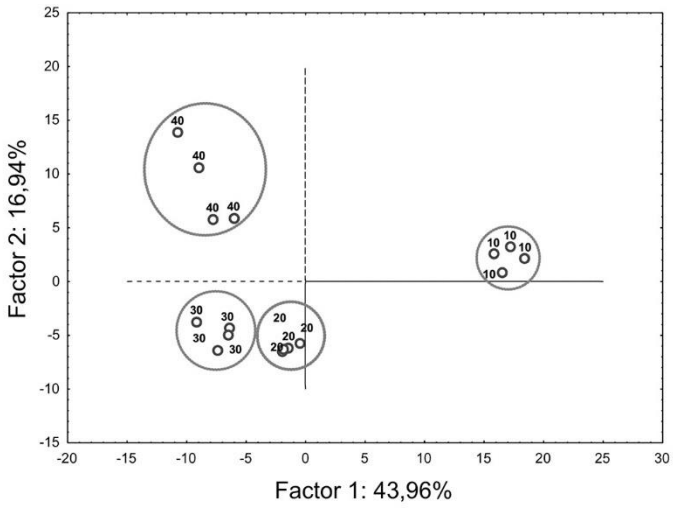



**Figure 5.** Principal component analysis (PCA) showed that all samples of the same seed development point cluster together. For PCA loadings the normalized intensity value for each characterized DAP was used. The first axis, factor 1, accounted for 43,96% of the variance, while factor 2 accounted for 16,94% of total variance between samples.

### 3.3.1 Protein metabolism

The most represented functional category was "protein" [MapMan Bin Code (BC 29)] with 141 DAPs. The majority of DAPs (57) were related to protein synthesis including 40S and 60S ribosomal proteins. While ELONGATION FACTOR 1-ALPHA (V7C080) and EUKARYOTIC TRANSLATION INITIATION FACTOR 3 SUBUNIT G (V7C6C6) showed decreased accumulation from the 10 to the 20 DAA, the ELONGATION FACTOR 1-ALPHA (V7BFQ4) showed enhanced accumulation for the same time points (**Supplementary Table 2**).

Forty-seven DAPs related to protein degradation were identified, being the majority (13) subunits of the 26S proteasome [e.g. 26S PROTEASOME NON-ATPASE REGULATORY SUBUNIT 1 (V7CQI1)]. DAPs related to ubiquitin pathway were also identified. The NEDD8-LIKE PROTEIN RUB2 (P0C031) showed decreased accumulation ($Log_2$ FC -1.12)

from the 10 to 20 DAA. The SKP1-LIKE PROTEIN (T2DMS9) had a $Log_2$ FC 0.96 from the 30 to 40 DAA.

Among proteases identified, the PROLINE IMINOPEPTIDASE (EC 3.4.11.5; V7BQN7) showed an increase ($Log_2$ FC 0.68) from the 30 to 40 DAA. The DNA DAMAGE-INDUCIBLE PROTEIN 1 (A0A0B2P2D7) showed a decreased accumulation ($Log_2$ FC -1.39) from the 10 to 20 DAA.



DAPs categorized as “protein targeting” (3), “amino acid activation” (3), “folding” (7) and “post-translational modification” (5) subcategories were identified.

Based on the relative abundance of all proteins categorized, average protein accumulation profiles were established for “protein” sub-categories: “protein synthesis, ribosomal proteins”, “protein synthesis, initiation, elongation and release”, “protein folding”, “protein amino acid activation”, “protein, targeting”, “protein, post translational modification”, “protein, degradation”, across time points studied (**Supplementary Figure 2**). The highest relative abundance was found at the 10 DAA samples. Throughout the time points studied a decrease in abundance was observed, irrespective of the sub-category studied. Nevertheless, few exceptions were noticed on the “targeting”, “folding” and “degradation” subcategories, in which the relative amounts tend to slightly increase in the last time point (**Supplementary Table 2**).

### 3.3.2 Amino acid and N metabolism

“Amino acid metabolism” (BC 13) encloses 18 DAPs related to amino acid synthesis and degradation. While the average relative abundance of DAPs included in “amino acid synthesis” subcategory decreased linearly from 10 DAA, the “amino acid degradation” subcategory slightly increased in the last time point (**Supplementary Figure 2**). Two proteins stood out with a different accumulation pattern: ARGININOSUCCINATE SYNTHASE,

CHLOROPLASTIC (EC 6.3.4.5; A0A0B2QMZ5 with a $Log_2$ FC 1.25 from the 10 to 20 DAA), and D-3-PHOSPHOGLYCERATE DEHYDROGENASE, CHLOROPLASTIC (EC 1.1.1.95; A0A0B2PBZ1 with a $Log_2$ FC 1.03 from the 30 to 40 DAA) (**Supplementary Table 2**).

GLUTAMINE SYNTHETASE (EC 6.3.1.2, V7B7Q6) and GLUTAMATE DEHYDROGENASE (V7CBZ3) were the only DAPs functionally categorized as "N-metabolism" (BC 12, **Supplementary Table 2**).



### 3.3.3 Tricarboxilic acid cycle and glycolysis metabolisms

Fourteen DAPs were categorized as belonging to the "tricarboxylic acid (TCA) cycle" (BC 8). On average TCA proteins decreased their abundance throughout SD until 30 DAA, increasing afterwards (**Supplementary Figure 2**). One of the exceptions is the CARBONIC ANHYDRASE (EC 4.2.1.1; V7C9Z9; $Log_2$ FC 3.06) which increase in abundance from the 10 to 20 DAA.

Eleven DAPs were classified into the "glycolysis" (BC 4) category. On average the abundance of these proteins decreased during SD (**Supplementary Figure 2**). While PHOSPHOGLUCOMUTASE (A0A0B2P5C9) abundance increases from the 10 to 20 DAA ($Log_2$ FC 2.35), the GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE (EC 1.2.1.-; V7D036) increase from 30 to 40 DAA ($Log_2$ FC 0.90) (**Supplementary Table 2**).

### 3.3.4 Major and minor carbohydrate metabolisms

Ten DAPs were categorized in "major carbohydrate metabolism" and 4 DAPs in "minor carbohydrate metabolism" (BC 2 and 3, respectively). Several enzymes of the "major carbon metabolism" category are involved in starch biosynthesis. These include GLUCOSE-1-PHOSPHATE ADENYLYLTRANSFERASE (V7BC72) and STARCH SYNTHASE (EC

2.4.1.-; Q9XIS6), which showed a significant increase in their abundance in the transition from 10 to 20 DAA ($Log_2$ FC 3.35 and 3.54, respectively, **Supplementary Table 2**). On average, starch synthesis proteins increased their abundance at 20 DAA, remaining stable afterwards (**Supplementary Figure 2**).



ISOAMYLASE-TYPE STARCH-DEBRANCHING ENZYME 3 (EC 3.2.1.68; A4PIT0) and STARCH SYNTHASE are only DA from the 10 DAA to the 20 DAA (**Figure 3**, **Supplementary Table 2**). The highest relative amount of SUCROSE SYNTHASE (EC 2.4.1.13; Q8GTA3; SuSy) was observed at 20 DAA, decreasing sharply afterwards (**Supplementary Table 2**).

Four DAPs were categorized in the “one-carbon (C1) metabolism” (BC 25) category. Among these, the FORMATE DEHYDROGENASE (D2DWA5) showed an enhanced accumulation from 10 to 20 DAA ($Log_2$ FC 2.84, **Supplementary Table 2**).

### 3.3.5 Stress and Redox metabolisms

Twenty-seven DAPs were categorized as “stress” (BC 20). Several DAPs were annotated as heat shock proteins, such as the HEAT SHOCK PROTEIN 83 (V7CJW6). While the tendency of these proteins is to decrease their relative abundance from 10 to 20 DAA, remaining stable afterwards (**Supplementary Figure 2**), others showed an opposite pattern. As an example, the COTYLEDONEOUS YIELDIN-LIKE PROTEIN (Q8H0C9) increases with a $Log_2$ FC 4.50 from 10 to 20 DAA, and $Log_2$ FC 2.56 from 20 to 30 DAA.

Among the 22 DAPs categorized as “redox” (BC 21), peroxiredoxins (EC 1.11.1.15; e.g. A0A0B2QLA9), SUPEROXIDE DISMUTASE (EC 1.15.1.1; T2DP08 and V5N8E1) and PROTEIN DISULFIDE-ISOMERASE (EC 5.3.4.1; e.g. V7AHC4) were found. Based on the average protein

abundance profiles established during SD, DAPs related with “redox” category increase in abundance throughout SD (**Supplementary Figure 2**). Proteins related with the glutathione-ascorbate cycle were annotated, such as the CYTOSOLIC GLUTATHIONE REDUCTASE (EC 1.8.1.7; A7XTY1) in which a $Log_2$ FC 1.01 from 30 to 40 DAA was recorded. Interestingly, the 1-CYS PEROXIREDOXIN (EC 1.11.1.15 / EC 1.11.1.7; A0A0B2RB28) stood out and was strongly accumulated ($Log_2$ FC 4.85) from 10 to 20 DAA.



### 3.3.6 RNA, DNA and Nucleotide metabolism

The functional categories “RNA” (BC 27) and “DNA” (BC 28) included 15 and 4 DAPs, respectively. The overall tendency showed that the average abundance profiles for DAPs in these categories decreases during SD, with a small increase towards 40 DAA (**Supplementary Figure 2**). In the “RNA” category, the NASCENT POLYPEPTIDE-ASSOCIATED COMPLEX SUBUNIT BETA (NAC, V7BWP6) had a decrease in abundance from 10 to 20 DAA ($Log_2$ FC -2.03) but increased in abundance from 30 to 40 DAA ($Log_2$ FC 1.18) (**Supplementary Table 2**).

DAPs categorized as “DNA” are related to DNA synthesis and structure. HISTONE H2A (V7BUR1) was more abundant at 10 DAA, decreasing to 20 and 30 DAA and finally increasing in abundance at 40 DAA ($Log_2$ FC of -2.08 from 10 to 20 DAA; and $Log_2$ FC 2 from 30 to 40 DAA). The PROGRAMMED CELL DEATH PROTEIN 4 (A0A0B2RMK9) also showed decreased accumulation between 10 to 20 DAA (**Supplementary Table 2**).

The HEAT SHOCK PROTEIN 83 (V7BEW3) and the ALLANTOINASE 1 (EC 3.5.2.5; T2DMK5) were the only DAPs categorized as “nucleotide metabolism” (BC 23), showing an abundance profile similar with “DNA” and “RNA” categories (**Supplementary Figure 2**).

### 3.3.7 Development

Eighteen DAPs were categorized in the “development” (BC 33) category, which includes storage or late embryogenesis abundant proteins. The overall analysis of the average of relative abundances of storage proteins showed an increase until 20 DAA, stabilizing afterwards **Supplementary Figure 2**). One PHASEOLIN (Q43632), three ALPHA PHASEOLIN (X5CHW7, X5D3J1 and P07219) and one LEGUMIN (F8QXP7) showed a high increase from 10 to 20 DAA ($Log_2$ FC between 6.36 and 11.75). The ALPHA-PHASEOLIN (X5CHW7) stood out since it abruptly decreases ($Log_2$ FC -4.02) between 30 to 40 DAA. GLUTELIN TYPE-A 1 (V7BYZ3) was the only protein that decreased its abundance during SD (**Supplementary Table 2**). GROUP 3 LATE EMBRYOGENESIS ABUNDANT PROTEIN (Q2N1E0) increased in abundance from 10 to 20 DAA in a $Log_2$ FC 5.76.

In the “miscellaneous” category (BC 26), PHYTOHEMAGGLUTININ (V5QN77 and V7C787) and ALPHA AMYLASE INHIBITOR-1 (A0T2V3) showed an accumulation profile similar to storage-related proteins.

### 3.3.8 Cell, cell wall and photosynthesis metabolism

“Cell” (BC 31) category encloses 21 DAPs related with cell organization, cell cycle and vesicle transport (**Supplementary Table 2**). On average the abundance of these proteins tends to decrease during SD (**Supplementary Figure 2**).

“Cell wall” (BC 10) category encloses 12 DAPs, mostly related with cell wall precursor synthesis. UDP-GLUCOSE 4-EPIMERASE (EC 5.1.3.2; A0A0B2Q9Z4) increases in abundance from 10 to 20 DAA, ($Log_2$ FC 3.75). UDP-GLUCOSE 6-DEHYDROGENASE (EC 1.1.1.22; V7C8C1) showed a sharp decrease from 30 to 40 DAA ($Log_2$ FC -4.81). The highest average

relative abundance for the DAPs included in this category was found at the 20 DAA decreasing afterwards (**Supplementary Figure 2**).

Eight DAPs belonging to the Calvin cycle and photorespiration metabolism were categorized as “photosynthesis” (BC 1). A general decrease in their abundance was seen until 30 DAA, with a small increase at the last time point (**Supplementary Table 2**). As example, the SERINE HYDROXYMETHYLTRANSFERASE (EC 2.1.2.1; V7BSE7) accumulation decreases from 10 to 20 DAA ($Log_2$ FC -2.35, **Supplementary Table 2**).



### 3.3.9 Signaling

Thirteen DAPs were included in the “signaling” (BC 30) category. The characterized proteins are G-proteins (e. g RAB GDP DISSOCIATION INHIBITOR ALPHA, A0A0B2ST35) and 14-3-3 proteins. The average protein abundance profile for this category showed an increase in abundance from 10 to 20 DAA, followed by a decrease through the remaining SD (**Figure 6**).

The RECEPTOR FOR ACTIVATED C KINASE 1 (RACK1; C8YZ71) slightly increases in abundance from 20 to 30 DAA ($Log_2$ FC 0.45), decreasing afterwards until 40 DAA ($Log_2$ FC -0.99).

### 3.3.10 Other metabolic categories

The category “secondary metabolism” (BC 16) encloses 7 DAPs of the mevalonic acid, flavonoids and glucosinolates pathways. One of the annotated proteins is the CHALCONE-FLAVONONE ISOMERASE FAMILY PROTEIN (V7AZH8) that decreases in abundance from 10 to 20 DAA in a $Log_2$ FC -2.36.

Five DAPs were categorized as “transport” (BC 34), 5 DAPs as “lipid metabolism” (BC 11) and 4 DAPs as “fermentation” (BC 5). The ALCOHOL DEHYDROGENASE 1 (EC 1.1.1.1; T2DLR9) stood out with a strong accumulation ($Log_2$ FC 5.11) from 10 to 20 DAA. The “mitochondrial electron

transport/ATP synthesis" (BC 9) category included 7 DAPs. Among them, the CYTOCHROME B-C1 COMPLEX SUBUNIT 7 (V7CXG4) increased in a $Log_2$ FC 2.12 from 10 to 20 DAA.

Four DAPs were categorized as "oxidative pentose phosphate pathway" (BC 7). An identical number of DAPs were categorized as "hormone metabolism" (BC 17). In the latter category, LIPOXYGENASE (V7BZK0; EC 1.13.11.-) increased in a $Log_2$ FC 10.31 from 10 to 20 DAA.



Categories with less than 2 DAPs were "co-factor and vitamin metabolism" (BC 18), "metal handling" (BC 15), "S-assimilation" (BC 14), "gluconeogenesis/glyoxilate cycle" (BC 6) and "tetrapyrrole synthesis" (BC 19). Twenty-one DAPs were categorized as "miscellaneous" (BC 26) including galactosidases, mannosidases, glutathione S transferases and lectins. Twenty-eight DAPs were classified as "not assigned" to any functional category in MapMan.

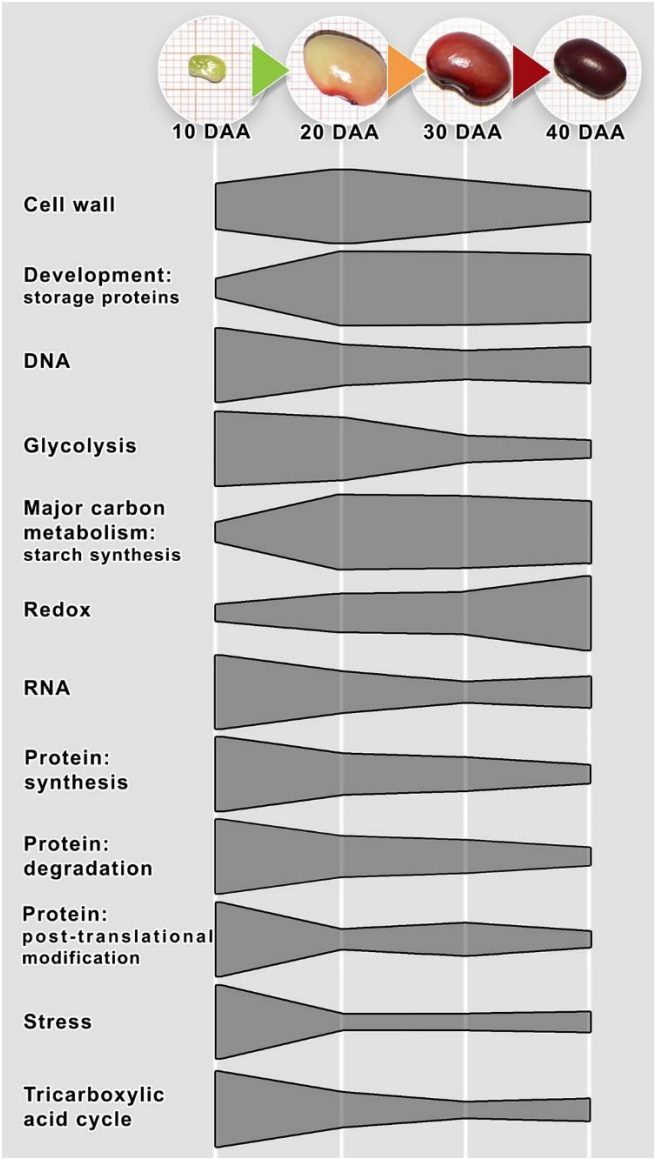



**Figure 6.** Schematic representation of the main changes in protein abundance profiles of selected MapMan functional categories during common bean seed development.

## 4. Discussion

In this work, a high-throughput gel-free (LC-MS/MS) proteomics study was conducted to extend our knowledge on the molecular mechanisms underlying SD in the common bean. Over 400 differentially accumulated proteins were identified and their profiles of relative abundance throughout

SD described, highlighting the diversity of molecular processes spanning from late embryogenesis until seed desiccation. “Protein” was the most represented MapMan functional category in our study. To the best of our knowledge, this is the first comprehensive seed proteome atlas available for common bean. The analysis of the protein abundance profiles (**Figure 5**, **Supplementary Table 2** and **Supplementary Figure 2**) together with the morphologic seed development characterization (**Figure 1**) allowed us to evidence the timeframe of the main metabolic pathways required for SD (**Figure 6**). These results are discussed in the following sections.



## 4.1 Early stages of seed development are defined by a metabolic shift for a high synthesis of cell structural components

Embryogenesis starts with a morphogenesis phase during which the embryo differentiates through several distinct stages (globular, heart, and torpedo) and ends at the cotyledonary stage when all embryo structures have been formed [34]. Previous studies in *P. vulgaris* seed development showed that around 12 days after pollination the embryo is in the cotyledonary stage [47]. Our results show that 10 DAA *P. vulgaris* seeds are at the late embryogenesis stage. At this stage there is still a high rate of cell division associated to embryo differentiation and morphogenesis, which demands for a high synthesis of cell structural components [18,20,34]. The accumulation of proteins belonging to the MapMan functional categories of “protein”, “glycolysis”, “TCA”, “DNA”, “RNA”, “cell” and “stress” metabolisms at 10 DAA confirmed the previous statement. Moreover, a decrease in abundance of several proteins with a role in cell division (tubulins and annexins) was found, while the accumulation of proteins involved in the synthesis of seed storage compounds on the transition from 10 to 20 DAA starts to be relevant. Such observations are in agreement with previous descriptions on other plant systems, namely *Medicago truncatula* [34] and rice [48].

Protein metabolism, with 141 annotated DAPs, was the most represented functional category (**Figure 4**). A relatively high accumulation of proteins associated with protein synthesis, folding, targeting and post-translational modification was found in the 10 DAA samples. Indeed, some of the translation initiation and elongation factors DA between 10 and 20 DAA in our study were previously described in *Lotus japonicus* during the first stage of SD [49].



Interestingly, an increased number and abundance of proteins related with degradation (proteolytic activity) was also observed at 10 DAA (**Figure 6**). This finding suggests that high protein turnover can occur simultaneously with protein biosynthesis during the early stages of SD, which has been previously described in *M. truncatula* [20,33]. Furthermore, the role of protein degradation and subsequent re-use of the amino acids in the synthesis of storage proteins cannot be neglected. Indeed, previous studies in *M. truncatula* showed the key role of proteolysis on nitrogen remobilization and the maintenance of an amino acid pool in the endosperm and seed coat for protein synthesis [33,34].

The ubiquitin/26S proteasome pathway is a major route for selectively degrading cytoplasmic and nuclear proteins in eukaryotes [50]. The regulatory role this proteolytic pathway in plant development is described [50–52]. Our study revealed ubiquitin-related proteins (e.g. UBIQUITIN CARBOXYL-TERMINAL HYDROLASE, SKP1-LIKE PROTEIN) with increased abundance at 10 DAA. Several subunits of the 26S proteasome are more abundant at 10 and 40 DAA, which reflects the late embryogenesis and seed desiccation stages. These overall results suggest the importance of ubiquitin/26S proteasome pathway regulating common bean SD.

Ubiquitin-NEDD8 has been implicated in the regulation of developmental processes such as embryogenesis [53,54]. NEDD8 can be conjugated with target proteins, such as the cullin subunits of cullin-RING E3 ubiquitin ligases inducing post-translational modifications [55]. UBIQUITIN-

NEDD8-LIKE PROTEIN RUB2 decreased significantly in abundance from 10 to 20 DAA. Based on this, it is tempting to speculate that several metabolic pathways could be regulated via neddylation at the early SD stages. Indeed, our STRING protein-protein interaction results for 10 vs. 20 DAA down accumulated (**Supplementary Figure 3**) are in agreement with this assumption. NEDD8 seems to interact with proteins related with glycolysis, TCA, protein degradation and stress. Nevertheless, more studies are needed to corroborate this hypothesis.



The high metabolic activity associated with embryo/seed differentiation and morphogenesis is energy-demanding. Consequently, it is not surprising that the increased accumulation of glycolysis and TCA cycle enzymes was found in the 10 DAA. Curiously, our results show an increased accumulation of ALCOHOL DEHYDROGENASE 1 from 10 to 20 DAA, suggesting that pyruvate may also be metabolized through alcoholic fermentation instead of following the TCA/oxidative phosphorylation pathway. Borisjuk & Rolletschek [56] and Rolletscheck [57] suggested that the development of relatively impermeable seed coat can induce hypoxia in developing seeds affecting gene expression levels, in vivo enzymatic activity, metabolite pool sizes and metabolic fluxes. Probably, this phenomenon is also observed in the early stages of SD in common bean. The accumulation of enzymes such as ALCOHOL DEHYDROGENASE and MALIC ENZYME may contribute to re-establish energy balance ensuring the proper time course of SD [58,59].

The intensive respiratory and metabolic activity seen can lead to an increase in ROS production [60]. In plants, ROS may damage cellular components and cause oxidative damage to DNA, leading to altered nucleotides, fragmentation and strand breaks [61]. ROS may also act as signalling molecules in development processes [62].

Chaperones have been previously described as relevant players for the correct folding and/or assembly of other proteins under stress conditions [63] and their involvement in oxidative stress response cannot be neglected. Our results suggest that, at some extent, oxidative stress may occur during SD and multiple evidences supports this hypothesis. Indeed, some components of the glutathione-ascorbate cycle, peroxiredoxins and chaperones were found DA from 10 to 20 DAA, suggesting a response to oxidative stress conditions. 1-CYS PEROXIREDOXIN also accumulates from 10 to 20 DAA. Probably this protein may act as a chaperone and/or a reducer of hydrogen peroxide [64]. We also observed LIPOXYGENASE accumulation from 10 to 20 DAA. Overexpression of lipoxygenase was previously related with a decrease in $H_2O_2$ in Arabidopsis [65].

The maintenance of genome integrity during zygotic embryo development is a crucial aspect with relevance in seed viability. Mechanisms must be activated to avoid the propagation of possible mutations during the intensive cell division rate associated with this developmental process.

In our study several proteins involved with DNA repair were DA. One of them is the PCNA which is more abundant at 10 DAA decreasing afterwards. This protein is a cofactor of DNA polymerase that can act as a scaffold protein recruiting proteins for DNA repair [66]. The HISTONE H2A and the DNA DAMAGE-INDUCIBLE PROTEIN 1 are also more abundant at 10 DAA. HISTONE H2A assures chromatin structure with impact in transcription, DNA repair, and chromosome segregation [67–70]. The DNA DAMAGE-INDUCIBLE PROTEIN 1 is a shuttle protein that takes the ubiquitinated protein to the 26S proteasome for degradation. The degradation of Ho endonuclease, which is part of a DNA damage response pathway is mediated by this process [71]. While genome integrity mechanisms have been widely characterized during seed germination or imbibition [72,73], the knowledge available about DNA damage and repair or chromatin remodelling during SD is scarce. Our work brings pioneer data

about how genome integrity (DNA repair/chromatin remodelling) is maintained during SD in legumes.

## 4.2 The accumulation of storage, signalling, starch synthesis and cell wall-related proteins is the molecular signature of seed filling



After the cessation of cell division, the seed storage compounds are synthesized in legumes [34]. On the 20 DAA, the accumulation of storage, signalling, starch synthesis and cell wall-related proteins stood out. This correlates with the high seed biomass accumulation observed from 10 to 20 DAA and maintained until 30 DAA (**Figure 1** and **Supplementary Figure 2**). Such fast continuous accumulation of biomass is characteristic of the seed filling stage, in which the synthesis of storage reserves occurs. [18,20,34,60]. Several enzymes involved in starch biosynthesis were strongly accumulated from 10 to 20 DAA (**Figure 6**, **Supplementary Table 2**) being maintained over the last SD stages. This reflects the high requirement for carbon and energy for rapid cell growth and synthesis of storage compounds. SUCROSE SYNTHASE (SuSy) has long been considered to be a biochemical marker for sink strength [74]. SuSy plays a central role in producing UDP-glucose (necessary for cell wall biosynthesis) and ADP-glucose (necessary for starch biosynthesis) [75]. In our results, a peak in the relative abundance of SuSy was found at 20 DAA, which may possibly correlate with starch accumulation characteristic of the seed filing stage. In plants, SNF1-RELATED KINASE (SnRK1) has been implicated in the control of carbohydrate and starch metabolism [76]. Coello *et al*. [77] demonstrated that SnRK1 activity is developmentally regulated in common bean seeds. SnRK1 activity exhibits a transient increase with the highest value at 20 DAA, which coincides with the beginning of protein and starch accumulation [77]. The accumulation of SuSy at 20 DAA seen in our results is in line with the previous statements.

Storage proteins such as PHASEOLIN, LEGUMIN, phytohemagglutinins, LECTIN and ALPHA AMYLASE INHIBITOR increased

in abundance from 10 to 20 DAA. Storage proteins constitute a source of energy and building blocks for the germinating seed [18,20,60] but are also relevant for the human dietary protein intake. For example, a single 7S globulin may also account for up to 50-55% of the total protein in dry bean [78]. Modulating the accumulation of storage proteins constitutes an exciting still open strategy to increase the nutritional value of pulses like common bean.



The accumulation of signalling proteins, like the 13-4-4 protein, and cell wall proteins, like the COTYLEDONEOUS YIELDIN-LIKE PROTEIN, followed the same accumulation pattern previously described for storage compounds. This reflects the overall stabilization in cell division and expansion, which characteristic of seed filling stage [18].

### 4.3 During maturation and desiccation an overall increase in proteins of the redox metabolism, protein degradation and folding and nucleic acids metabolism is seen

Seed length and fresh weight increased considerably up to the 30 DAA but a decline on these parameters was noticed from 30 to 40 DAA. This reduction is associated with the end of the seed growth and the onset of dehydration. This also marks the entrance in the quiescent state characteristic of orthodox seeds, such as those of *P. vulgaris*.

Mechanisms to ensure desiccation tolerance are activated halfway through the seed development, ensuring viability of the dried seeds [20]. Our data shows a strong increase in proteins related to the redox metabolism and, to a lesser extent, protein degradation/modification/folding and nucleic acids metabolism from 30 to 40 DAA.

The activity of chaperones have a crucial role on the protection of tissues under stress conditions [79]. The relative abundance of chaperones, as well as CHAPERONIN CPN60-2 and a subunit of T-complex protein 1,

members of the TCP-1/cpn60 chaperonin protein family suggest a possible role in the maintenance of cell structures, folding and storage of the synthesized proteins to ensure a successful germination, as described in *Pisum sativum* [80]. In our study, several forms of the PROTEIN DISULFIDE ISOMERASE (PDI) accumulated at 40 DAA. This enzyme that catalyzes disulfide formation and isomerization, ensures the correct folding of proteins [81]. In yeast, PDI expression was strongly induced under reducing agent [82], suggesting that this protein may play a major role in oxidative folding under stress conditions.



The accumulation of proteasome subunits and proteases was observed at 40 DAA, such as PROLINE IMINOPEPTIDASE. This suggests that some proteases may be stored in the mature seed, possibly to be used on the degradation of other storage proteins, during seed germination [83].

Also relevant during the late stages of SD is the accumulation of the HISTONE H2A. In this situation, this protein may exert a protective role on DNA repair, possibly rearranging the chromatin structure to cope with ROS generated during seed drying [70].

NAC proteins have a role in the assembly and transport of proteins synthesized de novo and in ribosome biogenesis [84]. A possible role of NAC in the assembly of the newly synthesized polypeptides has been previously described under salinity and drought [85,86]. We suggest that NAC may have a similar role during the seed desiccation, since an enhanced accumulation of NAC proteins was found at 40 DAA.

## 4.4 Post translational modification seems to play a role during seed development

Post-translational modifications (PTMs) seem to play an important role across *P. vulgaris* SD stages (**Supplementary Figure 2**). Our results suggested the relevance of ubiquitination and neddylation at the early SD

stages. New evidences for the role of phosphorylation in the regulation of SD were obtained in our study. The relative amount of RACK1 seems to be constant during SD until the transition to last stage (40 DAA), where a decrease in accumulation is noted.Our results suggest that RACK1 may play a central role in the molecular processes leading to the seed transition to a quiescent state. Indeed, STRING protein-protein interaction analysis (**Supplementary Figure 3**) seems to corroborate this assumption. In this analysis RACK1 interacts with proteins related to protein synthesis, possibly by regulating protein synthesis. RACK1 has been implicated in a range of biological functions, including protein translation, multiple hormonal responses, developmental processes, pathogen infection resistance and environmental stress responses, via a plethora of ligands that include kinases and phosphatases [90]. Speth *et al*. [89] also described a regulatory role of this protein in miRNA biogenesis. Nevertheless, more studies are needed to pinpoint the precise role of RACK1 and its mechanistic interactions on seed development.

Although several studies pointed out unclear roles of PTMs on seed development [91,92], our study unveils precise clues to further understand the regulation of seed development mediated by PTMs.

## 5. Conclusions

A gel-free high-throughput (LC-MS/MS) proteomics methodology was successfully used to comprehensively describe the proteome dynamics during seed development of *Phaseolus vulgaris*, one of the most consumed pulses worldwide. We identified over 400 DAPs and were able to characterize 255 of those proteins, in which the majority belongs to the “protein” category. Specific temporal protein abundance patterns were described highlighting the diversity of molecular processes spanning from the late embryogenesis to the seed desiccation stage.

An accumulation of proteins belonging to the MapMan functional categories of “protein”, “glycolysis”, “TCA”, “DNA”, “RNA”, “cell” and “stress” metabolisms were found at the early SD stages, reflecting an extensive metabolic activity. In the mid SD stages (20 and 30 DAA), the accumulation of storage, signalling, starch synthesis and cell wall-related proteins stood out, related with the synthesis of storage compounds. In the later SD stages, a strong increase in proteins related to the redox metabolism and, to a lesser extent, protein degradation/modification/folding and nucleic acids metabolism suggests that mechanisms for seed desiccation-resistance are activated.

Our study reveals precise clues to further understand the regulation of seed development mediated by PTMs. The understanding of the role of neddylation on SD constitutes a research topic innovative and challenging. New data about the maintenance of genome integrity (DNA repair/chromatin remodelling) during *P. vulgaris* SD was also unveiled. The data available enhances the understanding on the molecular mechanisms controlling seed development that may be used in the design and selection of common bean seeds with desired quality traits.

## 6. Acknowledgments

The authors acknowledge Dr. Steve Beebe from CIAT-CGIAR, Colombia, for kindly supplying the plant material in this study. Authors thank the financial support from Fundação para a Ciência e a Tecnologia (Lisbon, Portugal) through research projects PTDC/AGR-GPL/110224/2009, PTDC/AGR-TEC/3555/2012, PEst-OE/BIA/UI4046/2014, research unit “Biosystems & Integrative Sciences Institute” (UID/MULTI/04046/2013), research unit “GREEN-it: Bioresources for Sustainability" (UID/Multi/04551/2013), Plants for Life PhD grant of author JRP

(PD/BD/113474/2015) and present grant of SSA (SFRH/BPD/108032/2015). This work has been partly funded within the framework of PRIME-XS, grant number 262067, funded by the European Union 7th Framework Program. SSA also acknowledges a research contract in framework of project "Advanced Priming Technologies for the Lombardy Agro-Seed Industry-PRIMTECH' (Action 3, Code 2013-1727) sponsored by CARIPLO foundation and Regione Lombardia, Italy. The authors JRP and SSA are members of COST action FA 1306 - The quest for tolerant varieties - Phenotyping at plant and cellular level" to which networking support is acknowledged.



## 7. Supplementary Material

All supplementary materials are available at:
https://www.sciencedirect.com/science/article/abs/pii/S1874391916300586

**Supplementary Table 1. Percentage and number of differentially accumulated proteins included in MapMan functional categories.**

**Supplementary Table 2. MapMan functional categorization of differentially accumulated proteins across the time points studied.**

**Supplementary Figure 1. Volcano plots, showing P values (-log10) depicted against difference displaying the DAP during common bean seed development.** (A) Volcano plot showing P values (-log10) versus protein ratio 10/20 days. FDR=0.05; S=0.4. (B) Volcano plot showing P values (-log10) versus protein ratio 20/30 days. FDR=0.05, S=0.2. (C) Volcano plot showing P values (-log10) versus protein ratio 30/40 days. FDR=0.05, S=0.2.

**Supplementary Figure 2. Individual and composite protein abundance profiles of selected MapMan functional categories during common bean seed development.** The accumulation profiles for individual proteins (grey lines) are depicted. Individual accumulation profiles were calculated as the normalized intensity values minus the lowest intensity value observed in the study. The combined accumulation profiles for each category (red lines) are depicted. The combined accumulation profiles were calculated as the average of normalized intensity values of all DAPs in each category minus the lowest intensity value observed in the study (y axis) at each developmental stage (x axis). DAA: days after anthesis.

**Supplementary Figure 3. The protein-protein interaction network analysis among significantly accumulated proteins during during common bean seed development.** For the STRING prediction (confidence score = 0,9; 0 putative interactions), the proteins were separated in six groups: A) DAPs up accumulated in 10 DAA vs 20 DAA comparison; B) DAPs down accumulated in 10 DAA vs 20 DAA comparison; C) DAPs up accumulated in 20 DAA vs 30 DAA comparison, D) DAPs down accumulated in 20 DAA vs 30 DAA comparison; E) DAPs up accumulated in 30 DAA vs 40 DAA comparison and; F) DAPs down accumulated in 30 DAA vs 40 DAA comparison. DAA: days after anthesis.



## 8. Bibliography


1. Beebe, S. Common bean breeding in the tropics. In *Plant Breeding Reviews volume 36*; Janick, J., Ed.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, **2012**; Vol. 36, pp. 357–426 ISBN 9781118358566.

2. Araújo, S.S.; Beebe, S.; Crespi, M.; Delbreil, B.; González, E.M.; Gruber, V.; Lejeune-Henaut, I.; Link, W.; Monteros, M.J.; Prats, E.; *et al*. Abiotic stress responses in legumes: strategies used to cope with environmental challenges. *CRC. Crit. Rev. Plant Sci.* **2015**, *34*, 237–280.

3. Ribeiro, N.D.; Maziero, S.M.; Prigol, M.; Nogueira, C.W.; Rosa, D.P.; Possobom, M.T.D.F. Mineral concentrations in the embryo and seed coat of common bean cultivars. *J. Food Compos. Anal.* **2012**, *26*, 89–95.

4. Ghanbari, A.A.; Mousavi, S.H.; Pessarakli, M. Accumulation of reserve compounds in common bean seeds under drought stress. *J. Plant Nutr.* **2015**, *38*, 609–623.

5. Doria, E.; Campion, B.; Sparvoli, F.; Tava, A.; Nielsen, E. Anti-nutrient components and metabolites with health implications in seeds of 10 common bean (*Phaseolus vulgaris* L. and *Phaseolus lunatus* L.) landraces cultivated in southern Italy. *J. Food Compos. Anal.* **2012**, *26*, 72–80.

6. Schmutz, J.; McClean, P.E.; Mamidi, S.; Wu, G.A.; Cannon, S.B.; Grimwood, J.; Jenkins, J.; Shu, S.; Song, Q.; Chavarro, C.; *et al*. A reference genome for common bean and genome-wide analysis of dual domestications. *Nat. Genet.* **2014**, *46*, 707–713.

7. O'Rourke, J.A.; Iniguez, L.P.; Fu, F.; Bucciarelli, B.; Miller, S.S.;

Jackson, S.A.; McClean, P.E.; Li, J.; Dai, X.; Zhao, P.X.; *et al*. An RNA-Seq based gene expression atlas of the common bean. *BMC Genomics* **2014**, *15*, 866.

8. Pflieger, S.; Blanchet, S.; Meziadi, C.; Richard, M.; Thareau, V.; Mary, F.; Mazoyer, C.; Geffroy, V. The “one-step” Bean pod mottle virus (BPMV)-derived vector is a functional genomics tool for efficient overexpression of heterologous protein, virus-induced gene silencing and genetic mapping of BPMV R-gene in common bean (*Phaseolus vulgaris* L.). *BMC Plant Biol.* **2014**, *14*, 232.

9. Cabrera-Ponce, J.L.; López, L.; León-Ramírez, C.G.; Jofre-Garfias, A.E.; Verver-y-Vargas, A. Stress induced acquisition of somatic embryogenesis in common bean *Phaseolus vulgaris* L. *Protoplasma* **2014**, *252*, 559–570.

10. Collado, R.; Bermúdez-Caraballoso, I.; García, L.R.; Veitía, N.; Torres, D.; Romero, C.; Martirena-Ramírez, A.; Angenon, G. *Agrobacterium*-mediated transformation of *Phaseolus vulgaris* L. using indirect organogenesis. *Sci. Hortic. (Amsterdam).* **2015**, *195*, 89–100.

11. McClean, P.E.; Mamidi, S.; McConnell, M.; Chikara, S.; Lee, R. Synteny mapping between common bean and soybean reveals extensive blocks of shared loci. *BMC Genomics* **2010**, *11*, 184.

12. Hougaard, B.K.; Madsen, L.H.; Sandal, N.; de Carvalho Moretzsohn, M.; Fredslund, J.; Schauser, L.; Nielsen, A.M.; Rohde, T.; Sato, S.; Tabata, S.; *et al.* Legume anchor markers link syntenic regions between *Phaseolus vulgaris, Lotus japonicus, Medicago truncatula* and *Arachis*. *Genetics* **2008**, *179*, 2299–2312.

13. Porch, T.; Beaver, J.; Debouck, D.; Jackson, S.; Kelly, J.; Dempewolf, H. Use of wild relatives and closely related species to adapt common bean to climate change. *Agronomy* **2013**, *3*, 433–461.

14. Beebe, S.E.S.S.E.; Rao, I.M.I.I.M.I.I.M.I.; Blair, M.W.; Acosta-Gallegos, J.A. Phenotyping common beans for adaptation to drought. *Front. Physiol.* **2013**, *4*, 35.

15. Gupta, R.; Min, C.W.; Kim, S.W.; Wang, Y.; Agrawal, G.K.; Rakwal, R.; Kim, S.G.; Lee, B.W.; Ko, J.M.; Baek, I.Y.; *et al*. Comparative investigation of seed coats of brown- versus yellow-colored soybean seeds using an integrated proteomics and metabolomics approach. *Proteomics* **2015**, *15*, 1706–1716.

16. Muasya, R.M.; Lommen, W.J.M.; Muui, C.W.; Struik, P.C. How weather during development of common bean (*Phaseolus vulgaris* L.)

affects the crop's maximum attainable seed quality. *NJAS - Wageningen J. Life Sci.* **2008**, *56*, 85–100.

17. Clavijo Michelangeli, J.; Bhakta, M.; Gezan, S.; Boote, K.; Vallejos, C.E. From flower to seed: identifying phenological markers and reliable growth functions to model reproductive development in the common bean (*Phaseolus vulgaris* L.). *Plant. Cell Environ.* **2013**, *36*, 2046–58.

18. Coelho, C.; Benedito, V.; Maria, C.; Coelho, M.; Augusto, V.; Coelho, C.; Benedito, V.; Maria, C.; Coelho, M.; Augusto, V.; *et al*. Seed development and reserve compound accumulation in common bean (*Phaseolus vulgaris* L.). *Seed Sci. Biotechnol* **2008**, *2*, 42–52.

19. Mondoni, A.; Orsenigo, S.; Donà, M.; Balestrazzi, A.; Probert, R.J.; Hay, F.R.; Petraglia, A.; Abeli, T. Environmentally induced transgenerational changes in seed longevity: maternal and genetic influence. *Ann. Bot.* **2014**, *113*, 1257–1263.

20. Wang, W.-Q.; Liu, S.-J.; Song, S.-Q.; Møller, I.M. Proteomics of seed development, desiccation tolerance, germination and vigor. *Plant Physiol. Biochem.* **2015**, *86*, 1–15.

21. Deng, Z.Y.; Gong, C.Y.; Wang, T. Use of proteomics to understand seed development in rice. *Proteomics* **2013**, *13*, 1784–1800.

22. Peng, T.; Sun, H.; Qiao, M.; Zhao, Y.; Du, Y.; Zhang, J.; Li, J.; Tang, G.; Zhao, Q. Differentially expressed microRNA cohorts in seed development may contribute to poor grain filling of inferior spikelets in rice. *BMC Plant Biol.* **2014**, *14*, 196.

23. Jones, S.I.; Vodkin, L.O. Using RNA-Seq to profile soybean seed development from fertilization to maturity. *PLoS One* **2013**, *8*, e59270.

24. Liu, F.; Jensen, C.R.; Andersen, M.N. Drought stress effect on carbohydrate concentration in soybean leaves and pods during early reproductive development: Its implication in altering pod set. *F. Crop. Res.* **2004**, *86*, 1–13.

25. Jones, S.I.; Gonzalez, D.O.; Vodkin, L.O. Flux of transcript patterns during soybean seed development. *BMC Genomics* **2010**, *11*, 136.

26. Gu, Y.; Liu, Y.; Zhang, J.; Liu, H.; Hu, Y.; Du, H.; Li, Y.; Chen, J.; Wei, B.; Huang, Y. Identification and characterization of microRNAs in the developing maize endosperm. *Genomics* **2013**, *102*, 472–478.

27. Kang, M.; Zhao, Q.; Zhu, D.; Yu, J. Characterization of microRNAs expression during maize seed development. *BMC Genomics* **2012**,

*13*, 360.

28. Le, B.H.; Cheng, C.; Bui, A.Q.; Wagmaister, J.A.; Henry, K.F.; Pelletier, J.; Kwong, L.; Belmonte, M.; Kirkbride, R.; Horvath, S.; *et al*. Global analysis of gene activity during *Arabidopsis* seed development and identification of seed-specific transcription factors. *Proc. Natl. Acad. Sci.* **2010**, *107*, 8063–8070.

29. Gallardo, K.; Job, C.; Groot, S.P.; Puype, M.; Demol, H.; Vandekerckhove, J.; Job, D. Proteomic analysis of arabidopsis seed germination and priming. *Plant Physiol.* **2001**, *126*, 835–848.



30. Galland, M.; Job, D.; Rajjou, L. The Seed Proteome Web Portal. *Front. Plant Sci.* **2012**, *3*, 98.

31. Verdier, J.; Kakar, K.; Gallardo, K.; Le Signor, C.; Aubert, G.; Schlereth, A.; Town, C.D.; Udvardi, M.K.; Thompson, R.D. Gene expression profiling of *M. truncatula* transcription factors identifies putative regulators of grain legume seed filling. *Plant Mol. Biol.* **2008**, *67*, 567–580.

32. Verdier, J.; Lalanne, D.; Pelletier, S.; Torres-Jerez, I.; Righetti, K.; Bandyopadhyay, K.; Leprince, O.; Chatelain, E.; Vu, B.L.; Gouzy, J.; et al. A regulatory network-based approach dissects late maturation processes related to the acquisition of desiccation tolerance and longevity of *Medicago truncatula* seeds. *Plant Physiol.* **2013**, *163*, 757–774.

33. Gallardo, K.; Firnhaber, C.; Zuber, H.; Héricher, D.; Belghazi, M.; Henry, C.; Küster, H.; Thompson, R. A combined proteome and transcriptome analysis of developing *Medicago truncatula* seeds. *Mol. Cell. Proteomics* **2007**, *6*, 2165–2179.

34. Gallardo, K.; Le Signor, C.; Vandekerckhove, J.; Thompson, R.D.; Burstin, J. Proteomics of *Medicago truncatula* seed development establishes the time frame of diverse metabolic processes related to reserve accumulation. *Plant Physiol.* **2003**, *133*, 664–682.

35. Wang, M.; Verdier, J.; Benedito, V.A.; Tang, Y.; Murray, J.D.; Ge, Y.; Becker, J.D.; Carvalho, H.; Rogers, C.; Udvardi, M.; *et al*. LegumeGRN: a gene regulatory network prediction server for functional and comparative studies. *PLoS One* **2013**, *8*, e67434.

36. Verdier, J.; Thompson, R.D. Transcriptional regulation of storage protein synthesis during dicotyledon seed filling. *Plant Cell Physiol.* **2008**, *49*, 1263–1271.

37. Zargar, S.M.; Nazir, M.; Rai, V.; Hajduch, M.; Agrawal, G.K.; Rakwal, R. Towards a common bean proteome atlas: looking at the current

state of research and the need for a comprehensive proteome. *Front. Plant Sci.* **2015**, *6*, 1–4.

38. Marsolais, F.; Pajak, A.; Yin, F.; Taylor, M.; Gabriel, M.; Merino, D.M.; Ma, V.; Kameka, A.; Vijayan, P.; Pham, H.; *et al*. Proteomic analysis of common bean seed with storage protein deficiency reveals up-regulation of sulfur-rich proteins and starch and raffinose metabolic enzymes, and down-regulation of the secretory pathway. *J. Proteomics* **2010**, *73*, 1587–1600.



39. De La Fuente, M.; Borrajo, A.; Bermúdez, J.; Lores, M.; Alonso, J.; López, M.; Santalla, M.; De Ron, A.M.M.; Zapata, C.; Alvarez, G. 2-DE-based proteomic analysis of common bean (*Phaseolus vulgaris* L.) seeds. *J. Proteomics* **2011**, *74*, 262–267.

40. Natarajan, S.S.; Pastor-Corrales, M.A.; Khan, F.H.; Garrett, W.M. Proteomic analysis of common bean (*Phaseolus vulgaris* L.) by Two-Dimensional Gel Electrophoresis and Mass Spectrometry. *J. Basic Appl. Sci.* **2013**, 424–437.

41. Jorrín-Novo, J. V.; Pascual, J.; Sánchez-Lucas, R.; Romero-Rodríguez, M.C.; Rodríguez-Ortega, M.J.; Lenz, C.; Valledor, L. Fourteen years of plant proteomics reflected in Proteomics : Moving from model species and 2DE-based approaches to orphan species and gel-free platforms. *Proteomics* **2015**, *15*, 1089–1112.

42. Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26*, 1367–1372.

43. Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. a.; Olsen, J. V.; Mann, M. Andromeda: A peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **2011**, *10*, 1794–1805.

44. Tusher, V.G.; Tibshirani, R.; Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci.* **2001**, *98*, 5116–5121.

45. Usadel, B.; Poree, F.; Nagel, A.; Lohse, M.; Czedik-Eysenberg, A.; Stitt, M. A guide to using MapMan to visualize and compare Omics data in plants: A case study in the crop species, Maize. *Plant, Cell Environ.* **2009**, *32*, 1211–1229.

46. Thimm, O.; Bläsing, O.; Gibon, Y.; Nagel, A.; Meyer, S.; Krüger, P.; Selbig, J.; Müller, L.A.; Rhee, S.Y.; Stitt, M. MAPMAN: A user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* **2004**, *37*, 914–939.

47. Abid, G.; Sassi, K.; Muhovski, Y.; Jacquemin, J.-M.; Mingeot, D.; Tarchoun, N.; Baudoin, J.-P. Identification and analysis of differentially expressed genes during seed development using Suppression Subtractive Hybridization (SSH) in *Phaseolus vulgaris*. *Plant Mol. Biol. Report.* **2011**, *30*, 719–730.

48. Jami, S.K.; Clark, G.B.; Ayele, B.T.; Roux, S.J.; Kirti, P.B. Identification and characterization of annexin gene family in rice. *Plant Cell Rep.* **2012**, *31*, 813–825.

49. Nautrup-Pedersen, G.; Dam, S.; Laursen, B.S.; Siegumfeldt, A.L.; Nielsen, K.; Goffard, N.; Stærfeldt, H.H.; Friis, C.; Sato, S.; Tabata, S.; et al. Proteome analysis of pod and seed development in the model legume *Lotus japonicus*. *J. Proteome Res.* **2010**, *9*, 5715–5726.

50. Smalle, J.; Vierstra, R.D. The ubiquitin 26S proteasome proteolytic pathway. *Annu. Rev. Plant Biol.* **2004**, *55*, 555–590.

51. Smalle, J.; Kurepa, J.; Yang, P.; Emborg, T.J.; Babiychuk, E.; Kushnir, S.; Vierstra, R.D.; Read, C.; Smalle, J.; Kurepa, J.; *et al*. The pleiotropic role of the 26s proteasome subunit RPN10 in Arabidopsis growth and development supports a substrate-specific function in abscisic acid signaling. **2003**, *15*, 965–980.

52. Vierstra, R.D. The ubiquitin-26S proteasome system at the nexus of plant biology. *Nat. Rev. Mol. Cell Biol.* **2009**, *10*, 385–397.

53. Mergner, J.; Schwechheimer, C. The NEDD8 modification pathway in plants. *Front. Plant Sci.* **2014**, *5*, 103.

54. Bostick, M. Related to Ubiquitin 1 and 2 are redundant and essential and regulate vegetative growth, auxin signaling, and ethylene production in Arabidopsis. *Plant Cell Online* **2004**, *16*, 2418–2432.

55. Mergner, J.; Heinzlmeir, S.; Kuster, B.; Schwechheimer, C. DENEDDYLASE1 deconjugates NEDD8 from non-cullin protein substrates in *Arabidopsis thaliana*. *Plant Cell* **2015**, *27*, 741–753.

56. Borisjuk, L.; Rolletschek, H. The oxygen status of the developing seed. *New Phytol.* **2009**, *182*, 17–30.

57. Rolletschek, H. Legume embryos develop in a hypoxic environment. *J. Exp. Bot.* **2002**, *53*, 1099–1107.

58. Selinski, J.; König, N.; Wellmeyer, B.; Hanke, G.T.; Linke, V.; Neuhaus, H.E.; Scheibe, R. The plastid-localized NAD-dependent malate dehydrogenase is crucial for energy homeostasis in developing *Arabidopsis thaliana* seeds. *Mol. Plant* **2014**, *7*, 170–186.

59. Geigenberger, P. Response of plant metabolism to too little oxygen. *Curr. Opin. Plant Biol.* **2003**, *6*, 247–256.

60. Bailly, C.; Bailly, C. Active oxygen species and antioxidants in seed biology. *Seed Sci. Res.* **2004**, *14*, 93–107.

61. Roldán-Arjona, T.; Ariza, R.R. Repair and tolerance of oxidative DNA damage in plants. *Mutat. Res. - Rev. Mutat. Res.* **2009**, *681*, 169–179.

62. Møller, I.M.; Sweetlove, L.J. ROS signalling – specificity is required. *Trends Plant Sci.* **2010**, *15*, 370–374.

63. Finka, A.; Mattoo, R.U.H.; Goloubinoff, P. Meta-analysis of heat- and chemically upregulated chaperone genes in plant and human cells. *Cell Stress Chaperones* **2011**, *16*, 15–31.

64. Kim, S.Y.; Paeng, S.K.; Nawkar, G.M.; Maibam, P.; Lee, E.S.; Kim, K.-S.S.; Lee, D.H.; Park, D.-J.J.; Kang, S. Bin; Kim, M.R.; et al. The 1-Cys Peroxiredoxin, a regulator of seed dormancy, functions as a molecular chaperone under oxidative stress conditions. *Plant Sci.* **2011**, *181*, 119–124.

65. Lim, C.W.; Han, S.-W.; Hwang, I.S.; Kim, D.S.; Hwang, B.K.; Lee, S.C. The pepper lipoxygenase CaLOX1 plays a role in osmotic, drought and high salinity stress response. *Plant Cell Physiol.* **2015**, *56*, 930–942.

66. Moldovan, G.-L.L.; Pfander, B.; Jentsch, S. PCNA, the Maestro of the replication fork. *Cell* **2007**, *129*, 665–679.

67. Donà, M.; Mittelsten Scheid, O. DNA Damage Repair in the context of plant chromatin. *Plant Physiol.* **2015**, *168*, 1206–1218.

68. Kawashima, T.; Lorković, Z.J.; Nishihama, R.; Ishizaki, K.; Axelsson, E.; Yelagandula, R.; Kohchi, T.; Berger, F. Diversification of histone H2A variants during plant evolution. *Trends Plant Sci.* **2015**, *20*, 419–425.

69. Yelagandula, R.; Stroud, H.; Holec, S.; Zhou, K.; Feng, S.; Zhong, X.; Muthurajan, U.M.; Nie, X.; Kawashima, T.; Groth, M.; *et al*. The histone variant H2A.W Defines heterochromatin and promotes chromatin condensation in Arabidopsis. *Cell* **2014**, *158*, 98–109.

70. Moore, J.D.; Yazgan, O.; Ataian, Y.; Krebs, J.E. Diverse roles for histone H2A modifications in DNA damage response pathways in yeast. *Genetics* **2007**, *176*, 15–25.

71. Kaplun, L.; Ivantsiv, Y.; Kornitzer, D.; Raveh, D. Functions of the DNA

damage response pathway target Ho endonuclease of yeast for degradation via the ubiquitin-26S proteasome system. *Proc. Natl. Acad. Sci.* **2000**, *97*, 10077–10082.

72. Gursoy-Yuzugullu, O.; Ayrapetov, M.K.; Price, B.D. Histone chaperone Anp32e removes H2A.Z from DNA double-strand breaks and promotes nucleosome reorganization and DNA repair. *Proc. Natl. Acad. Sci.* **2015**, *112*, 7507–7512.

73. Balestrazzi, A.; Donà, M.; Macovei, A.; Sabatini, M.E.; Pagano, A.; Carbonera, D. DNA repair and telomere maintenance during seed imbibition: correlation of transcriptional patterns. *Telomere and Telomerase* **2015**, 2–5.

74. Braun, D.M.; Wang, L.; Ruan, Y.-L. Understanding and manipulating sucrose phloem loading, unloading, metabolism, and signalling to enhance crop yield and food security. *J. Exp. Bot.* **2014**, *65*, 1713–1735.

75. Sturm, A.; Tang, G.-Q. The sucrose-cleaving enzymes of plants are crucial for development, growth and carbon partitioning. *Trends Plant Sci.* **1999**, *4*, 401–407.

76. Jossier, M.; Bouly, J.-P.; Meimoun, P.; Arjmand, A.; Lessard, P.; Hawley, S.; Grahame Hardie, D.; Thomas, M. SnRK1 (SNF1-related kinase 1) has a central role in sugar and ABA signalling in *Arabidopsis thaliana*. *Plant J.* **2009**, *59*, 316–328.

77. Coello, P.; Martínez-Barajas, E. SnRK1 is differentially regulated in the cotyledon and embryo axe of bean (*Phaseolus vulgaris* L) seeds. *Plant Physiol. Biochem.* **2014**, *80*, 153–9.

78. Sathe, S.K. Dry Bean Protein Functionality. *Crit. Rev. Biotechnol.* **2002**, *22*, 175–223.

79. Wang, W.; Vinocur, B.; Shoseyov, O.; Altman, A. Role of plant heat-shock proteins and molecular chaperones in the abiotic stress response. *Trends Plant Sci.* **2004**, *9*, 244–252.

80. Wang, W.-Q.; Møller, I.M.; Song, S.-Q. Proteomic analysis of embryonic axis of *Pisum sativum* seeds during germination and identification of proteins associated with loss of desiccation tolerance. *J. Proteomics* **2012**, *77*, 68–86.

81. Wilkinson, B.; Gilbert, H.F. Protein disulfide isomerase. *Biochim. Biophys. Acta - Proteins Proteomics* **2004**, *1699*, 35–44.

82. Gasch, A.P.; Spellman, P.T.; Kao, C.M.; Carmel-Harel, O.; Eisen, M.B.; Storz, G.; Botstein, D.; Brown, P.O. Genomic expression

programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **2000**, *11*, 4241–4257.

83. Tan-Wilson, A.L.; Wilson, K.A. Mobilization of seed protein reserves. *Physiol. Plant.* **2012**, *145*, 140–153.

84. Kogan, G.L.; Gvozdev, V. a. Multifunctional nascent polypeptide-associated complex (NAC). *Mol. Biol.* **2014**, *48*, 189–196.

85. Karan, R.; Subudhi, P.K. Overexpression of a nascent polypeptide associated complex gene (SaβNAC) of *Spartina alterniflora* improves tolerance to salinity and drought in transgenic Arabidopsis. *Biochem. Biophys. Res. Commun.* **2012**, *424*, 747–752.



86. Kang, G.; Ma, H.; Liu, G.; Han, Q.; Li, C.; Guo, T. Silencing of TaBTF3 gene impairs tolerance to freezing and drought stresses in wheat. *Mol. Genet. Genomics* **2013**, *288*, 591–599.

87. Chen, J.-G. RACK1 mediates multiple hormone responsiveness and developmental processes in *Arabidopsis*. *J. Exp. Bot.* **2006**, *57*, 2697–2708.

88. Islas-Flores, T.; Guillén, G.; Sánchez, F.; Villanueva, M. a. Changes in RACK1 expression induce defects in nodulation and development in *Phaseolus vulgaris*. *Plant Signal. Behav.* **2012**, *7*, 132–134.

89. Speth, C.; Willing, E.-M.; Rausch, S.; Schneeberger, K.; Laubinger, S. RACK1 scaffold proteins influence miRNA abundance in Arabidopsis. *Plant J.* **2013**, *76*, 433–445.

90. Islas-Flores, T.; Rahman, A.; Ullah, H.; Villanueva, M.A. The Receptor for Activated C Kinase in plant signaling: tale of a promiscuous little molecule. *Front. Plant Sci.* **2015**, *6*, 1–19.

91. Dam, S.; Laursen, B.S.; Ornfelt, J.H.; Jochimsen, B.; Staerfeldt, H.H.; Friis, C.; Nielsen, K.; Goffard, N.; Besenbacher, S.; Krusell, L.; *et al*. The proteome of seed development in the model legume *Lotus japonicus*. *Plant Physiol.* **2009**, *149*, 1325–40.

92. Agrawal, G.K.; Hajduch, M.; Graham, K.; Thelen, J.J. In-depth investigation of the soybean seed-filling proteome and comparison with a parallel study of rapeseed. *Plant Physiol.* **2008**, *148*, 504–518.

# CHAPTER 3



# Maintaining genome integrity during seed development in *Phaseolus vulgaris* L.: evidence from a transcriptomic profiling study

**Adapted from:**

Parreira J, Balestrazzi A, Fevereiro P, Araújo S. 2018. Maintaining Genome Integrity during Seed Development in *Phaseolus vulgaris* L.: Evidence from a Transcriptomic Profiling Study. Genes 9, 463.

**Author contributions to the chapter:**

José Ricardo Torrão Dias Parreira Salvado participated in the experimental setup and design, performed the experiments except for Massive Analysis of 3'-cDNA Ends and corresponding analysis, analysed the data and wrote the chapter.

## Abstract

The maintenance of genome integrity is crucial in seeds, due to the constant challenge of several endogenous and exogenous factors. The knowledge concerning DNA damage response and chromatin remodelling during seed development is still scarce, especially in *Phaseolus vulgaris* L. A transcriptomic profiling of the expression of genes related to DNA damage response/chromatin remodelling mechanisms was performed in *P. vulgaris* seeds at four distinct developmental stages, spanning from late embryogenesis to seed desiccation. Of the 14,001 expressed genes identified using massive analysis of cDNA ends, 301 belong to the DNA MapMan category. In late embryogenesis, a high expression of genes related to DNA damage sensing and repair suggests there is a tight control of DNA integrity. At the end of filling and the onset of seed dehydration, the upregulation of genes implicated in sensing of DNA double-strand breaks suggests that genome integrity is challenged. The expression of chromatin remodelers seems to imply a concomitant action of chromatin remodelling with DNA repair machinery, maintaining genome stability. The expression of genes related to nucleotide excision repair and chromatin structure is evidenced during the desiccation stage. An overview of the genes involved in DNA damage response and chromatin remodelling during *P. vulgaris* seed development is presented, providing insights into the mechanisms used by developing seeds to cope with DNA damage.

**Keywords:** genome integrity; DNA damage response; chromatin remodeling; Phaseolus vulgaris; seed development

# Maintaining genome integrity during seed development in *Phaseolus vulgaris* L.: evidence from a transcriptomic profiling study



## 1. Introduction

Maintenance of genome integrity is particularly important to the seed phase of the plant lifecycle [1]. Chromatin integrity is constantly being challenged by environmental factors like drought and ionizing radiation or free radicals and alkylating agents generated by endogenous processes [2,3]. These agents cause a variety of DNA damage, including DNA base oxidation and alkylation, the formation of pyrimidine dimers and abasic sites, single- and double-strand breaks (SSBs and DSBs), and DNA inter-strand crosslinks, and therefore seriously threaten the integrity of the plant genome [2]. To maintain genome stability, all organisms have evolved DNA Damage Response (DDR) mechanisms that activate cell cycle checkpoints and DNA repair pathways and programmed cell death [4].

Two signal transducers of DNA breakage are the ATAXIA TELANGIECTASIA MUTATED (ATM) and ATM- AND RAD3-RELATED (ATR) kinases [5]. While ATR seems to be critical in the response to disturbances during the progression of DNA replication, the ATM kinase is activated in response to double-strand breaks (DSBs) [6]. The MRE11-RAD50-NBS1 (MRN) complex and REPLICATION PROTEIN A (RPA) complexes sense the damage and trigger activation of the ATM and ATR kinases. This leads to the transcription of the WEE1 kinase, triggering the intra-S checkpoint, and to the phosphorylation of SUPPRESSOR OF GAMMA RESPONSE 1 (SOG1), which induces the transcription of DDR

genes [7–9]. Downstream of the sensing mechanism, several DNA repair mechanisms can act, including base excision repair (BER), the nucleotide excision repair (NER), the mismatch repair (MMR), and double-strand break (DSB) repair [4]. Nevertheless, to detect and repair damaged DNA, DDR proteins need to overcome the barrier of condensed chromatin to gain access to the lesion in the DNA [10]. Multiple studies in eukaryotes, including plants, have evidenced a strong link between DDR and chromatin structure stability [10–13].



DNA lesions must be repaired to avoid cytotoxic and genotoxic consequences that adversely affect plant growth and development [14]. Seed germination has been widely exploited as a biological system to study DDR in plants. Effective DNA repair during early germination steps, such as imbibition, has been linked to enhanced seed vigour [14], while other works highlight the role of DDR in maintaining seed longevity and viability [1,15]. While genome integrity mechanisms have been extensively characterized during seed germination, the current knowledge concerning the link between DDR and chromatin remodelling during seed development (SD) is still scarce. Embryogenesis, with a high rate of cell division, as well as desiccation, in which orthodox seeds lose water while maintaining viability, are developmental stages prone to DNA damage. Indeed, a significant proportion of the gene expression and metabolic signatures of seed desiccation resemble those associated with seed germination, implying that seeds already start to be equipped for germination during desiccation [16]. Consequently, disturbances during seed development may impact subsequent seed physiology, as desiccation, dormancy, longevity, and germination [17]. As an example, an essential role for the apurinic/apyrimidinic endonucleases APE1L and APE2, involved in DNA repair during embryo development has been described in *Arabidopsis* [18]. Also, a differential accumulation of DNA repair-related proteins, such as RAD23-like protein, has been described during seed development of

*Brassica campestris* L. [19]. Nevertheless, more studies targeting different phenological stages are needed to provide a comprehensive picture of the different DDR mechanisms activated to ensure the maintenance of genome integrity in seeds.

The common bean (*Phaseolus vulgaris* L.) is among the most important grain legumes for human consumption worldwide [20]. Recently, our team unveiled new clues about the maintenance of genome integrity/DDR during *P. vulgaris* seed development, such as the differential accumulation of the PROLIFERATING CELL NUCLEAR ANTIGEN (PCNA) and DNA-DAMAGE INDUCIBLE PROTEIN 1, using a proteomic approach [21]. Genome integrity is, however, maintained as a result of intricate actions of multiple players and an extended characterization of DDR/chromatin remodelling mechanisms occurring during seed development in this species needs to be accomplished.



To extend our knowledge on this topic, we conducted a transcriptome study to explore the transcriptomic landscape of the DDR/chromatin remodelling mechanisms activated during SD in *P. vulgaris*. Our research hypothesis assumes that the expression of genes involved in DDR is regulated during SD to maintain genome integrity and seed viability upon desiccation. Our study, conducted at four distinct SD time points, presents an overview of the genes involved in the timeframe of these processes while providing new insights into the mechanisms that seeds likely use to cope with DNA damage during development.

## 2. Materials and Methods

### 2.1. Plant Material

Plants from the *Phaseolus vulgaris* genotype SER 16 (SER), kindly supplied by Dr. Steve Beebe of the International Center for Tropical

Agriculture (CGIAR, Cali, Colombia), were used in this study. Seeds were germinated in the dark, on water-soaked paper in Petri dishes, at 27 °C for two days, followed by three days at 23 °C. Seedlings were transferred to watered vermiculite trays and kept in a growth chamber under the following conditions: 50% to 60% humidity, photoperiod of 12/12 h, day/night at 25/16 °C, respectively, and average light intensity of 400 µmol $m^{-2}$ $s^{-1}$. One week later, seedlings were transferred to 2.5 L pots with a (3:1) mixture standard "terra de Montemor" commercial soil (Horto do Campo Grande, Lisboa, Portugal) and vermiculite, respectively. Potted plants were maintained in growth chambers under the previously described conditions and watered three times per week during the whole assay. Flowers were tagged at anthesis, when the flower opens, and seeds were harvested at 10, 20, 30, and 40 days after anthesis (DAA) as described in [21]. Harvested seeds were immediately frozen in liquid nitrogen and stored at –80 °C until further use.



## 2.2. RNA Extraction, Quantification, and Quality Assessment

For total RNA isolation, frozen seeds were ground to a fine powder in liquid nitrogen using a mortar and pestle. An optimized version of the method of Chang et al. (1993) [22] was used for RNA isolation. Hexadecyltrimethylammonium bromide (CTAB), Ethylenediaminetetraacetic acid disodium salt dihydrate (EDTA), chloroform and sodium chloride (NaCl) were purchased from Merck (Darmstadt, Germany); Polyvinylpyrrolidone (PVP), β-mercaptoethanol and spermidine from Sigma-Aldrich (St. Louis, MO, USA), Tris-HCl from Carl Roth (Karlsruhe, Germany); isoamyl alcohol and lithium chloride from BDH Prolabo (Llinars del Vallès, Spain).

Briefly, 900 µL RNA Extraction Buffer [(2% (w/v) CTAB, 2% (w/v) PVP, 100 mM Tris-HCl pH 8, 30 mM EDTA, 2 M NaCl, 0.5g/L spermidine and 3% β-mercaptoethanol (v/v))] was added to about 0.1 g of ground sample and incubated for 30 min at 65 °C. Two consecutive extractions with chloroform:isoamyl alcohol (24:1) were performed. In order to precipitate

RNA, a LiCl solution was added to the supernatant to a final concentration of 3 M and incubated on ice for 30 min. RNA was recovered by centrifugation at 20,238 rcf for 20 min at room temperature. Two washes of the RNA pellets were performed with 70% and 100% absolute ethanol at –20 °C. Pellets were left to air-dry and diluted in 40 µL of cold Milli-Q RNAse-free water. Trace amounts of DNA contamination were removed with an Ambion® TURBO™ DNase (Life Technologies, Carlsbad, CA, USA) following the manufacturer´s instructions. RNA quantification was performed using a NanoDrop$^{tm}$ 2000c Spectrophotometer (Thermo Fisher Scientific Inc., Waltham, MA, USA). Extracted RNA purity was estimated based on the A260/280 and A260/230 ratios absorbance ratios and was approximately 2 before DNAse treatment. RNA integrity was assessed by electrophoresis in a 2.0% agarose gel, stained with SYBR® Safe (Life Technologies). The absence of DNA contamination was verified by a standard polymerase chain reaction (PCR) using primers for the *P. vulgaris DEHYDRIN* mRNA, complete cds (gi|1326160) Forward 5′- AAGGAGAGGCGAAGGAGAAG -3′; and Reverse 5′- ACCAAACCCCACAACACAAC -3′. Samples were stored at –80 °C until needed.

## 2.3. Massive Analysis of 3’-cDNA Ends and Data Analysis

Massive analysis of 3’-cDNA ends (MACE) libraries, each representing one of the four time points corresponding to 10, 20, 30, and 40 DAA, were prepared using an equimolar pool of RNA from four biological replicates per time point. Each biological replicate corresponds to seeds harvested from an individual plant. MACE libraries were prepared and sequenced by GenXPro GmbH (Frankfurt am Main, Germany) following in-house developed protocols [23,24]. Briefly, poly-adenylated mRNA was isolated from 1 µg of the large fraction of total RNA using Dynabeads® mRNA Purification Kit (Life Technologies GmbH Invitrogen, Life Technologies). First- and second-strand synthesis of cDNA was performed using SuperScript® III First-Strand Synthesis System (LifeTechnologies

GmbH), with modified bar-coded 5′-end biotinylated poly-T adapters suitable for the Illumina Hiseq2000 flow cell (Illumina, San Diego, CA, USA). Subsequently, the cDNA was fragmented to yield 250 base pair (bp) fragments. The 3′-ends of the fragmented cDNA were captured with streptavidin beads, while PCR bias-proof technology 'TrueQuant' was used by ligation of TrueQuant adapters (GenXPro GmbH) to distinguish PCR copies from original copies [25,26]. The barcoded samples were sequenced simultaneously in one lane of an Illumina Hiseq2000 with 1 × 100 bps.



Analysis of MACE libraries started by removing low-quality sequence-bases using "cutadapt" (https://github.com/marcelm/cutadapt/) [27]. Poly(A)-tails were clipped by an in-house Python-Script. The reads were aligned to 'Pvulgaris_442_v2.0.fa.gz' (from Phytozome v12.0 *P. vulgaris*) using Bowtie 2 [28]. This tool maps reads depending on certain parameters (i.e., quality) and calculates thresholds for each sequence. The annotation information was taken from the files "Pvulgaris_442_v2.1.gene.gff3" and "Pvulgaris_442_v2.1.annotation_info.txt" (*Phaseolus vulgaris* v2.1, U.S. Department of Energy Joint Genome Institute, Phytozome v12.0: http://phytozome.jgi.doe.gov/).

Normalization was performed using DEGseq R-package version 1.16.0 [29]. Genes were considered expressed (EGs) when they present a raw read value number >50 in at least one library. A semi-quantitative analysis of the changes in EGs was performed calculating the fold changes between two consecutive time points. (10 DAA vs. 20 DAA, 20 DAA vs. 30 DAA, 30 DAA vs. 40 DAA). All raw sequencing data have been deposited in NCBI Sequence Read Archive (SRA) database with the SRA accessions: SRR6466368, SRR6466367, SRR6466366 and SRR6466365.

## 2.4. Primers and Probe Design

Primers/probes for eight candidate genes selected from the DDR pathway were constructed: *SUPPRESSOR OF GAMMA RESPONSE 1*

(*SOG1*; Phvul.003G027100), *ATAXIA-TELANGIECTASIA MUTATED* (*ATM*; Phvul.006G003000), *GAMMA HISTONE VARIANT H2AX* (*G-H2AX*; Phvul.006G095800), *DNA REPAIR AND MEIOSIS PROTEIN* (*MRE11*) (*MRE11*; Phvul.005G085700), *DNA REPAIR-RECOMBINATION PROTEIN* (*RAD50*; Phvul.001G266800), *NIJMEGEN BREAKAGE SYNDROME 1* (*NBS1*; Phvul.008G242800), *WEE1 KINASE HOMOLOG* (*WEE1*; Phvul.001G204900), and *NUCLEOSOME ASSEMBLY PROTEIN 1;2* (*NAP1;2*; Phvul.003G135500) (**Table 1**). Primers and TaqMan® probes sequences were designed following the Primer Express ® Software Version 3.0 (Applied Biosystems, Foster City, CA, USA) guidelines and using Primer-BLAST (https://www.ncbi.nlm.nih.gov/tools/primer-blast). Primers and TaqMan® probes were synthesized by Life Technologies. Conserved domain sequences were avoided to primer design in order to increase the specificity. Whenever possible, primer selection parameters were: primer size range of 20–26 bp, amplification product size range of 50–150 bp; primer melting temperature of 58–60 °C; primer GC content of 30–60%; primer with no more than two G/C in the last five 3'-end nucleotides and no more than three G nucleotides runs within the sequences. TaqMan® probes design followed the same criteria, except size between 18 and 30 bp and melting temperature 68–70 °C. Probes were labeled with FAM™ or VIC® dye on the 5'-end and Non-Fluorescent Quencher (NFQ) on the 3'-end (**Table 1** and **Supplementary Table 1**).

**Table 1**. List of selected genes and description of the primers and probes constructed to be used for digital PCR.

| GENE ID | Gene Symbol | Description | Primer Forward Sequence (5' –3' ) | Primer Reverse Sequence (5' –3' ) | MGB Probe Sequence (5' –3' ) | Dye |
|---|---|---|---|---|---|---|
| Phvul.003G027100 | *SOG1* | *SUPPRESSOR OF GAMMA RESPONSE 1* | TGGGACAGTGAGTCACAGAA | GAGCATAAACAGAAAGACCAGGAT | CTGGGAGACTGGCTGTGGAGGAGAT | FAM™ |
| Phvul.006G003000 | *ATM* | *ATAXIA-TELANGIECTASIA MUTATED* | TGGACTCAGATCAGGCATTGA | CACCAAAATCAGTGTCACCTCTT | AGCAGGCGGCAATGGATGTGGTT | FAM™ |
| Phvul.006G095800 | *G-H2AX* | *GAMMA HISTONE VARIANT H2AX* | GGTGAGGAATGATGAGGAACTG | ACTCTTGTGAAGCAGATCCAA | CCGTTCGCAATGGTGACAGACCCC | FAM™ |
| Phvul.005G085700 | *MRE11* | *DNA REPAIR AND MEIOSIS PROTEIN (MRE11)* | CCACCTCGGGTATATGGAGAA | AAATCTCTTCAAAGGCGTGGAA | ATGAGGTGCGCCGCCACGACT | FAM™ |
| Phvul.001G266800 | *RAD50* | *DNA REPAIR-RECOMBINATION PROTEIN (RAD50)* | TGATGGTATGCGGCAAATGTTT | GCTGAACTAGTAGCCTTCACTCTT | TGCCCTTGCTGTGAACGCCCT | VIC™ |
| Phvul.008G242800 | *NBS1* | *NIJMEGEN BREAKAGE SYNDROME 1* | CAAGGTTGATGATAATGAAACTGGAA | GAGTGTGTGCCTTTCTGAAACAT | TGCTGTCTGGTGCAGTGCTTACGCT | VIC™ |
| Phvul.001G204900 | *WEE1* | *WEE1 KINASE HOMOLOG* | CTCATTCCTCTCAACCAACCA | GTGAGCACAACGCACGAT | CCTCCGTTTCCTGCTTCCAGAACCC | VIC™ |
| Phvul.003G135500 | *NAP1;2* | *NUCLEOSOME ASSEMBLY PROTEIN 1;2* | CTTTCACCTCTGCAATGAGTAAC | CCGCTCTATTTTCCTCGTTGA | AGGACACCTTCAACGTCGCCGATCT | VIC™ |

## 2.5. cDNA Synthesis and QuantStudio™ 3D Digital PCR

Four biological replicates per time point were used for the digital PCR (dPCR) assay. Reverse transcription was performed on the same RNA samples used for MACE sequencing. Briefly, 600–900 ng of RNA was reverse-transcribed using oligo dT18 primers and the Promega ImProm-II$^{tm}$ Reverse Transcription System (Promega, Madison, WI, USA) according to the manufacturer´s instructions. Resulting cDNAs were diluted to a final concentration of 10 ng/μL.



After assessing efficiency and specificity, dPCR was used to quantify the candidate gene expression. QuantStudio™ 3D Digital PCR 20K Chip Kit v2 (Thermo Fisher Scientific Inc.) was used to perform the dPCR according to the manufacturer's instructions. The standard cycling protocol was run on a GeneAmp PCR system 9700 thermal cycler (Thermo Fisher Scientific Inc.): 10 min at 96 °C, 40 cycles of 2 min at 60 °C and 30 s at 98 °C, final extension step of 2 min at 60 °C. QuantStudio™ 3D Digital PCR Instrument was used to process the chips, by capturing the image and performing a quantitative analysis of the target DNA concentration by interpolation of the fraction of positive reactions, labeled by FAM and VIC probes, with a Poisson distribution [30]. Data were then transferred to the Thermo Fisher Cloud where the QuantStudio™ 3D Analysis Suite™ Software further evaluated the chip quality. The confidence level was set at 95%, desired precision at 10%, in the Poisson Plus algorithm version 4.4.10. The theoretical confidence interval for each probe and time point was calculated by multiplying the obtained copies/μL ($n$ = 4) by the corresponding precision value.

## 2.6. Bioinformatic Analysis

Functional characterization was performed using the MapMan web tools (http://www.plabipd.de/portal/mercator-sequence-annotation) [31–33]. Unspliced gene sequences of all expressed genes were obtained using BioMart in Phytozome v.12 (https://phytozome.jgi.doe.gov/) to create a

mapping file for the Mercator pipeline. A BLAST analysis was conducted on histone superfamily protein gene sequences via BLASTN (https://blast.ncbi.nlm.nih.gov/) using megablast algorithm, against the Nucleotide collection (nr/nt) database, in Viridiplantae.

Cytoscape [34] software (Version 3.6.1) was used to visualize the molecular interaction networks. To ease the analysis, molecular interaction networks were established using the EGs from MapMan "DNA" category between two consecutive time points (10 vs. 20 DAA, 20 vs. 30 DAA and 30 vs. 40 DAA) with a two-fold change in expression levels. For the same EGs, complementary functional categories resulting from MapMan categorization were retrieved. The inputs for the network analyses were the MapMan BinName as the target node and the Gene ID as the source node.



A comparison between DNA-related genes and their respective proteins found in Parreira *et al*., 2016 [21] was established. Proteins differentially accumulated were identified from the seed samples collected in an independent experiment at the same time points as the ones studied herein. The expression profiles were established using the $Log_2$ of the average normalized intensity values for protein abundance, and the $Log_2$ of the normalized expression values for the genes.

## 3. Results

### 3.1. Global Overview of the Gene Expression during Seed Development

Four MACE libraries (10, 20, 30, and 40 DAA) were prepared to capture the time frame of transcriptome changes underlying major SD stages. Seed samples harvested at 10 DAA represent the late embryogenic stage, in which evidence of a high rate of cell division associated with embryo

differentiation and morphogenesis has been described by us in a previous study [21]. At 20 DAA seeds are at the maturation/filling stage and an increased biomass accumulation is seen as a result of the synthesis of storage reserves. The 30 DAA represents the end of the filling stage marked by the end of biomass accumulation and the beginning of seed dehydration up until 40 DAA, when seeds are desiccated.



MACE sequencing of the four cDNA libraries resulted in 41.6 million 94-bp reads, with an average of 10.4 million reads/sample (**Table 2**).

**Table 2.** Characterization of the constructed massive analysis of 3'-cDNA ends (MACE) libraries. Total raw, cleaned, and mapped reads of each sequenced sample on the Illumina Hiseq2000 are described. Percentage of mapped reads was calculated using the number of mapped reads/cleaned reads. DAA: days after anthesis.

| Pooled sample ID | Raw reads | Cleaned reads | Mapped Reads | % Mapped Reads |
|---|---|---|---|---|
| **10 DAA** | **10880000** | **8000000** | **7818197** | **97.73%** |
| **20 DAA** | **10767653** | **7876850** | **7633075** | **96.91%** |
| **30 DAA** | **11648490** | **8477795** | **8296129** | **97.86%** |
| **40 DAA** | **8255161** | **5772840** | **5518643** | **95.60%** |

A total of 14,001 genes were expressed and identified among all the samples analysed in this study (**Supplementary Table 3**), representing 51.04% of the total number of loci described in the *Phaseolus vulgaris* v2.1 genome available at Phytozome v12.0 [35]. Sequences identified as *LOW-MOLECULAR-WEIGHT CYSTEINE-RICH 68* (*LCR68*; Phvul.005G071300), *LECTIN RECEPTOR KINASE A4.3* (*LECRKA4.3*; Phvul.004G158100) and *CUPIN FAMILY PROTEIN* (*PAP85*; Phvul.007G059775) were amongst those with highest total raw read counts, suggesting no ribosomal RNA (rRNA) contamination during library preparation.

MapMan functional categorization identified the major biological and metabolic processes occurring during SD. “Protein” [BinCode (BC) 29] is the most representative category with 14.80 % of assigned EGs, followed by “RNA” (BC 27; 10.65%), “signaling” (BC 30; 5.58%), “miscellaneous” (BC 26; 4.61%), “transport” (BC 34; 3.64%), “cell” (BC 31; 3.53%) and “development” (BC 33; 2.70%) (**Supplementary Table 2**). The functional category “not assigned” accounts for 34.09% of the EGs.



## 3.2. Genes Involved in DNA Damage Response and Chromatin Remodeling during Seed Development

Genes were retrieved from MACE datasets associated with DDR (sensing, signal transduction, and repair) and chromatin remodeling, expressed during the time frame of SD. To accomplish this goal, we started to investigate EGs belonging to the MapMan functional category of “DNA” (BC 28) (**Supplementary Table 4**). Among the 301 EGs identified in the “DNA” category, 67.11% (202 EGs) were assigned to the DNA.synthesis/chromatin structure (BC 28.1) sub-category, 15.61% (47 EGs) were assigned to DNA repair (BC 28.2), and 20.60% (62 EGs) were assigned to DNA.unspecified (BC 28.99). A semi-quantitative analysis of the changes in the expression of these genes was performed, calculating the fold changes between two consecutive time points. Among the 301 EGs identified, 194 showed fold changes higher than 2 between 10 DAA and 20 DAA, 63 between 20 DAA and 30 DAA, and 67 between 30 DAA and 40 DAA (**Figure 1**). In the transition between 10 DAA to 20 DAA, 96.9% of the “DNA” category genes were found to be downregulated. On the other hand, on the transition between 30 DAA and 40 DAA the expression of 73.13% of EGs was upregulated.

For each sub-category, gene expression profiles for EGs were established (**Figure 2**). The EGs belonging to the “DNA.repair” subcategory

show relatively lower expression when compared to the normalized read values observed in “DNA.synthesis/chromatin structure”.



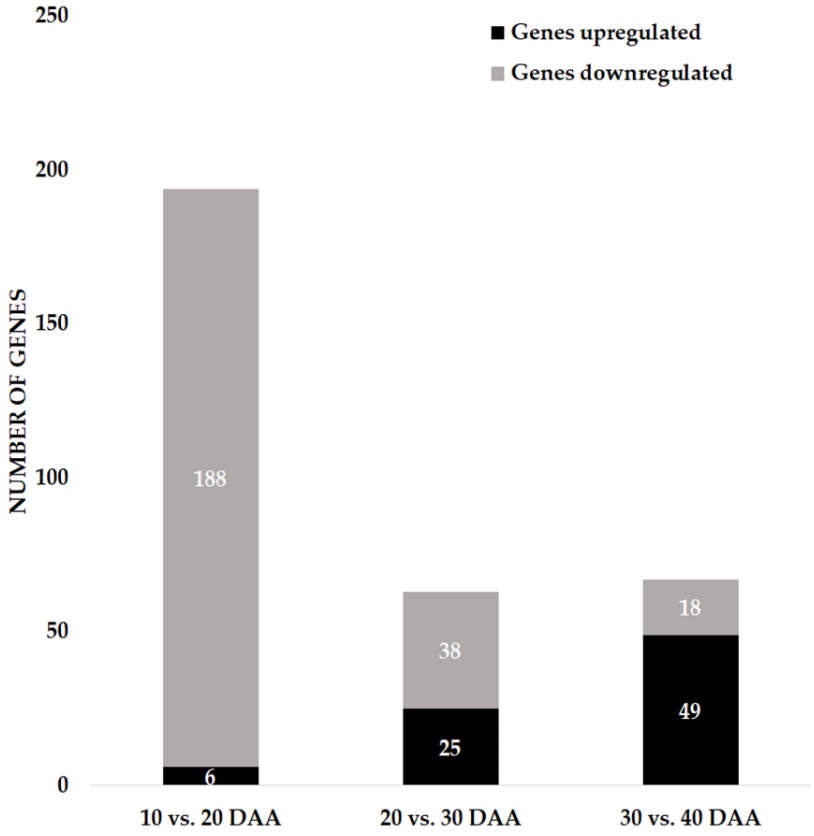



**Figure 1.** The number of up- and downregulated expressed genes belonging to DNA functional category in developing seeds of *P. vulgaris* with a minimum of two-fold change in the expression. DAA: days after anthesis.

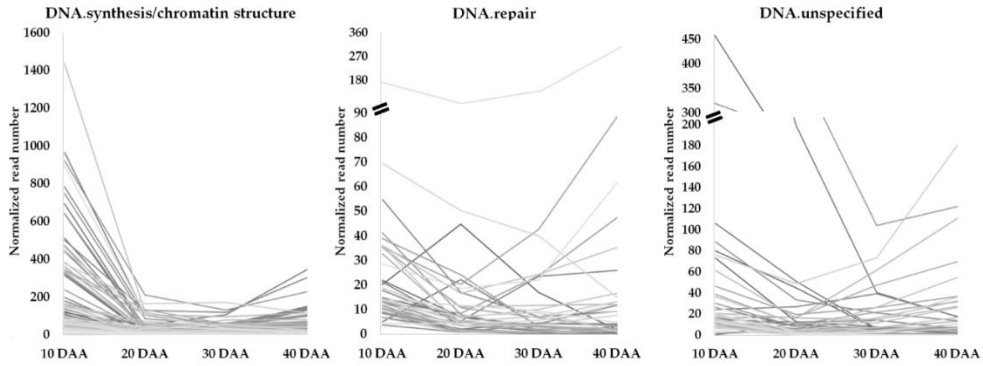


**Figure 2.** Profiles of expressed genes classified in the MapMan functional "DNA" subcategories during *P. vulgaris* seed development. Grey lines depict expression profiles for each individual gene in the subcategory with at least two-fold change in the expression.

Forty-four EGs encoding histones were identified, representing 21.8% of EGs included in the subcategory "DNA.synthesis/chromatin structure". The *GAMMA VARIANT OF HISTONE H2AX* (*G-H2AX*; Phvul.006G095800) is highly expressed at 10 DAA, decreasing strongly its expression levels in the next stage studied (**Supplementary Table 4**). Increase in the expression of some members of the *HISTONE SUPERFAMILY PROTEIN* were noticed on the transition from 30 to 40 DAA. The BLAST analysis conducted on histone superfamily protein gene sequences revealed that the majority has strong homology with predicted members of H3 (H3.2) and H4 family (see **Supplementary Table 5** for more information).

A decrease in expression from 10 to 20 DAA was observed for the majority of genes categorized in this "DNA" subcategory (**Figure 2**, such as the *NUCLEOSOME ASSEMBLY PROTEIN 1;2* (*NAP1;2*; Phvul.009G231400), the *SPO11/DNA TOPOISOMERASE VI, SUBUNIT A PROTEIN* (*BIN5*; Phvul.007G162700) the *RPA70-KDA SUBUNIT B* (*RPA70B*; Phvul.003G184400) and the *TOPOISOMERASE II* (*TOPII*; Phvul.005G024200). The expression of *MCM2*, *MCM3*, *MCM4*, *MCM5*, *MCM6*, *MCM7* and *MCM9 MINICHROMOSOME MAINTENANCE (MCM2/3/5) FAMILY PROTEIN* (Phvul.003G16100, Phvul.002G194200, Phvul.011G041300, Phvul.003G175800, Phvul.009G245600, Phvul.001G212600 and Phvul.006G204900, respectively), related with DNA replication [36], show also high expression at 10 DAA, decreasing afterwards. The *CHROMATIN REMODELING FACTOR17* (*CHR17*; Phvul.003G101700) and the *ORIGIN RECOGNITION COMPLEX 1* (*ORC1A*; Phvul.L009243) are also highly expressed at 10 DAA but the expression also decreases strongly to 20 DAA. Nevertheless, some

exceptions to this trend were noticed. That is the case of the *RESTRICTION ENDONUCLEASE TYPE II-LIKE SUPERFAMILY PROTEIN (RAD1*; Phvul.002G001900), which increases strongly between 20 and 30 DAA and, the *SMR (SMALL MUTS RELATED) DOMAIN-CONTAINING PROTEIN* (Phvul.006G068100), which increases from 30 to 40 DAA (**Supplementary Table 4**).



Within the "DNA.repair" subcategory, the *MUTL PROTEIN HOMOLOG 3* (*MLH3*; Phvul.002G116900) expression decreases from 10 to 20 DAA and increases afterward between 20 and 30 DAA (**Supplementary Table 4**). On the contrary, the *MUTS HOMOLOG 2* (*MSH2*; Phvul.003G177100), *MUTS HOMOLOG 6* (*MSH6*; Phvul.001G212500) and *MUTS HOMOLOG 7* (*MSH7*; Phvul.004G162000) decrease in abundance from 10 to 20 DAA (**Supplementary Table 4**). The *LEUCINE-RICH REPEAT (LRR) FAMILY PROTEIN* (*DNA-DAMAGE-REPAIR/TOLERATION PROTEIN 100 - DRT100*; Phvul.011G212100) has the highest expression observed at 20 DAA, decreasing afterward. The *DNA-DAMAGE-REPAIR/TOLERATION PROTEIN* (*DRT102*; Phvul.006G099800) also shows a peak in expression at 20 DAA, decreasing afterwards, namely on the 30 to 40 DAA. Still, in the "DNA.repair" subcategory, the *REPLICON PROTEIN A2* (*RPA2*; Phvul.003G145200) and *RAS ASSOCIATED WITH DIABETES PROTEIN 51* (*RAD51*; Phvul.003G126800) showed decreased expression from 10 to 20 DAA. Three *DNA GLYCOSYLASE SUPERFAMILY PROTEINS* (Phvul.003G156200, Phvul.005G045900 and Phvul.003G197200) have a high expression at 10 DAA, decreasing to 20 DAA. Interestingly, part of the MRN complex, the *DNA REPAIR AND MEIOSIS PROTEIN* (*MRE11*; Phvul.005G085700) and the *DNA REPAIR-RECOMBINATION PROTEIN* (*RAD50*; Phvul.001G266800), show high expression at 10 DAA, decreasing to 20 DAA (**Supplementary Table 4**).

Under the "DNA.unspecified" subcategory, the *CHROMATIN REMODELING FACTOR CHD3 (PICKLE)* (*PKL*; Phvul.001G046100) shows

a strong decrease in its expression from 10 to 20 DAA (**Supplementary Table 4**). The WHIRLY 2 (WHY2; Phvul.006G106800) was also highly expressed at 10 DAA (**Supplementary Table 4**).

Other genes not categorized as "DNA" and implicated in DNA damage response mechanisms were also retrieved from our MACE datasets (**Supplementary Table 3**). As an example, we found that the expression of the *HOMOLOG OF DNA MISMATCH REPAIR PROTEIN MSH3* (Phvul.007G069100), categorized as "Signalling.G-proteins," increases from 30 to 40 DAA. Expression of *RELATED TO UBIQUITIN 1* (*RUB1*; Phvul.005G060100), categorized as "Protein.degradation," decreases from 10 to 20 DAA. The expression of the *SUPPRESSOR OF GAMMA RESPONSE 1* (*SOG1*, Phvul.003G027100), categorized as "Development.unspecified," shows a strong decrease from 10 to 20 DAA, increasing afterwards. The expression of *ATM* (Phvul.006G003000) and *WEE1* (Phvul.001G204900), two kinases implicated in transduction of DNA damage that belong to the MapMan "Protein.postranslational modification" category, showed strong decrease in their expression in the transition from 10 to 20 DAA.

We also retrieved a collection of EGs related to base excision repair (BER), nucleotide excision repair (NER), homologous recombination (HR), mismatch repair (MMR), and non-homologous end joining (NHEJ) based on the categorization done by [4] (**Supplementary Table 6**). The analysis of this data shows a higher number of NER-related genes was found when compared with the number of genes from other repair mechanisms throughout SD (**Supplementary Table 6**). The expression profiles of these genes were established (**Figure 3**) and, importantly, a trend showing an increase of gene expression on the transition from 30 to 40 DAA was observed on genes related to NER.

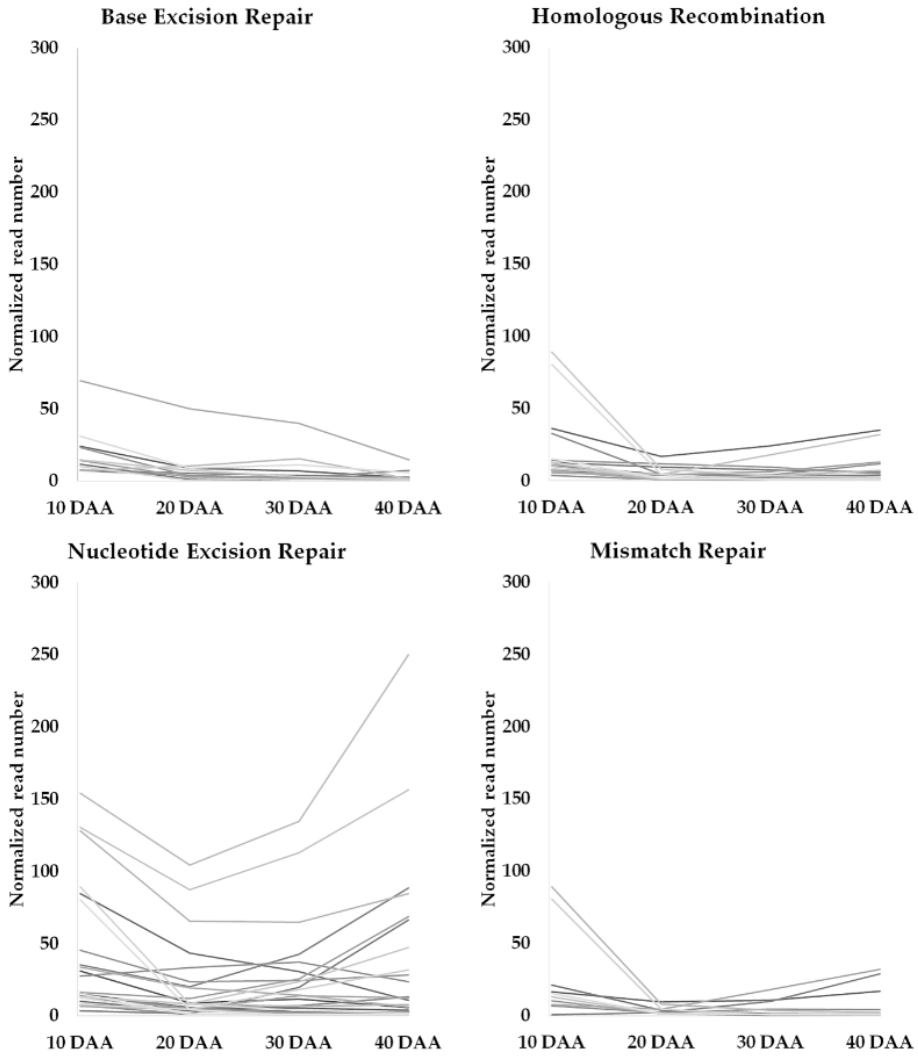



**Figure 3.** Profiles of expressed genes related with different DNA repair mechanisms (classification based on description made by Sampinato [4]). DAA: days after anthesis. Grey lines depict expression profiles for each individual gene in the DNA repair mechanisms.

## 3.3. Quantification by QuantStudio™ 3D Digital PCR

The MACE study provided evidence that the expression of genes implicated in DDR and chromatin remodelling occurs at relatively low levels and changes during seed development. To corroborate these assumptions, the expression levels of eight genes involved in DNA damage sensing were quantified by dPCR in four biological replicates. The transcript copy number

per microliter (Cn/µL) variation of the eight genes studied during the SD is shown in **Figure 4**.

For all tested genes, the highest Cn/µL, was observed at 10 DAA. From 10 to 20 DAA, all genes show a strong decrease in expression, being *WEE1* the most downregulated gene with a $Log_2$ fold change (FC) –4.89. Interestingly, an upregulation of *SOG1* and *MRE11* at 30 DAA was noticed, showing a $Log_2$FC 3.05 and 1.81, respectively. To a lesser extent, this was also observed for *NAP1;2*, *RAD50*, and *NBS1* (**Figure 4**).



A positive correlation ($R^2$ = 0.829) between the $Log_2$ of the expression values obtained by the two approaches (dPCR and MACE) was established for the eight genes analyzed, using the data from the four studied time points (**Supplementary Figure 1**). This high correlation seems to support the accuracy of the MACE analysis.

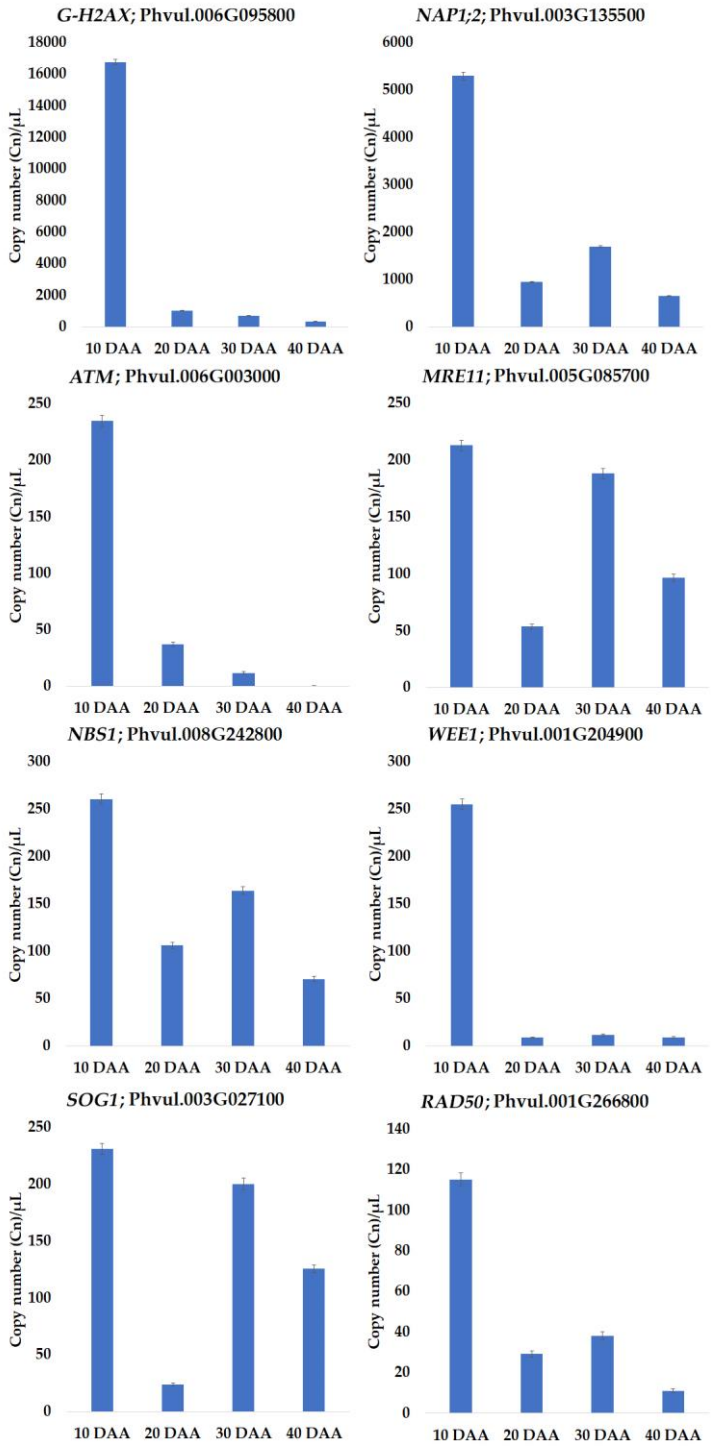



**Figure 4.** Bar plots represent mRNA copy number per microliter (Cn/µL) of biological quadruplicates for the eight genes selected for digital PCR. The Cn/µL was calculated using

QuantStudio™ 3D Analysis Suite™ assuming a Poisson distribution; error bars correspond to the theoretical confidence interval. The precision of quantification of ATM at 40 DAA and WEE1 at 20 and 40 DAA was higher than 10%. Y-axis: Cn/µL; X-axis: days after anthesis (DAA).

## 3.4. Changes in Transcriptomic Profiles Are in Accordance with Proteome Changes



To increase the robustness of our study, we investigated whether the changes detected in the transcriptome agreed with the changes observed in the proteome, using data from a previous study [21]. Only six proteins related with DNA metabolism were identified and their accumulation profiles were compared with the expression profiles of their protein-coding genes, *HISTONE H2A 2* (Phvul.005G090400; V7BUR1), *NUCLEOSOME ASSEMBLY PROTEIN 1;2* (*NAP1;2*) (Phvul.003G135500; V7CBA4), *MA3 DOMAIN-CONTAINING PROTEIN* (Phvul.003G207600; V7CDQ6), *DEAD/DEAH BOX RNA HELICASE FAMILY PROTEIN* (Phvul.009G093800; V7AUN1), *PROLIFERATING CELLULAR NUCLEAR ANTIGEN 1* (Phvul.006G137800; A1XCU7), and *UBIQUITIN FAMILY PROTEIN* (Phvul.006G060300; V7BL48). As observed in **Figure 5** there is a reasonable agreement between the transcriptomic and proteomic profiles obtained in the two independent experiments.

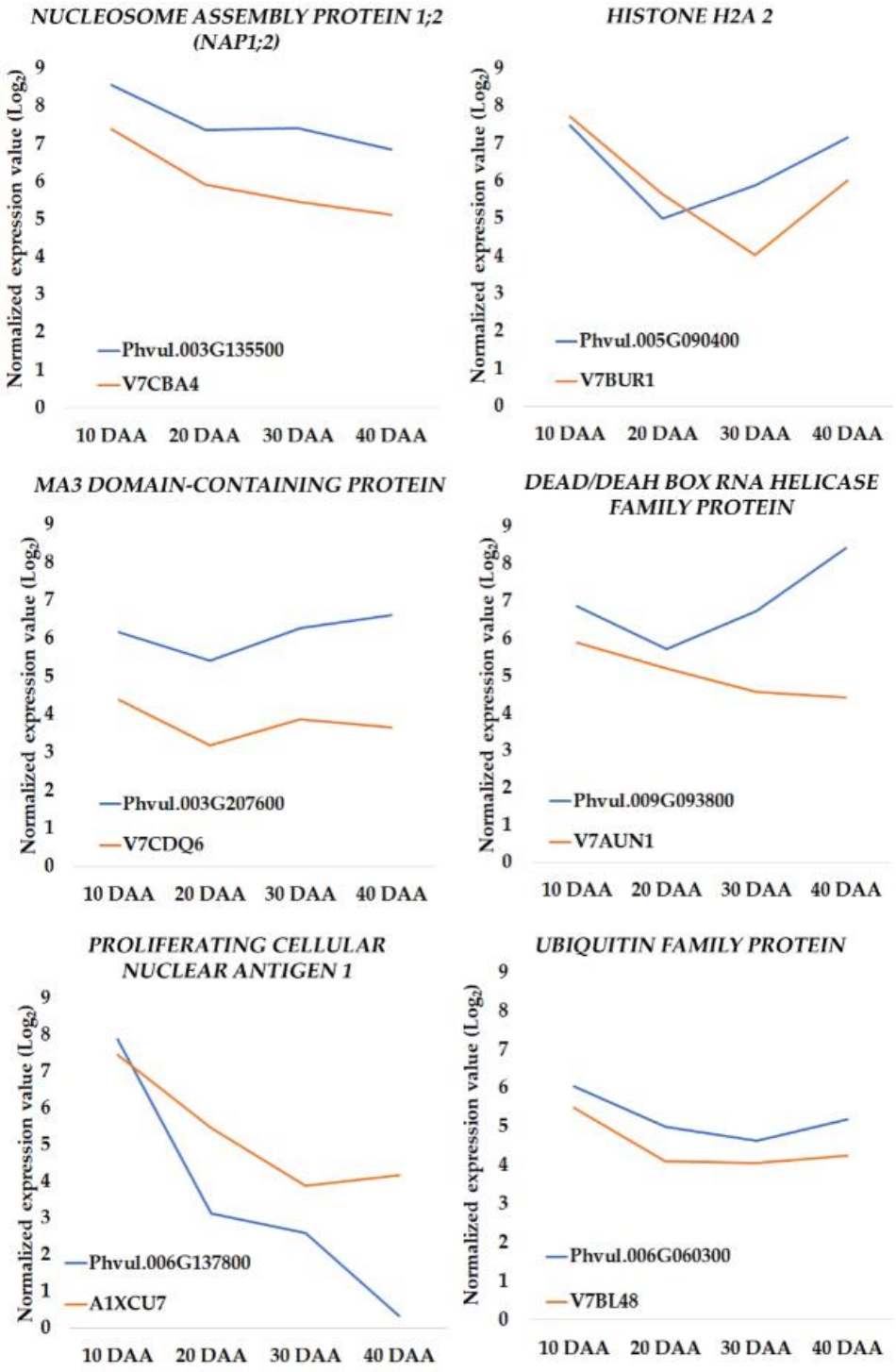



**Figure 5.** Abundance profiles of proteins [21], and expression of their correspondent genes, related to the DNA metabolism. Line in blue: $Log_2$ of the normalized expression values of the gene. Line in orange: $Log_2$ of the average normalized intensity values of the protein.

### 3.5. Network Analysis

Network analysis was conducted using EGs belonging to the “DNA” MapMan category that presented at least 2-fold changes between consecutive seed development stages. This analysis provides an integrative visualization of transcriptomic datasets allowing the identification of EGs that have more than one MapMan functional annotation besides “DNA”, thus evidencing additional functions.



#### 3.5.1. Network Analysis from 10 to 20 Days After Anthesis

Network analysis was conducted with the 194 EGs between 10 DAA and 20 DAA. Several connections between “DNA” functional category and other functional categories such as “RNA.regulation of transcription", "TCA/org transformation.TCA”, “protein.synthesis” were unveiled, reflecting possible interactions between different metabolic pathways (**Figure 6A**).

The previously described *ORC1A* (Phvul.L009243), involved in DNA replication [37], and the *HELICASE IN VASCULAR TISSUE AND TAPETUM* (*HVT1*; Phvul.008G196300 and Phvul.006G027200) bridges “DNA.unspecified” with “DNA.synthesis/chromatin structure” subcategory. Other DNA helicases as *RECQ4A* (Phvul.006G216200) and *RECQ HELICASE SIM* (*RECQSIM*; Phvul.006G082600), involved in maintenance of genome integrity [38,39], also connect these two above mentioned functional subcategories. The *DNA GYRASE B2* (*GYRB2*; Phvul.001G123100), which plays a role in the control of DNA topology [40], also bridges “DNA.synthesis/chromatin structure” to “DNA.repair”.

Other functional categories were found connected with the “DNA” categories and sub-categories. As an example, the *SEC14P-LIKE PHOSPHATIDYLINOSITOL TRANSFER FAMILY PROTEINS* (Phvul.011G042800 and Phvul.009G006000), related with membrane trafficking [41], is connecting “DNA.unspecified” to “protein.targeting” (BC

29.3) and to “transport.misc” (BC 34.99). The “redox.dismutases and catalases” (BC 21.6) is connected to “DNA.unspecified” via the *FE SUPEROXIDE DISMUTASE 2* (*FSD2*; Phvul.007G135400).

Among other different connections established, we noticed that the *FAR1-RELATED SEQUENCE 9* (*FRS9*; Phvul.008G293200), a putative negative regulator specific to phyB signaling [42], bridges “DNA.unspecified” to “signalling.light” (BC 30.11) and “RNA.regulation of transcription” (BC 27.3). On the other side, the *CHROMATIN REMODELING FACTOR17* (*CHR17*; Phvul.003G101700) also bridges “RNA.regulation of transcription” with “DNA.synthesis/chromatin structure”. *RPA70B*, previously described in section ‘3.2’ bridges “RNA.processing” (BC 27.1) with “DNA.synthesis/chromatin structure”. The *RAD21/REC8-LIKE FAMILY PROTEIN* (*SYN3*; Phvul.005G038800) is an essential gene for megagametogenesis [43] links “development.unspecified” (BC 33.99) to “DNA.synthesis/chromatin structure” in our networks. Interestingly, EGs categorized as “DNA.synthesis/chromatin structure” (e.g., *HISTIDYL-TRNA SYNTHETASE 1*, Phvul.001G179200; *KINASE INTERACTING (KIP1-LIKE) FAMILY PROTEIN*, Phvul.007G060600 and Phvul.002G043800) were also categorised as “protein” subcategories (BC 29.1; BC 29.4). “Signalling.receptor kinases” (BC 30.2) and “DNA.repair” is bridged by *DRT100*, already described in section 3.2.



### 3.5.2. Network Analysis from 20 to 30 Days after Anthesis

The network analysis of 20 to 30 DAA, constructed with 63 EGs, revealed that the subcategory of “DNA.repair” is no longer connected to “DNA.synthesis/chromatin structure” (**Figure 6B**). Nonetheless, the *HVT1* (Phvul.006G027200) bridges “DNA.unspecified” with “DNA.synthesis/chromatin structure” as observed for the 10 to 20 DAA comparison.

The *FE SUPEROXIDE DISMUTASE 2* (*FSD2*; Phvul.007G135400), involved in ROS scavenging [44] bridges the "redox.dismutases and catalases" (BC 21.6) to "DNA.unspecified". The *SEC14P-LIKE PHOSPHATIDYLINOSITOL TRANSFER FAMILY PROTEIN* (Phvul.009G061300) bridges "protein.targeting" (BC 29.3) and "transport.misc" (BC 34.99) with "DNA.unspecified". Also, the *LEUCINE-RICH REPEAT (LRR) FAMILY PROTEIN* (Phvul.005G036600) and the *DRT100* (Phvul.011G212100) are bridging "signalling.receptor kinases" (BC 30.2) with "DNA.repair".



### 3.5.3. Network Analysis from 30 to 40 Days after Anthesis

The last network established with the 67 EGs of the "DNA" functional category also revealed no connection between the "DNA.synthesis/chromatin structure" and the "DNA.repair" (**Figure 6C**). As seen in a previous network (**Figure 6A**), the *HVT1* (Phvul.008G196300) bridges "DNA.unspecified" with "DNA.synthesis/chromatin structure". The *FSD2* (Phvul.007G135400) bridges "redox.dismutases and catalases" (BC 21.6) with "DNA.unspecified", while the *SEC14P-LIKE PHOSPHATIDYLINOSITOL TRANSFER FAMILY PROTEIN* (Phvul.011G042800) bridges "protein.targeting" (BC 29.3) and "transport.misc" (BC 34.99) with "DNA.unspecified". Among other connections established, the *TRF-LIKE 9* (*TRFL9*; Phvul.001G232700) that encodes for a telomeric DNA-binding protein [45] bridges "DNA.unspecified" with "RNA.regulation of transcription" (BC 27.3). Interestingly, the *FRS9* (Phvul.008G293200) bridges "DNA.unspecified" with "RNA.regulation of transcription" (BC 27.3) and "signalling.light" (BC 30.11).

A

B
DNA synthesis/chromatin structure
DNA repair
DNA unspecified
signaling receptor kinases
redox dismutases and catalases
transport misc
protein targeting

C

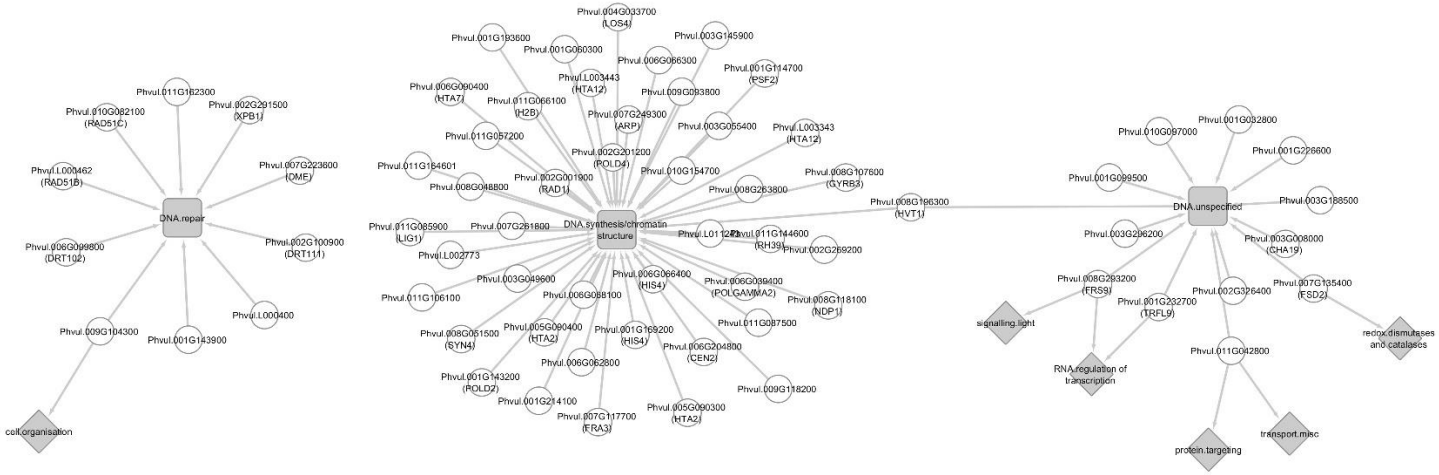



**Figure 6**. Network analysis of expressed genes between two consecutive seed development stages using the expressed genes (EGs) belonging to MapMan functional category " DNA" that changed expression at least 2-fold. For the same EGs, complementary functional categories resulting from MapMan categorization are also evidenced. Grey squares represent MapMan functional subcategories of " DNA", gray diamonds represent other MapMan functional categories beside " DNA" and white circles represent EGs with changed expression in a comparison. Comparisons between time points are shown: (A) 10 DAA vs. 20 DAA; (B) 20 DAA vs. 30 DAA; (C) 30 DAA vs. 40 DAA. DAA: days after anthesis.

## 4. Discussion

We characterized the transcriptomic landscape of *P. vulgaris* seeds at the stages of late embryogenesis (10 DAA), early (20 DAA) and late filling (30 DAA), and seed desiccation (40 DAA). While the maintenance of genome integrity has been thoroughly investigated in seed germination [12,14,15] or in seed response to priming agents [46], information is still scarce in relation to seed development.



Despite the sequencing of a single pooled MACE library for each time point analyzed, the identification of expressed genes implicated in DNA damage response (DDR) and chromatin remodeling in the time frame of seed development was possible. Our MACE approach did not allow the assessment of biological variance contribution [47], but still provided new insights into the mechanisms that seeds likely use to cope with DNA damage to maintain genome integrity and seed viability upon desiccation when grown under optimal growth conditions. Still, we cannot disregard that, when the plant matures under field conditions, with a certain level of environmental disturbance (e.g., abiotic stresses), different molecular responses may occur at the seed level. More studies would be needed to elucidate these aspects.

Among the different DDR components found expressed during seed development, we focused our attention on those acting upstream DNA repair, like the DSB sensing and signal transduction components (**Supplementary Figure 2**). Digital PCR (dPCR) allows absolute transcript quantification, even at low expression levels, due to its improved sensitivity [48,49] and precision, especially in low-concentration samples [48,50–52]. In this work we have used a chip-based platform, the QuantStudio 3D Digital PCR to quantify expression levels of eight genes involved in DSB sensing and signal transduction. This methodology was previously successfully applied to plants [53]. Although different technologies can report different expression levels for the same gene [47], our dPCR results showed a high positive correlation with

the MACE ones (**Supplementary Figure 1**). Moreover, a similarity between the transcriptomic and proteomics profiles was obtained (**Figure 5**), when gene expression data are compared with proteomic data from an independent experiment carried with four biological replicates for the same *P. vulgaris* seed development time points [21]. This supports the accuracy of the information provided by the transcriptomic study hereby presented.



With the inherent limitations described previously, our MACE study highlights a qualitative timeframe of molecular events associated with the maintenance of genome integrity during seed development. DNA damage sensing and the different repair mechanisms seem to be activated during early SD stages, when cell division and differentiation occurs. Chromatin structure and nucleotide excision repair seem to be relevant during seed dehydration, evidencing the activation of seed protection mechanisms that could play a major role on seed viability. Additionally, the molecular interaction networks established evidence other functional categories putatively related to DNA metabolism that may act in a concerted way during SD. These results are discussed in the following sections.

## 4.1. Tight Control of DNA Damage Seems to Occur during Seed Development

A generic downregulation of genes belonging to the MapMan “DNA” category was observed in the transition between 10 to 20 DAA. The 10 DAA time point reflects a late embryogenesis stage, in which a high rate of cell division has been reported [21,54–56] and consequently an increased expression of DNA replication and repair factors is expected. Our results are in agreement with this. As stated previously, the detection of DNA double-strand breaks (DSB) is necessary to initiate DSB repair. Our MACE study provided us indications that the expression of DSB sensing and signal transduction components may change during SD and we corroborated this information by digital PCR. As one example, the protein kinase *ATAXIA-*

*TELANGIECTASIA MUTATED* (*ATM*, Phvul.006G003000; **Supplementary Table 3**), a DNA damage-inducible protein kinase that phosphorylates a plethora of substrates participating in DNA damage response [7], was found to be highly expressed at 10 DAA, decreasing afterward, as seen in our dPCR results (**Figure 4**). Waterworth *et al*. (2016) reported the crucial role played by ATM in genome safeguarding during seed germination [1], our findings suggest a similar role for ATM during the first stages of seed development.



Besides ATM, the DSB sensing mechanism requires the complex interaction of several other DDR components. In *Arabidopsis*, the activation of ATM is dependent on a functional MRN (MRE11-RAD50-NBS1) complex [7,57], whereas ATM activates the phosphorylation of the DDR master regulator SUPPRESSOR OF GAMMA RESPONSE 1 (SOG1). The latter modulates the transcriptional response during DDR, triggering activation of cell cycle checkpoints and eventually programmed cell death [8,9,58]. Also, the *WEE1* kinase expression is regulated in an ATM-dependent manner, inhibiting the cell cycle upon activation of the DNA integrity checkpoint [59]. On the other side, the histone H2AX is also phosphorylated by ATM, in response to DNA damage, resulting in γH2AX foci, which are the sites for recruitment of DDR factors [3,57,60] (**Supplementary Figure 2**). Chaperones, as the NAP1-RELATED PROTEIN 2, a member of NUCLEOSOME ASSEMBLY PROTEIN-1 (NAP1) family of histone H2A/H2B have been described as having a role in nucleosome assembly and histone trafficking during DNA repair [61]. As observed for *ATM*, relevant components of the DSB sensing mechanism, the MRN complex genes, *SOG1*, and *WEE1* have a high expression at 10 DAA in our dPCR results (**Figure 4**). Similar results were obtained for the *GAMMA HISTONE VARIANT H2AX* and *NAP1;2*. It is not surprising that a tight control of DNA integrity is needed, due to the high rate of cellular division and metabolic activity characteristic of this stage [21,55]. Indeed, even in the absence

external stresses, DNA damage can arise from endogenous ROS, metabolic by-products, and breaks induced during DNA replication [14,15]. Another interesting aspect is the major upregulation of *SOG1*, *NAP1*, *MRE11*, *NBS1* and *RAD50* observed at 30 DAA in the dPCR profiling. From our previous study [21], we found that seed fresh weight decreased from 30 DAA, an aspect associated with the end of the seed growth and the onset of dehydration. Numerous reports highlights that the high metabolic activity in developing seed and drying of mature seed results in the production of ROS, such as superoxide radical ($O^{2-}$) and hydrogen peroxide ($H_2O_2$) [62,63]. In *Helianthus annuus*, the $H_2O_2$ content is quite high at the beginning of seed development, probably because the moisture content is high enough to allow metabolic activities [62]. In this context, we could suggest that upregulation in DSB sensing components at 30 DAA maybe be a consequence of the production and accumulation of ROS capable of genotoxic damage. In *Medicago truncatula* seeds, we have seen that upregulation of DNA response components occurs, being these implicated in the maintenance of genome stability in response to the genotoxic stresses applied [12,64]. Although, the expression levels of the histone *H2AX* and *ATM* seems to decrease slightly from 20 to 40 DAA, we cannot discard their role in DDR, since these components act at the post-translational level [7].

Other components of the response to DNA damage seem to be expressed during *P. vulgaris* seed development. One of these components is the PROLIFERATING CELLULAR NUCLEAR ANTIGEN (PCNA), a scaffold protein required to recruit DNA repair components at the damaged sites [65]. We found evidence that *PCNA1* (Phvul.006G137800) is highly expressed at 10 DAA, decreasing afterward during seed development. Although we did not validate the expression of this gene by dPCR, we noticed the same trend of accumulation for this protein (A1XCU7) based on the results of previous proteomic study conducted on similar seed samples [21] (**Figure 5**; **Supplementary Table 3**). In humans, PCNA is also described as interacting with MUTS proteins [66], heterodimers composed of distinct

MUTS homologs (MSH) subunits involved in mismatch repair mechanism (MMR) [4]. MutSα (MSH2/MSH6) and MutSβ (MSH2/MSH3) are eukaryotic mismatch recognition proteins that preferentially process base–base and small insertion/deletion (ID) mispairs, respectively [67]. We have detected considerable expression of the *MSH2* and *MSH6* subunits at 10 DAA. The expression of the *MSH3* subunit increases remarkably in the transition from 30 DAA to 40 DAA, when the seed starts to desiccate and cellular components are prone to oxidative damage. The presence of MUTS and PCNA homologs in plant tissues suggest that an interaction between these components may occur, as described with humans, but, to the best of our knowledge, this has not yet been described in plants and still deserves further validation.



The REPLICATION PROTEIN A (RPA), an essential regulator of eukaryotic DNA metabolism, has a key role in DDR, binding to single-stranded DNA (ssDNA) and coordinating the recruitment and exchange of genome maintenance factors [68]. RPA is involved in the sensing of lesions that cause stalled replication forks, as well as in NER, MMR and HR pathways [4,68,69]. Similar to what was observed for other DDR genes, and based on limited evidence from the MACE study, the expression of several RPA subunits, such as *RPA70B*, is high at 10 DAA and decreases to 20 DAA.

One question that remains to be elucidated is whether the profile of expression of DDR genes found at late embryogenesis is similar to the one found in other non-embryogenic proliferating tissues such as the shoot apical meristem (SAM) and root apical meristem (RAM). In rice, Kimura *et al*. [70] compared the expression patterns of DDR genes in proliferative tissues (SAM and RAM) and non-proliferative tissues as mature leaves. Among others, *OsPCNA*, *OsRPA32*, and *OsORC1* were found to be expressed in both proliferating tissues but not mature leaves. Interestingly, our results show a high expression of *PCNA1*, *ORC1A* and *RPA* subunits at 10 DAA (**Supplementary Table 3**), which suggests a similarity in DNA damage responses between

tissues where cell division and differentiation are actively occurring. Nevertheless, more comprehensive studies are required to further validate this evidence.

Another aspect revealed by this study is the peak expression at 20 DAA for the *DNA-DAMAGE REPAIR/TOLERATION 100 PROTEIN* (*DRT100*), which matches the early filling and change in seed coat coloration. Our network analysis in the transition from 10 to 20 DAA revealed that *DRT100* bridges “DNA.repair” and “signaling.receptor kinases” Mapman functional categories. DRT100 is a putative homolog of RecA chloroplastidial protein, a protein that plays important roles in DNA damage repair [71]. The overexpression of grape *DRT100* (Vv*DRT100-L*), in *Arabidopsis*, enhanced DNA repair under UV-B irradiation [72], suggesting for a specific role of VvDRT100-L in the repair/reduction of abasic sites and SSBs possibly arising from oxidative stress. The role of DRT100 in the context of seed development as well as the occurrence of DNA damage during this developmental transition is still unclear. However, we have evidence that oxidative damage may occur during early filling and seed dehydration in *P. vulgaris*. A strong accumulation ($Log_2$ FC 4.85) of the 1-CYS PEROXIREDOXIN enzyme (EC 1.11.1.15/EC 1.11.1.7; A0A0B2RB28) from 10 to 20 DAA and, at less extent from 30 to 40 DAA, was previously observed [21]. It is worth noting that 1-CYS PEROXIREDOXIN acts not only to relieve mild oxidative stresses but also as a molecular chaperone under severe stress conditions during seed germination and plant development [73].



Nucleotide excision repair has been described as the most versatile system for dealing with the DNA damage accumulated on desiccation and seed aging [74]. Our data show that the expression of different NER genes increases from 30 to 40 DAA (**Figure 3, Supplementary Table 6**), as a molecular signature of the seed desiccation process. It remains to be answered if this increase is triggered by the accumulation of DNA damage during this phase or

if this is a protective mechanism by which the seed is being prepared to deal with this type of damages during early germination as suggested by [14].

## 4.2. Role of Chromatin and Chromatin Remodelling in DNA Damage Repair during Seed Development



Decompaction and subsequent restoration of the starting chromatin structure in conjunction with DDR create another level of complexity in genome maintenance regulation [75]. Chromatin consists mainly of nucleosomes, spherical octamers formed by two molecules of each of the four histone types H2A, H2B, H3, and H4 with 1.7 turns of DNA wrapped around their surface and sealed by H1 linker histones [13]. Several reports highlight that, besides their role in chromatin structure, histones play a major role in the regulation of gene transcription and DNA repair [13].

Numerous genes encoding for different HISTONE H2A proteins, GAMMA HISTONE VARIANT H2AX (G-H2A), H2B proteins, H4 and H3 were detected in our data. Generally, a strong decrease in their expression was detected in the transition from late embryogenesis to early filling (10 to 20 DAA) suggesting that the enrolment of these proteins during embryogenesis was relevant. Considering the role of histone proteins in packaging DNA into chromatin, the synthesis of histones obviously needs to be tightly linked to DNA synthesis [76]. In the previous section, we discussed the role played by H2A.X, a variant of the H2A subunit conserved between eukaryotes, in the context of DNA damage signalling during early seed development. Increases in the expression of some members of the *HISTONE SUPERFAMILY PROTEIN* with strong homology with predicted members of *H3* (*H3.2*) and *H4* family were noticed on the transition from seed maturation to dehydration (30 to 40 DAA, **Supplementary Table 4**), which also marks the embryo entrance in the quiescent state characteristic of orthodox seeds, such as those of *P. vulgaris*. It is difficult to suggest a DDR role associated with the accumulation of those transcripts during seed dehydration but we cannot rule

out that the accumulation of *H3/H4* transcripts, and therefore proteins, may influence the condensation state of the chromatin. In yeast, elevated histone levels are correlated with tighter chromatin structure, which leads to a transcriptional silencing of genome regions that might be involved in lifespan extension [77]. Based on these previous statements, we could speculate that high histone accumulation at 10 and at 40 DAA could affect chromatin compactness during embryogenesis and dehydration, regulating gene transcription during seed development. In agreement with this hypothesis, the expression profiles of genes encoding for storage proteins such as the PAP85 CUPINS and RMLC-LIKE CUPINS PROTEIN families (**Supplementary Table 3**) and the profiles of some proteins implicated in carbohydrate metabolism, as the STARCH BRANCHING ENZYME (EC 2.4.1.18, Q9XIS5), GLUCOSE-1-PHOSPHATE ADENYLYLTRANSFERASE (EC 2.7.7.27; V7C329) [21] are opposite to the expression profiles of histone superfamily proteins.

Chromatin remodelers change the contacts between histones and DNA in the nucleosome, allowing them to arrange the distribution pattern of nucleosome spacing along the chromatin, thus playing a role in activation or repression of gene expression [78]. Previous works focused on seed germination have reported that chromatin remodelers act in concert with DNA repair machinery to maintain seed genome stability, essential for a successful germination [79]. Beside the changes on the expression of genes encoding for histones, we have also noticed interesting changes in the expression of genes encoding for chromatin remodelers as the *CHROMATIN REMODELING FACTOR17* (*CHR17*) and the *CHROMATIN REMODELING FACTOR CHD3* (*PKL*) during *P. vulgaris* seed development. The expression of *CHR17* was high at late embryogenesis (10 DAA) and subsequently decreases on early filling (20 DAA). Interestingly, the network analysis conducted for the same time points unveiled that *CHR17* bridges “DNA.synthesis/chromatin structure” and “RNA.regulation of transcription” MapMan functional categories (**Figure 6A**). CHR17 belongs to the imitation

switch (ISWI)-type chromatin remodelers, which, in *Arabidopsis*, seem to be involved in sliding nucleosomes in gene bodies, with impact in gene transcription [78]. Based on these assumptions, it is not surprising to observe higher expression levels of *CHR17* during early seed development, when embryo and remaining seed compartment cells are actively dividing and differentiating. The expression of *PKL* also decreases in the transition from late embryogenesis to the early filling stage. This gene, a member of the CHD subfamily II with ATP-dependent chromatin remodelling activity [80], is involved in several key physiological processes such as floral transition mediated by *LEAFY* (*LFY*) expression in *Arabidopsis* [81]. PKL was also described as a repressor of the transcription factor (TF) LEAFY COTYLEDON1 (LEC1) upon embryogenesis [82,83]. LEC1 is part of the LEC1/ ABSCISIC ACID INSENSITIVE3 (ABI3), FUSCA3 (FUS3), and LEAFY COTYLEDON2 (LEC2) (AFL) regulatory network that regulates seed filling, maturation, and storage compound synthesis [82,83]. The *PKL* expression profile is opposite to the observed for storage proteins during seed development [21], such as LEGUMIN (F8QXP7) and PHASEOLIN (Q43632), an aspect that is in agreement with the regulatory function proposed for PKL on the LEC1/AFL regulatory network [82–84].



Reports highlight that histone post-translational modifications (PTMs), e.g., acetylation and methylation of histone H3 and H4, are crucial in the regulation of DNA replication and transcription, as well as DNA repair [85]. Other post-translational modifications, as neddylation (Ubiquitin-NEDD8), appear to play a major role in the regulation of developmental processes during embryogenesis in plants [86,87]. We detected a high expression level of *RUB1*, a gene encoding a ubiquitin-NEDD8-like protein transition at late embryogenesis (10 DAA) decreasing afterward (**Supplementary Table 3**). In our previous proteomic analysis [21], we speculated that several metabolic pathways acting during early SD could be regulated via neddylation, since the accumulation of ubiquitin-NEDD8-like

protein RUB2 was high at 10 DAA. Evidence from a study conducted in human cells found that histone H4 was polyneddylated in response to DNA damage and NEDD8 was conjugated to the N-terminal lysine residues of H4 [88]. Nevertheless, more studies are needed to understand the role of neddylation in modulating DDR responses during seed development.

## 5. Conclusions



Mechanisms of genome integrity maintenance are active during seed development in *Phaseolus vulgaris*. The results show that most genes related to the DNA damage response and chromatin remodelling are more expressed during late embryogenesis when cell division and differentiation occurs. Among them, the expression of the *PCNA1*, the *RPA* and the *RUB1* were found highly expressed at 10 DAA. Genes involved in chromatin structure and remodelling, such as *H2AX*, the *CHR17* and the *CHD3* were also found highly expressed at the same time point. This suggests relevant mechanisms acting, possibly involved in the modulation of DNA damage response arising from rapid cell division and high metabolic activity characteristic of embryo morphogenesis.

Evidence of the activation of seed protection mechanisms during seed desiccation was also found. An upregulation of *SOG1*, *NAP1*, *MRE11*, *NBS1* and *RAD50*, implicated in the sensing of DNA double-strand breaks and signal transduction, was observed at 30 DAA, suggesting that genome integrity is also challenged at the onset of seed dehydration. An increase of the expression of genes associated with nucleotide excision repair was also noticed in the transition from 30 DAA to 40 DAA. An increase in the expression of some *HISTONE SUPERFAMILY PROTEIN* members was also observed during seed desiccation, possibly affecting chromatin compactness while regulating gene transcription.

Despite the evidence provided by this study, there are still open questions that need to be addressed. It remains to be understood the

molecular mechanisms that are underlying the temporal regulation of gene expression seen during seed development. Although it could be interesting to investigate DDR and chromatin remodelling mechanisms very early in seed development, more studies need to be focused on seed protection mechanisms activated during desiccation, since they could play a major role in seed viability. In due time, the molecular resources generated in this study would provide new targets that could be used to breed seeds with improved seed viability.



## 6. Acknowledgments

The authors acknowledge Dr. Steve Beebe (CIAT-CGIAR, Cali, Colombia) for kindly supplying seeds of the SER16 genotype of *P.vulgaris* used in this study. The authors thank the DOE Joint Genome Institute for pre-publication access to the annotated *Phaseolus vulgaris* genome v2.1. The authors J.R.P., S.S.A., and A.B. are members of COST action "FA1306—The quest for tolerant varieties—Phenotyping at the plant and cellular level," whose networking support is acknowledged. The authors acknowledge the financial support from Fundação para a Ciência e a Tecnologia (Lisbon, Portugal) through research projects PTDC/AGR-GPL/110224/2009, research unit "GREEN-it: Bioresources for Sustainability" (UID/Multi/04551/2013), Plants for Life PhD grant of author JRP (PD/BD/113474/2015) within the scope of the PhD program Plants for Life (PD/00035/2013) and a postdoctoral grant to SSA (SFRH/BPD/108032/2015). The financial support from the Italian Ministry of Education, University and Research (MIUR) through Dipartimenti di Eccellenza Program (2018–2022)—Dept. of Biology and Biotechnology "L. Spallanzani", University of Pavia to author AB is also acknowledged

## 7. Supplementary Materials

All supplementary materials are available at:
https://www.mdpi.com/2073-4425/9/10/463

**Supplementary Figure 1. Scatter plot with the $Log_2$ expression values for all candidate genes at all time points.** The $Log_2$ expression value for each technique was calculated from the normalized read count (in the transcriptomic dataset) and the transcripts copy number per microliter of the biological quadruplicates (in the dPCR dataset).



**Supplementary Figure 2. Schematic representation of the double strand break (DSBs) sensing mechanism in plants.** DSBs are sensed by the MRN[ DNA REPAIR AND MEIOSIS PROTEIN (MRE11) – DNA REPAIR-RECOMBINATION PROTEIN (RAD50) – NIJMEGEN BREAKAGE SYNDROME 1 (NBS1)] complex. Activation of ATAXIA-TELANGIECTASIA MUTATED (ATM) depends on MRN complex. The active ATM phosphorylates SUPPRESSOR OF GAMMA RESPONSE1 (SOG1), which in turn induces a transcriptional response that includes DNA damage response genes. The kinase ATM is also able to induce the transcription of *WEE1 KINASE HOMOLOG* (*WEE1*), triggering a DNA integrity checkpoint. The GAMMA HISTONE VARIANT H2AX is also phosphorylated by ATM, in response to DNA damage, resulting in γH2AX foci, which are the sites for recruitment of DDR factors. The NAP1-RELATED PROTEIN 2, a member of NUCLEOSOME ASSEMBLY PROTEIN-1 (NAP1) family of histone H2A/H2B have been described as having a role in histone trafficking during DNA repair. Black arrows represent induction/activation. Black T represents repression/inactivation. Grey arrow represents a putative involvement in the mechanism of histone trafficking during DNA repair plants (adapted from [3,7,61]).

**Supplementary Table 1. Primer/probe details used for dPCR.**

**Supplementary Table 2. Percentage and number of expressed genes included in MapMan functional categories.**

**Supplementary Table 3. List of expressed genes (raw value > 50 at least in one sample) with the respective gene ID, description, normalized and raw read count, Log2FC and MapMan functional category.**

**Supplementary Table 4. List of expressed genes (raw >50 in at least one sample) categorized as "DNA" MapMan functional category.** To ease visualization, the gene were grouped according "DNA" subcategories Duplicates inside each subcategory were removed.

**Supplementary Table 5. Blast results for histone superfamily genes identified within "DNA.synthesis/chromatin structure".** BLAST analysis was done using BLASTN (https://blast.ncbi.nlm.nih.gov/), megablast algorithm, against nucleotide collection (nr/nt) database, in Viridiplantae.

**Supplementary Table 6. List of expressed genes identified as associated with different DNA repair mechanisms (classification based on description made by Sampinato [4]).** BER: Base excision repair; NER: Nucleotide excision repair, HR: Homologous recombination, MMR: Mismatch repair, NHEJ: non-homologous end-joining, DAA: days after anthesis.



## 8. Bibliography


1. Waterworth, W.M.; Footitt, S.; Bray, C.M.; Finch-Savage, W.E.; West, C.E. DNA damage checkpoint kinase ATM regulates germination and maintains genome stability in seeds. *Proc. Natl. Acad. Sci.* **2016**, *113*, 9647–9652.

2. Roy, S. Maintenance of genome stability in plants: repairing DNA double strand breaks and chromatin structure stability. *Front. Plant Sci.* **2014**, *5*, 487.

3. Manova, V.; Gruszka, D. DNA damage and repair in plants – from models to crops. *Front. Plant Sci.* **2015**, *6*, 885.

4. Spampinato, C.P. Protecting DNA from errors and damage: an overview of DNA repair mechanisms in plants compared to mammals. *Cell. Mol. Life Sci.* **2017**, *74*, 1693–1709.

5. Ciccia, A.; Elledge, S.J. The DNA damage response: making it safe to play with knives. *Mol. Cell* **2010**, *40*, 179–204.

6. Abraham, R.T. Cell cycle checkpoint signaling through the ATM and ATR kinases. *Genes Dev.* **2001**, *15*, 2177–2196.

7. Hu, Z.; Cools, T.; De Veylder, L. Mechanisms used by plants to cope with DNA damage. *Annu. Rev. Plant Biol.* **2016**, *67*, 439–462.

8. Yoshiyama, K.O. SOG1: a master regulator of the DNA damage response in plants. *Genes Genet. Syst.* **2015**, *90*, 209–216.

9. Yoshiyama, K.O.; Kimura, S.; Maki, H.; Britt, A.B.; Umeda, M. The role of SOG1, a plant-specific transcriptional regulator, in the DNA damage response. *Plant Signal. Behav.* **2014**, *9*, e28889.

10. Tsukuda, T.; Fleming, A.B.; Nickoloff, J.A.; Osley, M.A. Chromatin remodelling at a DNA double-strand break site in *Saccharomyces cerevisiae*. *Nature* **2005**, *438*, 379–383.

11. Ramirez-Parra, E.; Gutierrez, C. E2F Regulates FASCIATA1, a chromatin assembly gene whose loss switches on the endocycle and activates gene expression by changing the epigenetic status. *Plant Physiol.* **2007**, *144*, 105–120.

12. Pagano, A.; Araújo, S. de S.; Macovei, A.; Leonetti, P.; Balestrazzi, A. The seed repair response during germination: disclosing correlations between DNA repair, antioxidant response, and chromatin remodeling in *Medicago truncatula*. *Front. Plant Sci.* **2017**, *8*, 1–14.

13. Donà, M.; Mittelsten Scheid, O. DNA damage repair in the context of plant chromatin. *Plant Physiol.* **2015**, *168*, 1206–1218.

14. Balestrazzi, A.; Confalonieri, M.; Macovei, A.; Donà, M.; Carbonera, D. Genotoxic stress and DNA repair in plants: Emerging functions and tools for improving crop productivity. *Plant Cell Rep.* **2011**, *30*, 287–295.

15. Waterworth, W.M.; Bray, C.M.; West, C.E. The importance of safeguarding genome integrity in germination and seed longevity. *J. Exp. Bot.* **2015**, *66*, 3549–3558.

16. Angelovici, R.; Galili, G.; Fernie, A.R.; Fait, A. Seed desiccation: a bridge between maturation and germination. *Trends Plant Sci.* **2010**, *15*, 211–218.

17. Leprince, O.; Pellizzaro, A.; Berriri, S.; Buitink, J. Late seed maturation: drying without dying. *J. Exp. Bot.* **2016**, *68*, erw363.

18. Murphy, T.M.; Belmonte, M.; Shu, S.; Britt, A.B.; Hatteroth, J. Requirement for abasic endonuclease gene homologues in arabidopsis seed development. *PLoS One* **2009**, *4*, e4297.

19. Li, W.; Gao, Y.; Xu, H.; Zhang, Y.; Wang, J. A Proteomic analysis of seed development in *Brassica campestri* L. *PLoS One* **2012**, *7*, e50290.

20. Leitão, S.T.; Dinis, M.; Veloso, M.M.; Šatović, Z.; Vaz Patto, M.C. Establishing the bases for introducing the unexplored portuguese common bean germplasm into the breeding world. *Front. Plant Sci.* **2017**, *8*, 1–18.

21. Parreira, J.R.; Bouraada, J.; Fitzpatrick, M.A.; Silvestre, S.; Bernardes da Silva, A.; Marques da Silva, J.; Almeida, A.M.; Fevereiro, P.; Altelaar, A.F.M.; Araújo, S.S. Differential proteomics reveals the

hallmarks of seed development in common bean (*Phaseolus vulgaris* L.). *J. Proteomics* **2016**, *143*, 188–198.

22. Chang, S.; Puryear, J.; Cairney, J. A simple and efficient method for isolating RNA from pine trees. *Plant Mol. Biol. Report.* **1993**, *11*, 113–116.

23. Zawada, A.M.; Rogacev, K.S.; Müller, S.; Rotter, B.; Winter, P.; Fliser, D.; Heine, G.H. Massive analysis of cDNA Ends (MACE) and miRNA expression profiling identifies proatherogenic pathways in chronic kidney disease. *Epigenetics* **2014**, *9*, 161–172.



24. Yakovlev, I.A.; Lee, Y.; Rotter, B.; Olsen, J.E.; Skrøppa, T.; Johnsen, Ø.; Fossdal, C.G. Temperature-dependent differential transcriptomes during formation of an epigenetic memory in Norway spruce embryogenesis. *Tree Genet. Genomes* **2014**, *10*, 355–366.

25. Eveland, A.L.; McCarty, D.R.; Koch, K.E. Transcript profiling by 3'-untranslated region sequencing resolves expression of gene families. *Plant Physiol.* **2007**, *146*, 32–44.

26. Torres, T.T.; Metta, M.; Ottenwalder, B.; Schlotterer, C. Gene expression profiling by massively parallel sequencing. *Genome Res.* **2007**, *18*, 172–177.

27. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **2011**, *17*, 10.

28. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–9.

29. Wang, L.; Feng, Z.; Wang, X.; Wang, X.; Zhang, X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* **2010**, *26*, 136–138.

30. De St. Groth, S.F.F. The evaluation of limiting dilution assays. *J. Immunol. Methods* **1982**, *49*, R11–R23.

31. Usadel, B.; Poree, F.; Nagel, A.; Lohse, M.; Czedik-Eysenberg, A.; Stitt, M. A guide to using MapMan to visualize and compare Omics data in plants: A case study in the crop species, Maize. *Plant, Cell Environ.* **2009**, *32*, 1211–1229.

32. Thimm, O.; Bläsing, O.; Gibon, Y.; Nagel, A.; Meyer, S.; Krüger, P.; Selbig, J.; Müller, L.A.; Rhee, S.Y.; Stitt, M. MAPMAN: A user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* **2004**, *37*, 914–939.

33. Lohse, M.; Nagel, A.; Harter, T.; May, P.; Schroda, M.; Zrenner, R.;

Tohge, T.; Fernie, A.R.; Stitt, M.; Usadel, B. Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. *Plant. Cell Environ.* **2014**, *37*, 1250–1258.

34. Shannon, P. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13*, 2498–2504.

35. Goodstein, D.M.; Shu, S.; Howson, R.; Neupane, R.; Hayes, R.D.; Fazo, J.; Mitros, T.; Dirks, W.; Hellsten, U.; Putnam, N.; *et al*. Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Res.* **2012**, *40*, D1178–D1186.

36. Rizvi, I.; Choudhury, N.R.; Tuteja, N. *Arabidopsis thaliana* MCM3 single subunit of MCM2–7 complex functions as 3′ to 5′ DNA helicase. *Protoplasma* **2016**, *253*, 467–475.

37. Masuda, H.P.; Ramos, G.B.A.; De Almeida-Engler, J.; Cabral, L.M.; Coqueiro, V.M.; Macrini, C.M.T.; Ferreira, P.C.G.; Hemerly, A.S. Genome based identification and analysis of the pre-replicative complex of *Arabidopsis thaliana*. *FEBS Lett.* **2004**, *574*, 192–202.

38. Wiedemann, G.; van Gessel, N.; Köchl, F.; Hunn, L.; Schulze, K.; Maloukh, L.; Nogué, F.; Decker, E.L.; Hartung, F.; Reski, R. RecQ helicases function in development, DNA repair, and gene targeting in *Physcomitrella patens*. *Plant Cell* **2018**, *30*, 717–736.

39. Bagherieh-Najjar, M.B.; de Vries, O.M.H.; Kroon, J.T.M.; Wright, E.L.; Elborough, K.M.; Hille, J.; Dijkwel, P.P. Arabidopsis RecQsim, a plant-specific member of the RecQ helicase family, can suppress the MMS hypersensitivity of the yeast sgs1 mutant. *Plant Mol. Biol.* **2003**, *52*, 273–84.

40. Evans-Roberts, K.M.; Breuer, C.; Wall, M.K.; Sugimoto-Shirasu, K.; Maxwell, A. *Arabidopsis thaliana* GYRB3 Does Not Encode a DNA Gyrase Subunit. *PLoS One* **2010**, *5*, e9899.

41. Bankaitis, V.A.; Mousley, C.J.; Schaaf, G. The Sec14 superfamily and mechanisms for crosstalk between lipid metabolism and lipid signaling. *Trends Biochem. Sci.* **2010**, *35*, 150–160.

42. Lin, R. Arabidopsis FHY3/FAR1 gene family and distinct roles of its members in light control of Arabidopsis development. *Plant Physiol.* **2004**, *136*, 4010–4022.

43. Yuan, L.; Yang, X.; Auman, D.; Makaroff, C.A. Expression of epitope-tagged SYN3 cohesin proteins can disrupt meiosis in Arabidopsis. *J. Genet. Genomics* **2014**, *41*, 153–164.

44. Myouga, F.; Hosoda, C.; Umezawa, T.; Iizumi, H.; Kuromori, T.; Motohashi, R.; Shono, Y.; Nagata, N.; Ikeuchi, M.; Shinozaki, K. A heterocomplex of iron superoxide dismutases defends chloroplast nucleoids against oxidative stress and is essential for chloroplast development in Arabidopsis. *Plant Cell Online* **2008**, *20*, 3148–3162.

45. Hwang, M.G.; Kim, K.; Lee, W.-K.K.; Cho, M.H. AtTBP2 and AtTRP2 in Arabidopsis encode proteins that bind plant telomeric DNA and induce DNA bending in vitro. *Mol. Genet. Genomics* **2005**, *273*, 66–75.



46. Ventura, L.; Donà, M.; Macovei, A.; Carbonera, D.; Buttafava, A.; Mondoni, A.; Rossi, G.; Balestrazzi, A. Understanding the molecular pathways associated with seed vigor. *Plant Physiol. Biochem.* **2012**, *60*, 196–206.

47. Hansen, K.D.; Wu, Z.; Irizarry, R.A.; Leek, J.T. Sequencing technology does not eliminate biological variability. *Nat. Biotechnol.* **2011**, *29*, 572–573.

48. Ma, J.; Li, N.; Guarnera, M.; Jiang, F. Quantification of plasma miRNAs by digital PCR for cancer diagnosis. *Biomark. Insights* **2013**, *8*, BMI.S13154.

49. Kim, T.G.; Jeong, S.-Y.; Cho, K.-S. Comparison of droplet digital PCR and quantitative real-time PCR in mcrA-based methanogen community analysis. *Biotechnol. Reports* **2014**, *4*, 1–4.

50. Hindson, C.M.; Chevillet, J.R.; Briggs, H.A.; Gallichotte, E.N.; Ruf, I.K.; Hindson, B.J.; Vessella, R.L.; Tewari, M. Absolute quantification by droplet digital PCR versus analog real-time PCR. *Nat. Methods* **2013**, *10*, 1003–1005.

51. Mukaide, M.; Sugiyama, M.; Korenaga, M.; Murata, K.; Kanto, T.; Masaki, N.; Mizokami, M. High-throughput and sensitive next-generation droplet digital PCR assay for the quantitation of the hepatitis C virus mutation at core amino acid 70. *J. Virol. Methods* **2014**, *207*, 169–177.

52. Bharuthram, A.; Paximadis, M.; Picton, A.C.P.P.; Tiemessen, C.T. Comparison of a quantitative Real-Time PCR assay and droplet digital PCR for copy number analysis of the CCL4L genes. *Infect. Genet. Evol.* **2014**, *25*, 28–35.

53. Santos, C.; Duarte, S.; Tedesco, S.; Fevereiro, P.; Costa, R.L. Expression profiling of *Castanea* genes during resistant and susceptible interactions with the oomycete pathogen *Phytophthora cinnamomi* Reveal possible mechanisms of immunity. *Front. Plant Sci.*

**2017**, *8*, 515.

54. Gallardo, K.; Le Signor, C.; Vandekerckhove, J.; Thompson, R.D.; Burstin, J. Proteomics of *Medicago truncatula* seed development establishes the time frame of diverse metabolic processes related to reserve accumulation. *Plant Physiol.* **2003**, *133*, 664–682.

55. Coelho, C.; Benedito, V.; Maria, C.; Coelho, M.; Augusto, V.; Coelho, C.; Benedito, V.; Maria, C.; Coelho, M.; Augusto, V.; et al. Seed development and reserve compound accumulation in common bean (*Phaseolus vulgaris* L.). *Seed Sci. Biotechnol* **2008**, *2*, 42–52.

56. Wang, W.-Q.; Liu, S.-J.; Song, S.-Q.; Møller, I.M. Proteomics of seed development, desiccation tolerance, germination and vigor. *Plant Physiol. Biochem.* **2015**, *86*, 1–15.

57. Amiard, S.; Charbonnel, C.; Allain, E.; Depeiges, A.; White, C.I.; Gallego, M.E. Distinct roles of the ATR kinase and the Mre11-Rad50-Nbs1 complex in the maintenance of chromosomal stability in Arabidopsis. *Plant Cell* **2010**, *22*, 3020–3033.

58. Yoshiyama, K.O.; Kobayashi, J.; Ogita, N.; Ueda, M.; Kimura, S.; Maki, H.; Umeda, M. ATM-mediated phosphorylation of SOG1 is essential for the DNA damage response in Arabidopsis. *EMBO Rep.* **2013**, *14*, 817–822.

59. De Schutter, K.; Joubes, J.; Cools, T.; Verkest, A.; Corellou, F.; Babiychuk, E.; Van Der Schueren, E.; Beeckman, T.; Kushnir, S.; Inze, D.; *et al*. Arabidopsis WEE1 kinase controls cell cycle arrest in response to activation of the DNA integrity checkpoint. *Plant Cell* **2007**, *19*, 211–225.

60. Amiard, S.; Depeiges, A.; Allain, E.; White, C.I.; Gallego, M.E. Arabidopsis ATM and ATR kinases prevent propagation of genome damage caused by telomere dysfunction. *Plant Cell* **2011**, *23*, 4254–4265.

61. Zhou, W.; Zhu, Y.; Dong, A.; Shen, W.-H. Histone H2A/H2B chaperones: from molecules to chromatin-based functions in plant growth and development. *Plant J.* **2015**, *83*, 78–95.

62. Bailly, C.; El-Maarouf-Bouteau, H.; Corbineau, F. From intracellular signaling networks to cell death: the dual role of reactive oxygen species in seed physiology. *C. R. Biol.* **2008**, *331*, 806–814.

63. Naithani, S.; Nonogaki, H.; Jaiswal, P. Exploring crossroads between seed development and stress response. In *Mechanism of Plant Hormone Signaling under Stress*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2017; Vol. 2, pp. 415–454.

64. Pagano, A.; de Sousa Araújo, S.; Macovei, A.; Dondi, D.; Lazzaroni, S.; Balestrazzi, A. Metabolic and gene expression hallmarks of seed germination uncovered by sodium butyrate in *Medicago truncatula*. *Plant. Cell Environ.* **2019**, *42*, 259–269.

65. Moldovan, G.-L.L.; Pfander, B.; Jentsch, S. PCNA, the maestro of the replication fork. *Cell* **2007**, *129*, 665–679.

66. Iyer, R.R.; Pluciennik, A.; Genschel, J.; Tsai, M.-S.S.; Beese, L.S.; Modrich, P. MutLα and proliferating cell nuclear antigen share binding sites on MutSβ. *J. Biol. Chem.* **2010**, *285*, 11730–11739.

67. Tian, L.; Gu, L.; Li, G.-M.M. Distinct nucleotide binding/hydrolysis properties and molar ratio of MutSα and MutSβ determine their differential mismatch binding activities. *J. Biol. Chem.* **2009**, *284*, 11557–11562.

68. Maréchal, A.; Zou, L. RPA-coated single-stranded DNA as a platform for post-translational modifications in the DNA damage response. *Cell Res.* **2015**, *25*, 9–23.

69. Aklilu, B.B.; Soderquist, R.S.; Culligan, K.M. Genetic analysis of the Replication Protein A large subunit family in Arabidopsis reveals unique and overlapping roles in DNA repair, meiosis and DNA replication. *Nucleic Acids Res.* **2014**, *42*, 3104–3118.

70. Kimura, S.; Tahira, Y.; Ishibashi, T.; Mori, Y.; Mori, T.; Hashimoto, J.; Sakaguchi, K. DNA repair in higher plants; photoreactivation is the major DNA repair pathway in non-proliferating cells while excision repair (nucleotide excision repair and base excision repair) is active in proliferating cells. *Nucleic Acids Res.* **2004**, *32*, 2760–2767.

71. Rowan, B.A.; Oldenburg, D.J.; Bendich, A.J. RecA maintains the integrity of chloroplast DNA molecules in Arabidopsis. *J. Exp. Bot.* **2010**, *61*, 2575–2588.

72. Fujimori, N.; Suzuki, N.; Nakajima, Y.; Suzuki, S. Plant DNA-damage repair/toleration 100 protein repairs UV-B-induced DNA damage. *DNA Repair (Amst).* **2014**, *21*, 171–176.

73. Kim, S.Y.; Paeng, S.K.; Nawkar, G.M.; Maibam, P.; Lee, E.S.; Kim, K.-S.S.; Lee, D.H.; Park, D.-J.J.; Kang, S. Bin; Kim, M.R.; et al. The 1-Cys peroxiredoxin, a regulator of seed dormancy, functions as a molecular chaperone under oxidative stress conditions. *Plant Sci.* **2011**, *181*, 119–124.

74. Boubriak, I.; McCready, S.; Osborne, D.J. DNA structure and seed desiccation tolerance. In *Plant Desiccation Tolerance*; Blackwell

Publishing Ltd: Oxford, UK, 2000; pp. 215–249.

75. Giglia-Mari, G.; Zotter, A.; Vermeulen, W. DNA damage response. *Cold Spring Harb. Perspect. Biol.* **2011**, *3*, a000745–a000745.

76. Nelson, D.M.; Ye, X.; Hall, C.; Santos, H.; Ma, T.; Kao, G.D.; Yen, T.J.; Harper, J.W.; Adams, P.D. Coupling of DNA synthesis and histone synthesis in S phase independent of cyclin/cdk2 activity. *Mol. Cell. Biol.* **2002**, *22*, 7459–72.

77. Feser, J.; Truong, D.; Das, C.; Carson, J.J.; Kieft, J.; Harkness, T.; Tyler, J.K. Elevated histone expression promotes life span extension. *Mol. Cell* **2010**, *39*, 724–735.

78. Li, G.; Liu, S.; Wang, J.; He, J.; Huang, H.; Zhang, Y.; Xu, L. ISWI proteins participate in the genome-wide nucleosome distribution in Arabidopsis. *Plant J.* **2014**, *78*, 706–714.

79. Balestrazzi, A.; Donà, M.; Macovei, A.; Sabatini, M.E.; Pagano, A.; Carbonera, D. DNA repair and telomere maintenance during seed imbibition: correlation of transcriptional patterns. *Telomere and Telomerase* **2015**, 2–5.

80. Ho, K.K.; Zhang, H.; Golden, B.L.; Ogas, J. PICKLE is a CHD subfamily II ATP-dependent chromatin remodeling factor. *Biochim. Biophys. Acta - Gene Regul. Mech.* **2013**, *1829*, 199–210.

81. Fu, X.; Li, C.; Liang, Q.; Zhou, Y.; He, H.; Fan, L.-M. CHD3 chromatin-remodeling factor PICKLE regulates floral transition partially via modulating LEAFY expression at the chromatin level in Arabidopsis. *Sci. China Life Sci.* **2016**, *59*, 516–528.

82. Verdier, J.; Thompson, R.D. Transcriptional regulation of storage protein synthesis during dicotyledon seed filling. *Plant Cell Physiol.* **2008**, *49*, 1263–1271.

83. Zhai, L.; Xu, L.; Wang, Y.; Zhu, X.; Feng, H.; Li, C.; Luo, X.; Everlyne, M.M.; Liu, L. Transcriptional identification and characterization of differentially expressed genes associated with embryogenesis in radish (*Raphanus sativus* L.). *Sci. Rep.* **2016**, *6*, 21652.

84. Sreenivasulu, N.; Wobus, U. Seed-development programs: a systems biology–based comparison between dicots and monocots. *Annu. Rev. Plant Biol.* **2013**, *64*, 189–217.

85. Liu, X.; Yang, S.; Zhao, M.; Luo, M.; Yu, C.-W.; Chen, C.-Y.; Tai, R.; Wu, K. Transcriptional repression by histone deacetylases in plants. *Mol. Plant* **2014**, *7*, 764–772.

86. Mergner, J.; Schwechheimer, C. The NEDD8 modification pathway in plants. *Front. Plant Sci.* **2014**, *5*, 103.

87. Bostick, M. Related to Ubiquitin 1 and 2 Are redundant and essential and regulate vegetative growth, auxin signaling, and ethylene production in arabidopsis. *Plant Cell Online* **2004**, *16*, 2418–2432.

88. Ma, T.; Chen, Y.; Zhang, F.; Yang, C.-Y.Y.; Wang, S.; Yu, X. RNF111-dependent neddylation activates DNA damage-induced ubiquitination. *Mol. Cell* **2013**, *49*, 897–907.

# CHAPTER 4



# MicroRNAs expression dynamics reveal post-transcriptional mechanisms regulating seed development in *Phaseolus vulgaris* L.

**Adapted from:**

Parreira JR, Cappuccio M, Balestrazzi A, Fevereiro P, Araújo S de S. 2021. MicroRNAs expression dynamics reveal post-transcriptional mechanisms regulating seed development in *Phaseolus vulgaris* L. Horticulture Research 8, 18.

**Author contributions to the chapter:**

José Ricardo Torrão Dias Parreira Salvado participated in the experimental setup and design, performed the experiments except for sRNA-Seq and Degradome analysis and part of RT-qPCR experiments, analysed the data and wrote the chapter.

## Abstract

The knowledge on post-transcriptional regulation mechanisms implicated in seed development (SD) is still limited, particularly in one the most consumed grain legumes, *Phaseolus vulgaris* L.. We explore for the first time, the miRNA expression dynamics in *P. vulgaris* developing seeds. Seventy-two known and 39 new miRNAs were found expressed in *P. vulgaris* developing seeds. Most of miRNAs identified were more abundant at 10 and 40 days after anthesis, suggesting that late embryogenesis/early filling and desiccation were SD stages in which miRNA action is more pronounced. Degradome analysis and target prediction identified targets for 77 expressed miRNAs. While several known miRNAs were predicted to target *HD-ZIP*, *ARF*, *SPL* and *NF-Y* transcription factors families, most of new miRNAs targets were predicted to encode for functional proteins. MiRNAs-targets expression profiles evidenced that these miRNAs could timely tune distinct seed developmental stages. MiRNAs more accumulated at early SD stages were implicated on regulating the end of embryogenesis, postponing the seed maturation program, storage compound synthesis and allocation. MiRNAs more accumulated at late SD stages could be implicated on seed quiescence, desiccation tolerance and longevity with still uncovered roles in germination. The miRNAs herein described represent novel *P. vulgaris* resources with potential application in future biotechnological approaches to modulate the expression of genes implicated in legume seed traits with impact in horticultural production systems.

**Keywords:** Common bean, miRNAs, Seed development, Degradome

# MicroRNAs expression dynamics reveal post-transcriptional mechanisms regulating seed development in *Phaseolus vulgaris* L.



## 1. Introduction

The common bean, *Phaseolus vulgaris* L., is one of the most important grain legume worldwide due to its high protein content, dietary fibre, and essential vitamins and minerals [1]. Besides, common bean immature pods are also consumed as vegetables constituting a relevant fresh commodity with increasing world production [2].

Seed development (SD) is a complex process that starts with a double fertilization and usually comprises three stages in orthodox seeds: embryogenesis, filling and desiccation [3]. We described the main molecular and metabolic mechanisms underlying SD in *P. vulgaris* by analysing proteome and transcriptomic changes in seeds harvested at 10, 20, 30, and 40 days after anthesis (DAA) [4,5]. The 10 DAA matched the late embryogenesis, supported by evidence of high metabolic activity, including cell division and DNA synthesis. The 20 DAA matched the mid filling/maturation stage, in which the synthesis and accumulation of storage compounds was evidenced. The 30 DAA reflected the onset of seed desiccation and activation of protection mechanisms to ensure embryo quiescence and viability, connected with genome integrity maintenance [5]. At 40 DAA, the seed was found desiccated. Both studies highlighted a timeframe of molecular and metabolic events occurring, but the molecular mechanisms controlling the observed temporal changes in gene expression remain to be understood.

Post-transcriptional regulation mediated by miRNAs controls numerous developmental processes in plants, including SD in legumes [6,7]. MiRNAs are small non-coding RNAs (sRNAs, ~21-22 nt) that act as negative post-transcriptional regulators of gene expression, either by transcript cleavage or translation repression [8]. In *P. vulgaris*, numerous studies have characterized the changes in sRNAs and miRNAs populations in different organs and growth conditions [9–15]. Evidences on the role played by some of these sRNAs in stress adaptation or symbioses were provided. As examples, pvu-miR399a was found implicated in phosphorus (P)-deficiency signalling in common bean roots [16], while miR319d has been implicated in nodule development [11]. In this important pulse, the few studies that have addressed seeds were focused in one or few members of miRNAs families and on specific time-points [9,13]. Importantly, none of these studies provided comprehensive overview of miRNA abundances and repressed targets during SD. Upon availability, this knowledge can be used to support the development of approaches to improve *P. vulgaris* seed traits, including the accumulation of storage compounds or enhanced germination.



To address this gap of knowledge, a small RNA sequencing (sRNA-Seq) approach was used to identify miRNAs expressed during the main SD stages, spanning from late embryogenesis to seed desiccation. Degradome analysis and a target prediction algorithm were used to identify or predict, respectively, miRNA targets. RT-qPCR was used to validate the sequencing results of a selected group of miRNAs and expression profiles of their targets. Using molecular network analyses, integrated overviews of the main miRNAs acting and target functional categories under miRNA regulation during SD were also provided.

## 2. Materials and methods

### 2.1 Plant material

*P. vulgaris* genotype SER 16 seeds were germinated as described [4]. Seedlings were transferred to watered vermiculite trays and maintained in a growth chamber under the following conditions: 50% relative humidity, photoperiod of 16/8 hours, at 25/18°C day/night temperature, with an average light intensity of 300 µmol $m^{-2}$ $s^{-1}$. One week later, seedlings were transferred to 2.5L pots with a (2:1:1) mixture standard "terra de Montemor" commercial soil (Horto do Campo Grande, Lisboa, Portugal), peat and vermiculite, kept on above environmental conditions and watered 3 times/week. Flowers were tagged, ovaries were sampled before anthesis (0 days) and seeds were sampled at 5, 10, 15, 20, 25, 30, 35 and 40 DAA. Two groups of samples were obtained, one used to measure seed length, fresh and dry weight to characterize the SD process and another immediately frozen in liquid nitrogen and stored at -80ºC.



### 2.2 Total RNA isolation

Frozen ovaries and seeds were ground to a fine powder in liquid nitrogen using a mortar and pestle. RNA isolation was conducted as described [5] with few modifications as the inclusion of 2% β-mercaptoethanol (v/v) on RNA extraction buffer and purifications were performed first with phenol: chloroform:isoamyl alcohol (25:24:1), followed by another with chloroform: isoamyl alcohol (24:1). Total RNA was overnight precipitated (-20ºC) by adding 6.7 µL of 3M Sodium acetate (pH=5.2) per 100µL recovered and two volumes of -20º C absolute ethanol. The Ambion® TURBO™ DNase (Life Technologies, Carlsbad, CA, U.S.A) was used to remove DNA contamination. RNA quality was assessed using a NanoDrop$^{tm}$ 2000c Spectrophotometer (Thermo Fisher Scientific Inc., Waltham). RNA purity was estimated based on the A260/280 and A260/230 ratios absorbance ratios and was approximately 2 before DNAse treatment. Qubit®

2.0 Fluorometer (Thermo Fisher Scientific Inc.) with RNA BR Assay Kit was used to quantify the RNA. Extracted RNA integrity was assessed by electrophoresis in a 2.0% agarose gel. DNA contamination absence was verified by standard PCR and samples were stored at -80°C.

## 2.3 sRNA libraries construction and high-throughput sequencing



Twelve small RNA (sRNA) libraries were constructed from three biological replicates at 10, 20, 30 and 40 DAA. Each biological replicate represents seed samples collected from an independent potted plant. Libraries construction and sequencing was performed by LC Sciences (Houston, Texas, USA). SRNA libraries were generated using the Illumina Truseq™ Small RNA Preparation kit (Illumina, San Diego, USA) following the Illumina's TruSeq™ Small RNA Sample Preparation Guide (15004197 C, Illumina Inc., Part #1004239 Rev. A, 2008; Catalog # RS-930-1012, Part # 15004197 Rev. B, January 2011). Cluster generation was performed on Illumina's Cluster Station using the purified cDNA libraries. Sequencing was performed using an Illumina Genome Analyzer IIx. Illumina's Sequencing Control Studio (Version 2.8) was used to obtain the raw sequencing reads (40 nts), following real-time sequencing image analysis and base-calling by Illumina's Real-Time Analysis (Version 1.8.70).

## 2.4 sRNA data analysis and identification of known and novel miRNAs

The sequencing data was analysed using the ACGT101-miR v4.2 (LC Sciences) [17,18]. This software sorts raw sequencing reads into unique families, removing adapter dimers, low complexity sequences, junk and repeats. Additionally, it also removes sequences mapped against reference database files for other noncoding RNAs (RFam database; http:// rfam.janelia.org, repetitive sequences (Repbase; http://www.girinst.org/repbase) or mRNAs (*P. vulgaris* v2.1, Phytozome v12.0; DOE-JGI and USDA-NIFA, http://phytozome.jgi.doe.gov/). A detailed

description of the pipeline and criteria applied to identify miRNAs and predict their secondary structure can be found at **Supplementary Table 3**.

The unique sequences between 15 and 32 bases were mapped to plant pre-miRNAs available in miRBase v21.0 [19] to identify known and novel miRNAs. The sequences that mapped to *P. vulgaris* miRNAs/pre-miRNAs, which also mapped against the *P.vulgaris* genome, were identified as known miRNAs. Additionally, sequences that mapped to other plant pre-miRNAs and these pre-miRNAs further mapped to *P. vulgaris* genome were also considered known miRNAs. The remaining sequences that aligned against *P. vulgaris* genome and present hairpin structure that contains the sequences were considered predicted miRNAs. Known and predicted miRNAs were filtered by the mean normalised read number ≥ 100 in at least one time point were considered expressed and kept for further analysis. RNAfold software (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi) was additionally used on provided extended miRNA sequences to corroborate the potential of selected new miRNAs to form a stable hairpin [20].



With the release of miRBase v22.1 (http://www.mirbase.org), the sequences of all selected miRNAs were re-checked, filtering by a maximum of two mismatches between the obtained miRNAs and mature miRNAs in miRBase v22.1. This was made with the purpose to clarify if, by example, any of the previously predicted new miRNAs (miRBase v21) has been already described in the latest miRBase release. Mismatches were restricted to a maximum of one mismatch in the 5' region, or 2 in the 3' region, and no mismatches in position 10 and 11 nt. Also, restriction was assumed for overhangs with more than 2 nucleotides. Valid miRNAs sequences mapping to a known precursor hairpin opposite to the annotated mature miRNA were considered novel 5p- or 3p-derived miRNA candidates. To ease their description, miRNAs within the known group were denominated as 'pvu-' followed by the specific miRNA name described in miRBase. The remaining sequences are described as 'miR' for new miRNAs .

An ANOVA followed by a pairwise mean comparisons (T-test) were applied on normalized read counts to assess differential miRNA expression, using a similar approach as described in other studies [21,22]. A Benjamini-Hochberg correction was applied to the analysis made. A hierarchical clustering analysis was performed in Heatmapper (http://www.heatmapper.ca). A principal component analysis (PCA) was performed using standardized (Z-score) miRNAs normalized read values as variables. Additionally, a regression analysis was conducted on the same variables to extract significant correlations. Only correlations with a P-value ≤ 0.05 and r = Pearson correlation above 0.75 or below -0.75 were considered. PCA and product-moment correlation analyses were conducted using Statistica, version 6 (Statsoft).



## 2.5 Degradome sequencing

Four qualitative degradome libraries were constructed for 10, 20, 30 and 40 DAA samples by LC Sciences to identify miRNA target transcripts. Each library consists of an equimolar pool of total RNA from three biological replicates per timepoint. Sequencing was performed on the Illumina HiSeq 2500. The data was analysed using the ACGT101-DGD (LCSciences - https://www.lcsciences.com/documents/sample_data/degradome_sequencing/DGD_html_report_DEMO.html) pipeline. Raw sequencing reads were obtained using Cutadapt [23] and in-house perl scripts were used to remove adaptors and low quality reads. CleaveLand v4.3 [24] was used to identify potentially cleaved targets from degradome sequencing. The degradome reads were mapped to the *P. vulgaris* mRNA (Phytozome v.12). CleaveLand classifies the targets into five categories according to degradome sequencing and cleaved sequences abundance relative to the overall profile of degradome tags matching the target. Only targets within categories 0, 1 and 2 with T plot *P* value < 0.05 and whose miRNAs were found expressed in our datasets were kept for discussion. The target annotation information was taken from *Phaseolus vulgaris* v2.1.

## 2.6 Target prediction and bioinformatic analysis

Mature sequences for all expressed miRNAs were aligned against the target transcript library of *P. vulgaris* 442_v2.1 using the web-based Plant small RNA target server (psRNATarget; http://plantgrn.noble.org/psRNATarget). Default parameters for Schema V2 (2017 release) were used, except the maximum expectation value (E) set to 2, to lower false positive prediction. BLASTN (https://blast.ncbi.nlm.nih.gov/) was used on target coding sequences that presented no annotation in *P. vulgaris*. Megablast algorithm was used against the Nucleotide collection (nr/nt) database, in *Viridiplanta*e. Results with Query cover > 80% and E value < $2e^{-39}$ were selected.



For functional categorization, coding sequences from degradome and predicted target transcripts were obtained using BioMart (Phytozome v.12, *P. vulgaris* v2.1). These sequences were used to create a mapping file for the Mercator pipeline (http://mapman.gabipd.org/app/mercator) to perform a MapMan functional categorization analysis [25].

Cytoscape [26] software (Version 3.7.1) was used to visualize molecular interaction networks made between expressed miRNAs and their target functional categories. MiRNAs were used as source nodes, MapMan functional category as target nodes and, to ease visualization, genes that encode target transcripts are edges connecting the nodes. MiRNAs with no predicted target identified were not included.

## 2.7 Reverse transcription quantitative PCR

Two known and seven new miRNAs and their respective targets were selected for qRT-PCR validation in samples at 0, 5, 10, 15, 20, 25, 30, 35, 40 DAA. Three biological replicates, from independent plants, were used. The two-tailed RT-qPCR method was used to quantify the miRNA expression

[27]. Specific miRNAs forward, reverse and two-tailed RT primers were designed following published guidelines (**Supplementary Table 16**). The two tailed RT reaction was performed using qScript Flex cDNA Kit (QuantaBio, Beverly, USA) in multiplex, whenever possible. Briefly, 110 ng RNA was used per reaction with 0.1 μM of each primer, in an adjusted total volume of 10 μL. The RT-reaction was run in a T100™ Thermal Cycler (BioRad, Hercules,CA, USA) following the published protocol [27] conditions: 25ºC for 45 min, followed by 85ºC for 5 min.



The expression profile of ten targets was quantified by RT-qPCR (**Supplementary Table 17**). Each primer pair specificity was examined via Primer-BLAST (https://www.ncbi.nlm.nih.gov/tools/primer-blast/) and melting curves obtained during the assay. The RT reaction was performed with 400 ng RNA using ImProm-II™ Reverse Transcriptase (Promega, Madison, USA) in a T100™ Thermal Cycler, following the manufacturer's instructions.

Both miRNA and target qPCR reactions were run on PikoReal Real-Time PCR System (Thermo Fisher Scientific Inc., Waltham, MA, U.S.A.). The reactions were performed in a total of 10 μL using PerfeCTa SYBR® Green SuperMix (QuantaBio, Beverly, USA) containing 1 μM of each primer in the miRNA assays, and 0.2 μM in the target assays. Reactions were performed onto same samples for both assays. Efficiency of the reactions was calculated using LinRegPCR [28] and Pfaffl method was used to calculate relative expression [29]. Pvu-miR166n and pvu-miR1510b were used as normalizers for miRNAs qPCRs, while the eukaryotic Release Factor 1 (eRF1) family protein (PEL1) and RING/U-box superfamily protein (XERICO) for target qPCRs.

## 2.8 Target expression analysis

The expression profile of selected predicted or degradome validated targets was established by mining Massive Analysis of cDNA Ends (MACE)

datasets (NCBI Sequence Read Archive accessions: SRR6466368, SRR6466367, SRR6466366 and SRR6466365) previously described [5] and proteomic datasets (PRIDE: PXD002254) previously described [4]. The transcriptomic and proteomic data was produced from seed samples collected at the same time points herein studied and come from independent experiments.



## 3. Results

### 3.1 Small RNA profiles and miRNA identification in *Phaseolus vulgaris* developing seeds

To identify miRNAs expressed during *P. vulgaris* SD, twelve independent sRNAs libraries were generated using total RNA extracts from 3 biological replicates per time point (10, 20, 30 and 40 DAA). A total of 230,953,396 raw reads were obtained, representing in average 19,246,116 raw read per library (**Supplementary Table 1**). Of these, 175,032,524 clean reads were mappable and used for miRNA identification. The most abundant sRNA reads were 24 nt in length, representing 29.90% of the total mappable reads, followed by 21-nt RNAs with 13.5% (**Figure 1a, Supplementary Table 2**).

The applied bioinformatic pipeline identified 8,848 unique miRNAs, of these 141 were known miRNAs (total of 7,044,944 reads) and 8,707 were considered predicted miRNAs (total of 7,507,000 reads) based on mapping against miRBase v.21 (**Supplementary Table 3**). The 111 miRNAs that showed a mean normalised read number higher than 100 at one SD time point were kept for further analysis and considered expressed in *P. vulgaris* seeds, being 51 miRNAs described as known and 60 predicted.

With the release miRBase 22.1, these miRNAs were rechecked and a final list of 72 known and 39 new miRNAs were obtained (**Table 1 and 2**,

**Supplementary Table 4 and 5**). Among predicted miRNAs, only miRNAs that presented the following secondary structure prediction criteria as: (1) number of base pair (bp) in stem region (≥22), (2) free energy (dG in kCal/mol ≤-17), (3) length of hairpin (up and down stem + terminal loop ≥56), (4) length of terminal loop (≤82), (5) number of allowed biased bulges in mature region (≤2), (6) number of basepair (bp) in mature or mature* region (≥14), (7) percentage of small RNA in stem region (pm) (≥80) and (8) minimal folding energy index (MFEI ≥ 0.80) were considered as new miRNAs (**Supplementary Table 3**)**.** Additionally, MFE and secondary structure of the extended sequences (**Supplementary Figure 1**) were predicted using RNAFold. MFE for the hairpin structures of the new miRNA precursors was lower than -17 kcal/mol (**Supplementary Table 5**). The calculated MFEI ranged from 0.8 to 2.2, with an average of 0.91 (**Table 2**), matching values described for miRNAs [30]. Such combined analysis showed sequences are distinguished from other coding or non-coding RNAs and their secondary structures miRNAs were considered stable.

**Table 1.** Known miRNAs found expressed during seed development in *Phaseolus vulgaris*. The miRNA annotation in miRBase v21, v22.1 and the miRNA name in this study is provided. MiRNA family (Gene family) and stem-loop strand was obtained from miRbase 22.1 (for more details see Supplementary Table 4).

| miRNA (miRBase v21) | miRNA (miRBase v22.1) | miRNA family | miRNA ID | Sequence (5´-3´) | Length (nt) | Stem-loop strand |
|---|---|---|---|---|---|---|
| mtr-miR1510a-3p_1ss1CT | mtr-miR1510a-3p_1ss1CT | MIR1510 | pvu-miR1510a | TGGAGGATTAGG TAAAACAAC | 21 | **3p** |
| gma-miR1515a | gma-miR1515a | MIR1515 | pvu-miR1515a | TCATTTTGCGTG CAATGATCTG | 22 | **5p** |
| gma-MIR156k-p3 | zma-miR156d-3p_1SS21GA_R+1 | MIR156 | pvu-miR156d | GCTCACTTCTCTT TCTGTCAACT | 23 | **3p** |
| mtr-miR156b-3p_L-1_1ss5CT | mtr-miR156c-3p_L-1 | | pvu-miR156c | GCTTACTCTCTAT CTGTCACC | 21 | **3p** |
| gma-miR156a | mdm-miR156i | | pvu-miR156i | TGACAGAAGAGA GTGAGCAC | 20 | **5p** |
| gma-miR156a_R+1 | gma-miR156s | | pvu-miR156s | TGACAGAAGAGA GTGAGCACT | 21 | **5p** |
| gma-miR156b_L+1R-1 | mdm-miR156w | | pvu-miR156w | TTGACAGAAGAG AGAGAGCAC | 21 | **5p** |
| gma-miR156c | mdm-miR156ab | | pvu-miR156ab | TTGACAGAAGAT AGAGAGCAC | 21 | **5p** |

**Table 1. Cont.**

| miRNA (miRBase v21) | miRNA (miRBase v22.1) | miRNA family | miRNA ID | Sequence (5´-3´) | Length (nt) | Stem-loop strand |
|---|---|---|---|---|---|---|
| pvu-miR159a.1 | pvu-miR159a.1 | MIR159 | pvu-miR159a.1 | TTTGGATTGAAGGGAGCTCTA | 21 | **3p** |
| vun-MIR319a-p5 | mdm-miR319b-5p_L+1 | | pvu-miR319b | TGAGCTTTCTTCAGTCCACTC | 21 | **5p** |
| vun-miR319b_L-2R+2 | gma-miR319q | | pvu-miR319q | TGGACTGAAGGGAGCTCCTTC | 21 | **3p** |
| gma-miR319a | gma-miR319a | | pvu-miR319a | TTGGACTGAAGGGAGCTCCC | 20 | **3p** |
| gma-miR159a-5p_R-1 | gma-miR159a-5p_R-1 | | pvu-miR159a | GAGCTCCTTGAAGTCCAATT | 20 | **5p** |
| gma-miR160a-5p | ata-miR160a-5p | MIR160 | pvu-miR160a | TGCCTGGCTCCCTGTATGCCA | 21 | **5p** |
| gma-MIR162c-p5 | stu-miR162b-5p | MIR162_1 | pvu-miR162b | GGAGGCAGCGGTTCATCGATC | 21 | **5p** |
| gma-miR162a_R+1 | vvi-miR162 | | pvu-miR162 | TCGATAAACCTCTGCATCCAG | 21 | **3p** |
| mes-miR164a | mes-miR164a | MIR164 | pvu-miR164a | TGGAGAAGCAGGGCACGTGCA | 21 | **5p** |

**Table 1. Cont.**

| miRNA (miRBase v21) | miRNA (miRBase v22.1) | miRNA family | miRNA ID | Sequence (5´-3´) | Length (nt) | Stem-loop strand |
|---|---|---|---|---|---|---|
| pvu-miR166a_L+1R-2 | gma-miR166k_L-1 | MIR166 | pvu-miR166g | CTCGGACCAGGC TTCATTCC | 20 | **3p** |
| gma-miR166a-5p_L-1R+1 | gma-miR166a-5p_L-1R+1 | | pvu-miR166d | GAATGTTGTCTG GCTCGAGGA | 21 | **5p** |
| gma-miR166i-5p_R+1 | gma-miR166i-5p_R+1 | | pvu-miR166i | GGAATGTCGTCT GGTTCGAGA | 21 | **5p** |
| gma-MIR166e-p5 | mtr-miR166e-5p | | pvu-miR166e | GGAATGTTGGCT GGCTCGAGG | 21 | **5p** |
| gma-miR166a-5p | aly-miR166c-5p | | pvu-miR166c | GGAATGTTGTCT GGCTCGAGG | 21 | **5p** |
| pvu-miR166a_L+1R-1 | bdi-miR166e-3p_1SS1CG | | pvu-miR166m | GTCGGACCAGGC TTCATTCCC | 21 | **5p** |
| gma-miR166a-3p_1ss21CA | aly-miR166b-3p_1SS21CA | | pvu-miR166b | TCGGACCAGGCT TCATTCCCA | 21 | **3p** |
| gma-miR166a-3p | pvu-miR166a | | pvu-miR166a | TCGGACCAGGCT TCATTCCCC | 21 | **3p** |
| pvu-miR166a_1ss21CG | gma-miR166j-3p | | pvu-miR166j | TCGGACCAGGCT TCATTCCCG | 21 | **3p** |

**Table 1. Cont.**

| miRNA (miRBase v21) | miRNA (miRBase v22.1) | miRNA family | miRNA ID | Sequence (5´-3´) | Length (nt) | Stem-loop strand |
|---|---|---|---|---|---|---|
| pvu-MIR166a-p3 | gma-miR166k_R-1 | MIR166 | pvu-miR166k | TCTCGGACCAGGCTTCATTC | 20 | **3p** |
| pvu-miR166a_L+2R-2 | gma-miR166k | | pvu-miR166h | TCTCGGACCAGGCTTCATTCC | 21 | **3p** |
| ahy-miR167-3p | ptc-miR167g-3p | MIR167_1 | pvu-miR167g | AGATCATGTGGCAGTTTCACC | 21 | **3p** |
| gma-miR167a_R-1 | mdm-miR167b_R-1 | | pvu-miR167b | TGAAGCTGCCAGCATGATCT | 20 | **5p** |
| gma-miR167c_R+1 | ath-miR167d | | pvu-miR167d | TGAAGCTGCCAGCATGATCTGG | 22 | **5p** |
| gma-miR167e_R+1 | mdm-miR167j | | pvu-miR167j | TGAAGCTGCCAGCATGATCTTA | 22 | **5p** |
| gma-MIR168a-p3 | ptc-miR168b-3p | MIR168 | pvu-miR168b | CCCGCCTTGCATCAACTGAAT | 21 | **3p** |
| ath-miR168b-3p_R+1_2ss5TC10TC | aly-miR168a-3p_R+1 | | pvu-miR168c | CCCGCCTTGCATCAACTGAATT | 22 | **3p** |
| gma-miR168a | bna-miR168a | | pvu-miR168a | TCGCTTGGTGCAGGTCGGGAA | 21 | **5p** |

**Table 1. Cont.**

| miRNA (miRBase v21) | miRNA (miRBase v22.1) | miRNA family | miRNA ID | Sequence (5´-3´) | Length (nt) | Stem-loop strand |
|---|---|---|---|---|---|---|
| gma-miR169k | gma-miR169k | MIR169_2 | pvu-miR169k | CAGCCAAGAATG ACTTGCCGG | 21 | **5p** |
| vun-miR169 | zma-miR169c-5p | | pvu-miR169c | CAGCCAAGGATG ACTTGCCGG | 21 | **5p** |
| gma-miR169a_R+1 | gma-miR169a_R+1 | | pvu-miR169g | CAGCCAAGGATG ACTTGCCGGA | 22 | **5p** |
| vun-MIR169-p3 | zma-miR169a-3p_L+1R-1_1SS8TC | | pvu-miR169b | CGGCAAGTGGTT CTTGGCTAC | 21 | **3p** |
| gma-miR169l-3p_L-2_1ss14TC | zma-miR169a-3p_R-1 | | pvu-miR169a | GGCAAGTTGTTC TTGGCTAC | 20 | **3p** |
| gma-miR171j-5p | ath-miR171a-5p | MIR171_1 | pvu-miR171a | TATTGGCCTGGT TCACTCAGA | 21 | **5p** |
| gma-miR171m | gma-miR171m | | pvu-miR171m | TTGAGCCGCGTC AATATCTCA | 21 | **3p** |
| pvu-miR2118 | pvu-miR2118 | MIR2118 | pvu-miR2118 | TTGCCGATTCCA CCCATTCCTA | 22 | **3p** |
| mtr-miR319c-5p_L-1R+1_1ss8CT | mtr-miR319c-5p_L-1R+1_1SS8CT | MIR319 | pvu-miR319r | GAGTTCTTTGCA GCCCAAAGC | 21 | **5p** |

**Table 1. Cont.**

| miRNA (miRBase v21) | miRNA (miRBase v22.1) | miRNA family | miRNA ID | Sequence (5´-3´) | Length (nt) | Stem-loop strand |
|---|---|---|---|---|---|---|
| gma-miR319g_L+1R-2 | vvi-miR319e | MIR319 | pvu-miR319e | TTTGGACTGAAGGGAGCTCCT | 21 | **3p** |
| gma-miR390b-5p | gma-miR390b-5p | MIR390 | pvu-miR390b | AAGCTCAGGAGGGATAGCACC | 21 | **5p** |
| ath-miR390a-5p | ghr-miR390a | | pvu-miR390a | AAGCTCAGGAGGGATAGCGCC | 21 | **5p** |
| ath-miR390a-3p_R-1 | aly-miR390a-3p_R-1 | | pvu-miR390c | CGCTATCCATCCTGAGTTTC | 20 | **3p** |
| gma-miR393a_R+1 | gma-miR393e | MIR393 | pvu-miR393e | TCCAAAGGGATCGCATTGATCC | 22 | **5p** |
| gma-miR394a-5p_R-2 | gma-miR394c-5p_R-2 | MIR394 | pvu-miR394c | TTGGCATTCTGTCCACCT | 18 | **5p** |
| gma-miR394a-5p | cpa-miR394a | | pvu-miR394b | TTGGCATTCTGTCCACCTCC | 20 | **5p** |
| gma-miR394a-5p_R+1 | vvi-miR394a_R-1_1SS21AT | | pvu-miR394a | TTGGCATTCTGTCCACCTCCT | 21 | **5p** |
| gma-MIR395f-p5_1ss21GA | aly-miR395e-5p_2SS19TC20TA_L+1R-1 | MIR395 | pvu-miR395e | AGTTCCTCTGAGCACTTCACA | 21 | **5p** |

**Table 1. Cont.**

| miRNA (miRBase v21) | miRNA (miRBase v22.1) | miRNA family | miRNA ID | Sequence (5´-3´) | Length (nt) | Stem-loop strand |
|---|---|---|---|---|---|---|
| gma-miR395a_1ss18AG | gma-miR395a_1ss18AG | MIR395 | pvu-miR395a | CTGAAGTGTTTG GGGGAGCTC | 21 | **3p** |
| gma-miR395d | gma-miR395g | | pvu-miR395g | TGAAGTGTTTGG GGGAACTTT | 21 | **3p** |
| gma-miR396b-3p | mtr-miR396a-3p | MIR396 | pvu-miR396b | GCTCAAGAAAGC TGTGGGAGA | 21 | **3p** |
| gma-miR396a-3p_L+1 | sly-miR396a-3p | | pvu-miR396i | GTTCAATAAAGC TGTGGGAAG | 21 | **3p** |
| gma-miR396a-5p | mes-miR396a | | pvu-miR396a | TTCCACAGCTTT CTTGAACTG | 21 | **5p** |
| gma-miR396b-5p | atr-miR396e | | pvu-miR396e | TTCCACAGCTTT CTTGAACTT | 21 | **5p** |
| gma-miR398a_L+2R-2 | stu-miR398b-3p_L+1R-1 | MIR398 | pvu-miR398b | TTTGTGTTCTCAG GTCACCCC | 21 | **3p** |
| zma-miR399f-5p_1ss22AT | zma-miR399f-5p_1ss22AT | MIR399 | pvu-miR399f | GGGCAACTTCTC CTTTGGCAGT | 22 | **5p** |
| mtr-miR399b | osa-miR399i | | pvu-miR399i | TGCCAAAGGAGA GCTGCCCTG | 21 | **3p** |

**Table 1. Cont.**

| miRNA (miRBase v21) | miRNA (miRBase v22.1) | miRNA family | miRNA ID | Sequence (5´-3´) | Length (nt) | Stem-loop strand |
|---|---|---|---|---|---|---|
| pvu-miR399a | pvu-miR399a | MIR399 | pvu-miR399a | TGCCAAAGGAGA GTTGCCCTG | 21 | **3p** |
| gma-miR403a | gma-miR403a | MIR403 | pvu-miR403a | TTAGATTCACGC ACAAACTTG | 21 | **3p** |
| gma-miR408a-3p | gma-miR408c-3p | MIR408 | pvu-miR408c | ATGCACTGCCTC TTCCCTGGC | 21 | **3p** |
| vun-MIR482-p5 | pvu-miR482-5p_R-1 | MIR482 | pvu-miR482 | GGAATGGGCTGA TTGGGAAGC | 21 | **5p** |
| gma-miR482b-3p | gma-miR482b-3p | | pvu-miR482b | TCTTCCCTACAC CTCCCATACC | 22 | **3p** |
| vun-miR482_L-2 | vun-miR482_L-2 | | pvu-miR482a | TTCCCAATTCCG CCCATTCCTA | 22 | **3p** |
| PC-5p-11154_803 | gma-MIR530a_R+1_1SS16AG | MIR530 | pvu-MIR530a | TGCATTTGCACC TGCGCTTTG | 21 | **5p** |
| gma-MIR399h-p5 | gma-miR399m_L+1R-1_2SS5TA20TG | | pvu-miR399m | AGGGCACCTCTC TCCTGGCAG | 21 | **5p** |
| gma-MIR156f-p3 | csi-miR156g-3p_2SS8AC12TC | | pvu-miR156g | GCTCTCTCTTCC TCTGTCATC | 21 | **3p** |

**Table 1. Cont.**

| miRNA (miRBase v21) | miRNA (miRBase v22.1) | miRNA family | miRNA ID | Sequence (5´-3´) | Length (nt) | Stem-loop strand |
|---|---|---|---|---|---|---|
| pvu-miR1514a | pvu-miR1514a | | pvu-miR1514a | TTCATTTTGAAAA TAGGCATTG | 22 | **5p** |

**Table 2.** New miRNAs found expressed during seed development in *Phaseolus vulgaris*.

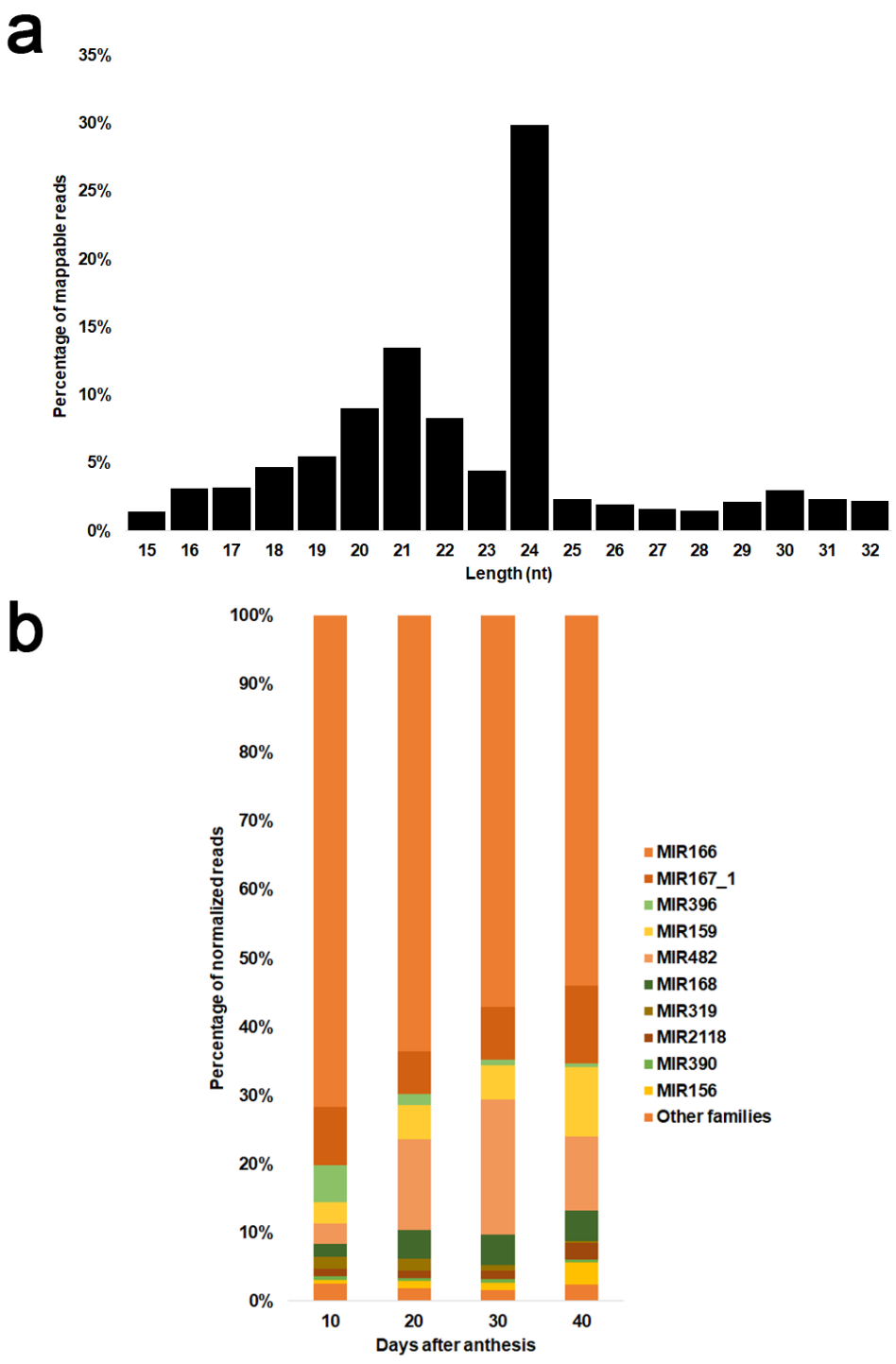
MFEI: Minimum folding free energy index, MFE: Minimal Free Energy

| miRNA ID | Sequence (5´-3´) | Length (nt) | MFE (kcal/mol) | MFEI |
|---|---|---|---|---|
| **miR_1** | TTGTTTTTCCTATTCCACCAAT | 22 | -41.8 | 1.2 |
| **miR_2** | TTTAAGAATTTCAGTTATGC | 20 | -41 | 1.2 |
| **miR_3** | TAACTGAATATTCTTAAAG | 19 | -41 | 1.2 |
| **miR_4** | GCTCTCTATATTTCTGTCATC | 21 | -48 | 1 |
| **miR_5** | TGCTGCTAGTTCATGGATACC | 21 | -83.7 | 1.1 |
| **miR_6** | CATGTGCCCCTCTTCCCCATC | 21 | -55.3 | 0.8 |
| **miR_7** | GGCAAGTTGGCCTTGGCTATA | 21 | -69.1 | 1 |
| **miR_8** | AAGTAGAGTGCAGCCAAGGAT | 21 | -63.5 | 1 |
| **miR_9** | TCGTCCTGAGACCACATGAGA | 21 | -61.4 | 0.9 |
| **miR_10** | GGGCAATGGCTTCTTTGGCAGT | 22 | -59.6 | 0.8 |
| **miR_11** | AACCTTGGTGACTAATTAGATACC | 24 | -83.3 | 2.2 |
| **miR_12** | TTCTTTCAAACAGGCCCTGAG | 21 | -48.1 | 1.1 |
| **miR_13** | AATAGAATTCTAGATTAGAAGATC | 24 | -37.9 | 0.8 |
| **miR_14** | AGAGACCGAGACACGTCATGACGT | 24 | -34.1 | 1 |
| **miR_15** | CCCGTGCGACCAAAATAAATTATT | 24 | -54.5 | 0.8 |
| **miR_16** | TATGATTTCCTTTGCTTCCTC | 21 | -72.6 | 1.3 |
| **miR_17** | TCTTCTCTCTATTGTCACCTT | 21 | -52 | 0.9 |
| **miR_18** | CATTTGTTTCTTTCTCTCCTTATA | 24 | -50.7 | 1.2 |
| **miR_19** | AATGGAGCATAGATATTGAATATA | 24 | -82.5 | 1.5 |
| **miR_20** | TTAGATTTTAAAATTTGGGAC | 21 | -17 | 1.3 |
| **miR_21** | TGGATGAGACGCTCATTTGAG | 21 | -45.5 | 1.1 |
| **miR_22** | GACACGGACACATCATTTAAGAGA | 24 | -30.8 | 1 |
| **miR_23** | AAACGAATTTTCAACGTGGACTGT | 24 | -35.2 | 0.9 |
| **miR_24** | ACAAGAGGCAGAAAGTAGAGTG | 22 | -57.9 | 1 |
| **miR_25** | CATCAAGATTGTATACAACTCT | 22 | -93.6 | 1.5 |
| **miR_26** | CAAATGAGTATTCCATCTACA | 21 | -45.5 | 1.1 |
| **miR_27** | ATGAAAATCCAGAACTCATAACTC | 24 | -36.4 | 0.9 |
| **miR_28** | CTAATCAAGGAAATCACAGTAG | 22 | -33.5 | 0.8 |
| **miR_29** | CGTTCCTGCATGGGGGCACCA | 21 | -87.2 | 1.1 |
| **miR_30** | GCCGAGAATTGAGATCCATAACTC | 24 | -52.5 | 0.8 |
| **miR_31** | AAAATAAATTATTATGGTCGTCGT | 24 | -54.5 | 0.8 |
| **miR_32** | GGTTTGTGCGTGAATCTGACG | 21 | -44.88 | 1.1 |
| **miR_33** | TAACTGAATATTCTTAAAGCCT | 22 | -41 | 1.2 |
| **miR_34** | AAGGGTTTCTATCAGAGTTTA | 21 | -48.6 | 0.8 |
| **miR_35** | TGTAAAACCCTGAATCCAAACCAT | 24 | -79.8 | 1 |
| **miR_36** | CAATGAGGGCATGTTGTAGGC | 21 | -54.5 | 1 |
| **miR_37** | GGGCATATCTTCTTTGGCACA | 21 | -56.6 | 0.9 |
| **miR_38** | ACCAGAGTTTCCATTTTAACATAT | 24 | -71 | 1.7 |
| **miR_39** | TAAGATGAAAACAAGGATACACT | 23 | -66.4 | 0.9 |

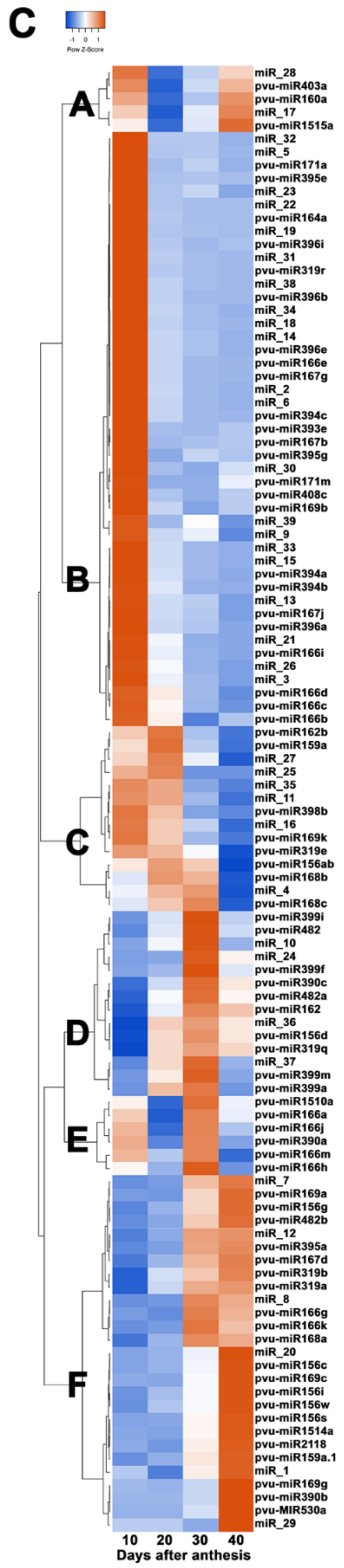
### 3.2 Twenty-five known miRNA families were found expressed during seed development

Seventy-two known miRNAs were identified, belonging to 25 miRNA families (**Table 1**). MIR166 is the most represented family based on the highest percentage of reads (**Figure 1b, Supplementary Table 6**), having the highest number of isoforms (11). Pvu-miR166a and pvu-miR166h are the most expressed miRNAs showing no relevant changes on their abundance (**Supplementary Figure 2**). Families MIR167_1 and MIR482 also presented high read number. MIR156 and MIR169 also stood out with different expression profiles during SD. A peak of expression was observed at 30 DAA for MIR399, MIR482 and MIR168, while on MIR396 the isoform expression decreases along SD (**Supplementary Figure 2, Supplementary Table 4**).

The abundance of 60 known miRNAs showed significant differences among all time-points (ANOVA adj. P-value ≤ 0.05) (**Supplementary Table 4**). Among those, 23 were differentially expressed (DE) between 10 and 20 DAA (adj. P-value ≤ 0.05) and none in the other comparisons established. Between 10 and 40 DAA, 44 known miRNAs were found DE (**Supplementary Figure 3**).




**MicroRNAs expression dynamics reveal post-transcriptional mechanisms regulating seed development in *Phaseolus vulgaris* L.**


C
Row Z-Score
A
B
C
D
E
F
miR_28
pvu-miR403a
pvu-miR160a
miR_17
pvu-miR1515a
miR_32
miR_5
pvu-miR171a
pvu-miR395e
miR_23
miR_22
pvu-miR164a
miR_19
pvu-miR396i
miR_31
pvu-miR319r
miR_38
pvu-miR396b
miR_34
miR_18
miR_14
pvu-miR396e
pvu-miR166e
pvu-miR167g
miR_2
miR_6
pvu-miR394c
pvu-miR393e
pvu-miR167b
pvu-miR395g
miR_30
pvu-miR171m
pvu-miR408c
pvu-miR169b
miR_39
miR_9
miR_33
miR_15
pvu-miR394a
pvu-miR394b
miR_13
pvu-miR167j
pvu-miR396a
miR_21
pvu-miR166i
miR_26
miR_3
pvu-miR166d
pvu-miR166c
pvu-miR166b
pvu-miR162b
pvu-miR159a
miR_27
miR_25
miR_35
miR_11
pvu-miR398b
miR_16
pvu-miR169k
pvu-miR319e
pvu-miR156ab
pvu-miR168b
miR_4
pvu-miR168c
pvu-miR399i
pvu-miR482
miR_10
miR_24
pvu-miR399f
pvu-miR390c
pvu-miR482a
pvu-miR162
miR_36
pvu-miR156d
pvu-miR319q
miR_37
pvu-miR399m
pvu-miR399a
pvu-miR1510a
pvu-miR166a
pvu-miR166j
pvu-miR390a
pvu-miR166m
pvu-miR166h
miR_7
pvu-miR169a
pvu-miR156g
pvu-miR482b
miR_12
pvu-miR395a
pvu-miR167d
pvu-miR319b
pvu-miR319a
miR_8
pvu-miR166g
pvu-miR166k
pvu-miR168a
miR_20
pvu-miR156c
pvu-miR169c
pvu-miR156i
pvu-miR156w
pvu-miR156s
pvu-miR1514a
pvu-miR2118
pvu-miR159a.1
miR_1
pvu-miR169g
pvu-miR390b
pvu-MIR530a
miR_29
10 20 30 40
Days after anthesis

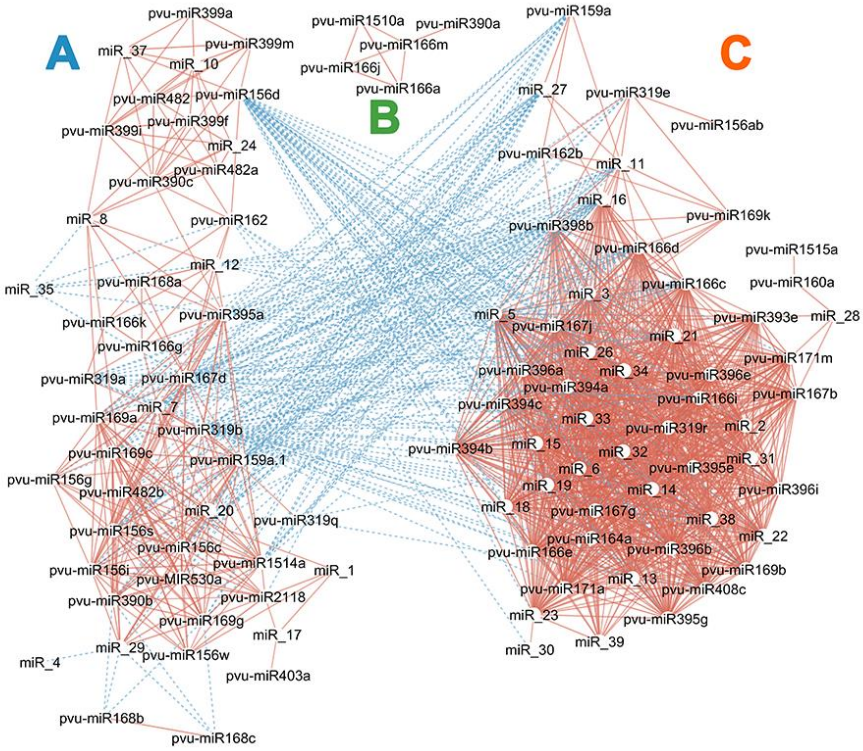



**Figure 1.** Characterization of sRNA-Seq data obtained from developing seeds of Phaseolus vulgaris at 10, 20, 30 and 40 days after anthesis (DAA). (a) Length distribution of mappable sRNA reads found expressed. (b) Percentage of the total amount of normalized reads for miRNA families expressed. The percentage was calculated as the sum of the average normalized reads of the miRNAs in each family, at each time-point, divided by the total average normalized reads for all known miRNAs at that time-point. (c) Heatmap depicting known and new miRNAs expression profiles. The heatmap was made using the average normalized read number for each miRNA at each timepoint. Pearson as distance measurement method and Complete Linkage as clustering method were applied. (d) Molecular interaction network of expressed miRNAs. The normalized read values from sRNA-Seq data were used on the correlation analysis. Only significant (P-value ≤ 0.05) correlations higher than 0.75 or lower than -0.75 are depicted. MiRNAs were used as source nodes (white circles) and the colour of the lines (edges) are red for a positive correlation or blue for negative correlation.

## 3.3 Thirty-nine new miRNAs were found expressed during seed development

Thirty-nine new miRNAs were identified, 17 with length of 21 nt and the remaining with length between 19 and 24 nt (**Table 2, Supplementary Table 5**). Some new miRNAs have aligned, with more mismatches than those allowed by our criteria, against known mature miRNAs/stem-loops available in miRbase v22.1 **(Supplementary Table 5)**. These observations provide insights into the families that these miRNAs might belong to. For instance, miR_28 partially aligned to gma-miR1509a, belonging to the MIR1509 family, while miR_33 partially aligned against gma-miR1512a-5p belonging to the MIR1512 family.

The abundance of 32 out of 39 new miRNAs, showed significant differences among all time-points (ANOVA adj. P-value ≤ 0.05) (**Supplementary Table 5**). Among those, 14 were DE between 10 and 20 DAA (adj. P-value ≤ 0.05) and none in the other comparisons stablished. Between 10 and 40 DAA, 23 new miRNAs were found DE (**Supplementary Figure 3**).

## 3.4 A high number of miRNAs accumulates preferentially during late embryogenesis and desiccation

PCA analysis conducted with all expressed miRNAs showed that the replicates of each studied stage clustered together, discriminating well the 4 experimental groups (**Supplementary Figure 4**). Hierarchical clustering highlighted six major miRNA abundance clusters with different abundance peaks along SD (**Figure 1c**). MiRNAs grouped in cluster A have high abundance at 10 and 40 DAA, while those on cluster B have the highest abundance at 10 DAA, decreasing afterwards. Cluster C grouped miRNAs with lowest abundance at 40 DAA. Cluster D groups miRNAs with highest abundance at 30 DAA, while Cluster E groups miRNAs with high abundance

at 10 and 30 DAA. In cluster F, miRNAs with the highest expression at 30 and/or 40 DAA were found.

Correlation networks were made with the 106 miRNAs that presented a strong and significant correlation ($R^2 \leq -0.75$ and $R^2 \geq +0.75$, P-value $\leq 0.05$) in their expression (**Figure 1d, Supplementary Table 7)**. Among others, miRNA groups A and C have inverse correlations between them. Most of group A miRNAs had the highest abundance at 40 DAA, while for group C most of the miRNA had the highest abundance at 10 DAA. Several miRNAs showed more than 50 significant correlations with other miRNAs, such as pvu-miR166d, miR_16, pvu-miR398b, pvu-miR166c and miR_26.



### 3.5 Validation of sRNA sequencing data by RT-qPCR

Two known (pvu-miR399a, pvu-miR156i) and seven new DE miRNAs (miR_6, miR_11, miR_16, miR_18, miR_29, miR_33 and miR_38) were selected for RT-qPCR validation of sequencing results and to examine their expression pattern along eight SD stages studied (**Supplementary Figure 5**). A strong positive correlation ($0.71 \leq R^2 \leq 0.99$) between sRNA-Seq and RT-qPCR data (**Supplementary Table 8**) was evidenced for most miRNAs, with exception of miR_29.

### 3.6 Targets for eight expressed miRNAs were validated through degradome analysis

A total of 124,993,267 raw reads were obtained from degradome sequencing, with a total of 3,048 miRNA/cleaved target pairs identified after CleaveLand v4.3 analysis (**Supplementary Tables 9 and 10**). Of these, 133 were found in categories between 0 and 2 with a significant (P-value < 0.05) association between miRNA and target (**Supplementary Table 10**). Based on our miRNA expression threshold (100 normalized reads), only 10 miRNA:cleaved target pairs were considered in this study (**Supplementary Table 11 and Supplementary Figure 6**). As examples, the *DEHYDRIN*

*FAMILY PROTEIN (RAB18)* is a validated miR18 target, while *DEAD BOX RNA HELICASE (PRH75)* is a validated miR_6 target.

### 3.7 Targets for seventy-three miRNA were predicted

To complement degradome information, targets for 73 miRNAs were predicted using psRNATarget (**Supplementary Table 12**). Several predicted targets for known miRNAs encode for transcription factors (TFs) (**Supplementary Table 12**). Members of the *SQUAMOSA PROMOTER BINDING PROTEIN-LIKE* (*SPL2*, *SPL4*, *SPL8*, *SPL9*, *SPL10* and *SPL12*) were predicted targets of several MIR156 members. Members of the HOMEODOMAIN-LEUCINE ZIPPER (HD-ZIP) family (*HB-8, PHB*, *REV* and *CNA*) were predicted targets of the MIR166. Nuclear Factor Y family TFs (*NF-YA1, NF-YA2, NF-YA3, NF-YA8, NF-YA9* and *NF-YA10*) were described as putative targets of MIR169. NAC family TFs (*NAC1, NAC100, NTL9* and *CUC2*) were predicted targets of pvu-miR164a and pvu-miR1514a, while the *ABA-INDUCIBLE BASIC/HELIX-LOOP-HELIX–TYPE TRANSCRIPTION FACTOR* and zinc knuckle (CCHC-type) family protein were predicted targets of pvu-miR530a. *AUXIN RESPONSE FACTORS* (*ARF10*, *ARF16* and *ARF17*) were found targeted by pvu-miR160a, while GRAS TFs family were found targeted by pvu-miR171m. *SPOROCYTELESS/NOZZLE (SPL/NZZ)* was a predicted target of pvu-miR159a.1. The *HYDROXYPROLINE-RICH GLYCOPROTEIN FAMILY PROTEIN (HRGP)* and *MYB DOMAIN PROTEIN 33 (MYB33)* were predicted targets of pvu-miR319a. *HRGP* is also a predicted target of pvu-miR319q.

Transcripts implicated in phosphate metabolism, such as *PHOSPHATE 2* (*PHO2*) and *PHOSPHATE TRANSPORTER 1;1* (*PHT1;1*) were predicted targets of pvu-miR339a and pvu-miR399i. *SUCROSE-PROTON SYMPORTER 2* (*SUT1*), the *ATP SULFURYLASE 1* (*APS*) and the *SULFATE TRANSPORTER 2;1* (*SULTR2*) were predicted targets of the MIR395 members, while the *CAROTENOID CLEAVAGE DIOXYGENASE 1* (*NCED1*) was predicted target of MIR167 members. Most of the new miRNAs

identified were predicted to target transcripts encoding functional proteins (**Supplementary Table 12**). No miRNAs putatively targeting TFs implicated in SD such as LEC1, LEC2, L1L, ABI3 and FUS3[31], despite the changes in their abundance along SD (**Supplementary Figure 7**).

No agreement between degradome and target prediction analysis was found. Hits on the same targets identified by degradome were observed only after running psRNAtarget with less restrictive parameters (E values >5) and inputting both miRNAs and target sequences (**Supplementary Table 13**).



Molecular interaction networks were established between miRNAs grouped in the same expression cluster (**Figure 1c**) and respective target MapMan functional category to investigate if any target functional category is preferentially regulated (**Figure 2; Supplementary Table 14 and 15**). The six networks generated evidenced a relation between miRNA expression profiles and putative functions repressed by them during SD. In some cases (e.g. pvu-miR395e), the same miRNA was found to target multiple transcripts belonging to different functional categories. Focusing on the most relevant networks, the network constructed with miRNAs more accumulated at 10 DAA (**Figure 2b**), shows a high number of miRNAs putatively repressing targets categorized as “Protein”, “RNA”, “Transport” and “Not assigned”. MiRNAs targeting functional categories “Hormone metabolism”, “S-assimilation” and “Secondary metabolism” were also observed, suggesting metabolic processes likely modulated on late embryogenesis/early seed filling.

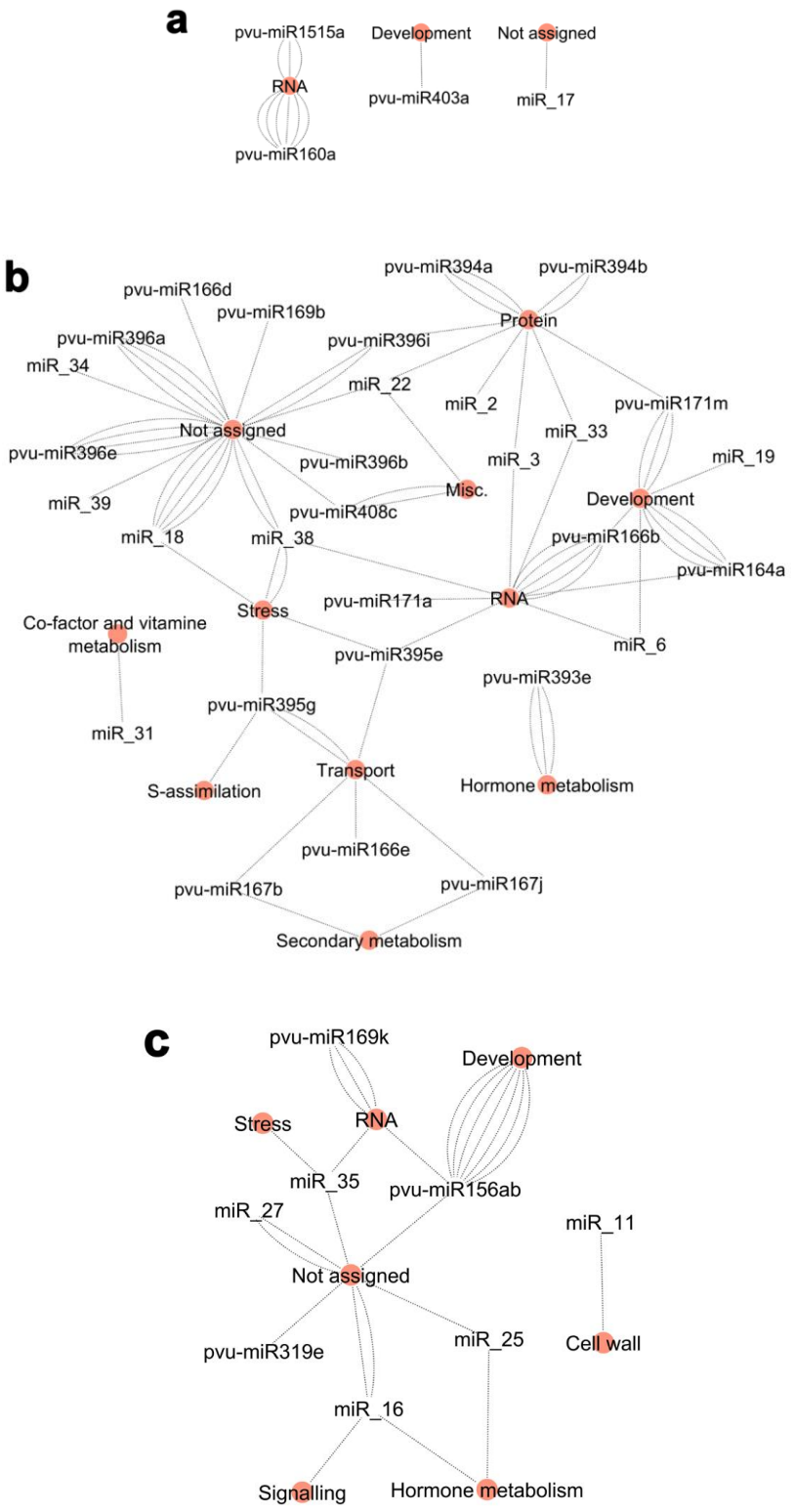
The network constructed with miRNAs more accumulated 40 DAA (**Figure 2e**) showed that putatively repressed targets are mainly categorized as “Development”, “RNA” and "Stress" and to a less extent “Protein” and “not assigned”. This network evidences that several TFs of the SPL family (MIR156) categorized as “Development” and NF-YAs (MIR169) or HD-ZIP

(MIR166) categorized both as “RNA”, as well as, several disease resistance proteins and receptors categorized as “Stress” are being potentially repressed at this stage.

## 3.8 Evidences of the relevance of identified miRNAs on seed development



Since miRNAs promote the cleavage/inhibition of target mRNAs, the analysis of the expression of miRNAs and their predicted/validated targets profiles can provide indirect evidence of their action. Our results evidenced time-points, in which target down-regulation mediated by miRNAs might be occurring (**Figure 3 and 4**). Five MIR166 members were predicted to target HD-ZIP TFs such as REV, PHB, CNA and HB8 in *P. vulgaris* seeds. As one example, the high accumulation of miR166g contrasted with the relative low expression levels of its targets *CNA*, *HB-8*, *PHB* and *REV* (**Figure 3d**). Although pvu-miR166e is more accumulated at 10 DAA, its expression significantly decreases at 20 DAA (P-value ≤0.05). At the same timepoint, its degradome validated target *RAN1* shows a low abundance (**Figure 3m**).




**MicroRNAs expression dynamics reveal post-transcriptional mechanisms regulating seed development in *Phaseolus vulgaris* L.**

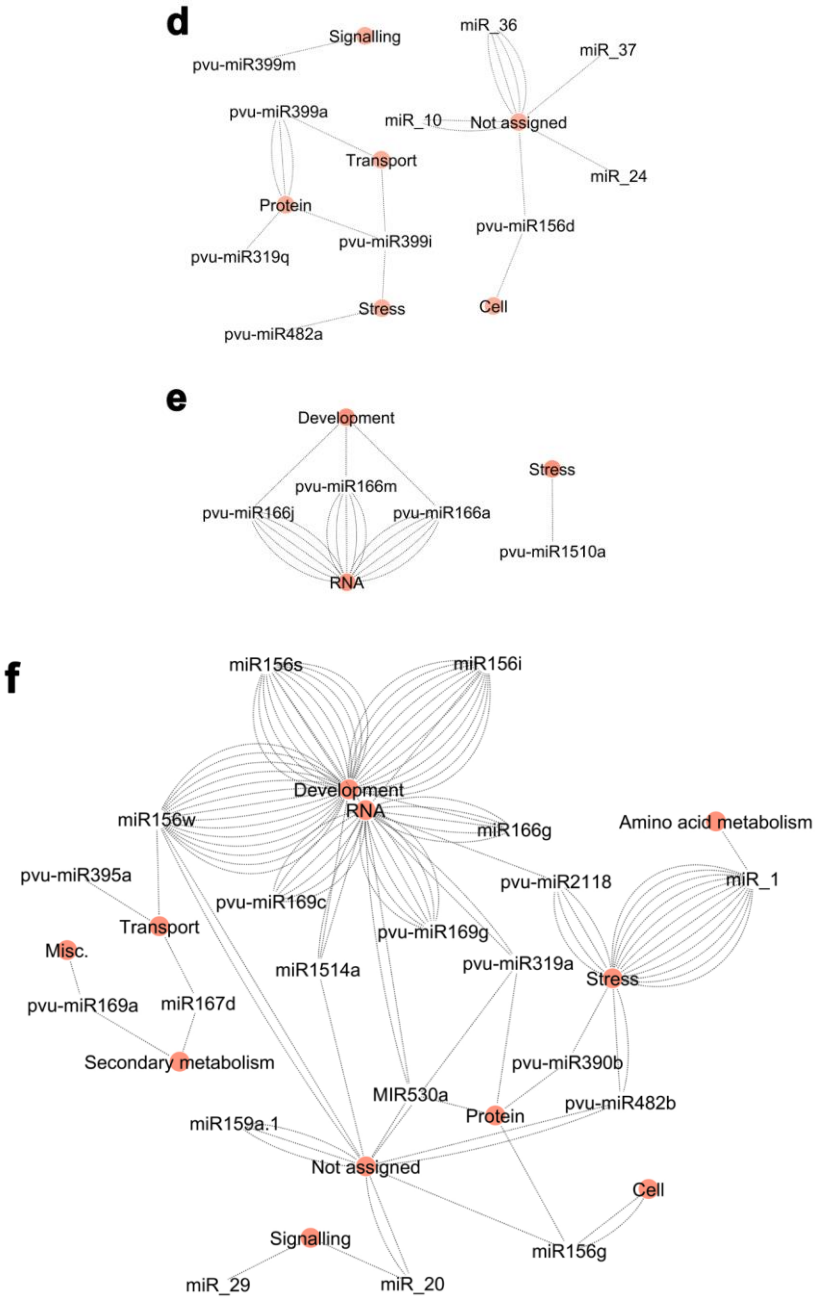
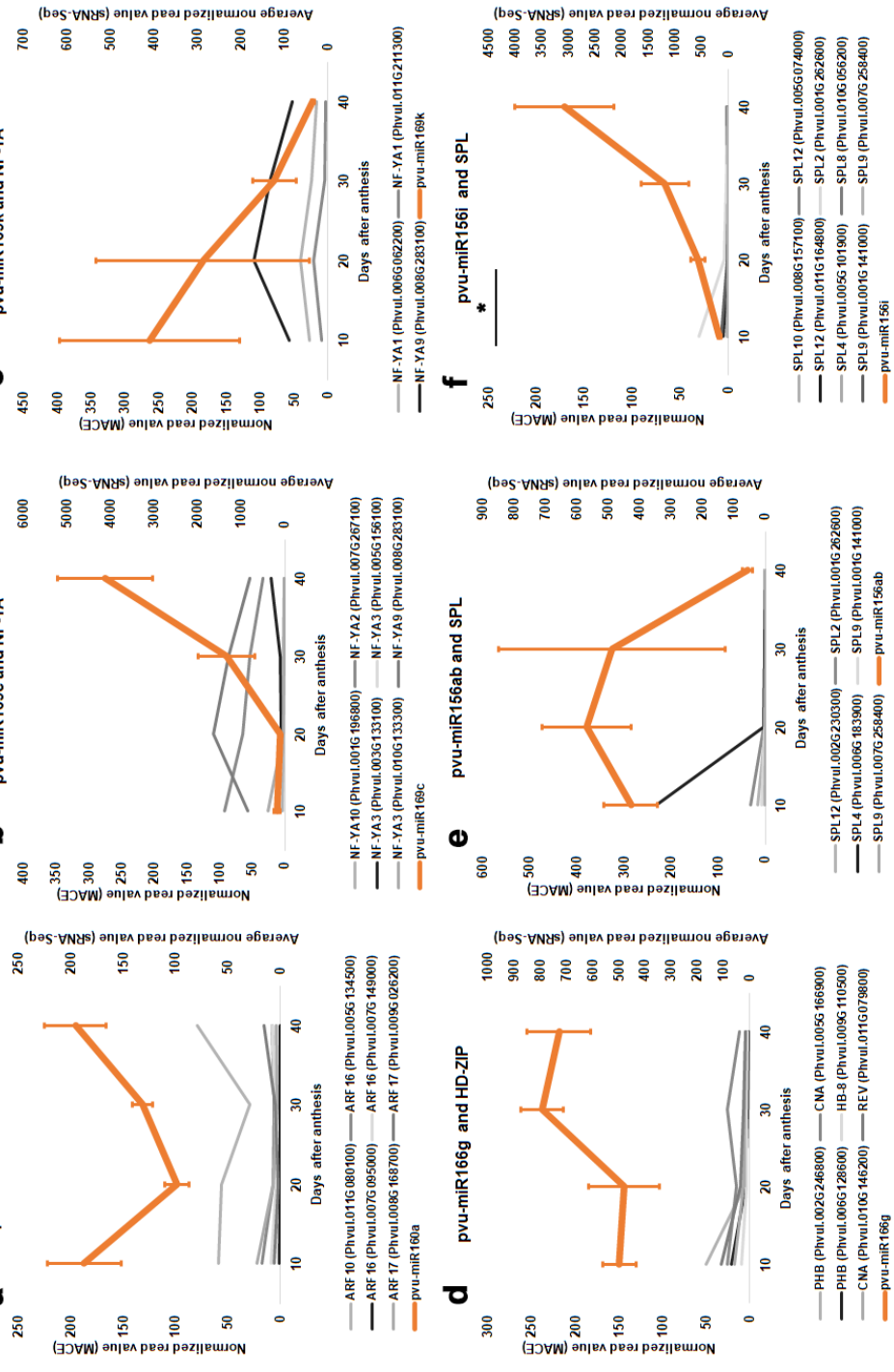
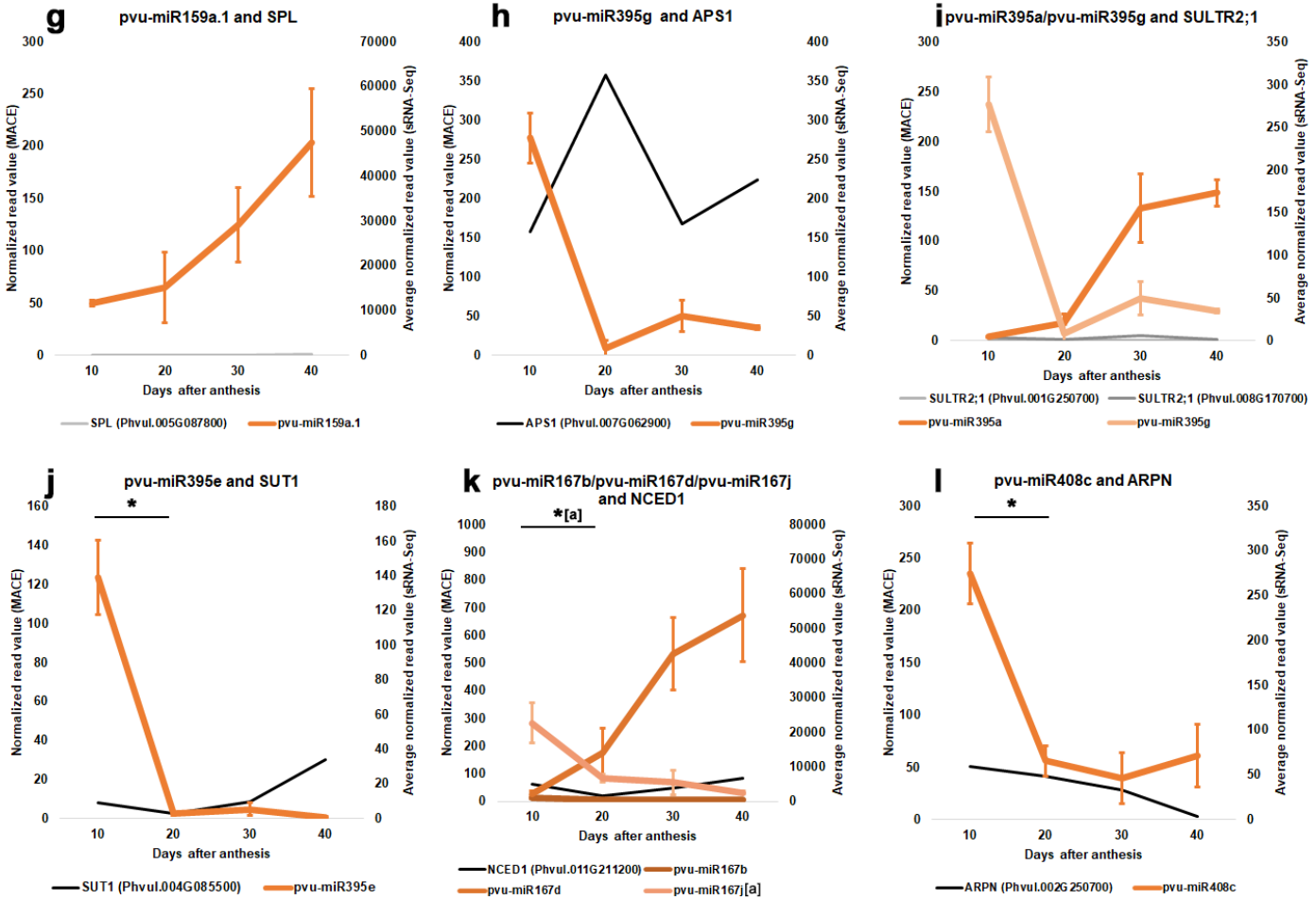



**Figure 2.** Molecular interaction network of expressed miRNAs connected to MapMan functional categories of their target genes. Each network was produced with the miRNAs from the clusters obtained using Heatmapper. (a) Cluster A. (b) Cluster B. (c) Cluster C. (d) Cluster D. (e) Cluster E. (f) Cluster F. Lines represent genes that encode the target transcripts and connect miRNAs (white circles) to Mercator functional categories of the targets (orange circles).




**MicroRNAs expression dynamics reveal post-transcriptional mechanisms regulating seed development in *Phaseolus vulgaris* L.**


g
pvu-miR159a.1 and SPL
Normalized read value (MACE)
Average normalized read value (sRNA-Seq)
Days after anthesis
SPL (Phvul.005G087800)
pvu-miR159a.1
h
pvu-miR395g and APS1
Normalized read value (MACE)
Average normalized read value (sRNA-Seq)
Days after anthesis
APS1 (Phvul.007G062900)
pvu-miR395g
i
pvu-miR395a/pvu-miR395g and SULTR2;1
Normalized read value (MACE)
Average normalized read value (sRNA-Seq)
Days after anthesis
SULTR2;1 (Phvul.001G250700)
SULTR2;1 (Phvul.008G170700)
pvu-miR395a
pvu-miR395g
j
pvu-miR395e and SUT1
*
Normalized read value (MACE)
Average normalized read value (sRNA-Seq)
Days after anthesis
SUT1 (Phvul.004G085500)
pvu-miR395e
k
pvu-miR167b/pvu-miR167d/pvu-miR167j and NCED1
*[a]
Normalized read value (MACE)
Average normalized read value (sRNA-Seq)
Days after anthesis
NCED1 (Phvul.011G211200)
pvu-miR167b
pvu-miR167d
pvu-miR167j[a]
l
pvu-miR408c and ARPN
*
Normalized read value (MACE)
Average normalized read value (sRNA-Seq)
Days after anthesis
ARPN (Phvul.002G250700)
pvu-miR408c

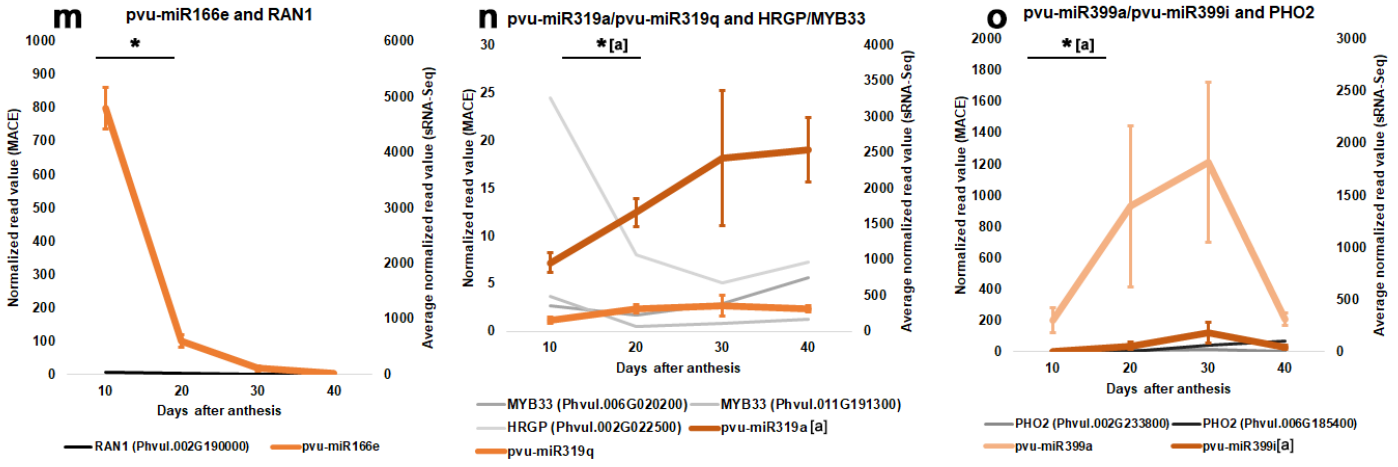
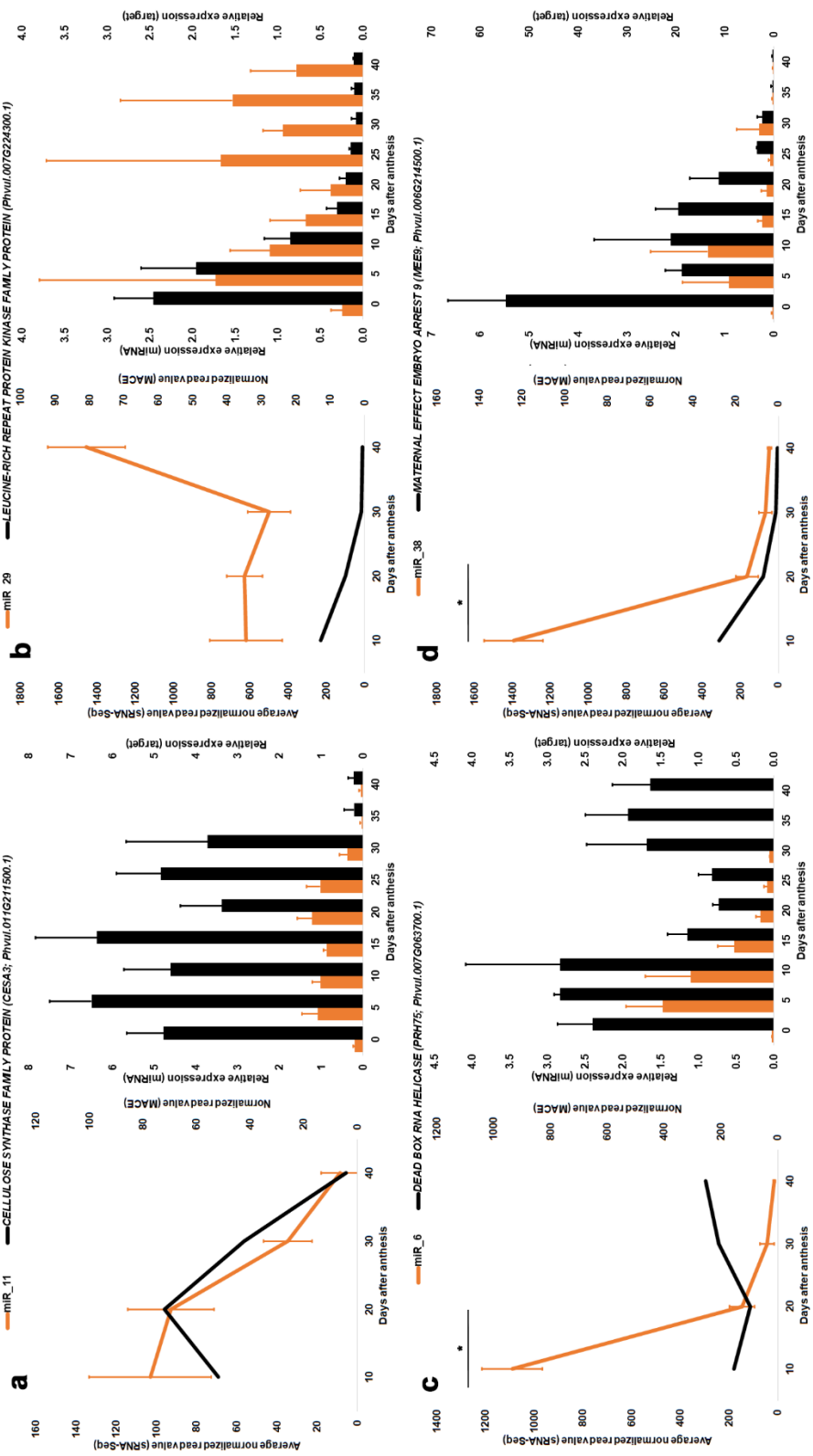
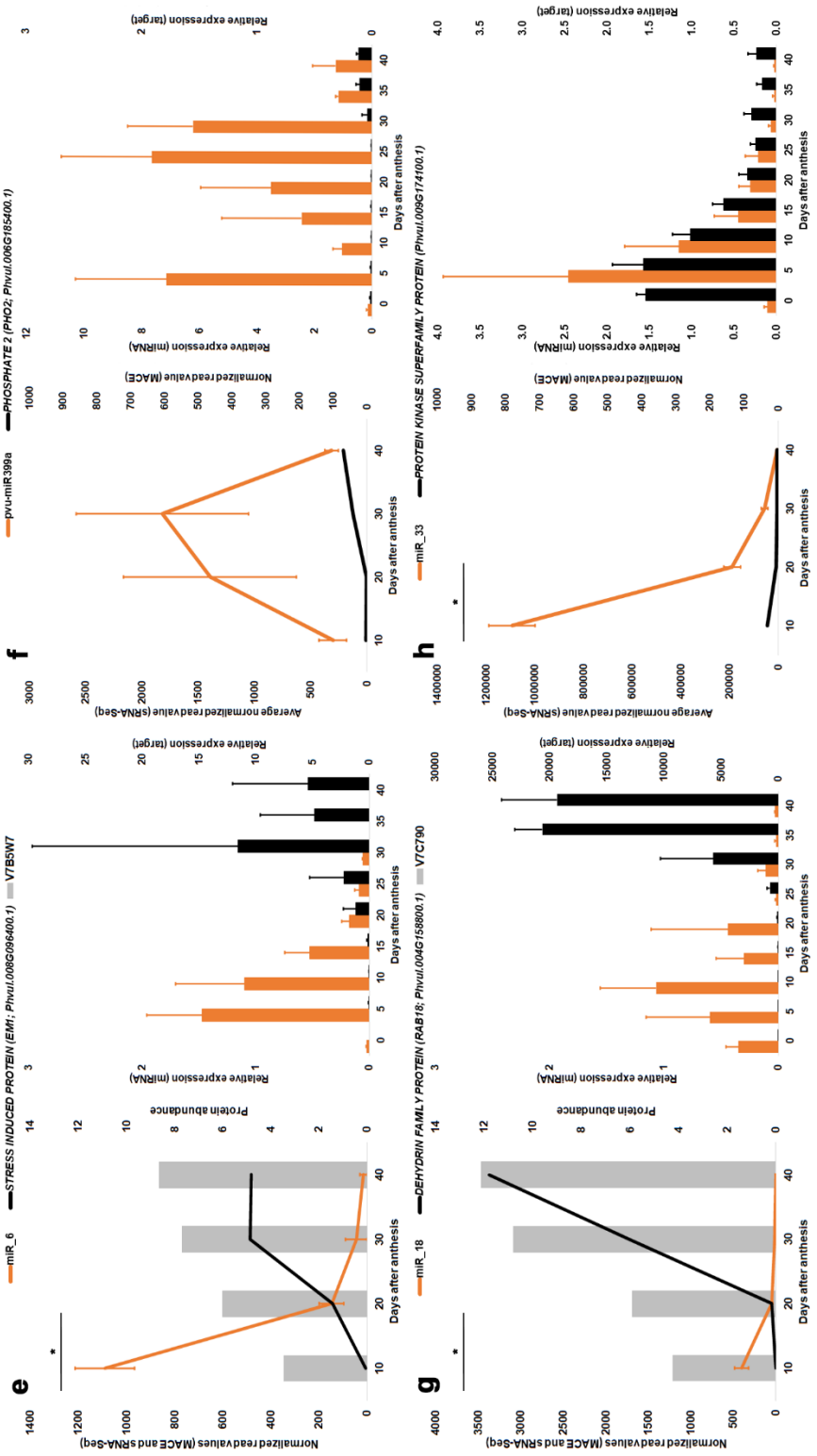



**Figure 3.** Expression profiles of selected miRNAs and targets during seed development in Phaseolus vulgaris. The expression profiles of the miRNAs were obtained using the average normalized read values from sRNA-Seq analysis at 10, 20, 30 and 40 days after anthesis (DAA). Error bars show the standard deviation from three sequenced biological replicates and the asterisk indicates a statistically significant difference (adj P-value ≤ 0.05) between consecutive timepoints. When more than one miRNA is depicted, the asterisk is followed by letters inside square brackets indicating the DE miRNA. The target expression profiles were produced using the normalized read value for target genes at the same timepoints retrieved from MACE datasets available in Parreira *et al*. (2018).


a
miR_11
CELLULOSE SYNTHASE FAMLY PROTEIN (CESA3; Phvul.011G211500.1)
b
miR_29
LEUCINE-RICH REPEAT PROTEIN KINASE FAMLY PROTEIN (Phvul.007G224300.1)
c
miR_6
DEAD BOX RNA HELICASE (PRH75; Phvul.007G063700.1)
d
miR_38
MATERNAL EFFECT EMBRYO ARREST 9 (MEE9; Phvul.006G214500.1)
Average normalized read value (sRNA-Seq)
Normalized read value (MACE)
Relative expression (miRNA)
Relative expression (target)
Days after anthesis
*




**MicroRNAs expression dynamics reveal post-transcriptional mechanisms regulating seed development in *Phaseolus vulgaris* L.**

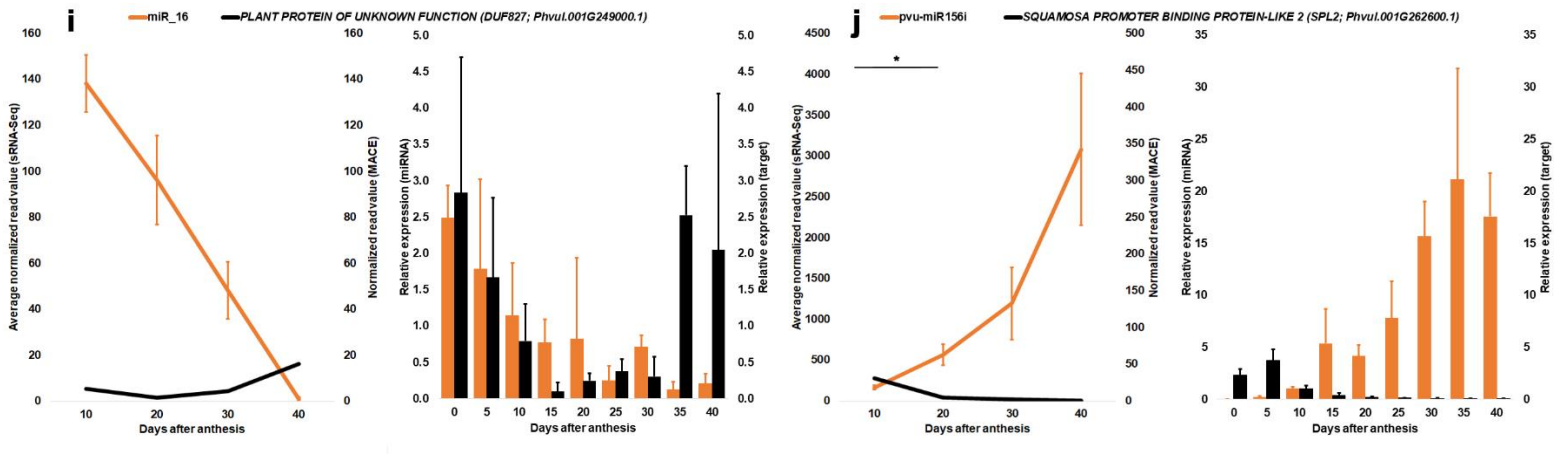



**Figure 4.** Expression profiles of selected miRNAs and targets differentially expressed during seed development in Phaseolus vulgaris. (Left panel) Expression profiles of the miRNAs were obtained using the average normalized read values from sRNA-Seq analysis at 10, 20, 30 and 40 days after anthesis (DAA). Error bars show the standard deviation from three sequenced biological replicates, the asterisk indicates a statistically significant difference (adj P-value ≤ 0.05) between consecutive timepoints. The target expression profiles were produced using the normalized read value for target genes at the same timepoints retrieved from MACE datasets available in Parreira et al. (2018). For e. and g., the protein average normalized intensity values at the same timepoints retrieved from the datasets available in Parreira et al. (2016). (Right panel) Relative expression (RT-qPCR) profiles were obtained for miRNA:target pairs in the ovary (herein defined as 0 DAA) and at subsequent eight seed development time-points.

The highest abundances of miR169k and miR169c were seen at 10 DAA and 40 DAA, respectively, while abundance of their target NF-YA9 remained low on these points (**Figure 3b and c**). NF-YA1, another predicted miR169k target, also showed low abundances at 10 DAA.

Along SD, a progressive accumulation of pvu-miR156i was observed, while a contrasting decrease on its targets *SPL2, SPL9, SPL12* abundance was observed at same time-points (**Figure 4j and Figure 3f**). Similar miRNA and target expression profiles were seen for pvu-miR159.1a and pvu-miR319a during SD, while the expression of their targets was found very low (**Figure 3g and Supplementary Table 4**). Indeed, significant (P-value ≤ 0.05) increases in the pvu-miR156i ($Log_2$ FC=4.23), pvu-miR159.1a ($Log_2$ FC=2.03), pvu-miR319a ($Log_2$ FC=1.40) and pvu-miR319q ($Log_2$ FC=1.01) abundances were observed when comparing 10 DAA values with 40 DAA ones (**Supplementary Table 4**).

Small RNA-Seq data showed a high accumulation of miR_6 (targeting *EM1* and *PRH75*), miR_18 (targeting *RAB18)*, miR_33 (targeting *PROTEIN KINASE SUPERFAMILY*) and miR_38 (targeting *MEE9)* at 10 DAA but decreasing significantly (P≤0.05) at 20 DAA (**Figure 4c, d, e, g and h**). The expression of their respective targets was found low at 10 DAA. RT-qPCR data showed that the accumulation of miR_6, miR_33 and miR_38 occurred earlier than 10 DAA, namely at 5 DAA (**Figure 4c, e, g and h**).

Numerous miRNAs targeting transcripts implicated in storage compounds and mineral allocation, signalling and homeostasis were found DE during SD. The pvu-miR395e, predicted to target *SUT1*, was strongly accumulated at 10 DAA contrasting with the low abundances seen for *SUT1* (**Figure 3j**). At 20 DAA, pvu-miR395e abundances were significantly (P-value ≤ 0.05) lower than those observed at 10 DAA ($Log_2$ FC=-5.62). Pvu-miR399a and pvu-miR399i, predicted to target *PHO2*, were found gradually accumulated till the seed start to dehydrate (30 DAA), decreasing afterwards

(**Figure 3o**). Indeed, for pvu-miR399i, a significant increase (P-value ≤ 0.05) in abundances were seen in the transition from 10 to 20 DAA ($Log_2$ FC=3.90) (**Figure 3o**). Pvu-miR408c, predicted to target *ARPN*, was found strongly accumulated at 10 DAA but its expression decreased significantly (P≤0.05) afterwards (**Figure 3l**).

Several miRNAs implicated in phytohormones metabolism have been identified. The pvu-miR160a, predicted to target *ARF10*, *ARF16* and *ARF17* has relatively high abundance along SD, while the expression of its targets remains low (**Figure 3a**). Pvu-miR167d and pvu-miR167j were found highly accumulated at 40 DAA and 10 DAA, respectively. This contrasted with the lower abundances seen for the predicted target *NCED1* (**Figure 3k**). *EM1* and *RAB18* are degradome validated targets of miR_6 and miR_18, respectively. Both miRNAs are highly accumulated at 10 DAA but their abundances strongly decreased significantly (P-value ≤ 0.01) at 20 DAA ($Log_2$ FC≤-6). Contrastingly, the transcript and proteomic abundances of their targets remains relatively low for the same time point (**Figure 4e and g**).



## 4. Discussion

Herein, we provide the first comprehensive overview of miRNAs accumulation dynamics during *P. vulgaris* SD, spanning from late embryogenesis to seed desiccation. Seventy-two known miRNAs, belonging to 25 families, were found expressed. Thirty-nine new miRNAs were identified contributing to expand the current number of *P. vulgaris* miRNAs described. Although members of the MIR159, MIR160, MIR166, MIR399 and MIR319 have been previously described in other *P.vulgaris* tissues [10,11,13,16], our study expands the evidences of their accumulation in developing seeds. For instance, the pvu-miR399a was found in common bean roots [16], the pvu-miR166b in seedling leaves [14] or the miR156 in

the leaves, roots and nodules [10]. Interestingly, some of the new miRNAs identified in this study have also been reported previously in *P. vulgaris* such as miR_32, miR_12, miR_18 and miR_19 identified as novel miRNAs by Formey *et al.* (2015) [11]. Importantly, none of these above-mentioned studies have provided comprehensive overview of miRNA abundances and repressed targets during SD in this important pulse as the present work herein described.



Degradome analysis and target prediction identified targets for 77 expressed miRNAs. Principal component analysis (**Supplementary Figure 4**), hierarchical clustering and expression correlation networks analysis (**Figure 1d and e**) of miRNA abundances highlight a timeframe for miRNA accumulation, in which different miRNA groups are accumulated at these stages. Based on abundances, 10 and 40 DAA were the timepoints where the miRNAs action seems more pronounced. The miRNA correlation analysis made (**Figure 1d**) evidenced two main clusters negatively correlated representing most miRNAs accumulated at early or late SD. MiRNAs/targets expression profiles provided biological evidences on relevance of these regulatory modules during SD, particularly in tuning distinct developmental stages in seeds. Nevertheless, given the complexity of regulatory networks since miRNAs can repress multiple targets, functional validation studies are needed to corroborate suggested roles in *P. vulgaris* developing seeds. This is particularly relevant to clarify if other regulatory mechanisms might be responsible for the observed down regulation of gene expression. A schematic overview of main seed developmental processes putatively regulated at post-transcriptional level is provided on **Figure 5.**

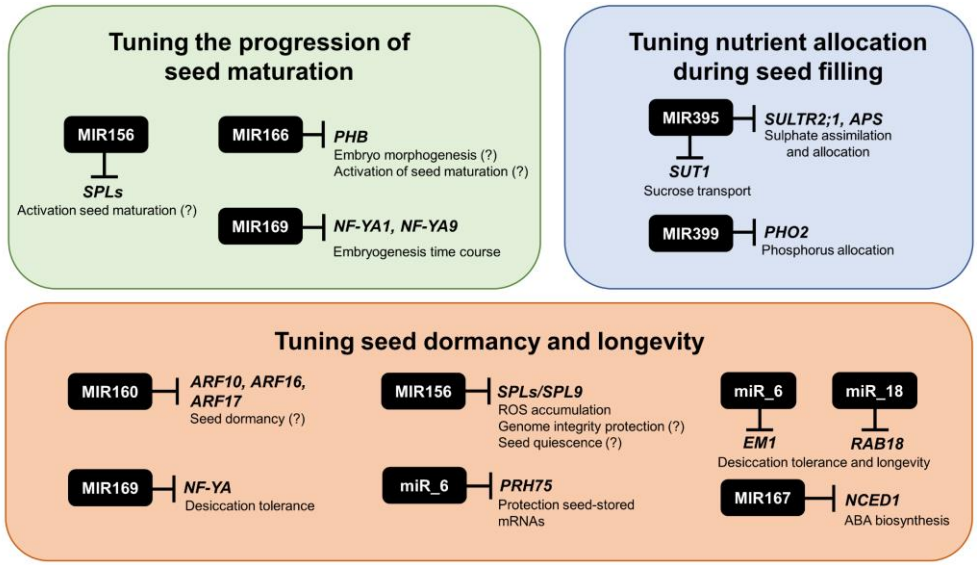



**Figure 5.** Schematic overview of main processes putatively regulated at post-transcriptional level by identified miRNAs during seed development in *Phaseolus vulgaris.*

## 4.1 Tuning the progression of seed maturation

During *P. vulgaris* SD, a global HD-ZIP TFs repression driven by the accumulation of MIR166 members is suggested. MIR166 was the most expressed family in this study (**Figure 1b**) and is possibly regulating *REV*, *PHB*, *CNA* and *HB8* expression during all studied points. The miR166-mediated repression of *PHB*/*PHV* in early embryogenesis is involved embryo symmetry and tissues patterning along the apical-basal axis [32]. An increased abundance of HD-ZIPs TFs at early SD was expected, since PHB is a direct activator of LEC2, which controls legume seed filling [33]. At 10 DAA, our MACE data evidenced that *PHB* abundances are already relatively low (**Figure 3d**) with residual *LEC2* abundances detected (**Supplementary Figure 7**). This suggests that the seed filing program activation might have occurred earlier than 10 DAA, the first sampling point for MACE and sRNA-seq. The small seed size increase seen from 5 from 10 DAA supports this. In other species, the increase in MIR166 abundances during SD progression [34] suggests a MIR166 role during seed maturation and quiescence. In *P.*

*vulgaris* developing seeds, the biological relevance of the high abundance for many MIR166 members along SD is still puzzling deserving further investigation.

NF-YA, NF-YB, and NF-YC are subunits of the NF-Y heterotrimer TFs that binds to the CCAAT box in target promoter [35]. LEC1 (NF-YB9) and its closely related homolog LEC1-like (L1L, NF-YB6) are seed filing master regulators in legume and non-legume species [31,36]. A putative pvu-miR169k-mediated repression of *NF-A1* and *NF-A9* was seen at 10 DAA (**Figure 3c**). NF-YA1 and NF-YA9 repression [37] has been implicated in defining embryogenesis time course. Our data supports a similar role this regulatory module.



MiR156-mediated repression *SPL10* and *SPL11* during early stages of embryogenesis, prevents the precocious expression of maturation genes [38]. FUS3 triggers the expression MIR156 members implicated in the *SPL10* and *SPL11* regulation [39]. At 10 DAA, the abundances of pvu-miR156i, *SPL10* and other SPLs, as well as seed filing regulators LEC2 and FUS3 are almost residuals (**Figure 3f**, **Supplementary Figure 7**), suggesting that activation of the seed filing program occurred earlier as described for MIR166. The majority of MIR156 members accumulates at the later SD (**Supplementary Figure 2**) suggesting a global repression of SPL expression at 40 DAA. A MIR156 role in seed quiescence with implication on germination performance is suggested, since MIR156-SPL regulatory module has been implicated on juvenile to adult transition phase of vegetative development [40].

The miR159 mediated silencing of a family of genes encoding R2R3 MYB TFs referred to as “GAMYB” or “GAMYB-like” implicated in gibberellins metabolism was found relevant to prevent abnormal vegetative development [41]. Contrary to what was described for the most of land plants, in our data GAMYB homologues as MYB33 were predicted to be targets of miR139 as

described for *Marchantia polymorpha* [42]. The miRNA accumulation profiles miR319/miR159 members along SD allow us to speculate that the repression of their targets is relevant during seed maturation and quiescence, but more studies are needed to clarify this.

## 4.2 Tuning nutrient allocation during seed filing



We obtained indirect evidence that sucrose metabolism might be regulated at the post-transcriptional level in *P.vulgaris* seeds. In this species, the transporter *SUT1* , a sucrose/H+ symporter, is highly expressed in seed coat cells and seems to be involved in sucrose efflux from the coat to the seed apoplasm [43]. Our results suggest that pvu-miR395e may repress *SUT1* expression at embryogenesis, being this repression released when the seed enters filling (**Figure 3j**). In agreement with this, our previous proteomic results showed that the accumulation of proteins implicated in sugar metabolism, such as starch synthesis, and cell wall-related proteins occurs in the 10 to 20 DAA transition [4].

MiRNAs implicated in phosphorus (P) or sulphur (S) signaling and homeostasis were identified in our seed samples. In *P. vulgaris* seeds, phytic acid is the main stored form of P, which accumulates concomitantly with dry matter accumulation [44]. While miR399-mediated *PHO2* down-regulation has been implicated with P uptake on roots and translocation to shoots [45], little is known about the MIR399 role in seeds. Our study showed that MIR399 abundances (**Figure 3o, Supplementary Table 4 and Supplementary Figure 2**) increases until 30 DAA, with significant changes observed for pvu-miR399i from 10 to 20 DAA. Such miRNA abundances profiles are in agreement with the seed filing observed our seed samples (**Supplementary Figure 5**).These results, together with the low *PHO2* expression levels observed, suggest a regulatory role for MIR399 to facilitate P translocation to the maturing seeds.

MiR395 is involved in of S assimilation and allocation regulation by targeting *APS* genes and *SULTR2;1* respectively [46]. In Arabidopsis reproductive tissues, the low affinity sulphate transporter SULTR2;1 is specifically expressed in siliques bases and veins or funiculus [47]. In agreement with this, we have found a relatively low expression of SULTR2;1 in developing *P. vulgaris* seeds contrasting with strong MIR395 accumulation (**Figure 3i**). In Arabidopsis, miR395 facilitates sulphate accumulation by targeting *APS* genes during sulphate starvation [46]. Our results highlight MIR395-mediated repression of *APS1* at 10 DAA (**Figure 3h**), suggesting a role of this regulatory module on S accumulation between end of embryogenesis and early filling. In *P. vulgaris* seeds, the contents of sulphur amino are sub-optimal for human nutrition [48]. The modulation of *SULTR2;1* expression in seeds by pvu-miR395 accumulation could be a promissory approach to tune sulphate flux and enhance synthesis of S-containing storage proteins.

### 4.3 Tuning seed dormancy and longevity

Auxins regulates Arabidopsis seed dormancy via a concerted action with ABA signaling pathway [49]. These authors suggested a feedback loop where ABI3-mediated repression of *MIR160B* gene, which targets *ARF10* and *ARF16,* allows the establishment of dormancy. In *P. vulgaris* seeds, the results obtained are puzzling since a potential MIR160-mediated repression of *ARF10*, *ARF16* and *ARF17* was seen along all SD (**Fig 3a**). Also, ABI3 expression levels where found increased at 30 DAA (**Supplementary Figure 7**), when the seed starts to dehydrate and become dormant. Altogether, this supports that other regulatory mechanisms of gene expression might be implicated in seed dormancy.

*NCED1* encodes 9-cis-epoxycarotenoid dioxygenase, a relevant enzyme in ABA biosynthesis, whose expression was negatively correlated with ABA accumulation under water deprivation in rice [50]. Our results,

suggest that some members of MIR167 potentially repress NCED1 expression (**Figure 3k**), and by this way contribute to tune ABA levels during *P. vulgaris* SD.

While it is generally accepted that ABA regulates the expression of many LEA genes, it has been recently shown that LEAs can be direct targets of ABI3 [51], as *EM1* [49]. LEA accumulation at late seed maturation and dehydration was implicated in embryo protection [52]. EM1 and RAB18 belong to LEA family and were found accumulated in *P. vulgaris* seeds [4]. *EM1* and *RAB18* were validated targets of miR_6 and miR_18, respectively (**Supplementary Table 11**). The high miRNA abundances seen at 10 DAA contrasts target transcript and protein accumulation profiles **(Figure 4e and g).** In orthodox seeds, desiccation tolerance is linked to seed longevity and EM1 role on this has been recently highlighted [49]. Reduced *RAB18* abundance was implicated in a decrease in seed longevity [53]. Also, the existence of DRE and ABRE cis-acting elements in the upstream region some MIR169 members targeting NF-YA suggests a role in desiccation tolerance [54]. Indeed, an upregulation of several MIR169 members under drought stress was observed in soybean [55]. In wheat, NF-YA1 decreases expression in drought stressed leaves [56]. Based on our MIR169 accumulation profiles (**Supplementary Figure 2**), it would be interesting to further investigate if NF-YA repression at later SD stages represents an adaptative seed response to desiccation with impact on seed viability.



Other mechanisms implicated in seed viability and posterior germination traits seem to be regulated at the post-transcriptional level, which includes the mechanisms to cope with oxidative damage. As one example, the MIR156/SPL9 regulatory module was implicated in the regulation of ROS accumulation, since miR156 overexpression mutants had ROS accumulation suppressed [57]. During seed filing and dehydration of *P. vulgaris* seeds, we evidenced the occurrence of oxidative damage [4] and its effects on the maintenance of embryo genome integrity [5]. Members of the

MIR156 increase expression throughout SD (**Supplementary Figure 2**). For pvu-miR156i, this contrasts with the lower abundances of its predicted SPL9 target (**Figure 3f**). This suggests that MIR156/SPL9 regulatory module can contribute to tune the ROS levels during SD and indirectly contribute to the maintenance of genome integrity. DEAD-box RNA helicases play various roles in RNA metabolism including among others ribosome biogenesis, mRNA splicing and translation [58]. Degradome analysis (**Supplementary Table 11**) validated *PRH75,* a nucleolar DEAD-box RNA helicase [59], *as* miR_6 target. MiR_6 expression is triggered on the onset of SD since expression levels were residuals on ovary tissues (**Figure 4c**). The contrasting miRNA/target accumulation profiles at 10 DAA, suggest a miR_6-mediated regulation of *PRH75* expression at late embryogenesis. This observation is intriguing since mutations in *PRH75* are lethal during embryogenesis in Arabidopsis [60]. During seed dormancy induction, seed-stored mRNAs for germination seem to be protected through their association with ribosome complexes that can include RNA binding proteins (RBP), such a DEAD-BOX RNA proteins [61]. A potential PHR75 involvement in this mechanism is suggested, based on its accumulation at later stages of SD.



## 5. Conclusion

We have identified 72 known and 39 new miRNAs expressed in developing seed of *P. vulgaris*. Degradome analysis and target prediction identified targets for 77 expressed miRNAs. Several transcription factor families, such as *HD-ZIP*, *ARF*, *SPL* and *NF-Y*, were identified as targets of known miRNAs, while most of new miRNAs targets were predicted to encode for functional proteins. Based on abundances, late embryogenesis/early filing (10 DAA) and desiccation (40 DAA) where stages in which miRNA action seems more pronounced. Altogether, the results support that these

miRNAs timely tune distinct seed developmental stages. They were found implicated on controlling embryogenesis time course, postponement of seed filing and maturation program, seed desiccation tolerance and longevity, with still uncovered roles on germination. The miRNAs herein described represent novel resources with potential application in future biotechnological approaches to modulate the expression of genes implicated in legume seed traits, including storage compound accumulation or seed viability. They could be exploited, even in combination, to simultaneously target more than one seed trait and be eventually translated to other agronomically relevant legumes with impact in agricultural and horticultural production systems.

## 6. Acknowledgements

The authors acknowledge Dr. Steve Beebe (CIAT-CGIAR, Colombia) for kindly supplying SER 16 *P. vulgaris* seeds. Funding from Fundação para a Ciência e a Tecnologia (Lisbon, Portugal) is acknowledged through research projects PTDC/AGR-GPL/110224/2009, research unit “GREEN-IT” (UID/Multi/04551/2020), J.R.P. PhD grant (PD/BD/113474/2015) from PhD program Plants for Life (PD/00035/2013) and S.S.A PhD holder contract (DL57). A.B. acknowledges the Italian Ministry of Education, University and Research (MIUR) through Dipartimenti di Eccellenza Program (2018–2022) for Department of Biology and Biotechnology “L. Spallanzani”, University of Pavia. M.C. acknowledges the grant by “Project North South Traineeship” Consortium from ERAMUS+/KA1 program, co-funded by European Commission and MIUR.

## 7. Supplementary Material

All supplementary materials are available at:
https://www.nature.com/articles/s41438-020-00448-0

**Supplementary Table 1. Characterization of the high-throughput sequenced sRNA libraries.**

**Supplementary Table 2. Length distribution of mappable reads obtained using sRNA-Seq.**

**Supplementary Table 3. Complete output of the ACGT101-miR v4.2 (LC Sciences) pipeline used for miRNA analysis.**



**Supplementary Table 4. Known miRNAs found expressed during seed development in *P. vulgaris*.** The miRNA annotation in miRBase 21 and 22.1 is provided. miRNA family and strand were obtained from miRBase v22.1 (Gene family). Sequence, length, average, normalized and raw read value for each sample and time-point are provided. Statistical analysis is provided and was performed using DeSeq2.

**Supplementary Table 5. New miRNAs found expressed during seed development *in P. vulgaris*.** The miRNA annotation from LC-Sciences is provided as well as the miRNA ID used in this study. MiRNA length, secondary structure, MFE, MFEI, extended sequence, average, raw and normalized read value for each sample and time-point are provided. Statistical analysis is provided and was performed using DeSeq2.

**Supplementary Table 6. Total number of normalized reads per time point in each miRNA gene family.** The miRNA family (Gene family) was obtained from miRbase 22.1 using the reference miRNAs (Table S2).

**Supplementary Table 7. Product-moment correlation analyses made with 111 miRNAs expressed during seed development in *P. vulgaris*.** Pairwise correlations $-0.58 \leq R2 \geq 0.58$ are significant at the $P \leq 0.05$ level. Correlation analyses were conducted using Statistica, version 6 (Statsoft).

**Supplementary Table 8. Correlations between datasets.** Correlation values between the Log2 of the average normalized read value for each miRNA (small RNA-seq) and the Log2 of the average miRNA expression ratio (RT-qPCR).

**Supplementary Table 9. Characterization of the high-throughput degradome libraries.**

**Supplementary Table 10. Full degradome data of the 10, 20, 30 and 40 days after anthesis samples from *P. vulgaris* seed development, obtained using CleaveLand v4.3.**

**Supplementary Table 11. Selected degradome miRNA/target pairs within categories 0, 1 and 2, with a p-value < 0.05. Degradome category and p-value, for each pair, was calculated using CleaveLand v4.3 (Addo-Quaye *et al*., 2009).**

**Supplementary Table 12. Putative targets for miRNAs obtained via psRNATarget Server (Schema V2, 2017 release) using the mature sequences of the selected miRNAs against the cDNA library of *Phaseolus vulgaris* v2.1 (Phytozome 12).** Default parameters were used, with the exception of Expectation: 2.



**Supplementary Table 13. MiRNAs identified in the degradome dataset and respective targets that were not found in accordance with the target psRNAtarget prediction (expectation (E) values ≤2).** psRNATarget Server (Schema V2, 2017 release) was used, with default parameter, except for Expectation:20. Hits on the same targets identified by degradome were observed after running psRNAtarget with less restrictive parameters (E values >5) and inputting both miRNAs and target sequences.

**Supplementary Table 14. Percentage of mercator functional categories of the target genes found in this study.** The percentage was calculated as the sum of the known or new expressed miRNAs in each functional category, divided by the total number of known or new expressed miRNAs, respectively.

**Supplementary Table 15. Members of each cluster performed using Heatmapper and their respective targets and Mercator category and subcategory.**

**Supplementary Table 16. Details of the two-tailed miRNA primers used for the relative quantification of gene expression using RT-qPCR.**

**Supplementary Table 17. Details of the predicted target primers used in the relative quantification of gene expression using RT-qPCR.**

**Supplementary Figure 1. Secondary (hairpin) structure of extended sequences for the candidate miRNAs identified in the study were obtained using RNAFold (http://rna.tbi.univie.ac.at//cgi-bin/RNAWebSuite/RNAfold.cgi).** The calculated Minimal Folding Energy (MFE) of the new miRNA precursors was lower than -17/kcal/mol. The sequence of the mature miRNA is highlighted in blue, while the sequence of the mature miRNA* is highlighted in orange, if found in our sequencing data.

**Supplementary Figure 2. Expression profiles of all expressed miRNAs within selected MIR families during seed development in *Phaseolus vulgaris*.** The expression profiles of the miRNAs were obtained using the Log2 average normalized read values from sRNA-Seq results obtained from samples collected at 10, 20, 30 and 40 days after anthesis.

**Supplementary Figure 3. Venn diagrams exhibiting the number of miRNAs differentially expressed (DE) during seed development in *Phaseolus vulgaris*.** Venn diagrams were constructed using a freely available web-based tool (http://bioinformatics.psb.ugent.be/webtools/Venn/) with differentially expressed (DE) miRNAs ANOVA and T-test (adj. P ≤ 0.05) between studied timepoints. Since no DE miRNAs were found between 20 vs 30 DAA and 30 vs 40 DAA, they were not included in this analysis. The number of DE miRNAs in each comparison established (10 vs 20 and 10 vs. 40 DAA) are shown between parentheses.



**Supplementary Figure 4. Principal component analysis of all expressed miRNAs during seed development in *Phaseolus vulgaris*.** A principal component analysis (PCA) was performed using Statistica, version 6 (Statsoft) and the standardized (Z-score) normalized read values of the expressed miRNAs at 10, 20, 30 and 40 DAA were used as variables. (DAA - Days After Anthesis)

**Supplementary Figure 5. Characterization of the *Phaseolus vulgaris* seed development.** Seed morphology, seed fresh weight, dry weight and length were measured at eight consecutive timepoints along the seed development stage studied

**Supplementary Figure 6. Target plot (t-plots) of representative validated *Phaseolus vulgaris* miRNA targets in different categories as confirmed by degradome sequencing.**

**Supplementary Figure 7. Expression profiles of the members of the LAFL regulatory network during seed development in *Phaseolus vulgaris*.** Expression profiles were established using the normalized read value for each target genes at 10, 20, 30 and 40 days after anthesis retrieved from MACE datasets available in Parreira *et al.* (2018). *AP2/B3-LIKE TRANSCRIPTIONAL FACTOR FAMILY PROTEIN* (*ABI3*; Phvul.008G122100); *LEC1 HISTONE SUPERFAMILY PROTEIN* (*LEC1*; Phvul.006G057800); *AP2/B3-LIKE TRANSCRIPTIONAL FACTOR FAMILY PROTEIN* (*FUS3*; Phvul.001G086800); *NUCLEAR FACTOR Y, SUBUNIT B6* (L1L; Phvul.007G234800 and Phvul.003G087000); *AP2/B3-LIKE TRANSCRIPTIONAL FACTOR FAMILY PROTEIN* (*LEC2*; Phvul.002G229800)

## 8. Bibliography

1. Polania, J.A.; Chater, C.C.C.; Covarrubias, A.A.; Rao, I.M. *Phaseolus* species responses and tolerance to drought. In *The Plant Family Fabaceae. Biology and Physiological Responses to Environmental Stresses*; Hasanuzzaman, M., Araújo, S., Gill, S.S., Eds.; Springer Singapore: Singapore, **2020**; pp. 319–336 ISBN 9789811547515.

2. Vaz Patto, M.C.; Araújo, S.S. Positioning Portugal into the context of world production and research in grain legumes. *Rev. Ciências Agrárias* **2016**, *39*, 471–489.

3. Kang, Y.; Li, M.; Sinharoy, S.; Verdier, J. A Snapshot of functional genetic studies in *Medicago truncatula*. *Front. Plant Sci.* **2016**, *7*, 1175.

4. Parreira, J.R.; Bouraada, J.; Fitzpatrick, M.A.; Silvestre, S.; Bernardes da Silva, A.; Marques da Silva, J.; Almeida, A.M.; Fevereiro, P.; Altelaar, A.F.M.; Araújo, S.S. Differential proteomics reveals the hallmarks of seed development in common bean (*Phaseolus vulgaris* L.). *J. Proteomics* **2016**, *143*, 188–198.

5. Parreira, J.; Balestrazzi, A.; Fevereiro, P.; Araújo, S. Maintaining genome integrity during seed development in *Phaseolus vulgaris* l.: evidence from a transcriptomic profiling study. *Genes (Basel).* **2018**, *9*, 463.

6. Yu, L.; Guo, R.; Jiang, Y.; Ye, X.; Yang, Z.; Meng, Y.; Shao, C. Genome-wide identification and characterization of novel microRNAs in seed development of soybean. *Biosci. Biotechnol. Biochem.* **2019**, *83*, 233–242.

7. DeBoer, K.; Melser, S.; Sperschneider, J.; Kamphuis, L.G.; Garg, G.; Gao, L.; Frick, K.; Singh, K.B. Identification and profiling of narrow-leafed lupin (*Lupinus angustifolius*) microRNAs during seed development. *BMC Genomics* **2019**, *20*, 135.

8. Yu, Y.; Jia, T.; Chen, X. The 'how' and 'where' of plant microRNAs. *New Phytol.* **2017**, *216*, 1002–1017.

9. Arenas-Huertero, C.; Pérez, B.; Rabanal, F.; Blanco-Melo, D.; De la Rosa, C.; Estrada-Navarrete, G.; Sanchez, F.; Covarrubias, A.A.; Reyes, J.L. Conserved and novel miRNAs in the legume *Phaseolus vulgaris* in response to stress. *Plant Mol. Biol.* **2009**, *70*, 385–401.

10. Valdés-López, O.; Yang, S.S.; Aparicio-Fabre, R.; Graham, P.H.; Reyes, J.L.; Vance, C.P.; Hernández, G. MicroRNA expression profile in common bean (*Phaseolus vulgaris*) under nutrient deficiency stresses and manganese toxicity. *New Phytol.* **2010**, *187*, 805–818.

11. Formey, D.; Iñiguez, L.P.; Peláez, P.; Li, Y.-F.; Sunkar, R.; Sánchez, F.; Reyes, J.L.; Hernández, G. Genome-wide identification of the *Phaseolus vulgaris* sRNAome using small RNA and degradome sequencing. *BMC Genomics* **2015**, *16*, 423.

12. Pelaez, P.; Trejo, M.S.; Iniguez, L.P.; Estrada-Navarrete, G.; Covarrubias, A. a; Reyes, J.L.; Sanchez, F. Identification and

characterization of microRNAs in *Phaseolus vulgaris* by high-throughput sequencing. *BMC Genomics* **2012**, *13*, 83.

13. Contreras-Cubas, C.; Rabanal, F. a; Arenas-Huertero, C.; Ortiz, M. a; Covarrubias, A. a; Reyes, J.L. The *Phaseolus vulgaris* miR159a precursor encodes a second differentially expressed microRNA. *Plant Mol. Biol.* **2012**, *80*, 103–115.

14. Patwa, N.; Nithin, C.; Bahadur, R.P.; Basak, J. Identification and characterization of differentially expressed *Phaseolus vulgaris* miRNAs and their targets during mungbean yellow mosaic India virus infection reveals new insight into *Phaseolus*-MYMIV interaction. *Genomics* **2019**, *111*, 1333–1342.

15. Han, J.; Xie, H.; Kong, M.L.; Sun, Q.P.; Li, R.Z.; Pan, J.B. Computational identification of miRNAs and their targets in *Phaseolus vulgaris*. *Genet. Mol. Res.* **2014**, *13*, 310–322.

16. Valdés-López, O.; Arenas-Huertero, C.; Ramírez, M.; Girard, L.; Sánchez, F.; Vance, C.P.; Luis Reyes, J.; Hernández, G. Essential role of MYB transcription factor: PvPHR1 and microRNA: PvmiR399 in phosphorus-deficiency signalling in common bean roots. *Plant. Cell Environ.* **2008**, *31*, 1834–1843.

17. Wei, Z.; Liu, X.; Feng, T.; Chang, Y. Novel and conserved microRNAs in Dalian Purple Urchin (*Strongylocentrotus nudus*) identified by next generation sequencing. *Int. J. Biol. Sci.* **2011**, *7*, 180–192.

18. Li, M.; Xia, Y.; Gu, Y.; Zhang, K.; Lang, Q.; Chen, L.; Guan, J.; Luo, Z.; Chen, H.; Li, Y.; *et al*. MicroRNAome of porcine pre- and postnatal development. *PLoS One* **2010**, *5*, e11541.

19. Griffiths-Jones, S. The microRNA Registry. *Nucleic Acids Res.* **2004**, *32*, D109–D111.

20. Meyers, B.C.; Axtell, M.J.; Bartel, B.; Bartel, D.P.; Baulcombe, D.; Bowman, J.L.; Cao, X.; Carrington, J.C.; Chen, X.; Green, P.J.; *et al*. Criteria for annotation of plant microRNAs. *Plant Cell* **2008**, *20*, 3186–3190.

21. Zhong, M.; Huang, F.; Luo, R.; Lv, Y.; Ali, U.; Sheng, Z.; Tang, S.; Wei, X.; Hu, P. The effect of cadmium on the microRNAome, degradome and transcriptome of rice seedlings. *Plant Growth Regul.* **2020**, *90*, 15–27.

22. Zhu, H.; Chen, C.; Zeng, J.; Yun, Z.; Liu, Y.; Qu, H.; Jiang, Y.; Duan, X.; Xia, R. MicroRNA528, a hub regulator modulating ROS homeostasis via targeting of a diverse set of genes encoding copper-

containing proteins in monocots. *New Phytol.* **2020**, *225*, 385–399.

23. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **2011**, *17*, 10.

24. Addo-Quaye, C.; Miller, W.; Axtell, M.J. CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. *Bioinformatics* **2009**, *25*, 130–131.

25. Lohse, M.; Nagel, A.; Herter, T.; May, P.; Schroda, M.; Zrenner, R.; Tohge, T.; Fernie, A.R.; Stitt, M.; Usadel, B. Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. *Plant. Cell Environ.* **2014**, *37*, 1250–1258.

26. Shannon, P. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13*, 2498–2504.

27. Androvic, P.; Valihrach, L.; Elling, J.; Sjoback, R.; Kubista, M. Two-tailed RT-qPCR: a novel method for highly accurate miRNA quantification. *Nucleic Acids Res.* **2017**, *45*, e144–e144.

28. Ruijter, J.M.; Ramakers, C.; Hoogaars, W.M.H.H.; Karlen, Y.; Bakker, O.; van den Hoff, M.J.B.B.; Moorman, A.F.M.M. Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. *Nucleic Acids Res.* **2009**, *37*, e45–e45.

29. Pfaffl, M.W. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res.* **2001**, *29*, 45e – 45.

30. Zhang, B.H.; Pan, X.P.; Cox, S.B.; Cobb, G.P.; Anderson, T.A. Evidence that miRNAs are different from other RNAs. *Cell. Mol. Life Sci.* **2006**, *63*, 246–254.

31. Verdier, J.; Thompson, R.D. Transcriptional Regulation of storage protein synthesis during dicotyledon seed filling. *Plant Cell Physiol.* **2008**, *49*, 1263–1271.

32. Smith, Z.R.; Long, J.A. Control of Arabidopsis apical–basal embryo polarity by antagonistic transcription factors. *Nature* **2010**, *464*, 423–426.

33. Manan, S.; Ahmad, M.Z.; Zhang, G.; Chen, B.; Haq, B.U.; Yang, J.; Zhao, J. Soybean LEC2 regulates subsets of genes involved in controlling the biosynthesis and catabolism of seed storage substances and seed development. *Front. Plant Sci.* **2017**, *8*, 1–16.

34. Wang, X.G.; Ju, Z.H.; Hou, M.H.; Jiang, Q.; Yang, C.H.; Zhang, Y.; Sun, Y.; Li, R.L.; Wang, C.F.; Zhong, J.F.; *et al*. Deciphering

transcriptome and complex alternative splicing transcripts in mammary gland tissues from cows naturally infected with *Staphylococcus aureus* mastitis. *PLoS One* **2016**, *11*, e0159719.

35. Frontini, M.; Imbriano, C.; Manni, I.; Mantovani, R. Cell-Cycle regulation of NF-YC nuclear localization. *Cell Cycle* **2004**, *3*, 205–210.

36. Jo, L.; Pelletier, J.M.; Harada, J.J. Central role of the LEAFY COTYLEDON1 transcription factor in seed development. *J. Integr. Plant Biol.* **2019**, *61*, 564–580.



37. Mu, J.; Tan, H.; Hong, S.; Liang, Y.; Zuo, J. Arabidopsis transcription factor genes NF-YA1, 5, 6, and 9 play redundant roles in male gametogenesis, embryogenesis, and seed development. *Mol. Plant* **2013**, *6*, 188–201.

38. Nodine, M.D.; Bartel, D.P. MicroRNAs prevent precocious gene expression and enable pattern formation during plant embryogenesis. *Genes Dev.* **2010**, *24*, 2678–2692.

39. Wang, F.; Perry, S.E. Identification of direct targets of FUSCA3, a key regulator of Arabidopsis seed development. *Plant Physiol.* **2013**, *161*, 1251–1264.

40. Xu, M.; Hu, T.; Zhao, J.; Park, M.-Y.; Earley, K.W.; Wu, G.; Yang, L.; Poethig, R.S. Developmental functions of miR156-regulated SQUAMOSA PROMOTER BINDING PROTEIN-LIKE (SPL) genes in *Arabidopsis thaliana*. *PLOS Genet.* **2016**, *12*, e1006263.

41. Zheng, Z.; Wang, N.; Jalajakumari, M.; Blackman, L.; Shen, E.; Verma, S.; Wang, M.-B.; Millar, A.A. miR159 represses a constitutive pathogen defense response in tobacco. *Plant Physiol.* **2020**, *182*, 2182–2198.

42. Tsuzuki, M.; Nishihama, R.; Ishizaki, K.; Kurihara, Y.; Matsui, M.; Bowman, J.L.; Kohchi, T.; Hamada, T.; Watanabe, Y. Profiling and characterization of small RNAs in the liverwort, *Marchantia polymorpha*, belonging to the first diverged land plants. *Plant Cell Physiol.* **2016**, *57*, 359–372.

43. Zhou, Y.; Qu, H.; Dibley, K.E.; Offler, C.E.; Patrick, J.W. A suite of sucrose transporters expressed in coats of developing legume seeds includes novel pH-independent facilitators. *Plant J.* **2007**, *49*, 750–764.

44. Coelho, C.M.M.; Benedito, V.A.; Figueira, A.; Vitorello, V.A.; Azevedo, R.A. Variation in the enzyme activity and gene expression of myo-inositol-3-phosphate synthase and phytate accumulation during seed

development in common bean (*Phaseolus vulgaris* L.). *Acta Physiol. Plant.* **2007**, *29*, 265–271.

45. Lin, S.-I.; Chiang, S.-F.; Lin, W.-Y.; Chen, J.-W.; Tseng, C.-Y.; Wu, P.-C.; Chiou, T.-J. Regulatory network of microRNA399 and PHO2 by systemic signaling. *Plant Physiol.* **2008**, *147*, 732–746.

46. Liang, G.; Yang, F.; Yu, D. MicroRNA395 mediates regulation of sulfate accumulation and allocation in *Arabidopsis thaliana*. *Plant J.* **2010**, *62*, 1046–1057.

47. Awazuhara, M.; Fujiwara, T.; Hayashi, H.; Watanabe-Takahashi, A.; Takahashi, H.; Saito, K. The function of SULTR2;1 sulfate transporter during seed development in *Arabidopsis thaliana*. *Physiol. Plant.* **2005**, *125*, 95–105.

48. Marsolais, F.; Pajak, A.; Yin, F.; Taylor, M.; Gabriel, M.; Merino, D.M.; Ma, V.; Kameka, A.; Vijayan, P.; Pham, H.; *et al*. Proteomic analysis of common bean seed with storage protein deficiency reveals up-regulation of sulfur-rich proteins and starch and raffinose metabolic enzymes, and down-regulation of the secretory pathway. *J. Proteomics* **2010**, *73*, 1587–1600.

49. Tian, R.; Wang, F.; Zheng, Q.; Niza, V.M.A.G.E.; Downie, A.B.; Perry, S.E. Direct and indirect targets of the arabidopsis seed transcription factor ABSCISIC ACID INSENSITIVE3. *Plant J.* **2020**, *103*, 1679–1694.

50. Changan, S.S.; Ali, K.; Kumar, V.; Garg, N.K.; Tyagi, A. Abscisic acid biosynthesis under water stress: anomalous behavior of the 9-cis-epoxycarotenoid dioxygenase1 (NCED1) gene in rice. *Biol. Plant.* **2018**, *62*, 663–670.

51. Jo, L.; Pelletier, J.M.; Hsu, S.W.; Baden, R.; Goldberg, R.B.; Harada, J.J. Combinatorial interactions of the LEC1 transcription factor specify diverse developmental programs during soybean seed development. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117*, 1223–1232.

52. Battaglia, M.; Olvera-Carrillo, Y.; Garciarrubio, A.; Campos, F.; Covarrubias, A.A. The Enigmatic LEA Proteins and Other Hydrophilins. *Plant Physiol.* **2008**, *148*, 6–24.

53. Hundertmark, M.; Buitink, J.; Leprince, O.; Hincha, D.K. The reduction of seed-specific dehydrins reduces seed longevity in *Arabidopsis thaliana*. *Seed Sci. Res.* **2011**, *21*, 165–173.

54. Martin, R.C.; Liu, P.-P.; Goloviznina, N. a; Nonogaki, H. microRNA, seeds, and Darwin?: diverse function of miRNA in seed biology and plant responses to stress. *J. Exp. Bot.* **2010**, *61*, 2229–2234.

55. Xu, M.Y.; Zhang, L.; Li, W.W.; Hu, X.L.; Wang, M.-B.; Fan, Y.L.; Zhang, C.Y.; Wang, L. Stress-induced early flowering is mediated by miR169 in *Arabidopsis thaliana*. *J. Exp. Bot.* **2014**, *65*, 89–101.

56. Stephenson, T.J.; McIntyre, C.L.; Collet, C.; Xue, G.-P. Genome-wide identification and expression analysis of the NF-Y family of transcription factors in *Triticum aestivum*. *Plant Mol. Biol.* **2007**, *65*, 77–92.

57. Yin, H.; Hong, G.; Li, L.; Zhang, X.; Kong, Y.; Sun, Z.; Li, J.; Chen, J.; He, Y. miR156/SPL9 regulates reactive oxygen species accumulation and immune response in *Arabidopsis thaliana*. *Phytopathology* **2019**, *109*, 632–642.

58. Rocak, S.; Linder, P. DEAD-box proteins: the driving forces behind RNA metabolism. *Nat. Rev. Mol. Cell Biol.* **2004**, *5*, 232–241.

59. Lorković, Z.J.; Herrmann, R.G.; Oelmüller, R. PRH75, a new nucleus-localized member of the DEAD-box protein family from higher plants. *Mol. Cell. Biol.* **1997**, *17*, 2257–2265.

60. Nayak, N.R.; Putnam, A.A.; Addepalli, B.; Lowenson, J.D.; Chen, T.; Jankowsky, E.; Perry, S.E.; Dinkins, R.D.; Limbach, P.A.; Clarke, S.G.; et al. An Arabidopsis ATP-dependent, DEAD-box RNA helicase loses activity upon IsoAsp formation but is restored by PROTEIN ISOASPARTYL METHYLTRANSFERASE. *Plant Cell* **2013**, *25*, 2573–2586.

61. Bai, B.; van der Horst, S.; Cordewener, J.H.G.; America, T.A.H.P.; Hanson, J.; Bentsink, L. Seed-stored mRNAs that Are specifically associated to monosomes are translationally regulated during germination. *Plant Physiol.* **2020**, *182*, 378–392.

# CHAPTER 5



# General Discussion, Conclusions, and Future Perspectives

**Author's contributions to the chapter:**

José Ricardo T. D. Parreira Salvado wrote this chapter based on the previous sections and the referred bibliography.

# General Discussion, Conclusions, and Future Perspectives

## 1. An overview



Seed development is one of the most relevant developmental processes in seed-producing crops. Seeds are essential to plants and, also essential as food/feed for animals, a source of nutrients, playing important agronomic and socio-economic roles. Still, seed development is a complex process that involves numerous molecular players and complex regulatory networks. As the final seed traits are influenced by both the genotype and adaptive changes to the environment [1], it is essential to identify the molecular players and regulatory networks acting during the developmental process.

*Phaseolus vulgaris*, the common bean, is a grain legume from the Fabaceae family that plays an important agronomical role due to its capacity for fixing atmospheric nitrogen and, nutritionally, being a good source of proteins, carbohydrates, dietary fibres, vitamins, and minerals (**Chapter I, Section 1.1.1**). While all the published reports have revealed an emerging role for common bean as one of the best agricultural model crops to study seed biology, preceding the onset of this PhD limited insights into the molecular mechanisms that shape *P. vulgaris* seed development were available (**Chapter I, Section 2.2**). There was no comprehensive description of the proteins and genes accumulated at the different developmental stages, or the acting molecular mechanisms and regulatory networks.

The broad goal of this PhD thesis was to describe the molecular and regulatory mechanisms underlying seed development in *P. vulgaris*. In order to achieve this goal several objectives were defined: 1) distinguish the timeframe and main seed developmental stages in which molecular analysis would be conducted; 2) identify the main molecular mechanisms and

metabolic pathways underlying seed development; 3) characterize the genome integrity maintenance players and mechanisms during this developmental process, and; 4) Understand the role of post-transcriptional regulatory mechanisms mediated by miRNAs in the developing seeds of *P. vulgaris*.



Herein, a comprehensive molecular analysis using distinct omics approaches, Proteomics, and Coding and Non-Coding Transcriptomics, was presented to unravel the molecular players and regulatory networks underlying seed development in the common bean. Indeed, the combined use of new high-throughput methodologies for transcriptome and proteome profiling allows the identification of a larger number of molecules, with higher specificity and accuracy, providing a deeper overview of the molecular complexities of a variety of organisms and developmental processes [2,3]. To accomplish this PhD work, a phenotypic characterisation was initially developed for the SER 16 line that allowed the identification of the time frame at which the molecular analysis would be conducted by harvesting seeds at: late embryogenesis/early filling (10 and 20 days after anthesis – DAA), late maturation (30 DAA), and desiccation (40 DAA).

The major outcomes achieved within this body of work are: 1) globally, new knowledge on *P. vulgaris* seed development and molecular resources that potentially impact the resulting seed traits; 2) the main molecular mechanisms and metabolic pathways, stage specific proteomic signatures and insights about the putative role of posttranslational modifications via a comprehensive description of protein abundance dynamics across developmental stages using gel-free high throughput proteomics methodology, Liquid Chromatography-tandem Mass Spectrometry (LC-MS/MS); 3) new information regarding DNA integrity maintenance, including DNA damage repair and chromatin remodelling mechanisms that ensure genome integrity in developing seeds via a qualitative description of gene expression dynamics across developmental

stages using high-throughput transcriptomics methodology, Massive Analysis of 3'-cDNA Ends (MACE), coupled with Digital PCR (dPCR); 4) new insights about the role of post-transcriptional regulation of gene expression acting during seed development via an exhaustive portrayal of the miRNAs expression dynamics, and their validated and predicted targets, at the main developmental stages using sRNA-Seq and degradome methodology. These outcomes will be further explored in the following sections.



## 2. Differential proteomics reveals the hallmarks of seed development in common bean (*Phaseolus vulgaris* L.)

In the second section of this thesis (**Chapter II**) we described the main molecular mechanisms and metabolic pathways, stage specific proteomic signatures and insights about the putative role of posttranslational modifications were provided by means of comprehensive description of protein abundance dynamics across developmental stages using gel-free high throughput proteomics methodology (LC-MS/MS). This comprehensive and systematic approach was necessary at the proteome level to create a 'proteome atlas' of the developing seed as well as pinpoint proteomic profiles for each developmental stage. Over 400 differentially accumulated proteins were identified, of which 255 were annotated. For the first time, the time frame of protein accumulation was described, highlighting the molecular processes occurring from late embryogenesis until the desiccated seed.

During early seed development, at 10 DAA, our results reflect an extensive metabolic activity. We observed an increased abundance of proteins categorised as "protein", "glycolysis", "TCA", "DNA", "RNA", "cell" and "stress" Mapman functional categories at late embryogenesis. Indeed, at this stage, there is still a high rate of cell division associated with embryo differentiation and morphogenesis, which demands a high synthesis of cell structural components [4–6]. Protein metabolism was the most represented

functional category, and several proteins appeared highly abundant at this time point. Curiously, an increased number and abundance of proteins related to degradation (proteolytic activity) was also observed, suggesting that high protein turnover can occur simultaneously with protein biosynthesis.

Ubiquitin and UBL (ubiquitin-like) modifiers are small proteins that covalently modify other proteins to alter their properties or behaviours. Ubiquitin-NEDD8, an UBL, has been implicated in the regulation of developmental processes such as embryogenesis [7,8]. NEDD8 can be conjugated with target proteins, such as the cullin subunits of cullin-RING E3 ubiquitin ligases inducing post-translational modifications [9]. In Arabidopsis, the combined knock-out of *RUB1* (*RELATED TO UBIQUITIN1*) and *RUB2*, two of the three NEDD8 coding genes in *Arabidopsis*, results in embryonic growth arrest at the two-cell stage [8]. In line with this observation, our results suggest that neddylation occurs at early seed development in *P. vulgaris*, possibly regulating several metabolic pathways. Indeed, we observed that UBIQUITIN-NEDD8-LIKE PROTEIN RUB2 decreased significantly in abundance from 10 to 20 DAA, while our STRING protein-protein interaction analysis showed an interaction between RUB2 with proteins related to glycolysis, TCA, protein synthesis and degradation, and stress.



Other highlights from the transition between late embryogenesis and early maturation include an increased accumulation of ALCOHOL DEHYDROGENASE 1, suggesting a role for alcoholic fermentation, for this developmental stage, besides the TCA/oxidative phosphorylation pathway. Another interesting aspect observed was the potential role of genome integrity maintenance mechanisms during zygotic embryo development. We have observed the PROLIFERATING CELL NUCLEAR ANTIGEN (PCNA), HISTONE H2A, and DNA DAMAGE-INDUCIBLE PROTEIN 1 with increased abundance at 10 DAA. These proteins might have a relevant role in the maintenance of genome integrity during the intensive cell division rate associated with this developmental stage. Knowledge of genome integrity

mechanisms, including DNA damage and repair or chromatin remodelling, was scarce for seed development. As such, our work brought pioneer data about how genome integrity is maintained during seed development in legumes.

During mid-seed development, at 20 and 30 DAA, our results correlate with the increased biomass accumulation observed phenotypically. Such fast continuous accumulation of biomass is characteristic of the seed filling stage, in which the synthesis of storage reserves occurs [4,5]. Indeed, we observed the accumulation of storage, signalling, and starch synthesis proteins associated with synthesis of storage compounds.



Several enzymes involved in starch biosynthesis were accumulated and maintained over the last seed development stages. This reflects the high requirement for carbon and energy for rapid cell growth and synthesis of storage compounds. SUCROSE SYNTHASE (SuSy) plays a central role in producing UDP-glucose (necessary for cell wall biosynthesis) and ADP-glucose (necessary for starch biosynthesis) [10]. In our results, a peak in the relative abundance of SuSy was found at 20 DAA, which may correlate with starch accumulation a characteristic of the seed filling stage. Also, storage proteins such as phaseolins and legumins increased in abundance from 10 to 20 DAA. Modulating the accumulation of storage proteins is an exciting and still open strategy to increase the nutritional value of pulses like common bean.

During the last developmental stage transition, our results provide clues over potential mechanisms activated for seed desiccation resistance. Indeed, an increase in proteins related to redox metabolism and, to a lesser extent, protein degradation/modification/folding and nucleic acids metabolism was seen later during seed development. The accumulation of proteasome components such as 26S PROTEASOME NON-ATPASE REGULATORY SUBUNIT 1 and proteases such as PROLINE

IMINOPEPTIDASE was observed at 40 DAA. This suggests that some proteases may be stored in the mature seed, possibly to be used on the degradation of other storage proteins during seed germination. Again, the increased HISTONE H2A abundance suggests it may exert a protective role on maintenance of genome integrity, possibly rearranging the chromatin structure to cope with reactive oxygen species (ROS) generated during seed drying [11].



Overall, the results included in chapter 2 provided not only a proteome atlas of the developing common bean seeds but also revealed the metabolic hallmarks of each seed's developmental stage. Furthermore, clues about the putative role of PTMs and new insights into the maintenance of genome integrity were unveiled. Taken together, these results constitute important knowledge and open research lines for the study of seed development in legumes.

## 3. Maintaining genome integrity during seed development in *Phaseolus vulgaris* L.: evidence from a transcriptomic profiling study

Maintenance of genome integrity is crucial in seeds due to the constant challenge of several endogenous and exogenous factors in the DNA. Clues about the mechanisms that contribute to maintain genome integrity during seed development were obtained on the proteomics study (**Chapter II**) but the knowledge was still limited. In the third section of this thesis (**Chapter III**) we provided new information regarding DNA integrity maintenance, including DNA damage repair and chromatin remodelling mechanisms that ensure genome integrity in developing seeds via a qualitative description of gene expression dynamics across developmental stages using high-throughput transcriptomics methodology (MACE and dPCR).

Of the 14,001 expressed genes, we focused on 301 genes categorised as “DNA” MapMan category. Overall, DNA damage sensing and the different repair mechanisms seem to be activated during early seed development. It is not surprising that a tight control of DNA integrity is needed, due to the high rate of cellular division and metabolic activity characteristic of this stage [5] (**Chapter II**). This can be observed by the higher expression of genes within the “DNA” category at 10 DAA, following the abundance profiles established for the proteins in the same functional category (**Chapter II**).



Insights regarding the DNA damage response acting during seed development were unveiled. Indeed, DNA double-strand breaks (DSB) sensing and signal transduction components were found to be more expressed in early seed development within our MACE results, further supported by dPCR. One of these components, the protein kinase *ATAXIA-TELANGIECTASIA MUTATED* (*ATM*) is a DNA damage-inducible protein kinase that phosphorylates a plethora of substrates participating in DNA damage response [12]. A crucial role in genome safeguarding during seed germination was previously reported for this kinase [13], suggesting a central role also during seed development. In agreement with this, other components of the DSB sensing mechanism, including *DNA REPAIR AND MEIOSIS PROTEIN* (*MRE11*), *DNA REPAIR-RECOMBINATION PROTEIN* (*RAD50*), and *NIJMEGEN BREAKAGE SYNDROME 1* (*NBS1*) were also found more expressed at early in seed development. Interestingly, upregulation of *SUPPRESSOR OF GAMMA RESPONSE 1* (*SOG1*), *NUCLEOSOME ASSEMBLY PROTEIN 1;2* (*NAP1;2*), *MRE11*, *NBS1,* and *RAD50* observed in the dPCR profiling at 30 DAA. From **Chapter II**, we found that seed fresh weight decreased from 30 DAA, an aspect associated with the end of the seed growth and the onset of dehydration. Reports have highlighted the high metabolic activity in developing seeds and drying of mature seeds result in the production of ROS [14,15] suggesting that upregulation of this mechanism can be a consequence of the production and accumulation of

ROS capable of genotoxic damage. Other components of the DNA damage response (DDR) were found more expressed at early seed development, including genes that are associated with Nucleotide Excision Repair (NER), mismatch repair mechanism (MMR), or even homologous recombination (HR).

Other interesting findings were described in this chapter. One of those findings is the similarity between expressed genes from the shoot apical meristem and root apical meristem and developing seeds analysed in this chapter. In rice, Kimura *et al*. [16] compared the expression patterns of DDR genes in proliferative tissues (SAM and RAM) and non-proliferative tissues as mature leaves. Among others, *OsPCNA*, *OsRPA32*, and *OsORC1* were found to be expressed in both proliferating tissues but not mature leaves. Interestingly, our results show a high expression of *PCNA1*, *ORC1A* and RPA subunits at 10 DAA, which suggesting a similarity in DNA damage responses between tissues where cell division and differentiation are actively occurring.



Another aspect revealed is the peak expression at 20 DAA for the *DNA-DAMAGE REPAIR/TOLERATION 100 PROTEIN* (*DRT100*), which matches the early filling and change in seed coat coloration. In *Arabidopsis*, the overexpression of grape *DRT100* (Vv*DRT100-L*) enhanced DNA repair under UV-B irradiation [17], suggesting a role of VvDRT100-L in the repair/reduction of abasic sites and SSBs possibly arising from oxidative stress. Interestingly, we have evidence that oxidative damage may occur during early filling and seed dehydration in *P. vulgaris*. A strong accumulation of the 1-CYS PEROXIREDOXIN enzyme from 10 to 20 DAA and, to less extent from 30 to 40 DAA, was previously observed (**Chapter II**). It is worth noting that 1-CYS PEROXIREDOXIN acts not only to relieve mild oxidative stresses but also as a molecular chaperone under severe stress conditions during seed germination and plant development [18].

Finally, our data show an increase in the expression of different NER genes from 30 to 40 DAA. Nucleotide excision repair has been described as the most versatile system for dealing with the DNA damage accumulated on desiccation and seed ageing [19]. It remains to be answered if this increase is triggered by the accumulation of DNA damage during this phase or if this is a protective mechanism by which the seed is being prepared to deal with this type of damages during early germination.



The role of chromatin and chromatin remodelling in DNA Damage Response (DDR) during seed development cannot be neglected. Numerous genes encoding for different HISTONE H2A proteins, GAMMA HISTONE VARIANT H2AX (G-H2A), H2B proteins, H4 and H3 were detected more expressed at 10 DAA. Noteworthy, members of the *HISTONE SUPERFAMILY PROTEIN*, with strong homology with predicted members of *H3* (*H3.2*) and *H4* family, increased in expression from 30 to 40 DAA. While it is difficult to propose a role associated with the accumulation of those transcripts during seed desiccation, the accumulation of *H3/H4* transcripts, and therefore proteins, may influence the condensation state of the chromatin.

Chromatin remodelers act in concert with DNA repair machinery to maintain seed genome stability, during germination [20]. Our results highlighted the *CHROMATIN REMODELING FACTOR CHD3* (*PKL*) with a decrease in expression from 10 to 20 DAA. This gene is involved in several key physiological processes such as floral transition mediated by *LEAFY* (*LFY*) expression in *Arabidopsis* [21]. PKL was also described as a repressor of the transcription factor (TF) LEAFY COTYLEDON1 (LEC1) upon embryogenesis [22,23]. LEC1 is part of the LEC1/ ABSCISIC ACID INSENSITIVE3 (ABI3), FUSCA3 (FUS3), and LEAFY COTYLEDON2 (LEC2) (LAFL) regulatory network that regulates seed filling, maturation, and storage compound synthesis [22,23]. Indeed, the *PKL* expression profile is opposite to the observed abundance of storage proteins during seed

development (**Chapter II**), such as LEGUMIN and PHASEOLIN, an aspect that is in agreement with the regulatory function proposed for PKL on the LEC1/AFL regulatory network [22,23].

We also observed high expression of *RUB1*, a gene encoding a ubiquitin-NEDD8-like protein at late embryogenesis (10 DAA). In **Chapter II**, we speculated that several metabolic pathways acting during early seed development could be regulated via neddylation, since the accumulation of RUB2 was high at 10 DAA. Evidence from a study conducted in human cells found that histone H4 was polyneddylated, in response to DNA damage, and NEDD8 was conjugated to the N-terminal lysine residues of H4 [24], further validating the need for more studies to understand the role of neddylation in modulating DDR responses during seed development.



Overall, this chapter provided insights into some mechanisms implicated in the of genome integrity and DNA damage response acting during seed development in *P. vulgaris*. Furthermore, these results support the observations made during **Chapter II**, while complementing and adding to those molecular resources that could be used to breed seeds with improved seed traits.

## 4. MicroRNAs expression dynamics reveal post-transcriptional mechanisms regulating seed development in *Phaseolus vulgaris L.*

In the fourth section of this thesis (**Chapter IV**), we provided new insights about the role of post-transcriptional regulation of gene expression acting during seed development via an exhaustive portrayal of the miRNAs expression dynamics at the main developmental stages using a high throughput non-coding transcriptomics methodology, sRNA-Seq. Additionally, a degradome analysis and a target prediction algorithm were

used to identify or predict, respectively, miRNA targets. Taken together, these results uncovered functional categories under potential miRNA regulation during this developmental process.

The previous sections (**Chapter II** and **III**) allowed a characterisation of the proteomic and transcriptomic landscapes during seed development, highlighting a time-frame of molecular and metabolic events, spanning from late embryogenesis to the desiccated seed. Still, molecular mechanisms controlling the observed temporal changes in gene expression remained to be understood. The work developed under the scope of this chapter aimed to address this gap of knowledge and elucidate the role of post-transcriptional regulation mediated by miRNAs on this process.



Post-transcriptional regulation mediated by miRNAs controls numerous developmental processes in plants, including seed development in legumes [25,26]. In *P. vulgaris*, a total of 72 known miRNAs, from 25 miRNA families, and 39 new miRNAs were identified and constitute the first comprehensive overview of miRNAs dynamics during seed development in this species. Based on their abundances, 10 and 40 DAA are the time points where the action of the miRNA seems more pronounced. These results are in line with the previous **Chapter II** and **III** where we have found a higher abundance of proteins and transcripts, suggesting these timepoints as crucial for seed development. Degradome analysis and target prediction found targets for 77 expressed miRNAs. The expression profiles of selected targets were established by mining the datasets produced in **Chapter III**.

Identified miRNAs seem to regulate developmental stage progression. A global repression of HOMEODOMAIN-LEUCINE ZIPPER (HD-ZIP) family of transcription factors (TFs) driven by the accumulation of MIR166 members is suggested. Indeed, MIR166 was the most expressed family and is possibly regulating *REV*, *PHB*, *CNA,* and *HB8* expression during all studied points. An increased abundance of HD-ZIPs TFs at early

seed development was expected since PHB is a direct activator of LEC2, which controls legume seed filling [27]. At 10 DAA, our MACE datasets evidenced that *PHB* abundances are already relatively low with residual *LEC2* abundances detected. This suggests that the seed filing program activation might have occurred earlier than 10 DAA, the first sampling point for MACE and sRNA-seq. Interestingly, high abundance for many MIR166 members along seed development was seen, but its biological relevance is still an open question.



MiR156-mediated repression *SPL10* and *SPL11* during early stages of embryogenesis, prevents the precocious expression of maturation genes [28]. At 10 DAA, the abundances of pvu-miR156i, *SPL10,* and other SPLs, as well as seed filing regulators LEC2 and FUS3 are almost residuals, suggesting that activation of the seed filing program occurred earlier as described for MIR166. The majority of MIR156 members accumulate at the later seed development suggesting global repression of SPL expression at 40 DAA. A MIR156 role in seed quiescence with implication on germination performance is suggested since the MIR156-SPL regulatory module has been implicated in the juvenile to the adult transition phase of vegetative development [29].

Another relevant miRNA is the miR169k that putatively represses *NF-A1* and *NF-A9* at 10 DAA. NF-YA1 and NF-YA9 repression [30] has been implicated in defining embryogenesis time course and our data supports a similar role in this regulatory module during *P. vulgaris* seed development.

A role for miRNAs in nutrient allocation during maturation was revealed. Indirect evidence that sucrose metabolism might be regulated at the posttranscriptional level in *P. vulgaris* seeds was obtained. The transporter *SUT1* , a sucrose/H+ symporter, is highly expressed in seed coat cells in this species and seems to be involved in sucrose efflux from the coat to the seed apoplasm [31]. Our results suggest that pvu-miR395e may repress *SUT1* expression at embryogenesis, being this repression released

when the seed enters filling. Indeed, our previous proteomic results (**Chapter II**) showed that the accumulation of proteins implicated in sugar metabolism, such as starch synthesis and cell wall-related proteins occur from 10 to 20 DAA.

MiRNAs implicated in phosphorus (P) or sulphur (S) signalling and homeostasis were identified in our seed samples. In *P. vulgaris* seeds, phytic acid is the main stored form of P, which accumulates concomitantly with dry matter accumulation [32]. While miR399-mediated *PHO2* down-regulation has been implicated with P uptake on roots and translocation to shoots [33], little is known about the MIR399 role in seeds. Our work showed that an increase in MIR399 abundance until 30 DAA, with significant changes observed for pvu-miR399i from 10 to 20 DAA. Such abundance profiles are in agreement with the stage transition to seed filling observed (**Chapter II**). These observations, together with the low *PHO2* expression levels observed, suggest a regulatory role for MIR399 to facilitate P translocation to the maturing seeds.

MiR395 is involved in S assimilation and allocation regulation by targeting *APS* genes and *SULTR2;1*, respectively [34]. We observed a relatively low expression of *SULTR2;1* in developing *P. vulgaris* seeds contrasting with strong MIR395 accumulation. Our results also highlight MIR395-mediated repression of *APS1* at 10 DAA, suggesting a role of this regulatory module on S accumulation between the end of embryogenesis and early filling. In *P. vulgaris* seeds, the contents of sulphur amino are sub-optimal for human nutrition (**Chapter I**). The modulation of *SULTR2;1* expression in seeds by pvu-miR395 accumulation could be a promissing approach to tune sulphate flux and enhance the synthesis of S-containing storage proteins.

Seed dormancy and longevity seem to be regulated at the posttranscriptional level. Auxins regulate Arabidopsis seed dormancy via a

concerted action with ABA signalling pathway [35]. In this report, the authors suggest a feedback loop where the ABI3-mediated repression of *MIR160B* gene, which targets *ARF10* and *ARF16,* allows the establishment of dormancy. In *P. vulgaris* seeds, the results obtained are puzzling since the putative MIR160-mediated repression of *ARF10*, *ARF16,* and *ARF17* was suggested across all seed development. Also, ABI3 expression levels were found to increase at 30 DAA, when the seed starts to dehydrate and becomes dormant. Altogether, this supports that other regulatory mechanisms of gene expression might be implicated in seed dormancy.



It has been recently shown that LEAs can be direct targets of ABI3, such as *EM1* [35]. EM1 and RAB18 belong to LEA family and were found accumulated in *P. vulgaris* seeds (**Chapter II**). Our degradome data validated *EM1* and *RAB18* targets of the new miRNAs miR_6 and miR_18, respectively. The high miRNA abundance seen at 10 DAA contrasts target transcript and protein accumulation profiles**.** In orthodox seeds, desiccation tolerance is linked to seed longevity, and EM1 role in this has been highlighted [35]. Reduced *RAB18* abundance was implicated in a decrease in seed longevity [36].

Mechanisms implicated in seed viability and posterior germination traits seem to be regulated at the posttranscriptional level, which includes mechanisms to cope with oxidative damage. The MIR156/SPL9 regulatory module was implicated in the regulation of ROS accumulation, since miR156 overexpression mutants had ROS accumulation suppressed [37]. During seed filing and dehydration of *P. vulgaris* seeds, we evidenced the putative occurrence of oxidative damage (**Chapter II**) and its effects on the maintenance of embryo genome integrity (**Chapter III**). Members of the MIR156 increase expression throughout seed development. For pvu-miR156i, this contrasts with the lower abundances of its predicted target, SPL9. This suggests that MIR156/SPL9 regulatory module can contribute to

tune the ROS levels during seed development and indirectly contribute to the maintenance of genome integrity.

Overall, this chapter presented the first comprehensive analysis of miRNAs and their predicted and validated targets, during seed development in *P. vulgaris*. These results reveal putative points of posttranscriptional regulation of gene expression, with impact on seed developmental stage transition, desiccation tolerance, longevity, and other biologically relevant processes. These are novel resources with potential application in future biotechnological approaches to modulate gene expression implicated in legume seed traits, including storage compound accumulation or seed viability.



## 5. Concluding remarks and future perspectives

Despite the substantial molecular profiling data produced in **Chapters II**, **III,** and **IV**, which used four timepoints related to the main seed developmental stages, the research presented here still lacks functional characterisation to unravel the precise roles of the proteins, transcripts, and miRNAs. This is due to the reduced number of tools for functional gene validation, particularly tissue culture of *P. vulgaris* that has repeatedly been found to be difficult, still hampering the application of genetic engineering studies (**Chapter I**). These studies are relevant to assess the potential for the definition the molecular targets for future molecular breeding applications to engineer improved seed traits in the common bean.

A molecular characterization with focus on prior timepoints, then those addressed in this study, can be performed to understand the histodifferentiation of the embryo and the time frame of the molecular events regulating the transition to seed maturation. Neddylation, or NEDD8-modification, is a post-translational modification that seems to be occurring

during early seed development in *P. vulgaris*. Our results evidenced a putative interaction with proteins related with several metabolisms. It can be interesting to understand its impact during the early stage of development. Also, the LAFL regulatory network has been described as promoting embryogenesis and seed maturation, among other processes (**Chapter I**). Our results suggest that this network is being regulated in developing common bean seeds. Understanding the protein and transcript dynamics of the members of the LAFL network, and their regulators, can reveal the timeframe of action and putative role in this species.



It has been shown that mutant common bean lines with decreased phaseolin and major lectins, show an increased sulphur amino acid content [38]. During seed maturation, it would be interesting to have a deeper understanding on the molecular mechanisms of storage protein and aminoacids synthesis and accumulation, since it can impact the seeds quality and nutritional content, potentially leading to a more balanced amino acid composition. Our results have evidenced this potentiality, but also evidenced the need to conduct on the earlier stages of seed development, to understand the molecular shift from embryogenesis to seed filling. Another interesting perspective would be to deepen the study of the later stages of seed development, revealing which molecular mechanisms impact seed viability during desiccation. Modulation of the expression of miRNAs involved in desiccation tolerance or seed quiescence is still an open research line, as these parameters not only impact seed viability but also agronomically important traits like germination and vigour. Indeed, in **Chapter III** we have explored, for the first time in common bean, genome integrity maintenance mechanisms during seed development. It would be interesting to investigate how these mechanisms are modulated during germination. Although our study was a focused analysis on DNA Mercator functional category, it could be of relevance to address other functional categories, such as Stress and Redox to identify genes involved the protection mechanisms used by the seeds at early development or during desiccation. In **Chapter IV** we have

observed a high abundance of 24 nucleotide reads in our samples. The 24 nt Pol IV-dependent siRNAs transcriptionally silence transposons in RNA-directed DNA methylation (RdDM) process [39]. This process has been shown to impact seed development *Brassica rapa* [39] and could be interesting to study this aspect in *P. vulgaris* developing seeds. Also, the genotype used in this thesis was SER 16, which is described as a drought-tolerant genotype with high remobilisation of photosynthates to the seed (**Chapter I**). It would be interesting to compare our molecular results with other alternative genetic backgrounds and under other relevant conditions, such as drought or heat stress.



Throughout this body of work, an integrative omics approach was used to describe the molecular and regulatory mechanisms underlying seed development in *P. vulgaris*. Together, these studies on the dynamics of proteins, as well as transcriptomes of the developing seeds, produced a large repository of molecular information that contributes to close the existing gap of knowledge on seed development, especially on this species. The herein identified proteins, transcripts and miRNAs were associated with, or implicated in, the regulation of several biological processes, such as nutrient allocation, which impact the final seeds' quality traits. Other identified molecules can have a potential role on seed viability and germination. Not only these novel molecular resources can be extended to other legume species and shed light in new lines of research, but they can also be used in new biotechnology-based approaches to improve legume seeds and meet the current and future food security challenges.

## 6. Bibliography

1. Mondoni, A.; Orsenigo, S.; Donà, M.; Balestrazzi, A.; Probert, R.J.; Hay, F.R.; Petraglia, A.; Abeli, T. Environmentally induced transgenerational changes in seed longevity: maternal and genetic influence. *Ann. Bot.* **2014**, *113*, 1257–1263.

2. Rani, B.; Sharma, V.K. Transcriptome profiling: methods and applications- A review. *Agric. Rev.* **2017**, *38*.

3. Aslam, B.; Basit, M.; Nisar, M.A.; Khurshid, M.; Rasool, M.H. Proteomics: Technologies and their applications. *J. Chromatogr. Sci.* **2017**, *55*, 182–196.

4. Wang, W.-Q.; Liu, S.-J.; Song, S.-Q.; Møller, I.M. Proteomics of seed development, desiccation tolerance, germination and vigor. *Plant Physiol. Biochem.* **2015**, *86*, 1–15.



5. Coelho, C.; Benedito, V.; Maria, C.; Coelho, M.; Augusto, V.; Coelho, C.; Benedito, V.; Maria, C.; Coelho, M.; Augusto, V.; *et al*. Seed development and reserve compound accumulation in common bean (*Phaseolus vulgaris* L.). *Seed Sci. Biotechnol* **2008**, *2*, 42–52.

6. Gallardo, K.; Le Signor, C.; Vandekerckhove, J.; Thompson, R.D.; Burstin, J. Proteomics of *Medicago truncatula* seed development establishes the time frame of diverse metabolic processes related to reserve accumulation. *Plant Physiol.* **2003**, *133*, 664–682.

7. Mergner, J.; Schwechheimer, C. The NEDD8 modification pathway in plants. *Front. Plant Sci.* **2014**, *5*, 103.

8. Bostick, M. Related to Ubiquitin 1 and 2 are redundant and essential and regulate vegetative growth, auxin signaling, and ethylene production in Arabidopsis. *Plant Cell Online* **2004**, *16*, 2418–2432.

9. Mergner, J.; Heinzlmeir, S.; Kuster, B.; Schwechheimer, C. DENEDDYLASE1 deconjugates NEDD8 from Non-cullin protein substrates in *Arabidopsis thaliana*. *Plant Cell* **2015**, *27*, 741–753.

10. Sturm, A.; Tang, G.-Q. The sucrose-cleaving enzymes of plants are crucial for development, growth and carbon partitioning. *Trends Plant Sci.* **1999**, *4*, 401–407.

11. Moore, J.D.; Yazgan, O.; Ataian, Y.; Krebs, J.E. Diverse roles for histone H2A modifications in DNA damage response pathways in yeast. *Genetics* **2007**, *176*, 15–25.

12. Hu, Z.; Cools, T.; De Veylder, L. Mechanisms used by plants to cope with DNA Damage. *Annu. Rev. Plant Biol.* **2016**, *67*, 439–462.

13. Waterworth, W.M.; Footitt, S.; Bray, C.M.; Finch-Savage, W.E.; West, C.E. DNA damage checkpoint kinase ATM regulates germination and maintains genome stability in seeds. *Proc. Natl. Acad. Sci.* **2016**, *113*, 9647–9652.

14. Bailly, C.; El-Maarouf-Bouteau, H.; Corbineau, F. From intracellular

signaling networks to cell death: the dual role of reactive oxygen species in seed physiology. *C. R. Biol.* **2008**, *331*, 806–814.

15. Naithani, S.; Nonogaki, H.; Jaiswal, P. Exploring crossroads between seed development and stress response. In *Mechanism of Plant Hormone Signaling under Stress*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2017; Vol. 2, pp. 415–454.

16. Kimura, S.; Tahira, Y.; Ishibashi, T.; Mori, Y.; Mori, T.; Hashimoto, J.; Sakaguchi, K. DNA repair in higher plants; photoreactivation is the major DNA repair pathway in non-proliferating cells while excision repair (nucleotide excision repair and base excision repair) is active in proliferating cells. *Nucleic Acids Res.* **2004**, *32*, 2760–2767.

17. Fujimori, N.; Suzuki, N.; Nakajima, Y.; Suzuki, S. Plant DNA-damage repair/toleration 100 protein repairs UV-B-induced DNA damage. *DNA Repair (Amst).* **2014**, *21*, 171–176.

18. Kim, S.Y.; Paeng, S.K.; Nawkar, G.M.; Maibam, P.; Lee, E.S.; Kim, K.-S.S.; Lee, D.H.; Park, D.-J.J.; Kang, S. Bin; Kim, M.R.; et al. The 1-Cys peroxiredoxin, a regulator of seed dormancy, functions as a molecular chaperone under oxidative stress conditions. *Plant Sci.* **2011**, *181*, 119–124.

19. Boubriak, I.; McCready, S.; Osborne, D.J. DNA structure and seed desiccation tolerance. In *Plant Desiccation Tolerance*; Blackwell Publishing Ltd: Oxford, UK, 2000; pp. 215–249.

20. Balestrazzi, A.; Donà, M.; Macovei, A.; Sabatini, M.E.; Pagano, A.; Carbonera, D. DNA repair and telomere maintenance during seed imbibition: correlation of transcriptional patterns. *Telomere and Telomerase* **2015**, 2–5.

21. Fu, X.; Li, C.; Liang, Q.; Zhou, Y.; He, H.; Fan, L.-M. CHD3 chromatin-remodeling factor PICKLE regulates floral transition partially via modulating LEAFY expression at the chromatin level in Arabidopsis. *Sci. China Life Sci.* **2016**, *59*, 516–528.

22. Verdier, J.; Thompson, R.D. Transcriptional regulation of storage protein synthesis during dicotyledon seed filling. *Plant Cell Physiol.* **2008**, *49*, 1263–1271.

23. Zhai, L.; Xu, L.; Wang, Y.; Zhu, X.; Feng, H.; Li, C.; Luo, X.; Everlyne, M.M.; Liu, L. Transcriptional identification and characterization of differentially expressed genes associated with embryogenesis in radish (*Raphanus sativus* L.). *Sci. Rep.* **2016**, *6*, 21652.

24. Ma, T.; Chen, Y.; Zhang, F.; Yang, C.-Y.Y.; Wang, S.; Yu, X. RNF111-

dependent neddylation activates DNA damage-induced ubiquitination. *Mol. Cell* **2013**, *49*, 897–907.

25. Yu, L.; Guo, R.; Jiang, Y.; Ye, X.; Yang, Z.; Meng, Y.; Shao, C. Genome-wide identification and characterization of novel microRNAs in seed development of soybean. *Biosci. Biotechnol. Biochem.* **2019**, *83*, 233–242.

26. DeBoer, K.; Melser, S.; Sperschneider, J.; Kamphuis, L.G.; Garg, G.; Gao, L.; Frick, K.; Singh, K.B. Identification and profiling of narrow-leafed lupin (*Lupinus angustifolius*) microRNAs during seed development. *BMC Genomics* **2019**, *20*, 135.

27. Manan, S.; Ahmad, M.Z.; Zhang, G.; Chen, B.; Haq, B.U.; Yang, J.; Zhao, J. Soybean LEC2 regulates subsets of genes involved in controlling the biosynthesis and catabolism of seed storage substances and seed development. *Front. Plant Sci.* **2017**, *8*, 1–16.

28. Nodine, M.D.; Bartel, D.P. MicroRNAs prevent precocious gene expression and enable pattern formation during plant embryogenesis. *Genes Dev.* **2010**, *24*, 2678–2692.

29. Xu, M.; Hu, T.; Zhao, J.; Park, M.-Y.; Earley, K.W.; Wu, G.; Yang, L.; Poethig, R.S. Developmental functions of miR156-regulated SQUAMOSA PROMOTER BINDING PROTEIN-LIKE (SPL) genes in *Arabidopsis thaliana*. *PLOS Genet.* **2016**, *12*, e1006263.

30. Mu, J.; Tan, H.; Hong, S.; Liang, Y.; Zuo, J. Arabidopsis transcription factor genes NF-YA1, 5, 6, and 9 play redundant roles in male gametogenesis, embryogenesis, and seed development. *Mol. Plant* **2013**, *6*, 188–201.

31. Zhou, Y.; Qu, H.; Dibley, K.E.; Offler, C.E.; Patrick, J.W. A suite of sucrose transporters expressed in coats of developing legume seeds includes novel pH-independent facilitators. *Plant J.* **2007**, *49*, 750–764.

32. Coelho, C.M.M.; Benedito, V.A.; Figueira, A.; Vitorello, V.A.; Azevedo, R.A. Variation in the enzyme activity and gene expression of myo-inositol-3-phosphate synthase and phytate accumulation during seed development in common bean (*Phaseolus vulgaris* L.). *Acta Physiol. Plant.* **2007**, *29*, 265–271.

33. Lin, S.-I.; Chiang, S.-F.; Lin, W.-Y.; Chen, J.-W.; Tseng, C.-Y.; Wu, P.-C.; Chiou, T.-J. Regulatory network of microRNA399 and PHO2 by systemic signaling. *Plant Physiol.* **2008**, *147*, 732–746.

34. Liang, G.; Yang, F.; Yu, D. MicroRNA395 mediates regulation of sulfate accumulation and allocation in *Arabidopsis thaliana*. *Plant J.*

**2010**, *62*, 1046–1057.

35. Tian, R.; Wang, F.; Zheng, Q.; Affri Gnago Epse Niza, V.M.; Bruce Downie, A.; Perry, S.E.; Niza, V.M.A.G.E.; Downie, A.B.; Perry, S.E. Direct and indirect targets of the Arabidopsis seed transcription factor ABSCISIC ACID INSENSITIVE3. *Plant J.* **2020**, *103*, tpj.14854.

36. Hundertmark, M.; Buitink, J.; Leprince, O.; Hincha, D.K. The reduction of seed-specific dehydrins reduces seed longevity in *Arabidopsis thaliana*. *Seed Sci. Res.* **2011**, *21*, 165–173.



37. Yin, H.; Hong, G.; Li, L.; Zhang, X.; Kong, Y.; Sun, Z.; Li, J.; Chen, J.; He, Y. miR156/SPL9 regulates reactive oxygen species accumulation and immune response in *Arabidopsis thaliana*. *Phytopathology* **2019**, *109*, 632–642.

38. Marsolais, F.; Pajak, A.; Yin, F.; Taylor, M.; Gabriel, M.; Merino, D.M.; Ma, V.; Kameka, A.; Vijayan, P.; Pham, H.; *et al*. Proteomic analysis of common bean seed with storage protein deficiency reveals up-regulation of sulfur-rich proteins and starch and raffinose metabolic enzymes, and down-regulation of the secretory pathway. *J. Proteomics* **2010**, *73*, 1587–1600.

39. Grover, J.W.; Kendall, T.; Baten, A.; Burgess, D.; Freeling, M.; King, G.J.; Mosher, R.A. Maternal components of RNA-directed DNA methylation are required for seed development in *Brassica rapa*. *Plant J.* **2018**, *94*, 575–582.

# Funding Acknowledgement

This work was supported by Fundação para a Ciência e a Tecnologia (FCT), with funding from Fundo Social Europeu and national funding from MEC, via a PhD Grant (PD/BD/113474/2015) awarded to José Ricardo Torrão Dias Parreira Salvado. The Plants for Life PhD Program (PD/00035/2013), research unit “GREEN-IT” (UID/Multi/04551/2020), and research projects PTDC/AGR-GPL/110224/2009 and PRIME-XS 262067 are also acknowledged.



**ITQB NOVA, Avenida da República, 2780-157 Oeiras, Portugal** | Tel +351 214 469 100 | itqb.unl.pt