Universidade do Minho
Escola de Ciências

Tomás Christopher Abreu Werner

The unveiling of the ancestral function of *MYB-LIKE* genes in *Marchantia polymorpha*

The unveiling of the ancestral function of *MYB-LIKE* genes in *Marchantia polymorpha*

Tomás Christopher Abreu Werner

UMinho | 2022

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR
COMPETE
QREN QUADRO DE REFERÊNCIA ESTRATÉGICO NACIONAL
UNIÃO EUROPEIA
Fundo Europeu de Desenvolvimento Regional

fevereiro de 2022

Universidade do Minho
Escola de Ciências

Tomás Christopher Abreu Werner

The unveiling of the ancestral function of *MYB-LIKE* genes in *Marchantia polymorpha*

Dissertação de Mestrado
Mestrado em Genética Molecular

Trabalho efetuado sob a orientação de
Professora Doutora Maria Manuela Ribeiro Costa
Doutor Rómulo Sacramento Sobral

fevereiro 2022

Acknowledgements

It's been a complicated year. I've learnt a great deal about what I'd like to do with my life and who I am. And because of that I wouldn't change this year, and I especially would not change the people who've helped me throughout, to whom I dedicate these words.

First, but not least, to Professor Manuela for teaching me all she could, not giving up on me despite all the antics and believing in me. It was a privilege and a pleasure to work with you.

To my main man Rómulo, for answering every single question I had with a patience and wisdom beyond your years. Thank you for humouring my early morning hypothetical quandaries and for unwittingly helping me understand what I want from life.

To Sara, my mini mentor and companion. Thank you for putting up with all my peculiarities, for being a shoulder to gossip on when waiting for results and for pushing me to work harder. Your effect on me and on this thesis is immeasurable and I haven't enough paragraphs to thank you.

To Ana "Astúrias" Maria for her stellar performance as someone who doesn't like me. If I weren't as astute, I would've missed the subtle caring you bury. Thank you for being a friend.

To camarada Alhinho, my first point of contact in the lab. Thank you for getting me settled, showing me the ropes, and discussing jazz and high fantasy, among other varied cultural subjects.

To Ana Marques. Although we've known each other for years, this felt like the first time we truly met. Thank you for bringing joy to the lab and never failing to put a smile on our faces.

To everyone in the lab and department, for accommodating me in your labs, technical support and helping me clean my messes. An honourable mention to Angélica, Sr. Luís and Inês.

To Patrícia. I wouldn't be here without you. I have never known such a caring and hilarious person. The mere thought of you would've helped but to have you there by my side was more than I could've ever asked for. Thank you for making me laugh when I needed to and for everything.

To my parents for making me who I am and for, despite not entirely understanding what I was doing in the lab or what I'm doing with my life, supporting me, and believing in me every step of the way. I love you.

Finally, to all friends, acquaintances, and pedestrians that have touched my life, thank you.

This work fell under the project PTDC/BIA-PLA/1402/2014, "EvoMod- Origin and Evolutionary establishment of a transcriptional module controlling flower asymmetry" and was supported by UID/MULTI/04046/2013 centre grant from FCT, Portugal (to BioISI).

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

Resumo

Título: "A revelação da função ancestral de genes *MYB-LIKE* em *Marchantia polymorpha*"

Redes de regulação genética são mecanismos essenciais e complexos que controlam vários sistemas em organismos e consistem de interações entre genes e produtos de genes. O módulo regulatório DDR foi descrito pela primeira vez em *Antirrhinum majus* como parte do mecanismo molecular que determina a assimetria floral da planta. Os genes *MYB-LIKE* que compõem o módulo, *DIV*, *DRIF* e *RAD*, estão envolvidos em interações DNA-proteína e proteína-proteína e foram identificados em várias espécies de plantas. Investigação com plantas superiores é complicada pelo número elevado de cópias de genes e pela complexidade de interações. Plantas ancestrais sofreram menos duplicações genómicas ao longo da evolução e então são seres mais simples de estudar, como a hepática *Marchantia polymorpha* que tem recentemente ressurgido como uma espécie modelo. Genes homólogos de *DIV* e *DRIF* foram encontrados em *M. polymorpha* e a planta foi escolhida como modelo para estudar a função ancestral e a evolução das proteínas DIV e DRIF.

Nesta dissertação, foram analisados os fenótipos de plantas *M. polymorpha* mutantes e com sobre expressão de genes do DDR e as funções dos genes *DIV* e *DRIF* em *M. polymorpha* ficaram mais esclarecidas. Ao nível molecular, um protocolo foi desenvolvido e otimizado para a expressão heteróloga das proteínas MpDIV1, MpDIV2 e MpDRIF para uso em estudos de interação. Análise filogenética foi utilizada como forma de revelar a evolução de *DIV* e *DRIF* em diferentes espécies de alga e foi revelada uma nova potencial história para a evolução primitiva dos genes. Destas formas e com a preparação de construções de sobre expressão, o caminho foi preparado para investigação futura focada em desvendar a função ancestral dos genes *MYB-LIKE* e poderá levar a uma melhor compreensão de como redes de regulação genética podem funcionar e evoluir.

Palavras-chave: Evolução; Função ancestral; *Marchantia polymorpha*; Módulo regulatório DDR; *MYB-LIKE*.

Abstract

Title: "The unveiling of the ancestral function of *MYB-LIKE* genes in *Marchantia polymorpha*"

Gene regulatory networks are complex and essential mechanisms that control various systems in organisms and consist of interactions between genes and gene products. The DDR regulatory module was first described in *Antirrhinum majus* as a part of the molecular mechanism that determines flower asymmetry. The *MYB-LIKE* genes that compose the module, *DIV*, *DRIF* and *RAD*, are involved in DNA-protein and protein-protein interactions and have been identified in many different species of plants. Studies in higher plants are complicated by the elevated number of gene copies and by the complexity of interactions. Ancestral plants have suffered less genome duplications and are thus much simpler organisms to study, such as the basal liverwort *Marchantia polymorpha* which has recently remerged as a model organism. Homologs of *DIV* and *DRIF* have been found in *M. polymorpha* and the species was chosen as a model in which to study the ancestral function and the evolution of the DIV and DRIF proteins.

In this thesis, *M. polymorpha* plant phenotypes for mutants and plants with overexpression of DDR genes were analysed and the functions of the *DIV* and *DRIF* genes of *M. polymorpha* were further clarified. At the molecular level, a protocol for the heterologous expression of the MpDIV1, MpDIV2 and MpDRIF proteins was developed and optimised for use in interaction studies. Phylogenetic analysis was employed to uncover the evolution of *DIV* and *DRIF* throughout different algal species and findings revealed a new potential story for the early evolution of these genes. In these ways and with the preparation of overexpression constructs, the way was paved for future studies into unveiling the ancestral function of the *MYB-LIKE* genes and could lead to a greater understanding of how gene regulatory networks could function and evolve.

Keywords: Ancestral function; DDR regulatory module; Evolution; *Marchantia polymorpha*; *MYB-LIKE*.

Table of contents

List of abbreviations and acronyms

A...	Absorbance at *** nanometers
APS	Ammonium persulfate
At*RAD*	*Arabidopsis thaliana RADIALIS*
bp	Base pairs
Cas9	CRISP-associated protein 9
cDNA	Complementary DNA
CLE	*CLAVATA3/EMBRYO SURROUNDING REGION-RELATED*
CRISPR	Clustered regularly interspaced short palindromic repeats
CTAB	Cetrimonium bromide
CYC	*CYCLOIDEA*
DDR	DIV-DRIF-RAD
DICH	*DICHOTOMA*
DIV	*DIVARICATA*
DRIF	*DIV-and-RAD-Interacting Factor*
DTT	Dithiothreitol
DUF	Domain of unknown function
EAR	Ethylene-responsive element binding factor-associated amphiphilic repression
EDTA	Ethylenediaminetetraacetic acid
EGT	Endosymbiotic gene transfer
EMSA	Electrophoretic mobility shift assay
EtOH	Ethanol
g	Relative centrifuge force
gDNA	Genomic DNA
GR	Glucocorticoid receptor
GST	Glutathione S-transferase
gRNA	Guide RNA
GTE	Glucose-Tris-EDTA buffer

h	Hour
H_2O_{up}	Ultra purified water
HGT	Horizontal gene transfer
IPTG	Isopropyl β-D-1-thiogalactopyranoside
kDa	Kilodalton
LB	Lysogeny broth
min	Minute
Mp*ARF*	*Marchantia polymorpha AUXIN RESPONSE FACTOR*
Mp*CLE*	*Marchantia polymorpha CLAVATA3/EMBRYO SURROUNDING REGION-RELATED*
Mp*CLV*	*Marchantia polymorpha CLAVATA*
Mp*EF1α*	*Marchantia polymorpha ELONGATION FACTOR1α*
Mp*GCAM1*	*Marchantia polymorpha GEMMA CUP-ASSOCIATED MYB*
mRNA	Messenger RNA
MW	Molecular weight
OD…	Optical density at *** nanometers
ON	Overnight
PCR	Polymerase chain reaction
PEG	Polyethylene glycol
RAD	*RADIALIS*
RNase	Ribonuclease
rpm	Revolutions per minute
RT-qPCR	Real-time quantitative polymerase chain reaction
SDS	Sodium dodecyl sulphate
TAE	Tris-acetate-EDTA buffer
TE	Tris-EDTA buffer
TEMED	Tetramethylethylenediamine
TEN	Tris-EDTA-NaCl buffer
UV	Ultraviolet
WT	Wild type

1. Introduction

Life is complex. Throughout time, its complexity has increased as life has evolved ever increasingly complex functions, many of which are regulated by intricate networks of genes. These regulatory networks have evolved in a myriad of ways and involve many different types of interactions. It is important to study and understand how they have come to be, because of how crucial they are in the regulation of the multitude of processes that shape life as we know it.

1.1. Gene Regulatory Networks

Gene regulatory networks consist of complex interactions between genes that regulate a variety of processes. These interactions can occur between proteins and proteins and between DNA and proteins and they manage many processes in organisms. Transcription factors play a big part in these networks as interconnecting factors due to their role in the regulation of gene expression. The duplication of transcription factors and the repurposing of these through mutation is one of the ways that the expansion of existing networks and the development of new networks is promoted (Voordeckers et al., 2015).

1.1. DDR regulatory module

A regulatory module composed of transcription factors that is thought to have originated by duplication and alteration of genes is the DDR regulatory module (Raimundo et al., 2018). This module consists of a plant specific MYB type family of proteins. It has been studied in different plant species such as *Arabidopsis thaliana, Solanum lycopersicum* (Machemer et al., 2011) and *Populus trichocarpa* (Petzold et al., 2018). This module is composed by genes known as *DIVARICATA* (*DIV*), *RADIALIS* (*RAD*) and *DIV-and-RAD-Interacting-Factor* (*DRIF*).

The interaction between these genes was described in *Antirrhinum majus*, found to be part of the regulatory mechanism for determining flower asymmetry (Corley et al., 2005; Raimundo et al., 2013). *A. majus* flowers are dorsoventrally asymmetric and have five petals with morphological differences between the two dorsal, the two lateral and the ventral petal (**Figure 1**). The ventral morphology of flowers was found to be promoted by *DIV*, a gene that codes for a MYB family transcription factor, with two different MYB domains. *DIV* is however expressed in the entire floral meristem (Galego & Almeida, 2002), indicating that DIV activity must be repressed in the dorsal

region of the meristem. The genes *CYCLOIDEA* (*CYC*) and *DICHOTOMA* (*DICH*) are not part of this module, but they interact with it in the *A. majus* flower. They encode for transcription factors that belong to the TCP family of genes and their expression promotes dorsal morphology (Luo et al., 1996; Luo et al., 1999) by post-transcriptionally repressing the action of DIV (Galego & Almeida, 2002).

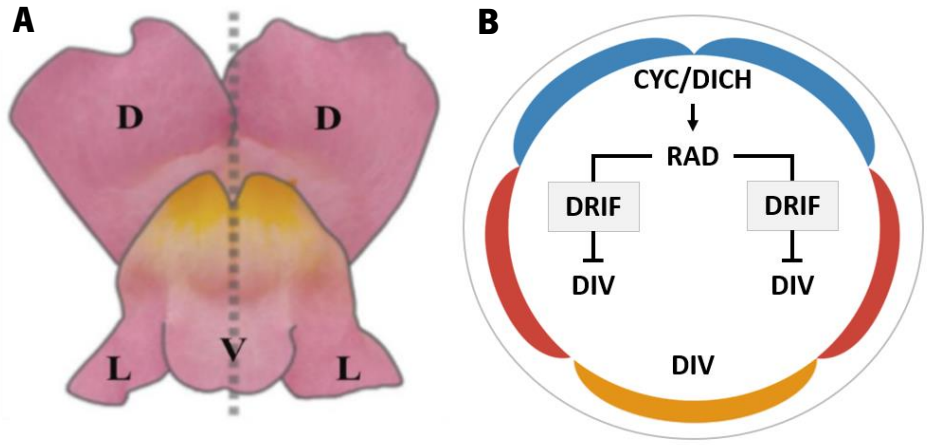

Figure 1. *Antirrhinum majus* flower with schematic model for the genetic interactions that determine flower asymmetry. A - *A. majus* flowers showing bilateral symmetry (axis represented by dotted line) and the five petals, the two dorsal petals (D), the two lateral petals (L) and the ventral petal (V). **B** - Diagram of *A. majus* flower that shows that *CYC* and *DICH* are expressed in the dorsal domain (blue) and promote expression of *RAD* that binds to DRIF in the dorsal and lateral domain (red), inhibiting DIV activity, which is expressed in the entire flower. In the ventral domain (yellow) RAD is not present and DIV functions normally (Adapted from Corley et al., 2005).

CYC and *DICH* are expressed in the dorsal part of the floral meristem and are known to promote the expression of *RAD*, a gene that codes for a protein with a single MYB domain. RAD inhibits the activity of DIV in the dorsal and lateral regions of the flower (Corley et al., 2005) (**Figure 1**), leading to dorsoventral asymmetric flowers. It was found that RAD and DIV compete to bind to another MYB protein, DRIF (Raimundo et al., 2013; Machemer et al., 2011). In the dorsal region of the flower, RAD and DRIF bind to each other, inhibiting DIV activity, which is dependent on forming a complex with DRIF (**Figure 2**). These three proteins and their interactions with each other are what is referred to as the DDR (DIV-DRIF-RAD) regulatory module and constitute part of a larger gene regulatory network that regulates floral development and morphology in *Antirrhinum majus*.

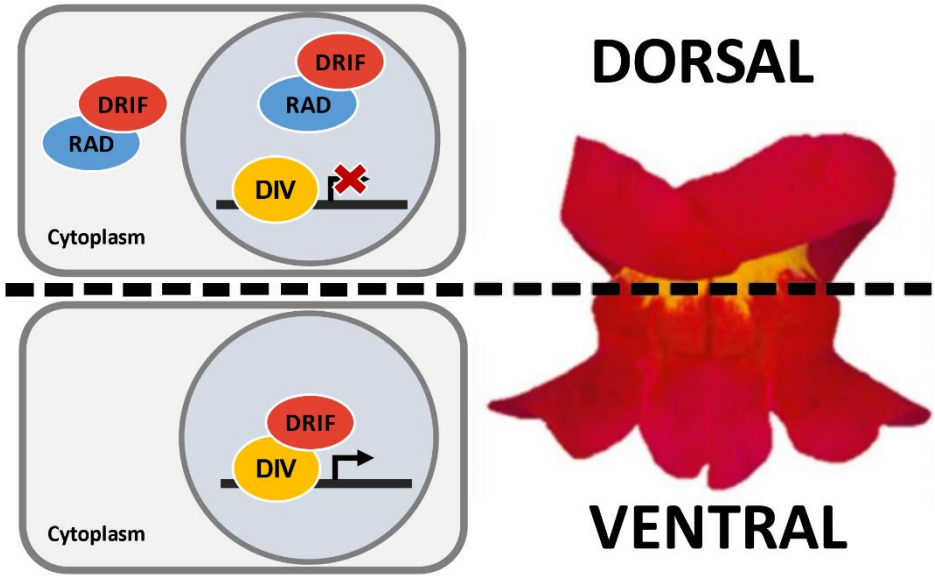

Figure 2. Proposed model for the antagonistic relationship of RAD and DRIF in the dorsal region of *Antirrhinum majus* flowers in which DIV and RAD compete to bind with DRIF. It is proposed that, in the ventral region of the flower, DIV binds to DRIF, promoting its nuclear localization and creating the DIV-DRIF protein complex, regulating expression and leading to petals with ventral identity. In the dorsal region, RAD is present and binds to DRIF, disrupting the formation of the DIV-DRIF complex, leading to petals with dorsal identity. (Adapted from Raimundo et al., 2013)

1.2. DDR module proteins

All the proteins that constitute the DDR regulatory module, DIV, DRIF and RAD, contain at least one MYB domain. The MYB family of transcription factors is present in a variety of eukaryotes, including animals, fungi, and plants, including algae. While in animals there is less variety of function, mostly being involved in regulating cell division and differentiation (Lipsick, 1996), the MYB transcription factors of plants are involved in controlling a variety of functions, from regulating cell cycle and meristem formation to controlling biosynthetic pathways and secondary metabolism (Jin & Martin, 1999; Rosinski & Atchley, 1998).

The DDR module MYB domains are considered to be distinct from other MYB domain proteins. MYB domains typically have three Tryptophan amino acids that are separated by other amino acids and are flanked by basic amino acids, Histidine (H), Lysine (K) and Arginine (R). The Tryptophan residues create a hydrophobic scaffolding that maintains the helix-loop-helix structure of the domain. The MYB domains of the DDR module proteins are described as MYB-like due to variation of the characteristic Tryptophan aromatic residues (Wang et al., 1997). In the three proteins part of the DDR regulatory module one of the three Tryptophan residues is replaced by a Tyrosine amino acid, another aromatic, hydrophobic residue (**Figure 3**). In the case of DIV and

RAD, the third Tryptophan of the MYB domains is replaced by Tyrosine while in the MYB domain of DRIF, the second residue is replaced.

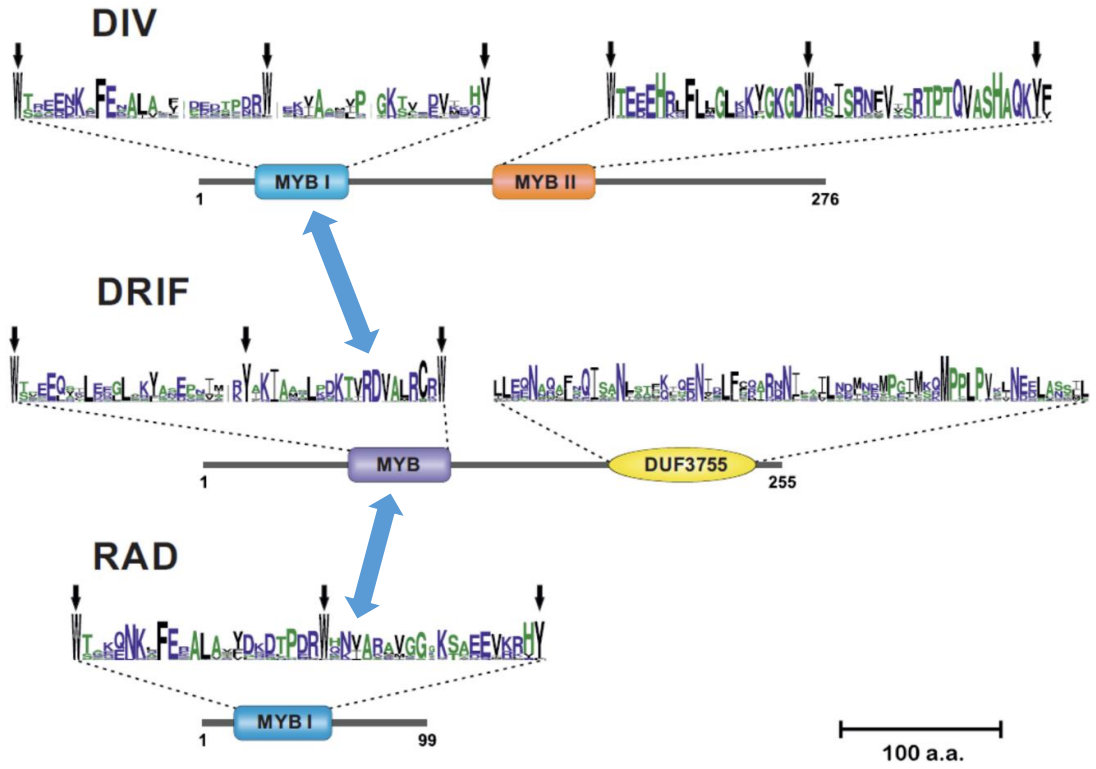

Figure 3. Representation of the conserved domains of the DIV, DRIF and RAD families of proteins. Domains are represented and named on the protein sequence and are matched with a generated sequence logo representing conservation of aminoacids. Sequence logos prepared using sequences from angiosperm species *Amborella trichopoda, Oryza sativa, Solanum lycopersicum, Antirrhinum majus* and *Arabidopsis thaliana*. Blue arrows point to domains known to bind to each other. Black arrows point to characteristic aromatic residues of MYB domains (Adapted from Raimundo et al., 2018).

DIV is a transcription factor that has two MYB domains: the N-terminal is a MYB/SANT protein binding MYB domain (MYBI) while the C-terminal is a DNA-binding MYB domain (MYBII). The DNA-binding domain is denominated as a SHAQKYF MYB domain due to a SHAQKYF amino acid sequence motif preserved in the domain, in which the characteristic third aromatic residue is contained (Almeida et al., 2018).

RAD is a small protein with a single MYB domain, which is very similar to the MYBI protein binding domain of DIV (Corley et al., 2005). It is considered to interact with DRIF, sequestering the protein and interfering with DIV function (Machemer et al., 2011; Raimundo et al., 2013).

DRIF is composed by a C-terminal domain of unknown function (DUF) known as DUF3755, that has recently been shown to interact with proteins of the WOX and KNOX families in the poplar, *P. trichocarpa* (Petzold et al., 2018), and a N-terminal MYB/SANT domain involved in protein binding, responsible for DRIF's interaction with DIV and RAD (Machemer et al., 2011).

1.3. DDR module evolution

In a recent study by Raimundo et al. (2018), DDR module genes were identified across all major land plant groups and in green algae species (**Figure 4**) and it was observed that the DRIF and DIV proteins of the green algae, *Klebsormidium nitens*, interact with each other, implying conservation of function. While these genes were not found in more distantly related algal groups such as Rhodophyta, several different genes containing the MYB-like DNA binding domain MYBII were found in a few species of red algae (**Figure 4**). It was determined that the DIV and DRIF protein families evolved, probably in ancestral green algae, via duplications of a pre-existing MYB domain, and that RAD later evolved in gymnosperms by duplication of the MYBI protein interacting domain of DIV (**Figure 5**).

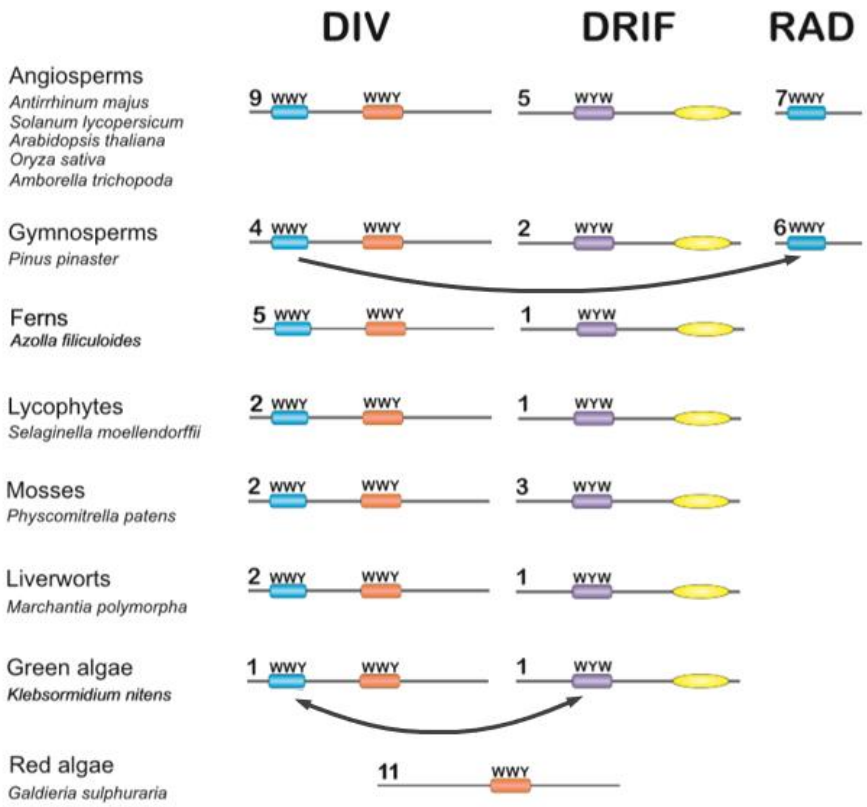

Figure 4. Representation of DIV, DRIF and RAD homolog proteins and their domains at several key evolutionary points, from red algae to angiosperms. The numbers indicate the average number of homologs of the gene present in the species. The arrows point to domain duplication events. Homolog domains are colour coded (Adapted from Raimundo, et al., 2018).

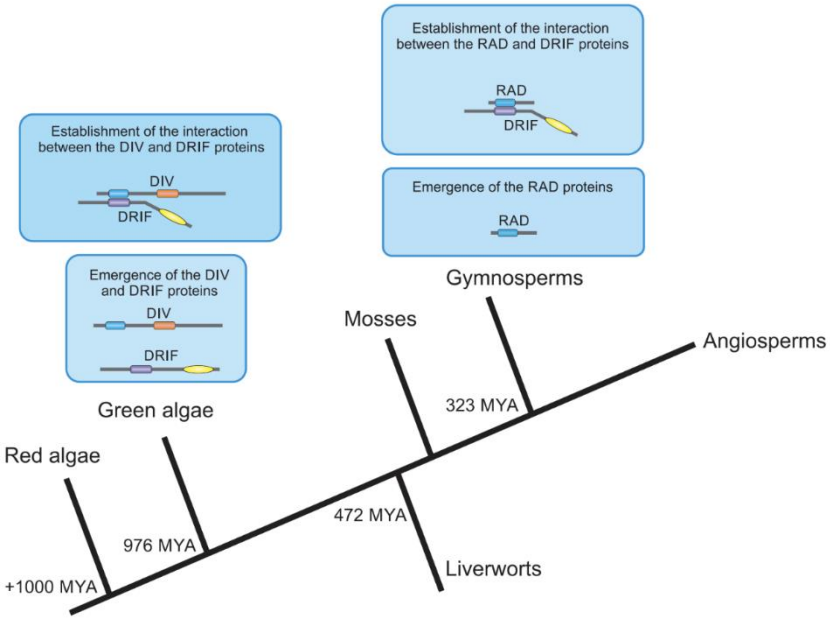

Figure 5. Scheme of the establishment of the DDR module protein families and their interactions over evolution. DIV and DRIF emerged and their interaction was established in green algae. RAD emerged in gymnosperms and its interaction with DRIF was established (Raimundo, et al., 2018).

Other groups of algae exist, other than Rhodophyta, that are closely related to the green lineage, such as Glaucophyta and Cryptophyta (**Figure 6**). Species of these groups, and of other more distantly related groups, have had genomes sequenced and published and, in more recent years, integrated into JGI databases (Grigoriev et al., 2020). It could prove enlightening to search for DIV and DRIF homologs in these species since it could help better define the early evolutionary history of *DIV* and *DRIF* before they would eventually form the DDR regulatory module with the evolution of *RAD* in gymnosperms.

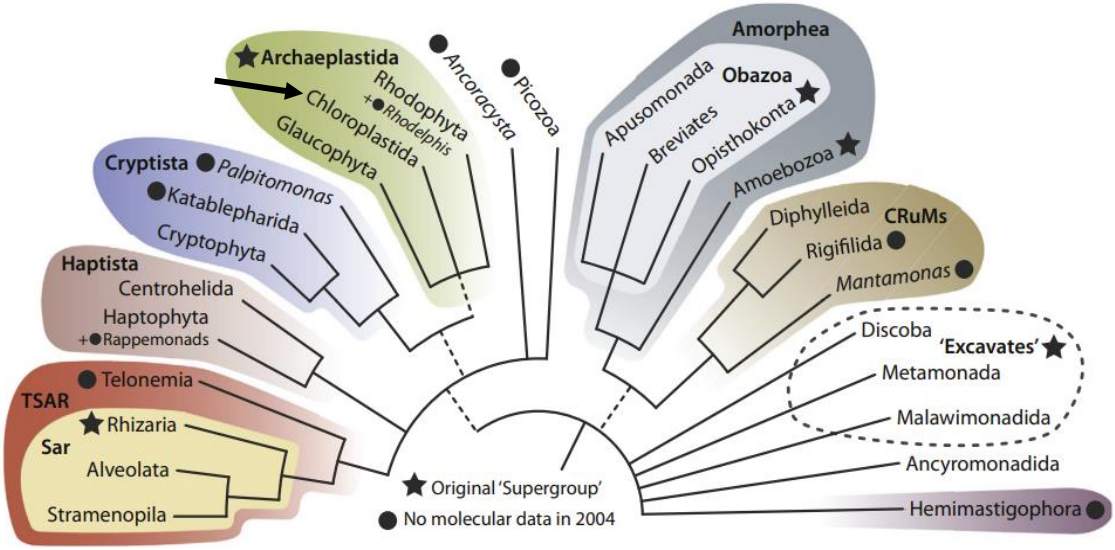

Figure 6. The tree of eukaryotes. Representation of the recently established phylogenetic relationship between the newly outlined supergroups. Arrow points to group in which DDR module genes have been found as of now (Chloroplastida) (Burki et al, 2020).

Studying the evolution of gene regulatory networks can prove difficult. Research has typically focused on higher organisms which, over time, have increased in complexity due to accumulation of genomic and gene duplications (De Smet & Van de Peer, 2012). To get around this issue, ancestral species can be used to study these networks. In the case of the ancestral function of the DIV and DRIF genes, *Marchantia polymorpha*, a species of liverwort that was found to have two homologs of *DIV* and one homolog of *DRIF* and thus could be used to study the ancestral function of some of the DDR genes.

1.4. *Marchantia polymorpha*

M. polymorpha is a species of liverwort (bryophytes subgroup) that is commonly distributed throughout temperate regions. Liverworts are basal non-vascular land plants and thought to be part of the first plants to evolve to live on land (**Figure 7**), as evidenced by fossilized spores that were found to belong to liverworts from the Ordovician period (470 m.y.). Records of liverworts in

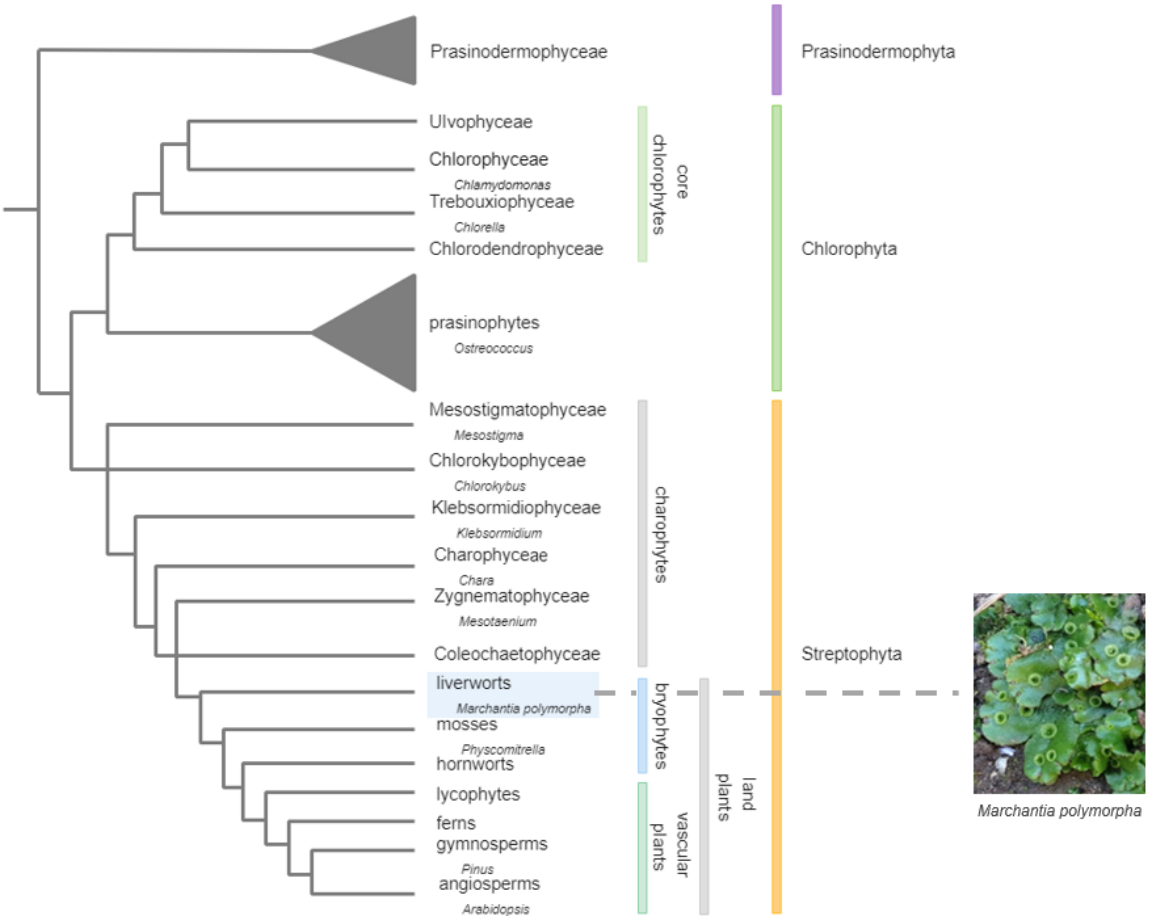

Figure 7. Phylogenetic tree presenting the evolution of Chloroplastida, from green algae to vascular plants. The position of Marchantia polymorpha as part of one of the earliest land plant groups is highlited with a blue box and an image of the plant (Adapted from Leliaert et al., 2012, Bowman, et al.,2017 and Li, Wang, et al., 2020)

literature date back to the ancient Greek civilisation where around 400 B.C. the first Herbals (compendiums of all accumulated knowledge about plants and their medicinal properties) were written (Bowman, 2016). The name liverwort originated in the Middle Ages from the practice of using plants that resembled human body parts to treat ailments affecting those parts and the thalloid body of liverworts like *M. polymorpha* was thought to be akin to the human liver. In the 1800s and early 1900s, *M. polymorpha* was widely used in scientific plant research, especially in studies of morphological and physiological responses to external environmental factors. However, in the late 20th century, *M. polymorpha* became relatively forgotten as a model plant. As genetics and molecular biology grew as fields of research, new model species were adopted, such as the angiosperm *Arabidopsis thaliana* and, in the bryophyte Division, the moss *Physcomitrella patens* (Shimamura, 2016).

More recently, *M. polymorpha* has re-emerged as a model species (Bowman et al., 2016; Bowman et al., 2017; Flores-Sandoval et al., 2018; Furumizu et al., 2018; Romani et al., 2018; Montgomery et al., 2020; Naramoto et al., 2020) with the development and adaptation of molecular research tools and techniques such as various transformation protocols with the use of *Agrobacterium*, gene targeting and editing techniques involving CRISPR-Cas9 (Ishizaki, Nishihama, Yamato, et al., 2015) and the recent completion of the *M. polymorpha* genome project, making it easier to perform research with the species.

1.5. *M. polymorpha* life cycle

The life cycle and basic morphology of *M. polymorpha* was recently reviewed in detail by Shimamura (2016) (**Figure 8**), the review serving as the basis for this short summary.

The dominant body of liverworts occurs during the haploid gametophyte generation. In the case of *M. polymorpha*, a complex thallus serves as the main plant body. Plants originate from a unicellular haploid spore that germinates and develops into an initial “protonema” with rhizoids, resembling a smaller, simpler thallus. The protonema develops into the thallus via organized cell divisions of a single apical cell and cell differentiation into various tissues.

New plants can also arise from gemmae. Gemmae are the product of asexual reproduction in *M. polymorpha* and are discoid groups of cells produced in gemma cups on the dorsal side of the thallus. When the gemmae are expelled from the cup by water drops and come into contact with the earth, they develop rhizoids and dorsoventrality, eventually forming a new plant that is genetically identical to the parent plant.

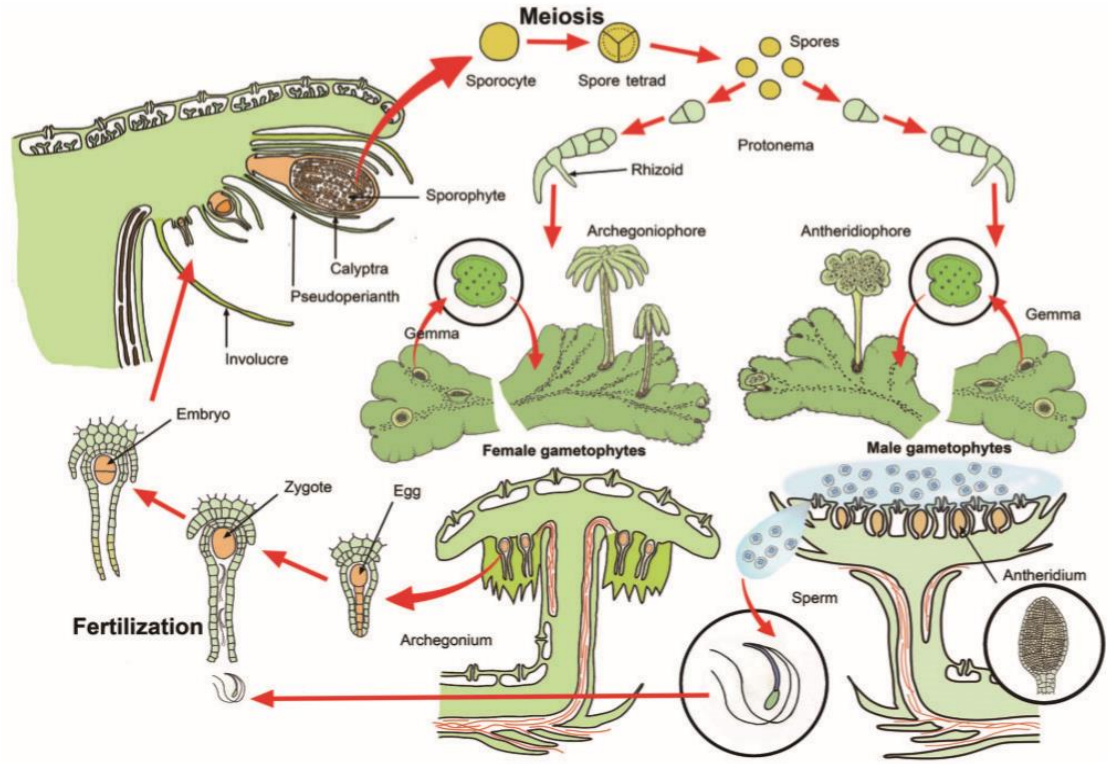

Figure 8. Representation of the life cycle of *Marchantia polymorpha*. Haploid spores grow into plants which can either reproduce sexually or assexually via gemmae that grow into fully formed plants once on soil. Sexual reproduction involves sperms produced in the antheridiophore being trasnported in water to the archegoniums, produced in the archegoniophores, and fertilizing the egg. The egg then grows and forms the sporophyte, within which the sporocytes are present. Each sporocyte suffers meiosis and forms four spores, each becoming a plant and repeating the cycle (Shimamura, 2016).

The production of spores occurs in the sporophyte generation, which is located and developed on specialized sexual structures. *M. polymorpha* is dioecious and so each plant either has male or female gametangia. The gametangia form on sex specific umbrella-shaped sexual structures, the gametangiophores. The male gametes move through water, and through the rhizoids (Pressel & Duckett, 2019), to reach the female archegonia and fertilize the egg forming the zygote, the first diploid cell of the sporophyte generation. These features of *M. polymorpha's* life cycle allow for easy manipulation in a research environment, ease of crossing and ease propagation, increasing the speed and effectiveness of genetic research.

1.6. Advantages of *M. polymorpha* as a model species

Overall, research has focused on model organisms that tend to be higher plants such as angiosperms, as flowering plants tend to be more economically important due to their use in agriculture. The focus on flowering plants creates difficulties in research. Higher plant genomes are more complex and can have many homologs of any one gene, leading to functional redundancy

and the possible evolution of new functions, further adding to the complexity of gene networks (De Smet & Van de Peer, 2012). As a possible strategy to get around these issues, studies can be and have been conducted using early land plants, which have reduced genome complexity.

Many of the advantages of using *M. polymorpha* as a model species are derived from its short life cycle and the dominance of the haploid phase. (**Figure 8**). Haploidy is a big advantage of using liverworts in genetic research because having only one copy of each gene reduces the time to obtain mutants, as it is not necessary to confirm whether the mutation is heterozygous or homozygous. *M. polymorpha* can reproduce both sexually and asexually, through spores and through gemmae, respectively. Plants can then be crossed to produce single cell haploid spores, which facilitates transformation as the protocols can be carried out using a single cell. The option of sexual reproduction also means that transformants can be crossed with each other to obtain double or even triple mutants. Through asexual reproduction, mutant lines can be easily propagated in laboratory conditions without requiring the development of sexual structures.

The recently published genome of *M. polymorpha* (Bowman et al., 2017) will be a valuable tool in the future to increase efficiency of studies into gene networks and allow for discovery of new, related genes through bioinformatic tools. As a basal land plant, its genome is small (about 230 Mb) (Berger et al., 2016) and has not suffered many genome wide duplications, meaning that it has a lower amount of gene copies and a less complex network of interactions with lower redundancy between genes than other model plants (Shimamura, 2016).

1.7. DDR module in *M. polymorpha*

In actuality, the DDR regulatory module does not exist in *M. polymorpha.* As described above, only *DIV* and *DRIF* homologs were found in this species and so the regulation resulting from competition between *DIV* and *RAD* does not naturally occur. The function of the *DIV* and *DRIF* genes present in the liverwort is largely unknown, as research into the module in this plant is fairly recent. Coelho (2019) worked to uncover the ancestral functions of these genes, having created mutant knockout lines for Mp*DIV1/2* and Mp*DRIF* using CRISPR-Cas9 technology, overexpression lines and performed expression analysis at a spatial and temporal level.

Initial results suggested that the MpDIVs are involved in the development of the thallus, the main body of the plant. Mp*DIVs* are believed to be involved in the regulation of cellular expansion and/or proliferation and, additionally, Mp*DIV2* could possibly have a role in controlling plant shape. The function of Mp*DRIF* remains unknown.

Much work remains to be done to uncover the role these ancestral genes play in the development and life cycle of *M. polymorpha*. Further mutant phenotype analysis, more detailed studies into the interactions occurring between the different proteins of the module, such as protein interaction analyses, and into other proteins that may be involved in or derived from the module.

1.8. Objectives

Gene regulatory networks are complex systems that can be involved in influencing a huge variety of different mechanisms in organisms. Within the plants group, the more recent groups, such as angiosperms and gymnosperms, have additional genome complexity and thus more complex networks of genes. Many of these genes have duplicates, adding to the complexity of interactions occurring in these plants. As a strategy to understand the essential function of the DDR regulatory module, the genes from this system can be studied in less complex organisms with fewer gene copies, such as *M. polymorpha*. Additionally, to add evolutionary context to findings in *M. polymorpha*, homologs of these genes can be studied in algae ancestral to land plants.

Overall, this thesis aims to gain further understanding of the ancestral function of Mp*DIVs* and Mp*DRIF* and to learn more about the early evolutionary history of the proteins that would eventually form the DDR regulatory module.

To further unveil the ancestral function of the *DIV* and *DRIF* homologs of *M. polymorpha*, a more classical approach will be used, with phenotypical analysis of knockout mutant and overexpression plant lines plants over early development. To uncover the evolutionary history of *DIV* and *DRIF*, a phylogenetic analysis of the earliest *DIV* and *DRIF* homologs in various eukaryotic species will be carried out. Finally, to learn more about the ancestral function of MpDIVs at the molecular level, interaction studies will be employed to determine whether the DIV proteins of *M. polymorpha* bind to the same DNA sequences as those of angiosperms.

2. Materials and methods

2.1. Biological material

2.1.1. Plant material

The *M. polymorpha* ecotype BoGa was provided by Sabine Zachgo's laboratory (University of Osnabrück, Germany). *M. polymorpha* plants were grown on Gamborg medium at half strength [1.582 g L^{-1} Gamborg B5 medium vitamins (Duchefa); 1.4% (w/v) plant agar (Duchefa)] in tall petri dishes (100 x 20 mm, Greiner bio-One) under long-day conditions (16 h in light/ 8 h in dark) at 20 °C with light intensity varying between 40-45 µmol m^{-2} s^{-1}.

2.1.1.1. Gemmae sterilisation protocol

Gemmae from *M. polymorpha* were used for plant propagation and were sterilised for growth on medium. Sterilisation started with collection of 1-3 gemma cups from plants into a sterile tube and submerged in 1 mL of sterilisation solution (0.2-0.5% (v/v) sodium hypochlorite and 0.1% (v/v) Triton X-100), mixed by vortex and, after 30 seconds, the supernatant was removed. The gemmae were then washed three times with 1 mL of ultrapure water (H_2O_{up}), which was vortexed, left for 1 min and then removed. Gemmae were transferred to plates with Gamborg B5 medium at half strength.

2.1.2. Bacterial material

Bacterial strains were used in a variety of methods including cloning procedures, protein expression and plant transformation. The different strains used in this work are described in **Table 1**.

2.2. Bacterial transformation

2.2.1. *Escherichia coli*

2.2.1.1. Preparation of competent cells (large scale)

A single *Escherichia coli* colony was used to inoculate 5 mL of lysogeny broth (LB) liquid media (10 g L^{-1} NaCl; 10 g L^{-1} Tryptone; 5 g L^{-1} yeast extract) (Bertani, 2013) and the culture was

incubated overnight at 37 °C and 200 rpm. In an Erlenmeyer flask, 200 mL of LB media were then inoculated with 1 mL of the 5 mL culture and incubated at 37 °C and 200 rpm for between 2-3 h or until the OD_{600} (optical density at 600 nm) of the culture reached 0.25. The culture was then transferred into four 50 mL tubes and the cells were pelleted by centrifugation at 4 °C and 3000 *g* for 5 min. After the supernatant was discarded, each pellet was resuspended in 16 mL of cold and sterile 0.1 M $MgCl_2$, the tubes were placed and maintained on ice for 30 mins, after which the cells were pelleted and resuspended in 5 mL of cold and sterile TG salts solution (75 mM $CaCl_2$; 6 mM $MgCl_2$; 15% (v/v) glycerol). Cells were again pelleted and resuspended in 1.5 mL TG salts solution and kept on ice for between 4-24 h. Finally, 100 µL of suspended cells were distributed into aliquots, frozen in liquid nitrogen, and stored at -80 °C.

Table 1. List of microorganism strains used with use case, species and genotype.

Use	Species	Strain	Genotype	Reference
Cloning procedures	*Escherichia coli*	DH10β	F- *mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) φ80*lac*ZΔM15 Δ*lac*X74 *rec*A1 *end*A1 *ara*D139 Δ(*ara-leu*)7697 *gal*U *gal*K λ- *rps*L(Str^R) *nup*G	Durfee et al., 2008
Heterologous expression	*Escherichia coli*	Rosetta™ (DE3)pLysS	F- *ompT hsdS*$_B$(r_B- m_B-) *gal dcm* (DE3) pLysSRARE (Cam^R)	Novagen
		BL21(DE3)-R3-pRARE2	F- *ompT hsdS*$_B$(r_B- m_B-) *gal dcm* (DE3) pRARE2 (Cam^R)	Novagen
		OverExpress™ C43(DE3)	F– *ompT hsdSB (rB- mB-) gal dcm* (DE3)	Miroux & Walker, 1996
Plant transformation	*Agrobacterium tumefaciens*	C58C1 (GV2260)	pTiB6S3ΔT-DNA	Deblaere et al., 1985

2.2.1.2. Quick preparation of competent cells (small scale)

A single *E. coli* colony was used to inoculate 5 mL of LB media and incubated overnight at 37 °C and 200 rpm. 100 µL were then taken and used to inoculate 4.9 mL of LB medium, diluting 1:50, and incubated for 2h at 37 °C. The whole medium was then centrifuged at 3000 *g* for 1 min, the supernatant was discarded and 2 mL of ice cold 0.1M $CaCl_2$ were added to the pellet and used to resuspend the cells, which were then incubated on ice for 20 min. Cells were then centrifuged again at 3000 *g* for 1 min and 500 µL of ice cold 0.1M $CaCl_2$ were used to resuspend the pellet and were distributed into 100 µL aliquots for later transformation.

2.2.1.3. *E. coli* transformation

A 100 µL aliquot of competent cells was taken from storage, 50-100 ng of DNA were added to it and the mixture was incubated on ice for 30 min. The cells were then submitted to a heat-shock by being placed at 42 °C for 45-60 s, followed by 2 more min on ice. 900 µL of LB were added to the tube and the cells were incubated for 1 h at 37 °C and 200 rpm. The cells were pelleted by centrifugation at 4000 *g* for 1 min, approximately 900 µL of supernatant were discarded and the remaining 100 µL of resuspended cells were spread on LB-agar plates (LB medium; 1.5% (w/v) agar) containing the proper selective antibiotics. Plates were incubated overnight at 37 °C.

2.2.2. *Agrobacterium tumefaciens*

2.2.2.1. Preparation of competent cells

A single *A. tumefaciens* colony was used to inoculate 5 mL of LB media with the appropriate antibiotics and the culture was incubated overnight at 28 °C and 200 rpm. 50 µL of the initial culture were then used to inoculate 50 mL of LB media which was incubated overnight at 28 °C and 200 rpm until the OD_{600} of the culture was between 0.5-1. The culture was then transferred to a 50 mL tube and cooled on ice for 10 min, centrifuged at 3000 *g* for 6 min at 4 °C. The supernatant was then discarded, and the pellet was rinsed with 1 mL of cold and sterile 20 mM $CaCl_2$. After the cells being once again pelleted, the supernatant was discarded, and the pellet resuspended in 1 mL of cold and sterile 20 mM $CaCl_2$. Cells were distributed in aliquots of 100 µL into pre-chilled tubes, frozen in liquid nitrogen and stored at -80 °C.

2.2.2.2. *Agrobacterium tumefaciens* transformation

A 100 µL aliquot of *A. tumefaciens* competent cells was taken from storage and 1 µg of DNA was added to the tube, which was then mixed by inversion and frozen in liquid nitrogen for 5 min. The cells were the incubated at room temperature for 10 min, 900 µL of LB media was added, and the tube was incubated at 3 h at 28 °C and 200 rpm. The cells were pelleted by centrifugation at 4000 *g* for 1 min, part of the supernatant (approximately 900 µL) was discarded and the remaining 100 µL with resuspended cells were spread on LB-agar plates with the proper selective antibiotics. Plates were incubated for 48 h at 28 °C.

2.3. DNA methods

2.3.1. Plant DNA extraction

2.3.1.1. Mixer mill protocol

Plant DNA was extracted from samples of 3–4-week-old thallus (±5 cm^2), which were placed into 1.5 mL tubes with 400 µL of extraction buffer (200 mM Tris-HCl pH 7.5; 250 mM NaCl; 25 mM EDTA; 0.5% (w/v) SDS) and two small metal bearings. Using a mixer mill (RETSCH MM 400), the tissue was ground by the bearings and homogenized with the buffer with two cycles for 45 s at 30 Hz. Phases were separated by centrifuging the tube at 21000 *g* for 5 min and 300 µL of supernatant were transferred to a new tube. 300 µL of isopropanol were then added to the tube to precipitate nucleic acids and the tube was once again centrifuged at 20000 *g* for 5 min. The pellet was then rinsed with 500 µL of 70% (v/v) ethanol and air dried. The samples of DNA were resuspended in 30-100 µL of H_2O_{up} and stored at -20 °C.

2.3.1.2. Maceration protocol

Samples of 3–4-week-old thallus (±5 cm^2) were sectioned from the plants and placed in a tube that was then frozen in liquid nitrogen. The frozen tissue was then ground with a mortar and pestle until reduced to powder. 500 µL of extraction buffer (100 mM Tris-HCl pH 8.0; 1.4 M NaCl; 20 mM EDTA; 2% (w/v) CTAB; 0.2% (v/v) β-mercaptoethanol) were added to the powder, still in the mortar, and further ground until homogenized. Once the solution melted, it was transferred to a tube and incubated at 65 °C for 20-30 mins. 500 µL of chloroform were then mixed in and the tube was centrifuged at 13000 *g* for 5 min. The supernatant was carefully transferred to a new tube to which 330 µL of isopropanol 100% were added and mixed. The tube was centrifuged at 13000 g for 10 mins, the supernatant was removed, and the pellet was air dried before it was suspended in 30 µL of H_2O_{up}. Samples were stored at -20 °C.

2.3.2. Isolation of plasmid DNA

2.3.2.1. *E. coli* miniprep protocol

A single *E. coli* colony was used to inoculate 10 mL of LB media with the appropriate antibiotics and the culture was incubated overnight at 37 °C and 200 rpm. Then, 1 mL of the

incubated overnight culture was placed in a tube and centrifuged at 21000 *g* for 1 min, the supernatant was discarded, and the process was repeated twice more in the same tube. The resulting pellet was resuspended in 100 µL of cold GTE solution I (50 mM glucose; 100 mM EDTA; 25 mM Tris-HCl pH 8.0). 200 µL of GTE solution II (1% (w/v) SDS; 0.1 M NaOH) were added from a freshly prepared solution and were mixed slowly by successive inversion of the tube. 150 µL of GTE solution III (3 M $C_2H_3KO_2$; 5 M CH_3COOH), were then added, mixed by slow inversion and the tube was place on ice and incubated for 15 min before being centrifuged at 21000 *g* for 15 min. The supernatant was retrieved to a new tube and centrifuged again. The supernatant was then retrieved, 1 mL of cold ethanol (EtOH) at 100% (v/v) was added to the tube which was then centrifuged at 21000 *g* for 15 min and at 4 °C. The pellet was rinsed with 50 µL of TE (10 mM Tris-HCl; 1 mM EDTA pH 8.0) with (10 ng mL^{-1}) RNase and the tube was incubated at 37 °C for 5 min, vortexed and incubated again at 37 °C for 15 min. After incubation, 30 µL of 20% (v/v) PEG 4000 and 2.5 M NaCl were added to the tube that was then vortexed and left on ice for 1-5 h. The tube was then centrifuged for 5 min at 21000 *g* at 4 °C, the supernatant was discarded, and the pellet was washed with 500 µL of cold 70% (v/v) EtOH. After a final centrifugation, the supernatant was carefully removed, and the pellet was air dried before being resuspended with 30 µL H_2O_{up}. The DNA solutions were stored at -20 °C.

2.3.2.2. *E. coli* NZYTech miniprep protocol

A single colony of *E. coli* was used to inoculate 10 mL LB supplemented with antibiotic and incubated overnight at 37 °C and 200 rpm. Then, 1 mL of the culture was added to a tube and centrifuged at 15000 *g* for 1 min and the supernatant was discarded. The pellet was then resuspended in 250 µL of chilled A1 buffer and vortexed. 250 µL of A2 buffer were then added, the contents of the tube were mixed by inversion and incubated at RT for 2-3 min, after which, 300 µL of A3 buffer were added and mixed by inversion. The tubes were centrifuged at 15000 *g* for 5 min, the supernatant was collected into a spin column in a 2 mL collection tube, the column was centrifuged at 11000 *g* for 1 min and the flowthrough was discarded. 500 µL of AY buffer were added to the spin column, it was centrifuged at 11000 *g* for 1 min and the flowthrough was discarded. 600 µL of A4 buffer were added, it was centrifuged at 11000 *g* for 1 min, the flowthrough was discarded, the spin column was moved into a new, dry 2 mL collection tube and was centrifuged at 11000 *g* for 2 min. The dry spin column was then placed in a tube and 30 µL

of H_2O_{up} were added to the spin column. The tube was then centrifuged at 11000 *g* for 1 min and the DNA stored at -20 °C.

2.3.3. DNA sample concentration and quality estimation

The concentration of DNA was estimated via UV spectrophotometry using the *Nanodrop* (NanoDrop® ND-1000 UV-Vis Spectrophotometer, Thermo Fisher Scientific) to measure absorbance at 260 nm (A_{260}). One unit of absorbance at 260 nm was assumed to correspond to 50 µg mL^{-1} of DNA. The quality of DNA samples was evaluated by the ratios of A_{260}/A_{280} and A_{260}/A_{230} also obtained using the *Nanodrop*. Pure DNA samples were expected to have ratio values of A_{260}/A_{280}=1.8 and A_{260}/A_{230}= 2 to 2.2.

2.3.4. Agarose gel electrophoresis

An agarose gel was prepared with 1% (w/v) agarose and *GreenSafe Premium* (NZYtech) (0.5x final concentration) was added to the gel before solidification to stain nucleic acids loaded onto the gel. It was then submerged in 0.5x TAE buffer (0.02 M Tris; 95 mM acetic acid; 50 mM EDTA, pH 8.0) and DNA samples were loaded onto it. If the DNA samples were from Polymerase chain reactions (PCR) performed using *NZYTaq II 2× Green Master Mix* (Nzytech), then they were directly applied onto the gel. If not, then a 5x loading buffer (20% (w/v) Ficoll; 0.3% (w/v) Tartrazine; 125 mg mL^{-1} Xylene Cyanol) were added to the samples. *NZYDNA Ladder III* (Nzytech) was used as molecular size marker when necessary. The gel was run at 100-120 V, until the yellow part of the xylene cyanol dye reached the end of the gel. DNA bands in the gel were visualised and photographed on a UV transilluminator.

2.3.5. DNA purification

2.3.5.1. PEG purification protocol

This protocol was used to purify DNA samples from *att*B PCR products. The full resulting 50 µL product of the PCR were mixed with 150 µL of TE buffer (1 mM EDTA; 10 mM Tris-HCl pH 8.0) and 100 µL of 30% (w/v) PEG 8000 with 30 mM $MgCl_2$ and the tube was centrifuged at 10000 *g* for 15 min. The supernatant was then carefully removed, and the pellet was dissolved in 30 µL of H_2O_{up}. DNA samples were then quantified as described in 2.3.3 and stored until use at -20 °C.

2.3.5.2. Phenol/Chloroform DNA purification protocol

This protocol started with the addition of an equal volume of phenol-chloroform to the DNA sample, the tube was mixed by vortex and centrifuged at 13000 *g* for 10 min. The upper aqueous phase was carefully removed and placed in a new tube, where an equal volume of isopropanol and 1/10 of the total volume of 3 M sodium acetate pH 5.2 were added and mixed. The tube was then incubated at -20 °C for 1 h, centrifuged at 13000 *g* for 15 min, the supernatant was discarded and 200 µL of 70% (v/v) EtOH were added. The tube was once again centrifuged at 13000 *g* for 15 min and the pellet was air dried before being suspended in 30 µL of H_2O_{up}. DNA samples were then quantified as described above and stored until use at -20 °C.

2.3.6. Polymerase chain reaction (PCR) methods

DNA amplification by PCR was performed using the *T100™ Thermal Cycler* (Bio-Rad) thermocycler. All the primers used in this work are described in Annex A.

2.3.6.1. Amplification of DNA fragments from transgenic plants

Transgenic plants were genotyped by PCR amplification. Total PCR reaction mixtures were set up for a final volume of 10 µL by mixing 5 µL of *NZYTaq II 2× Green Master Mix*, 0.1-0.5 µM of forward and reverse primers and 50-100 ng of genomic DNA (gDNA). The PCR reaction was performed with the following settings: initial denaturation at 95 °C for 3 min; 35 cycles of 95 °C for 45 s, 55-58 °C annealing for 45 s and an extension step of 72 °C for 1-3 min; and one extension cycle at 72 °C for 5 min. The PCR products were stored at -20 °C or at 4 °C until analysis.

2.3.6.2. Amplification of DNA fragments for use in Gateway® cloning

The DNA fragments needed used for the Gateway® cloning procedures were amplified with *NZYTaq II 2× Green Master Mix* and as described in 2.3.6.1. A first reaction was performed to amplify the target sequence and a second reaction was performed to introduce the complete *att*B sequence in the fragment.

The first reaction mixture used 5 µL *NZYTaq II 2× Green Master Mix*, 0.1-0.5 µM of forward and reverse primers and 50-100 ng of coding DNA (cDNA) from wild type *M polymorpha*. The settings of the PCR were as follows: 95 °C of initial denaturation for 3 min; 35 cycles of 95 °C for 45 s, 59-60 °C annealing for 45 s and an extension step of 72 °C for 2-5 min; ending with one

extension cycle at 72 °C for 5 min. For the second reaction, the mixture was made to be 50 µL and used 25 µL of *NZYTaq II 2× Green Master Mix*, 0.1-0.5 µM of each *att*B adaptor (*att*B1 and *att*B2, Annex A) and 10 µL of PCR product from the first reaction. The settings were as follows: initial denaturation step at 95 °C for 3 min; 5 cycles of 95 °C for 45 s, 45 °C for 45 s and 72 °C for 2 to 5 min; 15 cycles of 95 °C for 45 s, 55 °C for 45 s and 72 °C for 2 to 5 min; and a final extension step at 72 °C for 5 min. The products of this PCR were then purified as described in 2.3.5.

2.3.6.3. Colony PCR amplification

Colony PCR amplification was used to confirm the presence of plasmids in transformed *E. coli* or *A. tumefaciens*. A single colony was collected from the selection plate and suspended in H_2O_{up}. The reaction mixture of 25 µL was composed by 12.5 µL *NZYTaq II 2× Green Master Mix*, 0.1-0.5 µM of both the forward and reverse primers and the suspended colony was used as the DNA template. The parameters for the PCR were as follows: initial denaturation step at 95 °C for 10 min; 35 cycles of 95 °C for 45 s, 55 °C annealing for 45 s and an extension step of 72 °C for 45 s to 5 min; and one final extension cycle at 72 °C for 5 min.

2.3.7. Gateway® cloning methods

Gateway® Technology is a universal cloning method that takes advantage of the site-specific recombination properties of the lambda bacteriophage to provide a highly efficient tool to transfer DNA sequences into multiple vector systems.

2.3.7.1. Cloning into pMpGWB

The pMpGWB series of vectors are Gateway® Binary Vectors specifically developed to simplify molecular analyses in *M. polymorpha* (Ishizaki, Nishihama, Ueda, et al., 2015).

2.3.7.2. Cloning into pMpGWB208, pMpGWB308, pMpGWB313, pMpGWB318 and pMpGWB321

Mp*DIV1*, Mp*DIV2* and Mp*DRIF* non-stop coding sequences (Mp*DIV1ns*, Mp*DIV2ns* and Mp*DRIFns*) were cloned using *Gateway® technology* into pMpGWB208, pMpGWB308, pMpGWB313, pMpGWB318 and pMpGWB321 vectors (Annex B) with the aim of overexpressing these genes with specific tag/fusion proteins in *M. polymorpha.* The coding sequences of the different genes were amplified as described in section 2.3.6.2. using primers 507 and 513 for

Mp*DIV1ns*, 509 and 514 for Mp*DIV2ns*, and 511 and 515 for Mp*DRIFns* (Annex A) with cDNA as source templates, were cloned into the donor vector pDONR201 and then into the destination pMpGWB vector.

2.3.7.3. Cloning into Gateway™ pDEST™15

Gateway™ pDEST™15 was used during this project in heterologous protein expression methods. During the cloning procedures, the Mp*DRIF* sequence was recombined from the donor vector, pDONR201, into the destination vector, pDEST™15, with the goal of achieving heterologous expression of the proteins in bacteria.

2.3.7.4. BP reaction

This first recombination reaction was performed to insert the DNA sequences with *att*B sequences into the donor vector. *att*B sequences were added to the DNA sequences as described 2.3.6.2. The reactions were performed for a total volume of 10 µL. Each reaction contained 1 µL of *BP clonase* enzyme mix (Invitrogen), equimolar amounts of donor vector and the *att*B-PCR product, and TE buffer to fill to 10 µL. Tubes were incubated at 25 °C for 1-2 h. The resulting plasmids were then used to transform *E. coli* DH10β competent cells. Once positive colonies were confirmed, plasmids were isolated by miniprep and stored at -20 °C.

2.3.7.5. LR reaction

This second recombination reaction was performed to transfer the DNA fragments present in the donor vector into the destination vector. The reactions were performed for a total volume of 10 µL. Each reaction contained 1 µL of *LR clonase* enzyme mix (Invitrogen), equimolar amounts of the donor vector with target DNA sequence and of destination vector and TE buffer to fill to 10 µL. Tubes were incubated at 25 °C for 1-2 h. The resulting plasmids were then used to transform *E. coli* DH10β competent cells. Once positive colonies were confirmed, plasmids were isolated by miniprep and stored at -20 °C.

2.4. Generation of *M. polymorpha* transgenic plants

2.4.1. Plant transformation

Transformation of *M. polymorpha* was performed using spores collected from mature sporangium in the laboratory of Sabine Zachgo.

To begin with, the mature sporangium were sterilized with a solution of 12% (v/v) NaClO and 10% (v/v) Triton X-100. The sporangium were thoroughly vortexed to release all spores from capsule. The spores were then incubated for 1-2 min, spun down, and resuspended in half-strength liquid Gamborg B5 medium (1.582 g L^{-1} Gamborg B5; 0.03% (w/v) glutamine; 0.1% (w/v) casamino amino acids; 2% (w/v) sucrose). 100 µL of this suspension were then used to inoculate 25 mL of liquid half-strength Gamborg B5 medium and cultured under white light (60 µmol photons $m^{-2}\,s^{-1}$) at 22 °C for 7 days and 130 rpm.

During the 7 days of culture, a single colony of *A. tumefaciens* carrying the construct was used to inoculate 5 mL of LB medium and incubated for 48 h at 28 °C and 170 rpm. The *A. tumefaciens* culture was then pelleted at 2000 *g* for 15 min, resuspended in 10 mL of liquid half-strength Gamborg B5 medium with 200 µM acetosyringone and incubated for 6h at 28 °C and 170 rpm. Then, 1 mL of the *A. tumefaciens* culture was added to the 25 mL culture of spores and were co-cultured for 3 days under white light (60 µmol photons $m^{-2}\,s^{-1}$) at 22 °C and 130 rpm. Spores were collected, washed three times with half strength Gamborg B5 medium and transferred to half strength Gamborg B5 medium with 1.4% (m/v) plant agar and the appropriate selection markers (100 mg L^{-1} chlorsulfuron; 100 mg L^{-1} gentamycin).

2.5. Protein Methods

2.5.1. Heterologous protein expression

Heterologous protein expression was carried out using three different expression strains of *E. coli* throughout this project, Rosetta™ (DE3)pLysS, BL21(DE3)-R3-pRARE2 and OverExpress™ C43(DE3) (**Table 1**), the genotype of which is described above in bacterial material. MpDIV1-pGEX, MpDIV2-pGEX and MpDRIF-pDEST15 were transformed into these strains as described above.

Two distinct expression protocols were used during optimization of the process. One of these utilized lactose induction, while the other used IPTG for induction of heterologous expression.

2.5.1.1. Lactose induction

A single colony of the *E. coli* expression strain selected was used to inoculate 5 mL of LB liquid medium with the appropriate antibiotics, depending on strain (**Table 1**) and vector used (Annex B), and the culture was incubated ON at 30 °C at 200 RPM. Then, 1 mL was taken from

the medium and centrifuged for 1 min at 4000 g. The pellet was resuspended in 10 mL of LB liquid medium with 20 g L^{-1} of lactose and incubated ON (16-22h) at 30°C.

2.5.1.2. IPTG induction

5 mL of LB liquid medium with the appropriate antibiotics were inoculated with a single colony of the *E. coli* expression strain and grown ON at 30°C. The next day, 1 mL was taken from the medium and centrifuged for 1 min at 4000 g. After the supernatant was discarded, the pellet was resuspended in and used to inoculate 10 mL of LB liquid medium with appropriate antibiotics and the cells were grown at 30°C until they reached the exponential growth phase (OD at 600 nm= 0.4-0.6). 1 mL was then taken and used to inoculate 9 mL of LB liquid medium with 0.4 mM IPTG and grown for 4 h.

2.5.2. SDS-PAGE

2.5.2.1. Sample preparation

Out of a post induction culture, 1 mL was removed and placed in a 1.5 mL tube, centrifuged at 21000 g for 5 min and the supernatant was removed. For preparation of samples with the total protein content of the cells, 100 µL of SDS-PAGE buffer 2x (125 mM Tris-HCl pH 6.8; 2% (w/v) SDS; 20% (v/v) Glycerol; 0.25% (w/v) Bromophenol Blue; 2% (v/v) DTT (dithiothreitol) (added immediately before use)) were added to resuspend the pellet, the samples were boiled for 10 mins and centrifuged at 21000g for 5 min to pellet cellular debris. For the preparation of samples in which the insoluble protein content was separated from the soluble proteins, 100 µL of Protein buffer (50 mM Tris-HCl pH 8.0; 10% (v/v) Glycerol; 0.1% (v/v) Triton X-100) were added and used to resuspend the pellet and the cells were burst by repeated freeze-thaw cycles. For this, the samples were placed in liquid nitrogen and thawed at room temperature 4-5 times. The tubes were then centrifuged at 21000 g for 5 min and the supernatant, and the pellet were separated for soluble and insoluble fractions respectively. To both the pellet and the supernatant 100 µL of SDS-PAGE buffer 2x were added, the pellet resuspended, and the supernatant homogenized with the solution. The samples were then boiled for 10 min and centrifuged at 21100 g for 5 min.

2.5.2.2. Protein Electrophoresis

Acrylamide gels were made to be 1.0 mm thick, were run in a Mini-PROTEAN® Tetra Cell Systems (Bio-Rad). The stacking gel (T (total monomer concentration)=4% (w/v), C (cross-linker

concentration)=2.5% (w/v); (per 10 mL: 1.3 mL of a 30% (w/v) Bis-acrylamide solution (30:0.8), 100 µL of a 10% (w/v) SDS solution, 2.5 mL of a 0.5 M Tris-HCl (pH 6.8) solution, 52 µL of a 10% (w/v) APS solution and 12 µL TEMED)) and the resolving gel (T=12% and C=2.5% (per 10 mL: 5 mL of a 30% (w/v) Bis-acrylamide solution (30:0.8), 100 µL of a 10% (w/v) SDS solution, 2.5 mL of a 1.5 M Tris-HCl (pH 8.8) solution, 52 µL of a 10% (w/v) APS solution and 16 µL TEMED)) were prepared and around 3 mL of resolving gel were placed into the gel holder, before filling to completion with water, to keep gel level as it solidified. Once solidified, water was removed and stacking gel was used to fill the gel holder and well defining combs were placed and removed once gel was solidified. In each well, 10 µL of sample were used from the top of the sample, avoiding the cellular debris pelleted beforehand. For molecular weight reference, 1 µL of either SDS-PAGE Molecular Weight Standard (Broad Range, Bio-Rad) or Precision Plus Protein™ Kaleidoscope™ (Bio-Rad) were added to each gel. The gels were submersed in SDS-PAGE Running Buffer (144 g L^{-1} Glycine; 10 g L^{-1} SDS; 30.3 g L^{-1} Tris Base) and were run at 200 V for 50-60 min).

2.5.2.3. Gel Staining

The gels were submersed in Coomassie staining solution (50% (v/v) Methanol; 10% (v/v) Acetic Acid; 0.25% (v/v) Coomassie Brilliant Blue B250) (Meyer and Lambert, 1965) solution for 1h30min under mild agitation. They were then transferred to de-staining solution (23% (v/v) Methanol; 9% (v/v) Acetic Acid) and were left under mild agitation until the surrounding solution seemed saturated, whereupon it was replaced by new de-staining solution. This was repeated until the polypeptide profiles were deemed to be visible enough in the gel. The remaining blue de-staining solution was filtered through activated charcoal, removing the pigment so that the solution could be reused.

2.5.3. Western blot

After SDS-PAGE, the gel was incubated in Transference buffer (25 mM Tris-HCl pH 8.3; 190 mM Glycine; 20% (v/v) Methanol) for 10 min. A nitrocellulose membrane cut to the size of the gel was placed in methanol for 10 seconds, followed by 5 min in deionized water and in transference buffer for 10 min. The polypeptides on the gel were then electrotransferred to the membrane on a Mini-PROTEAN® Tetra Cell Systems (Bio-Rad), using pre-chilled transference

buffer, The transference occurred during 1h at a constant current of 350 mA with agitation of the buffer and an ice pad to prevent overheating of the buffer around the gel.

2.5.3.1. Antibody binding

After transference, the membrane was blocked in TBST (20 mM Tris-HCl pH 7.4; 150 mM NaCl; 0.05% (v/v) TWEEN® 20) with 5% (w/v) milk powder for 1 h and then incubated in TBST 1x solution with milk powder and Anti-Glutathione-S-Transferase (GST) antibody (Sigma-Aldrich®) produced in rabbit at a dilution of 1:5000, and the membrane was then incubated in this solution for 2h. The membrane was then washed 3 times in TBST 1x for 5 min. Then a TBST 1x solution with dry milk and Anti-Rabbit IgG (whole molecule)–Peroxidase antibody (Sigma-Aldrich®) produced in goat at a dilution of 1:15000 was applied to the membrane for 1h at RT. Tthe membrane was placed in a clear plastic folder and immersed in 1 ml each of Bio-Rad® Clarity Western Peroxide Reagent and of Clarity Western Luminol/Enhancer Reagent and was kept hidden from light until it was visualized in a G:BOX Chemi XX6/XX9 (Syngene).

2.5.4. Electrophoretic Mobile Shift Assay (EMSA)

Electrophoretic Mobile Shift Assay (EMSA) (Garner & Revzin, 1981) was used to test the DNA binding ability of the MpDIV1 and MpDIV2 proteins and to test the binding between the MpDRIF protein and the MpDIV protein homologues. This technique relies on the different migration rates that free DNA probes and proteins complexed with DNA probes have in nondenaturing polyacrylamide gels to study protein-DNA and protein-protein interactions.

2.5.4.1. DNA probe preparation

To label the DNA oligonucleotides probes the fluorescent dye 6-carboxyfluorescein (6-FAM) was used. 6-FAM DNA probes (Sigma-Aldrich®) containing the sequence GATAA (Annex A), to which DIV proteins have previously been shown to bind. The DNA probe was prepared by mixing 10 µL of TEN buffer (1 mM EDTA; 100 mM NaCl; 10 mM Tris-HCl pH 8.0) with 20 µL of each complementary oligonucleotide in a proportion of 1:1 molar ratio and incubated for 10 minutes starting with a temperature of 95°C and cooled to 25°C by decreasing the temperature by 0.2°C every second.

2.5.4.2. Reaction mixture preparation

Reaction mixtures consisted of 500 ng poly *GC* solution, 10 ng DNA probe, 2 µL binding buffer (50 mM Tris-HCl pH 8.0; 500 mM NaCl; 5 mM EDTA; 0.5% (w/v) BSA; 10 mM DTT; 20% (v/v) glycerol), 2 µL of each protein used in the reaction taken from soluble fraction and finally H_2O_{up} to complete a total of 10 µL per reaction mixture.

2.5.4.3. Electrophoresis and visualization

Prepared 1.0 mm thick non-denaturing gels with 6% acrylamide concentration. Samples were added to the wells in volumes of 10 µL and the voltage was set to a constant 100 V for 60 minutes in TBE buffer (2 mM EDTA; 100 mM boric acid; 100 mM Tris-HCL pH 8.0). The TBE buffer was prechilled at 4°C and the gel was run with an ice pad in the tub to prevent overheating of the reaction. The gels were then visualized by UV fluorescence with a G:BOX Chemi XX6/XX9 (Syngene).

2.6. Bioinformatic methods

2.6.1. Sequence retrieval

The MpDIV2 and MpDRIF amino acid sequences (Annex C) were used to perform BLASTs using the PhycoCosm algal genomics research by the Joint Genome Institute BLAST tool in the PhycoCosm website (https://phycocosm.jgi.doe.gov/phycocosm/home). The performed BLASTs were *tblastn*, where protein sequences were used to search for similar translated nucleotide sequences and were done with all genomes present in PhycoCosm, which include genomes from several major groups of eukaryotes (TSAR supergroup; Excavata; Cryptista; Haptista; Archaeplastida (**Figure 6**)).

2.6.2. Phylogenetic analysis

MEGA version X software (version 10.2.4) was used for phylogenetic and molecular evolutionary analysis by building protein sequence alignments. *DIV* homolog alignments were built using the MUSCLE (Edgar, 2004) tool for multiple sequence alignment with the amino acid sequences of the MpDIV1 and MpDIV2 proteins and of all DIV homolog proteins found in algal species (Annex D). *DRIF* homolog alignments were also built using the MUSCLE tool (Annex D). The sequences used were the MpDRIF protein sequence and all DRIF homolog protein sequences

found in algal species (Annex D). Both alignments were carried out using the amino acids within the conserved domains. Amino acids outside the domains were not considered.

Evolutionary relationships were inferred by maximum likelihood under the WAG substitution model, assuming a gamma distribution and with 1000 bootstrap replicates using the MEGA version X software, producing phylogenetic trees.

3. Results

The overall objective of this thesis was to unveil the ancestral function and evolution of the *DIV* and *DRIF* genes that are part of the DDR regulatory module that have been found to regulate a variety of functions in angiosperm species.

To understand the function in the ancestral plant *M. polymorpha*, phenotypical analysis was performed with *DIV* and *DRIF* knockout and overexpressing mutants. Additionally, the *DIV* and *DRIF* homologs of *M. polymorpha* were cloned into the pMpGWB Gateway Cloning plasmids to be transformed into plants. The plasmids have different protein tags that are expressed fused to the cloned target genes to be used in future studies such as determining the subcellular localization of proteins with fluorescent proteins and determining whether DIV and DRIF proteins promote or repress expression by binding to repressive domains, among others.

To unveil more about the early evolutionary history of the DDR regulatory module, algae genomic resources were explored and used to trace the origin, evolution, and the conservation of the MYB domains of DIV and DRIF families. Homologs were searched for and identified in the genomes of various eukaryotic species, the identified peptide sequences were analysed and compared to the peptide sequences of MpDIV1, MpDIV2 and MpDRIF and phylogenetic analysis was conducted to uncover relationships between the proteins.

Finally, at the molecular level, it has yet to be proven that MpDIV1 and MpDIV2 bind to DNA, and more specifically, to the GATAA sequence, the DNA consensus binding site, to which *A. majus* DIV protein have been shown to bind to via EMSA (Raimundo et al., 2013). To replicate this experiment with *M. polymorpha* homologous proteins, the heterologous expression of Mp*DIV1*, Mp*DIV2* and Mp*DRIF* in *E. coli* was optimized to produce soluble, active protein for use in an EMSA.

3.1. Phenotype observation

One of the strategies employed to study the ancestral function of *DIV* and *DRIF* genes was to generate knockout mutant and overexpression lines of plants for Mp*DIV2* and Mp*DRIF*. These transgenic lines were obtained during a previous study (Coelho, 2019). During this thesis, plants from those lines were grown for 15 days and their development was followed and photographed (Leica DMC6200) and the length and width of the plants were later measured using the ImageJ software (Schneider et al., 2012) (**Figure 9**). Transgenic plant measurements were then compared with values obtained with WT plants grown and photographed simultaneously.

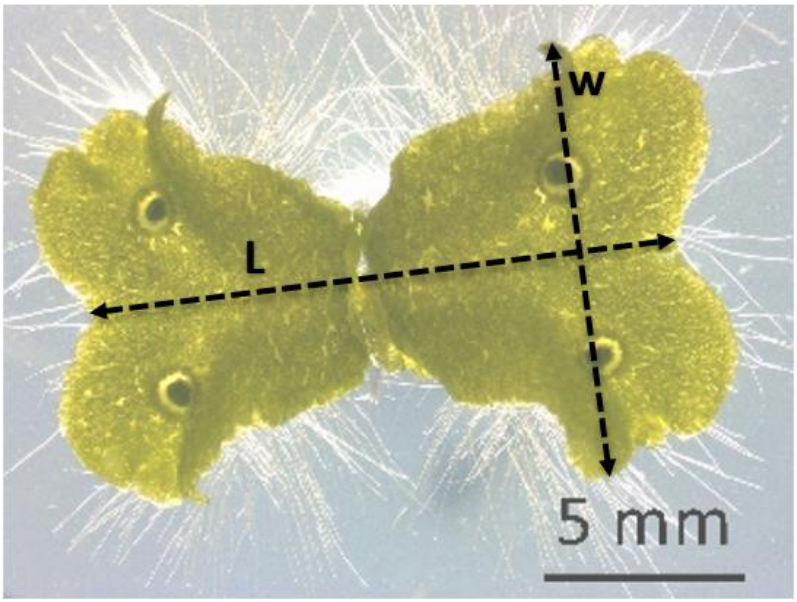

Figure 9. Example of how length (L) and width (W) of *M. polymorpha* plants were measured.

3.1.1. Knockout mutants

Knockout mutants for Mp*DIV1*, Mp*DIV2* and Mp*DRIF* genes were previously obtained using CRISPR/Cas9 technology. Using two guide RNA (gRNA), targeting either end of the target coding region of the gene intended for mutation, two double strand cuts are caused by the Cas9 protein, leading to deletions, insertions, or inversions, effectively knocking out or knocking down gene activity.

3.1.1.1. Mp*div2* knockout phenotype

M. polymorpha Mp*div2* plants were previously confirmed by sequencing. Two mutant lines Mp*div2* knockout lines #68 and #305 were analysed in this thesis. Knockout line 68 is one of several lines where deletion affected the MYBII, the DNA-protein interaction domain, while the sequence for the MYBI domain remained intact. The sequence of line #305 showed near complete deletion of the coding sequence, with loss of both functional domains.

Both the quantitative (**Figure 10**) and the visual analysis (**Figure 11**) show that Mp*div2* mutant plants are much smaller during the first 15 days of growth. Thallus width and length are both generally smaller in Mp*div2* plants than in WT plants. Between the plant lines, plants from knockout line #68 are smaller than plants from line #305.

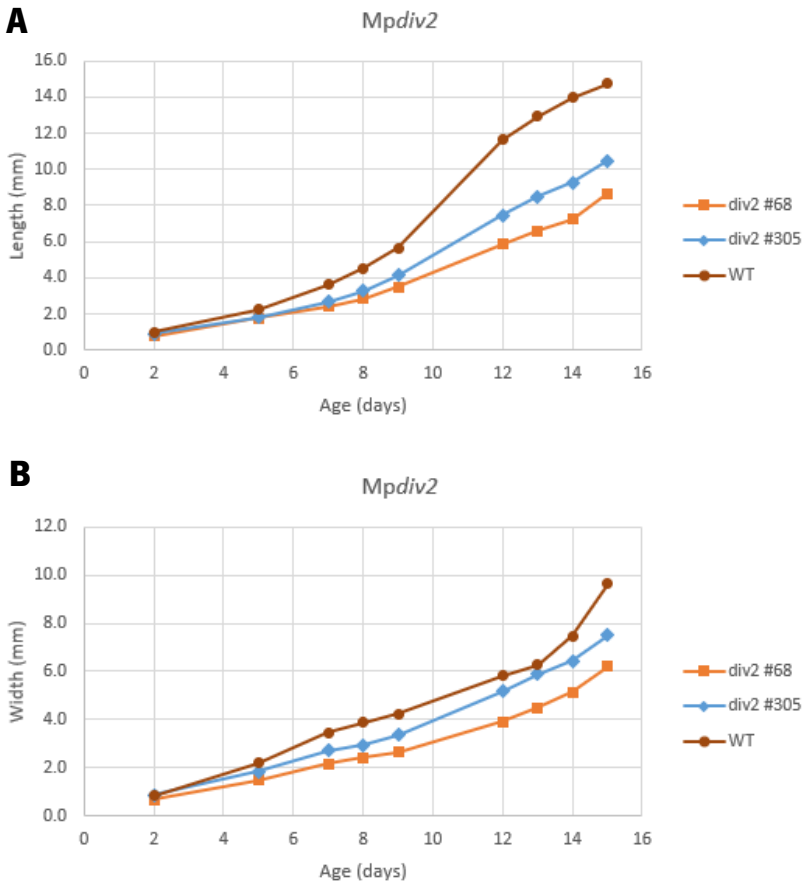

Figure 10. Analysis of the effect of Mp*div2* knockout on length and width of *M. polymorpha* plants. The thallus length and width of wild type and mutant Mp*div2* plants was measured for the first 15 days of development. **A -** The dark red line (circle) represents the wild type thallus length. The orange line (square) represents the thallus length of the Mp*div2* line #68 plants and the light blue line (diamond) represents the thallus length of Mp*div2* line #305 plants (n=4). **B -** The dark red line (circle) is the control that represents the wild type thallus width. The orange line (square) represents the thallus width of the Mpdiv2 line #68 plants and the light blue line (diamond) represents the thallus width of Mp*div2* line #305 plants (n=4).

M. polymorpha plants develop by growing radial branches that repeatedly bifurcate at the apex (Shimamura, M., 2015). Beginning as gemmae, the thallus is separated into two identical parts with an apical notch on each end. From the apical notches, these branches grow and extend until they bifurcate into two identical branches, and so on. On these branches, asexual reproductive tissues occurs by the development of gemmae cups. It is usual for one gemmae cup to form per branch, the first forming around the time the first bifurcation is initiated, around the 12th to 15th days of development (**Figure 11**), although sometimes gemmae cups can form earlier, on the body before the first bifurcation occurs. Considering these aspects of *M. polymorpha* growth, the bifurcations of Mp*div2* plants are less developed than in WT plants and gemmae cups do not form at the same time as WT plants. WT thallus branches curve upwards along the edges, making the plant body more three dimensional. Mp*div2* plants lack this curving and appear flatter than WT, even when in comparison with WT plants of similar size, for example 15-day old plants from line

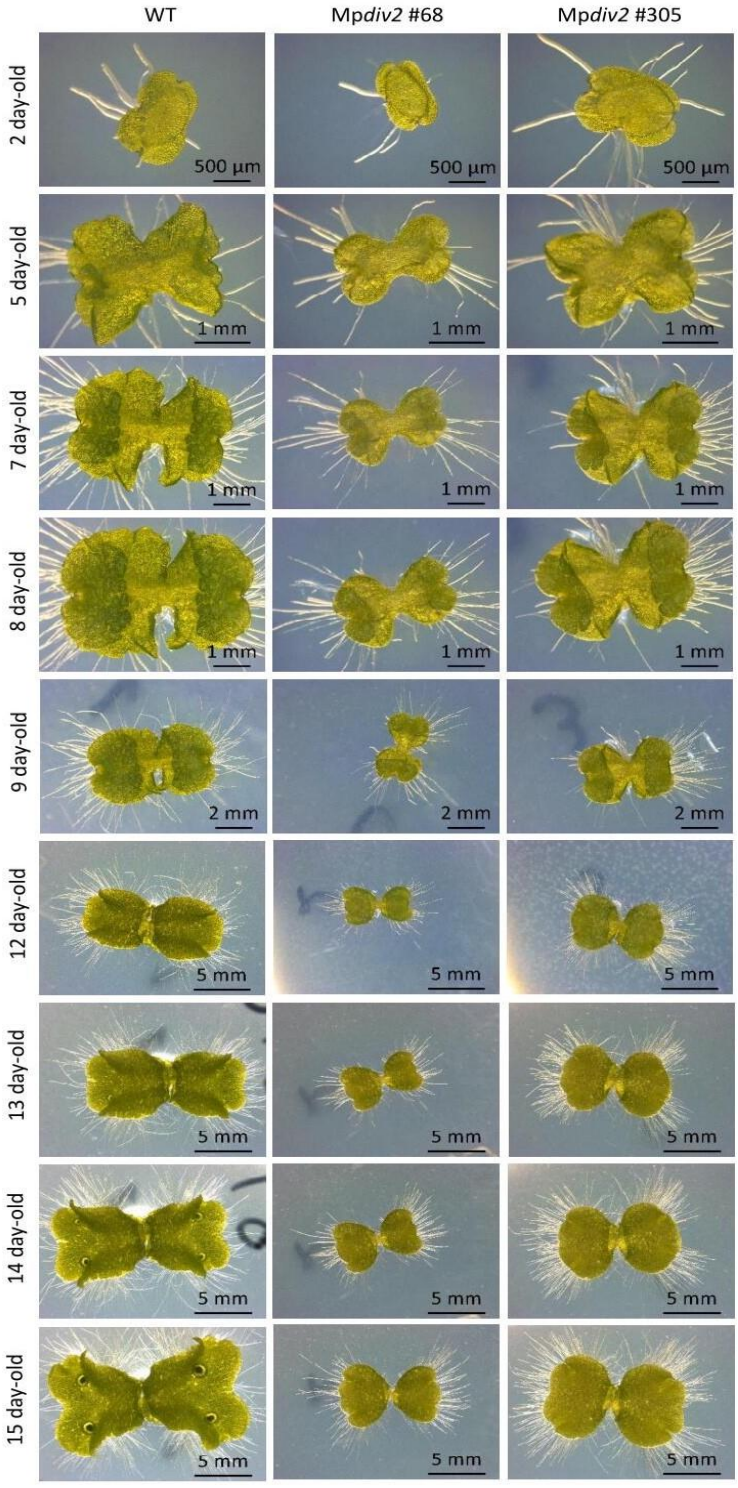

Figure 11. Visual analysis of Mp*div2* knockout mutation in *M. polymorpha* plants – Wild type plants and Mp*div2* knockout plant lines #68 and #305 were grown and observed over the first 15 days of development after gemmae propagation. Plants were grown on half strength Gamborg B5 medium under long-day conditions (16 h light/ 8 h dark) at 20 °C and with light intensity between 40-45 µmol·m^{-2} s^{-1}. Scale bar of: 0.5 mm for day 2; 1 mm from day 5 to 8; 2 mm for day 9; 5 mm from day 12 to 15.

#305 and 13-day old WT plants (**Figure 11**). In general, thallus and organ development are retarded.

3.1.1.2. Mp*drif* knockout phenotype

M. polymorpha Mp*drif* knockout plants were previously selected by sequencing of the Mp*DRIF* genomic sequence. Line #15 was successfully sequenced and had a deletion that affected the second and third introns, removing most of the first protein domain while line #23 was further analysed due to phenotypical alterations in plant shape, despite failed sequencing.

The quantitative analysis (**Figure 12**) of the Mp*drif* knockout mutants showed no apparently significant differences in size in comparison with wild type plants. The visual analysis (**Figure 13**) showed that during the first 9 days of development, no significative differences were

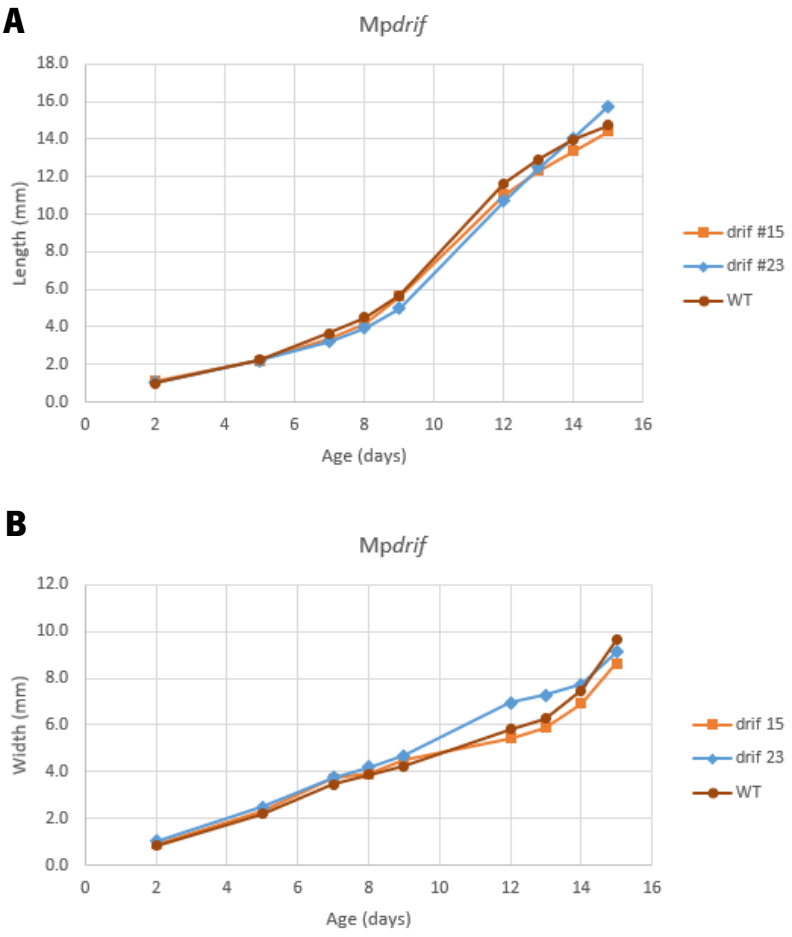

Figure 12. Analysis of the effect of Mp*drif* knockout on length and width of *M. polymorpha* plants. The thallus length and width of wild type and Mp*drif* mutant plants was measured for the first 15 days of development. **A -** The dark red line (circle) is the control that represents the wild type thallus length. The orange line (square) represents the thallus length of the Mp*drif* line #15 plants and the light blue line (diamond) represents the thallus length of Mp*drif* line #23 plants (line #15 - n=3; WT and line #23 - n=4). **B -** The dark red line (circle) is the control that represents the wild type thallus width. The orange line (square) represents the thallus width of the Mp*drif* line #15 plants and the light blue line (diamond) represents the thallus width of Mp*drif* line #23 plants (line #15 - n=3; WT and line #23 - n=4).

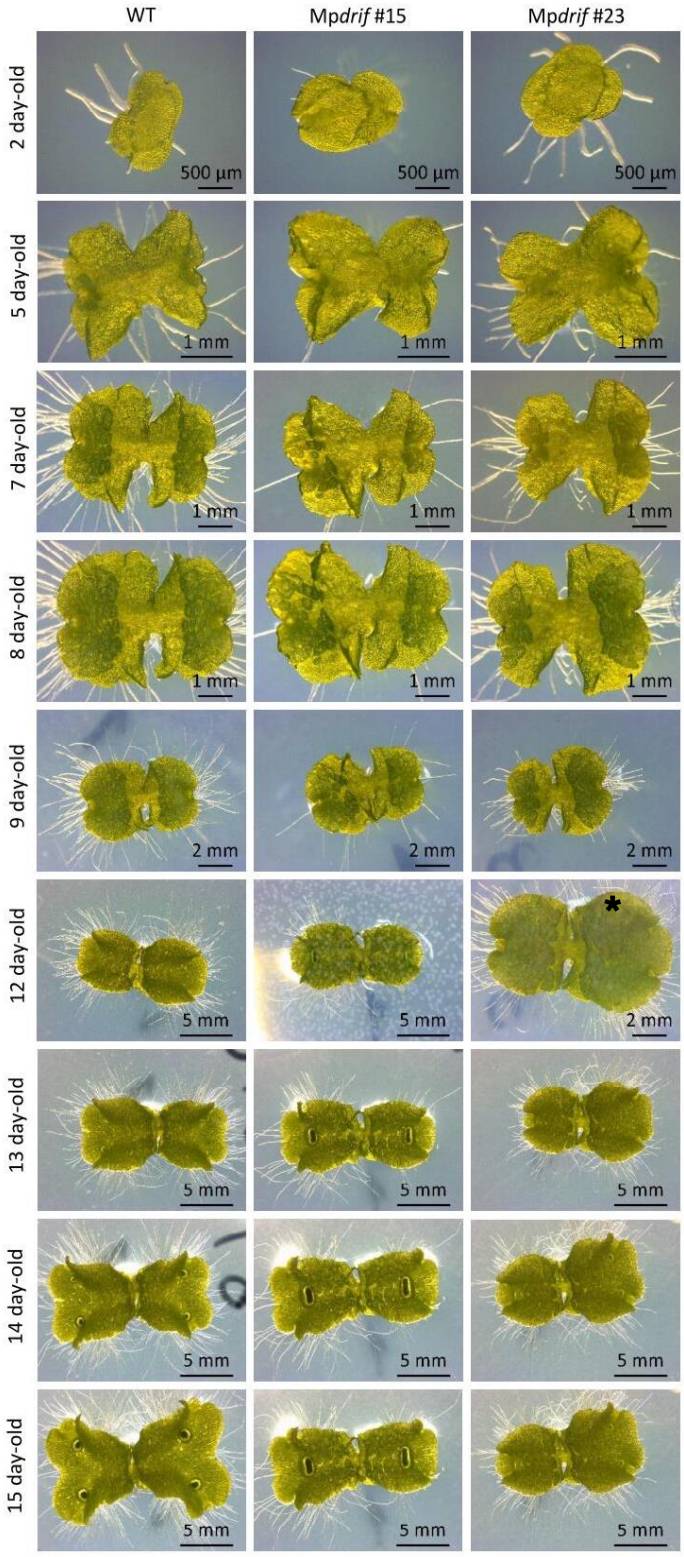

Figure 13. Visual analysis of Mp*drif* knockout mutation in *M. polymorpha* plants – Wild type plants and Mp*drif* knockout plant lines #15 and #23 were grown and observed over the first 15 days of development after gemmae propagation. Plants were grown on half strength Gamborg B5 medium under long-day conditions (16 h light/ 8 h dark) at 20 °C and with light intensity between 40-45 µmol·m^{-2} s^{-1}. Scale bar of: 0.5 mm for day 2; 1 mm from day 5 to 8; 2 mm for day 9; 5 mm from day 12 to 15 (*12 day-old Mp*drif* #23 has 2 mm scale bar).

visible between plant lines. From the 12th to the 15th day, knockout line #15 developed gemmae cups before the first bifurcation and both knockout lines show less bifurcation than wild type, leading to decreased width. Line #23 has late development of gemmae cups on the bifurcations, which themselves are late in development, but it also shows late development of a single gemmae cup on the left side, much like what occurred with line #15.

3.1.2. Overexpression analysis

Two distinct types of overexpression plant lines were developed. In one, developed in past studies, coding sequences of DDR module genes were inserted into vectors with the endogenous elongation factor 1α (Mp*EF1α*) promoter for ubiquitous expression. The resulting phenotypes are analysed in this thesis. The other type of overexpression plant line was developed during this thesis. In this one coding sequences of Mp*DIV1*, Mp*DIV2* and Mp*DRIF* were inserted into vectors with Mp*EF1α* promoters and with C-terminal tags with varied purposes (pMpGWB series of Gateway® Binary Vectors). For overexpression of proteins attached to various tags, constructs were prepared for this thesis using the pMpGWB208, pMpGWB308, pMpGWB313, pMpGWB318 and pMpGWB321 vectors.

3.1.2.1. Single gene overexpression

3.1.2.1.1. At*RAD2* overexpression phenotype analysis

As described previously, the DDR module consists of the interaction and regulation between the *DIV*, *DRIF* and *RAD* genes. *M. polymorpha* does not possess a *RAD* homolog as *RAD* is only present in gymnosperms and angiosperms (Raimundo et al., 2018). With the objective of confirming whether it was possible to establish the antagonistic effect of RAD over MpDIV in Marchantia, plants overexpressing AtRAD2 were obtained. At*RAD2* is one of the six *RAD* homologs of *A. thaliana* reported to have a phenotypic effect in *A. thaliana* (Hamaguchi et al., 2008).

Quantitative analysis (**Figure 14**) shows that At*RAD2*ox plant lines grow to be smaller in both length and width. Length especially is shorter than in wild type plants. The overexpression plants are smaller and bifurcation development is retarded. Gemmae cup development is absent in the 15 days of analysis. Visually (**Figure 15**), the phenotype is similar to that of Mp*div2* knockout mutant plants analysed. The plants are generally smaller and more rounded due to the late extension of the bifurcating thallus.

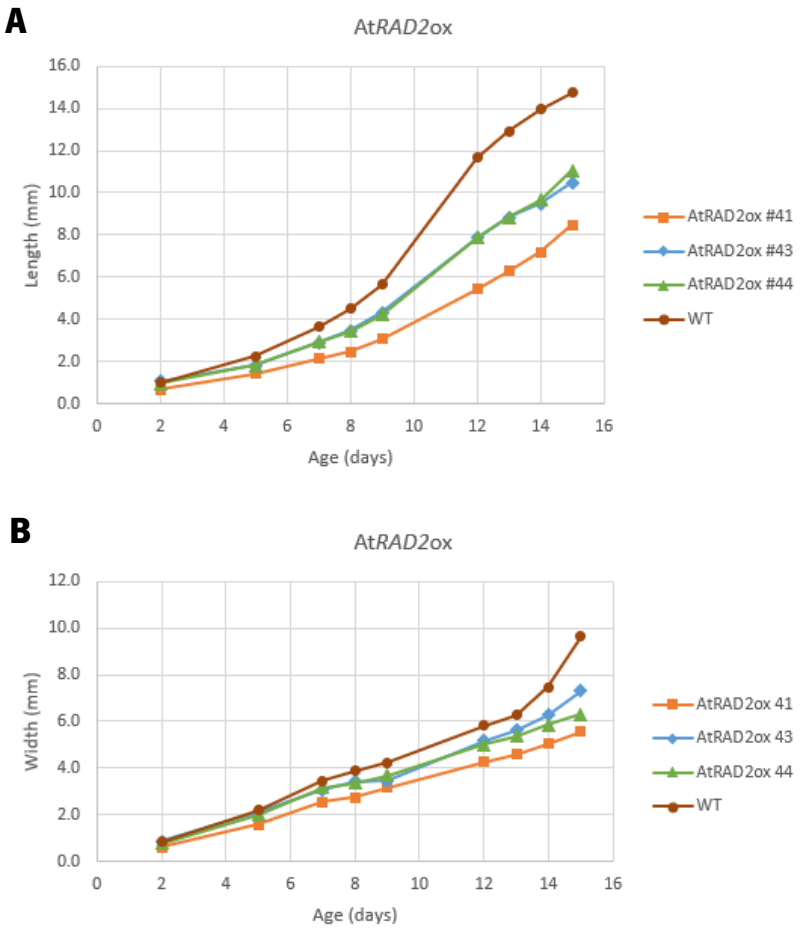

Figure 14. Analysis of the effect of At*RAD2* overexpression on length and width of *M. polymorpha* plants. The thallus length and width of wild type and At*RAD2* overexpression plants was measured for the first 15 days of development. **A -** The dark red line (circle) represents the wild type thallus length. The orange line (square) represents the thallus length of the At*RAD2*ox line #41 plants, the light blue line (diamond) represents the thallus length of the At*RAD2*ox line #43 plants and the light green line (triangle) represents the thallus length of At*RAD2*ox line #44 plants (n=4). **B -** The dark red line (circle) represents the wild type thallus width. The orange line (square) represents the thallus width of the At*RAD2*ox line #41 plants, the light blue line (diamond) represents the thallus width of the At*RAD2*ox line #43 plants and the light green line (triangle) represents the thallus width of At*RAD2*ox line #44 plants (n=4).

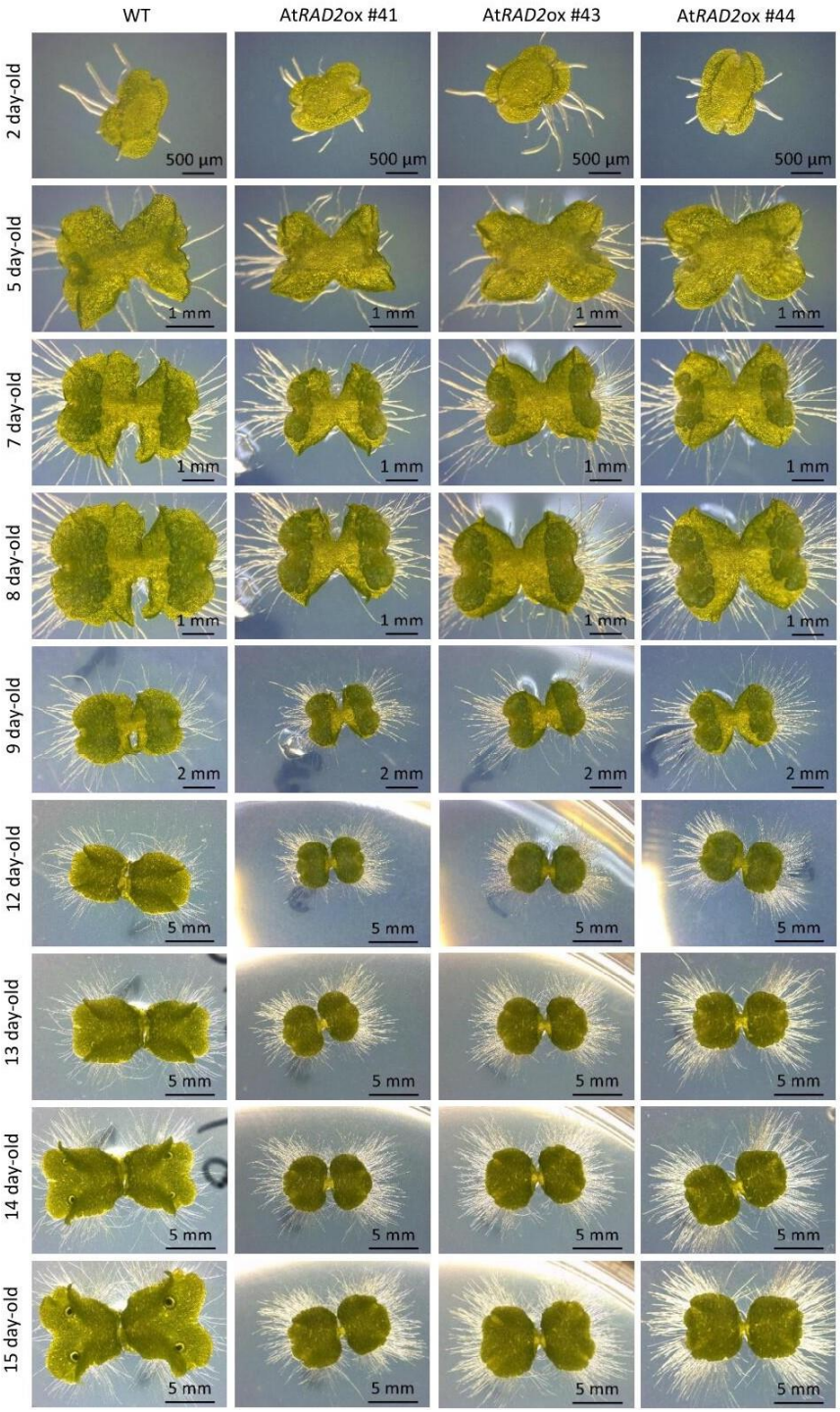

Figure 15. Visual analysis of At*RAD2* overexpression in *M. polymorpha* plants – Wild type plants and At*RAD2* overexpression plant lines #41, #43 and #44 were grown and observed over the first 15 days of development after gemmae propagation. Plants were grown on half strength Gamborg B5 medium under long-day conditions (16 h light/ 8 h dark) at 20 °C and with light intensity between 40-45 µmol·m^{-2} s^{-1}. Scale bar of: 0.5 mm for day 2; 1 mm from day 5 to 8; 2 mm for day 9; 5 mm from day 12 to 15.

3.1.2.1.2. Mp*DIV2* overexpression phenotype analysis

M. polymorpha plants were previously transformed with pMpGWB103-Mp*DIV2* for overexpression of Mp*DIV2* and screened for positive transformations.

Mp*DIV2*ox plants overall length is decreased in relation to WT plants (**Figure 16**). The decrease was more significant for line #3 and #11 and less significant for line #4. The width of lines #3 and #11 are increased in relation to WT plants while the width of plants in line #4 is near identical to WT widths. Visually (**Figure 17**), both Mp*DIV2*ox lines #3 and #11 show earlier bifurctation and increased growth of the bifurcations, leading to increased width. Line #4 and WT plant development is very similar, bifurcation beginning and developing relatively simultaneously

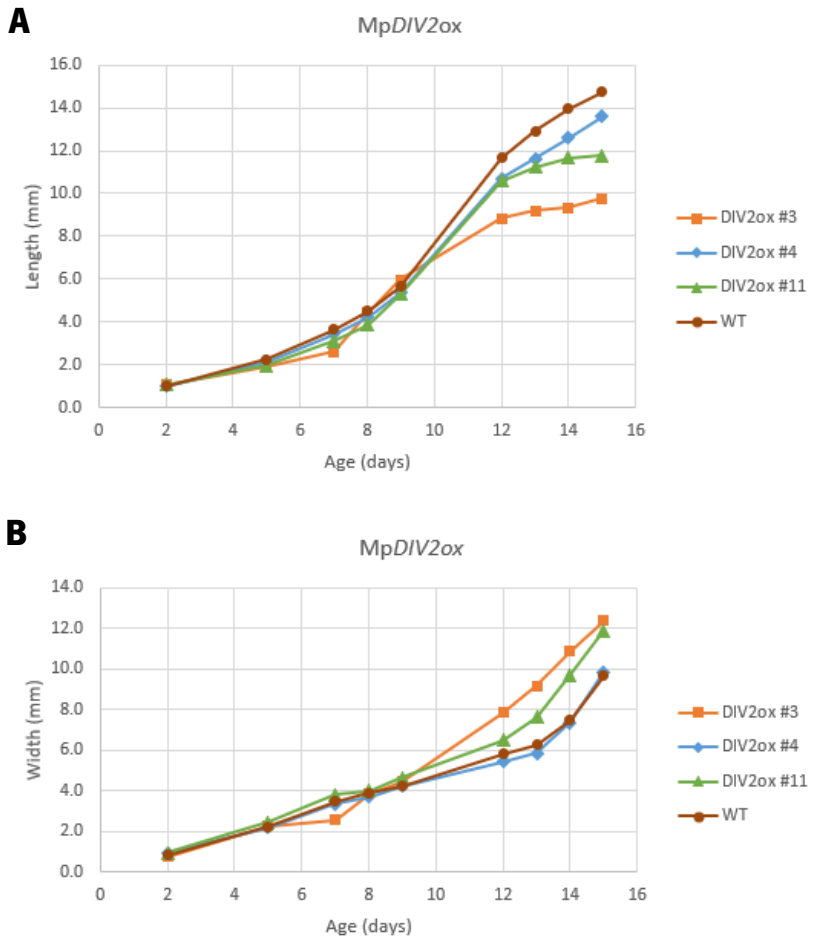

Figure 16. Analysis of the effect of Mp*DIV2* overexpression on length and width of *M. polymorpha* plants. The thallus length and width of wild type and Mp*DIV2* overexpression plants was measured for the first 15 days of development. **A -** The dark red line (circle) is the control that represents the wild type thallus length. The orange line (square) represents the thallus length of the Mp*DIV2*ox line #3 plants, the light blue line (diamond) represents the thallus length of the Mp*DIV2*ox line #4 plants and the light green line (triangle) represents the thallus length of Mp*DIV2*ox line #11 plants (line #11 - n=3; WT, line #3 and #4 - n=4). **B -** The dark red line (circle) is the control that represents the wild type thallus width. The orange line (square) represents the thallus width of the Mp*DIV2*ox line #3 plants, the light blue line (diamond) represents the thallus width of the Mp*DIV2*ox line #4 plants and the light green line (triangle) represents the thallus width of Mp*DIV2*ox line #11 plants (n=4).

and with development of gemmae cups around the 13th to 14th days, which did not occur in overexpression lines #3 and #11.

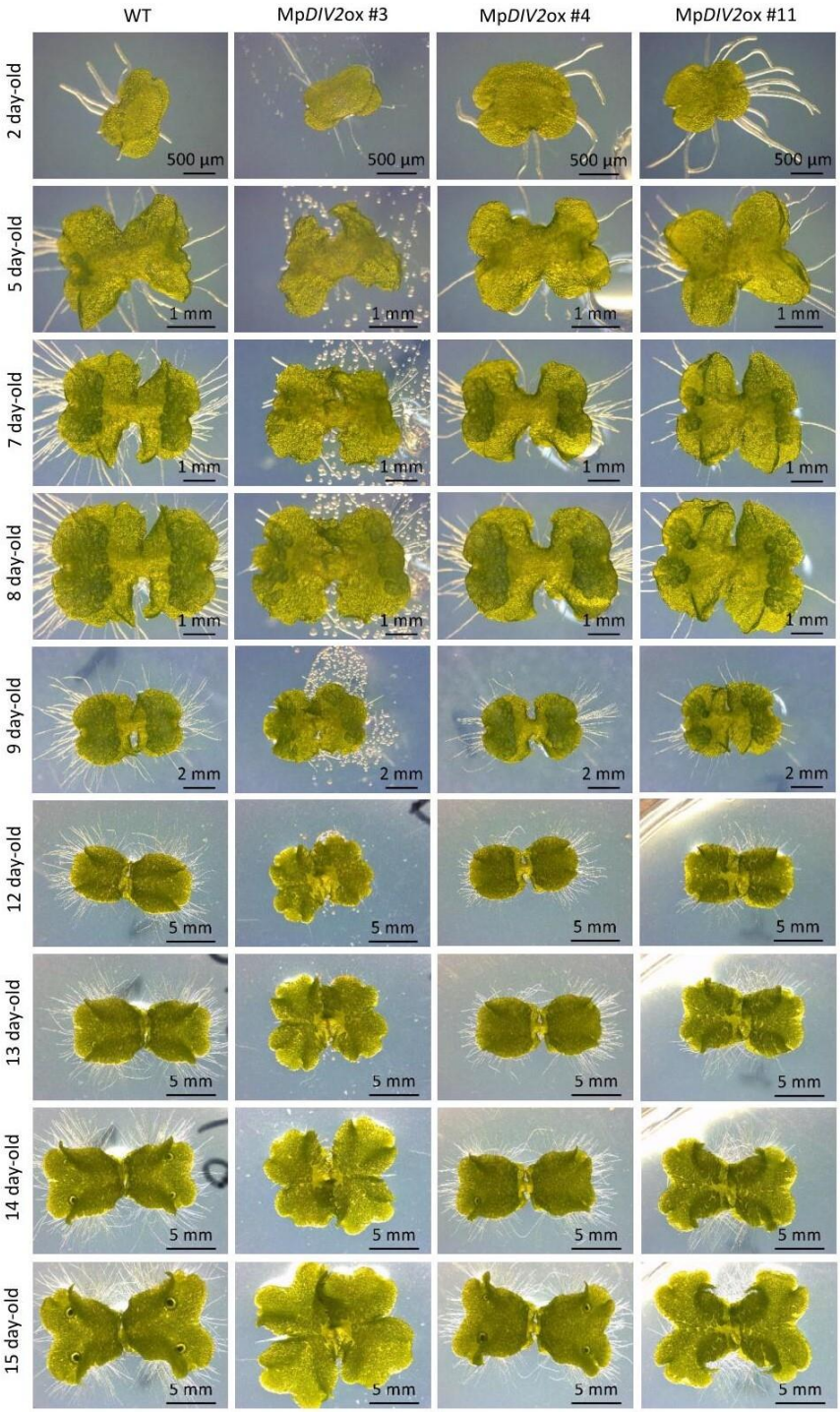

Figure 17. Visual analysis of Mp*DIV2* overexpression in *M. polymorpha* plants – Wild type plants and Mp*DIV2* overexpression plant lines #3, #4 and #11 were grown and observed over the first 15 days of development after gemmae propagation. Plants were grown on half strength Gamborg B5 medium under long-day conditions (16 h light/ 8 h dark) at 20 °C and with light intensity between 40-45 µmol·m^{-2} s^{-1}. Scale bar of: 0.5 mm for day 2; 1 mm from day 5 to 8; 2 mm for day 9; 5 mm from day 12 to 15.

3.1.2.1.3. Mp*DRIF* overexpression phenotype analysis

M. polymorpha plants were previously transformed with pMpGWB103-Mp*DRIF* for overexpression of Mp*DRIF* and screened for positive transformations.

Measurements (**Figure 20**) showed, in relation to WT plants, a slight decrease in length for line #5 and a larger decrease for line #2. The width of overexpression line plants is very similar to that of WT plants, except for the last few days of development or line #2, where width is decreased. Visually (**Figure 19**), development of Mp*DRIF*ox line #2 is retarded during the final few days analysed with slowed bifurcation. Line #5 plants present a more rounded, flattened shape to the thallus around the extremities while size and gemmae development remain similar to WT plants.

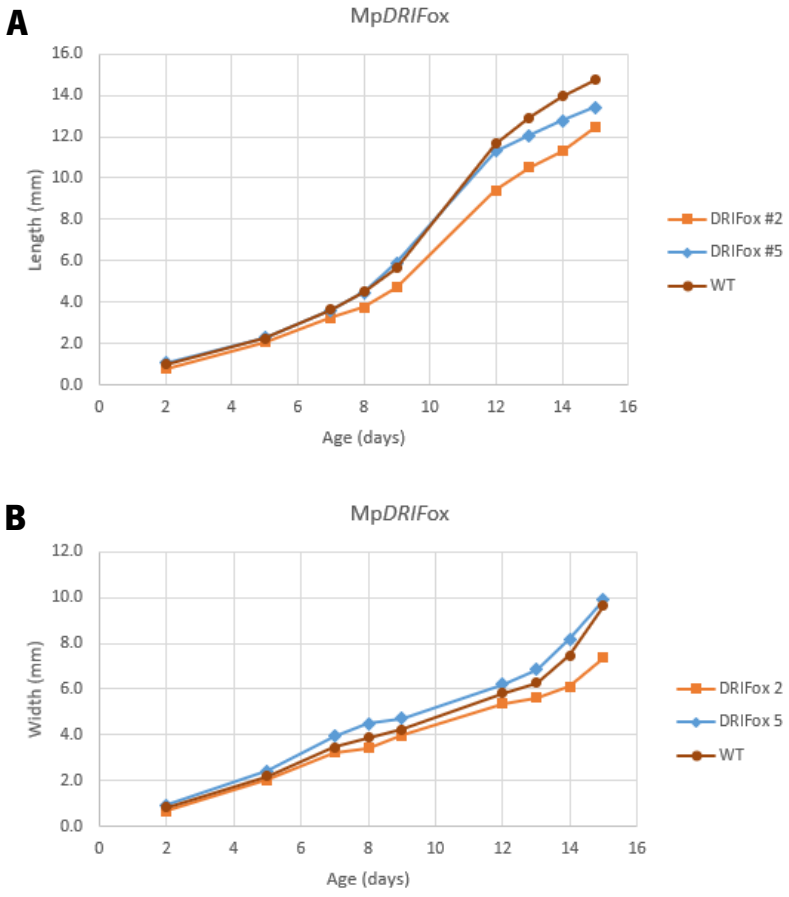

Figure 20. Analysis of the effect of Mp*DRIF* overexpression on length and width *M. polymorpha* plants. The thallus length and width of wild type and Mp*DRIF* overexpression plants was measured for the first 15 days of development. **A -** The dark red line (circle) is the control that represents the wild type thallus length. The orange line (square) represents the thallus length of the Mp*DRIF*ox line #2 plants and the light blue line (diamond) represents the thallus length of the MpDIV2ox line #5 plants (n=4). **B -** The dark red line (circle) is the control that represents the wild type thallus width. The orange line (square) represents the thallus width of the Mp*DRIF*ox line #2 plants and the light blue line (diamond) represents the thallus width of the Mp*DRIF*ox line #5 plants (n=4).

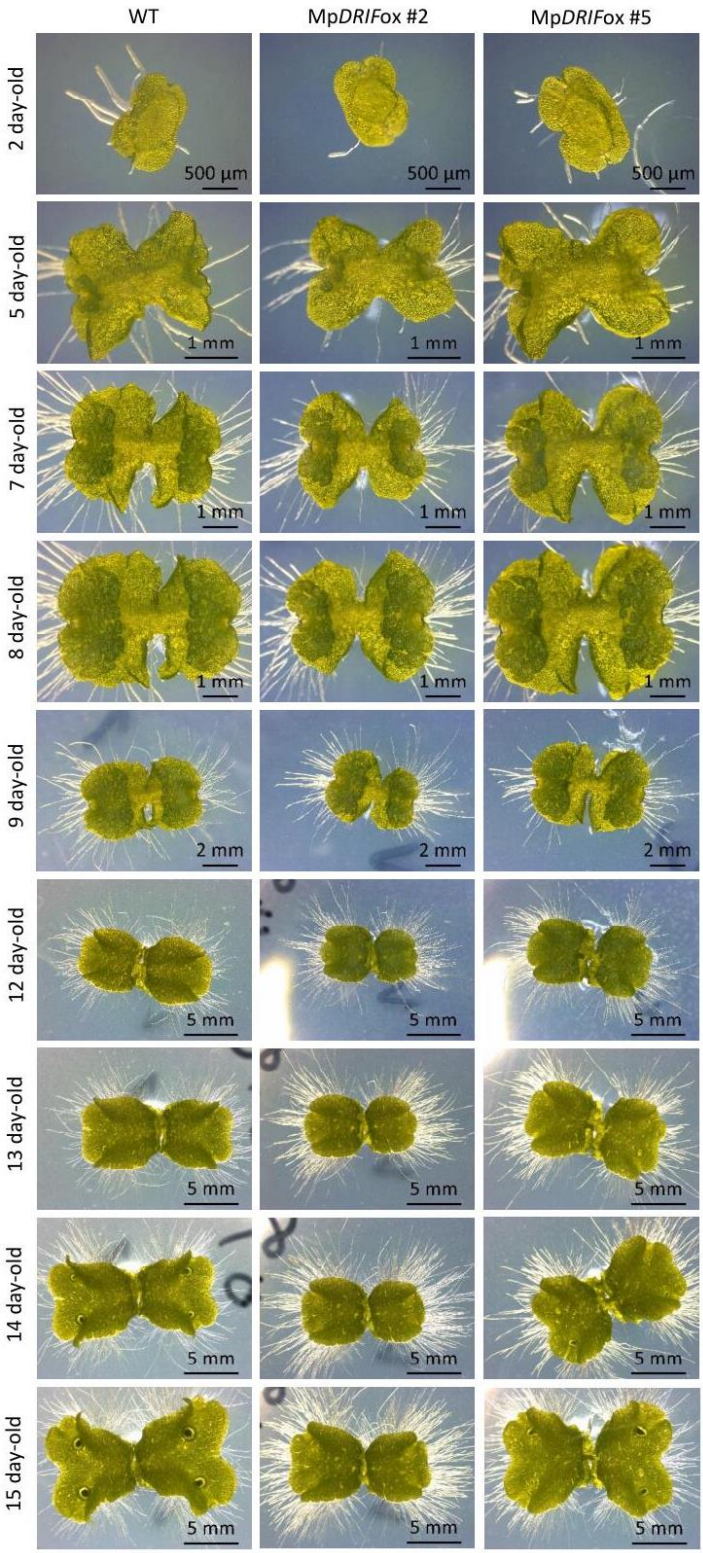

Figure 19. Visual analysis of MpDRIF overexpression in Marchantia polymorpha plants – Wild type plants and MpDRIF overexpression plant lines #2 and #5 were grown and observed over the first 15 days of development after gemmae propagation. Plants were grown on half strength Gamborg B5 medium under long-day conditions (16 h light/ 8 h dark) at 20 °C and with light intensity between 40-45 µmol·m^{-2} s^{-1}. Scale bar of: 0.5 mm for day 2; 1 mm from day 5 to 8; 2 mm for day 9; 5 mm from day 12 to 15.

3.1.2.2. Overexpression with tags

As stated previously, overexpression *Marchantia* lines were generated using some of the plasmids from the pMpGWB series of *M. polymorpha* gateway technology vectors (Annex B).

MpDIV1/2 and MpDRIF coding sequences were cloned into pMpGWB208 and pMpGWB308 with the aim of studying their cellular sub- localization. pMpGWB208 and pMpGWB308 both have the proMp*EF1α* and C-terminal Citrine tags, only differing in the plant antibiotic resistance where pMpGWB208 carries resistance to gentamycin and pMpGWB308 carries resistance to chlorsulfuron. The two different resistances were used to facilitate selection after cross. MpDIV1/2 and MpDRIF coding sequences were also cloned in vector pMpGWB313 that has the proMp*EF1α* and a C-terminal glucocorticoid receptor (GR) tag. The aim was to generate *M. polymorpha* mutant lines where the proteins of interest shuttle between the cytoplasm and nucleus could be tightly controlled.

pMpGWB318 has the proMp*EF1α* and a C-terminal modified EAR motif plant-specific repression domain showing strong repression activity (SRDX). Genes of interest were cloned into this plasmid with the aim of generating *M. polymorpha* mutant lines where the function of the proteins of interest would be replaced with strong repressive activity. pMpGWB321 has the proMp*EF1α* and a C-terminal SRDX domain and GR tag. Genes of interest were cloned into this plasmid with the aim of generating *M. polymorpha* mutant lines where the function of the proteins of interest would be replaced with strong repressive activity tightly controlled by the GR system.

As stated, the vectors have C-terminal tags, and thus the gene coding sequences inserted couldn't have a stop codon, otherwise the expressed fusion protein would not include the tag. The Mp*DRIFnonstop* coding sequence had been previously inserted into the pDONR201 gateway vector while for the Mp*DIV1* and Mp*DIV2* coding sequences, reverse primers were prepared to amplify gene sequences without a stop codon.

3.1.2.2.1. Mp*DRIF* overexpression cloning procedures

Mp*DRIFns* (Mp*DRIF* coding sequence without stop codon), individual LR recombination reactions were carried out between the pDONR201-Mp*DRIFns* construct and pMpGWB308, pMpGWB313, pMpGWB318 and pMpGWB321 and the product of the reactions was transformed into *E. coli*. Transformations were confirmed via colony PCR and gel electrophoresis with an expected sequence size of around 1000 bp (**Figure 20**).

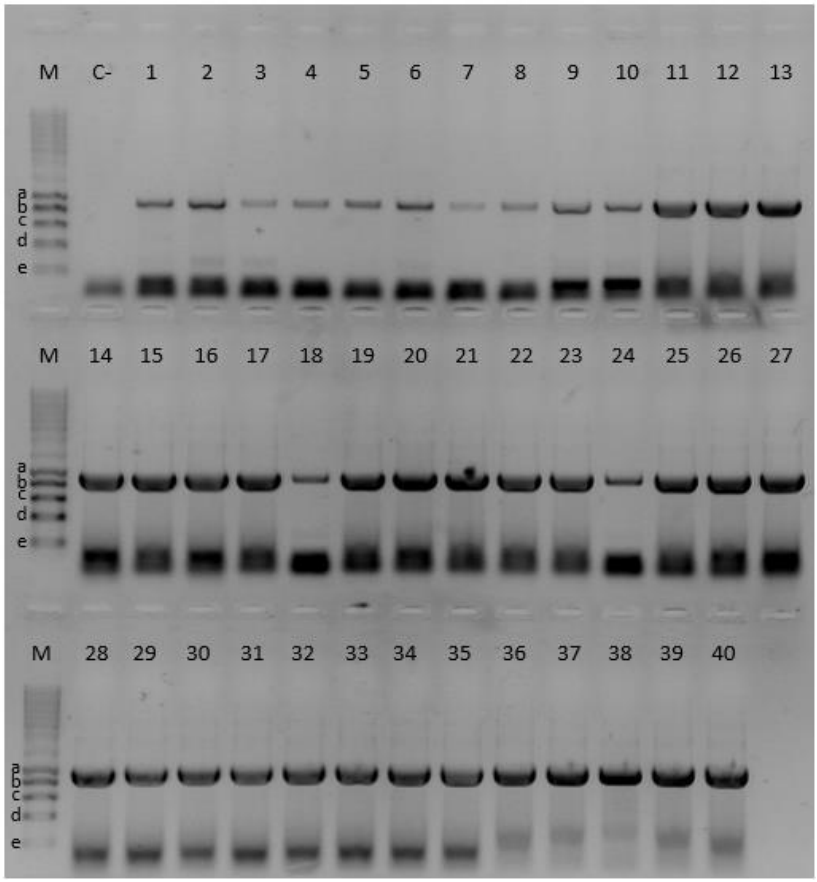

Figure 20. Colony PCR of *E. coli* transformed with LR reaction product of Mp*DRIFns* cloned into pMpGWB vectors. pDONR201-Mp*DRIFns* was the donor construct used. Mp*DRIFns* was recombined into pMpGWB308, pMpGWB313, pMpGWB318 and into pMpGWB321. Constructs were transformed into DH10-β *E. coli*. Positive transformations confirmed via colony PCR with primers 556 and 557 por Mp*DRIF*. Electrophoresis run at 120 V with 1.0% agarose concentration. **C-** - negative control; **1-10** – pMpGWB308-Mp*DRIFns* transformed colonies; **11-20** – pMpGWB313-Mp*DRIFns* transformed colonies; **21-30** – pMpGWB318-Mp*DRIFns* transformed colonies; **31-40** – pMpGWB321-Mp*DRIFns* transformed colonies; **M** – Molecular marker (NZYDNA Ladder III, Nzytech): a – 1000 bp; b – 800 bp; c – 600 bp; d – 400 bp; e – 200 bp.

A miniprep was then conducted to extract the successfully recombined pMpGWB constructs, which were then transformed into *Agrobacterium tumefaciens* for plant transformation. Transformations were confirmed via colony PCR and gel electrophoresis with an expected sequence size of around 1000 bp (**Figure 21**). Positively transformed *A. tumefaciens* were selected and used to transform *M. polymorpha* plants.

3.1.2.2.2. Mp*DIV1* and Mp*DIV2* overexpression cloning procedures

Utilizing appropriate primers, Mp*DIV1ns* and Mp*DIV2ns* coding sequences were amplified from a cDNA library and confirmed via gel electrophoresis (**Figure 22 - A**). The sequences, with an expected sequence size of around 1000 bp throughout the process, were then submitted to an *att*B PCR as preparation for gateway cloning procedures, which was analysed by gel electrophoresis (**Figure 22 - B**). Once confirmed, the product was purified and a BP recombination reaction was

carried out with the sequences, recombining them into the pDONR201 vector. The constructs were transformed into *E. coli* and positive recombinants were confirmed by colony PCR and gel electrophoresis (**Figure 22 - C**).

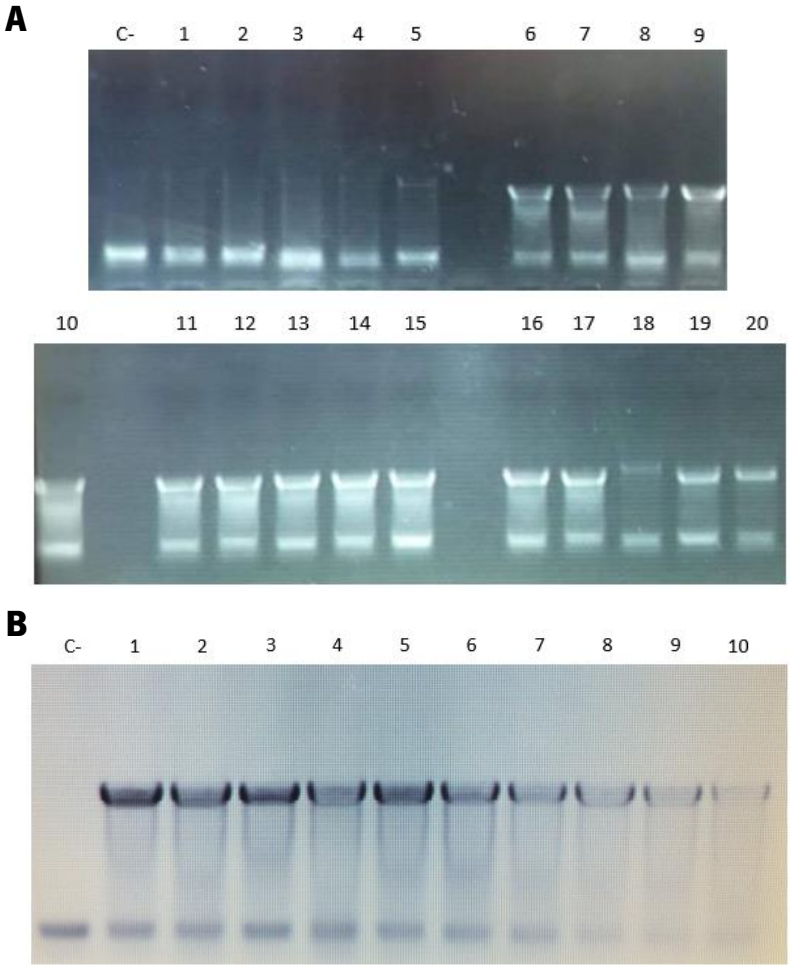

Figure 21. Colony PCR of *A. tumefaciens* transformated with pMpGWB308-Mp*DRIFns*, pMpGWB313-Mp*DRIFns*, pMpGWB318-Mp*DRIFns* and pMpGWB321-Mp*DRIFns*. Constructs were transformed into *A. tumefaciens*. Positive transformations confirmed via colony PCR with primers 556 and 557 for Mp*DRIF*. Electrophoresis run at 120 V with 1.0% agarose concentration. **A: C-** - negative control; **1-5** – pMpGWB308-Mp*DRIFns* transformed colonies; **6-10** – pMpGWB313-Mp*DRIFns* transformed colonies; **11-15** – pMpGWB318-Mp*DRIFns* transformed colonies; **16-20** – pMpGWB321-Mp*DRIFns* transformed colonies. **B:** Repetition of colony PCR for pMpGWB308-Mp*DRIFns* transformed *A. tumefaciens*. **C-** - negative control; **1-10** – pMpGWB308-Mp*DRIFns* transformed colonies.

A miniprep was carried out for the positive transformants and LR recombination reactions were performed between the extracted constructs (pDONR201- Mp*DIV1ns* and pDONR201-Mp*DIV2ns*) and the target pMpGWB plasmids (pMpGWB208, 313 and 318), the reaction product was then transformed into *E. coli*. Positive transformations were confirmed by colony PCR and gel electrophoresis with an expected sequence size of around 2700 bp (**Figure 24**). No positive transformations were achieved of the pMpGWB313-Mp*DIV1ns* construct and so the procedures were continued without this construct. The constructs were extracted from the positively

transformed colonies by miniprep and transformed into *A. tumefaciens*, which were then confirmed by colony PCR and gel electrophoresis still with with an expected sequence size of around 2700 bp (**Figure 23**). Positively transformed *A. tumefaciens* were selected and used to transform *M. polymorpha* plants.

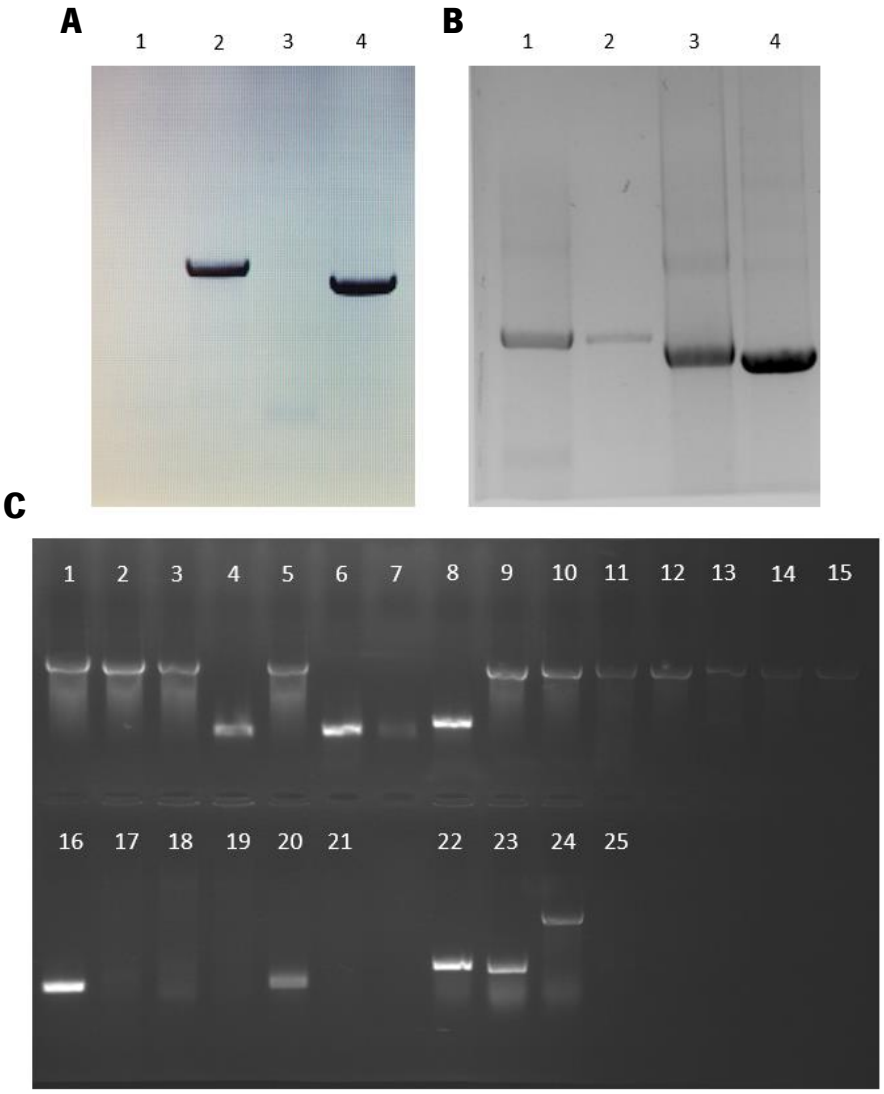

Figure 22. Preparation of Mp*DIV1ns* and Mp*DIV2ns* sequences with *att*B regions and colony PCR of *E. coli* transformed with BP product. A: Sequence amplification from cDNA library template. Primers 507 and 513 for Mp*DIV1ns* and 509 and 514 for Mp*DIV2ns* were used. **1** – Negative control with Mp*DIV1ns* primers; **2** - Mp*DIV1ns* sequence amplification; **3** – Negative control with Mp*DIV2ns* primers; **4** – Mp*DIV2ns* sequence amplification. **B:** PCR to add *att*B regions to Mp*DIV1ns* and Mp*DIV2ns* sequences. Primers Qs190 and Qs191 were used. **1** – Mp*DIV1ns* sequence with *att*B regions; **2** – Mp*DIV1ns* sequence amplification; **3** – Mp*DIV2ns* sequence with *att*B regions; **4** – Mp*DIV2ns* sequence amplification. **C:** BP product transformation colony PCR. Primers 509 and 514 for Mp*DIV2ns* 507 and 513 for Mp*DIV1ns* were used. **1-20** – Mp*DIV2ns* amplification from pDONR201-Mp*DIV2ns* transformed colonies; **21** – negative control with Mp*DIV2ns* primers; **22-24** – Mp*DIV1ns* amplification from pDONR201-Mp*DIV1ns* transformed colonies. **25** – negative control with Mp*DIV1ns* primers.

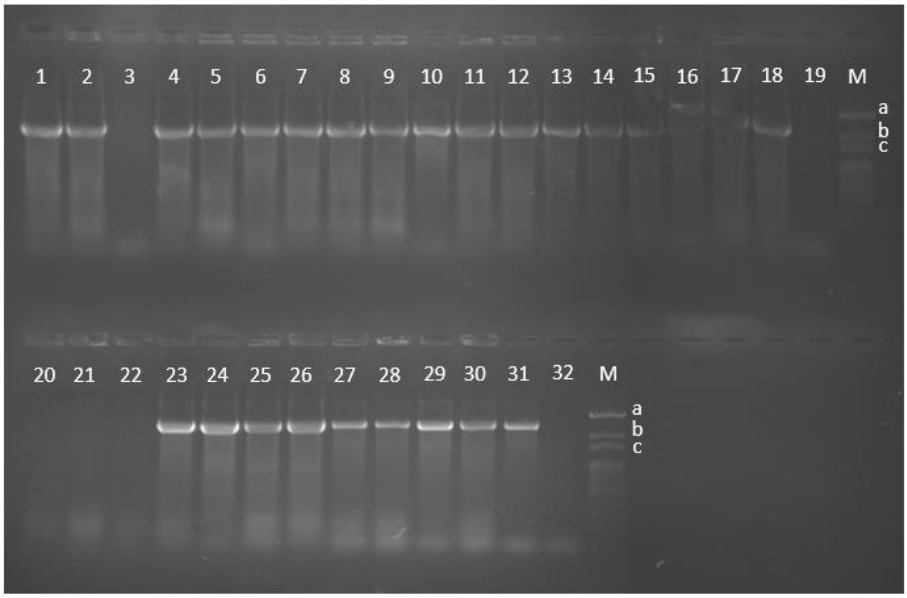

Figure 24. Colony PCR of *E. coli* transformed with LR products of Mp*DIV1* and Mp*DIV2* cloning into pMpGWB vectors. Primers 913 and 514 were used for Mp*DIV2* and primers 913 and 513 were used for Mp*DIV1*. **1-6** – pMpGWB208-Mp*DIV2ns* transformed colonies; **7-12** – pMpGWB313-Mp*DIV2ns* transformed colonies; **13-18** – pMpGWB318-Mp*DIV2ns* transformed colonies; **19** – negative control with Mp*DIV2* primers; **20-25** – pMpGWB208-Mp*DIV1ns* transformed colonies; **26-31** – pMpGWB318-Mp*DIV1ns*; **32** - negative control with Mp*DIV1* primers; **M** – Molecular marker (NZYDNA Ladder III, Nzytech): a – 3000 bp*; b* – 2500 bp; c – 2000 bp.

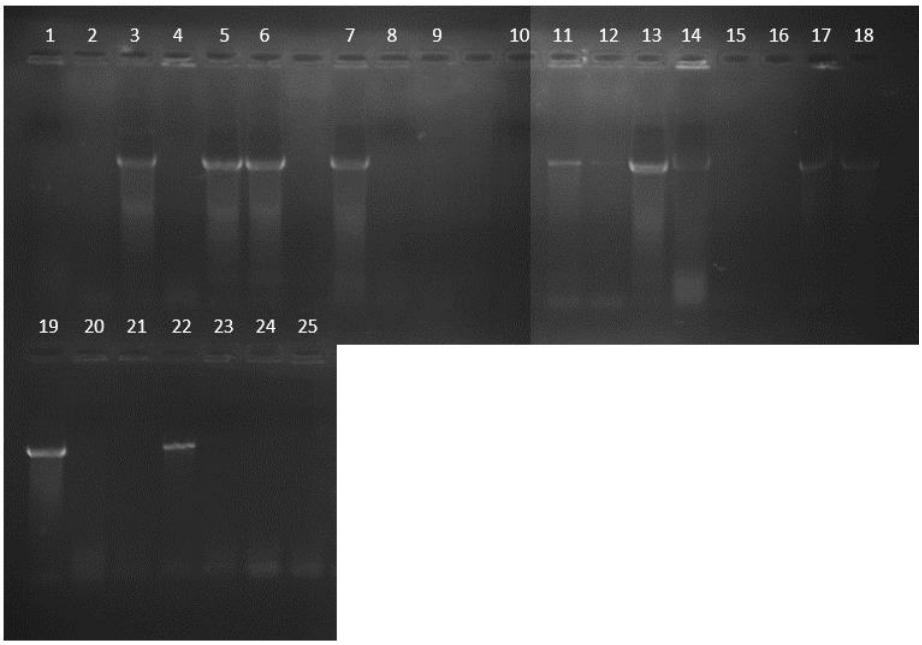

Figure 23. Colony PCR of *A. tumefaciens* transformed with LR products of Mp*DIV1* and Mp*DIV2* cloning into pMpGWB vectors. *A. tumefaciens* transformants colony PCR electrophoresis gel. Primers 913 and 514 were used for Mp*DIV2* and primers 913 and 513 were used for Mp*DIV1*. **1-6** – pMpGWB208-Mp*DIV1ns* transformed colonies; **6-7** – pMpGWB318-Mp*DIV1ns*; **8** – negative control with Mp*DIV1* primers; **9-15** – pMpGWB208-Mp*DIV2ns* transformed colonies; **16-18** – pMpGWB313-Mp*DIV2ns* transformed colonies; **19-24** – pMpGWB318-Mp*DIV2ns* transformed colonies; **25** – negative control with Mp*DIV2* primers.

3.2. Phylogenetic analysis

One of the main objectives of this thesis was to unveil the ancestral function of *DIV* and *DRIF*. By learning about the simpler, ancestral function of the genes, context would be created that would allow for better understanding of the eventual, more complex functions these genes have evolved in vascular plants, such as angiosperms.

Phylogenetic analysis was performed to uncover the early evolutionary history of the *DIV* and *DRIF* genes by analysing homolog DIV and DRIF sequences from various eukaryotic algal and protist species.

The eukaryotic tree of life has been traditionally divided into several supergroups (usually 5 to 8). In recent years the organization of these supergroups has changed significantly as new phylogenetic and classification techniques were developed (Burki et al., 2020). For the current thesis, the supergroups of interest are the supergroup Archaeplastida, that includes all plants and algae, Cryptista, Haptista, the clade TSAR and the phylum euglenozoan (**Figure 6**). The species of which genomes were used in a blast search belong to these groups (Grigoriev et al., 2020).

3.2.1. Sequence retrieval

Protein sequences homologous to MpDIV2 and MpDRIF were retrieved from the PhycoCosm resource for databases of algal genomes using the built-in BLAST tool. The complete protein sequences of MpDIV2 and MpDRIF were used as the query sequences within the *tblastn* alignment program, comparing the protein query sequences against translated nucleotide data from the genomic databases. The retrieved sequences were organized in FASTA format into files for DIV and DRIF homologs and aligned using MUSCLE. BLAST results revealed some sequences with only the SHAQKYF domain of DIV proteins (MYBII). These and sequences that were repeated were deleted from the list. The protein sequences utilized in the alignment are presented in Annex C.

DIV and DRIF homologs were both found in Viridiplantae, the group including all green algae and terrestrial plants (**Table 2**). The only major group, from within the green lineage, that was not represented was the class Ulvophyceae, which belongs to the core chlorophytes. DIV homologs were found outside the Viridiplantae group, in a species belonging to the Glaucophyta group and in two species belonging to the Cryptophyceae class within the Cryptista group.

Interestingly, no DIV or DRIF homologs were found in the red algae group, Rhodophyta, which belongs to the larger group Archaeplastida, along with Viridiplantae and Glaucophyta.

Table 2. List of DIV and DRIF protein homologs found in algae matched with corresponding species, class, and group. Blank cells with "------" indicate species where a homolog of the gene was not found. * - a homolog was found but excluded because of apparent sequencing issues

Group	Class	Species	DIV	DRIF
Prasinophytes	Mamiellophyceae	*Micromonas commoda*	MiccomDIV	MiccomDRIF
		Micromonas pusilla	MicpuDIV	MicpuDRIF
		Ostreococcus lucimarinus	OstluDIV	------
		Ostreococcus sp. RCC809	OstrcDIV	------
		Ostreococcus tauri	OsttaDIV	------
Core chlorophytes	Trebouxiophyceae	*Auxenochlorella protothecoides*	AuxprDIV	AuxprDRIF
		Botryococcus braunii	BotrbrauDIV	BotrbrauDRIF
		Chlorella sp. A99	------	ChloA99DRIF
		Chlorella sorokiniana	ChlosoDIV	ChlosoDRIF
		Chlorella variabillis	ChlvarDIV	ChlvarDRIF
		Coccomyxa subellipsoidea	CocDIV	CocDRIF
		Micractinium conductrix	------	MiccoDRIF
		Picochlorum renovo	PicreDIV	PicreDRIF
		Picochlorum soloecismus	PicsoDIV	PicsoDRIF
		Symbiochloris reticulata	SymretDIV	SymretDRIF
		Tetraselmis striata	TetstrDIV	TetstrDRIF
		Trebouxia sp. A1-2	------	TrebDRIF
	Chlorophyceae	*Chlamydomonas eustigma*	ChleuDIV1, ChleuDIV2	ChleuDRIF1, ChleuDRIF2
		Chlamydomonas reinhardtii	ChlreDIV	ChlreDRIF
		Chlamydomonas schloesseri	ChlscDIV	ChlscDRIF
		Chromochloris zofingiensis	ChrzofDIV	ChrzofDRIF
		Dunaliella salina	DunsalDIV	DunsalDRIF
		Edaphoclamys debaryana	EdadeDIV	EdadeDRIF
		Gonium pectorale	GonpecDIV	GonpecDRIF
		Tetrabaena socialis	TetsoDIV	TetsoDRIF
		Volvox carteri	VolcaDIV	VolcaDRIF
	Chloropicophyceae	*Chloropicon primus*	ChlpriDIV	ChlpriDRIF
Prasinodermophytes	Prasinodermophyceae	*Prasinoderma coloniale*	PracoDIV	PracoDRIF
Charophytes	Charophyceae	*Chara braunii*	ChabraDIV	*
	Chlorokybophyceae	*Chlorokybus atmophyticus*	ChlatDIV	ChlatDRIF
	Klebsormidiophyceae	*Klebsormidium nitens*	KlenitDIV	KlenitDRIF
	Mesostigmatophyceae	*Mesostigma viride*	MesovirDIV	MesovirDRIF
	Zygnematophyceae	*Mesotaenium endlicherianum*	MesenDIV	MesenDRIF
Glaucophyta	Glaucocystophyceae	*Cyanophora paradoxa*	CyaparDIV	------
Cryptista	Cryptophyceae	*Cryptophyceae sp.* CCMP2045	CryptoDIV	------
		Guillardia theta	GuithDIV1, GuithDIV2	------

Within the Mamiellophyceae class there are two distinct genus, *Ostreococcus* and *Micromonas*, in which DIV homologs were found but no DRIF homolog was found and within the Trebouxiophyceae class, a DRIF homolog was found in the species *Micractinium conductrix* and *Trebouxia sp.* while no DIV was found. These are the only occurrences of one homolog without the other within a certain species or genus in the Viridiplantae group. Outside this group, as stated above, DIV homologs were found in the *Cyanophora* genus of Glaucophyta and in the Cryptista group, where no DRIF homologs were found.

3.2.2. Alignment and phylogenetic tree construction

Once homologous proteins were retrived, the sequences were once again aligned using MUSCLE and parts of the sequences of DIV and DRIF protein homologs that did not belong to the functional domains of the protein were removed. As mentioned above, the functional domains of DIV proteins are the MYBI domain, responsible for protein interactions, and the MYBII domain, responsible for binding to DNA. The functional domains of DRIF proteins are the MYB domain, responsible for protein interaction with DIV proteins, and the DUF3755 domain, the exact function of which is unknown but has been reported to interact with proteins. The final alignment used sequences were the amino acids before, between and after the domains was removed. The partial sequences were composed of the MYBI and MYBII domains in the case of DIV sequences and the conjoined MYB and DUF3755 domains in the case of DRIF sequences.

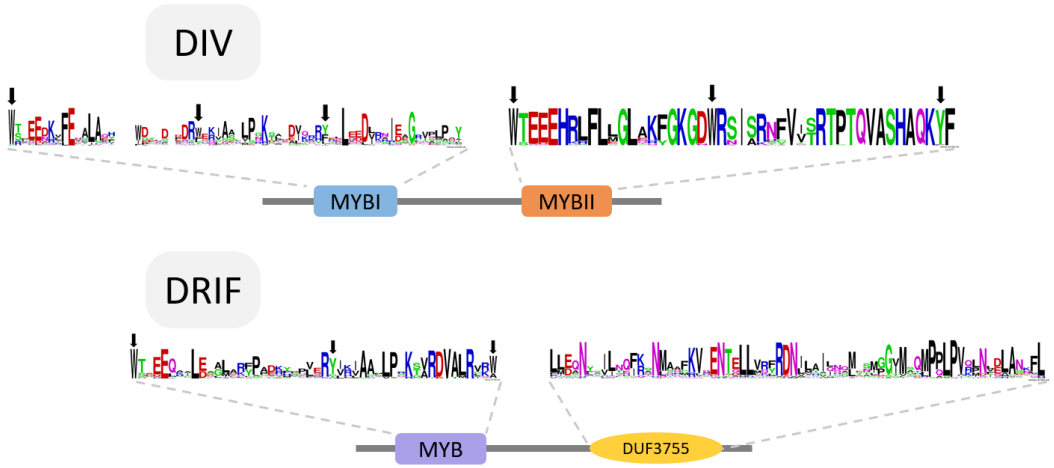

Figure 27. Structure of the DIV and DRIF protein families. Schematic representation of the general structure of the DIV and DRIF proteins from several species belonging to the Cryptophyta, Glaucophyta and Viridiplantae groups and from *M. polymorpha*. Domains are presented corresponding to the sequence logos that were generated based on the alignment depicted in Annex D. The degree of certainty of each amino acid position, calculated with conservation, is indicated by the height of the respective symbol. The conserved aromatic residues typical of the MYB domain topology are signalled with black arrows.

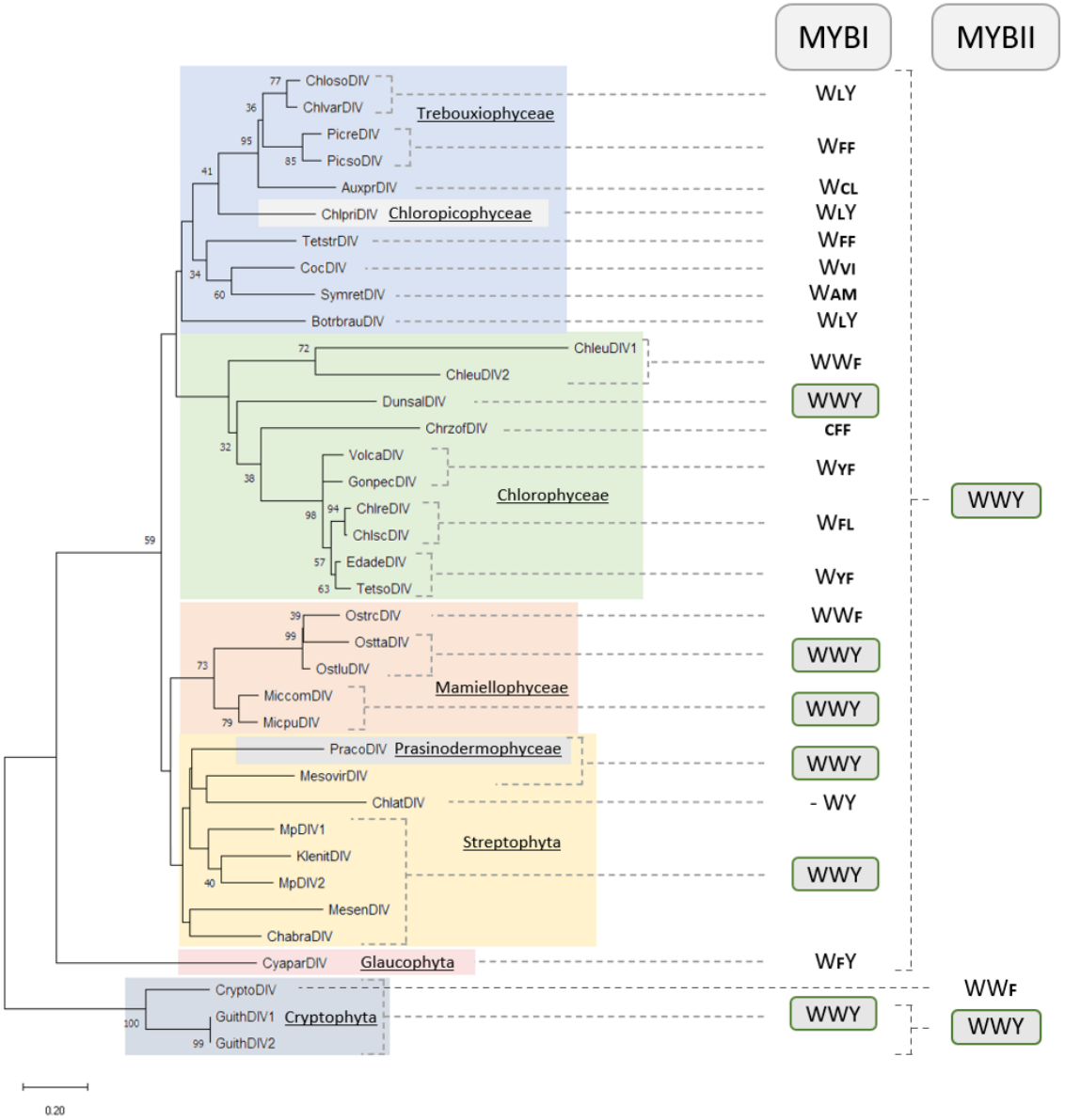

Figure 26. Evolutionary history and relationships of part of the DIV family of proteins. The phylogenetic tree was constructed from the alignment of protein domain sequences of DIV homologs from different algal species and *M. polymorpha* (*Micromonas commoda*, MiccomDIV; *Micromonas pusilla*, MicpuDIV; *Ostreococcus lucimarinus*, OstluDIV; *Ostreococcus sp.* RCC809, OstrcDIV; *Ostreococcus tauri*, OsttaDIV; *Auxenochlorella protothecoides*, AuxprDIV; *Botryococcus braunii*, BotrbrauDIV; *Chlorella sorokiniana*, ChlosoDIV; *Chlorella variabillis*, ChlvarDIV; *Coccomyxa subellipsoidea*, CocDIV; *Picochlorum renovo*, PicreDIV; *Picochlorum soloecismus*, PicsoDIV; *Symbiochloris reticulata*, SymretDIV; *Tetraselmis striata*, TetstrDIV; *Chlamydomonas eustigma*, ChleuDIV1 and ChleuDIV2; *Chlamydomonas reinhardtii*, ChlreDIV; Chlamydomonas schloesseri, ChlscDIV; *Chromochloris zofingiensis*, ChrzofDIV; *Dunaliella salina*, DunsalDIV; *Edaphoclamys debaryana*, EdadeDIV; *Gonium pectorale*, GonpecDIV; *Tetrabaena socialis*, TetsoDIV; *Volvox carteri*, VolcaDIV; *Chloropicon primus*, ChlpriDIV; *Prasinoderma coloniale*, PracoDIV; *Chara braunii*, ChabraDIV; *Chlorokybus atmophyticus*, ChlatDIV; *Klebsormidium nitens*, KlenitDIV; *Mesostigma viride*, MesovirDIV; *Mesotaenium endlicherianum*, MesenDIV). The conservation of the characteristic aromatic residues of the MYB domains is presented for each sequence. Two Tryptophan and one Tyrosine (WWY) are the expected amino acids. Smaller font size indicates the amino acid replacing the aromatic residue (F, Phenylalanine; L, Leucine; C, Cysteine; A, Alanine; M, Methionine). For the MYBI ChlatDIV, "–" indicates that in the position of the first Tryptophan there is no amino acid. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site.

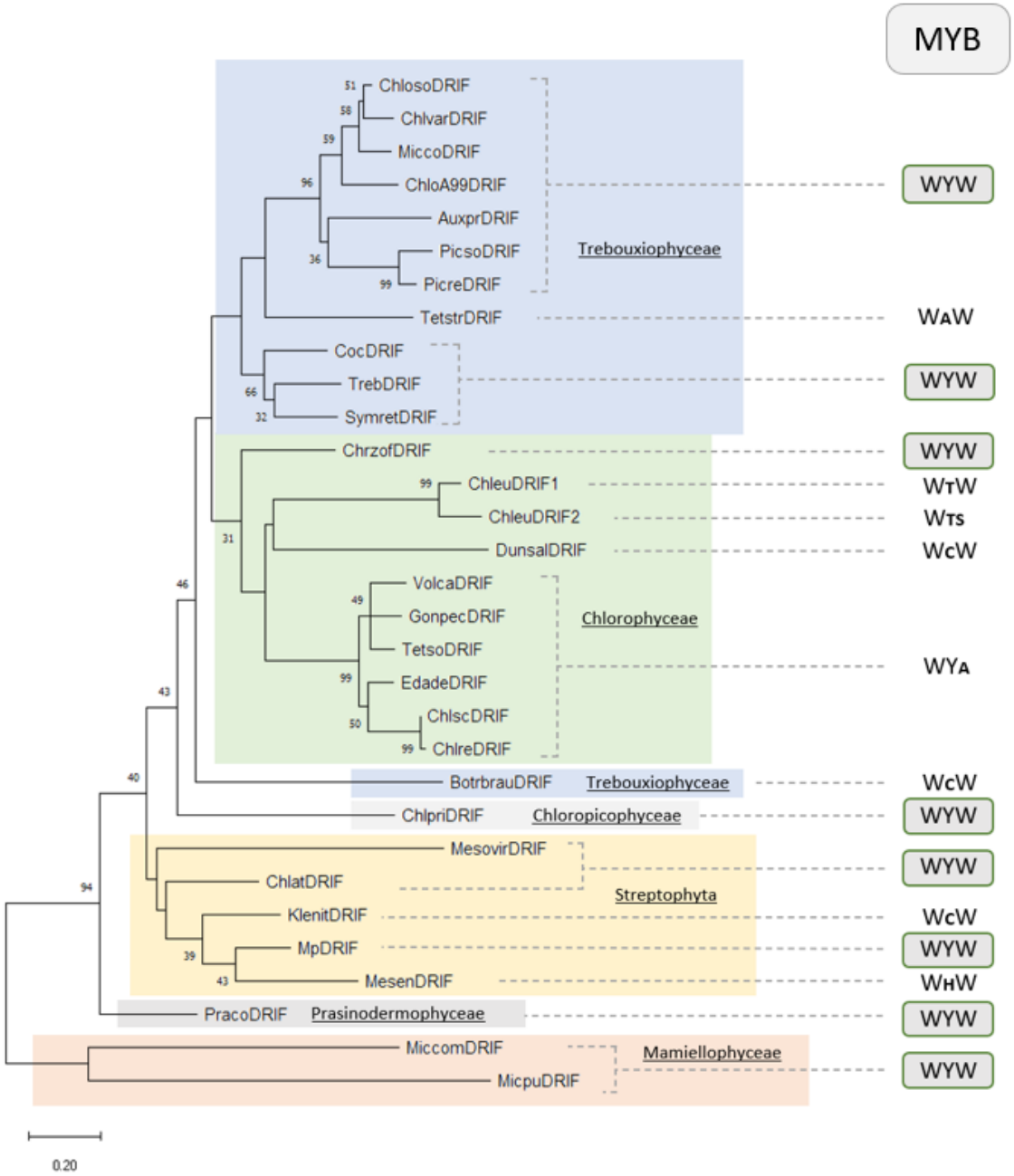

Figure 27. Evolutionary history and relationships of part of the DRIF family of proteins. The phylogenetic tree was constructed from the alignment of protein domain sequences of DRIF homologs from different algal species and *M. polymorpha* (*Micromonas commoda*, MiccomDRIF; *Micromonas pusilla*, MicpuDRIF; *Auxenochlorella protothecoides*, AuxprDRIF; *Botryococcus braunii*, BotrbrauDRIF; *Chlorella sp.* A99, ChloA99DRIF; *Chlorella sorokiniana*, ChlosoDRIF; *Chlorella variabillis*, ChlvarDRIF; *Coccomyxa subellipsoidea*, CocDRIF; *Micractinium conductrix*, MiccoDRIF; *Picochlorum renovo*, PicreDRIF; *Picochlorum soloecismus*, PicsoDRIF; *Symbiochloris reticulata*, SymretDRIF; *Tetraselmis striata*, TetstrDRIF; *Trebouxia sp.* A1-2, TrebDRIF ; *Chlamydomonas eustigma*, ChleuDRIF1 and ChleuDRIF2; *Chlamydomonas reinhardtii*, ChlreDRIF; *Chlamydomonas schloesseri*, ChlscDRIF; *Chromochloris zofingiensis*, ChrzofDRIF; *Dunaliella salina*, DunsalDRIF; *Edaphoclamys debaryana*, EdadeDRIF; *Gonium pectorale*, GonpecDRIF; *Tetrabaena socialis*, TetsoDRIF; *Volvox carteri*, VolcaDRIF; *Chloropicon primus*, ChlpriDRIF; *Prasinoderma coloniale*, PracoDRIF; *Chara braunii*, ChabraDRIF; *Chlorokybus atmophyticus*, ChlatDRIF; *Klebsormidium nitens*, KlenitDRIF; *Mesostigma viride*, MesovirDRIF; *Mesotaenium endlicherianum*, MesenDRIF). The conservation of the characteristic aromatic residues of the MYB domain is presented for each sequence. Two Tryptophan and one Tyrosine (WYW) are the expected amino acids. Smaller font size indicates the amino acid replacing the aromatic residue (F, Phenylalanine; L, Leucine; C, Cysteine; A, Alanine; M, Methionine; T, Threonine). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site.

A final alignment was carried out with the domain only peptide sequences (Annex D) and the conservation of amino acids between homologs was analysed with sequence logos (**Figure 27**). These alignments were also used to infer phylogenetic relationships between the homologous genes via construction of phylogenetic trees for DIV and DRIF homologs (**Figure 26** and **Figure 27**).

Alignment analysis, made clear with the sequence logos of **Figure 27**, revealed that the MYBI domain of DIV proteins showed lower conservation of amino acids than the MYBII domain. The aromatic residues characteristic of MYB domains have different degrees of conservation in the MYBI domain. The first Tryptophan (W) is almost completely conserved between species, while the second Tryptophan amino acid is conserved in only slightly above half of the species and the Tyrosine (Y) is conserved in about two thirds. The aromatic residues of the MYBII domain are nearly completely conserved with the only exception of the Tyrosine of the CryptoDIV (*Cryptophyceae* sp. CCMP2293) which was replaced by the aromatic residue Phenylalanine (F). Besides the aromatic residues, the surrounding amino acids of MYBII, including the SHAQKYF motif, are highly conserved as well has certain amino acids in the MYBI domain. Glutamic acids (E), Phenylalanine, Leucine (L) and Lysine (K) amino acids are highly conserved in certain positions, with even higher conservation than some of the characteristic aromatic residues.

Analysis of DRIF domain alignments showed that both the MYB and DUF3755 domains have lower conservation of amino acids than that of DIV MYBII domain, however, regarding the characteristic aromatic residues of the MYB domain of DRIF, they were highly conserved in comparison to the MYBI of DIV proteins. The first Tryptophan is completely conserved between species and the Tyrosine and final Tryptophan have near complete conservation. Besides the aromatic residues, certain amino acids are even more highly conserved in certain positions within the MYB domain, such as Glutamic acid (E), Leucine (L), Arginine (R) and Lysine (K), among others. The DUF3755 domain is less studied than the MYB domains and therefore characteristic amino acids have not been identified at this point in time. According to the Conserved Domain Database (Lu et al., 2019), the Asparagine (N) amino acid is highly conserved between species and that could indicate functional importance. In the sequence logos obtained from sequence alignments, it is apparent that among the DUF3755 domains of algae analysed, several positions are highly conserved such as Asparagine amino acids, along with various Leucine, Methionine (M), Proline (P) and Valine (V) amino acids, among others (**Figure 25**).

In the DIV phylogenetic tree, sequences belonging to the same group are generally grouped together with a few exceptions, although many of these exceptions are not statistically supported (**Figure 26**). Of note are the separation of the *Chlamydomonas eustigma* DIV sequences (ChleuDIV1 and ChleuDIV2), which were grouped separately to the remainder of Viridiplantae, including other *Chlamydomonas* species, in 72% of scenarios, the inclusion of the *Chloropicon primus* DIV protein (ChlpriDIV) within the group with all Trebouxiophyceae class proteins and the inclusion of PracoDIV, of *Prasinoderma coloniale*, within Streptophyta (**Figure 26**), even though recent findings consider the species part of a new phylum, Prasinodermophyta (Li, Wang, et al., 2020).

The DRIF phylogenetic tree generally presents sequences from species belonging to the same group together, with a few exceptions. The DRIF of *Botryococcus braunii* was significantly set apart from the remaining Trebouxiophyceae, which were more closely grouped with the class Chlorophyceae (**Figure 27**).

3.3. Heterologous Protein Expression

Another main goal of this thesis was to develop a protocol for the heterologous protein expression of the MpDIV1, MpDIV2 and MpDRIF proteins in *E. coli* and to experimentally determine whether the proteins of the *DIV* homologs of *M. polymorpha* bind to a specific DNA sequence and form a dimer with the MpDRIF.

pGEX-6P-1 was one of the vectors used for heterologous expression. It contains the *lac* promoter for control of expression, which is activated by lactose, or the analogue of lactose known as IPTG. The vector also contains the repressor gene *laclq*, which impedes the activation of expression when lactose or IPTG are absent. Besides expression machinery, the vector has an N-terminal Glutathione Signal Transferase (GST) tag for post expression processing and carries resistance to Ampicillin. This vector was used in the heterologous expression in various *E. coli* strains of the MpDIV1 and MpDIV2 proteins.

Gateway™ pDEST™15 was the other vector used for heterologous expression. It contains an N-terminal GST tag, carries resistance to Ampicillin and has the T7 promoter for control of expression. The T7 promoter, when used together with the appropriate expression strain of *E. coli*, can be induced with lactose and IPTG. This vector was used in the heterologous expression of the MpDRIF protein and was so chosen because the Mp*DRIF* coding sequence had already been

recombined into pDONR201, so a simple LR reaction was all that was needed to insert Mp*DRIF* into the expression vector.

3.3.1. Expression strain preparation – Cloning procedures

As stated above, constructs for Mp*DIV1* and Mp*DIV2* had already been prepared and used for expression. The expression vector for Mp*DRIF* was prepared during this project.

pDONR201 with Mp*DRIF* coding sequence had been previously prepared and transformed into *E. coli* DH10β. The pDONR201-Mp*DRIF* construct was extracted from the cells via miniprep and then used in a LR recombination reaction in which the target vector was pDEST™15. The results of the reaction were used in the transformation of three different strains of *E. coli*, Rosetta™ (DE3)pLysS, BL21(DE3)-R3-pRARE2 and OverExpress™ C43(DE3). Colony PCRs and agarose gel electrophoresis, with an expected sequence size of around 1000 bp, were used to confirm positive colonies as shown in **figure 28**.

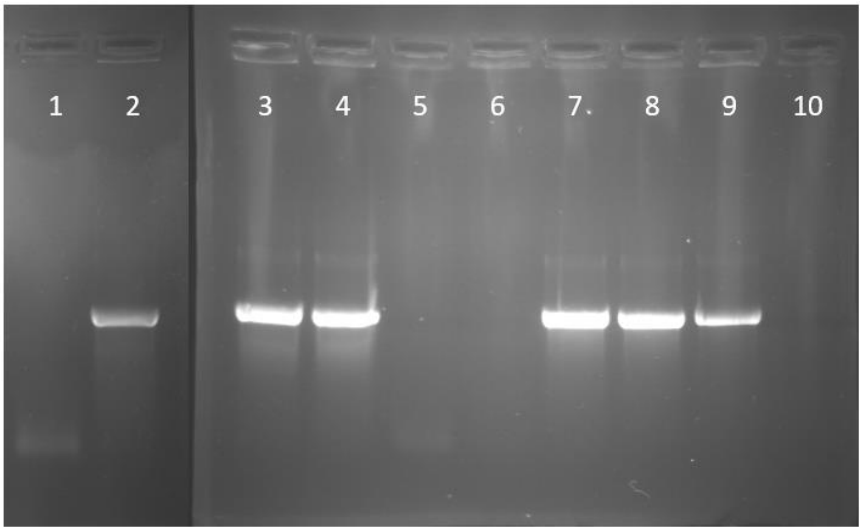

Figure 28. Colony PCR of *E. coli* transformed with the LR product of Mp*DRIF* cloning into pDEST15. Primers Qs190 and Qs191 or Qs190 and UMC512 were used for amplification of Mp*DRIF* sequence. **1** - negative control; **2-10** – pDEST15-Mp*DRIF* transformed colonies.

As mentioned above, the empty pGEX-6P-1 vector and constructs of pGEX-Mp*DIV1* and pGEX-Mp*DIV2* had been prepared and previously used in heterologous expression in the Rosetta™ (DE3)pLysS and BL21(DE3)-R3-pRARE2 strains. Transformations were performed in BL21(DE3)-R3-pRARE2 (pRARE2), OverExpress™ C43(DE3) (C43) and Rosetta™ (DE3)pLysS cells (Rosetta).

3.3.2. Expression protocol optimisation

Expression of the MpDIV1 and MpDIV2 protein had already been achieved previously (Almeida, 2019). Induction of expression was performed in the past with both lactose and IPTG but focused on analysis of IPTG induction. The following results for Mp*DIV* expression focused on lactose induction because it was believed it would lead to higher quantity of soluble protein. Heterologous expression of Mp*DRIF* had not been achieved in past works and so was attempted here. Expression protocols were attempted and optimized to eventually deliver soluble, and therefore potentially active, MpDIV1, MpDIV2 and MpDRIF proteins.

In the case of Mp*DRIF*, optimization focused on tuning the length of induction and the temperature at which it occurred. IPTG induction of expression with BL21(DE3)-R3-pRARE2 with pDEST15™-Mp*DRIF* was attempted with induction times of 2 or 4 hours, with different IPTG concentrations (0.4 mM and 1.0 mM) at 30°C or 37°C. The total fractions were analysed in the various resulting SDS-PAGEs pictured in **Figure 29** and **Figure 30**.

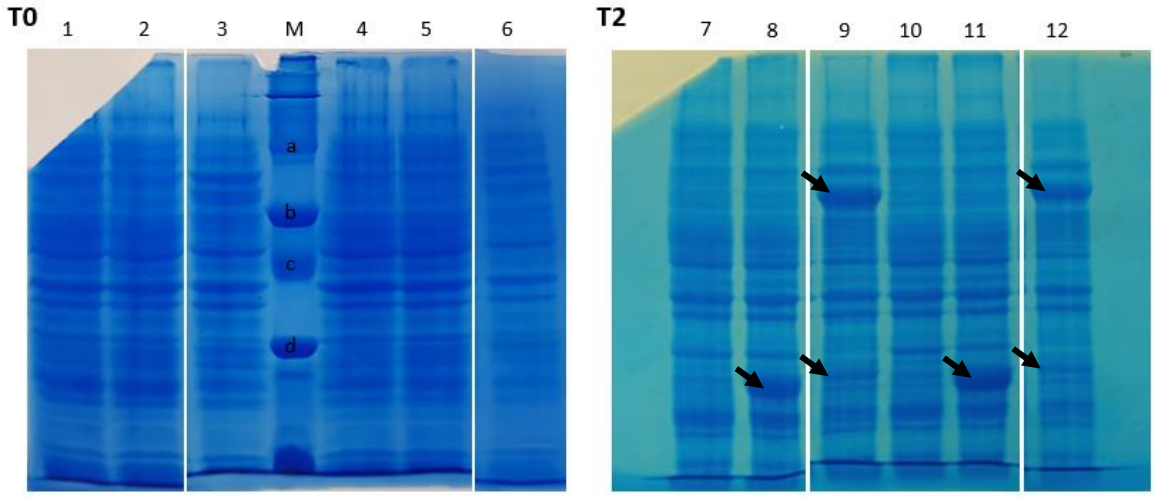

Figure 29. SDS-PAGE analysis of total fractions from heterologous expression of MpDRIF in BL21(DE3)-R3-pRARE2 *E. coli* before and after 2 hours of induction with two different concentrations of IPTG. SDS page acrylamide gel (T=12%, C=4%) was used to run samples with SDS-PAGE buffer. Staining with Coomassie Blue solution. T0 corresponds to analysis of samples taken from cells before induction occurred. T2 corresponds to analysis of samples taken from cells 2 hours after induction at 37°C. Black arrows point to bands potentially corresponding to expected heterologous proteins. Samples are total fraction of protein. Before induction: Grown at 30°C: **1** - pRARE2 without plasmid; **2** - pGEX; **3** - pDEST15-MpDRIF; Grown at 37°C: **4** - pRARE2; **5** - pGEX; **6** - pDEST15-MpDRIF; Induction at 37°C: IPTG 0.4 mM: **7** - pRARE2; **8** - pGEX; **9** - pDEST15-MpDRIF; IPTG 1.0 mM:**10** - pRARE2; **11** - pGEX; **12** - pDEST15-MpDRIF; **M** - SDS-PAGE Molecular Weight Standard (Broad Range, Bio-Rad); a - 97.4 kDa; b - 66.2 kDa; c - 45 kDa; d - 31 kDa.

The electrophoretic polypeptide profiles taken from cells before induction were analysed to serve as a negative control on induction of expression. As expected, there seemed to be no differences between pRARE2 cells without plasmids, with pGEX and with pDEST15-Mp*DRIF* since

induction had not yet occurred. Two hours after induction (T2) at 37°C, the analysis showed bands, in lane 9 and 12 of **Figure 29**, that could correspond to MpDRIF-GST protein and there appeared to be little difference in band intensity between induction with 0.4 mM IPTG and 1.0 mM IPTG. Four hours after induction (T4), the analysis suffered from clarity issues, possibly from high amount of sample being used in the wells, but this was resolved in the T2 analysis. It was still possible to

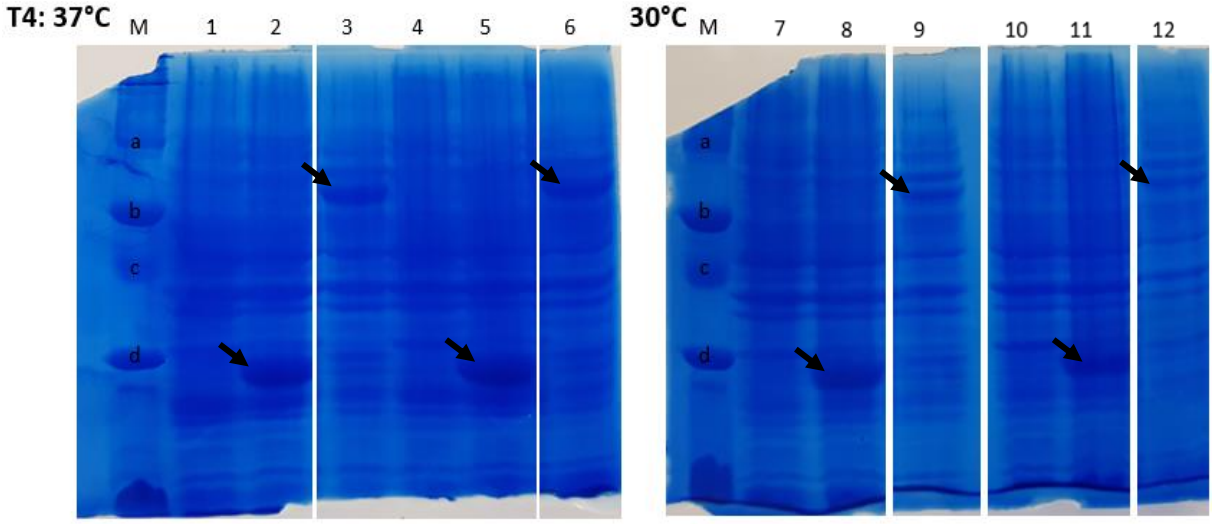

Figure 30. SDS-PAGE analysis of total fractions from heterologous expression of MpDRIF in BL21(DE3)-R3-pRARE2 *E. coli* after 4 hours of induction, comparing two different temperatures of induction and two different concentrations of IPTG. SDS page acrylamide gel (T=12%, C=4%) was used to run samples with SDS-PAGE buffer. Staining with Coomassie Blue solution. T4 correspond to analysis of samples taken from cells 4 hours after induction. Black arrows point to bands potentially corresponding to expected heterologous proteins. Samples are total fraction of protein. Induction at 37°C: IPTG 0.4 mM: **1** - pRARE2 without plasmid; **2** - pGEX; **3** - pDEST15-MpDRIF; IPTG 1.0 mM: **4** - pRARE2; **5** - pGEX; **6** - pDEST15-MpDRIF; Induction at 30°C: IPTG 0.4 mM: **7** - pRARE2; **8** - pGEX; **9** - pDEST15-MpDRIF; IPTG 0.4 mM: **10** - pRARE2; **11** - pGEX; **12** - pDEST15-MpDRIF; **M** - SDS-PAGE Molecular Weight Standard (Broad Range, Bio-Rad); a - 97.4 kDa; b - 66.2 kDa; c - 45 kDa; d - 31 kDa.

discern the presence of bands potentially corresponding to MpDRIF-GST in lanes 3, 6, 9 and 12 of **Figure 30**, especially when induction was carried out at 37°C. In all cases, bands potentially corresponding to MpDRIF-GST were at a molecular weight of above 66 kDa, which is well above MpDRIF-GST theorized MW of 58 kDa.

Analysis of the soluble fractions resulting from four hours of induction (**Figure 31**) showed that, despite induction at 30°C analysis having lower intensity MpDRIF-GST bands in the total and insoluble fractions (**Figure 30**), the lower temperature appears to lead to bands that could be target protein with increased intensity in the soluble fraction (Lanes 18 and 24 of **Figure 31**) when compared to 37°C. Further studies utilizing MpDRIF protein were carried out with the heterologous expression protocol with induction for 4 hours at 30°C with 0.4 mM IPTG.

For the expression of Mp*DIVs*, 5 g L^{-1}, 10 g L^{-1} and 20 g L^{-1} of lactose were used for overnight induction at 30 °C, using BL21(DE3)-R3-pRARE2 strains with Mp*DIV1*-pGEX and Mp*DIV2*-pGEX (**Figure 32**). In the results (lanes 4, 8, 14 and 15) MpDIV2-GST's molecular weight seems to coincide very closely to the 50 kDa band of the marker, which is below the expected molecular weight of the MpDIV2-GST fusion protein (59 KDa). The band that corresponds to MpDIV1-GST was also below the theorized molecular weight of the MpDIV1-GST fusion protein (63

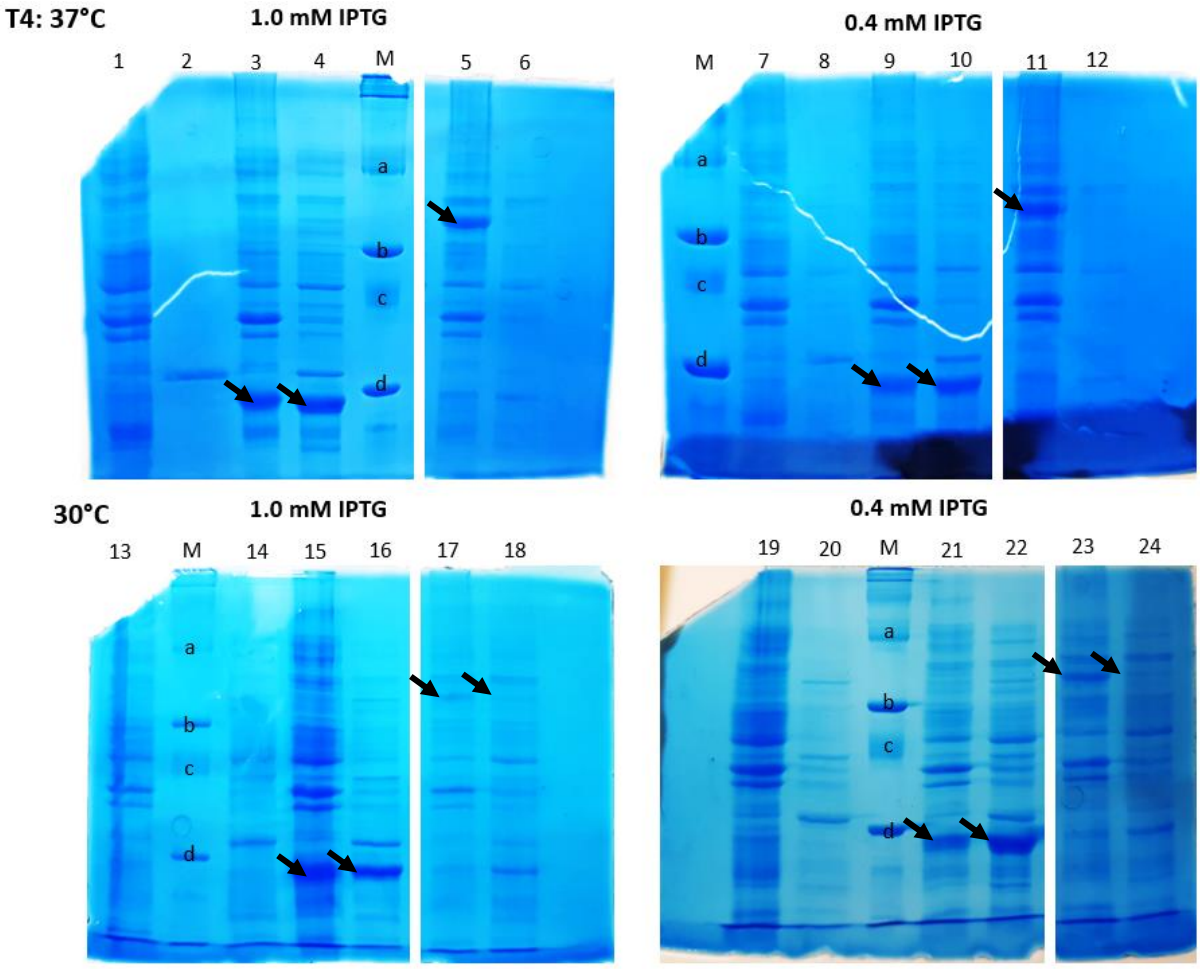

Figure 31. Comparison of the effect of different temperatures and concentration of IPTG on soluble and insoluble fractions resulting from heterologous expression of MpDRIF in BL21(DE3)-R3-pRARE2 *E. coli* after 4 hours of induction. SDS page acrylamide gel (T=12%, C=4%) was used to run samples with SDS-PAGE buffer. Staining with Coomassie Blue solution. T4 correspond to analysis of samples taken from cells 4 hours after induction. Black arrows point to bands potentially corresponding to expected heterologous proteins. Samples are total fraction of protein. Induction at 37°C: IPTG 1.0 mM: **1** - Insoluble fraction pRARE2 without plasmid; **2** - Soluble fraction pRARE2; **3** - Insoluble fraction pGEX; **4** - Soluble fraction pGEX; **5** - Insoluble fraction pDEST15-MpDRIF; **6** - Soluble fraction pDEST15-MpDRIF; IPTG 0.4 mM: **7** - Insoluble fraction pRARE2; **8** - Soluble fraction pRARE2; **9** - Insoluble fraction pGEX; **10** - Soluble fraction pGEX; **11** - Insoluble fraction pDEST15-MpDRIF; **12** - Soluble fraction pDEST15-MpDRIF; Induction at 30°C: IPTG 1.0 mM: **13** - Insoluble fraction pRARE2; **14** - Soluble fraction pRARE2; **15** - Insoluble fraction pGEX; **16** - Soluble fraction pGEX; **17** - Insoluble fraction pDEST15-MpDRIF; **18** - Soluble fraction pDEST15-MpDRIF; IPTG 0.4 mM: **19** - Insoluble fraction pRARE2; **20** - Soluble fraction pRARE2; **21** - Insoluble fraction pGEX; **22** - Soluble fraction pGEX; **23** - Insoluble fraction pDEST15-MpDRIF; **24** - Soluble fraction pDEST15-MpDRIF; **M** - SDS-PAGE Molecular Weight Standard (Broad Range, Bio-Rad); a- 97.4 kDa; b- 66.2 kDa; c- 45 kDa; d- 31 kDa.

kDa), appearing in **Figure 32** (lanes 3, 7, 11 and 12) at a height between the bands for 50 kDa and 37 kDa. This result, together with the existence of bands that correspond to GST cleaved from the MpDIVs (lanes 3, 4, 7, 8, 11, 12, 14 and 15), indicate that cleavage could be occurring within the proteins and not only between the MpDIVs and the GST tags. Relative to concentration of lactose, no significant band intensity differences were found between the 5 g L^{-1}, 10 g L^{-1} and 20 g L^{-1} samples.

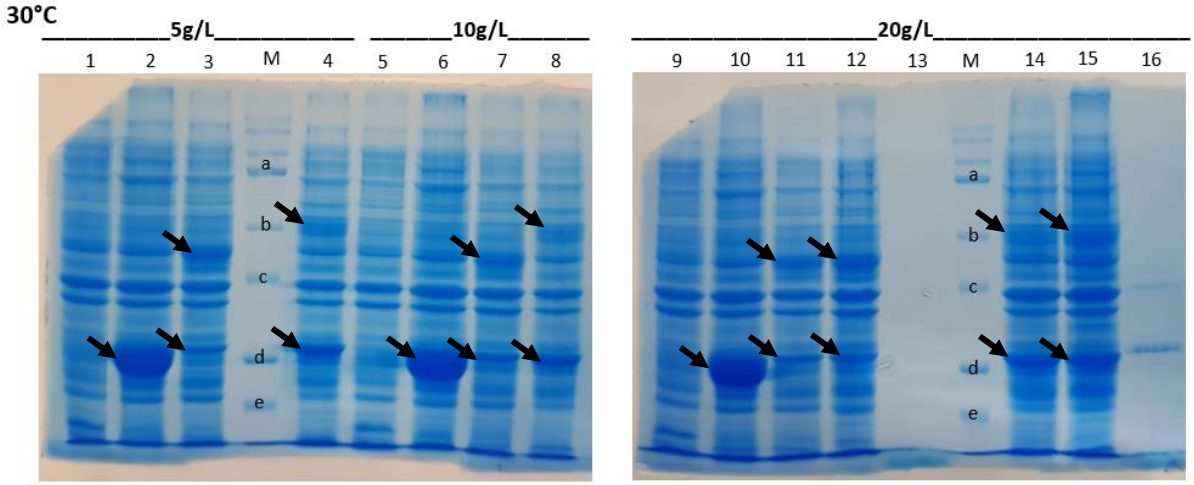

Figure 32. Comparison of the effect of different concentrations of lactose on heterologous expression of MpDIV1 and MpDIV2 in BL21(DE3)-R3-pRARE2 *E. coli*. Cells grown and expression induced overnight at 30°C. SDS page acrylamide gel (T=12%, C=4%) was used to run samples with SDS-PAGE buffer. Staining with Coomassie Blue solution. Black arrows point to bands potentially corresponding to expected heterologous proteins. Induction with lactose 5 g·L^{-1}: **1** - Total fraction pRARE2; **2** - Total fraction pGEX; **3** - Total fraction pGEX-MpDIV1; **4** - Total fraction pGEX-MpDIV2; Induction with lactose 10 g·L^{-1}: **5** - Total fraction pRARE2; **6** - Total fraction pGEX; **7** - Total fraction pGEX-MpDIV1; **8** - Total fraction pGEX-MpDIV2; Induction with lactose 20 g·L^{-1}: **9** - Total fraction pRARE2; **10** - Total fraction pGEX; **11** - Total fraction pGEX-MpDIV1; **12** - Insoluble fraction pGEX-MpDIV1; **13** - Soluble fraction pGEX-MpDIV1; **14** - Total fraction pGEX-MpDIV2; **15** - Insoluble fraction pGEX-MpDIV2; **16** - Soluble fraction pGEX-MpDIV2; **M** - Precision Plus Protein™ Kaleidoscope™ Prestained Protein Standards (Bio-Rad): a - 75 kDa; b - 50 kDa; c - 37 kDa; d - 25 kDa; e - 20 kDa.

In order to confirm whether the target proteins were in fact being expressed and whether cleavage was occurring during expression, a western blot was carried out for the protein profiles of pRARE2, pRARE2-pGEX, Mp*DIV1*, Mp*DIV2* and Mp*DRIF* (**Figure 33**).

The analysis showed antibody binding at various molecular sizes in most lanes. This is normal and could indicate that more washing steps are necessary, which in this case is more probable than non-specific binding since the antibody is not observed in the pRARE2 lane (lane 1). Antibody binding is otherwise concentrated on bands that could correspond to GST (lane 2), Mp*DIV1*-GST (lane 3), Mp*DIV2*-GST (lane 5) and to Mp*DRIF*-GST (lane 7). The Mp*DRIF* band is less distinguishable but there seems to be a band with increased intensity at the previously observed MW (arrow on lane 7).

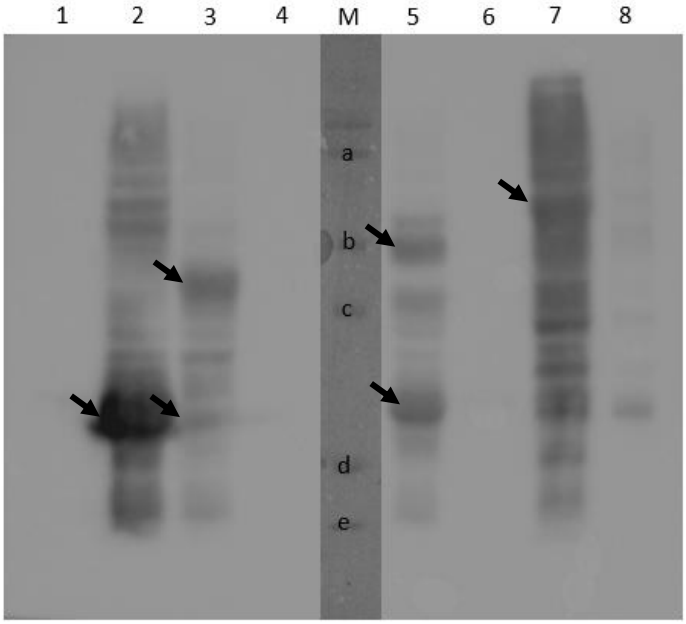

Figure 33. Western blot analysis of heterologous expression of MpDIV1, MpDIV2 and MpDRIF in BL21(DE3)-R3-pRARE2 *E. coli*. Samples with SDS-PAGE buffer run in SDS page acrylamide gel (T=12%, C=4%) and then transferred onto nitrocellulose membrane. Chemiluminescence observed with G:BOX Chemi/XX9 (Syngene). Induction of expression at 30°C, overnight with lactose 20 g·L^{-1} for MpDIVs, pGEX and pRARE2 and for 4 hours with IPTG 1.0 mM for MpDRIF. Black arrows point to bands potentially corresponding to expected heterologous proteins. **1** - Total fraction pRARE2; **2** - Total fraction pGEX; **3** - Total fraction pGEX-MpDIV1; **4** - Soluble fraction pGEX-MpDIV1; **5** - Total fraction pGEX-MpDIV2; **6** - Soluble fraction pGEX-MpDIV2; **7** - Total fraction pDEST15-MpDRIF; **8** - Soluble fraction pDEST15-MpDRIF; **M** - Precision Plus Protein™ Kaleidoscope™ Prestained Protein Standards (Bio-Rad): a - 75 kDa; b - 50 kDa; c - 37 kDa; d - 25 kDa; e - 20 kDa.

In this analysis, much like in the previous SDS-PAGE analysis (**Figure 32**), the MW of MpDIV1-GST and MpDIV2-GST bands is considerably below the theorized molecular weight, with the MpDIV1 band appearing between 50 and 37 kDa and the MpDIV2 band appearing at about 50 kDa. This and the presence of GST corresponding bands in lanes 3 and 5 indicate that cleavage must be occurring between the GST tag and the MpDIV proteins. Lanes 4, 6 and 8 correspond to the soluble fractions and, with the exception of some binding in lane 8, which corresponds to Mp*DRIF* expression, they appeared to have an imperceptible level of protein expression and no bands corresponding to the target protein.

In an attempt to solve both the issue of protein cleavage and the inefficiency in separation of the soluble fraction the constructs were newly transformed into the OverExpress™ C43(DE3) and Rosetta™ (DE3)pLysS *E. coli* strains and expression was induced with both strains.

Induction was carried out with the lactose protocol, overnight at 30 °C with a lactose concentration of 10 g L^{-1} while IPTG induction, with 0.4 mM IPTG, 4 h induction time at 30 °C, was used for cells with the Mp*DRIF*-pDEST™15 construct. Analysis was conducted with SDS-PAGE (**Figure 34**).

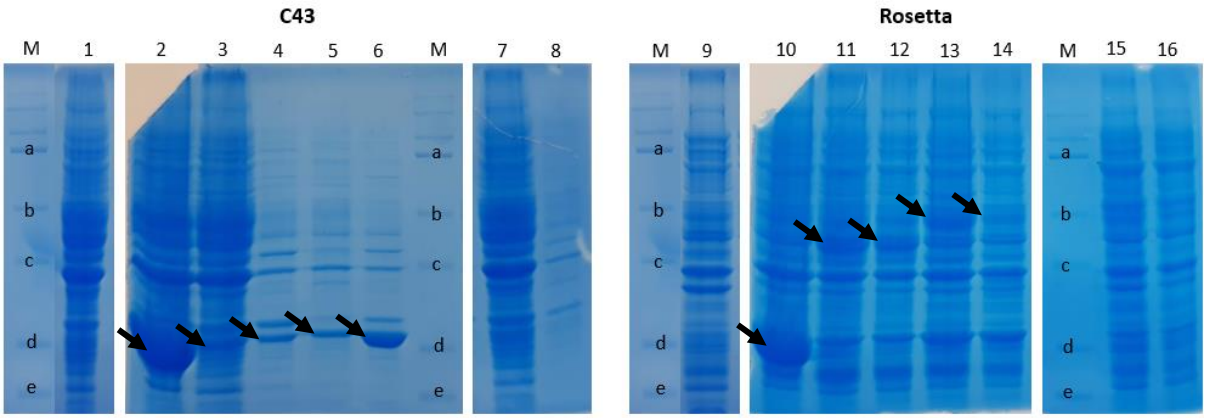

Figure 34. Comparison of heterologous expression of MpDIV1, MpDIV2 and MpDRIF in two different strains, OverExpress™ C43(DE3) and in Rosetta™ (DE3)pLysS *E. coli*. Cells grown and expression induced overnight at 30°C with lactose at 20 g·L^{-1}. SDS page acrylamide gel (T=12%, C=4%) was used to run samples with SDS-PAGE buffer. Staining with Coomassie Blue solution. Black arrows point to bands potentially corresponding to expected heterologous proteins. Induction in C43 strain: **1** - Total fraction C43; **2** - Total fraction pGEX; **3** - Total fraction pGEX-MpDIV1; **4** - Soluble fraction pGEX-MpDIV1; **5** - Total fraction pGEX-MpDIV2; **6** - Soluble fraction pGEX-MpDIV2; **7** - Total fraction pDEST15-MpDRIF; **8** - Soluble fraction pDEST15-MpDRIF; Induction in Rosetta strain: **9** - Total fraction Rosetta; **10** - Total fraction pGEX; **11** - Total fraction pGEX-MpDIV1; **12** - Soluble fraction pGEX-MpDIV1; **13** - Total fraction pGEX-MpDIV2; **14** - Soluble fraction pGEX-MpDIV2; **15** - Total fraction pDEST15-MpDRIF; **16** - Soluble fraction pDEST15-MpDRIF; **M** - Precision Plus Protein™ Kaleidoscope™ Prestained Protein Standards (Bio-Rad): a - 75 kDa; b - 50 kDa; c - 37 kDa; d - 25 kDa; e - 20 kDa.

Analysis of the C43 strain peptides shows definite expression of GST in both pGEX and the Mp*DIVs* samples but bands that could correspond to expression of MpDIV1, MpDIV2 and MpDRIF were not evident in the gel. This analysis also suffered from issues in sample quantities, the pGEX, Mp*DIV1* and Mp*DRIF* total fraction samples appeared to be over stained while the total fraction of Mp*DIV2* had a low quantity due to poor solubilization of the cell pellet in SDS-PAGE sample buffer (**Figure 34**).

Rosetta strain analysis shows the best results for soluble fraction, with bands of equal intensity to the total fraction of proteins. MpDIV1-GST and MpDIV2-GST bands are distinguishable in the gel, in both total and souble fractions (Lanes 11, 12, 13 and 14). MpDRIF-GST expression was not apparent and so Mp*DRIF* expression was attempted in Rosetta and pRARE2, for comparison, with the 0.4 mM IPTG induction protocol and analysed via SDS-PAGE (**Figure 35**).

This final analysis has bands corresponding to expressed Mp*DRIF*-GST in both strains (Lanes 3, 4 and 7). Comparison between soluble strains of the strains shows that Rosetta has the better results with higher heterologous protein content while the band in soluble pRARE2 strain (Lane 8) is nearly imperceptible and could be from the natural protein profile of the strain.

The final optimized protocols, believed to produce soluble and active protein, were overnight induction at 30°C, with at least 10 g L^{-1} lactose for Mp*DIV1* and Mp*DIV2* expression in

the Rosetta™ (DE3)pLysS strain and, for Mp*DRIF* expression in the Rosetta strain, induction with 0.4 mM IPTG at 30°C for 4 hours.

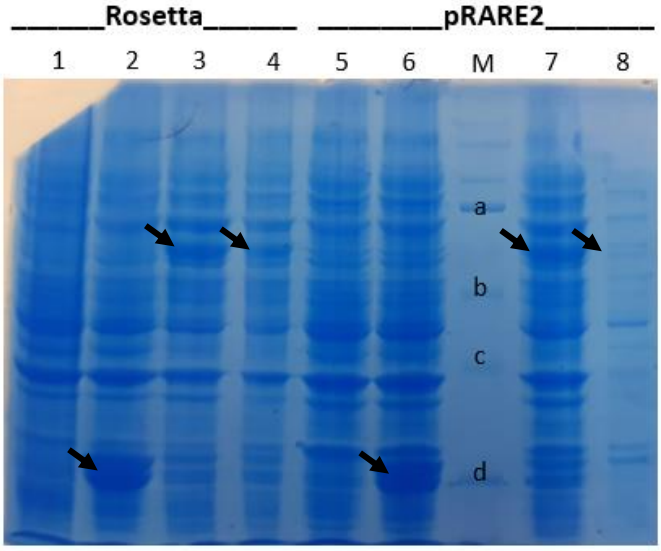

Figure 35. Comparison of soluble fractions resulting from heterologous expression of MpDIV1, MpDIV2 and MpDRIF in the two different strains, Rosetta™ (DE3)pLysS and BL21(DE3)-R3-pRARE2 *E. coli*. Expression induced at 30°C for 4 hours with 0.4 mM IPTG. SDS page acrylamide gel (T=12%, C=4%) was used to run samples with SDS-PAGE buffer. Staining with Coomassie Blue solution. Black arrows point to bands potentially corresponding to expected heterologous proteins. Induction in Rosetta strain: **1** - Total fraction Rosetta; **2** - Total fraction pGEX; **3** - Total fraction pDEST15-MpDRIF; **4** - Soluble fraction pDEST15-MpDRIF; Induction in pRARE2 strain: **5** - Total fraction pRARE2; **6** - Total fraction pGEX; **7** - Total fraction pDEST15-MpDRIF; **8** - Soluble fraction pDEST15-MpDRIF; **M** - Precision Plus Protein™ Kaleidoscope™ Prestained Protein Standards (Bio-Rad): a - 75 kDa; b - 50 kDa; c - 37 kDa; d - 25 kDa.

3.3.3. EMSA – Gel shift

The Electromobility Shift Assay was conducted to determine whether the *DIV* homologs of *M. polymorpha* bound to the GATAA sequence, the DNA consensus binding site of *Antirrhinum majus* DIV proteins (Raimundo et al., 2013). The assay was carried out with MpDIV1 and MpDIV2 obtained from heterologous expression in the Rosetta™ (DE3)pLys strain of E.coli. MpDRIF was not used until MpDIV binding to DNA could be confirmed.

For the EMSA, soluble, active protein is necessary and so a purification process was attempted with the protein profile samples of pGEX, Mp*DIV1* and Mp*DIV2* expression. As described in the methods, a glutathione resin-based protocol was used on the soluble fraction samples of protein expressed using the optimized lactose expression and the results were analysed by SDS-Page (**Figure 36**).

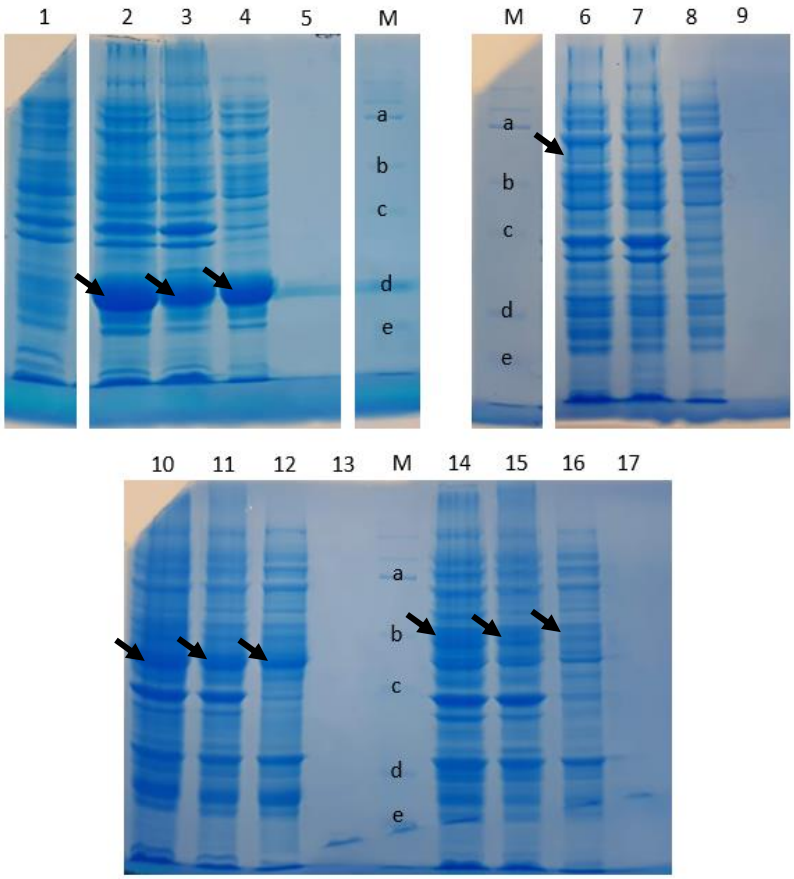

Figure 36. SDS-PAGE analysis of purification results of GST, MpDIV1, MpDIV2 and MpDRIF in Rosetta™ (DE3)pLys E.coli. Expression induced at 30°C, overnight with lactose at 20 $g \cdot L^{-1}$. SDS page acrylamide gel (T=12%, C=4%) was used to run samples with SDS-PAGE buffer. Staining with Coomassie Blue solution. Black arrows point to bands potentially corresponding to expected heterologous proteins. **1** - Total fraction Rosetta; **2** - Total fraction pGEX; **3** - Insolube fraction pGEX; **4** - Soluble fraction pGEX; **5** - Purified soluble fraction pGEX; **6** - Total fraction pDEST15-MpDRIF; **7** - Insoluble fraction fraction pDEST15-MpDRIF; **8** - Soluble fraction pDEST15-MpDRIF; **9** - Purified soluble fraction pDEST15-MpDRIF; **10** - Total fraction pGEX-MpDIV1; **11** - Insoluble fraction pGEX-MpDIV1; **12** - Soluble fraction pGEX-MpDIV1; **13** - Purified soluble fraction pGEX-MpDIV1; **14** - Total fraction pGEX-MpDIV2; **15** - Insoluble fraction pGEX-MpDIV2; **16** - Soluble fraction pGEX-MpDIV2; **17** - Purified soluble fraction pGEX-MpDIV2; **M** - Precision Plus Protein™ Kaleidoscope™ Prestained Protein Standards (Bio-Rad): a - 75 kDa; b - 50 kDa; c - 37 kDa; d - 25 kDa; e - 20 kDa.

Analysis showed no perceptible protein in the purification products except for a band present in pGEX purified sample that is at the correct MW for GST (Lanes 5 and adjacent M). However, this band is continuous across the gel, even intersecting the MW marker lane which indicates it could result from protein falling outside the wells and smearing across the gel and not from actual expression. Since purification was unsuccessful several times, non-purified protein samples were used for the EMSA method.

As described in 2.5.4, samples were prepared with a combination of 6-FAM labelled DNA probe, containing the GATAA sequence, to which DIV proteins have been shown to bind, with non-specific poly GC probe solution that proteins with non-specific binding should bind to, with binding buffer and with heterologous protein. A sample without protein and with the probe was run, as well as samples with probe and soluble fraction of Rosetta and of pGEX, all three as control. To evaluate binding activity, samples with the probe and individual MpDIV1, MpDIV2 were run. The resulting EMSA is pictured in **Figure 37**.

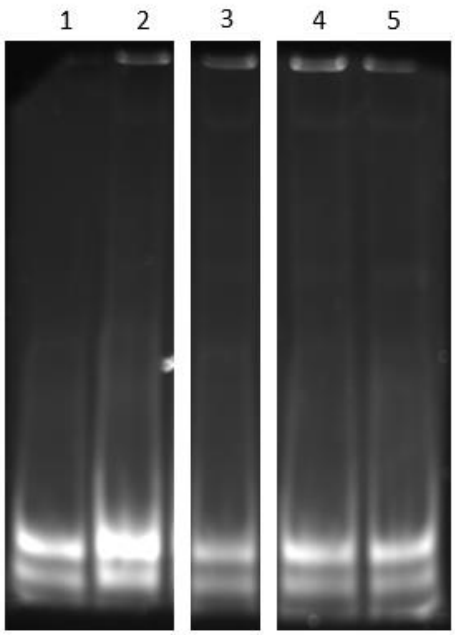

Figure 37. Electromobility Shift Assay (EMSA) of MpDIV1 and MpDIV2. Heterologous proteins were expressed in Rosetta™ (DE3)pLys strain of E.coli. Induction overnight with lactose at 20 $g{\cdot}L^{-1}$. Samples run on non-denaturing Bis-acrylamide gel (T=6% and C=2.5%). 6-FAM dyed DNA oligonucleotide probe. **1** - DNA probe without protein; **2** - Probe and MpDIV1; **3** - Probe and MpDIV2; **4** - Probe and Rosetta sample; **5** - Probe and pGEX sample.

The EMSA gel showed no apparent binding of antibodies. This could be because MpDIV1 and MpDIV2 do not bind to sequences with the GATAA motif, and that this specificity evolved at a later point. It could also be that the quantity of active protein used in the assay is too low. To find out more about what was going on, a western blot analysis was repeated (**Figure 38**). This analysis hadn't been performed for protein samples expressed in the Rosetta strain, which was used for the EMSA, and it would indicate more about the state of the proteins, and whether cleavage was still an issue.

First, non-specific binding does not seem to be an issue since the Rosetta sample (lane 1) has no bound antibody, except for a band at around the MW of GST which is continuous across the lanes (from lane 1 to lane M). In comparison with the previous western blot, performed with the pRARE2 strain, this analysis has less binding across lanes and specific bands are more

distinguishable. Additionally, antibody binding occurred in the soluble fractions of Mp*DIV1*, Mp*DIV2* and Mp*DRIF* (lanes 5, 7 and 9). Cleavage appears to still be an issue with the Rosetta strain.

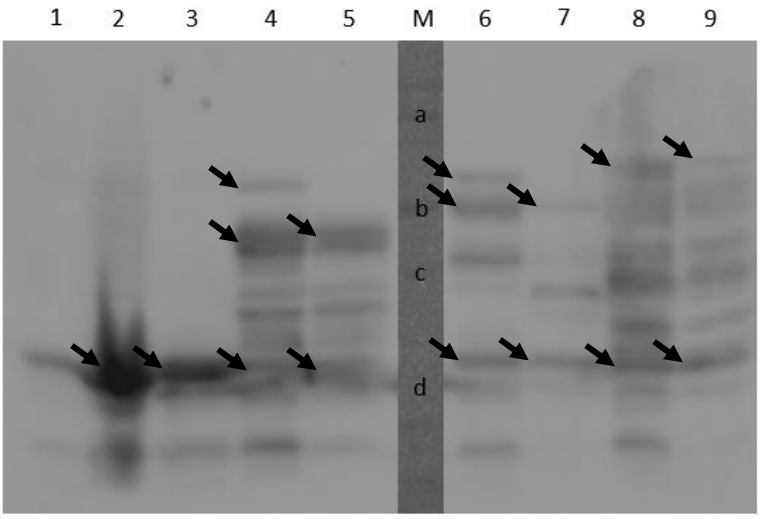

Figure 38. Evaluation of expressed protein condition via western blot analysis of expressed MpDIV1, MpDIV2 and MpDRIF in Rosetta™ (DE3)pLys *E. coli*. Samples with SDS-PAGE buffer run in SDS page acrylamide gel (T=12%, C=4%) and then transferred onto nitrocellulose membrane. Chemiluminescence observed with G:BOX Chemi/XX9 (Syngene). Induction of expression at 30°C, overnight with lactose 20 $g \cdot L^{-1}$ for MpDIVs, pGEX and pRARE2 and for 4 hours with IPTG 1.0 mM for MpDRIF. Black arrows point to bands potentially corresponding to expected heterologous proteins. **1** - Total fraction pRARE2; **2** - Total fraction pGEX; **3** - Soluble fraction pGEX; **4** - Total fraction pGEX-MpDIV1; **5** - Soluble fraction pGEX-MpDIV1; **6** - Total fraction pGEX-MpDIV2; **7** - Soluble fraction pGEX-MpDIV2; **8** - Total fraction pDEST15-MpDRIF; **9** - Soluble fraction pDEST15-MpDRIF; **M** - Precision Plus Protein™ Kaleidoscope™ Prestained Protein Standards (Bio-Rad): a- 75 kDa; b- 50 kDa; c- 37 kDa; d- 25 kDa.

In the case of Mp*DIV1* and Mp*DIV2*, the bands with highest intensity are between 50 and 37 kDa and at 50 kDa, respectively, which is, again, significantly below the theorized molecular weight. There are however bands above these in the total fraction lanes (lanes 4 and 6) with higher MW, closer to the theorized value.

Mp*DRIF* lanes had no one band that stood out in terms of intensity. The ones with highest intensity were one between 50 and 75 kDa, one slightly below 37 kDa and one above 25 kDa. The first could correspond to the MpDRIF-GST fusion protein, which has a theorized MW of 58 kDa. The second and third bands of notice could correspond to MpDRIF and GST, respectively, the fusion having been cleaved and resulted in separation of the MpDRIF protein and the GST tag.

These results indicate that the cleavage of the heterologous proteins is impairing the heterologous expression of soluble, active Mp*DIVs* and less so in the case of Mp*DRIF*.

4. Discussion

The DDR regulatory module influences many different, essential aspects of plant development in angiosperms. Study of the module could help us understand some of the many ways life has evolved to control some of its most complex processes. Angiosperms and higher plants, in particular, have accumulated greater developmental and genomic complexity over time. Thus, the use of more basal species and the study of ancestral genes can be a simpler way in which we can approach this biological question.

This thesis approaches the DDR regulatory module by studying the evolution and ancestral functions of the *DIV* and *DRIF* genes in *M. polymorpha.* Here, the results of the phenotypical analysis, plasmid cloning, heterologous expression and phylogenetic analysis are discussed in detail, conclusions are drawn, and future perspectives are considered.

4.1. *DIV* ancestral function further clarified

In this thesis, both knockout and overexpression of Mp*DIV2* in *M. polymorpha* plants were analysed via phenotypical observation and plant measurement. A previous analysis of knockout DIV plants (Coelho, S., 2019) found smaller plants and suggested that Mp*DIV2* was involved in promoting cell proliferation and/or expansion and in controlling the shape and size of the thallus. In the current analysis of Mp*div2* mutant lines, the results indicate the same conclusion. The two analysed lines of Mp*div2* plants have retarded growth and are of similar size and shape, although plants from line #305 are slightly larger, length and width wise. As stated previously, mutant plant line #68 has a complete deletion of the MYBII DNA binding domain of MpDIV2 and line #305 has a complete deletion of both MYB domains. This begs the question, could this difference be responsible for the difference, albeit small, in size between the mutant lines? A possible explanation for the slight discrepancy in length and width could be that, by losing the DNA binding domain, the MpDIV2 protein of plants from line #68 essentially becomes a RAD protein, retaining only the domain theoretically responsible for binding to MpDRIF. This potential new protein could still be able to bind to MpDRIF, if the MYBI domain were structurally intact, and thus impede it from binding to the remaining DIV homolog, MpDIV1. Previous studies using Mp*DIV1* overexpression and knockout plants showed that this gene might be involved in regulating cell proliferation or expansion, although the effects on plant size and shape did not appear to be as significant as with loss of Mp*DIV2* (Coelho, S., 2019). Thus, a partial loss of Mp*DIV1* function by abduction of MpDRIF

proteins by the MYBI domain of MpDIV2 could be a possible explanation for the difference in size between the plants of line #68 and of line #305.

The analysis of Mp*DIV2* overexpression further confirms Mp*DIV2* role in promoting cell proliferation and/or expansion and in regulating thallus development. The overexpression of Mp*DIV2* led to plants with increased width due to accelerated development and extended bifurcations. Both lines #3 and #11 have increased extension of the bifurcation branches. On the 15th day of growth, the branches of the plants of line #3 (**Figure 17. Visual analysis of MpDIV2 overexpression in *M. polymorpha* plants** – Wild type plants and MpDIV2 overexpression plant lines #3, #4 and #11 were grown and observed over the first 15 days of development after gemmae propagation. Plants were grown on half strength Gamborg B5 medium under long-day conditions (16 h light/ 8 h dark) at 20 °C and with light intensity between 40-45 µmol·m-2 s-1. Scale bar of: 0.5 mm for day 2; 1 mm from day 5 to 8; 2 mm for day 9; 5 mm from day 12 to 15.) appear to be bifurcating for the second time with little extension occurring beforehand. Usually, in WT plants, the first bifurcation occurs and then the resulting branches extend before a second bifurcation occurs. In the case of line #11, the branches resulting from the first bifurcation appear slightly wider than WT branches which could indicate early second bifurcation; however, this could be due to accelerated development and increased cell expansion and/or proliferation. The third line of *MpDIV2ox* plants analysed (#4), were visually and analytically very similar to WT. Although the plants are a little smaller in terms of length, their development appears near identical, bifurcation and development of gemma cups occurring simultaneously. These results indicate that overexpression of Mp*DIV2* in these plants is being somehow repressed or silenced. Increased expression of MpDIV2 in this line should be confirmed in future by RT-qPCR, for example.

Results show that Mp*DIV2ox* affects gemma cup development. Although development of bifurcation is accelerated, gemmae cups do not develop at all in the first 15 days. This could indicate that MpDIV2 somehow promotes development of some cells and tissues but represses the development of others. It could be happening via signalling pathways involving numerous genes involved in *M. polymorpha* development such as *CLE* family genes Mp*CLE1*, Mp*CLE2* and Mp*CLV1* shown to regulate meristem activity and size (Hirakawa et al., 2020; Hirakawa et al., 2019) and the Mp*GCAM1* R2R3-MYB transcription factor and auxin response factor Mp*ARF* which appear to be necessary for gemmae development (Yasui et al., 2019; Kato et al., 2017).

Overall, analysis of Mp*DIV2ox* in *M. polymorpha* plants and of knockout mutants indicate that Mp*DIV2* has a role in the promotion of thallus development and growth while simultaneously repressing the development of the gemmae responsible for asexual reproduction.

4.2. *DRIF* ancestral function remains uncertain

The exact way in which *DRIF* functions has largely remained unclear. It is known that DRIF proteins bind to DIV proteins and that these bind to specific DNA motifs. Bound together, they are thought to have a regulatory effect on expression of different genes that regulate a variety of processes. This effect is potentially carried out via the DUF3755 domain of DRIF, a domain of unknown function but that has been shown to interact with proteins of the WOX and KNOX families (Petzold et al., 2018). WOX and KNOX transcription factors are known to regulate a variety of functions related to development in higher plants such as stem cell and meristem maintenance and organ formation (Hake et al., 2004; van der Graaff et al., 2009). In non-vascular plants like *Physcomitrella patens* and *M. polymorpha* WOX type genes have been shown to regulate stem cell formation (Sakakibara et al., 2014) and KNOX type genes appear regulate the alternation between the haploid and diploid phases (Dierschke et al., 2021; Sakakibara et al., 2008).

In *M. polymorpha* plants, the exact function of DRIF is unknown but it is known to interact with MpDIV and, since its function is believed to be tied to its interaction with MpDIV, it is thought that it might be involved in promoting cell proliferation and/or expansion and in controlling thallus development. Mp*drif* knockout mutant plant analysis showed some phenotypical differences with wild type plants. Even though both the length and width of the plants measured was similar between mutant plants and wild type, the visual analysis showed that the two distinct mutant lines were analysed, line #15 and line #23, have slight phenotypical variations with WT plants and even between each other, but they generally appear to be slightly smaller and have an altered shape, in relation to WT plants. The plants of line #15 do not have much variation in size until the 15th day of development and, measurement wise, there are no more considerable differences. The phenotypical differences of this line are present in the shape and gemmae development. Normal *M. polymorpha* development occurs with sequential dichotomous branching with the first gemmae cups developing on the first bifurcation branches, usually leading to four initial gemmae cups. Although early development of two gemmae cups before bifurcation extension has been observed in WT plants, with line #15 it occurs consistently on almost all plants. Potentially related, plants from this line also show a retarded extension of the bifurcations, leading to the difference of width

on day 15. Sequencing of line #15 revealed a partial deletion of introns and a complete deletion of the second exon, which corresponds to part of the MYB domain of MpDRIF. Out of the three characteristic aromatic residues, the first Tryptophan would be the only one unaffected, as it is coded by the first exon. An initial idea for what could be happening in mutant plants from this line is that the MYB domain that binds to DIV is be mostly lost, and binding most likely does not occur, while the DUF3755 domain might still be able to interact with proteins and have a regulatory effect in plant development. However, this does not appear to be the case. If the remaining exons and introns of Mp*DRIF* in line #15 are spliced normally, the reading frame for the remainder of the protein after the first translated exon could be altered, with the direct connection of the first and third exon, and the protein would have an eventual early stop codon, completely altering and essentially erasing the DUF3755 domain. This would mean that plants of line #15 could result from a complete loss or alteration of function of the MpDRIF protein. However, there is a possibility that the DUF3755 domain could remain intact and functional or even that the remaining MYB domain is somehow active.

Plants from line #23 appeared to have a different phenotypical difference to WT plants in comparison to plants from line #15. They do not present the early development of two gemma cups, instead showing slowed development of gemmae cups in comparison to WT plants. Even in cases when a single gemmae cup develops before bifurcations, these are less developed. Unfortunately, sequencing of the Mp*DRIF* of plants of line #23 was unsuccessful but the phenotype was analysed nonetheless because significant phenotypical differences with WT plants were noted during screening. Thus, why the phenotype of line #23 is so different can only be supposed. It could be that one of the domains of MpDRIF was eliminated while the other was left unchanged, a complete deletion of the protein, losing even the beginning of the first domain, or changing only part of a domain, leading to an alteration of function.

Due to the lack of clear knockout mutants of Mp*DRIF*, other strategies are being employed. In an attempt to clarify the effect loss of Mp*DRIF* has on plant development, plants were transformed with pMpGWB318 vectors for overexpression of Mp*DRIF-SRDX.* SRDX is an EAR-repression domain which is fused to transcription factors to replace whatever regulatory effect they may have by repression of expression (Mahfouz et al., 2011). Although analysis of these plants is not complete, preliminary results show smaller, rounded plants, quite similar to the phenotype of Mp*div2* mutants. This similarity could potentially indicate that MpDRIF somehow regulates MpDIV2

function and that MpDIV2 promotes expression of different genes as opposed to repressing, although future analysis and further phenotypic analysis is required.

Mp*DRIF*ox plants show minimal alterations to shape. The only apparent phenotype is a flattened and rounded thallus, which is similar to Mp*div2* mutants. This result is unexpected. Under the current idea of the DDR module, DRIF and DIV proteins work together to regulate expression of different unknown genes and so it was believed that loss of one protein would have a similar phenotype to loss of the other, as would overexpression. This could just be a similarity between phenotypes and not be functionally related since the Mp*DRIF* overexpression plants are not significantly smaller than WT. However, it could also mean that how DIV and DRIF interact and how they collectively and individually affect plant development could be a more complex system. For example, it could be that an overabundance of DRIF protein somehow negates the effects of Mp*DIV2* function, perhaps via a negative feedback loop involving DRIF or it could indicate that DRIFs function is to regulate a possible inherent function of DIV.

In general, results obtained provide evidence that Mp*DRIF* is involved in regulating cell expansion and/or proliferation and in regulating the plants development and that its function could be tied to Mp*DIVs* and that MpDRIF regulates their activity. However, further study is necessary to determine its exact function in *M. polymorpha*.

4.3. *Arabidopsis RAD* overexpression affects development of *M. polymorpha*

The bryophyte ancestral land plant *M. polymorpha* has two *DIV* homologs, one *DRIF* homolog and no *RAD* homologs. The *RAD* gene was identified in gymnosperms and angiosperms and not in earlier plants, while the *DIV* and *DRIF* genes have been identified in algae and, in this very thesis, have been identified in eukaryote groups outside of green algae. The DDR regulatory module of higher plants self regulates via the antagonistic relationship between DIV and RAD proteins in which they compete to bind the DRIF protein. Since *M. polymorpha* does not have a *RAD* gene, this antagonism doesn't naturally occur in these species.

To learn more about the evolution of the relationship between DIV and DRIF and to see if antagonism could be established in *M. polymorpha*, the *RAD2* of *Arabidopsis thaliana* was overexpressed in *M. polymorpha* plants and the phenotype analysed. Plants showed decreased growth, and the shape of the thallus altered, quite similar to the phenotype observed in Mp*div2* knockout mutant plants, which has been extensively analysed by Coelho, S. (2019) and was

analysed in this thesis, as discussed above. The similarity between phenotypes indicates that the loss of *DIV2* and the overexpression of *RAD* individually have a comparable effect on cell proliferation and/or expansion the regulation of plant development.

At*RAD2*ox results could indicate that the DRIF MYB domain, responsible for interaction with the MYBI of DIV, is highly conserved over land plant evolution, since AtRAD2 might be interacting with MpDRIF, sequestering it and impeding binding of MpDRIF with MpDIV. However, the sequestering of MpDRIF alone doesn't seem to explain the similarity between Mp*div2* and At*RAD2*ox plant phenotypes. If this were the case, Mp*drif* mutant plants would have a phenotype more similar to that of Mp*div2* plants. A possible explanation is that RAD proteins are known to interact with other proteins. AtRAD2 has been potentially identified as interacting with various transcription factors from different families outside the DDR module (Trigg et al., 2017) and with MAP kinases (Popescu et al., 2009), known to regulate hormone response and plant development (Pearson et al., 2001) in *A. thaliana*. This means that At*RAD2*ox could be affecting plant development by binding to proteins other than MpDRIF that also interact with the MpDIV protein binding domain, MYBI. In a mechanism similar to what occurs with the DDR module of *A. majus*, overexpressed RAD proteins could be sequestering several different proteins that typically interact with the MpDIVs. By antagonizing the potential interactions, DIV function may be more compromised than with simple loss of Mp*DRIF*, leading to a similar effect on development to the loss of Mp*DIV2*.

From an evolutionary point of view, the phenotypic results of At*RAD2*ox show that RAD-DIV antagonism could be conserved between species millions of years of evolution apart. Evidence shows that liverworts existed around 470 mya (Shimamura, 2016) while angiosperms evolved between 250 and 140 mya (Sauquet et al., 2017), meaning that over at least 200 million years of evolution and the MYBI domain of DIV remains conserved enough to the point where AtRAD2 could have similar interaction targets in *M. polymorpha*.

The main conclusions that could be taken from these results are that MpDIV2 function might be at least partially dependant on binding with a variety of other proteins, including MpDRIF, that may regulate its function and that the MYB domains of the DDR module responsible for protein interaction are highly conserved across land plants, from a functional point of view.

Future studies could attempt to identify the potential other proteins with which MpDIVs may be interacting. To begin with, homologs should be searched for in the genome of *M. polymorpha* of the genes with which AtRAD2 has been shown to interact or at least genes of the

same families and then it could be determined whether interaction between these and MpDIV proteins occurs. This could be done via techniques such as Yeast-2-Hybrid (Y2H) screening (Lin & Lai, 2017), which could also be used to identify other potential interacting proteins.

4.4. Heterologous protein expression susceptible to cleavage in *E. coli*

Efforts to perform an Electromobility Shift Assay in order to ascertain whether MpDIV1 and MpDIV2 bind to DNA proved overall unsuccessful. However, the optimization of expression protocols produced promising results. The heterologous expression of Mp*DRIF* was achieved for the first time and the optimization of the physical and biological parameters of expression lead to heterologous expression of Mp*DIV1*, Mp*DIV2* and Mp*DRIF* in a soluble form, as shown by the SDS-PAGE and Western Blot analysis results. Although protein fused with GST was present in low quantities, if present at all, the quantity of obtained soluble protein was increased, relative to initial results. The main issue that seems to be affecting obtention of properly tagged, soluble protein is cleavage occurring between the DDR module proteins and the GST tag.

SDS-PAGE and western blot results presented bands of heterologous protein at unexpected molecular weights. In the case of Mp*DRIF* expression, initial results presented the band that probably corresponds to the MpDRIF-GST fusion has a MW (at least 66 kDa) several kDa above the expected 58 kDa while following results, using a different MW marker, present MpDRIF-GST between 50 and 75 kDa, which could be above 66 kDa but appears to be closer to the middle of the two, and thus closer to the theorized MW. There are several explanations for why the initial results show the potential MpDRIF-GST fusion at such a high MW. The most likely of these is that the MW marker used was faulty, but it would have to be specifically for the 66 kDa band and above as the MW of GST run as a positive control is correctly just below the 31 kDa band of the marker. This is the most likely option because when using a newer molecular marker, the observed MW of the band probably corresponding to MpDRIF-GST is closer to the theorized MW of the fusion protein.

In the case of Mp*DIV1* and Mp*DIV2*, the bands with highest intensity were below the theorized kDa for both protein fusions and the bands corresponding to GST had an elevated intensity in comparison to MpDRIF lanes, indicating that cleavage could be separating the proteins from the GST tag. If this is the case, the MW observed for MpDIV1 and MpDIV2 are above the theorized MW of the proteins (36.6 and 32.3 kDa, respectively). If cleavage is occurring between GST and the proteins then the higher intensity bands should correspond to the MW of the proteins

and of GST however the bands are at higher MW, between 37 and 50 kDa for MpDIV1 and around 50 kDa for MpDIV2.

What is possible is that cleavage is occurring to a specific region present in both DIV proteins that leads to a bigger C-terminal cut in the MpDIV1 protein than in the MpDIV2. This is supported by the fact that these bands are bound by antibodies in the western blot analysis, meaning that GST, or at least part of GST, is probably still fused with the DIV proteins.

In the future, to further the optimization process and eventually perform a EMSA without issues, a different tag, such as a His-tag, could be used, does assessing whether the cleavage occurring to the DDR module proteins in the *E. coli* model is due to GST or the proteins themselves. If this proves unsuccessful, a second option could be to change cell model and use yeast species for heterologous expression.

4.5. What to expect from pMpGWB constructs

The cloning procedures carried out with the pMpGWB series of gateway vectors were performed with the intention of fusing the DIV and DRIF proteins of *M. polymorpha* with different tags and overexpressing the constructs in WT plants.

The Citrine tags present in the pMpGWB208 and pMpGWB308 vectors should allow for the localization of the MpDIV and MpDRIF proteins at the intracellular level. By knowing more specifically where these proteins are present in *M. polymorpha*, greater understanding or new ideas for their ancestral function could arise. In the future, different tags with different emittance wavelengths should be obtained and combined with the proteins so that the localization of MpDIVs and MpDRIF could be compared *in planta*. For example, considering that overexpression of At*RAD2* appears to have a phenotypic effect on *M. polymorpha* plants, At*RAD2* could be combined with a fluorescent tag, and it could be confirmed whether antagonism is occurring, where AtRAD2 is binding MpDRIF and impeding it from entering the nucleus and interaction with MpDIV proteins, thus indicating that the phenotype could be caused by deregulation of MpDIV. Studies of this kind have been performed before with *A. majus* DDR proteins transformed in tobacco leaves (Raimundo et al., 2013). They transformed *Nicotiana benthamiana* leaves with *A. majus RAD*, *DIV* and *DRIFs* fused to fluorescent proteins with different colours. They observed that when only DIV and DRIF proteins were present, both were located only in the nucleus. When RAD was present, DRIF proteins were located both in the nucleus and the cytoplasm, as were RAD proteins, indicating that RAD

was sequestering DRIF proteins in the nucleus and into the cytoplasm. A similar idea could be attempted in *M. polymorpha*.

GR is commonly used to control the activation of plant transcription factors, which becomes dependant on the presence of the synthetic steroid hormone dexamethasone (dex) (Picard et al., 1990; Schena et al., 1991). When the hormone is absent, GR is localized in the cytoplasm and so is any transcription factor to which it is fused, effectively deactivating the transcription factor until dex is added to the medium in which plants are growing. As of now, MpDRIF-GR and MpDIV2-GR constructs were transformed into WT *M. polymorpha* plants so overexpression of the genes can be controlled and eventually activated at later points in development to hopefully learn more about Mp*DIV* and Mp*DRIF* function in controlling plant development. In the future, the constructs will be transformed into mutant plants with the corresponding knocked out gene, to evaluate whether overexpression of the gene in a mutant background can retrieve WT or overexpression phenotype at later points in plant development.

The pMpGWB318 vectors have SRDX tags that are fused to the C-terminal end of the inserted proteins. Mp*DIV1*, Mp*DIV2* and Mp*DRIF* were inserted cloned into the vector for overexpression fused with SRDX tags. As discussed in 4.2, SRDX is an EAR-repression domain that is commonly used together with transcription factors to create chimeric repressors (Mahfouz et al., 2011), in theory, replacing the usual activity of the transcription factor with repression. The intention of having transformed *M. polymorpha* with the cloned constructs is to learn more about the function of the different transcription factors. In theory, if MpDIV1, MpDIV2 and MpDRIF proteins promote expression of the genes they potentially interact with, then overexpression with SRDX should yield phenotypes similar to that of knockout mutant plants for the genes and could equate to a loss of function. Of course, the opposite is also true and if the transcription factors analysed with this technique turn out to repress expression, then plants with these constructions will have phenotypes similar to what occurs when that transcription factor is overexpressed. The preliminary results discussed in 4.2 serve as early preview of the potential use for this tool. By overexpressing Mp*DRIF-SRDX* in *M. polymorpha* plants, a phenotype was observed similar to loss of Mp*DIV2*, potentially indicating that MpDIV2 and MpDRIF naturally have a non-repressive effect on gene expression. With this vector we can learn more about the nature of the DIV and DRIF proteins of *M. polymorpha* and know more about their function.

With the pMpGWB321 vectors, Mp*DIV1*, Mp*DIV2* and Mp*DRIF* were fused with both the GR and SRDX tags, fused together. Essentially, it is a combination of the later in development

activation potential of the GR tag and the repression caused by SRDX. As such, transformation of plants with this construct was done with the intention of inducing repression at later points of development in the hopes of uncovering more about the function of DIV and DRIF proteins and specifically how that function changes and evolves during plant development.

Overall, this part of the project aims to open a variety of avenues through which the function of the DDR module in *M. polymorpha* can be further studied and unveiled, via a multidisciplinary approach that has already begun to show promise.

4.6. MYBII of *DIV* homologs is highly conserved throughout evolution

Across land plants, there is evidence to support that the MYB domains of DDR module responsible for protein interaction are highly conserved. However, sequence analysis of DIV and DRIF protein homologs outside of land plants, found throughout the other groups of Archaeplastida and within Cryptista, revealed that conservation of amino acids in the MYB protein binding domains was relatively low, especially compared to the conservation of the MYBII domain of DIV homologs. Even excluding the characteristic SHAQKYF motif and aromatic residues, the rest of the MYBII domain is nearly all conserved in more than 90% of the species analysed, indicating massive selective pressure to keep the MYBII domain conserved. The characteristic aromatic residues were all 100% conserved except for the Tyrosine (Annex D, **figure D-1**, position 140) contained in the SHAQKYF motif of CryptoDIV (from the unclassified species *Cryptophyceae sp.*) which has been replaced with a Phenylalanine residue. This substitution is the most common for Tyrosine, as the difference between the aromatic residues is a hydroxyl group present in Tyrosine but absent in Phenylalanine (Betts & Russell, 2003), so function may be maintained.

4.7. *DIV* homologs of Chlorophyte algae may have altered function

As shown, the MYBI domain of DIV proteins analysed is not as conserved as the MYBII domain. More detailed analysis of the amino acids revealed that the characteristic aromatic residues that are used to define MYB domains are not very conserved in the MYBI protein binding domain. This is especially apparent in the Chlorophyte algae of the group Trebouxiophyceae, where Tryptophan and Tyrosine residues are, in many cases, replaced by non-aromatic amino acids (Annex D, **figure D-1**, positions 1, 31 and 52). When Tryptophan or Tyrosine are replaced with

the other or with Phenylalanine there is a higher chance that amino acid function could be retained than when replaced by non-aromatic amino acids. Replacements with non-aromatic amino acids are present in MYBI domains from Chlorophyceae species but they are less frequent and the alterations to the aromatic residues are conserved between closely related species, which is not present in DIVs from Trebouxiophyceae species. Within Trebouxiophyceae, when DIV protein sequences were retrieved, no *DIV* homolog was found in the species *Micractinium conductrix* and *Trebouxia sp.*, however both species have *DRIF* homologs. These species could have lost DIV at some point in the past. This and the fact that the aromatic residues of the protein binding domain of DIVs from Trebouxiophyceae are less conserved indicates that there is a low selective pressure to maintain the DIV protein binding domain and that it may have evolved to the point where it no longer interacts with DRIF and has lost function. This combined with the loss of the aromatic residues could mean that some new function may exist, independent of the Tryptophan and Tyrosine amino acids and that it may no longer involve binding to DRIF. Since the MYBI domain has lost its characteristic amino acids in many of these DIV proteins, they may even have to be considered new non-DIV proteins after further investigation.

Curiously, the Trebouxiophyceae DRIF proteins have highly conserved characteristic aromatic residues in the MYB domain (Annex D, **figure D-2**, positions 2, 59 and 85), responsible for binding with DIV. If selective pressure exists to maintain this domain but the MYBI domain of DIV to which it binds has been altered, it could mean that, in these species, DRIF proteins may have evolved a new function, independent of DIV or that the MYB domain binds to other proteins other than DIV.

All of the ideas discussed here are potential explanations that will need further studies to ascertain their viability. Protein interaction studies, such as a yeast-two-hybrid assay, could be used to determine whether the DRIF and DIV proteins of the species in question are able to bind to one another and could be used to search for unknown proteins with which DIV and DRIF may interact.

4.8. *DIV* apparently older than *DRIF* and older than the green lineage

DRIF homologs were only found in Chloroplastida, the green lineage, while DIV homologs were found outside of Archaeplastida, in the Cryptista phylum. Until now, it was believed that DIV and DRIF could have originated together via duplications a pre-existing MYB domain (Raimundo et al., 2018). This could mean that DIV originated in a common ancestor of both the Cryptista and

Archaeplastida clades, which would place the evolution of DIV during the Paleoproterozoic, between 2200 and 1600 million years ago (mya) and could implicate that DRIF evolved from duplication of the DIV MYB domains, but this would need further evidence to back up. If DRIF evolved only in Chloroplastida, then its origin is placed during the Mesoproterozoic, between 1600 and 1000 mya (Strassert et al., 2021).

DIV is only known to interact with DRIF proteins. As such, the fact that *DIV* potentially originated before *DRIF* indicates that DIV proteins could have had had or still have an ancestral function that predates its interaction with DRIF and could indicate that the MYBI protein interacting domain of DIV could be able to interact with other proteins. This idea is supported by the results obtained from At*RAD2* overexpression in *M. polymorpha*. Besides this, it could also mean that DIV transcription factors can influence expression without additional ligands, just by binding to certain DNA sequences. Considering these hypotheses, a look at the relationship between the more ancestral *DIVs* and the conservation of the MYBI domain between these could reveal more about their feasibility.

According to the phylogenetic tree analysis (**Figure 26**) the *DIVs* of Cryptophyta and Glaucophyta species are grouped together and apart from those of Chloroplastida species, corresponding to the phylogenetic relationship and taxonomy of the species. Analysis of the MYBI domains of these DIV proteins (**Figure 27**, **figure 26**, Annex D) shows that the DIV of the Glaucophyta species *Cyanophora paradoxa*, CyaparDIV, has the main aromatic residues conserved, except for the central Tryptophan, which was substituted for a Phenylalanine residue. While substitution of a Tyrosine with a Phenylalanine is typically of little consequence, Tryptophan is unique in terms of chemistry and size which could mean that, even when substituted by another aromatic residue, function is lost (Betts & Russell, 2003). The DIVs of the Cryptophyte species, GuithDIV1 and GuithDIV2, and CryptoDIV, have all three characteristic aromatic residues. So, while loss of function could have occurred in CyaparDIV due to loss of the Tryptophan or in all three proteins due to other substituted amino acids, it is less likely to be an option. In fact, these ancestral DIVs have a higher conservation of these characteristic aromatic residues than many of the species of Chloroplastida analysed, species in which DRIF homologs have been found.

There is, however, a third group of algal species analysed, within Chloroplastida, in which *DIV* is present and *DRIF* is not. The genomes of three species of Ostreococcus were analysed and all three had a *DIV* homolog and none had a *DRIF* homolog. Analysis of the amino acid sequence of the MYBI domain of the three species indicates that *Ostreococcus tauri* has all three

characteristic aromatic while both OstluDIV and OstrcDIV Phenylalanine substitutions in place of the Tyrosine residue, which is known to be a common replacement that is likely to have little consequence. Ostreococcus species are commonly known as one of the smallest eukaryotic species and thus have many adaptations such as substantial gene loss, including a great variety of transcription factors present in the genomes of land plants and in other green algae species (Palenik et al., 2007). This could explain the loss of *DRIF* in this group.

Overall, the analysis of the MYBI domains of the analysed species indicates that there is no correlation between conservation of the characteristic Tryptophan and Tyrosine residues in the MYBI of *DIV* homologs and the presence of a *DRIF* homolog in the species and that the conservation of these amino acids remains elevated when *DRIF* is not present. This last observation points to another possible conclusion. That ancestral *DIV* homologs present in algal species, even outside the green lineage, might have a function independent of *DRIF*, specifically involving the MYBI domain.

However, without more information on whether interactions between the DIV and DRIF proteins are conserved in algal species, without knowledge of proteins other than DRIF that interact with ancestral DIVs and without more genomes from Cryptophyta and Glaucophyta species, or ancestral species to these, it is not yet possible to conclude with certainty whether *DIV* evolved before *DRIF* and whether has an ancestral function independent of *DRIF.*

4.9. *DIV* lost in red algae lineages

Archaeplastida is a commonly accepted clade that includes three major groups that originated from a single endosymbiotic event, these being Rhodophyta, Glaucophyta and Chloroplastida. Genomes of species from the three groups were analysed and, in both Chloroplastida and Glaucophyta species, *DIV* homologs were found while none were found in the genomes of seven different red algal species present on the Phycocosm resource (Grigoriev et al., 2020). Since Chloroplastida and Glaucophyta have common origin with red algae the first idea is that *DIV* evolved in an ancestor common to Chloroplastida and Glaucophyta but not common to the Rhodophyta lineage. This could be the case but, as was discussed above, *DIV* homologs were found in the genomes of two species in the Cryptista supergroup, which is a monophyletic group outside of but considered to be a sister group to the Archaeplastida supergroup (Burki et al., 2020).

The presence of *DIV* homologs in Cryptophyta species can be explained a few ways. It could be that horizontal gene transfer (HGT) from a species with a *DIV* homolog. This option is less

likely because phylogenetic analysis (**Figure 26**) grouped the DIV proteins of the Cryptophyta species together and distinctly apart from those of Glaucophyta and Chloroplastida species and, if the *DIV* of Cryptophyta had originated via HGT from those species, they would be more closely grouped on the phylogenetic tree. Of course, there is the possibility that the HGT occurred before the establishment of the groups we have today, with ancestors of Glaucophyta and Cryptophyta for example, which would explain the positioning on the phylogenetic tree, meaning this explanation cannot be completely discarded.

Another possibility is that endosymbiotic gene transfer (EGT) occurred from ancestral red algae when secondary endosymbiosis occurred. EGT is a special case of HGT in which genes from the symbiont are transferred into the genome of the host cell (Henze et al., 2002). Cryptophyta are widely accepted as having originated from a secondary endosymbiosis event in which endosymbiosis occurred with a cell already containing plastids from a primary endosymbiosis event, in this case, a cell from the lineage of red algae (Burki et al., 2020). It is then possible that *DIV* evolved in an ancestor of all Archaeplastida, was present in the red algae involved in the secondary endosymbiotic event that originated the Cryptista supergroup, and that this *DIV* was transferred to the Cryptista genome. For this to have occurred, to our current knowledge, *DIV* would then have been subsequently lost in all red algae, which leads to the third hypothesis.

It is possible that *DIV* evolved in a common ancestor to the Cryptista and Archaeplastida supergroups and was subsequently lost in Rhodophyta (**Figure 39**). This possibility is supported by the theorized genome contraction event that occurred when the most recent common ancestor of all red algae adapted to extreme environments (Petroll et al., 2021). The genome contraction and subsequent gene loss that occurred in Rhodophyta would also support the EGT hypothesis for why *DIV* is found in Cryptista genomes and not in red algae.

With the current information, it is not possible to determine with certainty which of these occurred, although it seems more likely that the EGT hypothesis or the Rhodophyta loss of *DIV* hypothesis are correct. To differentiate between these and determine which is more likely a possible course of action would be to analyse the genomes of Cryptista ancestral species that do not possess plastids. If these were to have *DIV*, it would mean that the hypothesis that the *DIV* evolved in a common ancestor of Cryptista and Archaeplastida would be the most likely option, essentially discarding the EGT hypothesis, with the only viable alternative option of the three discussed, other than loss of *DIV* in Rhodophyta, being that of HGT having occurred even before the Cryptista supergroup evolved.

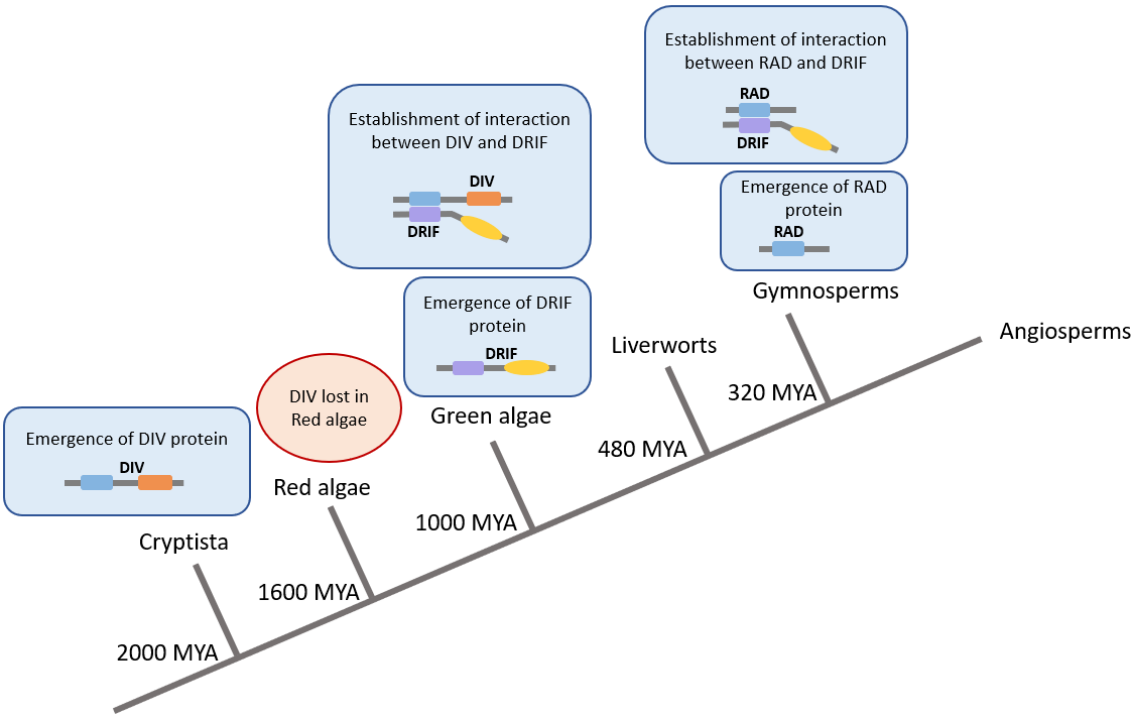

Figure 39. Schematic of possible evolution of the DIV, DRIF and RAD protein families. Representation of the evolution of the DDR module proteins based on new findings pertaining to early DIV and DRIF evolution. It is proposed that DIV proteins emerged in a common ancestor to the Cryptista and Archaeplastida supergroups and was lost in an ancestor of Rhodophyta (Red algae). DRIF emerged in Chloroplastida (Green algae) and the interaction between DIV and DRIF was established. RAD emerged in gymnosperms and the interaction between RAD and DRIF was established (Raimundo, et al., 2018).

4.10. Conclusions

The effort to uncover the ancestral function and the evolution of the DDR regulatory module is still ongoing as progress is made to understand this gene network. This thesis furthered many of the different approaches being employed and achieved several new results.

As was previously believed, analysis of plant phenotypes indicates that Mp*DIV2* promotes cell proliferation and/or expansion and influences plant shape in *M. polymorpha*. Phenotype analysis also indicated that Mp*DRIF* appears to have a role in regulating plant development, which could occur by influencing Mp*DIV* function. This was theorized to be the case but had yet to be observed until this thesis.

Results from overexpression of At*RAD2* in *M. polymorpha* plants indicate that MpDIV could be interacting with proteins other than MpDRIF. To my current knowledge, studies have yet to observed DIV proteins interacting with proteins other than DRIF proteins. Possible implications are that either these interactions exist and haven't been observed yet in higher plants or that ancestral DIV proteins interacted with and were regulated by various proteins and overtime became more

specialized, only interacting with DRIF. This analysis also indicated that the protein binding domain of DIV, MYBI, is highly conserved throughout plant evolution.

The preparation of Gateway Cloning pMpGWB constructs with *M. polymorpha DIV*s and *DRIF* was successful. They will be a useful tool in future studies and initial results have already begun to show promise. The different potential uses above discussed and many that could yet to be thought of will help in better understanding the ancestral function of *DIV* and *DRIF*.

Regarding the protein-DNA interaction analysis, heterologous expression of Mp*DIV1*, Mp*DIV2* in *E. coli* was improved and of Mp*DRIF* was achieved. Whether MpDIVs bind to DNA was not determined via EMSA due to issues with protein integrity. MpDIV proteins fused with GST appear to be susceptible to cleavage in *E. coli*, potentially losing function. As of now, work has begun experimenting with His tags and, if this were to have similar issues, perhaps eventually a different organism, such as a yeast could be considered for heterologous expression of MpDIVs and MpDRIF proteins.

Phylogenetic analysis revealed that, contrary to previous belief, *DIV* is probably older than *DRIF* and thus both did not arise simultaneously. *DIV* probably arouse in an ancestor to both the Archaeplastida and Cryptista clades before the evolution of *DRIF* in Chloroplastida. These results open the possibility that DRIF arose from a duplication event of a DIV protein. Until now, DIV and DRIF were thought to have arisen around the same time via duplication events of a single MYB domain. Results also imply that DIV must have had some earlier function, independent of DRIF, perhaps by binding with other proteins. Additionally, analysis indicated that *DIV* was lost in red algae, which could be due to a genomic contraction event in an ancestor of this group. In order to make the tree of DIV evolution clearer, DIV could be searched for in the genomes of ancestor species of the red algae and the Cryptista and both genes should be searched for in more genomes of the groups outside the green lineage.

This thesis answered many questions, some even unexpectedly, and lead to even more unanswered. Previous conclusions were further confirmed, protocols were optimized, old ideas were brought into question, new conclusions were drawn, and new questions have been formed. Overall, more is understood about the nature and evolution of *DIV* and *DRIF* and part of the way has been paved to discover the essential function of the DDR regulatory module.

5. Bibliography

- Almeida, M. J. B. de. (2019). Establishment and evolution of gene regulatory networks. In *repositorium.sdum.uminho.pt*. http://hdl.handle.net/1822/65181

- Almeida, M. J. B., Sobral, R. S., Romão, C. V., & Costa, M. M. R. (2018). Establishment and evolution of a gene regulatory network controlling flower asymmetry. *Free Radical Biology and Medicine*, *120*, S138. https://doi.org/10.1016/j.freeradbiomed.2018.04.454

- Berger, F., Bowman, J. L., & Kohchi, T. (2016). Marchantia. *Current Biology*, *26*(5), R186–R187. https://doi.org/10.1016/j.cub.2015.12.013

- Betts, M. J., & Russell, R. B. (2003). Amino-Acid Properties and Consequences of Substitutions. In *Bioinformatics for Geneticists* (pp. 311–342). https://doi.org/10.1002/9780470059180.ch13

- Bowman, J. L. (2016). A Brief History of Marchantia from Greece to Genomics. *Plant & Cell Physiology*, *57*(2), 210–229. https://doi.org/10.1093/pcp/pcv044

- Bowman, J. L., Araki, T., & Kohchi, T. (2016). Marchantia: Past, Present and Future. *Plant and Cell Physiology*, *57*(2), 205–209. https://doi.org/10.1093/pcp/pcw023

- Bowman, J. L., Kohchi, T., Yamato, K. T., Jenkins, J., Shu, S., Ishizaki, K., Yamaoka, S., Nishihama, R., Nakamura, Y., Berger, F., Adam, C., Aki, S. S., Althoff, F., Araki, T., Arteaga-Vazquez, M. A., Balasubrmanian, S., Barry, K., Bauer, D., Boehm, C. R., & Briginshaw, L. (2017). Insights into Land Plant Evolution Garnered from the Marchantia polymorpha Genome. *Cell*, *171*(2), 287-304.e15. https://doi.org/10.1016/j.cell.2017.09.030

- Burki, F., Roger, A. J., Brown, M. W., & Simpson, A. G. B. (2020). The New Tree of Eukaryotes. *Trends in Ecology & Evolution*, *35*(1), 43–55. https://doi.org/10.1016/j.tree.2019.08.008

- Coelho, S. (2019). *Study of ancestral genes in Marchantia polymorpha* [MSc].

- Corley, S. B., Carpenter, R., Copsey, L., & Coen, E. (2005). Floral asymmetry involves an interplay between TCP and MYB transcription factors in Antirrhinum. *Proceedings of the National Academy of Sciences*, *102*(14), 5068–5073. https://doi.org/10.1073/pnas.0501340102

- De Smet, R., & Van de Peer, Y. (2012). Redundancy and rewiring of genetic networks following genome-wide duplication events. *Current Opinion in Plant Biology*, *15*(2), 168–176. https://doi.org/10.1016/j.pbi.2012.01.003

- Deblaere, R., Bytebier, B., De Greve, H., Deboeck, F., Schell, J., Van Montagu, M., & Leemans, J. (1985). Efficient octopine Ti plasmid-derived vectors forAgrobacterium-mediated gene transfer to plants. *Nucleic Acids Research*, *13*(13), 4777–4788. https://doi.org/10.1093/nar/13.13.4777

- Dierschke, T., Flores-Sandoval, E., Rast-Somssich, M. I., Althoff, F., Zachgo, S., & Bowman, J. L. (2021). Gamete expression of TALE class HD genes activates the diploid sporophyte program in Marchantia polymorpha. *ELife*, *10*. https://doi.org/10.7554/elife.57088

- Durfee, T., Nelson, R., Baldwin, S., Plunkett, G., Burland, V., Mau, B., Petrosino, J. F., Qin, X., Muzny, D. M., Ayele, M., Gibbs, R. A., Csorgo, B., Posfai, G., Weinstock, G. M., & Blattner, F. R. (2008). The Complete Genome Sequence of Escherichia coli DH10B: Insights into the Biology of a Laboratory Workhorse. *Journal of Bacteriology*, *190*(7), 2597–2606. https://doi.org/10.1128/jb.01695-07

- Flores-Sandoval, E., Romani, F., & Bowman, J. L. (2018). Co-expression and Transcriptome Analysis of Marchantia polymorpha Transcription Factors Supports Class C ARFs as Independent Actors of an Ancient Auxin Regulatory Module. *Frontiers in Plant Science*, *9*. https://doi.org/10.3389/fpls.2018.01345

- Furumizu, C., Hirakawa, Y., Bowman, J. L., & Sawa, S. (2018). 3D Body Evolution: Adding a New Dimension to Colonize the Land. *Current Biology*, *28*(15), R838–R840. https://doi.org/10.1016/j.cub.2018.06.040

- Galego, L., & Almeida, J. (2002). Role of DIVARICATA in the control of dorsoventral asymmetry in Antirrhinum flowers. *Genes & Development*, *16*(7), 880–891. https://doi.org/10.1101/gad.221002

- Garner, M. M., & Revzin, A. (1981). A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic Acids Research*, *9*(13), 3047–3060. https://doi.org/10.1093/nar/9.13.3047

- Grigoriev, I. V., Hayes, R. D., Calhoun, S., Kamel, B., Wang, A., Ahrendt, S., Dusheyko, S., Nikitin, R., Mondo, S., Salamov, A., Shabalov, I., & Kuo, A. (2020). PhycoCosm, a comparative algal genomics resource. *Nucleic Acids Research*, *49*(D1), D1004–D1011. https://doi.org/10.1093/nar/gkaa898

- Hake, S., Smith, H. M. S., Holtan, H., Magnani, E., Mele, G., & Ramirez, J. (2004). THE ROLE OF KNOX GENES IN PLANT DEVELOPMENT. *Annual Review of Cell and Developmental Biology*, *20*(1), 125–151. https://doi.org/10.1146/annurev.cellbio.20.031803.093824

- Hamaguchi, A., Yamashino, T., Koizumi, N., Kiba, T., Kojima, M., Sakakibara, H., & Mizuno, T. (2008). A Small Subfamily ofArabidopsis RADIALIS-LIKE SANT/MYBGenes: A Link to HOOKLESS1-Mediated Signal Transduction during Early Morphogenesis. *Bioscience, Biotechnology, and Biochemistry*, *72*(10), 2687–2696. https://doi.org/10.1271/bbb.80348

- Henze, K., Martin, W., & Schnarrenberger, C. (2002). Endosymbiotic Gene Transfer. *Horizontal Gene Transfer*, 351–XII. https://doi.org/10.1016/b978-012680126-2/50034-7

- Hirakawa, Y., Fujimoto, T., Ishida, S., Uchida, N., Sawa, S., Kiyosue, T., Ishizaki, K., Nishihama, R., Kohchi, T., & Bowman, J. L. (2020). Induction of Multichotomous Branching by CLAVATA Peptide in Marchantia polymorpha. *Current Biology*, *30*(19), 3833-3840.e4. https://doi.org/10.1016/j.cub.2020.07.016

- Hirakawa, Y., Uchida, N., Yamaguchi, Y. L., Tabata, R., Ishida, S., Ishizaki, K., Nishihama, R., Kohchi, T., Sawa, S., & Bowman, J. L. (2019). Control of proliferation in the haploid meristem by CLE peptide signaling in Marchantia polymorpha. *PLOS Genetics*, *15*(3), e1007997. https://doi.org/10.1371/journal.pgen.1007997

- Ishizaki, K., Nishihama, R., Ueda, M., Inoue, K., Ishida, S., Nishimura, Y., Shikanai, T., & Kohchi, T. (2015). Development of Gateway Binary Vector Series with Four Different Selection Markers for the Liverwort Marchantia polymorpha. *PLOS ONE*, *10*(9), e0138876. https://doi.org/10.1371/journal.pone.0138876

- Ishizaki, K., Nishihama, R., Yamato, K. T., & Kohchi, T. (2015). Molecular Genetic Tools and Techniques for Marchantia polymorpha Research. *Plant and Cell Physiology*, *57*(2), 262–270. https://doi.org/10.1093/pcp/pcv097

- Jin, H., & Martin, C. (1999). Multifunctionality and diversity within the plant MYB-gene family. *Plant Molecular Biology*, *41*(5), 577–585. https://doi.org/10.1023/A:1006319732410

- Kato, H., Kouno, M., Takeda, M., Suzuki, H., Ishizaki, K., Nishihama, R., & Kohchi, T. (2017). The Roles of the Sole Activator-Type Auxin Response Factor in Pattern Formation of Marchantia polymorpha. *Plant and Cell Physiology*, *58*(10), 1642–1651. https://doi.org/10.1093/pcp/pcx095

- Leliaert, F., Smith, D. R., Moreau, H., Herron, M. D., Verbruggen, H., Delwiche, C. F., & De Clerck, O. (2012). Phylogeny and Molecular Evolution of the Green Algae. *Critical Reviews in Plant Sciences*, *31*(1), 1–46. https://doi.org/10.1080/07352689.2011.615705

- Li, L., Wang, S., Wang, H., Sahu, S. K., Marin, B., Li, H., Xu, Y., Liang, H., Li, Z., Cheng, S., Reder, T., Çebi, Z., Wittek, S., Petersen, M., Melkonian, B., Du, H., Yang, H., Wang, J., Wong, G. K.-S., & Xu, X. (2020). The genome of Prasinoderma coloniale unveils the existence of a third phylum within green plants. *Nature Ecology & Evolution*, *4*(9), 1220–1231. https://doi.org/10.1038/s41559-020-1221-7

- Lin, J.-S., & Lai, E.-M. (2017). Protein–Protein Interactions: Yeast Two-Hybrid System. *Methods in Molecular Biology*, *1615*, 177–187. https://doi.org/10.1007/978-1-4939-7033-9_14

- Lipsick, J. S. (1996). One billion years of Myb. *Oncogene*, *13*(2), 223–235.

- Lu, S., Wang, J., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Marchler, G. H., Song, J. S., Thanki, N., Yamashita, R. A., Yang, M., Zhang, D., Zheng, C., Lanczycki, C. J., & Marchler-Bauer, A. (2019). CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Research*, *48*. https://doi.org/10.1093/nar/gkz991

- Luo, D., Carpenter, R., Copsey, L., Vincent, C., Clark, J., & Coen, E. (1999). Control of Organ Asymmetry in Flowers of Antirrhinum. *Cell*, *99*(4), 367–376. https://doi.org/10.1016/s0092-8674(00)81523-8

- Luo, D., Carpenter, R., Vincent, C., Copsey, L., & Coen, E. (1996). Origin of floral asymmetry in Antirrhinum. *Nature*, *383*(6603), 794–799. https://doi.org/10.1038/383794a0

- Machemer, K., Shaiman, O., Salts, Y., Shabtai, S., Sobolev, I., Belausov, E., Grotewold, E., & Barg, R. (2011). Interplay of MYB factors in differential cell expansion, and consequences for tomato fruit development. *The Plant Journal*, *68*(2), 337–350. https://doi.org/10.1111/j.1365-313x.2011.04690.x

- Mahfouz, M. M., Li, L., Piatek, M., Fang, X., Mansour, H., Bangarusamy, D. K., & Zhu, J.-K. (2011). Targeted transcriptional repression using a chimeric TALE-SRDX repressor protein. *Plant Molecular Biology*, *78*(3), 311–321. https://doi.org/10.1007/s11103-011-9866-x

- Miroux, B., & Walker, J. E. (1996). Over-production of Proteins inEscherichia coli: Mutant Hosts that Allow Synthesis of some Membrane Proteins and Globular Proteins at High Levels. *Journal of Molecular Biology*, *260*(3), 289–298. https://doi.org/10.1006/jmbi.1996.0399

- Montgomery, S. A., Tanizawa, Y., Galik, B., Wang, N., Ito, T., Mochizuki, T., Akimcheva, S., Bowman, J. L., Cognat, V., Maréchal-Drouard, L., Ekker, H., Hong, S.-F., Kohchi, T., Lin, S.-S., Liu, L.-Y. D., Nakamura, Y., Valeeva, L. R., Shakirov, E. V., Shippen, D. E., & Wei, W.-L. (2020). Chromatin Organization in Early Land Plants Reveals an Ancestral Association between H3K27me3, Transposons, and Constitutive Heterochromatin. *Current Biology*, *30*(4), 573-588.e7. https://doi.org/10.1016/j.cub.2019.12.015

- Naramoto, S., Hata, Y., & Kyozuka, J. (2020). The origin and evolution of the ALOG proteins, members of a plant-specific transcription factor family, in land plants. *Journal of Plant Research*, *133*(3), 323–329. https://doi.org/10.1007/s10265-020-01171-6

- Palenik, B., Grimwood, J., Aerts, A., Rouzé, P., Salamov, A., Putnam, N., Dupont, C., Jorgensen, R., Derelle, E., Rombauts, S., Zhou, K., Otillar, R., Merchant, S. S., Podell, S., Gaasterland, T., Napoli, C., Gendler, K., Manuell, A., Tai, V., & Vallon, O. (2007). The tiny eukaryoteOstreococcusprovides genomic insights into the paradox of plankton speciation. *Proceedings of the National Academy of Sciences*, *104*(18), 7705–7710. https://doi.org/10.1073/pnas.0611046104

- Pearson, G., Robinson, F., Beers Gibson, T., Xu, B., Karandikar, M., Berman, K., & Cobb, M. H. (2001). Mitogen-Activated Protein (MAP) Kinase Pathways: Regulation and Physiological Functions. *Endocrine Reviews*, *22*(2), 153–183. https://doi.org/10.1210/edrv.22.2.0428

- Petroll, R., Schreiber, M., Finke, H., Cock, J. M., Gould, S. B., & Rensing, S. A. (2021). Signatures of Transcription Factor Evolution and the Secondary Gain of Red Algae Complexity. *Genes*, *12*(7), 1055. https://doi.org/10.3390/genes12071055

- Petzold, H. E., Chanda, B., Zhao, C., Rigoulot, S. B., Beers, E. P., & Brunner, A. M. (2018). DIVARICATA AND RADIALIS INTERACTING FACTOR (DRIF) also interacts with WOX and KNOX proteins associated with wood formation in Populus trichocarpa. *The Plant Journal*, *93*(6), 1076–1087. https://doi.org/10.1111/tpj.13831

- Picard, D., Schena, M., & Yamamoto, K. R. (1990). An inducible expression vector for both fission and budding yeast. *Gene*, *86*(2), 257–261. https://doi.org/10.1016/0378-1119(90)90287-2

- Popescu, S. C., Popescu, G. V., Bachan, S., Zhang, Z., Gerstein, M., Snyder, M., & Dinesh-Kumar, S. P. (2009). MAPK target networks in Arabidopsis thaliana revealed using functional

protein microarrays. *Genes & Development*, *23*(1), 80–92. https://doi.org/10.1101/gad.1740009

- Pressel, S., & Duckett, J. G. (2019). Do motile spermatozoids limit the effectiveness of sexual reproduction in bryophytes? Not in the liverwort Marchantia polymorpha. *Journal of Systematics and Evolution*, *57*(4), 371–381. https://doi.org/10.1111/jse.12528

- Raimundo, J., Sobral, R., Bailey, P., Azevedo, H., Galego, L., Almeida, J., Coen, E., & Costa, M. M. R. (2013). A subcellular tug of war involving three MYB-like proteins underlies a molecular antagonism inAntirrhinumflower asymmetry. *The Plant Journal*, *75*(4), 527–538. https://doi.org/10.1111/tpj.12225

- Raimundo, J., Sobral, R., Laranjeira, S., & Costa, M. M. R. (2018). Successive domain rearrangements underlie the evolution of a regulatory module controlled by a small interfering peptide. *Molecular Biology and Evolution*, *35*(12). https://doi.org/10.1093/molbev/msy178

- Romani, F., Reinheimer, R., Florent, S. N., Bowman, J. L., & Moreno, J. E. (2018). Evolutionary history of HOMEODOMAIN LEUCINE ZIPPER transcription factors during plant transition to land. *New Phytologist*, *219*(1), 408–421. https://doi.org/10.1111/nph.15133

- Rosinski, J. A., & Atchley, W. R. (1998). Molecular evolution of the Myb family of transcription factors: evidence for polyphyletic origin. *Journal of Molecular Evolution*, *46*(1), 74–83. https://doi.org/10.1007/pl00006285

- Sakakibara, K., Nishiyama, T., Deguchi, H., & Hasebe, M. (2008). Class 1 KNOX genes are not involved in shoot development in the mossPhyscomitrella patensbut do function in sporophyte development. *Evolution & Development*, *10*(5), 555–566. https://doi.org/10.1111/j.1525-142x.2008.00271.x

- Sakakibara, K., Reisewitz, P., Aoyama, T., Friedrich, T., Ando, S., Sato, Y., Tamada, Y., Nishiyama, T., Hiwatashi, Y., Kurata, T., Ishikawa, M., Deguchi, H., Rensing, S. A., Werr, W., Murata, T., Hasebe, M., & Laux, T. (2014). WOX13-like genes are required for reprogramming of leaf and protoplast cells into stem cells in the moss Physcomitrella patens. *Development*, *141*(8), 1660–1670. https://doi.org/10.1242/dev.097444

- Sauquet, H., von Balthazar, M., Magallón, S., Doyle, J. A., Endress, P. K., Bailes, E. J., Barroso de Morais, E., Bull-Hereñu, K., Carrive, L., Chartier, M., Chomicki, G., Coiro, M., Cornette, R., El Ottra, J. H. L., Epicoco, C., Foster, C. S. P., Jabbour, F., Haevermans, A., Haevermans, T., & Hernández, R. (2017). The ancestral flower of angiosperms and its early diversification. *Nature Communications*, *8*(1). https://doi.org/10.1038/ncomms16047

- Schena, M., Lloyd, A. M., & Davis, R. W. (1991). A steroid-inducible gene expression system for plant cells. *Proceedings of the National Academy of Sciences*, *88*(23), 10421–10425. https://doi.org/10.1073/pnas.88.23.10421

- Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nature Methods, 9*(7), 671–675. https://doi.org/10.1038/nmeth.2089

- Shimamura, M. (2016). Marchantia polymorpha: Taxonomy, Phylogeny and Morphology of a Model System. *Plant & Cell Physiology*, *57*(2), 230–256. https://doi.org/10.1093/pcp/pcv192

- Strassert, J. F. H., Irisarri, I., Williams, T. A., & Burki, F. (2021). A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-22044-z

- Trigg, S. A., Garza, R. M., MacWilliams, A., Nery, J. R., Bartlett, A., Castanon, R., Goubil, A., Feeney, J., O'Malley, R., Huang, S. C., Zhang, Z. Z., Galli, M., & Ecker, J. R. (2017). CrY2H-seq: a massively multiplexed assay for deep-coverage interactome mapping. *Nature Methods*, *14*(8), 819–825. https://doi.org/10.1038/nmeth.4343

- van der Graaff, E., Laux, T., & Rensing, S. A. (2009). The WUS homeobox-containing (WOX) protein family. *Genome Biology*, *10*(12), 248. https://doi.org/10.1186/gb-2009-10-12-248

- Voordeckers, K., Pougach, K., & Verstrepen, K. J. (2015). How do regulatory networks evolve and expand throughout evolution? *Current Opinion in Biotechnology*, *34*, 180–188. https://doi.org/10.1016/j.copbio.2015.02.001

- Wang, Z. Y., Kenigsbuch, D., Sun, L., Harel, E., Ong, M. S., & Tobin, E. M. (1997). A Myb-related transcription factor is involved in the phytochrome regulation of an Arabidopsis Lhcb gene. *The Plant Cell*, *9*(4), 491–507. https://doi.org/10.1105/tpc.9.4.491

- Yasui, Y., Tsukamoto, S., Sugaya, T., Nishihama, R., Wang, Q., Kato, H., Yamato, K. T., Fukaki, H., Mimura, T., Kubo, H., Theres, K., Kohchi, T., & Ishizaki, K. (2019). GEMMA CUP-ASSOCIATED MYB1, an Ortholog of Axillary Meristem Regulators, Is Essential in Vegetative Reproduction in Marchantia polymorpha. *Current Biology*, *29*(23), 3987-3995.e5. https://doi.org/10.1016/j.cub.2019.10.004

Annex A : Primers

Table 3. List of all primers used during this thesis with the identifying names and numbers, the sequence, and the situations of use.

Name	Sequence	Use
DOMYB1 6F Fw	[6FAM]CCTTTTAGGATGAGATAAGACTATTCTCATTCTGA	EMSA
DOMYB1 6F Rv	[6FAM]AAGGTCAGAATGAGAATAGTCTTATCTCATCCTAA	EMSA
pMpEF1alfa Fw (913)	AAAAAGCAGGCTTGCAAATGAGTCACACACATTG	Gateway cloning/Colony PCR
MpDIV1attb Fw (507)	AAAAAGCAGGCTTGCGGATGGCAGCACCCTC	Gateway cloning/Colony PCR
MpDIV1NSattb Rv (513)	AGAAAGCTGGGTTATGATGCATGGCAGGCTGTG	Gateway cloning/Colony PCR
MpDIV2attb Fw (509)	AAAAAGCAGGCTCACTGATGGCAACAACCGTC	Gateway cloning/Colony PCR
MpDIV2NSattb Rv (514)	AGAAAGCTGGGTTAGCGTTCGCAGACTGGG	Gateway cloning/Colony PCR
MpDRIFattb Fw (511)	AAAAAGCAGGCTTCAGAATGGCGGGCTCCG	Gateway cloning/Colony PCR
MpDRIFNSattb Rv (515)	AGAAAGCTGGGTGTTACGTTTGTGAAGTTGGAGACG	Gateway cloning/Colony PCR
MpDRIFattbRv (512)	AGAAAGCTGGGTGTTACGTTTGTGAAGTTGGAGAC	Gateway cloning/Colony PCR
attB1 Fw (Qs190)	GGGGACAAGTTTGTACAAAAAAGCAGGCT	Gateway cloning/Colony PCR
attB1 Rv (Qs191)	GGGGACCACTTTGTACAAGAAAGCTGGGT	Gateway cloning/Colony PCR
MpDRIF pGBT gr Fw (556)	TGACTGTATCGCCGGAATTCCCGGGGATCCAAATGGCGGGCTCCGTCGG	Colony PCR
MpDRIF pGBT gr Rv (557)	AGAAATTCGCCCGGAATTAGCTTGGCTGCAGTTACGTTTGTGAAGTTGGAGAG	Colony PCR
MpDIV2 Fw (455)	ATGGCAACAACCGTCGC	Sequencing
MpDIV2 Rv (456)	TTAAGCGTTCGCAGACTGGG	Sequencing
MpDRIF 78F1 Fw (407)	CTCGCTGCGTCGCCGGCCGTCAAT	Sequencing
MpDRIF 270F1 Fw (630)	CTCGTCCAGTTGTTACATGACCCA	Sequencing
MpDRIFintron Rv (768)	ACCTTCCGTGCCTACGAATCAT	Sequencing

Annex B: Vector Maps

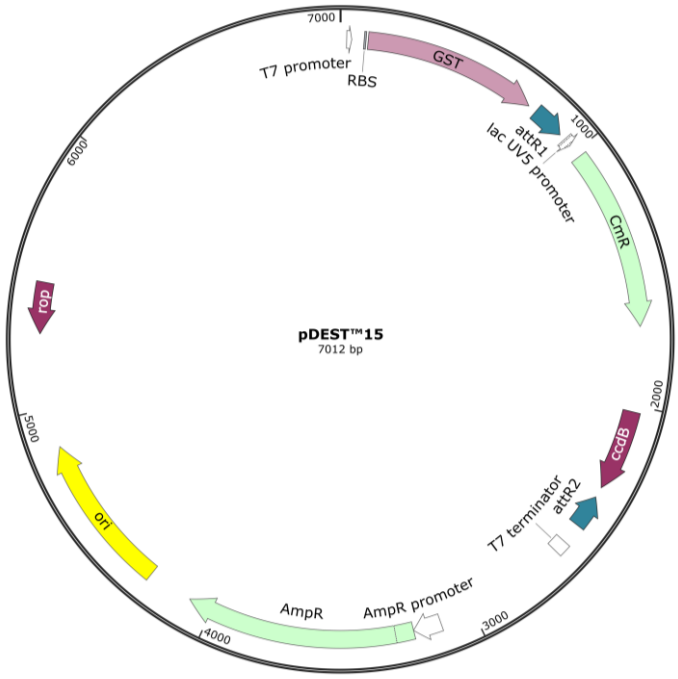

Figure B-1. pDEST™15 vector map. Gateway destination bacterial expression vector with T7 promoter, N-terminal GST tag and genes for resistance to ampicillin and chloramphenicol. Expression of target gene is induced with lactose or IPTG. Vector was used in heterologous protein expression of MpDRIF.

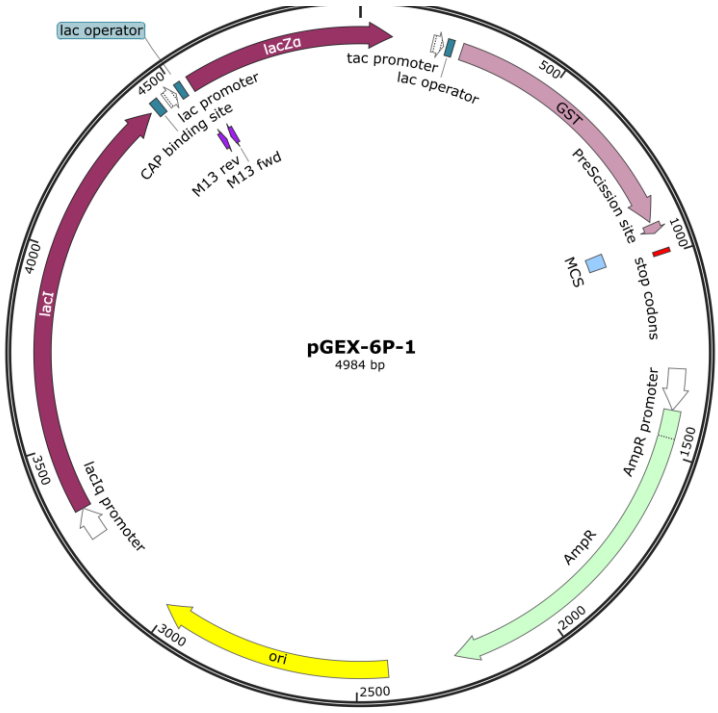

Figure B-2. pGEX-6P-1 (pGEX) vector map. Bacterial expression vector with tac promoter and lac operon, an N-terminal GST tag and gene for resistance to ampicillin. Expression of target gene is induced with lactose or IPTG. Vector was used in heterologous protein expression of MpDIV1 and MpDIV2.

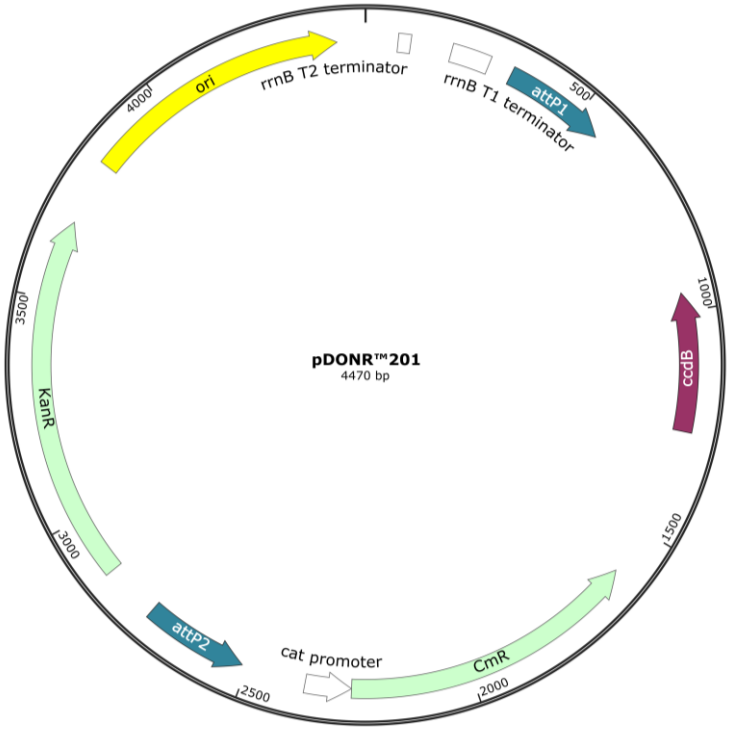

Figure B-3. pDONR™201 vector map. Gateway donor vector with genes for resistance to kanamycin and chloramphenicol. This vector was used to create entry clones for target genes.

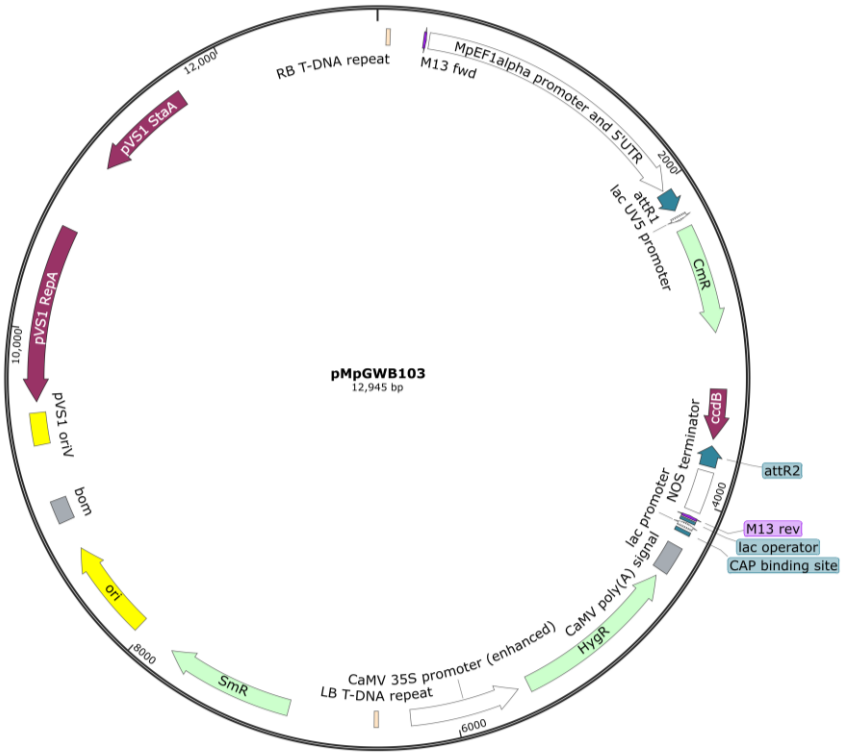

Figure B-4. pMpGWB103 vector map. Gateway destination vector with the strong promoter Mp*EF1α* and with genes for resistance to hygromycin, chloramphenicol and spectinomycin. This vector was used for previously prepared overexpression assays.

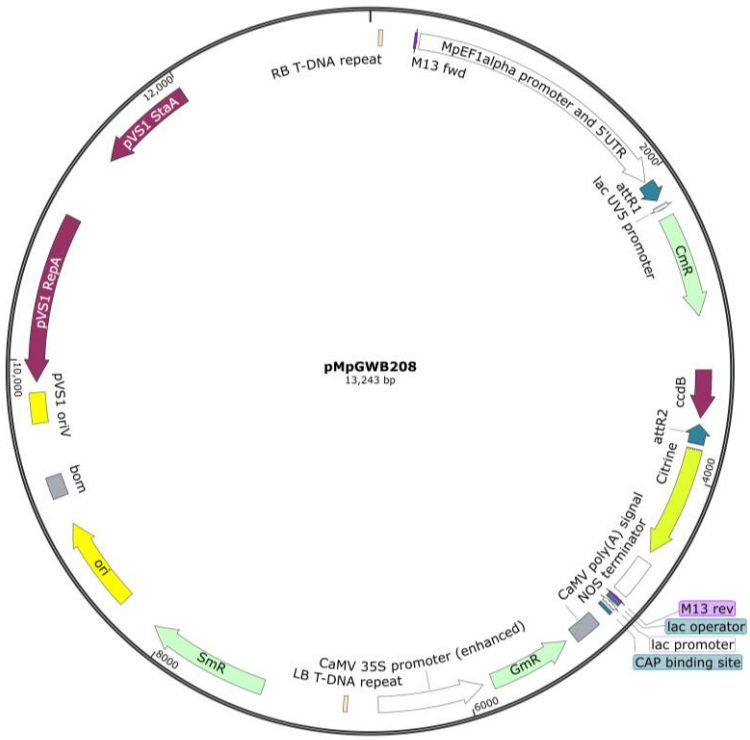

Figure B-5. pMpGWB208 vector map. Gateway destination vector with the strong promoter Mp*EF1α*, a C-terminal Citrine tag and with genes for resistance to gentamicin, chloramphenicol and spectinomycin. This vector was used for overexpression assays.

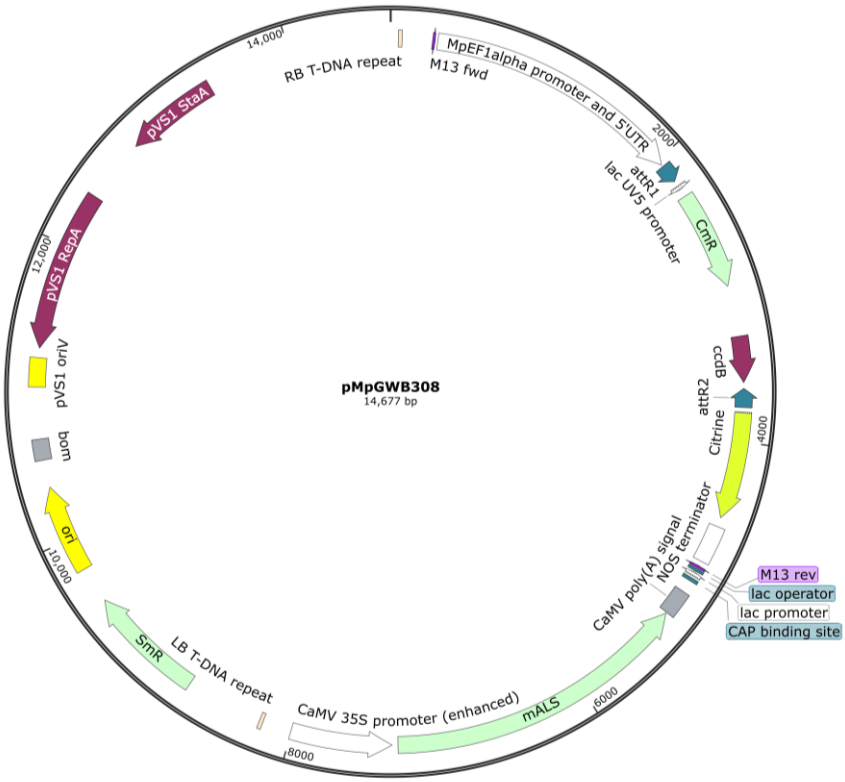

Figure B-6. pMpGWB308 vector map. Gateway destination vector with the strong promoter Mp*EF1α*, a C-terminal Citrine tag and with genes for resistance to chlorsulfuron, chloramphenicol and spectinomycin. This vector was used for overexpression assays.

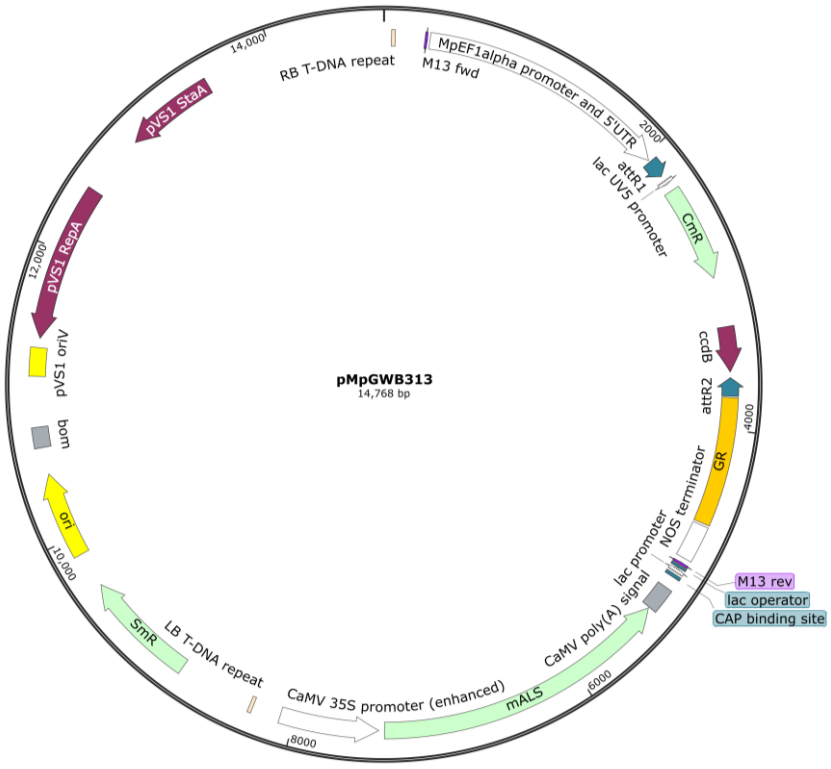

Figure B-7. pMpGWB313 vector map. Gateway destination vector with the strong promoter Mp*EF1α*, a C-terminal Glucocorticoid Receptor (GR) tag and with genes for resistance to chlorsulfuron, chloramphenicol and spectinomycin. This vector was used for overexpression assays.

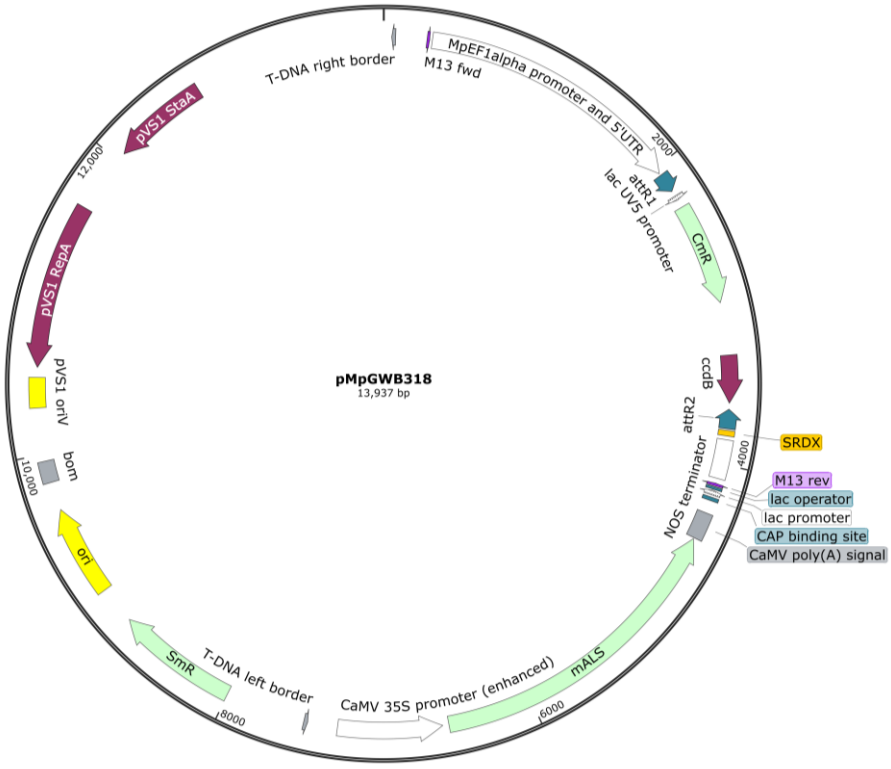

Figure B-8. pMpGWB318 vector map. Gateway destination vector with the strong promoter Mp*EF1α*, a C-terminal SRDX tag and with genes for resistance to chlorsulfuron, chloramphenicol and spectinomycin. This vector was used for overexpression assays.

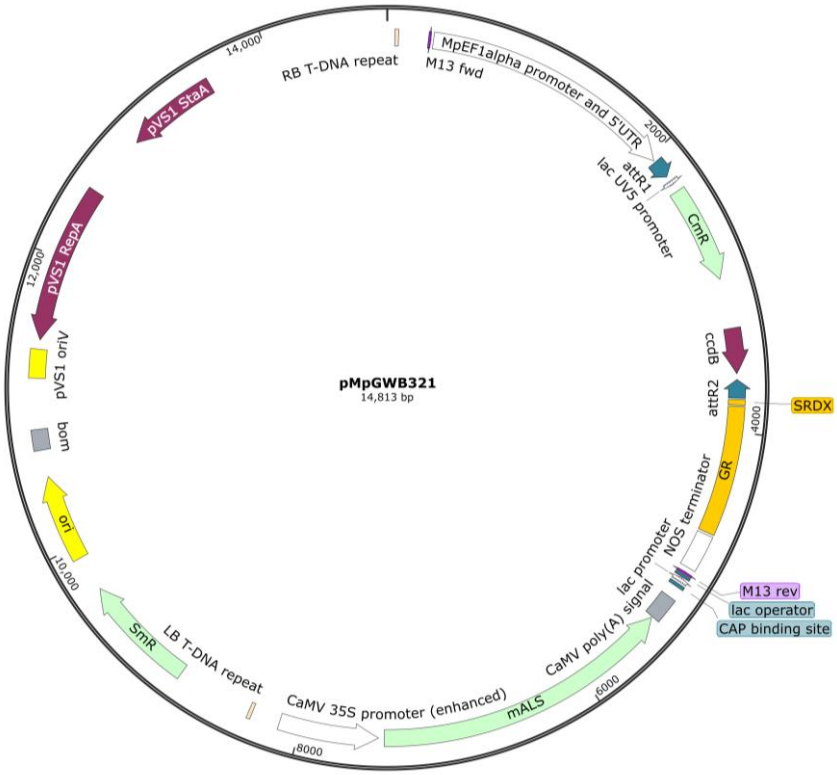

Figure B-9. pMpGWB321 vector map. Gateway destination vector with the strong promoter Mp*EF1α*, C-terminal SRDX and Glucocorticoid Receptor (GR) tag and with genes for resistance to gentamicin, chloramphenicol and spectinomycin. This vector was used for overexpression assays.

Annex C: Protein sequences

Cryptophyta

Crytptophyceae sp. CCMP2045

>CryptoDIV

MAFFPNSNGGGLPAGNSPLAGVGGVLTPLGMNGGGMDGGGFSPAFATGMHGAFPLALQQPAVRDIQM
GNSGENSGTRWSKEEHAQFVTALEEYGVGSTGNEWNLMAQAVGKTEADVKIHAQQYFLKLERERQVPAE
NMLQPVSQGLPGMAKSAFMIPHEQGSEGKGAGQGVNGTVWTVLEAQLFEEKLAEVDPDSETRWQQIAAS
LPEKSPEDVKAHYKWLQRLLRSRGAGEVSPHDGGGRKGKDKGKQETHGLSWTEEEHCRFLEGLERFGK
GDWRNISKHCVVTRTPTQVASHAQKFFVRQQNAAKKQDKRRSSIHDITTAAKGESKVEHNQSGNAKNLL
DSPVNAPSFPSETPRNMFGMPLAEGLGGAADGSLMTPGVSPYDNGGMLLSQLQESLLKGNAEQSANDS
SWTGGENTLSVTTPKGGDATPKIKAAKKQH*

Guillardia theta

>GuithDIV1

MSNGTKFTREEHMKFLRALDELDSNINGNEWEKIAKEVGKSENEVKVHAQQYFLKLERERRIPTENVLSS
DQNMSSQAMQPYMSSSFIVPFGGELSSSSNDPTQSKPQGVVWTPEEARIFEDKISEIDPNDDDRWMRIAS
LLPNKSADDVQSYYTWLQNLLRARGAGQSSSSPIDQATGKKSGKEKGKLETHGLSWTEEEHRRFLEGLE
RFGKGDWRNISKHCVVTRTPTQVASHAQKYFVRQQNAAKKKEKRRNSIHDITPSSIKTYWSGGKEKEGSS
SPDEGNESQENENQQGTQGSSSNNTSGNVSGGASNQINPLKGEETSKGIFDSPTNNCLPPETPLGISLFS
SLLPSGSGPEASISSGLTPEASPFGTMTLSQLQAGLLKPDGA*

>GuithDIV2

MHGNFPFLLQSPSGRDPGGGSLPPDMSNGTKFTREEHMKFLRALDELDSNINGNEWEKIAKEVGKSEN
EVKVHAQQYFLKLERERRIPTENVLSSDQNMSSQAMQPYMSSSFIVPFGGELSSSSNDPTQSKPQGVVWT
PEEARIFEDKISEIDPNDDDRWMRIASLLPNKSADDVQSYYTWLQNLLRARGAGQSSSSPIDQATGKKSGK
EKGKLETHGLSWTEEEHRRFLEGLERFGKGDWRNISKHCVVTRTPTQVASHAQKYFVRQQNAAKKKEKR
RNSIHDITPSSIKTYWSGGKEKEGSSSPDEGNESQENENQQGTQGSSSNNTSGNVSGGASNQINPLKGE
ETSKGIFDSPTNNCLPPETPLGISLFSSLLPSGSGPEASISSGLTPEASPFGTMTLSQLQAGLLKPDGA*

Glaucophyta

Cyanophora paradoxa

>CyaparDIV

MESLPFPFPFSSGEHMLFLEALEIFGYTDNGGDEWLLVAQHIGSRSIQEVKEHAERYFRLLQSKNRNPPAA
FATPKLGLCTAHLENSLDALGTPIYSDEADWSREEEARFEEALAALDESDPQRFHKLSALLPGKSADECSN
HYHALLYDVARIERGHIALPAYQHSQFTFSWSGPIPGLPSTPGSSALAEAAQGTHSASGGSRSIGRSRRRG
APGSGSPHEGDRDKIGTKGMPWSEEEHRLFLSGLQKFGKGDWRNISRSFVVTRTPTQVASHAQKYFMRL
ANGAKEKRRSSIHDLTITPEQAAQLEKVAAQRSKDASPAPETADVDAPPRPRAGRRRSEPQIKFSMMTDG
GEEVQGPSGRAMTSWSPKSALGGGGGTGPAELACGRHAAMLAPPSAGGFRASFPNSSSSTPGGSPKPA
PEGPNSGGSSPTHGEDAPAPPIEGLNISGKRRRDLPDTLSPGAMLTDASIGVPSAAVYGNPAAVALPPLPS
PSPPPSSMSIS*

Prasinodermophyta

Prasinoderma coloniale

>PracoDIV

MPGVARAAGAPATTAAHLTPAERQASLQWSVEEERAFENALARHIDEPDTEQRWERVASAVGHKSAIDCK
RRYELLVEDVRNICAGRVPMPNYSNSDGERRKGIPWTEEEHRLFLLGLAKFGKGDWRSISRNFVISRTPTQ
VASHAQKYFIRLHSMNKDKRSRAEGAAARWAGTSQQGAQAQRQSGGAAAAAQQQQGPGFGAPPMGFQ
GQEVYGRQF*

>PracoDRIF

MLLKARASATRRAIIGGHPEVDPSTRPTSADGAPATVPAVASEPAAASGADVGAPPVAGDDATAAAPAAAT
TPAAVAAAPAEAAVPAAAVATDKPEGMAPEAGAAPTQTVPPASLAAPAKPAAAPKVAAPTKGPASRPPAKA
KAGPASKPPAGGKSAAAMKPSGGVAKAAPKKRPTTSGARKRTVKPASGGSKGRSGSKGDRSALASGAGG
SGTTSEVDVNGDVGVNGGGTRQGDGAGTGRGSGMHRNGEPPALQTIDWTRKEQQLLEDGLNKCPEDK
HMPLDRYIRIASMLPGKGVRDVALRVRWMSRKEQGKRRKADTDTGGSKKSGSRSREGGRVEKTSIFAMR
PPAMPAAGAQRAMPQGAGASMGDLDTRILGGEGIGGTTGRLLGENMQVINQIRHNLAQCKVQENLDLFR
HVRDNVMGINNSMTTMRGMMSHMPPLPVQLNEQLANVVLGALANGLSGPPNGLGGGNPPGMMGPGG
GNPNVMAPPGGGGVQHLSGGGGQLPAARQMSR*

Chlorophyta

Auxenochlorella protothecoides

>AuxprDIV

MLGGAGSSTPSGMVEEPWSKEEDKIFEDALAHVWDRVDRCERCATQLARRDVQSVRQRLEQLERDVMS
VEEGRVLLPNYAVPGESLSVAHLQKKVKSQETERRKGIPWTEEEHRLFLMGLAKYGKGDWRSISRNFVITR
TPTQVASHAQKYFIRLNSQSRKDKRRASIHDITSPAPPEPSLLAYGMGHHMAAPGMGMMHVTPGGVAAG
VPVMHPGALAHHRAMLQQRMGS

>AuxprDRIF

MSAPSSIAHWNVSLDWSEEEQRALETGLARFPPTQGSPLAQYVRIALALPRKSVREVALRARWTKQAALLK
KRRTEELESGESAAKRPLLHNPLLQVPQPSMPNWPGSMGGLGPIMMPSHPGPVPASAAPAPPPAQERA
TPAPGPGRGGASLGSLLEANLALLQQFKGNMTHFRVHDNTRLLVQFRDNILAVIKQMESMDGVMAQMP
QLPVRYAADVLNVDLANNFLPSQPLGLVPALSDPASLTHPSPASEGPGMVPLGNGSAAHAPPASGGGVAI
KREAETGTPNGVAGLQARPRVEAVKAEPTGGGAAQPTALAMADCRTAAAGEASMRDAGAGTTTGEAHGP
EARGTEGDP

Botryococcus braunii

>BotrbrauDIV

MANTSVSGGSGGGGKKGNFHPVKAEPGSVWTPEEEKTFELVFLENIDAPQQDRLHKVAARLPGKTLADIE
KYIEDLEADFKALDTGSIPLPPYRHSANEPSIPLTVQKPAVKAADQERRKGIPWTEEEHRLFLLGLGKFGKG
DWRSISRNYVTTRTPTQVASHAQKYFIRLNSQTKKDKRRSSIHDITTPTTDMTSPVMGDNRPPPPLLSVLS
VLSNPSS*

>BotrbrauDRIF

MMSIATVDNKGSIPLGNGINTGGPEGPLAQGVHCYEGPRLRTARNVYGVEWTGEEQKILEAALARLPAAM
SKLHRCLEIAKQLPRKLSIDVALRIQWMALRDQRKPKIDDASVGKAAPKPVRRDRSTSIFAVPAQGLPIGGL
PMHAQLGHVAHGTALGTPVGPLPGPAIVGQAMVPPGMMSAGAYMPQPPPHMSQPVPMHVGPIHPGQG
NRDDRPADTVDANDAEVAKFLEANYVILGHIKENMKHWNVAVNTGLLLNLRDNLRNIMRLMSNMPAMVE
MPPLPVQIDTQLADEWLPTVPQSVSSYNPLDMSNRHHQGMLPGMGMYPHMHQPAALPGQLMMAPVH
PLAGSTVPPVPGGIVAGVPVGVPVPALQDANVPAAPPGGPPMRMVGPAAVSSAPPLPQDIKPAIAPAFSVN
GTAEAKPPPPGPAPAPASAPAPAPPVTSAEPPAASPSLRQGLRPNRSRAAGIQSSRPSARSSKGS*

Chlorella sp. A99

>ChloA99DRIF

MAAPALPLGVRPPPGVLPPPLGTAGPQPGLPAPPQALMLPPAAWQVSLDWTEEEQRALEAALARYPPERV
PVAERYIKIAAMLPRKSVRDVTLRVRWTIQHQLGKKLGKVDHLPLPAGGGYAKAGAARAMGPGAAADVDG
AGAGSSIDGPVGQLLESNFAILNQFRSNMSTFKVQENTELLVQFRDNILAILSHMEGMGGVMAQMPQLPV
RLNVDLANNFLPSRPIASLAVSGQMPQPSLGPGMVPISGFGAGGEGAPPPQPPQPDGGGAAAGGGAPAG
PPPGAASSRALLGGAPALVRQDSSLIRQDSAALPFGKILLKQEG*

Chlorella sorokiniana

>ChlosoDIV

MMQLPPGLRPPGLRPGVVVDNWSVEEDRVLENALAQFWEHTDRLEKCASLLSRKDLAAVKRRYQQLED
DLRAIDMGRVQLPNYPVPGEALSVAQLQKKVKSQDTERRKGIPWTEEEHRLFLMGLAKYGKGDWRSISRN
FVITRTPTQVASHAQKYFIRLNSQNKKDKRRASIHDITTVAPAPSDPSASAAAPWSTVPITGANPAVAAVVAP
GAPPPAMPGMAGLPITGIPGMPGMPGLPRPQ*

>ChlosoDRIF

MQPVLPPGAAAALAAFPAAMAVALPPPAAGFVAPPSYPVSLDWTDEEQRALEAGMQRYPPDRFDVVQRYV
KIAAMLPRKSVRDVALRCRWTLNQQLLKKRKPGEALVPPAGALGGGAKKPLGAAPGLMPQQAPALPPVP
MMPGVAPAGAAAAAAVAPAEGPTQAVIAGPIAQLLESNFTILNEFRSNMADFKVHENTQLLVQFRDNILAII
NSMEAMGFRDNILAIINSMEAMGGVMAQMPQLPVRLNVDLANNFLPSRPANVPAYNLAMPPPQPALNAP
GMVPC*

Chlorella variabillis

>ChlvarDIV

MQAVRPALLRGAPVDAVWSTEEDKVFENALAQFWEHNDRLEKCASLLSRKDLPAVQRRYLQLEEDLKAID
CGRVQLPNYPVPGEALSVAQLQKKVKSQDTERRKGIPWTEEEHRLFLMGLAKYGKGDWRSISRNFVITRT
PTQVASHAQKYFIRLNSQNKKDKRRASIHDITTVAPTVGDHANGGAMGGGGSAPSFMSGVMSLTITGQNS
AVAAVVAPGAPAPPGGIAMSAGLAMACAAPSALPPGSMIPP*

>ChlvarDRIF

MQQSAMALPPPAPGYLAPPTWPVALDWTEEEQRALEAGVQRYPPDRFDMVQRYVKIAAMLPRKSVRDVA
LRVRWTVNQQLLKKRKPGEALMPMAPGAKKAAVPGGMLPPKAPTLPPVPMMPGMSALPPAAAAIPPES

PTHAVIGGPIAQLLEANFSILNEFRANMSEFKVGENTELLVAFRDNILAIINAMEGMGGVMAQMPQLPVRL
NVDLANNFLPARPASLPRHNLAMPPPQPALNAPGMVPLTEDCAGPGAGAVPPPPPPPAPGGVPGGGGV
PGGGPLGPLPGGGGMPFMGGSFGGAPTLIRQEQPVIVKKQEG*

Coccomyxa subellipsoidea

>CocDIV

MGVAANGTAPANAGLAAPKLGSTDWSVEEDKILESALAEFWDVDNRVDKIILKLPRKTKDLIKHRINLLEE
DVRNIESGKVPLPKYLASAEPASAAVGKSKASEQERRKGIPWTEEEHRLFLMGLAKFGKGDWRSISRSFVA
SHAQKYFIRLNSMNKKDKRRSSIHDITNPGGVSGDVTGMMPNGMGMQQGMMPMANGMVMQGGMAM
GTPGMMGLQGSMPMQPHG*

>CocDRIF

MGSLAPPGLIIAQPQLGENGGAPGPSGQTILQARGDTAYTTEWDSAEQAALDSALARFPADRHPPLERYVR
AAACLPKKNVRDVALRVAWLRATAAARKRKMADEANSKKQVRRERGQSIFAVQPKPMGGGVGHPMAASL
AAMPGPNMGMASGMPMPAPIVVQPHAGMAYAQPVVPLAPMPQLDDHGAGTVGGVGGPLAQPLEQNYAI
LNQFKQNMAAYKVNENTELLVRFRDNILTGVMQQMPPLPVRLNVELANNFLPKAVANGMCQFPYVPPSG
GMASGPMLGQQGSSAGMAPPSVATSAPVAAAASGAAPALPAVSAPALLPLPDSSPAFPPAQQQPQPAQL
QPPAQLQQPAQTAQSQQVPPQTMQLQQQPQAAQLEQQHLPPPASAAMPAGGTAAASAAAQVPPPAHRP
ANGLPVVSMPSPATPAPASASAALAAPVQQPQPLVPPPAPSNPVPAAAPQLPVPLATLDAVIKLEPVTTPVA
VPIVEQPAQPSTLQPPSLPAAVAPVQPKGEPQTSAAAPPQQLPLPSPAAADQPAPNPSAPPPPKEEPAAAA
PEPPPAKTPPAAPPSTPALVNGVESSEAPAAPAAPQPPQSDAAAAPACVTSSPKAGGVGTRSSSAAQPSP
GRPSRATRSMAAAKRNPSAAASPSTSKGQG*

Micractinium conductrix

>MiccoDRIF

MAAGYIAPPAFQVALDWSEEEQKALEAGLARYPADRFDFVQRYVKVAAMLPRKSVRDVALRARWTINQQL
LKKRKPGELVSGAGGGAQKSMGAGSMLPPKAPQLPPVPMMPGMSALPASAAMPINTPTNAIIGGPVAQL
LETNFTILNEFRSNMADFKVPENTQLLVQFRDNILAIINAMEAMGGVMAQMPQLPVRLNVDLANNFLPSR
PATMPAYNLAMPPPQPALNAPGMVPLTEDYGPGSGAVPSAANGMQQGGGGGGGAAAGGGAVPSSTLP
LPATLPSASMPFMGGSFGGAPTLIKQEQPVLSKKQDG*

Picochlorum renovo

>PicreDIV

MVSGTRSAGNGTKQQQQQKDAHDGSHAKRAEGKAADHAVAKGGWSEEEDRVFENSLAQYWDFPDRFE
KCASMLSRKNLTDVIARFKELDEDIRNIEMGRKKVKSQDTERRKGIPWTEEEHRLFLMGLAKYGKGDWRS
ISRNFVITRTPTQVASHAQKYFIRLNSQNKKDKRRASIHDITSVHPEYKKKPKKSTKKE*

>PicreDRIF

MQSSLIPTTQWGVSLEWTDEEQKSLETLMYRYAPERMDPVQRYVRIAAALPRKSVRDVALRVRWTMQQQ
LKRRAGDPGKAPMGIGGMGPGNPMMSMNPNLGIVTPPLLPLQSQDGAQTVDGPIAYLLDANLSILNQFR
TNMASFKVHENTQLLVQFRDNILQILHAMDNMGGVMTQLPPLPVKLNIDMANDFLPTRPTGIFAMDGMV
AIPPPPQPAMNVPGMVPLNGLGQTNQPSSWGQQHGGGQS*

Picochlorum soloecismus

>PicsoDIV

MLDQILKIDSTSGADGVVYRAKHHDGVEGGVAGSSGNIWSPEEDRVFENALAQFWDYPDRFEKCASMLS
KRNITDVIQRFKELDQDIREIELGRIQMPAYPVPGEALSISQLQKKVKSQDTERRKGIPWTEEEHRLFLMGL
AKYGKGDWRSISRNFVITRTPTQVASHAQKYFIRLNSQNKKDKRRASIHDITTVAPGGGVKKSAGGTSAKS
GPKGGK*

>PicsoDRIF

MGNSSSGGAGHADWGVSLEWTDEEQKSLELLMNRYTPERMDPVQRYVRIAAALPRKSVRDVALRVRWTT
QQRMKRRAGDGMKGGIGVSQNNPMMSMNPCLGTVTPPTLPLQGQDGRQTVDGPIAYLLDANLSILNQF
RTNMASFKVHENTQLLVQFRDNILQILHAMDSLGGVMAQLPQLPVRLNIEMANNFLPSRPVGVMAMDGV
VNVPPPPQPALNAPGMVPLNGLSQHAGQMPGTQPGGAGYSVPPTNAPPGWGS*

Symbiochloris reticulata

>SymretDIV

MTASAAAPVQLDTSDVPSPAEWSAEEDKALEVVLAEHYAAPDRAQKAAARLNRPLDAIQDRMTILQEDVN
NIEAGLIAFPKYDTNDIELSILRASKPATDQERRKGIPWTEEEHRLFLMGLAKFGKGDWRSISRNFVVTRTP
TQVASHAQKYFIRLNSMNKKDKRRSSIHDITSAGIPGGQDSTGVMNHSMPQMAPTMVPMNGAAMGMAT
GVPLHPHAPPMLHPMGLPGHAVPQ*

>SymretDRIF

MSKLAPSSISASSVRPPAVSNATVLPASMPPAPFPPATPEPTTRLRSGTACSISWSAEEQAALETAMVRYP
PDRFQPLERYLRIAANLTQKGARDVALRLKWMAACQAARKRQLSENDSSKQKQQQQFRRERGQSIFNIQP
KPSMPRNGVYGSQLNSSLDDHGSTSVGVVSGPIAHLLEQNYAVLNQFKQNMAQYKVNENTDLLVRFRDN
ILAILSQMNSMQGVMQQMPPLPVRMNIELATNFLPKTGSGAFPLGMPFGPPAVPGLVPGVGMQGFMHPV
GTPQTGPARSQPGTVPPAVHQPQQHGSAMQQPPMHQPPAADATPLQNGSTPAAAAPPLAAHIPQPPNC
APLIQMPLMRPMGQLPVGFPPGMAAGSYPLPPFLQNGLGPASSMSLHTPMQSALPASMFMPGGHGGN
PPLAAHMAQMSGSIALPVPHLQQAPQFQGQAVGHMPAVMPSLPPNSAPARMPVIKAEHS*

Tetraselmis striata

>TetstrDIV

MGDLRPVSWLPDGWSAEENSRFESLLAEHFDASDKFAKISAKLPGKTADAVRVRYNQLVEDMKNIEAGC
VEMPAYAQEEDEPMVHKPPKGVKASDQERRKGIPWTEDEHRLFLLGLAKFGKGDWRSISRTFVQTRTPT
QVASHAQKYFIRMNTMNKKDKRRSSIHDITGSNAAHEAAQLQAQMAGQAHLMGHAAPGHMLPPGAVM
SAHMQAPMGGPRPGVMYVQQQPMQPMQ*

>TetstrDRIF

MCAVEQDPMATGSKATSVTTKVEPSAGGTTGPSRPPAGPSVKAASNGGPASQVKAEPAGSGGRVDTGPS
AGPRPQNPEWGAEELKGLEAGMAKWPAAKHGLLERAVRIAGSLPGKSARDVALRLTWIAKGGAAGKRLK
KGEPEKGVKGAQRPPRAGSIFAVQPGPVAGGAGGAAGGAEGAMDGGPGGTVAQLLQNNFDIITNIRNNM
HQFKVNENTELLVRMRDNTLRILASMKQSEGVMSQMPELPVKMNLDLANSFLPKVMGQAPPMAPPPQP
QQHMVMQMGPNGPIALGGMPQAICVSGMSGDMSMPMMQMAPGQLQMMQMAGPGGPHHMRFVHA
PPPGQPQHSGPMLMAAAQHQQHQQQQQQQQQVGGGMH*

Trebouxia sp. A1-2

>TrebDRIF

MEEPYDYDAEALQAQAAWSLATSQGIHHSDSHAIAAALHIPAQAQGGGYGSSLAQTPTDAVNAVADEPWV
NDAVVIGGASAVLSPIIQWGHTILQSKPQGRVRRRVPRRNPRPSFPLQSLIYPLRGCRKAAPKFTLLSLTRLH
MTANGAGGATSKTSQKVTRAAPAMADTESKVKSAQQSRDPAEVATMSSEDALPTGTLKARGAAVYNAEW
TAHEQAALDQAAVKFPAERYQPFERYVRIAATLPRKGVRDVALRLRWLSQARKRKISEDGPNKRLRRDRCQ
SIFATQQKPPNVMGQWPQQQLPMPQLDDHGAQTVGAVGGLVAQLLEQNLHILYQYKQNMHQFKVQENT
ELLVRFRDNILAVLNQMNSMDGVMSQMPQLPVRMNVELANNFLPAASGASPMFPLGMPGMPPGMGN
MPSGPMPMNGYMGANAMGSAASMPMTNNGMPMSSNGMPPGSAPFTGQQQQQQQHHSNGASPNAG
GVVRAGPPPVTANAAAAAAGTAVALGPGQTKAPAQSASSSKGVAAKAPLKQPVVHNGRPAEQEQASTQPA

PAAGLNAVASAVPSGRPAGSVAAQPTPAVSVASKPQSAASKSLAASASKVAAGMASKAASAKSNKQVKPK
T*

Chlamydomonas eustigma

>ChleuDIV1

MIFATSAPLTSWSIEENKKFEKALAQHFHDEDRWTKISEHCPHKQIEDIIGQFEKLKMDLTRIQGSQPSAL
GMNKLTSQTKIESLKRAKTEPMDIPEQPRKGVSWTQQEHQKFLEGLEQYGKGNWRAISRDFVISRTPTQV
ASHAQKYFLRVTNTKSKRRSSIHDLARQGSYSLE*

>ChleuDIV2

MVLELVRLEEDRVRPWTFEETKAFELALAQHFNDVDKWVKIAAILPQKGIIEIQRHFRLLEEDLENIQAGKQ
MLPTMHIECPKRVVATKMKAVSETKSSGSGSSVENRKGVSWTAEEHRLFLLGLAQFGKGNWRSIANEAVL
TRTPTQVASHAQKYFLRLALSKDKREKRRASIHDMTHESEDFSLEAEGQRMGKPGMRVPLKKIQRVGGR*

>ChleuDRIF1

MGTSVDEGSGTESQVEEPGLSGTSGSDWTSEEQINLDQAILLYPADQYPTAFERTILVAALVPTRSAREVAL
RINWLSSKSSQKSHELKRRGSLPTSTLTRVSSLQAPSPNKSSNSKVFSRTSASKHGSVGQQPQSNSVPPT
SASDQQNFPPQLMLPPPLSSPLLPSLSVSCLFQSSSPPTNSNSSDDACVNIGVTSASAALSATDSISAKSS
APSPSGSVQPPVASVQSLIDQNYVILTNFKKNMQQCRVVVENTELLVRLRDNIVTCINQMGNLPSTTSTLPP
LPVQLNLELAGKFLPNKMVLPPMPSGMPPFSFNPALGPPPMMLPPGMPMPSPGMIPMPMLSGQPGLM
PPFPIGIPPPLMSMQLPTSTESTPPGFVPVPPSRHLPGPALLSQSVPLSAMGTPGNVTNAMSSAGMVPMS
AVLAPLVRQVDSQGRQGEP*

>ChleuDRIF2

MGTSVDEGSGTESLTEDAGLSGTSGCDWTREEQTSLDQALLLYPAHQYPIAFERTILVAALVPTRSARDVAL
RINCLSTKSCNKSQELKRRGTLPTSTLMRISSLQSPSPTKSPNSKSFSRASVPKYGGITQQPQSNKNVAPT
SASDQQNAPLQLTIPPPLSTPLLPALPASCLFQTSSPPTNSNSSDDGCVNMGVLATSTAIFEPDAASVKLS
SPSPSPCPSAAEQPALGKVQSLIDQNHGILINFKTNMQQCRVVVENTELLVRLRDNIVACLNQIGNLPSTVST
LPPLPVQLNLVLACKFLPHKMMLQSGVPPFSFNPALGPPPMMLPPALPLPAPGVIPMPLLGAQPGMMPP
FSMGLPPTLMNMQLPMSVEPTASGFVPSTS*

Chlamydomonas reinhardtii

>ChlreDIV

MAASFSISGDFACGQSTGHATFWRLEENKVFEVALARHYADVDRFERIASYLPNKTPNDIQKRLRDLEDDL
RRIDEGCNEGASAQSAPAATPARSEDSAPNAKRPKTDVPANGDRRKGVPWTEEEHRLFLLGLAKFGKGD
WRSIARNFVISRTPTQVASHAQKYFIRLNSMNKKDKRRASIHDITSPTLPASVANPAPTTGLAPAAASGKAT
SSLVQGATSSATTATSQPMAAAAAAAAAAFPAAAHVAAAAAAAAAAATSTTSVFAQLAMHGLAMQPVMQQ
AAAAAAAAGMMPQLNAAAAAAAAAGMPAPVLPNAAQYMVQV*

>ChlreDRIF

MASTAGAFPAVPIRVDAPVANTSSQNSVDKSTLREQPGAGGPAPAPTIASSASGDDFDADFELQLQGTTGS
DWTPEEINILESGLAQYPADKFTPVERYIKLAAILPSKTARDVALRVKACGLDERKGPGQESGAAGGGAAAK
GGRKGGGGRGSAKGGGGGGAGAAGSGGGGNGAGLGEDSSSGIPAALTQLMEQNYGILTQFKANMAAFK
VMENTELLMRYRDNLLGIQQQLASIGGIMGQMPPLPVTPNFDLANKFLPPGVKPPPGSTPTAPVAPAPPA
MPAAPPVLQPPPPPPPPAMPMPVPGQMPPGMASLMGMAAPPAPTPHPPPMPAPGSTPVGPPGASAAA
AAAAAAAAAAAAAAAASMPGMAAPTAAAPIPGMSMPGVVAPAVAPAVSPTPPPGPPVMPMMPPFSFNPAA
AAAAAAAAAAAGMPQPPGMPGAMPGTMGVMPSGMSMDPSSFFGAAGMPGMPGMPGVMPPQMMAGA
MNPAAAAAAAAAAMGGMGPGAPGLPPGFNPYAAMAAAAAPGMMGMPGMPGAPPPPGAMGAPPGMPD
GGAAAAAAAAMAAAAAAAAHHHHQQQQQQAAAVASMSRQGSAVQPMAMPVMPVPVKQETG*

Chlamydomonas schloesseri

>jgi|Chlsc1|3553|g407.t1 - ChlscDIV

MAASFSFSGDFASCPATGHATFWRLEENKVFEVALAKHYADADRFERIASYLPNKTPSDIQKRLRDLEDDL
RRIDEGCNEGASAQSPPAATQTRSEDSAPNAKRPKTDVPANGDRRKGVPWTEEEHRLFLLGLAKFGKGD
WRSIARNFVVSRTPTQVASHAQKYFIRLNSMNKKDKRRASIHDITSPTLPASVANPTPTTGLAPTAASGKTT
SSAVQGATSSATTATSQPMAAAAAAAAAAFPAAAHVAAAAAAAAAAATSTTSVFAQLAMHGLAMQPVMQQ
AAAAAAAAGMMPQLNAAAAASNAAAAAAAAAGVPAPAMPSTVPYMVQV*

>ChlscDRIF

MASTAGAFQAVPIRIDGPAAITSSQNSVDKGALRDQPGAGGPAPAPTIASSASGDDFDADFELQLQGTTGS
DWTPEEVSILESGLAQYPADKFTPVERYIKLAAILPSKTARDVALRVKACGLDERKGPGQESGAPGGGAATK
GGRKGGGGRGGAKGGGAGAAGAGGSSAGLGDDSSPGIPTALTQLMEQNYGILTQFKANMAAFKVMENT
ELLMRYRDNLLGIQQQLSSIGGIMGQMPPLPVTPNFDLANKFLPPGVKPPPGGAPAAAPAAAAPAPPAMP
VAPTPVPPPPPPAMPMPVPVPGQLPPGMAGLMGMVPPPTAPTPVPTPMPAPGTTPVGPPGGAAAAAAAA
AAAAAAAAAAASMPGMAAPAAPAPPMPGMTMPGVAAPAVAPAVSPTPPPAPPVMSMMPPFSFNPAAAA
AAAAAAAAAGMPQPPGMPGAMPGAMGVMPPGMSMDPSSFFGAAGMPGMPGMPGVMPPQMMAGAM

NPAAAAAAAAAAMGSMGAGAPGMPPGFNPYAAMAAAAPGMMGMPGMPGAPPPPGAMGAPPGMPDG
GAAAAAAAAAMAAAAAAAAQQQAQQQAAANMSRQGSAVQPMAMPVMPVPVKQETG*

Chromochloris zofingiensis

>ChrzofDIV

MLSCRDEEHKRFEIALAQFYRDPHRFQRIAELLPGKTLADIQLCFQRLQADVANIQEGRIQFTEYSGSGSDA
SEPPQKKLKDVTDRKKGVPWTEEEHRLFLMGLAKFGKGDWRNIARNYVVSRTPTQVASHAQKYFIRLNQI
NKVGPAITKRDKKRASIHDMAAVPEPATLPSAAAAPLAPGQTAATATAAATGPSTEPALATQPQAPHQQLV
ASQQQTHMAANPPLPVQQQQQQQQQYPVMQQQLTAPPGTAPIAVHLSAPPQQLLPLPAGMSLPLPHLL
PPMGMLLPHQMPLGMPMPPHPPFMVQM*

Dunaliella salina

>DunsalDIV

MKGSNWTFADSKALEVSLTAHYSKPDRWEHVQTCLPDKSFEDMEAYLHQLEDDIKSIEDGTTPLPPYAPL
PHPPQSIKDESAAAQRLLPAPKKSKTDCTGGSSSAAAAAAAAADRKKGVPWTEEEHKLFLQGLTKFGKGD
WRNIARTFVMTRTPTQVASHAQKYFIRLNSQNNKKDKRRASIHDITH*

>DunsalDRIF

MMSLEQPASWNPPGLATQQAAEQHYPQQQGSQVVDNREWREWSLDEHSTLCKLIEQDWFPQLTGVER
CLRLAAQLPQKTARDVALRLRWMATAGKQQDPGIAGNHPTGRPVSKRPRSRRRGSRQSSFKISQAMKDT
SSEGEEASEGDGSSGPQLPKAASMRAGTTHTDASCTVPSCNAPRPWQAGRAGVSRDPRATAEAGNSIISS
LVEQNYSILASFRANMAQAKVAENTELLLKYRDNIATALENMAAMPGVMSRMPALPVKPSLEAASRLLPP
MATKPGPRPGHIPFGRCGMGPRCQCPQRLTSNTPPPATGMPSMQLMPPPPPPPHDLPNPPPFSCMPP
PPISSMDLPRPPVASLPSMAPPSCPLPPPIVDPPLPHAMATFPPFQPPPPLQTPVSMPPSGGASNAATTH
SMPPPTLPMFPPPGTLPHPTPSFACSLPLQQQLHPQRGGTLTPLGPFAPPLFPARHHHHMPTSMPCMP
PHAPAQLGVVLPQQQQQQQQQQCSSHPKPCHPHLDQLHVHAPAELAQPHATALPGSMACSTPSMAQ
APRSPLHAQAPGSPLHVQQQQQQQQQQLKQQQQLQLQHQHGLHGLHHHAPPFSPSPQPLPPLQQQ
HALHSAAAHHHLMVQGAQGPFCPYPQVHPTSLPEVFHPPGSPIQPEDLYLPPHLPAPGPDPFLPRAPPQI
SATMPAAPAAAAAPPASPPAAAAEEPIFGTAGVKEEPRAASGPIFGAAAEEVEPHTAEESIFKAAAVKEELQ
PAEEPIFGTAAEEGEQPAEEAIFGTADEGEPYAAPESVFSKAAVMEEPQAAEEPIFGMAAEEGGQPAEEPIIG
TAADEGEPQAAAESIFGSLADEGEPQAAAAGPIPATAAEEEGLSNPAAAATEGDLHSAAVQPPTTPAGPAT
VGAGTQHNMAPFTPANGDLESRRKEVVQPAAGCNTLATQVGGQSPPTPAHAQGQQLSKGIRWVACSHC
CGPAAAQEYAVAIPLTALVPDAFVTNPKLARAPIMPGTHSHYARNLFP*

Edaphoclamys debaryana

>EdadeDIV

MSVSVSFSDDYVQPTGVATFWRLEENKVFEVALAKHYADADRYERIASYLPNKSANDIQKRFRELEDDLRR
IDEGCSESGSAQSAPTPAGRSDEQPAAKKPKTDVPANGDRRKGVPWTEEEHRLFLLGLAKFGKGDWRSIA
RNFVVSRTPTQVASHAQKYFIRLNSMNKKDKRRASIHDITSPTLPASVPNANPTTGLTPTAPVSSSKQATG
PAASAGSPPAPPAAPAPSLPLAAAATAAMFPSAAAAAAAAAAAAAAAAQAGGQLFAQLAMHGLSLPAVAAPS
ANAMVPMTMPTNFMVSV*

>EdadeDRIF

MDMPAPNTSSQGSADKSNTTLRDAQAGLPAPLGSSQSDDFDTDFELQLQGTTGSDWTQEELATLESALA
RFPADKYLPVERYIHVAASLPSKTARDVALRVKACGLDEKSSKGPEPAAGGKRKGGRAGAGGKSAATAAG
GTAAGGSTAGAWTTTDDSSSGIPVALTQLMEQNYSILTQFKSNMSAFKVMENTELLIRYRDNLMAIQQQLS
TVGGIMGQMPPLPVQPNFELASKFLPQGVGKTPGLAMAMPMSMPMPIAAAPLPPVAAPMPMPVPAAPS
PAAAAAAAAAASAAAAAASMSGASMGMAGAAPAAPPAVSPTPPPMAMMAAFPMGPPGIPGMPPGMPPG
MPPGTAMAQAPGSMMAPMGAMDPSAFMAAAAAAAAAGGMPGAMAGVMPGVMPGMHAGMPGMHGS
MPGVMPGTMPGTMPGLMPGTMPGGMPGGMPGMPGGMPGMQAFQFMGAMGAPGMLGCLPMGAPP
GGMQMGAPCMPDGGAAAAAAAAAAAAAAAAAAAQHAGMPGAMGALSMSRQGSAAQPMSMMVPHVPV
VKQEGM*

Gonium pectorale

>GonpecDIV

MSASFSISGDYVQPTGVATFWRLEENKVFEVALAKHYLDADRYERIAAYLPNKTANDIQKRFRELEPTKKPK
ADVPANGDRRKGVPWTEEEHRLFLLGLAKFGKGDWRSIARNFVVSRTPTQVASHAQKYFIRLNSMNKKD
KRRASIHDITSPTLPASVPNANPTTGLAPSGSSGAPSADGTAAAKGLTPAAAATAIAAAPAVAAATAATAAGA
QSAAAAMAAAAAAAAAAASSSVFAQALAMHGMAMPAAAAAAAAAANMAGMVPAMGMPPAPFMVQV*

>GonpecDRIF

MATGGLYSNVPRASDIGANTSSSSHAEQQGAGAVPANNSSASGEDFDADFELQLQGTTGSDWTPEELAT
LDSALARFPADKYPPVERYIHVAASLPSKTARDVALRVKACGLDEKGRRDDSAKRKGGGSAGARSVGQQG
GKGSGGAGGSAAAADDASPTVPAALTQFMEQNYSILIQFKSNMAAFKVMENTELLMRYRDNLMAIQQQL
STIGGIMGQMPPLPVQPNFDLANKFLPQGSMKLPSMPMAASGMHQNPAAAAAAAAAAAAAMPGAHGM
SVPGAASTAGAVAGPSAVPGSAPMPQMPPMMPPYPFNPAAAAAAGVGAPPGMPGAMMPPGMDPAAFF

GAAAGMPGAMGVMSMGGMAGAMNPAAAAAAAAAMGAAMPGAQFPFAMGAPGMMACMPMGPMAMP
NPAMPDGGAAAAAAAAAAAAAMAAQAQQAAMPALPSAASMSRQGSAVQPMNMAIPMNVSVKQEGAV*

Tetrabaena socialis

>TetsoDIV

MAAGFSLCWDYVQPSGMATFWRLEENKVFEVALAKHYLDEDRYERIASYLPNKSLGDVQKRFRELEDDLR
RIDEGCSEGASEQSSAEPSPTRSDENMSQQPSKKAKTDVPANGDRRKGVPWTEEEHRLFLLGLSKFGKG
DWRSIARNFVVSRTPTQVASHAQKYFIRLNSLNKKDKRRASIHDITSPTLPAHAPNANPTTGSPPAAPASP
PQPAAPPCSGPAPVHSSLPTQAQQAAAMGVFTSLALGMGMSMQQNMTPTLAMPTAPFMVSC*

>TetsoDRIF

MATNGAAGAVPFRPSGGTDPPATHTSSQSSADKSSLREQPAGAPANNGSSNSGEDFDADFELQLQGTTG
SDWTPEELTILESALARFPGDKYPPVERYIHVAASLPSKTARDVALRVKACGLDEKGRRPPAEDTAKRRAG
GGIPPRPGAPPGGKGAGAGAAGAAGAGGPVPGDDPSPGVPAALTQLMEQNYGILTHFKSNMAAFKVMEN
TELLMRYRDNLMAIQQQLASVGGIMGQMPPLPVQPNFDLASKFLPPGSGKPPGLSGPPAMPGALPTSAA
AVAAAAAAAAAAAASMPGMSMNPVGAAQGPPGTSAAQLSAPTAAPGPPVLPMMPGFPFNPVGMPGMG
QPHGMPGGMMAPMHLDSFFGASGGMPGGMPGMGVLPPSLMAGALNPAAAAAAMGMPGAQFPFGMG
GPGFMGCMPMGAGPMGGASMPDGGAAAAAAAAAAAAVAAAAAHQVGMPGGVPGGMPPGLSMSRQGS
AVQPMPQAMGLAACMPVNIKQEG*

Volvox carteri

>VolcaDIV

MTASFSISVDYVQPRGVATFWRLEENKVFEVALAKHFLDVDRYERIAAYLPNKTASDVQKRFRELEDDLRRI
EEDHDSASAQSAPSPAPRIDENPAKKPKADVPANGDRRKGVPWTEEEHRLFLLGLAKFGKGDWRSIARN
FVVSRTPTQVASHAQKYFIRLNSLNKKDKRRASIHDITSPTLPASAPNANPTTGILPNGAAGSTAAAAAAAA
AAAAAAVVKAPSAASAAPSPVPSSPAPPPPAAPLVSSPAQSAAAAMAAAAAAAAASTSSVFAQLAMHGMSI
PQSGMVPPMAMPSAPYMVQV*

>VolcaDRIF

MASTAATFPAVPARPNGNLDVQTANTSSQSSAEKSNNTLREQQPGAPANNGSSASGEDFDTDFELQLQG
TTGSDWTPEELVVLESALARFPADKYAPVERYIHVAASLPSKTARDVALRVKACGLDDKARRAGLEDSSKR
KAGGGVQTRGNTQQGGKGSGAGTAGAGGAGAAGSVDDSSPGVPLVLTQLMEQNYTILAQFKSNMAAFKV
MENTELLVRYRDNLLAIQQQLSSIGGTMGHMPALPVQPNFELASKFLPTGGLKLPPAAPLPPVGGAPAAA

MAPPPAMAAAAAIAATAAAAASTSLCAAAATPPTGGSVGMAVTGPSPVAPQVAASPPGAALAGASPVAPQP
SLAAPTPLLPMMPPFPFSPAAGAAGLGPVPGLPPLPQMASMDSFFGAAAAGMPGGLGVMPPQLMPGAL
NQAAAAAAAMGAAGIPGAQFPFAIGNPGMLGCMPLAGGGAMPAPAMPDAGTPGVAAAAGVAAAAAVQQV
ALPVGVQSSMAPAPSMSRQGSAIQPVGVPVTIKQENGI*

Chloropicon primus

>ChlpriDIV

MAVSDAAGSAQAQQSGGGNDASNGSAAKANWWSPQEDKVFERVLSEKFGERLQDILEEISKQIETKDME
AVRRRYEQLEEDIKNIEAGRVPLPNYADSGSVATSGSRKGGKGSNGKKDQHSERKKGIPWTEEEHRLFLL
GLEKFGKGDWRSISRNFVVSRTPTQVASHAQKYFIRLSSMNKRDKRRASIHDITSVNQADVQNMASVHAA
RANGENLAAAQNQQGKPIYGGGMVSMPPASYVSNMMPQSGPQQYV*

>ChlpriDRIF

MELTKWNEEEQHQLDLALKKFPSEKFSPLGRYIKISGLLPQKSVRDVALRVKWLSKREEKKKKKGSESAG
KRKQDKADKAAGAASWQRQQSPVGEKANQLYASITEVLDRNIVVIKQIQQNMMHNKVRENTDLLLKFRE
NLIKAQGVMTNTGGIMKQMPPLPAQVNQQLVQAVLPAKNA*

Micromonas commoda

>MiccomDIV

MTTMDPFPNFSLDGLGLELGRGSMGGIDSVIPYEHWTVDEDKHFETSLAQIGDLDSDDMWGQFSAHIPG
KSMVGLKRRFNLLQEDIKNIESGRVPLPHYENHDGVLNTEGVVAPAKVDTAPVAPAPATQTNSGGSNGSK
SSSKKKGGKAPAAKTSDQERRKGIPWTEEEHRLFLLGLAKFGKGDWRSISRNFVISRTPTQVASHAQKYFI
RLNSLNKKDKRRSSIHDITSVNGAGDSAPNSSQNGQPMPTMVPMQPMASGPMSGNGYYGAPMGNSMG
YMQTPGMQTMYVQQPM*

>MiccomDRIF

MTDPLDPSAAHDHHVGLDLHDDHHHGMEKFELPNLDENMAHFPDFDDDDFGVPDMKHADALHANG
DHHHDLNGVRAEMNGGAMARMNGSAMPPPANKTEEPNRVATIDRRSDVRSPVGRMDGDAAAEAVASG
SHAGTAGTNLYDAQWSAEEQAVLERGMETYGADEHKSLWRYIKIAATLPAKGVRDVALRMRWMSRRAGK
NGDGARGSKRKGVANGGGGDGDGGGKGGGGGGKGKKKASTKPPSVFSVGLTSPPARANGGVVNGHHG
AANGVANGHHHHAQHHSHLNGGYAQQTPMRQPPGTDGGGGGRMSMGYAHMVDHMGNVVQTPGGG
MMMQSPVPTGMNTARYSENMGPPRSMTNAIIGGGGDAGSHGRSNHGAMMSGSMMSAGGYGMSGGG
NGMNVGYVSAPVVYDGVYGHQTAGGGMVQQHGGVMQQGSWGGPSQDGGGGYVAFNPSGGMGMHPP

SHHGGMVMVDPTGARLVHYPTMQINQLEEHGGPGRVTGVVGDILAENVGLVSQIRGNMDAMKPPRGNL
ELLARFRDNLMAAKEHLAQEEGASQMPPLPVDIDHQLANQILPAPDTNARVDVKQAAAKSPAKGAKGAK
GGGAGGGRGGRGGRGGRGGRGGRGGRGKS*

Micromonas pusilla

>MicpuDIV

MVDAFNANFSLDGLGLELGRSPGYGGLDMLIQDQWTVEDDKLFENTLAQFGDLDGEDSWTQFGANVPG
KSMVGLKRRFNLLQEDIKNIESGRVPLPHYDARNDTAHQQMMQPAHHAVPIAQVAQSNPTGNAKASSKG
SSGHSPKKGGGSGANASKNGANGAKAKSAPAKTTDQERRKGIPWTEEEHRLFLLGLAKFGKGDWRSISR
NFVISRTPTQVASHAQKYFIRLNSMNKKDKRRSSIHDITSVKGSGKGGNAKGAANNDGGGSSHSGSDASA
NIGGTSAALPVAARGGMGHAGGGYYASGPVPVGFMQTPGMYGVNAHGGM*

>MicpuDRIF

MTTTTTTTTNAAQNNPPTTTTPPPPKRATNLYDVTWTNDEQATLEAGLDDPSPTPAGWTRVPGRESLWR
YVRIAARLPNKGVRDVAMRVRWMKRKGIKSGAAAAAGGGGNGTKTGKTTGGKKAASAKTKAPRSTRGQP
QPRPGGSTATTSMGTHAHAPPGTYGYDAHQQHPQQQHGGGYPAVAPRGGGGGGAADATLFSPPTAMG
LPSGMGARLRWGPSPNGTHQVGGGPTGTPAYAQQQQQHHQQHPHAQQHAAHHQHQGAAMMHGGG
AMHPYGGYVMQHPPSQHQHQLQHQPQHHHHGTFAYVAAGPQGPYHAGALQHGQPTTTHPAYHHMTQ
RQHQQLNGGDHIFGPQAGYAATSPLLAEEPVGGSAQDDVFFAWPGGMDGAGLGMAGMAGMGMGMDG
GGMGMGMMDEGVSDIFRDNAELATEISKNLQHGAADENVPLLARYRDNLAHASAAIEPAGGGVLVDGLN
PLTSPPGMLPGTTTTGTTSHVFIEEEPAGRASAGGPNAAGAAVVVVEEEEEEEDASPEGTMTTTTADAEGG
GGGGGGAGTRARVSPLRGASARARAGKKSPSKSPARPARKGATTSVVSSPSPRGGTRRTRGSSG*

Ostreococcus lucimarinus

>OstluDIV

WTFEEDKFFETSLAQYDGSWPITGDDYWGQLQEQMPQKGVHDLKNRFSKLEEDVRNIEAGLVQLPDYDD
DSDHHSKAAPKTGEQERRKGVPWTEEEHKLFLLGLNKFGKGDWRSISRNFVISRTPTQVASHAQKYFIRL
NSMSKKDNKRRSSIHDITS

Ostreococcus sp. RCC809

>OstrcDIV

MSTLDFGYGGVDAGFTNLGLLGSDSVSWTWEEDKFFETSIAQYDGSWPIMGDDYWSRLQEKMPQKGVQ
DLKDRFTRLEDDVRAIESGLVPLPDFEDDSDHSKPAPKTGEQERRKGVPWTEDEHRLFLLGLNKFGKGD
WRSISRNFVISRTPTQVASHAQKYFIRLNSMSKKDNNRRSSIHDITSPTPKSSG*

Ostreococcus tauri

>OsttaDIV

MSTLDFGYGDAGFMYSSLGLLGTDSVSWSFEEDKFFETNLAQYDGWPITGDDYWGQLQQQMPQKAVQE
LKDRYAKLKEDIREIESGFVSLPEYYDEGVDSEDYVTAEVSFAPMKTVKAQPAAPAVQAPAPAAPPAKKSK
NVPKTGDQERRKGVPWTEEEHRLFLLGLNKFGKGDWRSISRNFVVTRTPTQVASHAQKYFIRLNSMSKK
DNKRRSSIHDITSATGRD*

Streptophyta

Chara braunii

>ChabraDIV

MLSETRSSGEASAATTMGDSCLQRFSDSDGGGSDEGQNGNMSTSSGCHVSSPVRWTPQEDKLFEQALA
DVDENDEARWEKVAARLPGKSIDDLVRHYELLVEDIIMIDEGRLALPAYNATSSVSGEAGLDPGGCGGGAG
VVGGGGGSSDTTMVVPSSPGTTSSGGGGGGGGGGGVKKQSSKLSSLGKSAEQERRKGIPWTEEEHRLFL
LGLQKFGKGDWRSISRNFVISRTPTQVASHAQKYFIRLNSQNKDKRRSSIHDITSVSNGDAMSQSQGPITG
QPAVAPQPIAHPHVHSHASPPPIQPGLYMTSVGQPMGPLPTVMPIRPPPGHHPSRAHLARPVGMTGTGIP
MPHMAAYVPQPAMHH*

Chlorokybus atmophyticus

>ChlatDIV

MGGQTAGGGATAAGVDLKKFEEALAQVDENDAGRWEKVAALVPQMTPAEVQREYDRLCEDVQVLETGNV
PMADFRETTASTPVAMVARPLTPSGTPMTATPVDDALLAHGMSADRPPTANGRLAQERRKGVPWTEEEH
KRFLVGLTRFGKGDWRSISRECVITRTPTQVASHAQKYFIRLSSTGKDKRRSSIHDITSIGQDGLPRQTPASA
QMQGQPGVGVPIARAPSTGQSPADGVPTAGFAPAPIMGVPGTPVVHPGYTPQAQMVPAVPTQGVPVQYM
APAPAIRQ*

>ChlatDRIF

MTRNAPDNVKALACLVQQRKTWACGQVVAEWQEEDAEEDDEVKEKDGGEILCQGAAGGSNGAGVQLAM
AAAAAALVSSSSAGAAGSDAAGAFPDVGMFDHFDMGEDDVVGGASITALSSLLGVEQDGGQGGQDGTSG
LENVWGAAVGAEELLLEQKPIVPRCCLGTPPFLQDWTVDEQRILEEGLQRYPADKMSNVMRYVRIAAMLP
EKSVRDVALRARWMSRKDNGKRRKPLEESVVAKKPRDKKVQPAPHIVAKPLPPVPALLPVETPSPLEACD
GSETGRLLEMNTAVANTIKQNLVHCKVEENGALLARMRDNILAIMNGLTSMPGITSMPPLPIKMDLELAN
TYLPPSLPTMTSTS*

Klebsormidium nitens

>KlenitDIV

MEQEPALGMQTEENEKKEVKAVKWENEAASPQLMTSPETSSQGLDSSDLGSEPHFFDNGSIHMRFGDD
GVSSGHAGDLSGWTPSENKLFENALNMYGEEDEARWNNIAGQVPGKTPDEVKRKYEQLLEDVRAIESGR
VPVIAYGDPHKQTASEEEMMRDDVGNGVEMAMPASPGGTRRPKVGQRSSEQERRKGIAWSEEEHRLFLL
GLAKFGKGDWRSISRNFVVSRTPTQVASHAQKYFIRLNSISKDKRRNSIHDITSVNGQGGHGPSQRGIGGH
PPIAPNLHPGQNMYGHHMMMNPGHMRPMGPPMGVPVNHMGGPPHMQPYVTTGGHQ*

>KlenitDRIF

MALSSDFMPDPGDPSLAGPTPGEPQAAQLLQPKLEHPEPILEVGNGLQSTGVLPATSMWTSSDDQKANS
ANGGTASTSSSPSFGAGLAHPNGALLPSAGTLVAPGPIVPVSGMIPTPMSNGSAGPAGAQGPGMPVQMH
EPALAVEWTAEEQKLLEEGLTRFPGDKYSNIVRCIKIAAMLADKTVRDVAMRCRWMSKKEIGKRRKDEQQ
QSQTKKSKESKKDKPGDMHRPLAPARPPVPIMAPPPMPPLDDKIPPIGGPTGQLLDENVRMVNQIRQNL
ANCKIQENNELLVKFRDNITTIINGMTSMPGILSSMPPLPVKLNTPLADSFLPPSKNLPLPPPVTTPPK*

Mesostigma viride

>MesovirDIV

MAPSQTQCAAPVASMIRQSQPLSNSAQDMEWPPELDMLFEKTLAKYAEETGQKRWQKVASVLPNKTPD
DVSRRYELLVDDIDKIEMGLFPLPDYSDDDLSLQVDMSMNGAHRNLGVMMAAPCVGPINGMRMDGVSM
GMPMTEVELGGDMMGGDGGLAAGGNRGGGIRTKNKMPGHAPPKSSSEQERKKGIPWSEEEHRLFLLG
LQKFGKGDWRSISRNYVITRTPTQVASHAQKYFIRLNSMNKDKRRASIHDITSLGNAVDMSGMVGGGGGG
PGAPLHQGTPAPITGTHAGMVGARPPQQQQGMAPAGLPGQAKPMVTTNGVNPAHGMYHNGMMQPAP
QPPQQGMHGRPMAAPGAMGHMAPGHHPGQMQGMQQGNANGAAIGIPATVPAQGPMQGMSVGYLPQ
SQPQMMVQWKAPQ*

>MesovirDRIF

MVLDSLPVMVKQEHLPGDDVGDGHLALGLDADLVTVSDPLHPSNCNADAKLKAEPGAASHPGGGTAGQ
VPGADGTLAAPGAPLPANGGMTPVSQDHLVTGAGGAPAVSTANAMPGGNQNGPAVVAGALPSASPVNG
NSSNGVAGAHGGVNGGGPRPPSAEWTSEEQKILDDTLQALVAKGSPNGAAGSCTNGGPVPAGTLVERYL
LVAEKLPNKSVRDVAYRCRWLARQREANKRKKMGEDANALRKVASKKSLHPQDVGPMGASGGPMVTTP
MPNQAEAFPSNLLDQNAKVLQAIQLNLQNMKLQENIQQLSLCRDNILCIQNCQLPPLPVSIKMDACNLLL
APLAATAAPTAGPVPS*

Mesotaenium endlicherianum

>MesenDIV

MEASSQPSNISTDPKAEAGTDGSQCSECPVSDGGWAPCGWTTADDKLFETILAGFEKEKDINWDNIATKI
PGKKLEDIRKHYDMLVIDVGNIDAGLVQVPDIVMAGQEISPADEGGVHDLGTSQSPTAKKVGSSCRPQGLL
QPRAAAPPQGRTTEQERRKGIPWTEEEHRLFLLGLAKFGKGDWRSISRNFVTSRTSTQVASHAQKYFIRLN
SVSSKDKRRSSIHDMTSIHNGESGAAAASGPITGQPAVPGGIYAQGRPHAIGHVHPVQQMPPPPGHLPYG
ARAHLPRAMMASPGMPIHQMGYVSSPAMHL*

>MesenDRIF

MPDLAMSESPVPNTAAAATMSPSGNPLPNGDVSSAEVKSGSQELQLLHHPILSSVWTAEEQRILDDNLAK
FSDEQQYSSLMRHIKLAALLPEKTVRDVALRCKWLAKNESGKRKREEPVSSKKSSKDKKEKAGEVHKGA
STSRPATGLPLLLPPSVPPPNAEALSPDALKGKTKQLLDQNAQVILQIRTNFSAMKVQVLHII*

Marchantia polymorpha

>MpDIV1

MAAPSPGPSPSSPSTASTPIPASAAAAAAGPSAAPVVPTPSAVSAPTIPVIPAPDVAPAVSSSWTSEQDKLF
ENALAVYDEESPNRWDNVASMVPGKDAADVMKHYELLTEDVTSIDAGRVALPSYILPGSLSGADAAGEQS
DSSVSKNKAWSGQSPGVSASGTSGTVGGLERKSSSSKADQERRKGIPWTEEEHRSFLLGLAKFGKGDWR
SISRNFVISRTPTQVASHAQKYFIRLNSINKDKRRSSIHDITSVNSAGVEVMQGSPGPITGQSPSGTSTSGQP
LPHKTQPAFQGGMYVTPVGPGATTGLGTPVYPPGQMGYGVRGHMVRPGMGGPPMNMTHMTYSMPQP
AMHH*

>MpDIV2

MATTVAQTAGSPGQLPIPPPWTPKLDKLFEKALAIYDGDSPDRWEKIAAKLPGVDPTEVKKHYDRLIEDLN
SIETGRVALPNYHKSVGLMSSSDEDSPTSRKPVHGGHHGLGGSNANATNATSLPSGKAPSSKASDPERR
KGIPWSEEEHRLFLLGLAKFGKGDWRSISRNFVVSRTPTQVASHAQKYFIRLNSINNKDKRRSSIHDITSVN

DGDSLPQSPGPITGLPSPGAQWPRSGLQGASMYDMGGMGGPDQAIGGQMLMTPTGHPHHVPYGHVPV
MQGPPMAMQHMSYPMPQSANA*

>MpDRIF

MAGSVGNNSTTNSSAAATSASPAVNGNHSSMYNSNAQGASTQSTTTTINSGNNGISRPNGPATNGSGNG
TNSVASDQPPPLQLQLLHDPGITADWSSEEQATLDDGLTKFAGETSNLAKYIKIANLLPEKTVRDVAMRCR
WMTKKEIGKRRKPEDQNASKKNKDKKDKSDSMSTKAPTGHIRPGLSSYTAPTPNVDNDDGISNDAIGGT
TGQLLEQNSHVILQIRSNLAAMKLQENTELLVRFRDNICAILNGMTNMPGIMSQMPPLPVKLNTELADTIL
PKSLPQASPTSQT*

Annex D: Multiple sequence alignment

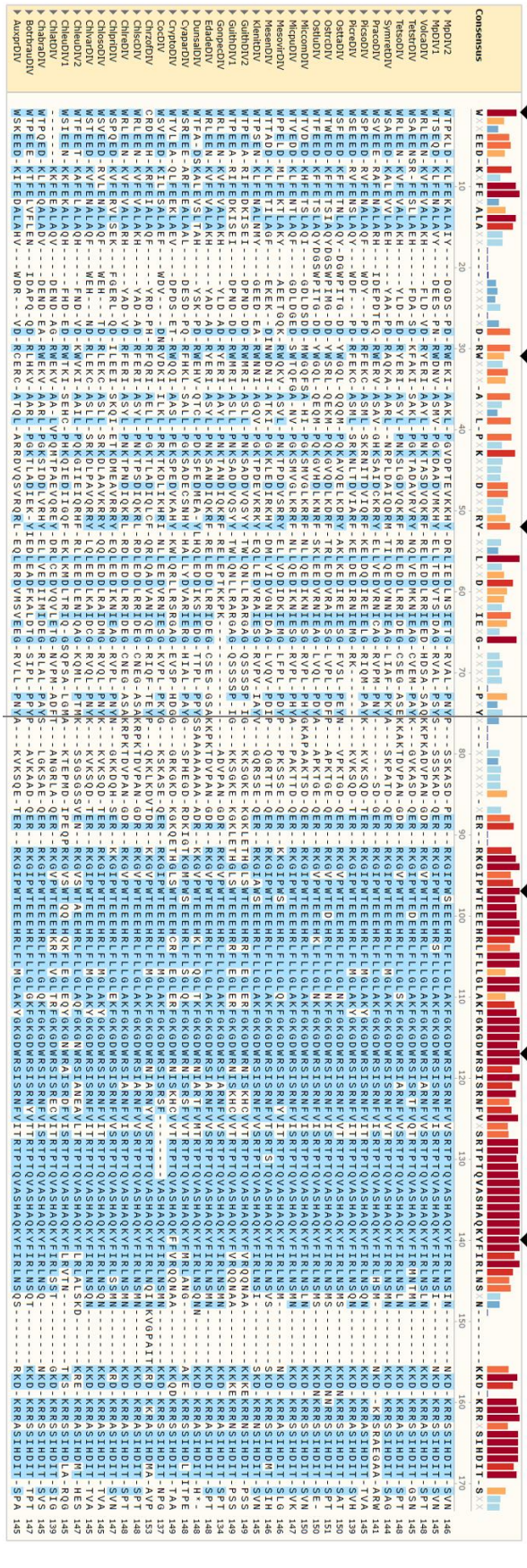

Figure D-1. Multiple sequence alignment of DIV homolog domains peptide sequences. The peptide sequences of the MYB domains of DIV homologs were isolated from the remainder of the protein and aligned by MUSCLE. Amino acids shaded with blue correspond to the consensus sequence, which presents amino acids that are conserved in at least 60% of sequences. The coloured bars above are bigger and redder the higher the conservation of the amino acid below. Arrows point to characteristic aromatic residues. Between sequences and consensus, alignment ruler numbers amino acid positions for reference.

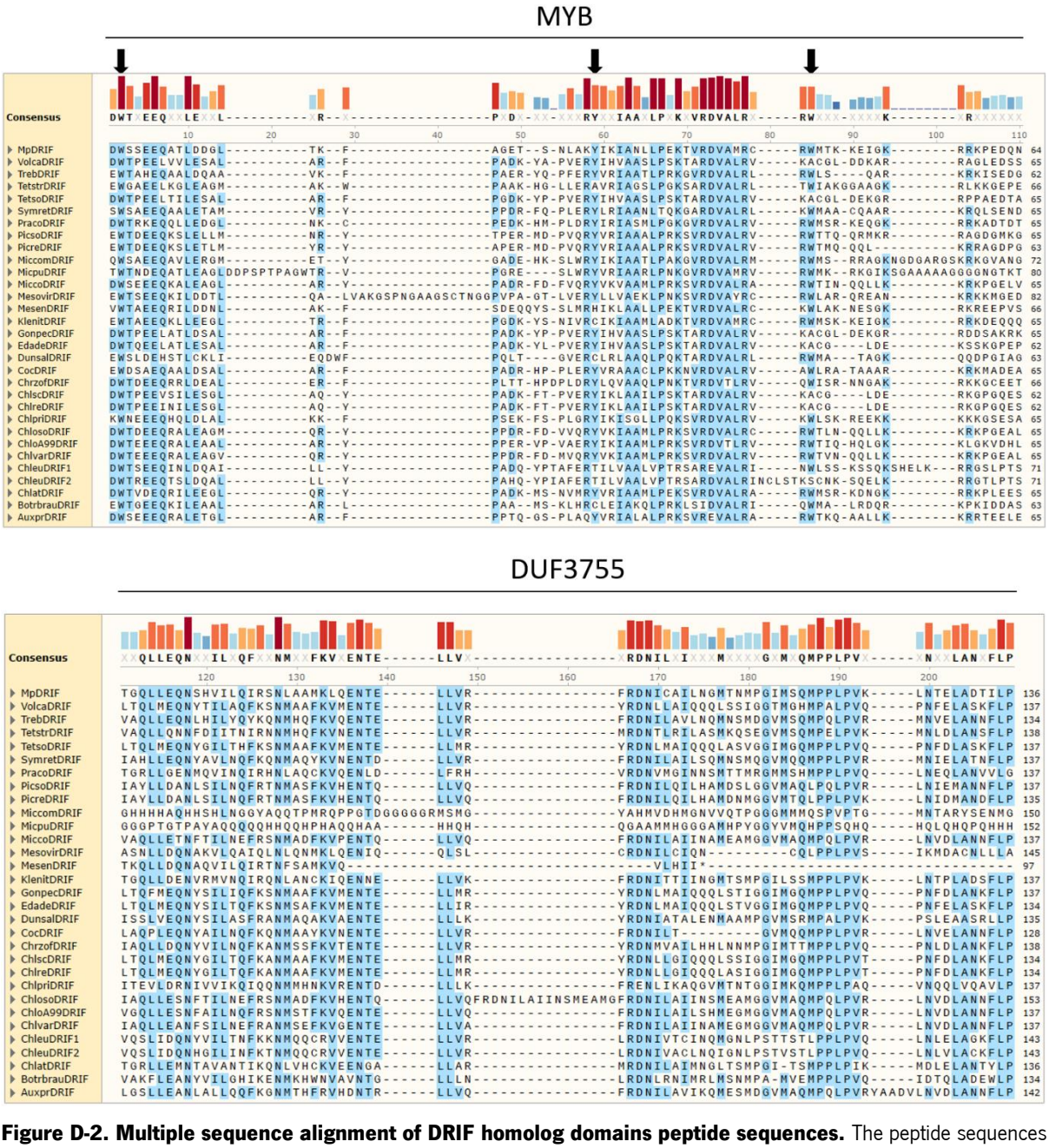

Figure D-2. Multiple sequence alignment of DRIF homolog domains peptide sequences. The peptide sequences of the MYB and DUF3755 domains of DRIF homologs were isolated from the remainder of the protein and aligned by MUSCLE. They are presented separated to improve legibility. Amino acids shaded with blue correspond to the consensus sequence, which presents amino acids that are conserved in at least 60% of sequences. The coloured bars above are bigger and redder the higher the conservation of the amino acid below. Arrows point to characteristic aromatic residues. Between sequences and consensus, alignment ruler numbers amino acid positions for reference.