STRESS RESPONSES OF THE FLAX GENOME: ACTIVATION OF TRANSPOSABLE ELEMENTS, DEFENSE GENES AND GENOME RESTRUCTURING AND DIVERSIFICATION.

By

Leonardo Galindo González

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor in Philosophy
in
Plant Biology

Department of Biological Sciences
University of Alberta

Abstract

Transposable elements (TEs) are DNA sequences that can move in the genome (transpose) via a DNA or a RNA intermediate. TEs are abundant in plant genomes and can be activated by stress. Their activity can result in structural and gene expression alterations that increase diversity within and among species.

Canada is a major exporter of flax, which is a source of valuable seed and stem fiber-derived bioproducts. To understand and improve characteristics of seeds and fibers, genomic approaches are now being applied. The sequencing of the flax genome, led by our lab group, showed a landscape with over 43,000 protein coding genes and more than 23% of TE genome coverage. The largest group of TEs were the Ty1-*copia* elements, and bioinformatic analysis indicated that many of them could still remain active. While these mobile elements are widespread in the flax genome, their influence in diversification, gene expression and genome restructuring is yet to be assessed.

In the current study we anayzed members of the Ty1-*copia* superfamily in flax cultivars, and potential elicitors of TE activation. One of these elicitors (fungal inoculation with *Fusarium oxysporum*) allowed us to analyze the general defense response of flax to this pathogen which constitutes a threat to flax cultivation. Finally, we designed a reverse genetics methodology to find mutations in genes of interest that could be related to phenotypic changes and used in the future to dissect TE-controlling mechanisms used by the host genome.

We first compared flax cultivars using TE-derived molecular markers, and found that retrotransposition events have occurred since breeding began and that TE polymorphisms

allowed us to separate flax types. Most TE insertions derived from these polymorphisms fell in close proximity or inside genes, and can potentially alter gene expression.

We then tested potential modulators of TE transcription including wounding, fungal extracts, fungal infection, and different plant tissues. The analyses with end-point PCR, quantitative reverse transcriptase PCR, and RNA-seq, showed little evidence that most treatments affected TE activation, but many TE families had high constitutive expression. TE expression across plant tissues resulted in differences that indicate that a better resolution on TE expression modulation can be found when studying meristems and reproductive tissues.

While inoculation with *F. oxysporum* did not alter TE expression, the RNA-seq used to survey TE changes gave additional information on gene regulation upon fungal infection. Most expected defense mechanisms were activated when flax was challenged with *F. oxysporum*: detection of fungal elicitors, signal transduction cascades, transcriptional reprogramming, activation of defense genes, hormonal signalling, and secondary metabolism modulation. However, the activation of certain genes involved in auxin regulation, cell growth, cell wall expansion, water and nutrient mobilization, plus the repression of major latex proteins, indicated possible manipulation of the host by the pathogen to facilitate infection. Many of the genes found related to plant defense constitute good candidates to analyze relationships between gene expression and disease resistance across flax cultivars. In the meantime, the modulation of unexpected genes opens a door to study cross-kingdom epigenetic manipulation mechanisms (e.g. small RNAs).

Finally, we designed a reverse genetics methodology to simultaneously test hundreds of flax mutagenized lines, to discover mutations in genes of interest, using next generation sequencing Ion Torrent technology. Several mutations were found in cell wall and metabolism

genes, but with no phenotypic effects. However, this methodology can be applied in the future to detect mutated genes involved in the process of epigenetic modification (e.g. methylation) that results in TE silencing, to test the effects on TE activation.

Preface

This thesis work is presented to fulfill the requirements for the degree of Doctor of Philosophy in Plant Biology.

Chapter 1 is based on an article that will be submitted for publication: Galindo-González L.; Mhiri C.; Deyholos M.K.; Grandbastien M.A. 2016. Ty1-*copia* elements in plants: engines of evolution. Target journal: Mobile DNA. Leonardo Galindo-González wrote the review; Corinne Mhiri, Michael Deyholos and Marie-Angèle Grandbastien, revised the manuscript.

Chapter 2 is based on an article accepted for publication: Galindo-González L.; Mhiri C.; Grandbastien M.A.; Deyholos M.K. 2016. Ty1-*copia* elements reveal diverse insertion sites linked to polymorphisms among flax (*Linum usitatissimum* L.) accessions. BMC Genomics (submitted June 8-2016: manuscript ID GICS-D-16-00881, 57 pages). Leonardo Galindo-González performed all experiments and analyses, and wrote the article; Corinne Mhiri and Marie-Angèle Grandbastien supervised the experiments and revised the article; and Michael Deyholos revised the manuscript.

Chapter 3 is based on unpublished research on the effects of different stress elicitors on TE transcription. Leonardo Galindo-González performed all experiments and analyses; Corinne Mhiri, Marie-Angèle Grandbastien and Michael Deyholos supervised the experiments.

Chapter 4 is based on a published article: Galindo-González L. & Deyholos M.K. 2016. RNA-seq transcriptome response of flax (*Linum usitatissimum* L.) to the pathogenic fungus *Fusarium oxysporum* f.sp. *lini*. Frontiers in Plant Science. 7:1766. Leonardo Galindo-González performed all experiments and analyses, and wrote the article; Michael Deyholos supervised the experiments and revised the manuscript.

Chapter 5 is based on a published article: Galindo-González L.; Pinzon-Latorre D.; Bergen E.A.; Jensen D.C,; Deyholos M.K. 2015. Ion torrent sequencing as a tool for mutation discovery in the flax (*Linum usitatissimum* L.) genome. Plant Methods. 11:19. Leonardo Galindo-González

designed the project, standardized the methodology for DNA extraction, performed all next generation sequencing experiments, performed the pilot bioinformatics analysis and wrote the article. David Pinzón-Latorre collected the plant material, performed DNA extractions, helped with bioinformatics analyses and revised the article. Erik Bergen helped with DNA extractions and preliminary PCR, performed bioinformatics analyses and revised the manuscript. Dustin Jensen wrote the Phyton script to automate the bioinformatics analysis. Michael Deyholos supervised the experiments and revised the manuscripts.

Chapter 6 contains the general discussion, ongoing research and future perspectives.

For my beloved Colombia,
that has endured over 50 years of violence.
May each one of us who were born on that soil,
work to bring peace and stability for all.

“Science is a way of thinking much more than it is a body of knowledge”
Carl Sagan

Acknowledgements

First I would like to thank Dr. Michael K. Deyholos for being a true mentor during my PhD. Mike fully supported all my decisions, provided valuable advice and discussion, and allowed me to grow as a scientist by giving me freedom to develop my own ideas, make mistakes, and provided guidance every time he considered it necessary. I also want to thank the members of my supervisory committee who provided timely help throughout the development of my degree. To Dr. Stephen Strelkov for his guidance in the world of plant pathology, both his technical and theoretical advice were of great value. To Dr. Jocelyn Hall for her interesting discussions in phylogenetics and for pushing me to understand the use of this discipline. To Dr. Corey Davis who not only supported me with technical aspects of my experiments, but acted as co-supervisor after Mike accepted an offer as Head of Biology in UBC Kelowna.

Likewise, my sincere gratitude goes to the members of the Deyholos lab for the valuable exchanges in their different levels of expertise. I would like to thank Dr. David Pinzón-Latorre with whom I worked side by side in several projects, and who provided me with technical support in many occasions. To Mary De Pauw, who helped me to settle in the lab in the first years I worked for Mike before starting my PhD. To Dr. Neil Hobson, who was unselfish in sharing all his knowledge in molecular biology. To Erik Bergen, for his valuable technical support and his diligence and discipline when helping in my projects. To Eva Fernandez for her unconditional friendship which goes on until today. I also want to thank Dr. Khalid Rashid from Agriculture and Agri-Food Canada, for supplying the fungal isolates necessary for the pathology experiments, and the members of the Cooke and Taylor lab for their technical support.

I greatly appreciate the support from the members of the Molecular Biology Research Unit. To Sophie Dang for her teachings and patience during my training with Ion Torrent technology. Thanks also to Cheryl Nargang for her technical support and conversations about life. I am deeply grateful with Troy Locke, because his help was key to most of what was accomplished in my PhD. Troy did not only provide his expertise in technical aspects, but was a guide and peer in important scientific discussions almost on a weekly basis. Troy went above and beyond his responsibilities and was both an excellent professional and friend.

My gratitude also goes to all the people in the Department of Biological Sciences, faculty and staff, who helped me further my scientific knowledge and collaborated in all administrative

aspects during my degree. Special thanks to Dr. Collen Cassady St. Clair who became a mentor in the later stages of my PhD; and to Chesceri Mason Gafuik, who in spite of always being swamped with work, would greet me with a smile and help me without hesitation.

Special thanks to my advisor, Dr. Marie Angèle Grandbastien, during my interniship at the Institut National de la Recherche Agronomique (INRA) in Versailles France, for all her teachings in the field of transposable elements and her continuing mentorship to this day. I also thank Dr. Corinne Mhiri, for all her time and dedication during my training in the institute, and to all the members of the institute for helping me to feel at home during my five-month visit.

I am grateful for the contributions of my funding agencies: Natural Sciences and Engineering Council of Canada (NSERC), Alberta Innovates Technology Futures (AITF), Total Utilization of Flax Genomics (TUFGEN), Labex Saclay Plant Sciences (SPS), and the Department of Biological Sciences and the University of Alberta.

A special thank you to my collaborators: Dr. Mauricio Quimbaya from Javeriana University in Colombia, Dr. Lina Quesada and Dr. Liliana Maria Cano from North Carolina State University, and Dr. Juan Jovel from the Faculty of Medicine at the University of Alberta, for their support in the development of projects related to my thesis work.

I want to thank my friends: Adriana Almeida, Adriana Arango, Mónica Molina, Rosa González, Duina Posso, Jaime Pinzón, Claudia Castillo, Nicolás Abarca, Jenny Fonseca, Catalina Solano, Andia Chaves, Eva Fernandez, Mónica Higuera, Po-Yuan Ku, Maryam Chamanifard, Tara Narwani, Troy Locke, Mattéa Bujold, Elisa Verma, Claudia Saenz Falchetti, Fernando Castellanos, Mark Solomons and Altin Kurti, for their love and support through all these years in Canada, and the unforgettable shared moments. A special thanks to Mauricio Quimbaya, who has supported me with his incomparable friendship and has also done everything in his hands to further my professional development.

Finally, I want to thank my family. Thanks to my parents, Nancy and Antonio, for teaching me about love and respect for others, for your unconditional support in all my decisions and for being with me in every step of the way. To my little one, Seb, because you have made me a better person through your love; you are my joy and reason to fight for a better world. And to my wife, Angelica, because you are the brightest person I have ever known, your light is always a guide in my life and touches everyone around you. You are the main reason I made it through this challenge, and I will forever hold in my heart all your love and support.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

4CL	Cynnamoyl-CoA reductase
AAP	Amino acid permease
ABA	Abscisic acid
ABC	ATP-binding cassette
ABRE	Abscisic acid responsive elements
Ac	Activator
ACO	1-aminocyclopropane-carboxylate oxidase
ACS	1-aminocyclopropane-carboxylate synthase
AFLPs	Amplified fragment length polymorphism
ALA	α-linolenic acid
ALS	Acetolate synthase
Alu	*Arthrobacter luteus* element
AOC	Allene oxide cyclase
AoPR	*Asparagus officinalis* pathogenesis-related
AOS	Allene oxide synthase
Aox	Alternate oxydase
ARM	Armadillo
ATAF	*Arabidopsis thaliana* Transcription Activation Factor
Avr	Avirulence
BARE	Barley retrotransposon
BGI	Beijing Genomics Institute
bHLH	basic-helix-loop-helix
BRCA	Breast cancer
BSA	Bovine serum albumin
bZIP	basic-region leucine zipper
C	Cytokinesis
C2H2-type zinc finger	Cysteine$_2$Histidine$_2$-type zinc finger
C4H	Cinnamic acid 4-hydroxylase
CACTA	CACTA motif

CAD	Cinnamyl alcohol dehydrogenase
CBL	Calcineurin B-like
CC	Cortical cells
CDA	Cell death and aging
CDC	Crop development center
CEBiP	Chitin elicitor binding protein
Cf	*Cladosporium fulvum* resistance gene
CHI	Chalcone isomerase
chit	Chitinase
CHS	Chalcone synthase
CIPK	CBL-interacting protein kinase
CL	Cluster
CLE	Cyclic peptide
CR	Cell rescue
CTAB	Cetyl trimethylammonium bromide
CTL	Chitinase-like
CU	Carbohydrate utilization
CUC	Cup-shaped cotyledon
CWB	Cell wall biogenesis
CYP	Cytochrome family protein
DArT	Diversity arrays technology
ddm	Decrease in DNA methylation
DEK protein	(DK are the initials of the leukemia patient from whom this gene was cloned)
DFR	Dihydroflavonol reductase
DHPLC	Denaturing high-performance liquid chromatography
DIRS	*Dictyostelium* intermediate repeat sequence
DMSO	Dimethyl sulfoxide
DOF	DNA-binding with one finger protein
DPI	Days post-inoculation
DR	Defense-related

Ds	Dissociation
DSB	Double strand break
DUF	Domain of unknown function
DYW	DYW motif
EDTA	Ethylenediaminetetraaceitc acid
EFIA	Elongation factor 1-α
EMS	Ethyl methanesulfonate
EREPB/AP	Ethylene-responsive element binding protein/Apetala
ERF	Ethylene response factor
ES	Extracellular secretion
ET	Ethylene
ETI	Effector-triggered immunity
ETIF	Eukaryotic translation initiation factor
Foln	*Fusarium oxysporum* f. sp. *lini*
FPKM	Fragments per kilobase of transcript per million fragments mapped
FS	Fiber spring
FW	Fiber winter
GAG	Group-specific antigen
GAPDH	Glyceraldehyde 3-phosphate dehydrogenase
GDSL	GDSL motif
GH	Glycosyl hydrolase
GO	Gene ontology
GOS	Golgi SNARE
GST	Glutathione s-transferase
hAT	Hobo Ac Tam3 element
HAT_KAT11	Histone acetyltransferase / lysine acetyltransferase 11
HCT	Shikimate quinate hydroxycynnamoyltransferase
HPL	Hydroperoxide lyase
HR	Hypersensitive response
HRE	Heat response element
HSF	Heat shock (stress) factor

IAA	Indol-acetic acid
IAOx	Indole-3-acetaldoxime
IAR	IAA-alanine resistant
ILL	IAA-leucine resistant
INRA	Institut National de la Recherche Agronomique (National Institute for Agronomic Research)
INT	Integrase
iPBS	Inter-primer binding site
IPTG	Isopropyl β-D-1-thiogalactopyranoside
IRAP	Inter-retrotransposon amplified polymorphism
IRE	Incomplete root hair elongation
ISP	Ion sphere particle
ISPs	Ion sphere particles
ISSR	Inter-simple sequence repeat
ITIS	Identification of transposon insertion sites
ITS	Internal transcribed spacer
JA	Jasmonate
JAZ	Jasmonate-zim-domain protein
LARD	Large retrotransposon derivative
LB	Lysogeny broth
LecRK	Lectin protein kinase
leuc_diox	Leucoanthocyanidin dioxygenase
LF	LTR_finder
LINE	Long interspersed repetitive element
LOX	Lypoxygenase
LRR	Leucine-rich repeat
LRRNT	Leucine-rich repeat N-terminal
LS	LTR_STRUC
LTP	Lipid transfer protein
LTR	Long terminal repeat
LysM RLK	LysM domain-containing receptor-like kinase

MAPK	Mitogen-activated kinase
MATE	Multidrug and toxic compound extrusion
mcn	molecule copy number
ME-Scan	Mobile element scan
met	DNA methyltransferase
MHC	Major histocompatibility complex
MIP	Major intrinsic protein
MITE	Miniature inverted repeat transposable element
MLP	Major latex protein
MSF	Major facilitator superfamily
Mu	Mutator
MULE	Mutator (Mu)-like element
MYB	Myeloblastosis
MYC	Myelocymatosis
NAC	NAM-ATAF1-CUC2
NAM	No apical meristem
NBS	Nucleotide binding site
NCBI	National Center for Biotechnology Information
NGS	Next generation sequencing
NJ	Neighbor joining
NOL	Non-ribosomal protein
NR	Nitrate reductase
nr	non-redundant
NSERC	Natural Sciences and Engineering Council of Canada
OPR	12-oxophytodienoate reductase
OS	Oil spring
OW	Oil winter
PAMP	Pathogen-associated molecular pattern
PBS	Primer binding site
PCD	Program cell death
PCF	Proliferating cell factor

PDA	Potato dextrose agar
PDR	Pea dispersed repeat
peroxid	Peroxidase
PGM	Personal genome machine
PI	Protease inhibitor
Pif	P instability factor
PIP	Plasma membrane intrinsic protein
Pit	Rice blast resistance-t
PLR	Pinoresinol-lariciresinol reductase
PME	Pectin methylesterase
PMEI	Pectin methylesterase inhibitor
pol	Polymerase
POX	Polyphenol oxidase
Ppcrt	*Pyrus pyrifolia* copia retrotransposon
PPR	Pentatricopeptide
PPT	Polypurine tract
PR	Pathogenesis-related / Protease
PRR	Pattern recognition receptors
PRX	Peroxidase
PSY	Phytoene synthase
PTI	PAMP-triggered immunity.
PVP	Polyvinylpyrrolidine
PYR	Pyruvate carboxylase
qPCR	Quantitative PCR
qRT-PCR	Quantitative reverse transcription PCR
R	Repeat region
RAB	Rabgap/TBC domain containing protein
RAPD	Random amplified polymorphic DNA
RBIP	Retrotransposon-based insertion polymorphism
RBO	respiratory burst oxidase
rdf	reduced fiber

REMAP	Retrotransposon-microsatellite amplified polymorphism
R-genes	Resistance genes
RH	Ribonuclease H
RIPK	RPMI-induced protein kinase
RLC_Lu	Retrotransposon LTR copia *Linum usitatissimum*
RLCK	Receptor-like cytoplasmatic kinase
RLK	Receptor-like kinase
RLP	Receptor-like protein
RmlC	dTDP (deoxythymidine diphosphates)-4-dehydrorhamnose 3,5-epimerase gene C-like
RNAse H	Ribonuclease H
RNI	Ribonuclease inhibitor-like
ROS	Reactive oxygen species
RPKM	Reads per kilobase of transcript per million mapped reads
RPMI	Resistance to *avrRpm1*
RPP	Resistance gene *Perenospora parasitica*
RT	Reverse transcriptase
RT-PCR	Reverse transcription PCR
SA	Salicylic acid
sad	Suppressor of ascus dominance
SAR	Systemic acquired resistance
SERK	Somatic embryogenesis receptor kinase
SH	SRC homology
sil	Silver locus
SINE	Short interspersed repetitive element
SNARE	SNAP receptor
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SPRI	Solid-phase reversible immobilization
SPS	Saclay Plant Sciences
SRA	Sequence read archive

SSAP	Sequence-specific amplification polymorphism
SSR	Simple sequence repeat
sth	*Solanum tuberosum* hypersensitivity
sub	subunit
TAIR	The *Arabidopsis* information resource
TAN	Tel1/ATM N-terminal motif
TAZ	Transcription adaptor putative zinc finger
Tb	Teosinte branched
Tc	Transposon *Caenorharbditis elegans*
TCP	Teosinte branched 1 – cycloidea – proliferating cell factor 1
TD	Transposon display
TEF	Transcription elongation factor
TEs	Transposable elements
TF	Transcription factor
Tfb	Transcription factor b
TFBS	Transcription factor binding site
TIP	Tonoplast intrinsic protein
TIR	Toll/Interleukin 1 receptor
TLC	Transposon *Lycopersicum chilensis*
Tnt	Transposon *Nicotiana tabaccum*
Tos	Transposon *Oryza sativa*
treh6p_synt	Trehalose 6 phosphate synthase
TSD	Target site duplication
Ttd	Transposon *Triticum durum*
Tto	Transposon tobacco
TUFGEN	Total Utilization of Flax Genomics
Ty1	Transposon yeast 1
Ty3	Transposon yeast 3
U3	3’ untranslated region
U5	5’ untranslated region
UBI	Ubiquitin

UBP1b	Oligourydilate binding protein 1B
UGT	UDP-glycosyltransferase
WRKY	WRKY motif
Xa	*Xanthomonas* disease resistance gene
X-gal	5-bromo-4-chloro-3-indolyl- β-D-galactopyranoside
XV	Xylematic vessels
ZZ type	Zinc-Zinc binding type

CHAPTER 1 – General introduction

This chapter is based on an article that will be submitted for publication: Galindo-González L.; Mhiri C.; Deyholos M.K.; Grandbastien M.A. 2016. Ty1-*copia* elements in plants: engines of evolution. Target journal: Mobile DNA.

1.1 Flax

Flax (*Linum usitatissimum* L.) is a species from the family Linaceae, which comprises over 300 related species [1], with 180 species belonging to the genus [2]. Flax and its relatives were among the first plants cultivated as crops, with records of use of *Linum bienne* dating 7000-8000 B.C. in Mesopotamia [3], but accounts of use of flax fibers used in hunter-gathering communities date to around 30,000 years ago [4]. According to the most recent data, selection for oil-associated traits preceded selection for fiber traits during domestication, but apparently various domestication events occurred to select for different traits [5]. Modern flax (*L. usitatissimum*) was probably domesticated from *L. bienne* Mill., which is also known as *Linun angustifolium* Huds. [3,5]. The most likely center of origin is the Indian subcontinent which maintains the greatest diversity of the genus [3].

Flax grows as a summer annual in temperate regions, but diverse flax types (fiber and oil or linseed) are grown as winter annuals in Europe [6]. The plant has a tap root, an erect main shoot, and a panicle-type inflorescence. The five-petal flowers occur in a variety of colours including blue, pink and red, with the former being the most common. Fruits are capsules that can bear a maximum of 10 seeds. The plant can grow between 20 to 150 cm high depending on the variety. Dense planting supresses branching and is used when growing fiber varieties, while low-density planting promotes branching and is preferred for seed production [6]. Fiber flax is produced by many countries especially in Europe, but most flax is produced for oil [7]. In this respect Canada is the top producer and exporter of flax seed with 875,000 t produced in 2014-2015, and 80% exported mainly to China, the U.S. and Europe [8].

Fibers produced from the stem are used in the textile industry, but current uses also include different types of composites. The seed contains oils that can be used for nutritional purposes and industrial products [6]. The most important characteristic of the flaxseed oil is its high content of α-linolenic acid (ALA) which can reach 55% of total oil in some varieties [9], but the seeds are also relatively rich in protein, fiber and lignans [10]. ALA and lignans have been reported to have beneficial effects in human health including protection against cardiovascular disease, diabetes, and some types of cancer [11–13].

The major flax diseases are: rust, wilt, pasmo, blights and rots, and these are generally due to fungi. Probably the two most damaging pathogens are *Melampsora lini* (rust), and *Fusarium oxysporum* f. sp. *lini* (wilt) [14]. Earlier in the 20th century, fusarium wilt was a big

problem in North America, but nearly all currently cultivated genotypes are at least moderately resistant, and this resistance is polygenic. *Fusarium* enters the plants through the roots and colonizes the water-conducting vessels, interfering with water uptake, causing wilting and finally death of the plant [14].

L. usitatissimum and its presumed wild ancestor (*L. bienne*) both have 30 chromosomes, while sister clades have 2n = 16 or 18 [15]. The divergence in chromosome number among flax-related species is due to genome duplications and loss of chromosomes. In fact paleopolyploidy events have been documented at 5-9 mya and at 20-40 mya [15,16]. The genome of *L. usitatissimum* (cultivar CDC Bethune) was recently sequenced and annotated, showing a total of 43,384 protein-coding genes [15], and over 23% of its sequence identified as transposable elements (TEs).

1.2 Plant transposable elements (TEs)

During the 1940's and 1950's Barbara McClintock, an American cytogeneticist, studying chromosomes in maize, noted a large number of mutable loci that affected variegation of different plant characteristics [17,18]. The changes that could be detected at the chromosome level included deletions, duplications and structural modifications. She believed that these unstable genes caused the same type of instability in all organisms, and with very basic genetic tools she discovered how the breaks occurring at one specific locus where related to changes in other loci, and that such changes had effects on the expression of neighboring genes. The mutable loci were apparently generated by the transposition of an element (Dissociation –*Ds*-) to a new location where the receptor locus became affected (mutable locus). The removal of this *Ds* element restored the action of the mutable loci, and the actual transposition of the *Ds* required an Activator (*Ac*) element which could itself also transpose and cause mutations [17,18]. The techniques in the time of McClintock did not allow her to isolate the DNA sequence of these elements, and the community of geneticists resisted the intricate complexity of her findings and she stopped publishing on the topic a few years later. The mobile elements would finally be isolated and characterized in 1983 (same year that McClintock received the Noble Prize in Physiology and Medicine), showing that in effect, the *Ds* element was almost identical to *Ac* but carried a deletion [19]. This deletion of the transposase enzyme (the enzyme necessary for transposition) made the *Ds* element dependant on the transposase from *Ac* for its transposition.

This clearly supported the results McClintock obtained 30 years earlier. Barbara McClintock saw transposable elements as modifiers of gene action upon stress (shock), but also, actors of genome restructuring through major genomic rearrangements [20]. The misinterpretation of her findings and statements resulted in people thinking that TEs by themselves could promptly change a genome so radically as to escape extreme events of stress; however, she only spoke generally about the potential of the elements and gave no time frame as to how they could influence changes during evolution [20,21].

When enough evidence accumulated, TEs were first viewed as junk or selfish DNA. The replicative and cumulative character of TEs (especially from prokaryotes which were well-characterized at that point), and their initial characterization as mutable agents resulted in thinking that mobile elements were part of a large fraction of junk DNA that did not confer any advantage to the host genome [22,23]. Under natural selection, a genome should accumulate genes that contribute to fitness, and TEs were apparently accumulating in large numbers without a direct advantage, which gained them their nickname of selfish DNA. Furthermore, the increased cost of replicating non-useful sequences would be detrimental for an organism bearing more selfish and junk DNA than useful DNA (e.g. gene coding) [23]. However, if the cell is considered as an environment for competition, DNA sequences would evolve and compete with the only objective of self-preservation, and not necessarily to give the host organism a short or long term phenotypic benefit. Under this theory called non-phenotypic selection, where mutation could increase the probability of survival of certain DNA sequences, TEs could appear and have no direct influence on organismal fitness [22].The characteristics of transposable elements, where many copies insert randomly in the genome with a low probability of selective advantage, fits this model of self-preservation [22]. However, even with the view of TEs as selfish DNA, it was suggested that, while most junk DNA would not confer any advantage, some insertions would fall at the right place and time to generate new, useful controlling mechanisms for genes, and that some of these events could become fixed by natural selection [22,23]. This inference is at least partly true today, as accumulating evidence shows that numerous TEs are highly related to genes, and can confer evolutionary advantages [21,24–27].

1.2.1 TE types

Transposable elements are now known as widespread components in plant genomes [28,29] and can comprise anywhere from over 14% of the genome in *Arabidopsis thaliana* [30], to 80% in maize [31]. They constitute DNA fractions that can move through the genome and create new insertions thanks to two basic mechanisms that correspond roughly to the two large classes of TEs. In class I elements (retrotransposons) transcribed mRNA encodes a protein (reverse transcriptase) that allows the mRNA to be retrotranscribed into a double stranded DNA molecule, and then a second protein (integrase) encoded on the mRNA creates cuts and integrates the double stranded DNA into a new genomic location. This leaves a copy of the mobile element at the original locus, and creates a new copy at a different locus, and therefore this mechanism is commonly known as copy and paste. In class II elements (DNA transposons), the encoded transposase enzyme excises the full transposon from one genomic location and creates new cuts in a different location to reinsert this sequence. In this sense this is known as a cut and paste mechanism. There are also additional TEs that generally fall into these two classes, but have slightly different enzymes and replication mechanisms (e.g. *DIRS* elements, *Helitrons*) [32].

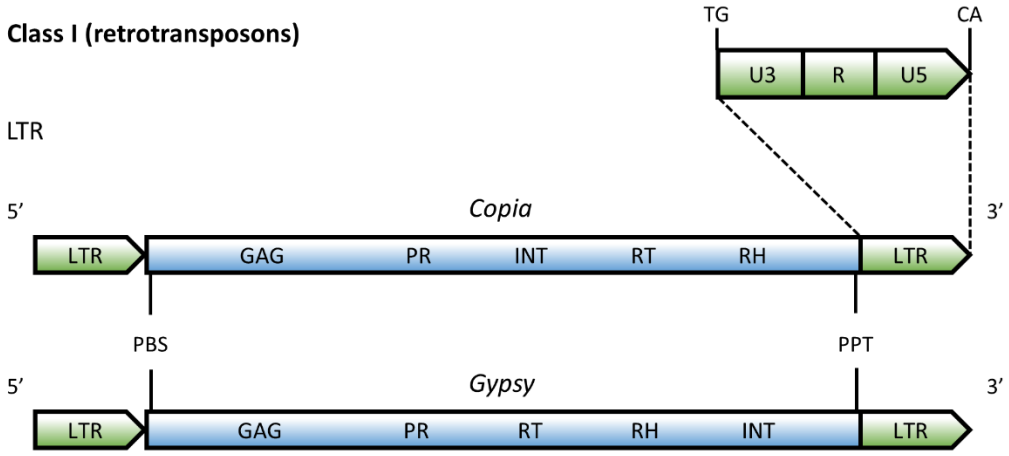
Long terminal repeat (LTR) retrotransposons are the predominant group of TEs in plants [28], and they can range in length from a few hundred basepairs (e.g. terminal-inverted repeat elements, which have lost internal coding domains [33]), to average and large LTR elements between approximately 2 kb and 15 kb, which bear partial or complete internal domains [34]. Some of the largest elements known as LARDs (Large retrotransposon derivatives), are a group of retrotransposons with no internal domains and usually large LTRs [35]. From all groups of TEs, LTR retrotransposons are most responsible for the C-value paradox, in which genome size is not correlated with physiological complexity of the organism [36,37]. Specific cases of TE-mediated genome expansion have been reported for species including maize, barley, rice and *Arabidopsis* [38–43].

The two most common superfamilies of LTR retrotransposon are the *copia*-type and the *gypsy*-type which encode proteins named: group-specific antigen (GAG), protease (PR), integrase (INT), reverse transcriptase (RT) and ribonuclease H (RH). In the *copia* elements the order of these domains is GAG-PR-INT-RT-RH, while in *gypsy* elements the order is GAG-PR-RT-RH-INT (**Figure 1.1**). In the process of retrotransposition, the retrotransposon is transcribed

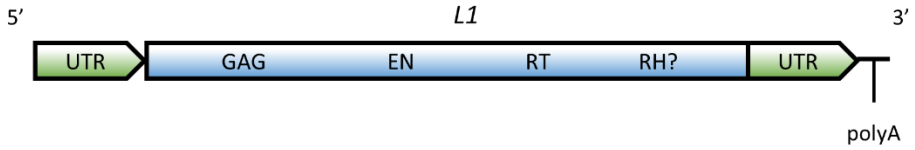
and translated to generate the necessary proteins for retrotranscription and transposition. The protease catalyzes the cleavage of the GAG-POL polyprotein (the polyprotein encompasses PR, INT, RT and RH); the GAG protein creates virus-like particles where proteins and mRNA from the retroelements are transitionally packed, and then the process of reverse transcription is initiated by the RT. The RH degrades the RNA template before synthesizing the double stranded extrachromosomal DNA that will be reintegrated by the action of the integrase protein, which generates staggered cuts in host DNA which are filled upon integration of the element creating target site duplications (TSDs) flanking the TE. The LTRs that encompass the domains can range in size from a few hundred bases in TRIM elements to around 4 kb in LARDS [33,35]. Because of the replication process, LTRs are identical at the time of insertion [44], and they contain three regions: a 3' untranslated region (U3), a repeat region (R) and a 5' untranslated region (U5) (**Figure 1.1**). Transcription initiates at the 5' end of R in the 5' LTR and terminates in the 3' end of R in the 3' LTR; the full retrotransposon element with the extra sections upstream and downstream from the repeat regions is reconstituted during generation of the double stranded extrachromosomal retrotransposon before reinsertion [44]. Additionally the two LTRs usually contain inverted repeats, 5' TG –CA 3', a primer binding site (PBS) and a polypurine tract (PPT), which are essential for DNA first and second strand synthesis respectively [44] (**Figure 1.1**). The LTRs contain promoter-like *cis*-regulatory motifs [45–54], which control transcription of the mobile elements in response to stimuli including microbial extracts, pathogen attack, tissue culture, wounding, polyploidization, and environmental stresses [46,52,55–65]. At the time of insertion, the LTRs at each end of a particular TE are identical [38], thus sequence variations between LTRs can be used as a molecular clock to determine time since insertion (assuming a constant rate of mutation). Although both superfamilies of LTR retrotransposons have often been found associated with genes, most *Gypsy* elements seem to have a close association with heterochromatic regions, while *Copia* elements generate a more random pattern of insertion and associate as well with gene-coding regions [66]. Non-LTR retrotransposons are constituted by LINEs (Long interspersed repetitive elements) and SINEs (Short interspersed repetitive elements). LINEs are several kilobases in length and have similar domains to LTR retrotransposons (**Figure 1.1**), but lack a protease and have an endonuclease instead of an integrase for reintegration of the retroelement into a new

Class I (retrotransposons)

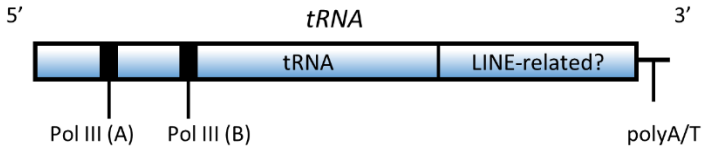
LTR



LINE



SINE



Class II (DNA transposons)

TIR

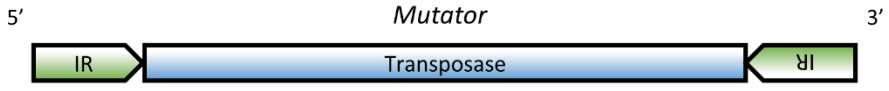


Figure 1.1 Structure of main plant transposable elements. LTR retrotransposons have long terminal repeats (LTRs) at their ends, which have the same sequence at the time of insertion. The LTR itself is divided into three sections: a 3' untranslated region (U3), a repeat region (R), and a 5' untranslated region (U5). Additionally, LTR retrotransposons have five domains coding for proteins: Group-specific Antigen (GAG), Protease (PR), Integrase (IN), Reverse transcriptase (RT), Ribonuclease H (RH). Non-LTR retrotransposons are classified as Long INterspersed repetitive Elements (LINEs) and Short INterspersed repetitive Elements (SINEs). The former also possesses GAG RT and sometimes RH domains, plus an Endonuclease (EN) and a poly A tract. SINEs are non-autonomous elements and having no coding domains usually transpose using enzymes from LINE elements. The have two RNA polymerase III binding sites a tRNA derived structure and sometimes a LINE related sequence, plus a poly A/T tract. Finally, a Terminal Inverted Repeat (TIR) DNA transposon shows two regions of inverted repeats on the ends (TIRs) and a central transposase enzyme. One superfamily of each group is depicted here: *Copia*, *Gypsy*, *L1*, *tRNA* and *Mutator*.

location [32,67]. They have untranslated regions that contain promoter activity and usually a poly A tail on their 3' end, but many elements are truncated on their 5' regions due their mechanism of transposition [67–69]. Upon integration, they also create staggered cuts in DNA which result in TSDs, but their different levels of truncation makes it difficult to localize them sometimes. LINEs are very abundant in mammalian genomes but seem rare or less investigated in plants [32]. SINEs are classified as retrotransposons not because they have a common origin, or have been created as deletion derivatives of other TEs but because their mechanism of transposition uses the enzymes of other retroelements, making them non-autonomous. They are usually generated as pseudogenes of RNA polymerase III transcripts (e.g. tRNA) and have internal Pol III motifs allowing them to be expressed [70]. The origin of their 3' sections is unclear but sometimes they contain LINE-like sequences [32] (**Figure 1.1**). These elements are also abundant in mammals (e.g. *Alu*), they are usually short (150-200 bp), and can also create TSDs [70]. LINEs and SINEs have not been studied in depth in plants, but a few of their characterized insertions were found to be closely associated with genes [67]. Class I elements in plants also have representatives of: i) *DIRS* TEs, which have GAG, PR, RT, RH and a tyrosine recombinase instead of INT, and are flanked by TIRs instead of LTRs; and ii) PLE elements, which have an endonuclease and a retrotranscriptase (more similar to a telomerase) and are flanked by LTR-like termini [32].

Class II DNA transposons are found in all organisms and are the major component of TE in prokaryotic genomes [71]. They range from a few hundred bases to 10-20 kb [32,71]. DNA transposons are divided into two subclasses. The first and most diverse contains six superfamilies represented in plants (*Tc1-Mariner, hAT, Mutator, P, Pif-Harbinger, CACTA*), all of which have a transposase domain and TIRs (**Figure 1.1**) [32]. The transposase enzyme recognizes the TIRs at both sites, creating cuts and then reintegrating in a novel site, where TSDs are generated.

Class II elements (especially non-autonomous) are often associated with gene regions [36,72]. For example, the most famous MITE (Miniature inverted repeat transposable element) elements (*Tourist* and *Stowaway*), which have no coding domains, were found to be closely associated with numerous genes, usually in their controlling regions or introns [73,74]. Elements derived from *Mutator* TEs, called MULEs (*Mutator*-like elements), are able to carry cellular genes, which is an important mechanism for gene evolution [75]. *Mutator* (*Mu*) elements are probably the most active TE-derived mutagenic system implemented in plants, due to the ability of these TEs to insert in gene regions [76], and most of the mutations caused by this superfamily are in fact caused by non-autonomous elements which have TIRs that have captured non-TE sequences.

Subclass II of DNA transposons comprises TEs that do not transpose via a cut and paste mechanism, but replicate without the need of causing a double strand break, and instead use displacement of one strand to initiate replication [32]. The two groups belonging to this subclass are *Helitrons* and *Mavericks. Helitrons* encode a tyrosine recombinase with a helicase domain, replication protein and have non-coding flaking regions. *Mavericks* encode an integrase, a domain for a packing ATPase, a cysteine protease and a DNA polymerase B. Between the ATPase and the cysteine protease they can have an additional ORF and the full elements are flanked by TIRs.

1.2.2 Stress activation of Ty1-*copia* elements

Ty1-*copia* plant retrotransposons can be activated upon exposure to stresses, including tissue culture, wounding, microbial elicitors and pathogen attack [56–58,62,64,77–79]. The potential of TEs to react to stress conditions and cause bursts of transposition, and the high rate of mutation and recombination associated with them, makes them important factors in diversification and potentially in speciation. While TEs may not themselves cause speciation,

they may change their abundance substantially after species have diverged, resulting in increased genomic polymorphism between species [41,42,80,81]. These effects can be seen for example following polyploidization, which triggers TE mobilization, and can result in expansion, reduction or rearrangement of TEs [61,82–84]. Polyploidization can also result in large epigenetic reprogramming [85–87], and probably accounts at least in part for activation and mobilization of many transposons, when the methylation transcriptional inhibition is lifted.

1.2.2.1 Polyploidization

Studies on polyploidization, which is an example of a genomic shock, have shown generational TE changes that result in genome restructuring and diversification. For example, a study of 17 putatively active LTR retrotransposons, which included both Ty3-*gypsy* and Ty1-*copia* elements, was conducted to assess the impact of polyploidization in *Aegilops* species. The results showed species-specific and TE-specific restructuring, and novel insertions after the polyploidization event, mainly dominated by families with evidence of recent activity [82]. A similar result was found when studying the *Tnt1 copia* element from tobacco in synthetic polyploids, which demonstrated transposition of younger elements in the allopolyploid progeny, resulting in local restructuring around insertion sites that included insertion-deletion events [61]. These *copia*-type young families are probably good starting points to study bursts of transposition upon other stresses, and to investigate potential epigenetic effects on numerous genes at the same time. Activation of TEs during polyploidization itself requires changes in the methylation status that can lift the transcriptional repression of the mobile elements which are often silenced under normal conditions [85,86]. For example, in a study on short-term hybridization of *Solanum* species, *copia* elements *Tnt1* and *Tto1* were detected in the hybrids, and the regions related to these elements became hypomethylated, which is congruent with their activation [88]. The TE-related evolutionary influence of polyploidy does not lie in bursts that increase copy number and augment genome size, but on the potential of the elements to generate alternate patterns of expression of surrounding genes, epigenetic reprogramming and points of recombination at insertion locations that can result in genome restructuring [89].

1.2.2.2 Other stresses

Ty1-*copia* elements respond to numerous external elicitors. The first Ty1-*copia* element (*Tnt1*) for which activity was proven was studied in tobacco [90]. Subsequent studies showed how different stress conditions including tissue culture, pathogens, pathogen elicitors, compounds related to plant defense, wounding, freezing and other abiotic stresses, were all elicitors that could activate this retrotransposon [45,48,55,58,62,64,77]. Another two *copia* retrotransposons that have been frequently studied because of their inducibility by stress conditions are *Tto1* from tobacco and *Tos17* from rice. Both of these TEs are activated by elicitors including tissue culture, wounding, methyl jasmonate and fungal elicitors [56,57,65,91,92]. Other Ty1-*copia* elements have also responded to diverse stress conditions. For example, in wheat *Td1a* is activated by light and stress, giving rise to new insertions [53,54]. The oat genome carries at least 10000 copies of *OARE-1* which can be induced by wounding, jasmonic and salicylic acid and by UV light [79]; in melon the TE *Reme1* is transcriptionally induced by UV light too [93]. In *A. thaliana* and other members of the Brassicaceae family, the *copia* retrotransposon *ONSEN* is transcribed upon heat stress and seems to be tightly controlled by siRNAs [46,94–96]; and an element with similarity to *ONSEN* in *Gossypium barbadense* named *GBRE-1* is also responsive to heat [97]. In *Solanum chilense* the promoter of the Ty1-*copia* retrotransposon *TLC1.1* seems to mediate responses to diverse signalling molecules (e.g. ethylene), salt stress and wounding, showing the TE is transcriptionally active [49,52]; likewise in strawberry, the Ty1-*copia FaRE1* has a promoter which can be activated by hormonal treatments [47,98].

The effects that can potentially be produced by transposition of these elements are not different from the ones produced by genome-wide reprogramming upon polyploidization. However, depending on the stress a smaller population or even a few TEs may be mobilized and therefore the possibility of generating adaptive evolution depends on the insertion site, and on the likelihood that the TE can be co-opted and fixed in the long term.

1.2.3 Ty1-*copia* gene regulation and genome restructuring

1.2.3.1 Control of TE activation: Are LTRs captured or capturers?

As seen from the examples above, TEs can be activated by biotic and abiotic stresses. The *cis*-acting elements embedded in the LTR sequences of retrotransposons contain motifs

resembling transcription factor binding sites and controlling regions that are also found in stress responsive genes. This immediately suggests either convergent evolution, or the co-option of the regulatory sequences either by the TEs or by the stress-responsive genes.

Studies of the function of LTRs as promoters two decades ago started highlighting the presence of these motifs in the response of TEs to elicitors. For example the examination of the transcriptionally active *Tnt1* from tobacco showed protoplast-specific activation sequences in the LTRs [99], and LTR-GUS fusions were regulated by microbial elicitors [55], and abiotic stresses [48]. Several other promoters have been examined since, for the presence of regulatory factors responding to different elicitors. For example, a 13-bp motif in the LTR of another tobacco *copia*-type element (*Tto1*), allows the element to respond to tissue culture, wounding, fungal elicitors and methyl jasmonate; inside this motif there are sequence sub-motifs that can be found in the promoters of phenylpropanoid biosynthetic genes, and can be controlled by transcription factors (e.g. *MYB*) involved in defense responses [50,92]. Likewise, the examination of the *TLC1.1* LTR showed specific ethylene responsive elements (PERE boxes - ATTTCAAA) [52], and other *cis*-elements that are involved in responses to methyl jasmonate, salicylic acid, abscisic acid, auxin and hydrogen peroxide [49]; such controlling elements are common to plant defense genes [49]. The transcription factor *ERF1* belonging to the EREPB/AP2 (ethylene-responsive element binding protein/Apetala2) group, modulates diverse responses in a similar way to *TLC1.1* [49], and is a key factor in hormone-mediated plant defense and also abiotic responses [100,101]. Interestingly *FaRE1* from strawberry also shows a response to hormonal treatment (auxin and ABA), with specific motifs characterized as responsive to these hormones, and including as well, the ethylene response element seen in *TLC1.1* with one ambiguity (AWTTCAAA) [47,98]. In Durum wheat, the LTR from the Ty1-*copia* retrotransposon *Ttd1a* contains a CAAT box which binds nuclear proteins in response to light and salt stress [53]. Finally, heat can also activate *copia* TEs; *ONSEN* a TE from *A. thaliana*, has a specific heat response element (HRE) which is bound by heat shock factors (HSFs) that are necessary for the activation of this retrotransposon [46].

Variation in the stress responsive motifs embedded in the LTRs happens even within TE families. For example, in the subfamilies of *Tnt1 copia* elements, the responses become a result of the variation in their LTR regions [45]. It seems therefore that TE LTRs, as happens with stress responsive gene promoters, have motifs that respond to different types of stresses and are

TE specific. For example analysis of *gypsy* and *copia* LTRs from sunflower, showed stress response elements that can be bound by MYB, MYC, WRKY transcription factors, but also other transcription factor binding sites of constitutive genes including Dof elements, and others that were light-responsive and tissue specific [102]. As a result, transcription of specific TEs may respond to specific combinations of stresses. The fact that retrotransposons have multiple stress-response elements in their LTR regions, and that a large percentage of plant genes and plant gene promoters carry TE-derived fragments [103], supports a complex evolutionary history of movement through the genome, where TEs become carriers of new regulatory units for themselves and for genes.

The stress inducible co-regulation of some TEs and some protein coding genes raises the question of how these mechanisms arose. In the case of the *copia* retrotransposon *ONSEN,* it was proposed that a heat response element (HRE) was recruited by the TE, allowing the TE to use the heat response machinery of the plant to become active [46]. *ONSEN* usually inserts nearby genes and can confer heat-mediated activation to such genes [94], which would support the view that normal gene promoters co-opt sequences from *ONSEN* LTRs for their own regulation. Furthermore, its insertion pattern close to genes seems to be widespread among Brassicaceae [95]. It is still left to investigate whether this pattern can be confirmed outside the Brassicaceae, since HREs are controlled by widespread heat shock transcription factors (HSFs) across many plant species [104], and are themselves short sequences (consensus sequence nTTCnnGAAn [46]) which could be easily produced by mutation in plant promoter regions. The question remains if *ONSEN* regulatory motifs were initially co-opted by genes instead of the TE recruiting the heat response and then expanding it to other genes.

Another example on how LTR TE regions can be co-opted is illustrated by the *Tcs1 copia* element which is inserted upstream of an MYB transcription factor named *Ruby* in blood oranges. The transcripts of the *Ruby* gene, which regulates anthocyanin biosynthesis, were shown to be controlled by the LTRs of this TE inserted upstream of the transcription factor [105]. Both complete and solo LTRs have been found in this region of the MYB gene, showing how TEs can gradually mutate and degrade, and how some TE insertions can become fixed as regulatory regions of normal genes. The next step in this evolutionary process of acquisition of TE sections by normal genes is shown by the promoter of an asparagus wound-inducible defense gene (*AoPR1*) [106] which contains several regions of high similarity to the complementary

sequence of the *Tto1* LTR promoter [92], including the 13-bp conserved region of this element (described above). While outside these conserved regions the identity of the original LTR has been completely lost, the process suggests that this *copia* element was inserted ancestrally in the promoter region of the gene, and only the remnants, which the gene uses as controlling motifs for its response, have persisted. These examples demonstrate how TE insertions can decay over time, but their controlling regions remain and become part of basic controlling mechanisms of host genes.

An additional example of regulation of host genes by Ty1-*copia* elements is the insertion of the retrotransposon *Hopscotch* in the regulatory region (over 50 kb away) of the maize domestication gene known as *teosinte branched1 (tb1)*, which accounts in part for the transformation of the maize progenitor (teosinte) into the plant we know today with strong apical dominance [107]. The insertion of the TE acts as an enhancer of a gene that favors apical growth by repressing branching. Such enhancing function by inserting upstream of genes, is also seen in some members of specific families of TEs in maize (including *copia*-type TEs like *raider* or *ubel*), which apparently upregulate gene expression in response to abiotic stresses [60]. Likewise a rice blast resistance gene allele (*Pit*) has the Ty1-*copia Renovator* element inserted in its promoter, and the 3' region of the TE enhances the expression of the gene in the resistant cultivar [108].

The examples above support the evolution of *copia*-derived sequences as controlling regions of normal genes, and many of their characteristics favor this interaction: i) the random insertion pattern of *copia* elements [66] allows them to insert close to genes, becoming an integral part of their promoters (**Figure 1.2**), or picking up motifs that become integral part of the LTR and can be further moved to new locations; ii) even when all the internal regions of a retrotransposon are lost by non-homologous recombination between two LTRs of a single element [109], the solo LTR is still available to be co-opted as a gene regulator; iii) LTRs themselves have multiple promoter features and in organisms like plants where ploidy varies and there are numerous paralogous genes, the insertion of an LTR gives the potential for gene control diversification; iv) rate of mutations in TEs is usually higher than in normal plant genes [110], which would also rapidly generate new motifs for gene regulation; v) retrotransposons are the most abundant type of TEs in plant genomes [28], and can reach over 80% of genome coverage in plants like maize [31], and therefore have a high chance of interacting with genes.

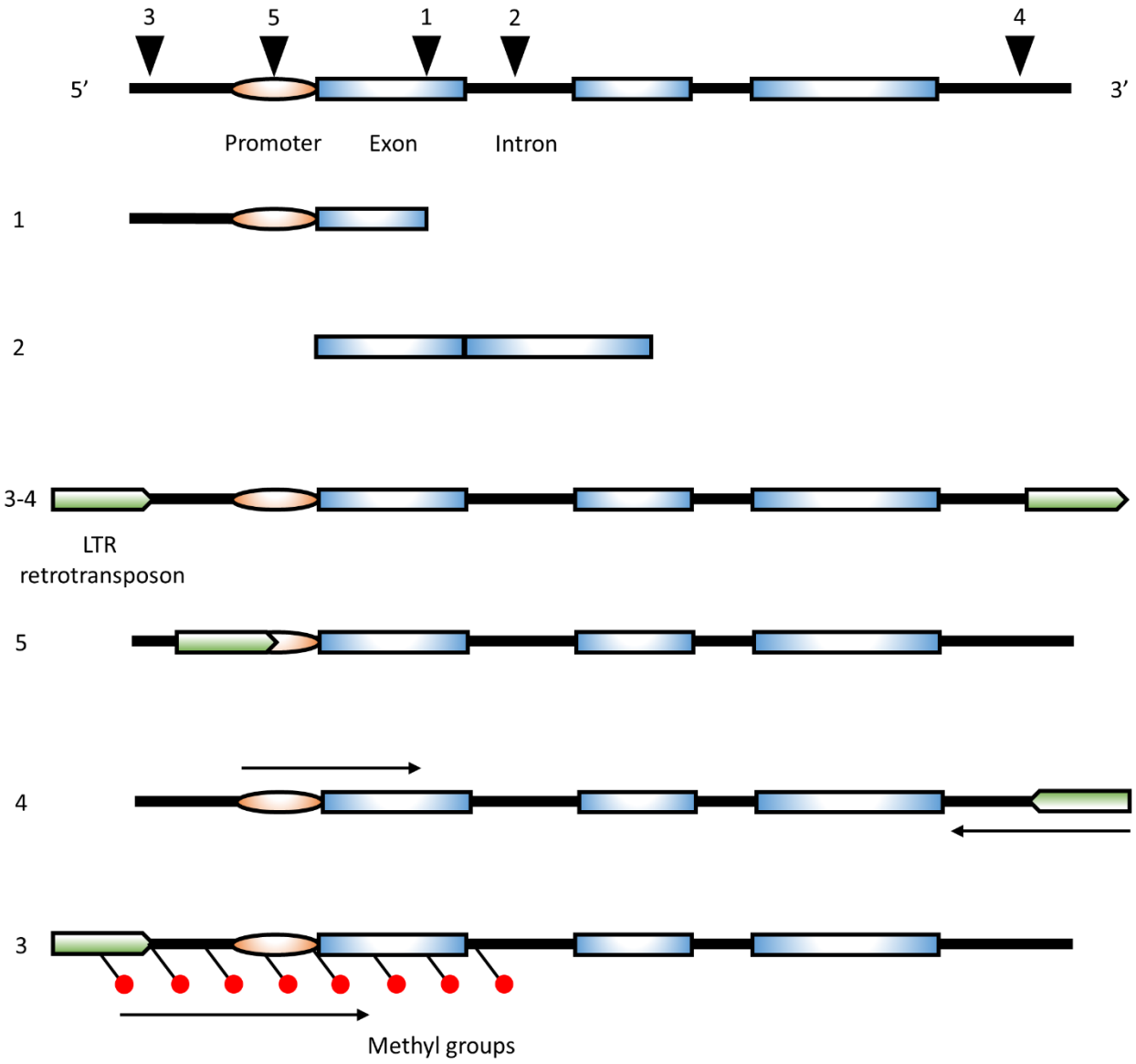

3
5
1
2
4
5’
3’
Promoter
Exon
Intron
1
2
3-4
LTR
retrotransposon
5
4
3
Methyl groups

Figure 1.2 Regulatory changes of LTR retrotransposons when inserting close to genes. The triangles indicate possible points of insertion of the TEs. When TE inserts in point 1 corresponding to an exon, the most probable outcome is gene disruption and premature termination of the transcripts, which in most occasions renders a non-functional protein. When TE inserts in point 2 corresponding to an intron, alternative splice forms can be generated, one of which can be the result of exon skipping. Simultaneous insertion of a similar TE in regions flanking a gene (insertions in 3 and 4), can result in large mRNA starting in one TE, going through the gene, and ending in the other TE. These read-through transcripts would start on the 3' LTR of one TE and find termination signals on the 5' LTR of the other TE. However, the insertion of solo LTRs, or the transcription of complete elements on both sides, or one side along with the internal gene (or genes) is also possible. Insertion in point 5 can result in alternative control of the gene by the cis-acting sequences of the LTR from the retrotransposon. These insertions do not necessarily have to be on the promoter region of the host gene, since TE-mediated controlling effects can happen from tens of kilobases away. When an insertion happens in point 4 in opposite transcription orientation of the gene, opposite read-through transcripts from the TE that reach the gene anneal in antisense orientation to the normal gene transcripts generating a double RNA molecule which can be used to initiate small RNA synthesis which in turn can be used to tag the gene for silencing. Under some circumstances, the methylation used to silence TEs, can be extended to the flanking regions, thus silencing nearby genes as when insertion happens in point 3.

As with exaptation of TE domains/genes, tracing the events of co-option of LTR regulatory regions requires an in-depth analysis of the TE fragments in gene promoter regions, and the establishment of the characteristics that make co-opted TE sequences different from conventional TEs. First, since co-opted sequences should have adaptive traits, they usually should not be found silenced in the genome. Second, mutagenesis could be used to demonstrate their role in gene regulation, similarly to when experiments are performed to find the function of other genes. However, due to the fast mutation rate of TEs [110], it is sometimes difficult to determine whether TE regulatory motifs have been co-opted since only their useful controlling motifs might remain, without trace of the rest of the TE (see example for gene *AoPR1* above), but if TEs insert in multiple gene family members and follow different selective and degeneration processes, the trace motifs would then be easier to find. The ultimate proof of co-option of TE-derived control is however only seen in the long term, when insertions become fixed and provide fitness to the population. A review on exaptation of TEs into cis-regulatory elements defined four levels of experimental evidence [111]: on the first level are the experiments that show the biochemical basis (binding of a TF to an LTR); on the second level

are experiments that show that the presence of the TF binding sites in the LTR can alter gene expression; the third level comprises changes in physiology and anatomy as a result of the insertion (e.g. the changes in maize branching pattern caused by the TE *Hopscotch* [107]); and the fourth level would involve establishing a relationship between the TE insertion with reproductive success, and therefore fitness.

1.2.3.2 Gene disruption and epigenetic control: changing gene functions.

The previous section demonstrated that TEs can influence the regulatory function of normal genes via incorporation of LTR sections as part of their promoters. But the regulatory power of TEs goes beyond the acquisition of promoter-like sequences by genes. TE insertions in introns can cause alteration in splicing patterns [112] (**Figure 1.2**), and exon insertions can be directly disruptive of gene function (**Figure 1.2**), but also, in some cases, be co-opted to generate new functions [26]. Likewise, the orientation and distance of the TE from genes can have an effect on the production of new genes, as well as on variation in gene expression. Read-through transcripts that reach the gene can result in chimeric gene/TE products [113], but also create anti-sense sequences that can repress the genes [114,115] (**Figure 1.2**). Furthermore, TEs are commonly methylated via small RNA targeting, and this methylation can be extended to surrounding genes, rendering them inactive [116,117] (**Figure 1.2**).

The classic example of a retrotransposon altering splicing patterns upon insertion into intronic regions came from the study of alleles of the maize *waxy* gene responsible for amylose biosynthesis. The presence of the LTR retrotransposon within the gene alters the recognition of the normal splicing sites, creating alternate patterns where exons surrounding the TE insertion can be spliced, and gene expression can be altered [112,118]. Another example where the insertion of a *copia* retrotransposon alters the transcripts is shown by a TE inserted in intron 1 of the recognition of *Perenospora parasitica* resistance gene *RPP7* in *A. thaliana* [119]. This interruption results in the generation of transcripts with an alternative polyadenylation site in the LTR, and these alternate transcripts are increased when specific histone marks tagged to the TE are repressed in mutants, demonstrating that epigenetic regulation of the inserted TE is important to produce normal gene transcripts. TE intron insertions can also abolish gene expression as was demonstrated for a MADS-box transcription factor bearing a LTR retrotransposon in introns 4 and 5 of two apple varieties, resulting in seedless phenotypes [120]. An opposite effect was seen

when a Ty1-*copia* insertion in an intron generated longer primary transcripts and potentially higher and ubiquitous transcription, for an alternate oxidase (*Aox*) gene in a specific grape cultivar [121]. Interestingly, this TE was found to be inserted in introns of at least 20 more genes, highlighting their potential regulatory power and raising questions concerning the mechanisms of potential bias for intronic insertions for this element. In fact, a study on alternative splicing in one of the cotton progenitors (*Gossypium raimondii*), noted that TEs were present in a large percentage (43%) of the introns retained in alternative spliced forms of genes, and that this phenomenon was fairly common among other plant species with similar genome sizes [122]. If this is true in most plants, and methylation status influences the production of alternate transcripts, TEs strongly regulate the production of normal gene transcripts and create new transcripts that can be tested for novel functions.

TE insertions in exons are expected to produce detrimental effects on genes, since they directly affect the reading frame, however if an insertion becomes lethal it would not be evident in natural populations, since the host would not survive. Therefore, if TE exonic insertions can be discovered, either they do not completely disrupt gene function, or they are found in genes that are not required for the survival of the organism (e.g. non-essential genes, haplosufficient genes, or functionally redundant paralogs). In fact, the first discovered tobacco TE was a Ty1-*copia* element inserted in the open reading frame of a nitrate reductase (NR) gene [90] and further experiments with insertions of this element in different exons of NR produced chimeric gene-TE products, which either failed to splice introns or had early termination signals (**Figure 1.2**), resulting in truncated chimeras [123]. However, the changes in transcription or the impairment of the gene did not necessarily result in undesirable phenotypic traits. In soybean, the insertion of a Ty1-*copia* element in exon 1 of one phytochrome A paralog, results in a stop codon after the insertion that produces a truncated protein conditioning insensitivity to long day flowering which is a case of adaptive evolution [124,125]. In glutinous rice the insertion of Ty3-*gypsy*, produces altered non-functional transcripts in granule-bound starch synthase, causing the desirable, glutinous rice phenotype [126].

Transcription that is initiated within a retrotransposon can sometimes continue past its normal termination sequences and continue into the flanking genomic sequence. In some circumstances, read-through transcription can result in transduction of the flanking genomic sequence, meaning that one or more genes are captured and moved to new genomic locations

(**Figure 1.2**). One of the first such examples of read-through from Ty1-*copia* elements into genes was found by studying the *Bs1* element of maize. While this element had features of an LTR retrotransposon, part of its sequence revealed similarity to a proton-translocating ATPase [113], showing it was the product of read-through transduction. These events caused exon shuffling and produced a novel, *Bs1*-derived gene that was normally transcribed and translated at a specific stage of reproductive development [127]. Even more dramatic is the transduction event triggered by the *Rider* Ty1-*copia* element, which read through a region of over 24 kb, moving several genes from chromosome 10 to chromosome's 7 *sun* locus in tomato [128]. The movement not only disrupted one gene in chromosome 7, but seems to be largely responsible for the elongated fruit shape of the Sun1642 variety, due in part to regulation of one of the transduced genes. In rice and sorghum, 1343 and 672 genes have been captured by LTR retrotransposons, and while the mobilized genes can become non-functional pseudogenes, some of them maintain expression and could evolve into new functions [129].

Read-through beyond TE boundaries can also affect epigenetic regulation of the flanking genes depending on the orientation of the TE. Analysis of read-through transcripts of the *Wis2-1A* from the LTR found several sections of genes in the opposite orientation relative to the TE [114], causing silencing of the genes. The study argued that the generation of antisense gene transcripts would result in double stranded RNA which would be processed into small RNAs for post-transcriptional gene silencing [130] (**Figure 1.2**). But gene silencing from nearby TEs does not always involve read-through transcripts. Pioneering studies on the role of TEs in heterochromatin and epigenetic control of genes, demonstrated that small interfering RNAs (siRNAs) tagged TEs for methylation and that this methylation could be extended to genes that were in close proximity of TEs [131] (**Figure 1.2**). Therefore, since silencing of TEs nearby genes would be deleterious for plants, purifying selection against methylated TEs close to genes seems to be a normal mechanism to avoid the harmful effects. In *Arabidopsis thaliana* it was found that a lower level of methylation was associated with TEs that were close to genes, and that methylated TEs close to genes were correlated with lower gene expression [116,132]. This later correlation was then validated in both *Arabidopsis thaliana* and its close relative *Arabidopsis lyrata* [117]; however, the genome of *A. lyrata* is 1.5 times larger than the genome of *A. thaliana*, and has three times the number of TE insertions, which causes more likelihood for TE-gene interaction. Because *A. lyrata* has many more members in each TE family, and new

bursts of transposition generate many identical copies of the same TE, the amount of siRNAs available for these multiple TEs (assuming enzyme supply is limited), makes it harder to silence all the elements and consequently places less epigenetic burden on nearby genes [117], and would allow also for some TEs to remain active, further increasing genome size. Furthermore, in *Arabidopsis thaliana* it was found than retrotransposons were targeted by siRNAs more frequently than DNA transposons, and that on average LTR retrotransposons represented younger insertions than non-LTR retrotransposons, and had a greater repressive effect on flanking genes. As in the experiment that compared the two *Arabidopsis* species, this later study also showed that when siRNAs tag unique TEs, then their repression of both the TE and the neighboring gene is increased [132].

Lately it has also been shown that TE transcripts can be processed into small RNAs that epigenetically control other genes in *trans* [133], which are not necessarily in close proximity of the TEs. One of the first examples of such regulation demonstrated that activation of the Ty3-*gypsy* retrotransposon *Athila* is linked to the production of small RNAs that are recruited to post-transcriptionally and translationally control a RNA-binding protein (*UBP1b*) involved in stress granule formation in *A. thaliana* [134]. After this study, 27 *A. thaliana* candidate genes were identified as possible targets in *trans* of siRNAs generated by TEs [135].

On the other hand, another study showed that intragenic TEs tend to be less methylated than intergenic TEs [136], but some methylation could still be detected on the intragenic TEs. However, in this study, the intragenic methylation marks remained restricted to the TE region, and did not spread to the exons. Instead the genes with intronic methylated TEs had relatively high transcription, and the epigenetic marks might be important for proper transcription of the host genes, promoting the correct splicing of the intron containing the TE [136].

In the previous paragraphs we showed how TEs close or inside genes can have different degrees of methylation and this methylation can be spread to the genes when the TE is nearby the gene or restricted to the TE when the TE falls in an intronic region. In specific cases the mechanisms by which TEs near genes are allowed to remain unmethylated, would also allow the TE to remain active, potentially resulting in additional transcription (possibly including read-through transcripts), and transposition, which would not benefit the host if new insertions disrupted other genes, but could alternatively duplicate genes or promote exon shuffling as raw material for new gene functions. Therefore, the epigenetic regulation of TEs becomes a trade-off

between silencing TEs to stop further transposition, and not silencing TEs close to genes because of potential silencing of adjacent genes, and the possibility of positive transduction events.

As evidenced by these examples, the influence of *copia* elements when inserting nearby or inside genes, is case specific. This is partly related to the variability in genomic context, but also due to the fact that even between close members of a TE family, changes start accumulating after a burst of transposition. Nevertheless, some generalizations can be made. Insertion of TEs in introns will mostly influence alternative splicing or premature termination. Methylation of intron-inserted TEs might also be important for correct splicing, and may help explain the origin of TE-derived introns. Early models proposed that introns originated as vestiges of retrotransposons, based on the fact that TEs had their own splicing mechanism [137], although Class II TEs were also shown to be conducive to producing intronic structures [138]. While there are several possible origins of introns, TEs seem to be an important part of their formation [139,140]. The examples showing how TEs can fall in different introns of the same gene family (e.g. *waxy* and *Aox* [112,121]), and produce different splice forms, demonstrate the ability of TEs to generate variants that can be filtered through natural selection. In the meantime, exon insertions usually result in transcription decrease or depletion, and although some of the characteristics resulting from these insertions may be selected by breeders, it is less likely that they would be naturally preserved. Finally, for TEs that are inserted nearby genes, their impact depends on distance and orientation, in the case of methylation spread or antisense transcripts respectively, but epigenetic regulation in *trans* by TEs far from genes is also possible. The origin of miRNAs in humans can be explained by the generation of the miRNA hairpin in the interface of two consecutive and opposing TEs and read-through transcription [141]. In plants, some miRNAs can be generated from already self-folding miniature inverted-repeat TEs (MITEs) [142], however juxtaposition of two inverted TE copies and inverted repeats of non-autonomous TEs (the ones which have lost the internal coding regions), can also give rise to the necessary hairpin that is processed into micro RNA, and therefore some miRNAs can also be derived from other types of TEs including retrotransposons [143]. The model of evolution of TE derived miRNAs to control host genes depends on the insertion of TEs in transcribed regions so the mRNA necessary for the generation of miRNA would be available. The generation of the stem loops originated from the TE would target the genes with the TE insertion, but since processing of the small RNAs could be imprecise, miRNAs would also be derived from the target gene. TE-

derived miRNAs would not only target one gene (or related family members) like it happens with gene derived miRNAs, but diverse genes associated with similar TE insertions, and miRNAs would increasingly be selected if they confer an advantage, with many of them eventually not showing traces of their TE origin [143].

1.2.3.3 Do plants really need TEs? The case for TEs in resistance gene evolution

LTR TEs can also help in gene evolution through mechanisms of recombination and gene duplication. Recombination is one of the mechanisms by which TE-mediated genome obesity is halted. Formation of solo-LTRs by recombination between LTRs of the same or different TEs eliminates internal domains of the TEs, and large DNA segments in-between retrotransposons [109,144], but TE-mediated recombination can also give rise to diverse genomic rearrangements [145]. Additionally, the ability of TEs to mobilize through the genome and to transduce genes or gene fragments can generate diversity through duplicated regions. Gene carrying capacity has been clearly identified in mutator like elements (Pack-MULEs) [75], but 672 and 1342 genes captured by LTR retrotransposons have also been detected in rice and sorghum respectively [129]. TE-mediated processes of gene movement, duplication and recombination can occur by several mechanisms: i) retroposition of a gene transcript, which involves normal transcription of a gene, but retrotranscription and insertion in a different genomic location using the retrotransposon enzymes (**Figure 1.3A**); ii) readthrough transcripts from LTR retrotransposon inserted close to genes (**Figure 1.3B**), which results in gene transduction [127,128]; alternatively retrotranscribed sequences could be inserted in double strand break (DSB) regions [146]; iii) recombination of TE sections flanking gene regions with TEs in different chromosomal locations (**Figure 1.3C**); iv) readthrough transcription of TE-gene regions, followed by retrotranscription and recombination with homologous regions elsewhere (**Figure 1.3D**); v) Recombination of two LTRs of similar elements flanking gene regions can delete the intervening sequence (**Figure 1.3E**); vi) an insertion in a gene can disrupt the gene or create alternative splicing (see **Figure 1.2**), but also create raw material for new recombination with only parts of a gene (**Figure 1.3F**).

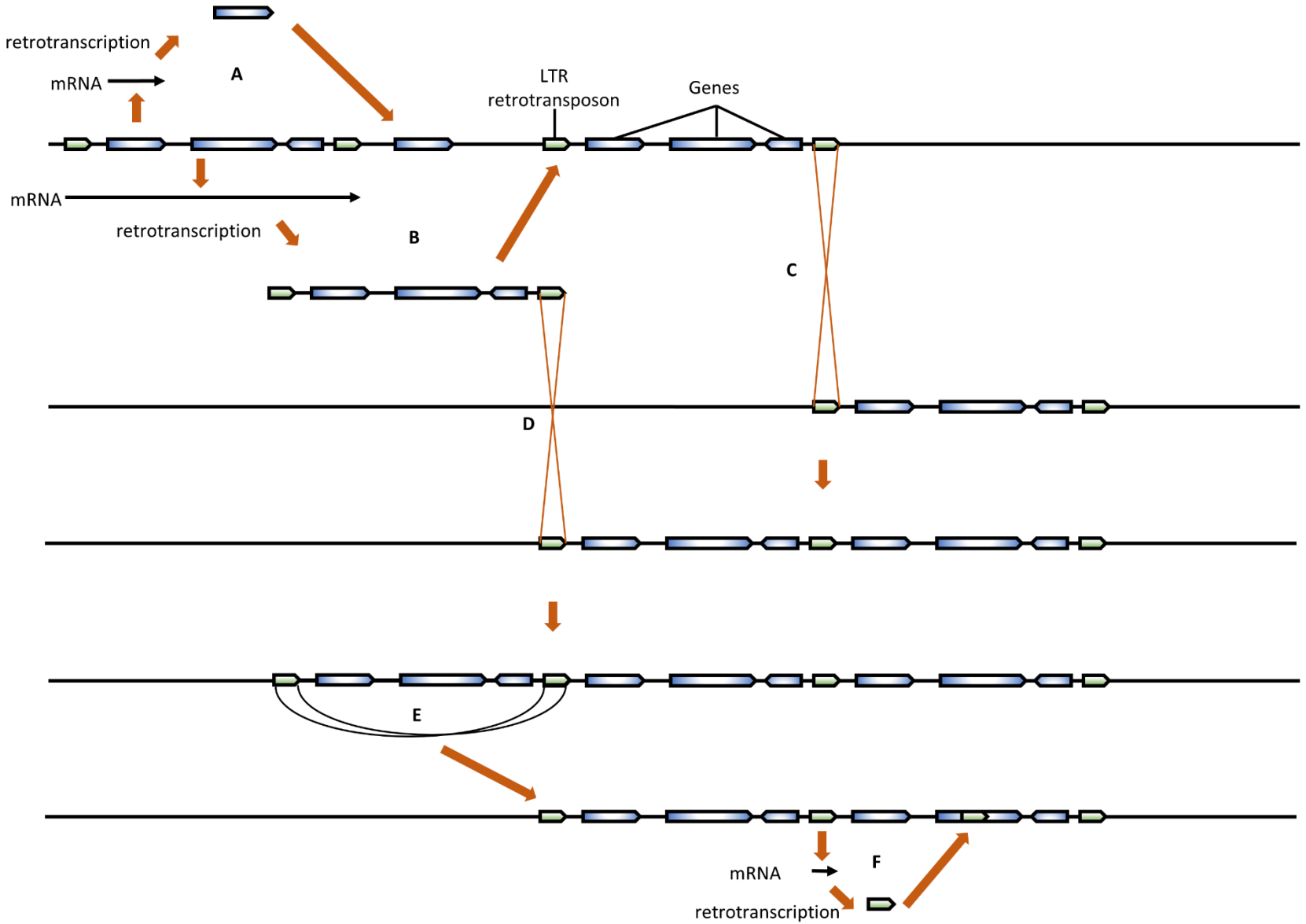

retrotranscription
mRNA
A
LTR
retrotransposon
Genes
mRNA
retrotranscription
B
C
D
E
mRNA
F
retrotranscription

Figure 1.3 Mechanisms of TE-mediated gene movement. A. A normal gene is transcribed, retrotranscribed and reinserted in a different genomic location by the TE enzymes (retroposition). B. A read-through transcription starts in one LTR retrotransposon, covers three normal genes and finishes in another LTR retrotransposon. The large transcript is retrotranscribed and inserted in a new genomic location. C. Non-homologous recombination between two retrotransposons (usually between their LTR sections) in different genomic locations produces a duplication of previously transduced gene sequences. D. Recombination of the double DNA strand produced by retrotranscription with another genomic location results in a similar duplication as in C. E. Non-homologous recombination between two similar TEs (usually through their LTR regions or with solo-LTRs), results in deletion of the genes between the TEs. F. A TE transposes inside a gene and generates a potential new spot for recombination of just specific sections of a gene.

Evolution of many gene families can be impacted from the effects of TE insertion and gene/exon shuffling induced by TEs. One interaction that has been speculated to occur in several species is TE-mediated evolution of resistance genes. Resistance genes (*R*-genes), are part of the machinery of perception and transduction of signals detected upon pathogen attack in plants. Initially *R*-genes were only defined as those involved in gene-for-gene interaction [147], and responsive to fungal specific virulence factors (effector-triggered immunity – ETI), but such distinction has now become blurry and genes involved with general perception of non-specific defense response elicitors (pathogen-associated molecular patterns triggered immunity – PTI) can also be classified in this category [148]. Resistance genes have specificities for the pathogen virulent factors, such that if the effectors can be recognized, then the plant can trigger its defenses. However, pathogens generate mutations in their effectors that allow them to escape detection by plants and favor infection; this occurs until the plant creates new *R*-genes that can detect the new effectors. This arms race in the plant-pathogen interaction implies a need for duplication, recombination and diversification of resistance genes to keep up with the pathogen's arsenal changes. In fact many of these disease resistance genes are found in clusters, where they seem to recombine and evolve rapidly [149]. Taking into account all the possible processes by which TEs can cause gene shuffling, duplication or recombination, it is easy to see how the rate of evolution of clusters of resistance genes enriched with TEs would benefit from the presence of mobile elements (**Figure 1.3**). For example, the leucine-rich repeat (LRR) domain is a repetitive amino acid motif common in some resistance genes. The co-location of TEs and LRR proteins could promote exon shuffling and unequal crossing-over resulting in duplication and expanding

the repertoire of resistance specificities. The human major histocompatibility complex (MHC) – which plays a central role in the acquired immune system - partly relies on the variability generated by retroelements clustered with the immunity proteins [150]; a similar mechanism, developed for plant immunity, would certainly favor the generation of new *R*-genes. In rice, transposable elements have been found to be related to clusters of the *Xa21* disease resistance gene family members [151]; most of these TEs are intergenic but two disrupt resistance genes, with one of them being a retrotransposon with similarity to *copia* elements. The TE insertions in genes result in premature transcript termination, but one insertion generates proteins that resemble tomato fungal resistance genes *Cf2* and *Cf9* [151], showing how even disruptive TEs give rise to novel transcripts with potential adaptive value. Although this study shows recombination events of the region were only related to highly conserved sections with high GC, it is not inconsistent to think TE transduction and recombination events could have or can potentially influence changes in these loci (**Figure 1.3**), since 17 TE sequences are found in the introns and 5' and 3' regions of *Xa21* genes, and the TEs have been active during the evolutionary history of this gene family [152]. In sorghum, the *Pc* locus contains three resistance genes that determine susceptibility in their dominant state to a toxin by the fungus *Periconia cicinata*. Mutations of the central gene of three consecutive paralogues, separated by two almost identical retrotransposons, confer resistance to the plants [153]. Many of the mutations in 13 different mutant lines can be explained by unequal recombination between the paralogue genes, but in one mutant the total deletion of the central gene was explained by the ample similarity of intergenic regions. Although the retrotransposons were not mentioned as the causes of this later event, there is proof that a high accumulation of retrotransposons in specific region results in recombination hotspots [154]. In *A. thaliana*, divergence in haplotypes of the *RPP5* resistance genes clusters depend on mutation, recombination but also retrotransposition [155]. The paralogues of these clusters are an amalgamation of combined gene segments of the *R*-genes largely dependent on duplicated and recombined LRR regions, and while most are non-functional, they may serve as evolving reservoirs of new resistance genes that could interact with novel virulent pathogen factors. The haplotype clusters themselves vary in the number of paralogues, intervening sequence and TEs, and some of the differences may be related to their levels of resistance to the oomycete *Peronospora parasitica* [155].

A combination of TE properties as recombination agents (**Figure 1.3**) and also as epigenetic triggers (**Figure 1.2**) could render them useful in the generation of new resistance genes upon pathogen infection. For example, if a cluster of resistance genes is populated by epigenetically silenced TEs via methylation and the silencing extends to the genes, the methylation could be lifted upon pathogen attack, leaving both the TEs and the resistance genes active. The *R*-genes would then be ready to be used for defense, while the active TEs could now be transcribed and transposed but also maybe transduce regions from the resistance genes providing raw material for recombination, exon shuffling or duplication. In fact, it is know that many defense genes including *R*-genes are under epigenetic repression through small RNA silencing, which is lifted upon pathogen attack [156,157]. An experiment where methylation was artificially reduced in rice, showed that methylation marks disappeared in a retrotransposon and a resistance gene. The genes were different clones and apparently unrelated, and while there was no trace of retrotransposon activation, the lack of methylation in the resistance gene resulted in its constitutive transcription [158]. While the retrotransposon in the later study was demethylated but not activated, a decrease in DNA methylation enzymes in *A. thaliana*, resulted in reduced methylation and increased copy numbers of *gypsy* and *copia* elements in a different study, demonstrating that methylation reduction can change the status of these TEs [136,159]. In *A. thaliana* methylation mutants showed marked resistance against *Pseudomonas syringae*, and decreases in methylation in response to salicylic acid (SA) were largely linked to TEs related to regulatory regions of defense genes. In some cases those genes were shown to be controlled by the methylation status of the associated TE, and were also upregulated when methylation was lifted [160]. These decreases in methylation were also associated with increases in the production of 21 nucleotide small RNAs, which were speculated to be involved in controlling surrounding vegetative cells or reproductive tissues (possibly conferring transgenerational resistance), but could also be regulating other defense genes [160,161]. Likewise, demethylation in *A. thaliana* mutants resulted in increased susceptibility to *Fusarium oxysporum*; the defense genes reacting to the infection were downregulated and shown to be enriched with TEs in their promoter regions [162]. The TEs were apparently the main target of the demethylases since the mutants showed CG hypermethylation on the TEs and surrounding sequences, thus providing an epigenetic mechanism of control of flanking sequences which included the many defense-related genes. Probably one of the clearest examples on how the insertion of a TE exerts epigenetic

control on a resistance gene comes from a co-opted *copia* in *Arabidopsis* [119]. The insertion of the retrotransposon in intron 1 of the *RPP7* resistance gene generates an alternative polyadenylation site. The TE is controlled epigenetically by histone marks, and a low level of methylation in the *copia* region correlates with the alternative transcript production, which cannot produce a functional resistance gene receptor; the normal resistance gene is restored when the TE marks are properly in place, demonstrating the importance of TE-mediated epigenetic control on the host gene. Furthermore, an investigation of methylation defective mutants using public data showed that hypomethylated intronic TEs are related to problems of transcription in the associated genes [136], showing that the condition presented by *RPP7* maybe more widespread than expected and that epigenetic marks in TEs inserted in intronic regions are important for proper transcription.

As shown in the previous sections, the interactions of TEs occur often when the mobile element inserts close or inside the gene, but can also occur from a considerable distance, as in the case of the Ty1-*copia* element *Hopscotch* which exerts its enhancer effect from 50 kb away [107], or when TE-derived small RNAs control distant genes [134]. The large abundance of TEs and especially of retrotransposons in plant genomes necessarily results in their interaction with genes over long evolutionary timescales. While arguments on their selfish character were stronger when TEs were first studied, the vision has changed towards them maybe having a dual character, where some of their behavior can still be characterized as selfish, but their co-option can be useful [21]. If a burst of transposition can be detected with current laboratory techniques, it is very seldom that an immediate effect on genes, the genome or the physiology of the plant, can be noticed, with rare exceptions like in the case of *mPing* TEs, which actively transpose from one generation to the next and can have an effect on gene transcription of adjacent genes [163]. However, not even in this particular case is it yet known if these changes will be fixed, and which ones will be beneficial in the long term. So it can be argued that in recent events of transposition, TEs can have a selfish character, but if changes are fixed in the population then their useful characteristics become evident and provide fitness.

1.2.4 TE-derived markers and their use in diversity and evolution studies

Because of their abundance, diversity, high rate of mutation and transposition characteristics, TEs have been used frequently as markers of intra- and inter-species

differentiation. For this, researchers have relied on the design of TE-based markers that include SSAPs, RBIPs, IRAPs, REMAPs and iPBS [164–169].

1.2.4.1 Sequence-specific amplification polymorphism (SSAP)

SSAPs rely on the amplification between a TE and its flanking region, after restriction digestion of the genomic DNA. The PCR takes place with a radioactively labeled primer that binds a retrotransposon section (usually the LTR), and a primer that binds to an adaptor that has been ligated to the restriction site (**Figure 1.4**) [165]. Such amplicons reveal polymorphisms between samples depending on the transpositional history of the TEs, when run on a high-resolution polyacrylamide gel.

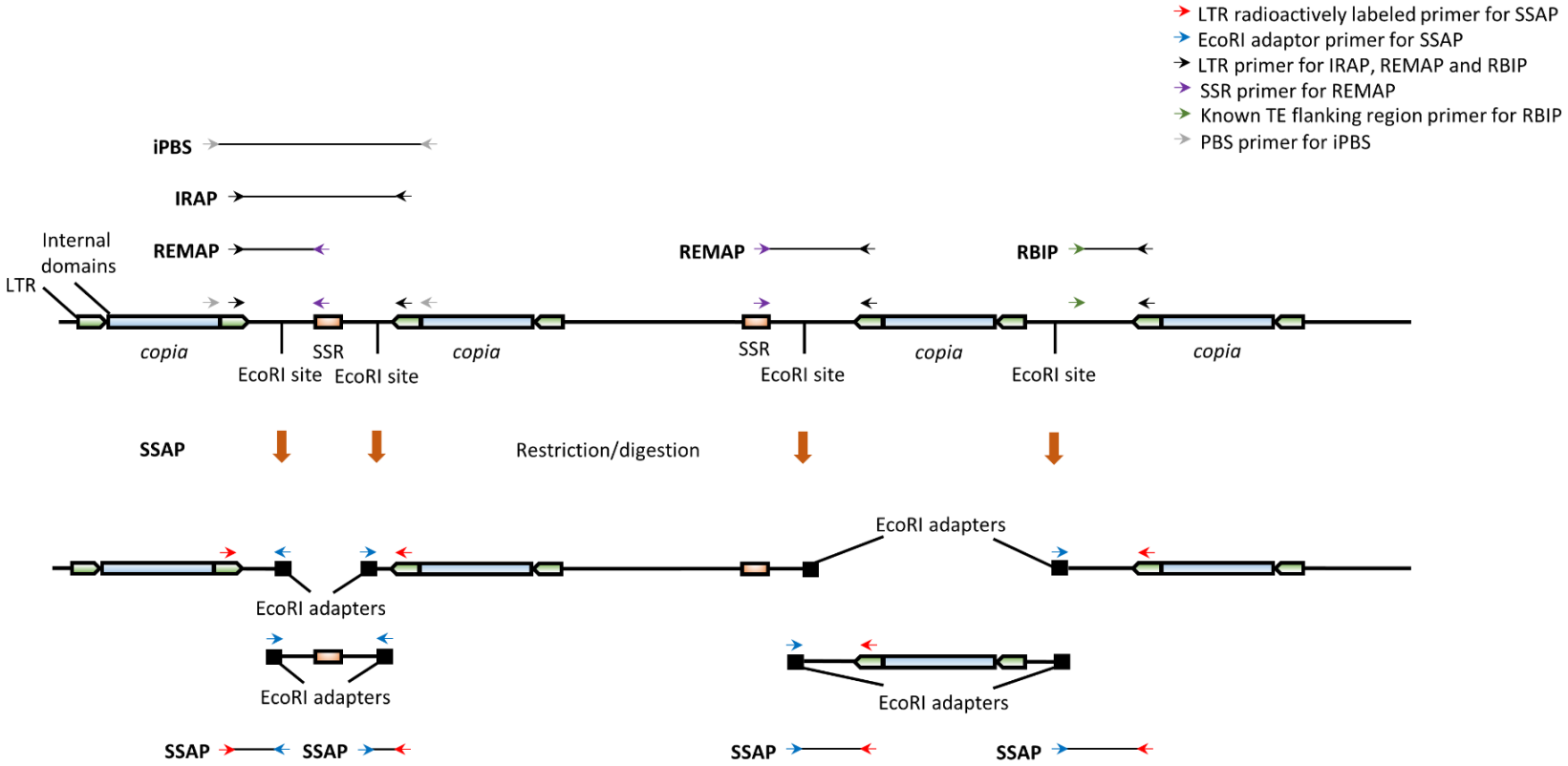

LTR radioactively labeled primer for SSAP
EcoRI adaptor primer for SSAP
LTR primer for IRAP, REMAP and RBIP
SSR primer for REMAP
Known TE flanking region primer for RBIP
PBS primer for iPBS
iPBS
IRAP
REMAP
REMAP
RBIP
Internal domains
LTR
copia
SSR
copia
SSR
copia
copia
EcoRI site
EcoRI site
EcoRI site
EcoRI site
SSAP
Restriction/digestion
EcoRI adapters
EcoRI adapters
EcoRI adapters
EcoRI adapters
SSAP
SSAP
SSAP
SSAP

Figure 1.4 TE-derived molecular marker techniques. In sequence specific amplification polymorphism (SSAP), DNA is digested with a restriction enzyme and adapters are ligated to the restricted fragments; then, DNA is amplified with a radioactively labeled TE-specific primer (usually from the LTR of the retrotransposon), and a primer that binds the adapter sequence. In inter-retrotransposon amplified polymorphism (IRAP), it is only necessary to design one TE-specific primer facing outward; the amplicons are produced when two retrotransposons are in close proximity and in opposite orientation. For retrotransposon-microsatellite amplified polymorphism (REMAP), one primer is designed on the TE and a second on a microsatellite or simple sequence repeat (SSR); when the TE is close to the SSR, then an amplicon is produced. Retrotransposon-based insertion polymorphism (RBIP) uses a TE-derived primer and the knowledge of the sequence surrounding the TE to design a second primer; when the TE is present then an amplicon is produced. If no amplicon is produced, further confirmation of the absence of the TE insertion is performed using a second primer on the other edge outside of the TE, which will produce another amplicon of expected size for a TE empty site (not shown). For inter-primer binding site (iPBS), the tRNA site used for priming retrotranscription is used to design a primer which is common to most LTR-retrotransposons. Similarly to IRAP, the intervening sequence between two PBS of two retrotransposon placed in opposite orientation will be amplified. The amplicons produced by each technique, have the respective technique name besides them in the diagram.

Possibly one of the most exploited TEs for diversity studies is the *copia*-like *BARE-1*, a high copy number and widespread retrotransposon originally isolated from barley [170,171]. An additional advantage of studying this TE is its active transposition [51,172,173], which potentially allows for the creation of natural variation even within closely related organisms (e.g. individuals from the same species). Since breeding searches for new allelic combinations, the natural variation created by an abundant TE like *BARE-1* allows for a selfing crop like barley, to partialy escape high genetic homogeneity [174], and may constitute additional desirable characteristics in breeding strategies. SSAPs were originally tested in two barley genotypes with *BARE-1*, revealing higher levels of polymorphism than amplified fragment length polymorphisms (AFLPs) [164]. Later, the same technique was applied to study genetic diversity of 103 barley cultivars [174]. The 150 polymorphic bands indicated variation among cultivars that followed distinct phenotypic characteristic, but also intra-cultivar variation showing that retrotransposons are useful markers of diversity and can be further used to test variability between individuals [174]. *BARE-1* also depicts how environmental factors may affect transposition rates [40], an idea that follows from McClintock's concept of genomic shock. The

study of *BARE-1* from wild barley in a sharp microclimatic gradient showed how the number of elements increased with altitude and drought stress on the plants [40]. LTRs from *BARE-1* contain abscisic acid responsive elements (ABRE) that allow this TE to be responsive to changes in water status. Additional studies of *BARE-1* copy number variation in barley and wheat, which allow to infer both activity and polymorphism, have been performed using quantitative PCR [173,175].

One of the first retrotransposons characterized (*Tnt1*) has also been used to study diversity. The study of accessions of tobacco and its progenitors using four populations of *Tnt copia*-like elements demonstrated separation of accessions and species, but TE-induced diversification of tobacco after its creation seemed scarce [176]. Another SSAP study with a *Capsicum annuum* and a set of related species demonstrated that the markers were able to clearly resolve species and cultivars, and were even congruent with geographic distributions [81].

SSAPs with Ty1-*copia* TEs have also been successfully used as a comparator to AFLPs and other molecular markers in cashew, artichoke, citrus, tomato and pepper [177–180], where the SSAPs demonstrated parallel or sometimes improved levels of polymorphisms; in strawberry and *Vicia* species they were used to discern polymorphisms between accessions [181–183], in blue agave, for investigating phylogenetic relationships among species and varieties [184], and in Asian pears, SSAPs allowed the detection of hybridization events [185].

1.2.4.2 Inter-retrotransposon amplified polymorphism (IRAP) and retrotransposon-microsatellite amplified polymorphism (REMAP)

Two other techniques that rely on unknown sites of TE insertions are IRAP and REMAP. In the former, primers are designed facing outward from the retrotransposons (in the LTR) to capture length polymorphisms from a region between two TEs situated close enough to generate an amplicon (**Figure 1.4**). In REMAP the same strategy is used but the amplifications are performed with a primer from the retrotransposon and another primer anchored on a simple sequence repeat (SSR) or microsatellite (**Figure 1.4**). While theoretically this technique could also yield SSR-SSR or IRAP products, this seldom happens [167].

IRAPs with primers from barley *copia* elements and other TEs, were successfully used to distinguish banana cultivars, and to follow the ancestry of their hybridization events [186]. Likewise in rice, *Tos17* derived primers were used with IRAP and REMAP to compare 51

accessions and assess the usefulness of these techniques in breeding strategies [187]. In sunflower, scoring of both cultivated and wild accessions of 39 species of *Helianthus* using IRAP with *copia*-like elements and other unclassified retrotransposons, demonstrated a decrease in the number of TE-derived polymorphic loci in domesticated accessions, but larger diversity in wild accessions and species [188]. In this latter case, not only was the IRAP technique useful for assessing diversity, but the selected retroelements were sufficiently well-conserved that they could be used across species, which is uncommon due to high mutation and recombination rates among TEs. This is also attributed to a potential recent divergence of this genus [188]. However, the number of unique bands in *Helianthus* species was so high that phylogenetic relationships could not be resolved from these markers due to lack of shared characters. This suggests a rapid turnover of TEs and active transposition influencing the divergence of the species. In a different study, unclassified LTR-retrotransposons were used for IRAP in accessions of flax (*Linum usitatissimum*) and other species from the genus, demonstrating a decrease in diversity in breeding lines and cultivars as compared to landraces. The technique was unable to draw a clear distinction between the main flax types (fiber and linseed), but showed a good agreement with the phylogeny for the species in the genus [189]. In eucalyptus species, the use of REMAP and IRAP with both *copia* and *gypsy* retrotransposons, showed more fragments and polymorphic bands in *Eucalyptus grandis*, indicating recent activity of most families in this species; furthermore, the most recently inserted families fell closer to gene-rich regions. However, a dendrogram of relationships of the different species was not in full agreement to phylogenetic relationships generated by other markers (e.g. DArT markers) [190]. Likewise, the use of IRAP markers from a combination of *copia* and *gypsy* elements in *Lilium* species, showed phylogenetic relationships that did not fully agree with previous reports using internal transcribed spacers (ITS) [191].

While retrotransposon-derived markers are a suitable tool for diversity and phylogenetic studies, unusual patterns that do not follow expected relationships can arise because of the nature of TEs. High rates of recombination and mutation can account for erroneously selecting members of distinct families for molecular marker use, bearing different evolutionary histories even when thinking that just specific families are being studied. Another possibility is that events of horizontal transfer can disrupt normal evolutionary patterns associated with the host. In a recent study evaluating 40 plant genomes, 26 of the genomes had at least one case of LTR

retrotransposon horizontal transfer [192]. The exchanges occurred in phylogenetically distant plants (e.g. palm and grapevine), with TEs being functional after the transfer and able to actively transpose in some cases. While the mode of transfer is still an enigma, it is possible that the close relationship of retrotransposons with retroviruses presents an alternative for movement if a LTR retrotransposon could find a way to create infectious particles, but also parasitism between plants or common pathogens could account for TE horizontal movement [192].

1.2.4.3 Retrotransposon-based insertion polymorphism (RBIP)

Different from SSAPs, IRAPs, and REMAPs, the design of RBIP markers requires previous knowledge of both the TE and the flanking sequence, since primers are designed from the retrotransposon and the specific sequences surrounding the insertion site [168] (**Figure 1.4**). Of the four types of markers, RBIPs are the only ones with true co-dominance, and since each amplification product results from a specific locus, this technique is more amenable for phylogenetic studies. Despite the fact that the technique may involve more work to discover the TE flanking regions that are necessary for primer design, additional information is gained on the insertion site preferences of the TEs and of the possible evolutionary and regulatory implications of each insertion.

The *PDR1 copia*-type retrotransposon from pea is the most studied TE in *Pisum*, and has been used repeatedly with RBIP. Sixty eight PDR insertions were tested in 47 *Pisum* accessions and compared with previous SSAP analyses; results showed that only certain insertions followed the phylogeny of the accessions and others did not, suggesting that introgression had occurred between different germplasm sources [193]. The study also demonstrated that most target sites for this TE were other TEs, with only 7% of the insertions being targeted to coding regions (without counting insertions into introns and locaitons in close proximity to genes). Another study evaluated 3020 *Pisum* accessions using RBIP mostly with retrotransposon *PDR1*, and the polymorphism of the markers followed the grouping of landraces, cultivars, domestication events and geographical distribution of the pea accessions [194].

Rice is another crop where RBIPs have been exploited. From the insertions of 179 retrotransposons from four families (including *copia* and *gypsy* retrotransposons) in two rice varieties representing the Indica and Japonica genomes, the authors concluded that the gene radiation of the two gene pools marking the divergence of the rice types dated to 200,000 years

ago, which is older than the established domestication date for rice during the Neolithic at 10,000 years ago; this conclusion argued for an independent domestication event for both rice types [195]. The analysis from these two rice cultivars was extended to 66 landraces (both from Japonica and Indica types) to evaluate 13 specific insertion events using RBIPs, which confirmed the double domestication event [195]. The availability of genome sequences nowadays allows extrapolating these techniques to *in*-silico analyses. Also in rice, three cultivated varieties were searched for RBIPs by aligning their syntenic regions and looking for presence or absence of the TE insertions; the analysis demonstrated how certain retrotransposons had irradiated after divergence of the cultivars but other TEs indicated events of introgression [196]. When analyzing one of the most representative Ty1-*copia* retrotransposons from rice (*Tos17*), the distribution of the copies throughout accessions of *Oryza sativa* showed good agreement with isozyme analyses; furthermore, sequenced insertion sites demonstrated a bias to stress response genes, and confirmation of polymorphism between accessions performed using RBIP demonstrated activity of these elements since domestication started [197].

Finally, RBIPs have also been exploited in pears (*Pyrus*), where 80 cultivars were tested with 22 RBIP markers showing the presence of the Ty1-*copia* retrotransposon (*Ppcrt4*) RBIPs in in Asian but not in European pears [198]. Another study with 25 RBIPs from *Ppcrt1* in 110 accessions of pear was mostly in agreement with previous *Pyrus* studies using markers like AFLPS, RAPDs and SSRs, which divided pears into two geographical groups: oriental an occidental pears [199].

1.2.4.4 Inter-primer binding site (iPBS)

iPBS is a more recent technique that uses the conserved tRNA binding site used for priming during retrotranscription in LTR retrotransposons [169] for amplification of the intervening sequence between two TE insertions. This includes the LTR since the PBS (primer binding site) flanks the LTR in the internal section of the retroelement, plus the flanking host DNA region between the two elements (**Figure 1.4**). This technique is not only useful to find polymorphisms between species, accessions or individuals, but also allows to isolate the LTR sequence, which can be used to develop more TE molecular markers; additionally, because of the high conservation and omnipresence of this region in retroelements, it allows for identification of both autonomous and non-autonomous TEs. As with techniques like IRAP, the amplification of

products using iPBS depends on how close the TEs are from each other, with the disadvantage that additional complete LTR sequences have to be covered to complete the amplification products, and this might pose PCR problems as LTRs and intervening sequence become longer. Also, because of expected proximity of TEs, most of these bands should insert away from gene-rich regions, where TEs are purified to avoid detrimental effects on genes. Finally, while the PBS in plants is mostly part of retrotransposons, other sequences might have the same binding site and therefore not all derived products will correspond to TEs unless selective bases are added at the end of the primers.

iPBS was used to isolate LTRs from *Helianthus annus* (sunflower) [188] and from *Linum usitatissimum* (flax) [189], to generate primers for IRAPs. As a molecular marker, iPBS was used to assess the diversity of 35 grape varieties. In this study, the use of 15 iPBS primers resulted in an average polymorphism of 86.3%, and good separation of the cultivated and wild varieties when building a dendrogram [200]. Likewise, 104 landraces and 34 field pea breeding lines from Turkey were evaluated with 12 iPBS primers yielding 76.4% polymorphic bands; the analysis showed that field peas did not display congruent patterns between these markers and geographical distribution [201]. Some additional studies have been performed with iPBS in conjunction with other molecular markers (see below).

1.2.4.5 Using diverse markers in the same study

Variation can also be analyzed by comparing several marker techniques at the same time, both derived from retrotransposons and from traditional molecular markers; the reason for this is that some markers may not yield enough variability to distinguish between closely related individuals, as is the case of recently diverging cultivars or varieties. For example, clementine oranges show minimal genetic variability when examined with SSRs, AFLPs and RAPDs (random amplification of polymorphic DNA), but IRAPs anchored in *copia* elements show higher frequency of polymorphism [202]. These retrotransposon insertions followed a pattern congruent with the diversification process of the oranges, where IRAP bands were lost from the older cultivars from which newer accessions had been generated. This pattern is likely a result of older insertions having more accumulated mutations that impair primer binding, or probably, purifying selection of TEs in older insertions. When IRAP and REMAP markers from Ty1-*copia* and *Ty3-gypsy* were used along RAPD and ISSR (inter-simple sequence repeat) markers to

assess diversity in the genus *Citrus*, all four markers – analyzed separated or together - gave similar relationships among the species [203]. In the *Coffea* genus, SSAPs were used along REMAPs and RBIPs to characterize 182 accessions using two Ty1-*copia* retrotransposons (*Nana* and *Divo*) [204]. While *Nana* was efficient in resolving species-level differences, *Divo* could be used for the lower taxonomic levels, which reflected their evolutionary history, with *Nana* inserting earlier and before the divergence of cultivated and wild species, and *Divo* being absent in the wild species. In another experiment using microsatellites, IRAPs and RBIPs, the microsatellites were the only markers able to clearly distinguish 25 varieties of pea. IRAPs were easy to design and use, but allowed the distinction of only 64% of the tested varieties, while RBIPs using the Ty1-*copia* PDR-1, were robust and easy to score, distinguishing 72% of the tested varieties [205]; another two studies of pea confirmed this trend where SSRs displayed more polymorphism and more discriminatory power than RBIPs [206,207]. SSRs, AFLPs and RAPDs were also used along REMAP to generate a linkage map in Japanese gentians which have great floricultural value; the retrotransposon derived primers for REMAPs were designed after using iPBS to isolate LTRs [208]. Nineteen linkage groups that included 30 REMAP markers plus 133 traditional markers were obtained, constituting a genetic resource for breeding programs. iPBS was also used along with ISSR to explore additional genetic diversity in wild species of *Cicer,* since the diversity in the cultivated species of chickpea (*Cicer arientinum*) is limited. iPBS with 10 primers produced 130 polymorphic bands, while the same number of primers used for ISSR produced 136 scorable bands [209] for 71 evaluated accessions (six species). iPBS showed high level of polymorphism, and confirmed that the cultivated species of chickpea possess lower genetic diversity.

1.2.4.6 NGS for studying TEs

Next generation sequencing (NGS) allows acquisition of information of millions of reads in a single experiment [210], sometimes in days, depending on the organism that is being sequenced. This has resulted in the nearly complete annotation of genome assemblies that include both genes and TEs, which in plants represent a substantial percentage of most genomes [28,29]. Therefore, analysis of polymorphisms by the classic TE molecular markers is steadily being replaced now by full genome-scale analysis of diversity based on TEs. The use of NGS in TE studies not only has the advantage of performing more complete studies, but also the bias for

studying specific families of TEs can now be removed, and a more complete picture of the evolutionary dynamics of different classes of TEs can be elucidated.

A similar procedure to RBIP was used by ME-Scan, to identify presence/absence of TE specific families, with previous knowledge of the TE-flanking region junctions and the flanking regions of the TE [211]. By generating indexed (barcoded) amplicons from different samples, it is possible to use NGS to assess which samples have the TE present in specific loci. This procedure was dependent on amplification of specific TEs, but the possibility of sequencing full genomes in addition to the reference genome, opened the way for evaluating all polymorphic TE insertions of any family between the samples.

Initially, the recognition of TE insertional polymorphisms could be detected by performing pair-end mapping to well-curated reference genome. For example, if someone wanted to see how TEs from a distinct variety of *Arabidopsis* aligned to the reference genome, a library of paired end fragments could be built so that the mapping of the two fragments could reveal either presence, absence or new insertions of a TE in the query genome (**Figure 1.5**). During the process of mapping if both paired-end reads match the genome at the expected distance from each other, one can infer there is no variation, and these pair-ends become non-informative. The remaining read pairs can be used to match at least one of the two paired-end reads of the query genome to either the genome or a database of TEs (or a TE which is in a different locus), indicating polymorphism between query and reference (**Figure 1.5**). Since hundreds or thousands of reads align in the different tested loci, algorithms designed under this premise would easily detect the variations. Ideally this whole process should be performed with paired-ends that are separated by at least 1-3 kb so that there are sufficient reads that can map to the TE and outside of it, since many TEs can be several kilobases long. Other characteristics of algorithms to distinguish TE polymorphisms between reference and query genomes include: i) reads that flank an annotated TE in the reference genome but demonstrate a distance between them that is less than expected if the insertion was present (paired-end read would have to be kilobases long depending on the TE type), ii) reads that span a junction between a TE and a flanking region which would have to be broken up to be mapped to the reference genome, are an indication of a TE in the query but not in the reference genome, and iii) enough read depth supporting these variations.

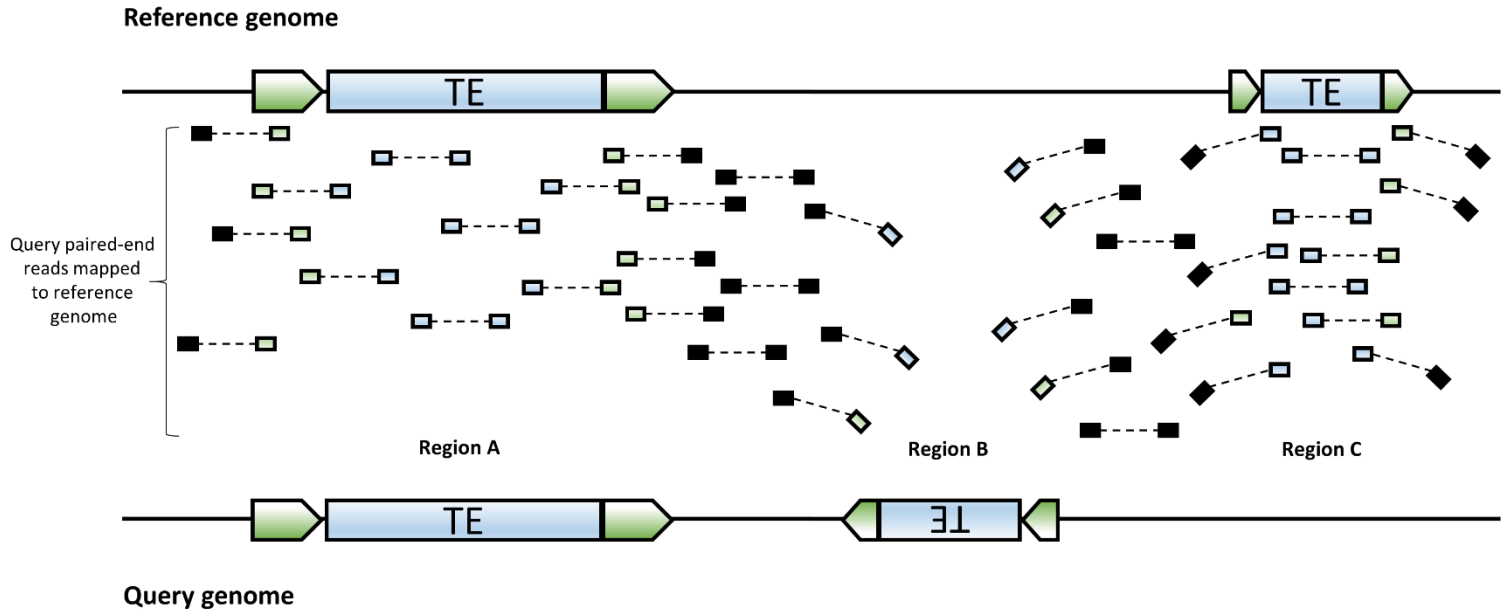

Reference genome
TE
TE
Query paired-end reads mapped to reference genome
Region A
Region B
Region C
TE
TE
Query genome

Figure 1.5 Paired-end read mapping for transposable element (TE) insertion polymorphism discovery. In region A, there is no change between the reference and query genomes and the paired end reads map perfectly. On region B, there is a new TE insertion in the query genome (alternatively the TE could have been removed from the reference) which is absent in the reference genome. One of the reads from the pair maps to the reference genome but the matching pair which matches a TE does not map to the reference at the expected distance between the pair; the hanging read however matches a TE elsewhere in the genome or could match a TE from a TE database. In region C, a TE has been removed from the query sequence (alternatively a novel TE insertion could have happened in the reference genome). Reads which match the TE in the reference probably correspond to a similar TE but located elsewhere in the query genome and for this reason the paired read outside the TE does not match the surrounding sequence.

A simple paired-end mapping approach was used for mapping TEs from a mutant line of rice to the reference genome. TE *de*-novo insertions in the query (mutant line) genome would have one non-TE read mapped unambiguously to the reference genome and another corresponding to a known TE located elsewhere, showing that a novel TE insertion was located in the mutant line [212]. Some of the first algorithms that used paired-end reads to find structural variations were adapted to find TE derived variations (e,g. VariationHunter [213]). One of the pioneering algorithms designed specifically for TE detection (T-lex) builds a database of the edge regions of TEs plus flanking sequences (TE-flanking region junctions) and maps the contigs of the query genome to these regions to find presence of the TEs. Alternatively a sequence is created with 100 bp of the flanking regions of the TEs to emulate the ancestral state of no insertion, and reads that map and expand this state from the query genome indicate absence of the TE [214]. Another TE-polymorphism discovery software, PAIR, was designed to discover Alu insertions under the same premise of two reads mapping discordantly, one of them mapping to a different location than the expected for its respective mate or pair-end read [215]. RelocaTE identifies reads containing TEs (using a database of TEs or specific TEs) plus flanking sequences, trims the TE sequence, and aligns the remaining sequence to a reference genome, to identify to potential sites of *de*-novo TE insertions. Two reads that overlap the alleged insertion site in the query genome, must align to the reference through the target site duplication which should be at the end of the reads after the TE sequence has been trimmed [216]. Recently more advanced algorithms that include different levels of confirmation for TE insertions have been

developed. For example, ITIS (Identification of Transposon Insertion Sites), first aligns the paired ends to the reference and a TE database, and discards read pairs that both align to the reference at the expected distance. Then uses the reads where at least one of the two maps to a TE or a junction of a TE/flanking sequence. The mapping to the reference reveals potential insertion sites. However ITIS also uses read-depth to support a real insertion event, and the possibility of the insertion being homozygous or heterozygous is also determined using the amount of reads mapped [217].

TE-derived markers have provided ample coverage in diversity and evolution studies for close to 20 years. The introduction of NGS and the rapid advances and decreasing costs of these technologies are already allowing to incorporate the basis of the marker development in genome-wide studies. While comparing *de*-novo assemblies for TE-based polymorphism is a possibility, one of the greatest hurdles in genome assembly is the repetitive character of TEs. Therefore, the presence of a well-curated reference genome constitutes a valuable tool to find TE-based polymorphisms in new individuals or varieties.

1.3 Overview

Evidenced by the background presented on this general introduction, TEs, and specifically Ty1-*copia* elements, exert mutational and regulatory influence on their insertion sites and flanking regions. Their high-relative abundance and diversity in most plant species is related with transpositional activity, and many TEs are currently active demonstrating that they are an important part of genome restructuring and evolution. The annotation of TEs in flax showed that over 23% of the genome was made of TEs, and 40% of such insertion were characterized as Ty1-*copia* elements, from which many represented recent insertions [218]. These characteristics indicated Ty1-*copia* elements might be most suitable for exploring TE-mediated changes in the flax genome. Because annotation of the flax genome is relatively recent, the influence of non-coding DNA and TEs on the flax genome is not known. Furthermore, it is not known if breeding practices and other commonly known stresses studied in flax can activate TEs and result in genome variability.

For the development of this thesis work I wanted to ask if members of the Ty1-*copia* superfamily had an influence on the diversification of flax, and on restructuring of the genome via mutational changes due to insertional polymorphism. My general hypothesis was that Ty1-

copia families with high similarity in their LTRs are potentially the most active, and should have an influence on flax diversification, and additionally be regulated upon common stresses modulating TEs in other plants (e.g. microbial elicitors, wounding). To answer this general question, I designed several experiments described below.

In Chapter 2, I compare several flax cultivars for TE-derived insertional polymorphisms of Ty1-*copia* families. I hypothesized that the conservation of LTR pairs and internal protein domains in certain Ty1-*copia* families would result in these families generating insertional polymorphisms among cultivars, and that their insertion would likely be in the flanking regions of genes as was previously predicted by bioinformatics approaches [218]. The analysis showed that the TE families have been recently active and are part of the molecular diversity of the cultivars. Additionally, the insertional patterns of the TEs showed that numerous TEs may have an influence on gene structure and function, but fall not only in flanking regions but also inside introns and exons.

In Chapter 3, I explore potential stress factors that may trigger TE activity in Ty1-*copia* families (some of which were used also in Chapter 2). I hypothesized that stress factors that have commonly been associated with activation of other Ty1-*elements* in other plants (e.g. fungal elicitors and wounding) [48,55,65,92] should have a regulatory effect on flax TE families. Since these families showed the highest probability of being active by bioinformatic analyses, and showed diversification patterns among cultivars, they were also likely to be active upon an external trigger. The stress factors used included wounding and microbial elicitors, which are common triggers of plant defense responses but also of TEs [45,55,58,59,65,77,219]. While the analyses showed no consistent pattern of responsiveness to these stresses, certain TE families demonstrated constitutive expression. TE families were tested for differences in expression among tissues, showing that TE expression may be tissue/organ-specific.

One of the elicitors tested in Chapter 3 as a potential TE activator was infection of flax plants with the pathogenic fungus *Fusarium oxysporum*. While no major changes in TE activity could be detected by this stress, we explored the transcriptome changes that underlie the molecular response of flax to the pathogen, which is one of the two most important pathogens of this species.

In Chapter 4, I show the disease progression of two flax cultivars with distinct levels of resistance to *Fusarium oxysporum,* and characterize the transcriptome response of the most

resistant cultivar. I hypothesized that flax would activate its defense response genes early (within the first two days post-inoculation –DPI-) as has been shown for early activation of specific genes and physiological processes in the *F. oxysporum*-flax pathosystem [220–224] to efficiently cope with the pathogen infection. Both the comparison of the two cultivars and the RNA-seq study demonstrated that molecular defense responses were activated mostly later (18 DPI). The RNA-seq transcriptome evidenced an array of genes involved in the defense response of the elite cultivar CDC Bethune, but also some groups of genes potentially manipulated by the invading pathogen. This study revealed numerous candidate genes that can be used as markers of resistance, but also established a wide molecular basis for the study of this and related pathosystems.

In Chapter 5, along with several collaborators, I developed a reverse genetics methodology to find rare variants in pooled mutant lines using next generation sequencing Ion Torrent technology. We hypothesized that by means of high-throughput sequencing and bioinformatics analyses, hundreds of samples from mutant flax lines could be evaluated at the same time to find rare variants in genes of interest, and that this methodology could be applied to chose mutants with genes that could have a phenotypic effect. The tests of this methodology were performed in genes of interest related to desirable flax traits (e.g. cell wall, metabolism), but can be applied to any gene. In the future it is expected that this methodology could be applied for example, to find mutations in genes related to the processing of epigenetic changes which control TEs in genomes. For example, by finding individuals with mutations that affect the function of genes processing the epigenetic marks that result in methylation of TEs, one can explore what families of TEs are activated, and establish a relationship to changes in methylation.

In Chapter 6, I give my concluding remarks and try to argue why TEs are an important part of flax evolution, and what avenues of research can be explored to further the research developed during this thesis.

CHAPTER 2 - Ty1-*copia* elements reveal diverse insertion sites linked to polymorphisms among flax (*Linum usitatissimum* L.) accessions.

This chapter is based on an article accepted for publication: Galindo-González L.; Mhiri C.; Grandbastien M.A.; Deyholos M.K. 2016. Ty1-*copia* elements reveal diverse insertion sites linked to polymorphisms among flax (*Linum usitatissimum* L.) accessions. BMC Genomics (submitted June 8-2016: manuscript ID GICS-D-16-00881, 57 pages).

2.1 Abstract

Initial characterization of the flax genome showed that Ty1-*copia* retrotransposons are abundant, with several members being recently inserted, and in close association with genes. Recent insertions have a potential for ongoing transpositional activity that can create genomic diversity among accessions, cultivars or varieties. The polymorphisms generated constitute a good source of molecular markers that may be associated with specific phenotypes if the insertions alter gene activity. Flax, where accessions are bred either for seed nutritional properties or for fibers, constitutes a good model for studying the relationship of transpositional activity with diversification and breeding. In this study, we estimated copy number and used a type of transposon display known as Sequence-Specific Amplification Polymorphisms (SSAPs), to characterize six families of Ty1-*copia* elements across 14 flax accessions. Polymorphic insertion sites were sequenced to find insertions that could potentially alter gene expression, and a preliminary test was performed with selected genes bearing TE insertions.

Quantification of six families of Ty1-*copia* elements indicated different abundances within and between flax accessions, which suggested a diverse transpositional history. SSAPs showed a high level of polymorphism in most of the evaluated retrotransposon families, with a trend towards higher levels of polymorphism in low-copy families. Ty1-*copia* insertion polymorphisms among cultivars allowed a general distinction between oil and fiber types, and between spring and winter types, demonstrating their utility in diversity studies. Characterization of polymorphic insertions revealed an overwhelming association with genes, with insertions disrupting exons, introns or within 1 kb of coding regions. A preliminary test on the potential transcriptional disruption by TEs of four selected genes evaluated in three different tissues, showed one case of significant impact of the insertion on gene expression.

We demonstrated that specific Ty1-*copia* families have been active since breeding commenced in flax. The retrotransposon-derived polymorphism can be used to separate flax types, and the close association of many insertions with genes defines a good source of potential mutations that could be associated with phenotypic changes, resulting in diversification processes.

2.2 Introduction

Transposable elements (TEs) are DNA fragments that can move between genomic locations using a cut and paste mechanism (DNA transposons), or a copy and paste mechanism via an RNA intermediate (retrotransposons). Transposition can result in alterations of gene expression and diversification between individuals, populations and species. TEs are commonly activated upon stresses that include tissue culture, wounding, microbial elicitors and pathogen attack [25,56,57,62,64,78,79,178]. Polyploidization (whether spontaneous or induced) also mobilizes transposable elements, resulting in genome restructuring, and genetic and epigenetic effects on gene activity [89]. Selective breeding can also affect TE activity. For example, in vegetatively propagated grape clones, TE insertional polymorphisms constitute the largest class of mutations [225]. Genetic diversity associated with TE polymorphisms has been commonly explored in plant varieties and species such as pepper and tomato [81,178], barley [174], strawberry [181], coffee [204], blue agave [184] and cashew [177].

We previously showed that more than 20% of the flax (*Linum usitatissimum*) genome is made of TEs [15,218]. Furthermore, the main superfamilies represented in the genome are LTR (Long Terminal Repeat) retrotransposons. The Ty1-*copia* elements are the dominant superfamily, and have numerous members which have been recently inserted, as inferred from their LTR similarity and gene domain conservation (at least 83 Ty1-*copia* elements have 100% LTR similarity) [218]. Furthermore, Ty1-*copia* elements have had increasing activity in the flax genome starting five million years ago [218]. These observations indicate that Ty1-*copia* elements could generate polymorphisms among closely related flax cultivars.

Flax is a valuable source of bioproducts derived from the seed (i.e. linseed) and stem fiber [226]. Its breeding for either seed or fiber traits in diverse climates has resulted in diverse cultivars and an array of agrobotanical characteristics that have been artificially selected [226]. While flax grown for human consumption (seeds are used for nutrition but also for oil derived industrial products), is the same species as the flax grown mainly to manufacture linen, they represent two different flax types (oil and fiber) and the products are usually obtained from cultivars (or accessions) that have been bred to have mainly one of the two characteristics [226]. Additionally, flax is a summer annual crop in temperate climates and is usually sown during spring, but winter cultivars have been bred that can be sown in the autumn in milder climates.

Flax is therefore an interesting system for studying the relationship of TEs to continuous and divergent selection practices.

The current study aims to uncover the impact of specific Ty1-*copia* retrotransposon families on diversification of flax cultivars. We measured the level of polymorphism among a set of flax cultivars, and analyzed their relationship using a TE-based marker system. Since TE insertions within genes are more likely to interfere with gene function, we characterized the nature of target sequences of polymorphic insertions to find out if they were closely associated with genes, and measured the effect of retrotransposon insertion on transcript expression in selected genes.

Several strategies have been devised to find TE insertional polymorphisms [227]. Previously, Inter-Retrotransposon Amplified Polymorphism (IRAP) was used to study flax cultivars and species [189]. Here we used a type of transposon display (TD), known as Sequence-Specific Amplification Polymorphism (SSAP) [164,165], to evaluate Ty1-*copia* retrotransposons insertion in 14 flax accessions of either oil or fiber types, and spring or winter types. In SSAP, TEs and flanking DNA are preferentially amplified using a PCR primer that anneals to a sequence specific to a particular TE family (usually an LTR), and a second primer that anneals to an adaptor ligated to a restriction enzyme site. Our study shows that families of flax Ty1-*copia* TEs, have high levels of polymorphism between cultivars, indicating recent activity since organized breeding commenced in the last century. While the copy number of each family did not vary greatly between cultivars, some families of TEs were consistently more abundant than others across multiple cultivars. Analysis of sequence insertion sites demonstrated that many of these Ty1-*copia* elements inserted within or in close proximity to genes. Finally, we found one case where an insertion of a TE in an exon of a Laccase gene decreased gene expression in roots. Our study demonstrates that TEs from the Ty1-*copia* group have been part of the diversification associated with breeding, and that they may play a role in modifying gene expression patterns in the flax genome, which can lead to diversified phenotypes.

2.3 Materials and Methods

2.3.1 Plant material

For determination of TE family copy numbers, eight plants from each of 14 flax cultivars or accessions (**Table 2.1**) were grown in a growth chamber at the University of Alberta under the

following settings: seeds sown in pods with a 50/50 soil/sand mix, 16 hours of light / 8 hours of dark (0.132 µMoles of light), 22°C, 50% humidity. Aerial sections (stems + leaves) were harvested after two weeks of growth and instantly frozen with liquid nitrogen in 2 mL tubes.

Table 2.1 Cultivars used for transposon display.

Cultivar	Type
Stormont Cirrus	fiber spring
Aurore	fiber spring
Belinka	fiber spring
Drakkar	fiber spring
Evea	fiber spring
Hermes	fiber spring
Violin	fiber winter
Adelie	fiber winter
Rdf[a]	oil spring
CDC Bethune	oil spring
Lutea	oil spring
Blizzard	oil winter
Oleane	oil winter
Oliver	oil winter

[a] *rdf* is a mutant derived from CDC Bethune and therefore cannot be classified as a cultivar *per-se*, and should be referred as an accession.

For SSAPs, 14 flax cultivars were used (**Table 2.1**). Plants were grown in greenhouse conditions (14 hours of light, 24°C day / 20°C night, 40% humidity) at the National Institute for Agronomic Research (INRA) in Versailles, France. Seeds were sown in pods with a 50/50 soil/sand mix, and left to grow for two weeks before aerial sections were collected in 2 mL tubes and instantly frozen in liquid nitrogen.

For testing the expression of genes bearing polymorphic TE insertions among cultivars, additional plants of each cultivar were grown in the same growth chamber at the University of

Alberta under the same conditions used for determination of TE copy number. Stems, 5-10 young leaves (including the apical meristem) and roots were harvested after 2-3 weeks of growth.

2.3.2 Nucleic acids extraction and cDNA synthesis

The samples for SSAPs were ground with a plastic pestle maintaining the tube in liquid nitrogen until achieving a fine powder. Samples for SSAP validation, transposon families copy number determination and gene expression were ground adding an autoclaved 5.6 mm stainless steel bead, and using a Retsch MM301 mixer mill (Retsch, Haan, Germany) with two cycles of 1 minute at 20 Hz. DNA extraction was performed using the DNeasy Plant Mini Kit (QIAGEN, Venlo, The Netherlands). Sample quantification was performed with a Nanodrop ND-1000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA).

RNA was extracted using the RNeasy Plant Mini Kit (QIAGEN, Venlo, The Netherlands), and quantity was assessed using a Nanodrop ND-1000 spectrophotometer (Thermo Scientific, Waltham, MA, USA). A DNAse treatment was performed for 30 minutes at 37°C after extraction with DNaseI (Thermo Scientific, Waltham, MA, USA). For cDNA synthesis 500 ng of DNAse treated RNA were used to perform reverse transcription using the RevertAid H Minus Reverse transcriptase under the manufacturer specifications and using oligo dT (18) (Thermo Scientific, Waltham, MA, USA). To test for residual contamination of DNA, a PCR was performed with primers from the eukaryotic translation initiation factor 3E (ETIF3E) which has constitutive expression in the tested tissues (**Table 2.2**). The PCR was run with 1X buffer, 2 mM $MgCl_2$, 0.2 mM dNTPs, 0.4 μM of each primer, 5 ng of cDNA and 1.5 units of Taq polymerase (Thermo Scientific, Waltham, MA, USA). Cycling conditions were 94°C for 2 minutes, followed by 35 cycles of 94°C for 30 seconds, 60°C for 30 seconds and 72°C for 1 minute, finalizing with an extension at 72°C for 5 minutes.

Table 2.2 Primers used to test the expression of reference genes.

Name	alias	sequence
Elongation factor 1-α	EF1A-fw	gctgccaacttcacatctca
Elongation factor 1-α	EF1A-rv	gatcgcctgtcaatcttggt
Eukaryotic translation initiation factor 3E	ETIF3E-fw	ttactgtcgcatccatcagc
Eukaryotic translation initiation factor 3E	ETIF3E-rv	ggagttgcggatgaggttta
Eukaryotic translation factor 5A	ETIF5A-fw	tgccacatgtgaaccgtact
Eukaryotic translation factor 5A	ETIF5A-rv	ctttaccctcagcaaatccg
Glyceraldehyde 3-phosphate dehydrogenase	GAPDH-fw[a]	gaccatcaaacaaggactgga
Glyceraldehyde 3-phosphate dehydrogenase	GAPDH-rv[a]	tgctgctgggaatgatgtt
Ubiquitin	UBI-fw	ctccgtggaggtatgcagat
Ubiquitin	UBI-rv	ttccttgtcctggatcttcg
Ubiquitin extension protein	UBI2-fw	ccaagatccaggacaaggaa
Ubiquitin extension protein	UBI2-rv	gaaccaggtggagagtcgat
Eukaryotic translation initiation factor 1	ETIF1-fw[a]	ctcaggtgatgcgaatgct
Eukaryotic translation initiation factor 1	ETIF1-rv[a]	aatccctcagccctacaagg

[a] Not from Huis et al., 2010 [228]

2.3.3 TE primers

Retrotransposon sequences were obtained from our previous study on transposable elements of flax [218]. To design Ty1-*copia* primers, family membership of a Ty1-*copia* element was defined with a threshold similarity of at least 80% in at least 80% of the aligned sequence, following previously established rules for family membership [32]. The comparison was performed on the 554 non-redundant reverse transcriptase (RT) domain sequences, which were first predicted using RepeatExplorer [229], and then used as input for CD-HIT-est [230] using an identity cutoff and minimal alignment coverages of 0.8. Families were named using the suggested designation of class, order and superfamily [32], followed by a species designation, and a number corresponding to the specific TE (e.g. Retrotransposon-LTR-Copia from *Linum usitatissimum* family 0, representative sequence 1 = RLC_Lu0-1). Selected families with evidence of recent insertion (high similarity among its LTRs and conserved domain proteins)

were selected for primer design. To calculate the insertion date of the TEs, first LTR pairs from each element where aligned using Clustal W [231] and the Kimura two-parameter method [232] was used to calculate nucleotide substitution. Then, the age of insertion was estimated as $t = K/2r$, where K corresponds to the nucleotide substitution per site and r corresponds to the nucleotide substitution rate, which in this case was taken from a previous study used for dating LTR retrotransposons in *Arabidopsis* [66]. The presence of the main protein domains in Ty1-*copia* elements: GAG (group-specific antigen), PR (protease), INT (integrase), RT (reverse transcriptase) and RNase H, was assessed using conserved domains from NCBI [233,234], and RepeatExplorer [229].

To design reverse transcriptase (RT) primers to assess TE copy number (see below), the RT nucleotide sequences from all members in each TE family were aligned using Clustal W [231] from MEGA v6 [235] with the following parameters: a gap opening penalty of 15 and a gap extension penalty of 6.66 for both pairwise and multiple alignments, DNA weight matrix – IUB, transition weight of 0.5, negative matrix off and delay divergent sequences that have less than 40% similarity. For each family, one representative sequence bearing conserved sites for primer design from all (or most) family members, was used as input for Primer3 [236,237] with the following parameters: primer size range between 18 and 24bp, temperature between 57 and 63°C, product size 100-200bp, and GC content between 40 and 60% (the rest of the parameters were left by default). Candidate representative sequences from each alignment were chosen based on preliminary bioinformatics analysis showing protein domain conservation, at least 4 out of 5 of expected domains (GAG, PR, INT, RT and RNase H) and high LTR similarity among all members of the family (not shown). The reference gene used to normalize copy number was ETIF1 (eukaryotic translation initiation factor 1), which has been previously tested in flax quantitative gene expression [228]. Selected RT primer pairs (**Table 2.3**), were aligned to the flax genome (CDC Bethune) using BLAST to get an estimate of the expected copy numbers per TE family. From two primer pairs designed per family (six families in total) for qPCR (see below), the one with a better standard curve was selected in each family.

Table 2.3 Reverse transcriptase (RT) primers to evaluate TE copy number.

Primer name	Sequence	Expected number of hits in CDC Bethune using BLASTn
RT-RLC_Lu0-a-2-fw	ggcccctataccaattagatgtg	24
RT-RLC_Lu0-a-2-rv	cttcttcagtctgcacaccat	
RT-RLC_Lu1-a-1-fw	ggagagacacaaggctaggc	21
RT-RLC_Lu1-a-1-rv	gacgtccatttgatataggggc	
RT-RLC_Lu2-a-2-fw	ttctcaccagtggcaaagat	24
RT-RLC_Lu2-a-2-rv	tcctcatccaagtctccatg	
RT-RLC_Lu6-a-1-fw	ttcagtcaaaggaagggcatc	47
RT-RLC_Lu6-a-1-rv	tcttcctccaaatcgccatg	
RT-RLC_Lu8-a-1-fw	tggtgacctgcatgaagaagt	18
RT-RLC_Lu8-a-1-rv	agtaccactgccttgatgct	
RT-RLC_Lu28-a-1-fw	tggaggagttacgagctttgg	45
RT-RLC_Lu28-a-1-rv	ccgtctgctctatatttaatggtg	
ETIF1-fw	ccttgtagggctgagggatt	1
ETIF1-rv	ctcatcaagaccaccagcaa	

To design primers for SSAPs, LTRs from all family members in each of the selected families were aligned following the same alignment parameters as for the RT sequences when designing primers to assess TE family copy number. After filtering largely divergent and redundant sequences from the alignment, the complete retrotransposon sequences were checked for the presence of internal *Eco*RI sites in order to minimize the chance of amplifying retrotransposon internal regions. The LTR alignments were then scanned for a region that is conserved among most aligned elements of the family, to maximize the generation SSAP bands. A representative sequence was selected in each family (**Table 2.4**), and the conserved region was then used as input for Primer3 [236,237] along with the primer corresponding to the *Eco*RI adapter with the following parameters: primer size range between 20 and 24bp, temperature between 55 and 62°C and GC content between 30 and 70% (the rest of the parameters were left by default). A total of 19 primers were designed for six TE families, but only seven were used for the final experiment (**Table 2.5**).

Table 2.4 Insertion age and domains of representative sequences from selected Ty1-*copia* families.

Identifiers			Insertion parameters			Domains[e]				
TE representative sequence scaffold location[a]	Cluster	TE name[b]	LTR identity	Kimura 2-parameter distance[c]	Insertion date[d]	1	2	3	4	5
S786_34727-39745	0	RLC_Lu0-1	99.8	0.004	133333.3	y	y	y	y	y
S147_473692-477955	1	RLC_Lu1-1	100	0	0	y	n	y	y	y
S1042_129095-134034	2	RLC_Lu2-1	100	0	0	y	y	y	y	y
S464_146788-151474	6	RLC_Lu6-1	99.2	0.006	200000	n	y	y	y	y
S98_763798-768930	8	RLC_Lu8-1	100	0	0	y	y	y	y	y
S272_1752754-1757140	28	RLC_Lu28-1	100	0	0	y	y	n	y	y

[a]Location in scaffold from flax draft genome available at phytozome [238,239]. The TEs were first annotated using LTR finder [218].
[b]e.g. RLC_Lu0-1 = Retrotransposon-LTR-Copia *Linum usitatissimum* family 0, TE representative sequence number 1 in the family.
[c]Calculated for LTR pairs.
[d]Calculated as t=K/2r. t is the insertion date in years, K is the nucleotide substitution between LTRs given by the kimura 2-parameter distance, r corresponds to the nucleotide substitution rate taken from reference [240].
[e]Protein domains identified with RepeatExplorer [229] and conserved domains from NCBI [233,234]. 1 = GAG, 2 = PR, 3 = INT, 4 = RT, 5 = RH. y = yes, n = no.

Table 2.5 LTR primers and adaptor sequences used for SSAP.

Primer name	Position in LTR	sequence
EcoRI Adaptor 1		ctcaggctcgtagactgcgatcc
EcoRI Adaptor 2		aattggatcgcag
EcoRI primer 00		gtagactgcgatccaattc
LTR-RLC_Lu0-primer3	5'LTR - reverse	gtagtaatccctacaaatcagg
LTR-RLC_Lu1-primer1	5'LTR - reverse	atacagcattcctcctactgac
LTR-RLC_Lu1-primer2	5'LTR - reverse	agatctgactgtatacaaataagg
LTR-RLC_Lu2-primer1	3'LTR - forward	tccttcttcttcgctttctctg
LTR-RLC_Lu6-primer3	3'LTR - forward	atggtattgcggtggtaatca
LTR-RLC_Lu8-primer1	3'LTR - forward	gcgttctgttaagtgatgaaga
LTR-RLC_Lu28-primer1	3'LTR - forward	aacccatgtacccctatgtaatc

2.3.4 Transposon family copy number

To find the absolute copy number of TEs from each of the families used in the SSAPs (see below), we performed qPCR on DNA samples from the 14 cultivars using reverse transcriptase (RT) TE primers (**Table 2.3**). The amplifications were then compared to standard curve dilutions of the cloned amplified RT fragments (see below).

PCR to amplify the RT regions to be cloned was performed with 1X Taq buffer, 2 mM of $MgCl_2$, 0.2 mM dNTPs, 0.2 µM of each primer and 1 unit of recombinant Taq polymerase (Thermo Fisher Scientific, Waltham, MA, USA). Cycling conditions were as follows: 94°C for 2 minutes followed by 35 cycles of 94°C for 30 seconds, 60°C for 30 seconds and 72°C for 1 minute, with a final extension at 72°C for 5 minutes. PCR was run on a 1% agarose gel at 90 V for 60 minutes and the expected amplicon size was assessed. Then, the bands were eluted using the Wizard SV gel and PCR clean-up system (Promega. Madison, WI, USA). Eluted products were quantified using a Nanodrop ND-1000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). For sequencing 150 ng of the eluted product was used along with a primer at a final concentration of 0.25 µM (forward or reverse primers corresponding to the same primers used for PCR). Sequencing reactions were performed with the BigDye terminator v3.1 cycle sequencing kit (Applied Biosystems - Thermo Fisher Scientific, Waltham, MA, USA)

using a 3730 Genetic Analyzer equipment (Applied Biosystems -Thermo Fisher Scientific, Waltham, MA, USA). Sequencing products were aligned with the original RT sequences to confirm that amplification products were as expected.

To clone the amplification products, ~5-8 ng of the insert (this varied depending on the amplicon size) were cloned into the PGEM-T vector II system (Promega, Madison, WI, USA) to create a 3:1 (insert:vector) molar ratio. Ligation products were transformed into JM109 high-efficiency competent cells (Promega, Madison, WI, USA), following the manufacturer recommendations. One hundred microliters of the transformed cultures were plated into LB-agar plates with 2% X-gal, 20% IPTG and 50 ng/µL of ampicillin. Cultures were incubated overnight (ON) at 37°C and white colonies were selected as positive for the insertion.

Selected colonies were grown in LB supplied with 100 ng/µL of ampicillin. Tubes were placed in a shaker at 200 rpm ON (minimum of 12 hours) at 37°C. Plasmids were extracted from concentrated bacterial cultures using the QIAprep Spin Miniprep Kit (QIAGEN, Venlo, The Netherlands) and concentrations were measured using a Nanodrop ND-1000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). To confirm the identity of the cloned products, inserts were reamplified from plasmids using the same conditions previously mentioned, and resequenced using 575 ng of plasmid and the generic T7 and Sp6 primers matching the vector.

Five nanograms of DNA from eight samples of each of the 14 cultivars, and a 1:10 8-serial dilution (5 to 5 x 10^{-7} ng) of each plasmid with the different TE family inserts, were used for qRT-PCR with 5 µL of SYBR green (Molecular Probes – Thermo Fisher Scientific, Waltham, MA, USA) and 2.5 µL of the mixed primer pair (3.2 µM), in a 10 µL reaction (three technical replicates per each sample or dilution). Samples were aliquoted in 384-well plates using a Biomek 3000 Laboratory Automation System (Beckman Coulter, Brea, CA, USA), and the qRT-PCR was run using a QuantStudio 6 Flex Real-Time PCR system (Applied Biosystems-Life Technologies, Carslbad, CA, USA). Cycling conditions were 94°C for 2 minutes, followed by 35 cycles of 94°C for 30 seconds, 60°C for 30 seconds and 72°C for 45 seconds. A melting curve stage was added: 95°C for 15 seconds, 60°C for 1 minute and 95°C for 15 seconds.

To find out the molecule copy number (mcn) in the dilution series, we used the amount of DNA from each point of the serial dilution and the size of the plasmid plus insert [241], and performed a $\log_{10}$ transformation. The standard curve was built by plotting the C_t values (average of technical replicates) against the log10 of mcn. The linear equation for the slope $y = mx+b$ was

used to determine the $\log_{10}$ mcn intercepts (x) for the C_t values of the eight replicates in each of the 14 cultivars for all primers tested. The $\log_{10}$ mcn values where then back-transformed using the power function (10^x). The same procedure was conducted to find out the copy numbers for the reference gene. The copy numbers of the reverse transcriptases of each family for each sample were normalized to the copy numbers of the reference gene (ETIF1) and the average absolute copy number and standard deviation (from the 8 replicates), were calculated and plotted for each cultivar and TE family.

Statistical differences for each TE copy number among cultivars were determined using the non-parametric test of Kruskal-Wallis, followed by multiple comparisons using Dunn's test, using GraphPad Prism version 6.0 (GraphPad Software, La Jolla California USA). Correlation coefficients were calculated in Excel for the relationship between the expected copy number of TEs in each family estimated using the primer pairs to BLAST against the flax genome, and either the calculated copy numbers from qPCR, or the number of scored bands in the SSAPs.

2.3.5 Sequence-Specific Amplification Polymorphism (SSAP)

One hundred nanograms of each DNA sample were used for restriction digestion at 37°C for 16 hours, with 10 units of *Eco*RI and supplemented with 0.03 mg of BSA and 1X restriction-ligation buffer (10 µL of the digestion were used to check the restriction in a 1% agarose gel). A 1 µM mixture of *Eco*RI adapter 1 and 2 (**Table 2.5**), were ligated to the digested ends for 16 hours, using 0.2 mM ATP, 1x restriction-ligation buffer and 0.004 units of T4 DNA ligase (Invitrogen, Carlsbad, CA, USA). Ligations were centrifuged and diluted with 80 µL of 1x TE. To confirm ligation efficiency, a cold PCR (with non-radioactively labelled TE primer) was performed using 1X Taq Buffer, 2 mM $MgCl_2$, 0.2 mM of each dNTP, 0.4 µM of the specific TE primer and 0.4 µM of the EcoRI primer 00 (**Table 2.5**), 1.5 units of recombinant Taq DNA polymerase (Thermo Fisher Scientific, Waltham, MA, USA) and 5 µL of the diluted restriction-ligation. Cycling conditions were 94°C for 5 minutes, followed by 35 cycles of 94°C for 30 seconds, 56°C for 30 seconds and 72°C for 1 minute, finalizing with an extension at 72°C for 5 minutes.

Retrotransposon primer labelling with P^{33} was performed using the LTR TE specific primers (**Table 2.5**) at a final concentration of 4 µM, 1X kinase buffer A, 0.5 units of T4 kinase (Thermo Fisher Scientific, Waltham, MA, USA) and 1 µCi of gamma ATP^{33} (PerkinElmer

Health Sciences, Boston, MA, USA). The cycling conditions to label the primer were: 1 hour at 37°C and 15 minutes at 70°C to inactivate the kinase enzyme (the oligo was kept at -20°C until used in the SSAP PCR).

SSAP PCR was performed with 1x buffer, 2 mM $MgCl_2$, 0.2 mM of each dNTP, 0.4 µM of adapter primer, 0.16 µM of specific radioactively labeled primer, 1.5 units of recombinant Taq DNA polymerase (Thermo Scientific, Thermo Fisher Scientific, Waltham, MA, USA) and 2.5 ng of the restriction-ligation product. Cycling conditions were as following: 94°C for 5 minutes followed by 13 cycles of 94°C for 30 seconds, 65°C for 30 seconds and 72°C for 2 minutes; then 25 cycles of 94°C for 30 seconds, 56°C for 30 seconds and 72°C for 2 minutes; finishing at 72°C for 10 minutes. After PCR the product was diluted 1:1 with 2X AFLP loading buffer and kept at -20°C until running the gel. PCR products were separated in 6% denaturing polyacrylamide gels on a Bio-Rad Sequi-Gen GT electrophoresis system (Bio-Rad, Hercules, CA, USA). After the run, the gel was dried and adhered to Whatman paper, and exposed from 1 to 3 days to Kodak Biomax XAR films (Carestream Health Inc., Rochester, New York, USA) and then developed for band scoring.

2.3.6 Band scoring and neighbor network

Exposed films displaying the SSAP band patterns were captured as images (.tif files) and used as input in GelAnalyzer [242] where bands were digitally scored as present = 1 or absent = 0. The scored bands were used to create a binary matrix utilized as input to generate a neighbor network with the SplitsTree4 software [243]. The parameters used in the program were: least square variance, excluded constant sites and uncorrected p-distance.

2.3.7 Band recovery and sequencing

To recover the polymorphic bands, the exposed film was overlaid on the original dried gels on Whatman (both film and gel were pinned previously on the corners to allow matching). A clean scalpel was used to cut the mapped band on surface of the gel-Whatman assembly, and the detached piece was placed in a 1.5 mL tube with 35 µL of nuclease free water. The band in water was vortexed for 1 minute and spun down for incubation at 37°C for 15-16 hours. The liquid was recovered to a new 1.5 mL tube and 5 µL were used for a PCR with 1X Taq buffer, 2 mM of $MgCl_2$, 0.2 mM dNTPs, 0.2 µM of each primer and 1 unit of recombinant Taq

polymerase (Thermo Fisher Scientific, Waltham, MA, USA). Primers for the PCR corresponded to the LTR specific primer for the band along with the EcoRI adapter primer (**Table 2.5**). Cycling conditions were as following: 94°C for 2 minutes followed by 35 cycles of 94°C for 30 seconds, 56°C for 30 seconds and 72°C for 2 minutes, and a final extension at 72°C for 10 minutes. The total PCR (25 µL) was run on a 1% agarose gel at 80V for 60 minutes and the bands were eluted using the Wizard SV gel and PCR clean-up system (Promega, Madison, WI, USA). Eluted products were quantified using a Nanodrop ND-1000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). For sequencing, 75 to 225 ng of the eluted product was used (depending on the band size) along with a primer at a final concentration of 0.25 µM (forward or reverse primers corresponded to the same primers used for PCR). Sequencing reactions were performed with the BigDye terminator v3.1 cycle sequencing kit (Applied Biosystems - Thermo Fisher Scientific, Waltham, MA, USA) using a 3730 Genetic Analyzer equipment (Applied Biosystems -Thermo Fisher Scientific, Waltham, MA, USA).

Sequences from eluted SSAP bands were compared to the flax genome deposited in phytozome [238,239] using blastn and Gbrowse to determine the insertion site of the TE. Once mapped on the genome, the IDs of the flax genes with associated TE insertions, were used to find the closest *Arabidopsis thaliana* ortholog from a database previously obtained by blasting flax transcripts against the peptide TAIR database (release 10). *Arabidopsis* ortholog IDs were then used to perform functional Gene Ontology (GO) classification using the GO annotation search from TAIR [244], to find out what categories of genes were predominantly affected by TE insertions. An additional enrichment analysis was performed using AgriGO [245,246] by comparing the *Arabidopsis* orthologs to the background of all *Arabidopsis* genes using a Fisher test, a Yekutieli multiple test adjustment and a minimum of 1 mapping read.

2.3.8 Validation of TE insertions

Once the TE insertions were mapped to the genome, primers were designed from the flanking regions of the transposable element insertion site to validate the polymorphism encountered with the initial SSAP gels. For the TEs that fell within genes we looked for paralogs and performed an alignment to select allele-specific primers which would not bind related genes. Twenty eight primers were designed with Primer3 [236,237] under the same parameters cited above to be compatible with the original LTR-derived primers, but with a product size range of

200-1000bp (**Table 2.6**), and were used to perform amplification in eight replicate plants in each one of the 14 cultivars. Five nanograms of DNA from each of eight samples per cultivar was used for PCR on 384-well plates using 1X Taq buffer, 2 mM of $MgCl_2$, 0.2 mM dNTPs, 0.2 µM of each primer and 1 unit of recombinant Taq polymerase (Thermo Scientific - Thermo Fisher Scientific, Waltham, MA, USA) in a 10 µL reaction. Cycling conditions were as following: 94°C for 2 minutes followed by 35 cycles of 94°C for 30 seconds, 60°C for 30 seconds (this temperature varied according to the primer used – see **Table 2.6**) and 72°C for 1 minute, with a final extension at 72°C for 5 minutes. Bands were visualized in 1% agarose gels run at 90V for 60 minutes.

Table 2.6 Primers used for the validation of the presence of the insertion extracted from SSAP profiles.

Primer	Sequence	Annealing temp. (°C) to run	Expected amplicon
val-RLC_Lu0-primer3-lutea-7	agttgaattctgaaatatatccac	54	463
val-RLC_Lu0-primer3-oleane-8	caaacatctggtgacttatctt	55	351
val-RLC_Lu0-primer3-lutea-12	ccagacattacagacaacaag	55	436
val-RLC_Lu0-primer3-blizzard-17	aaacactaagcaaccagag	54	314
val-RLC_Lu1-primer1-s.cirrus-18	caacgagggctgttcagt	58	395
val-RLC_Lu1-primer1-aurore-19	ctaatatgatgcactgaaccgc	58	759
val-RLC_Lu1-primer1-s.cirrus-23	aatcctgacagaaagaaccctt	58	575
val-RLC_Lu1-primer1-violin-26	ggtgcaattgtgtgtcttacc	58	856
val-RLC_Lu1-primer2-oleane-9	cattgaaaaagaaaccccaac	55	660
val-RLC_Lu1-primer2-lutea-12	gctgctcaaacttggtaaga	55	565
val-RLC_Lu1-primer2-blizzard-16	gcattgcaaacatcaaattcc	55	664
val-RLC_Lu1-primer2-oleane-18	catgatttgttggagcaagaa	55	409
val-RLC_Lu2-primer1-hermes-2	tgtaatggaagcctgccagc	60	411
val-RLC_Lu2-primer1-oleane-7	acggatattcaacgatttcaatag	57	557
val-RLC_Lu2-primer1-belinka-10	gaaatatacctgattccgcctg	58	640
val-RLC_Lu2-primer1-drakkar-13	cttgcaggttgtttgtcacc	59	762
val-RLC_Lu6-primer3-rdf-5	agcaaagtttgggatttctcaa	58	759

Primer	Sequence	Annealing temp. (°C) to run	Expected amplicon
val-RLC_Lu6-primer3-s.cirrus-8	atttcggaagcggaaccatc	59	463
val-RLC_Lu6-primer3-s.cirrus-9	gcaaacgctatgagtcagc	58	413
val-RLC_Lu6-primer3-lutea-16	ctgcctcgattcaagtcct	58	434
val-RLC_Lu8-primer1-oleane-7	cgttactattaccatcatcaccct	59	714
val-RLC_Lu8-primer1-oliver-15	tcttgtggcggactagaga	58	670
val-RLC_Lu8-primer1-drakkar-17	gatgcacaggtcagacgtt	58	954
val-RLC_Lu8-primer1-lutea-18	tgggaagatcaagctcaacc	58	539
val-RLC_Lu28-primer1-oliver-11	tcacgccacccaaggatt	60	580
val-RLC_Lu28-primer1-violin-12	cccctgctttcaataaaattcact	59	834
val-RLC_Lu28-primer1-lutea-14	agctcacgatgtcaaacagg	59	361
val-RLC_Lu28-primer1-oleane-20	acaaggtctttcattacagcaag	58	410

2.3.9 Expression of genes with TE insertions

We selected four genes to test their expression in five cultivars that were polymorphic for the respective TE insertion (**Figure 2.5**). The primer pairs per gene were named according to their gene of origin: Pyruvate carboxylase (PYR), Rabgap/TBC domain containing protein-1 (RAB1), Laccase-13-related (LAC), and Rabgap/TBC domain containing protein-2 (RAB2) (**Figure 2.5** and **Table 2.7**).

Table 2.7 Primers used to test expression changes of genes bearing TE insertions.

Gene name - ID	Name	sequence
Pyruvate carboxylase - Lus10022077	PYR-fw	ccacacttttgcttgaaatgataatg
	PYR-rv	tcgagttgaagttaatggatcgac
Rabgap/TBC domain containing protein - Lus10036500	RAB1-fw	gaaatctcctgctccaactgc
	RAB1-rv	tctggttgaaggttgaattgtgc
Laccase-13-related - Lus10026400	LAC-fw	gaacagccctcgttgccaa
	LAC-rv	gctgccataatcccctgctg
	RAB2-fw	aaggaaaccgtgtcatgctatttc

Rabgap/TBC domain containing protein - Lus10040349	RAB2-rv	caaatgttgaatctgccagtgat

cDNA from three tissues (leaf + apical meristem, stem and roots) from four biological replicates (different plants), was used to evaluate the primer pairs of each gene using qRT-PCR. Seven reference genes were tested for stability among tissues and replicates [228] (**Table 2.2**). While all seven genes were stable, the three with higher stability according to Bestkeeper [247] and GeNorm [248] were GAPDH (glyceraldehyde 3-phosphate dehydrogenase), ETIF5A (eukaryotic translation initiation factor 5 A), and EF1A (elongation factor 1-α). These were used to generate the geometric mean for relative quantification of the test genes using the ΔCt of the reference – the test gene. Statistical differences in each gene among cultivars were calculated using unpaired two-tailed *t*-tests after a Bonferroni correction for multiple comparisons using GraphPad Prism version 6.0 (GraphPad Software, La Jolla California USA).

Samples were aliquoted in 384-well plates (with three technical replicates per sample and tissue combination) using a Biomek 3000 Laboratory Automation System (Beckman Coulter, Brea, CA, USA), and the qRT-PCR was run using a QuantStudio 6 Flex Real-Time PCR system (Applied Biosystems-Life Technologies, Carslbad, CA, USA). Sample reactions were done in 10 µL with 5 µL of SYBR-green (Molecular Probes – Thermo Fisher Scientific, Waltham, MA, USA), 2.5 µL of the mixed primer pair (3.2 µM) and 2.5 µL of a 1:50 dilution of the synthesized cDNA. Cycling conditions were: 95°C for 2 minutes followed by 40 cycles of 95°C for 30 seconds, 60°C for 1 minute. A melting curve stage was added: 95°C for 15 seconds, 60°C for 1 minute and 95°C for 15 seconds.

2.4 Results

2.4.1 Comparison of TE copy number between flax accessions

To compare the abundance of TE families between flax cultivars, we designed reverse transcriptase (RT) primers (**Table 2.3**) from six selected Ty1-*copia* elements representative of six retrotransposon families, and used quantitative PCR (qPCR) to measure their abundance in 14 diverse flax accessions belonging to either oil or fiber, or spring and winter types (**Table 2.1**). The retrotransposon families were selected because previous analysis showed them to have LTRs with high similarity and conserved protein domains [218], suggesting the elements had been recently active, and may therefore be expected to be polymorphic between cultivars. Families

were named according to previously suggested conventions (see Methods [32]) as: RLC_Lu0, RLC_Lu1, RLC_Lu2, RLC_Lu6, RLC_Lu8 and RLC_Lu28. From each family a representative sequence which showed conserved sites for primer design among family members was selected (see additional selection characteristics of representative sequences in the Methods section). Four of the six representative sequences from the selected retrotransposon families had 100% similarity in their intraelement LTRs and two had LTRs over 99% similar, indicating insertion of these elements in the last 200,000 years (**Table 2.4**). Similarly, we identified the five expected protein domains from Ty1-*copia* elements in half of the representative sequences, and four domains in the other half (**Figure 2.1**).

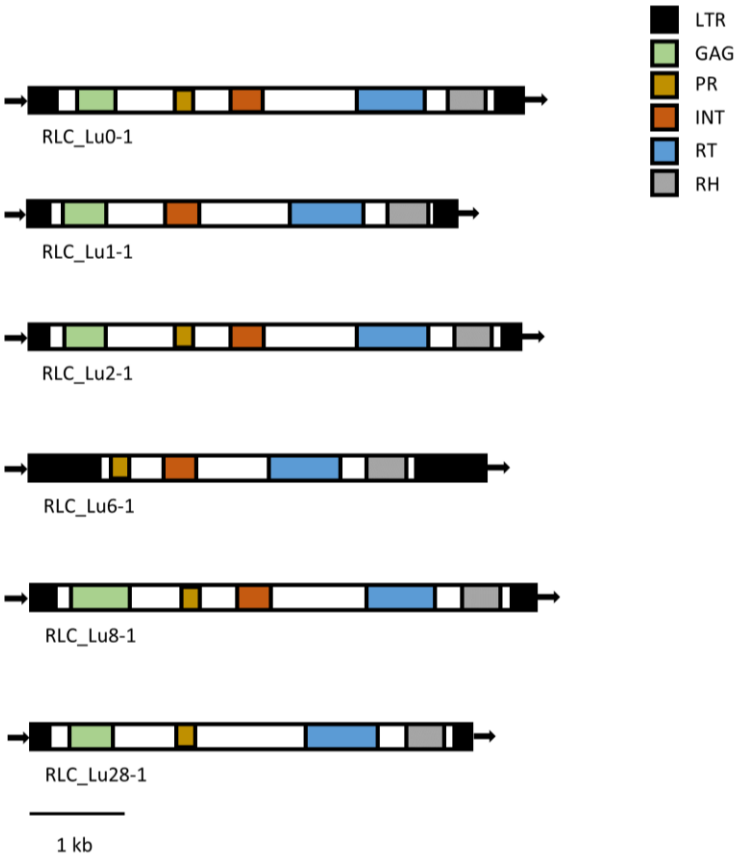


Figure 2.1 Diagrams of representative Ty1-*copia* TEs. Long terminal repeat (LTR), GAG domain (GAG), protease (PR), integrase (INT), reverse transcriptase (RT), RNase H (RH).

Quantitative PCR showed differences in TE family abundances (**Figure 2.2**). When averaged across cultivars, family RLC_Lu2 presented the lowest copy number per haploid genome (17.7), with families RLC_Lu1, RLC_Lu8 and RLC_Lu0 following with 22.5, 25.3 and

30.8 copies respectively. Finally, family RLC_Lu28 had 50.1 copies per haploid genome, while family RLC_Lu6 had the largest average copy number of all with 84.2 copies. For the CDC Bethune cultivar, which has an available whole genome assembly, the comparison of the quantitative PCR results with the expected copy number of TEs in each family estimated by BLAST alignments of the respective primer pairs (**Table 2.3**), showed a correlation of 0.85 demonstrating the validity of the quantitative PCR analysis. Non-parametric tests demonstrated that the variation in copy number for each family between cultivars was highly significant in all cases ($p \leq 0.0027$, Kruskal-Wallis test). Moreover, adjusted *p*-values (Dunn's test) for all pairwise comparisons showed significant differences between some accessions for all TE families tested with the exception of family RLC_Lu6 (**Figure 2.2**).

2.4.2 Identification of polymorphic TE insertions using SSAP

Having demonstrated significant variation in TE copy number between flax accessions, we used SSAP to identify individual insertions that were polymorphic between the accessions. We selected seven LTR primers that consistently amplified distinct bands from the same six Ty1-*copia* families used for copy number quantification (**Table 2.5**). Two primers were used for family RLC_Lu1 because they generated distinct patterns and resulted in additional polymorphisms. The seven primers were used to amplify DNA from each of the 14 flax accessions (**Table 2.1**), to generate the SSAP profiles (an example is shown in **Figure 2.3**). A total of 219 bands were scored, from which 140 were polymorphic (63.9% - **Table 2.8**). The

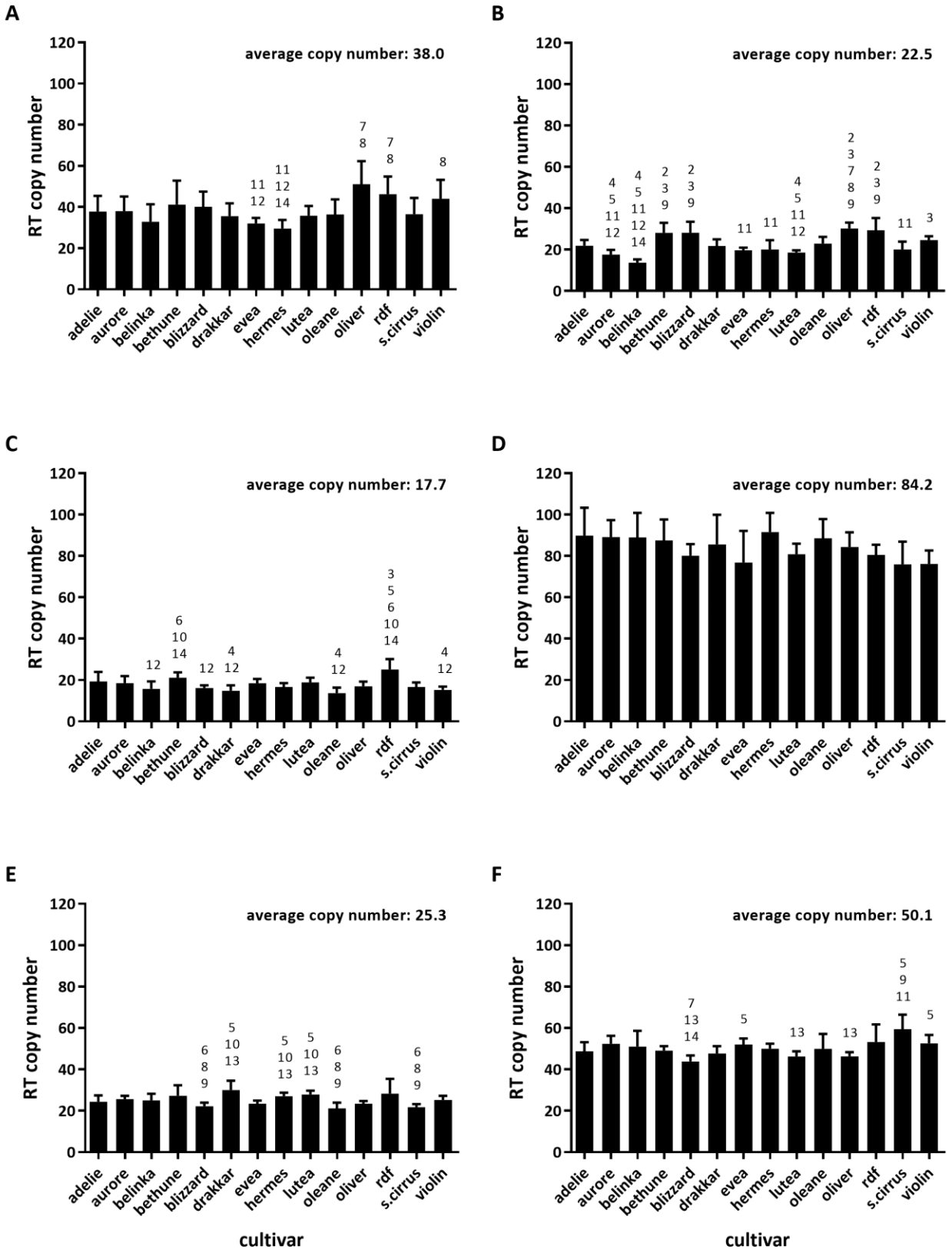

A
average copy number: 38.0
B
average copy number: 22.5
C
average copy number: 17.7
D
average copy number: 84.2
E
average copy number: 25.3
F
average copy number: 50.1
RT copy number
cultivar
adelie
aurore
belinka
bethune
blizzard
drakkar
evea
hermes
lutea
oleane
oliver
rdf
s.cirrus
violin

Figure 2.2 Absolute quantification of Ty1-*copia* retrotransposon families. The quantification was performed in 14 flax cultivars, based on amplification from their retrotranscriptase (RT) domains. The log10 of molecule copy number (mcn) was calculated using an online tool (see text) that accounts for plasmid+insert size. This value was used along Ct to generate standard curves to calculate molecule copy numbers for RTs, which were normalized to ETIF1 to find absolute copy number. Families depicted are: A. RLC_Lu0, B. RLC_Lu1, C. RLC_Lu2, D. RLC_Lu6, E. RLC_Lu8, F. RLC_Lu28. Error bars = standard deviation. Numbers above represent significant differences of the respective cultivar to other cultivars (Dunn's multiple comparison test $p \leq 0.05$) which are numbered as: 1. Adelie, 2. Aurore, 3. Belinka, 4. Bethune, 5. Blizzard, 6. Drakkar, 7. Evea, 8. Hermes, 9. Lutea, 10. Oleane, 11. Oliver, 12. rdf, 13. Stormont Cirrus. 14. Violin. The average copy number for all cultivars in each TE family is also indicated.

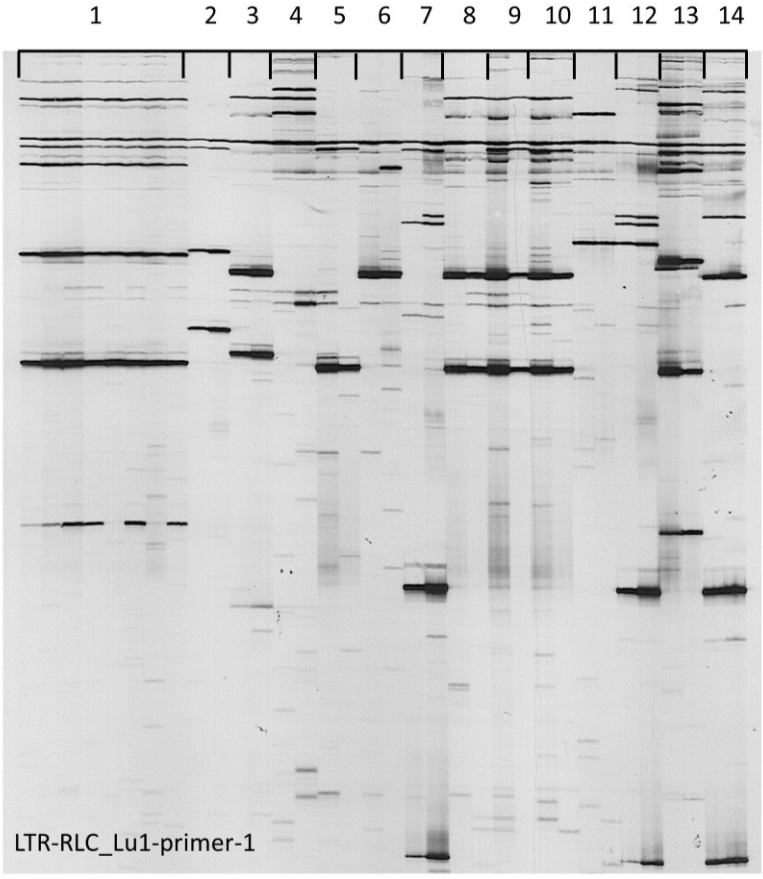


Figure 2.3 SSAP example of retrotransposon family RLC_Lu1. The SSAP was run for cultivars: 1. CDC Bethune, 2. Lutea, 3. Stormont Cirrus, 4. Adelie, 5. Aurore, 6. Belinka, 7. Blizzard, 8. Drakkar, 9. Evea, 10. Hermes, 11. Oleane, 12. Oliver, 13. rdf, 14. Violin.

primers with the lowest number of average bands were LTR-RLC_Lu1-primer1 and LTR_RCL_Lu1-primer2 (8.1 and 8.5 respectively); however, they also showed the highest rate of polymorphism (96.6% and 90% respectively). Conversely, LTR-RLC_Lu6-primer3 produced the highest number of bands across cultivars, with an average of 49.6, but showed the lowest rate of polymorphism (25%). The number of expected TEs from the CDC (Crop Development Center) Bethune genomic sequence analysis (**Table 2.3**) was also correlated ($r = 0.80$) with the number of scored bands in CDC Bethune (**Table 2.8**), showing consistency between methods.

SSAP bands were converted into a binary matrix (band presence = 1, absence = 0), which was used to construct a neighbor-net [243]. For the most part, oil (linseed)-types were more similar to each other than they were to fiber-types, with the exception of the winter fiber variety, Violin (**Figure 2.4**). A grouping pattern was also discerned for the dichotomy between spring and winter types, with the exception of Adelie, a winter fiber type which seemed closer to spring fiber types, and Lutea, a spring oil type which was closer to the winter oil types (**Figure 2.4**).

Table 2.8 SSAP band scoring and polymorphic bands.

Accession	LTR-RLC_Lu0-primer-3	LTR-RLC_Lu1-primer-1	LTR-RLC_Lu1-primer-2	LTR-RLC_Lu2-primer-1	LTR-RLC_Lu6-primer-3	LTR-RLC_Lu8-primer-1	LTR-RLC_Lu28-primer-1	Totals
rdf	18	12	13	12	50	8	20	133
Bethune	18	10	14	12	50	8	20	132
Lutea	14	4	6	12	48	8	20	112
Oleane	14	4	5	9	48	8	22	110
Blizzard	16	10	9	9	49	10	17	120
Oliver	17	10	10	11	49	8	20	125
Violin	16	12	11	12	49	10	24	134
Adelie	14	7	8	14	50	10	20	123
S. Cirrus	13	8	6	17	48	5	23	120
Evea	14	10	9	13	52	8	23	129
Drakkar	15	9	8	14	50	12	23	131
Hermes	14	8	9	13	50	8	24	126
Belinka	20	4	5	13	50	8	22	122
Aurore	18	6	6	11	51	9	23	124
Average number of bands	**15.8**	**8.1**	**8.5**	**12.3**	**49.6**	**8.6**	**21.5**	**17.8**
Scored positions	**28**	**29**	**30**	**22**	**56**	**22**	**32**	**219**
Polymorphic positions	**17**	**28**	**27**	**16**	**14**	**19**	**19**	**140**
%Polymorphism	**60.7**	**96.6**	**90.0**	**72.7**	**25.0**	**86.4**	**59.4**	**63.9**

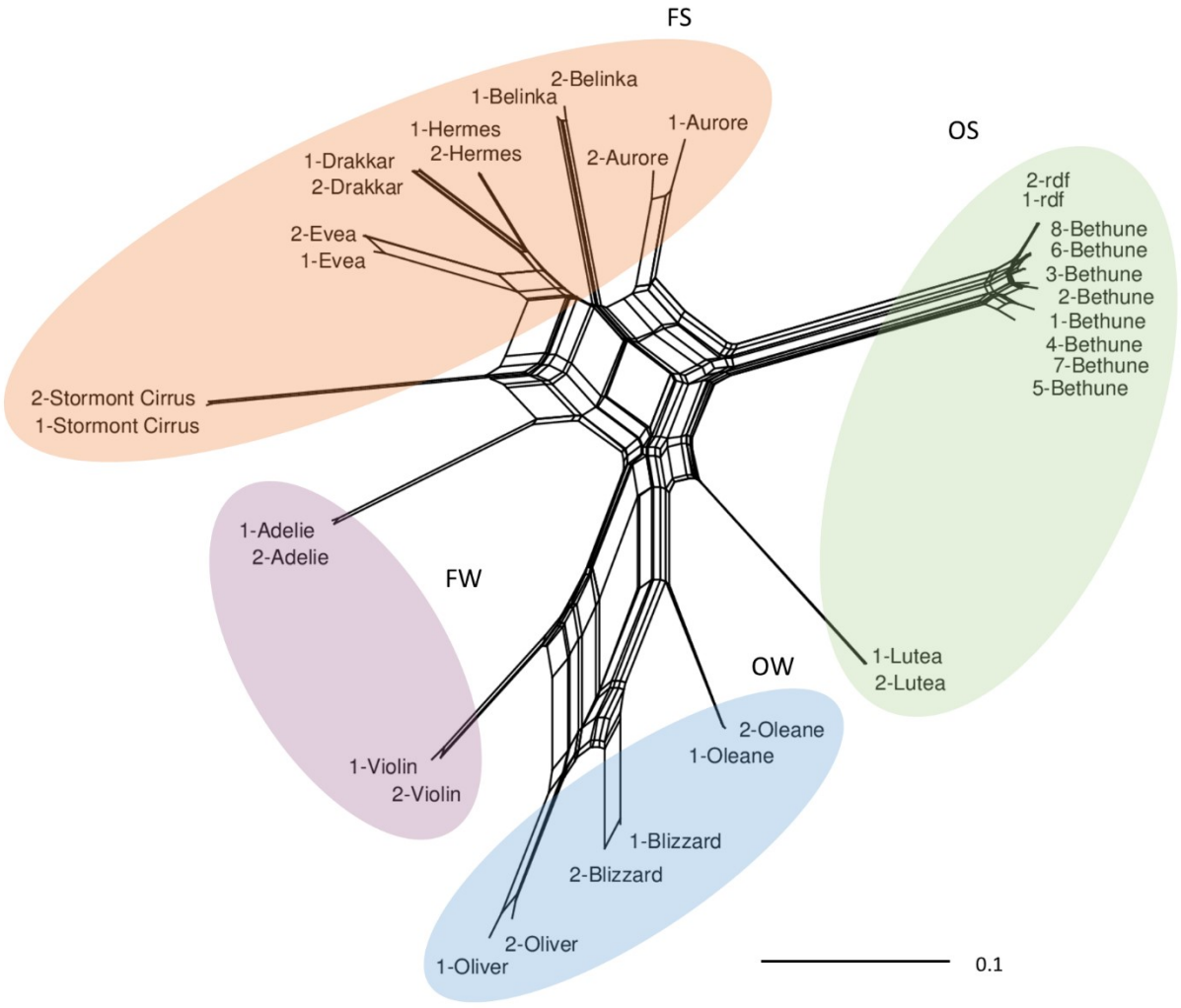


Figure 2.4 Neighbor net using 14 flax cultivars. Two biological replicates were used per cultivar with the exception of Bethune with eight replicates. The network was built using uncorrected p-distance. The colored groupings reflect different flax types: orange (fiber spring - FS), purple (fiber winter - FW), green (oil spring - OS), blue (oil winter - OW).

No single band could distinguish all linseed-types from all fiber-types, or all winter-types from spring-types, however, the definition of a cultivar as either a winter-type or spring-type can sometimes be ambiguous depending on the breeding program from which it originated [226]. Within the linseed-types, Bethune and *rdf* (reduced fiber) had the most SSAP sequenced bands in common, followed by Oliver and Blizzard (**Table 2.9**). Bethune and *rdf* had 4 bands which were only present in those two accessions: band 25 (LTR-RLC_Lu1-primer1), band 15 (LTR-RLC_Lu1-primer2), band 5 (LTR-RLC_Lu2-primer1), and band 4 (LTR-RLC_Lu8-primer1). At

the same time, Oliver and Blizzard had 2 bands that were exclusively in those two cultivars: band 17 (LTR-RLC_Lu0-primer3), and band 14 (LTR-RLC_Lu1-primer1).

In the case of the fiber-types, Evea, Drakkar, and Hermes shared the most bands. A band derived from LTR-RLC_Lu2-primer1 (band 10) was present in seven of the eight fiber-types tested and was absent from the linseed types (**Table 2.9**); the same was true for band 6 from LTR-RLC_Lu28-primer1 but the band was also present in *rdf* and Bethune (linseed types). One more band from LTR-RLC_Lu28-primer1 was common to eight fiber-types (band 12), although in this case, one linseed-type (Lutea) also had the band. One single band was present only in all spring-fiber types (LTR-RLC_Lu2-primer1, band 4), but this one was also present in Lutea (**Table 2.9**).

2.4.3 Analysis of flax genomic sequences targeted by polymorphic insertions

From the 140 polymorphic bands detected using SSAPs, 99 were successfully excised and sequenced. Most of the failed sequencing occurred with the highest molecular weight bands, which were either difficult to re-amplify or did not otherwise produce high quality sequences. Some of the resulting sequences were redundant, probably because the restriction enzyme used for the SSAPs found polymorphic sites in larger fragments, or because different cultivars had the same TE insertion, but accumulated mutations that generated a new restriction site, which resulted in bands with different electrophoretic mobility, but which represented the same insertion.

Table 2.9 Mapping of insertion sites of SSAP bands sequenced.

IDs		OS			OW			FW		FS						Annotation		
Copia family primer	band ID[a]	rdf	bet	lut	ole	bli	oli	vio	ade	sci	eve	dra	her	bel	aur	Phytozome match	Insertion site	Orientation gene / TE[d]
LTR-RLC_Lu0-primer3	5															Lus10014555 (NADH:ubiquinone reductase (non-electrogenic) / Ubiquinone reductase)	Intron 2	Same
	6															Lus10013587 (Protein virilizer homolog)	Intron 4	Same
	7															Lus10022077 (Pyruvate carboxylase)	Exon 8	Same
	8															Lus10036500 (Rabgap/TBC domain containing protein)	Intron 6	Same
	10															Intergenic	Intergenic	N/A
	11															Lus10043191 (Pyruvate dehydrogenase E1 component subunit beta, mitochondrial)	Intron 19	Opposite
	12															Lus10041231 (RRP12-like protein)	Intron 1	Opposite
	14															Lus10017623 (DNA ligase)	Intron 5	Opposite
	16															Lus10033880 (Transcription factor EMB1444-related)	Intron 5	Same
	17															Lus10009307 (Protein T01H10.8)	Intron 4	Opposite
LTR-RLC_Lu1-primer1	12															Intergenic	Intergenic	N/A

IDs		OS			OW			FW		FS						Annotation		
Copia family primer	band ID[a]	rdf	bet	lut	ole	bli	oli	vio	ade	sci	eve	dra	her	bel	aur	Phytozome match	Insertion site	Orientation gene / TE[d]
	13															Lus10033319 (Zinc finger, ZZ type (ZZ) // TAZ zinc finger (zf-TAZ) // Histone acetylation protein (HAT_KAT11))	Intron 12	Opposite
	14															Intergenic	Intergenic	N/A
	15															Lus10001114 (Alpha/beta hydrolases superfamily family)	Intron 3	Same
	16															Lus10016813 (Mitofilin).	Intron 4	Same
	18															Lus10026400 (Laccase-13-related)	Exon 3	Opposite
	19															Lus10040349 (Rabgap/TBC domain containing protein)	Intron 4	Opposite
	21															Lus10019489 (Telomere-length maintenance and DNA damage repair (TAN))	Intron 38	Opposite
	22															Lus10030711 (Spatacsin)	Intron 3	Same
	23															Lus10022840 (RNI-like superfamily protein)	Intron 7	Same
	24															The LTRs and an apparent degenerate internal TE region overlap a section from exon 3 to exon 5 of gene Lus10036612 (Formin-like protein 13)[c]	Exon 3	Opposite

IDs		OS			OW			FW		FS						Annotation		
Copia family primer	**band ID[a]**	**rdf**	**bet**	**lut**	**ole**	**bli**	**oli**	**vio**	**ade**	**sci**	**eve**	**dra**	**her**	**bel**	**aur**	**Phytozome match**	**Insertion site**	**Orientation gene / TE[d]**
	25															Lus10001216 (Cytoskeleton-associated protein 5), is 64 bp from TE[b]	Downstream of gene	Opposite
	26															Lus10034176 (anaphase-promoting complex subunit 5)	Intron 7	Same
LTR-RLC_Lu1-primer2	9															Lus10035905 (DEK protein)	Intron 3	Opposite
	12															Lus10025751 (Ubiquitin carboxyl-terminal hydrolase)	Intron 5	Same
	14															Lus10036170 (Tousled-like protein kinase)	Intron 15	Opposite
	15															Lus10026982 (Exocyst complex component 3)	Intron 19	Same
LTR-RLC_Lu2-primer1	2															Lus10030545 (CBL-interacting serine/threonine-protein kinase 2) is 5 bp from TE[b]	Downstream of gene	Opposite
	4															Lus10027426 (N-methylcoclaurine 3'-monooxygenase / N-methylcoclaurine 3'-hydroxylase)	Exon 2	Same
	5															Lus10040443 (Pinoresinol-lariciresinol reductase 3-related)	Exon 2	Same
	6															Intergenic	Intergenic	N/A
	7															Lus10020601 (DNA-directed RNA	Upstream of gene	Same

IDs		OS			OW			FW		FS						Annotation		
Copia family primer	band ID[a]	rdf	bet	lut	ole	bli	oli	vio	ade	sci	eve	dra	her	bel	aur	Phytozome match	Insertion site	Orientation gene / TE[d]
																polymerase II protein) is 295 bp from TE[b]		
	9															Lus10029082 (Uncharacterized protein) is 990 bp from TE[b]	Downstream of gene	Opposite
	10															TE is between genes Lus10035815 (PPR repeat (PPR) // PPR repeat family (PPR_2) // DYW family of nucleic acid deaminases (DYW_deaminase)) and Lus10035816 (RNA polymerase II transcription elongation factor Elongin/SIII, subunit elongin B) 178bp and 749 bp from them[b]	Upstream of two genes	TE opposite to first gene and in same orientation of second gene
	12															Intergenic	Intergenic	N/A
	13															Lus10001212 (Transcription factor Tfb2 (Tfb2) // Protein tyrosine kinase (Pkinase_Tyr))	Intron 2	Same
LTR-RLC_Lu6-primer3	5															Intergenic	Intergenic	N/A
	8															The complete 3' LTR and a partial degenerate 5' LTR from this element flank Lus10037467 (F-box domain	Containing one gene and downstream from another	TE opposite to gene inside and in same orientation

IDs		OS			OW			FW		FS						Annotation		
Copia family primer	band ID[a]	rdf	bet	lut	ole	bli	oli	vio	ade	sci	eve	dra	her	bel	aur	Phytozome match	Insertion site	Orientation gene / TE[d]
																protein), and a RNAse H domain was identified close to the 3' LTR; the element is also placed 13 bp from Lus10037468 (Domain of unknown function (DUF966))[b]		of second gene
	9															Intergenic	Intergenic	N/A
	13															The complete LTR overlaps intron 1 and exon 1 of gene Lus10013474 (Uncharacterized protein). The 5kb upstream of the 3'LTR give no indication of a complete TE so this could be a solo LTR[b,c]	Intron 1	Opposite
	15															Intergenic	Intergenic	N/A
	16															Intergenic	Intergenic	N/A
LTR-RLC_Lu8-primer1	4															Intergenic	Intergenic	N/A
	5															Lus10036983 (ATP dependent RNA Helicase)	Exon 1	Same
	6															Lus10020564 (Uncharacterized protein)	Exon 1	Same
	7															Lus10028760 (1-aminocyclopropane-1-carboxylate synthase 2-related)	Intron 2	Same
	9															Lus10030545 (CBL-interacting	Upstream of gene	Same

IDs		OS			OW			FW		FS						Annotation		
Copia family primer	band ID[a]	rdf	bet	lut	ole	bli	oli	vio	ade	sci	eve	dra	her	bel	aur	Phytozome match	Insertion site	Orientation gene / TE[d]
																serine/threonine-protein kinase 2) is 721 bp from TE[b]		
	10															Lus10016251 (TATA box-binding protein associated factor RNA polymerase I subunit B)	Unique exon	Opposite
	11															The 5'LTR overlaps with exon 3 of Lus10039295 (WRKY transcription factor 27-related)[c]	Exon 3	Same
	12															Lus10029998 (Clathrin coat assembly protein AP180)	Exon 1	Same
	13															Lus10009285 (GOS-28 Snare-related) is 7 bp from a partial section of TE LTR[b]	Upstream of gene	Opposite
	14															Lus10014756 (Polynucleotide 5'-hydroxyl-kinase NOL9)	Intron 9	Opposite
	15															Lus10031899 (Tetraticopeptide-like helical)	Exon 1	Same
	16															Lus10001449 (Glycerol-3-phosphate acyltransferase 2-related) is 349 bp from TE[b]	Downstream of gene	Same
	17															Lus10038912 (IRE Serine/threonine protein kinase)	Exon 3	Opposite

IDs		OS			OW			FW		FS						Annotation		
Copia family primer	band ID[a]	rdf	bet	lut	ole	bli	oli	vio	ade	sci	eve	dra	her	bel	aur	Phytozome match	Insertion site	Orientation gene / TE[d]
	18															Lus10033384 (Protein kinase domain (Pkinase) // Leucine rich repeat N-terminal domain (LRRNT_2) // Leucine rich repeat (LRR_8))	Exon 1	Same
LTR-RLC_Lu28-primer1	1															Lus10037058 (Neurolysin / neurotensin endopeptidase)	Exon 8	Opposite
	4															Lus10017405 (AAA-type ATPase domain-containing protein related)	Intron 16	Same
	6															Lus10041785 (Uncharacterized gene) is 574 bp from TE[b]	Downstream of gene	Opposite
	7															Lus10008548 (Variant SH3 domain protein (SH3_9)).	Intron 14	Opposite
	9															Intergenic	Intergenic	N/A
	11															Lus10001366 (Disease resistance protein related).	Intron 1	Same
	12															Lus10025655 (ARM repeat superfamily protein)	Intron 18	Opposite
	14															Lus10037943 (Limit dextrinase / R-enzyme)	Intron 5	Same
	20															Lus10001455 (Transcriptional regulator BRCA1)	Intron 7	Opposite

IDs		OS			OW			FW		FS						Annotation		
Copia family primer	band ID[a]	rdf	bet	lut	ole	bli	oli	vio	ade	sci	eve	dra	her	bel	aur	Phytozome match	Insertion site	Orientation gene / TE[d]
	22															The 3'LTR overlaps with exon 4 of Lus10008138 (Plant protein of unknown function DUF936). The 5'LTR is 53 bp from Lus10081137 (Uncharacterized protein)[b,c]	Upstream of gene	Opposite to both

[a]Refers to the number of the band identified. Sequences of the respective bands can be found in **Appendix 2.1**.
[b]When the TE was not inside an annotated gene the distance to the closest gene(s) was calculated. The distance was recorded if it was within 1kb of the gene. Distances of insertions not present in CDC Bethune (the reference genome) are inferred from the match of the flanking region.
[c]When the TE mapping overlapped with the phytozome annotation the insertion site was annotated as the 5'-most region of the TE.
[d]In this column N/A means non applicable.

After filtering for residual redundancies, sequences where the restriction site fell inside or just besides a TE, and sequences where no LTR could be identified, 66 unique insertion sites were found (**Table 2.9** and **Appendix 2.1**). Each insertion was classified according to its Ty1-*copia* family and was mapped to the genome assembly according to the annotation deposited in phytozome (**Table 2.9**). Of the 66 insertions, 14 (21.2%) interrupted annotated exons, 30 (45.5%) were in introns, 11 (16.7%) were within 1 kb of a gene opening reading frame (upstream or downstream), and 11 (16.7%) were characterized as intergenic (where the TE was inserted at a distance of more than 1 kb from any annotated gene). Altogether, more than 83.3% of the cloned TE insertions mapped within genes, or within 1 kb of a gene. For insertions in introns, exons, or within 1 kb of the CDS, the inferred transcription sense strand of the TE and gene were the same in 30 cases. Conversely, in 28 cases, the TE and associated gene were transcribed from opposite strands (**Table 2.9**).

To validate the results of the SSAPs, we conducted genomic PCR assays of 28 selected insertions (**Table 2.10**) on each of the 14 flax accessions. Because we had already sequenced and mapped the TE insertion sites, for each assay, one primer was designed to be complementary to the genomic DNA flanking the TE, and the second primer corresponded to the LTR primer used in SSAP for the respective family of the inserted TE. Nineteen (67.9%) of the insertions showed a perfect match or nearly perfect match of the polymorphisms initially assessed with SSAPs, while the rest had different levels of disagreement (**Table 2.10**).

Table 2.10 Comparison of selected SSAP band scores to PCR validation in 14 flax accessions. Selected SSAP band polymorphisms were selected for validation using conventional PCR (see Methods). 1 = present, 0 = absent, W = weak band at expected size, (?) = weak band at non-expected size, 1+L = expected band plus an additional lower band, 1+H = expected band plus an additional higher band, P = polymorphic among replicates of same cultivar.

			OS			OW			FW		FS						
Copia family primer	Band ID	Technique	rdf	bet	lut	ole	bli	oli	vio	ade	sci	eve	dra	her	bel	aur	Match[b]
LTR-RLC_Lu0-primer3	7	SSAP	0	0	1	0	0	0	0	0	0	0	0	0	0	0	PERFECT
		PCR	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
	8	SSAP	0	0	0	1	0	0	1	0	0	0	0	0	0	0	NO
		PCR	0	W	0	1	0	0	1	W	0	W	W	W	W	W	
	12	SSAP	0	0	1	0	0	0	0	0	0	0	0	0	0	0	PERFECT
		PCR	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
	17	SSAP	0	0	0	0	1	1	0	0	0	0	0	0	0	0	NEARLY
		PCR	0	0	0	0	1	1	P	0	0	0	0	0	0	0	PERFECT
LTR-RLC_Lu1-primer1	18	SSAP	0	0	0	0	0	0	1	0	1	1	1	1	1	0	PERFECT
		PCR	0	0	0	0	0	0	1	0	1	1	1	1	1	0	
	19	SSAP	1	P	0	0	0	0	0	1	1	1	P	0	0	P	NEARLY
		PCR	1	1	0	0	0	0	0	1	1	1	1	1	?	1	PERFECT
	23	SSAP	0	0	0	0	0	0	0	0	1	0	0	0	0	0	PERFECT
		PCR	0	0	0	0	0	0	0	0	1	0	0	0	0	0	
	26	SSAP	0	0	0	0	1	1	1	0	0	0	0	0	0	0	PERFECT
		PCR	0	0	0	0	1	1	1	0	0	0	0	0	0	0	
LTR-RLC_Lu1-primer2	9	SSAP	0	0	0	1	0	0	0	1	0	0	0	0	0	0	NO
		PCR	L	L	L	1+L	L	L	L	L	L	L	L	L	L	L	
	12	SSAP	1	1	1	0	1	0	1	0	0	1	0	1	0	1	NO
		PCR	1	1	1	0	1	0	1	0	0	0	0	1	0	0	
	16[a]	SSAP	0	0	0	0	P	1	1	0	0	0	0	0	0	0	PERFECT
		PCR	0	0	0	0	P	1	1	0	0	0	0	0	0	0	
	18[a]	SSAP	0	0	0	1	0	1	0	0	0	0	0	0	0	0	NEARLY
		PCR	0	0	0	1	0	0	0	0	0	0	0	0	0	0	PERFECT
LTR-RLC_Lu2-primer1	2	SSAP	1	1	1	0	0	0	1	1	0	0	1	1	0	0	PERFECT
		PCR	1	1	P	0	0	0	1	1	0	0	1	1	0	0	
	7	SSAP	0	0	0	1	1	1	1	0	0	0	0	0	0	0	PERFECT
		PCR	0	0	0	1	1	1	1	0	0	0	0	0	0	0	
	10	SSAP	0	0	0	0	0	0	0	1	1	1	1	1	1	1	PERFECT
		PCR	0	0	0	0	0	0	0	1	1	1	1	1	1	1	
	13	SSAP	1	1	0	0	0	0	0	1	1	0	1	0	1	0	PERFECT
		PCR	P	W	0	0	0	0	0	1	P	0	1	0	1	0	
LTR-RLC_Lu6-primer3	5	SSAP	1	P	0	0	0	0	0	0	0	1	1	1	1	1	NO
		PCR	1	?	P	?	0	1	P	0	P	1	1	1	1	1	
	8	SSAP	1	1	0	1	1	0	0	1	1	1	1	1	1	1	NO

		PCR	1	1	H	1+H	1+H	H	H	1+H	1	1	1	1	1	1	
	9	SSAP	1	1	0	1	1	1	1	1	1	1	1	1	1	1	NO
		PCR	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
	16	SSAP	1	1	1	0	0	0	1	1	1	1	1	1	1	1	NO
		PCR	P	1	1	1	1	P	P	1	P	1	1	1	1	1	
LTR-RLC_Lu8-primer1	7	SSAP	0	0	0	1	1	1	1	0	0	0	0	0	1	0	NO
		PCR	0	0	0	1	0	1	0	0	0	0	0	0	W	0	
	15	SSAP	0	0	0	0	0	1	1	1	0	0	0	0	0	0	NEARLY
		PCR	0	0	0	0	0	1	1	1	0	?	?	?	?	?	PERFECT
	17	SSAP	0	0	0	0	0	0	0	1	0	0	1	0	0	0	PERFECT
		PCR	0	0	0	0	0	0	0	1	0	0	1	0	0	0	
	18	SSAP	0	0	1	0	0	0	0	0	0	0	0	0	0	0	PERFECT
		PCR	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
LTR-RLC_Lu28-primer1	11	SSAP	0	0	0	1	0	1	1	0	1	0	0	0	0	0	NEARLY
		PCR	?	?	?	1	?	1	1	?	1	?	?	?	?	?	PERFECT
	12	SSAP	0	0	1	0	0	0	1	1	1	1	1	1	1	1	PERFECT
		PCR	0	0	1	0	0	0	1	1	1	1	1	1	1	1	
	14	SSAP	1	1	1	0	0	0	0	0	0	0	0	0	0	0	NO
		PCR	1	1	1	W	W	0	0	W	0	W	W	W	W	W	
	20	SSAP	0	0	1	1	0	0	0	0	0	0	0	0	0	0	PERFECT
		PCR	0	0	1	1	0	0	0	0	0	0	0	0	0	0	

[a] These two bands are not displayed in **Table 2.9** or **Appendix 2.1** since they were redundant to bands 13 and 15 from LTR-RLC_Lu1-primer1 respectively.
[b] Nearly perfect matches are defined as only having one mismatch, or perfect matches with additional weak bands at a different size from the expected band.

We used Gene Ontology (GO) categories to classify the genes that were found to be associated with polymorphic TE insertions. The genes represented 15 cellular components, 12 molecular functions, and 14 biological processes (**Table 2.11**). Nine genes where classified as responsive to stress, four in DNA or RNA metabolism, nine in cell organization and biogenesis, nine corresponded to protein metabolism, five were related to transcription, three to transport, five to development, six involved in signal transduction and one was related to electron transport or energy processes. None of the categories were enriched when the *Arabidopsis* orthologs were compared to the background of all annotated genes using AgriGO (data not shown).

Table 2.11 GO functional categories of flax closest orthologues in *Arabidopsis*. The flax gene sequences from Table 2.9 where used to search the *Arabidopsis* closest ortholog, and these were used for functional classification using Gene Ontology (see Methods).

Keyword Category	Functional Category	Annotation Count	Gene Count
GO Cellular Component	other cytoplasmic components	31	18
GO Cellular Component	other intracellular components	30	13
GO Cellular Component	nucleus	26	23
GO Cellular Component	other membranes	16	11
GO Cellular Component	mitochondria	10	8
GO Cellular Component	chloroplast	10	7
GO Cellular Component	plasma membrane	9	9
GO Cellular Component	cytosol	5	5
GO Cellular Component	Golgi apparatus	5	2
GO Cellular Component	other cellular components	4	4
GO Cellular Component	plastid	3	2
GO Cellular Component	extracellular	3	3
GO Cellular Component	cell wall	2	2
GO Cellular Component	unknown cellular components	2	2
GO Cellular Component	Endoplasmic reticulum	1	1
GO Molecular Function	other binding	24	21
GO Molecular Function	protein binding	16	11
GO Molecular Function	transferase activity	14	11
GO Molecular Function	nucleotide binding	11	11
GO Molecular Function	kinase activity	11	6
GO Molecular Function	other enzyme activity	9	8
GO Molecular Function	unknown molecular functions	9	9
GO Molecular Function	hydrolase activity	8	6
GO Molecular Function	DNA or RNA binding	4	4
GO Molecular Function	other molecular functions	2	2
GO Molecular Function	transcription factor activity	2	2

GO Molecular Function	nucleic acid binding	1	1
GO Biological Process	other cellular processes	64	27
GO Biological Process	other metabolic processes	50	24
GO Biological Process	response to stress	18	9
GO Biological Process	DNA or RNA metabolism	15	4
GO Biological Process	cell organization and biogenesis	13	9
GO Biological Process	protein metabolism	13	9
GO Biological Process	Transcription, DNA-dependent	11	5
GO Biological Process	other biological processes	10	7
GO Biological Process	response to abiotic or biotic stimulus	10	8
GO Biological Process	unknown biological processes	9	9
GO Biological Process	transport	9	3
GO Biological Process	developmental processes	7	5
GO Biological Process	signal transduction	6	6
GO Biological Process	electron transport or energy pathways	1	1

2.4.4 qRT-PCR analysis of selected genes with polymorphic TE insertions

To assess effects of TE insertions on gene expression, we selected four flax genes from **Table 2.9** that were expected to be constitutively expressed under normal conditions, so that any effects of TE insertion on gene expression could be detected. In making this selection, we relied partly on flax RNA-seq data of control plants from an experiment on the flax-fusarium interaction performed in our lab (Galindo-González & Deyholos, in preparation), and on comparisons to the presumptive *Arabidopsis* orthologs of our flax genes, since extensive transcript expression data is available for *Arabidopsis* (ThaleMine - [249,250]). Four flax genes were selected for qRT-PCR analysis: Pyruvate carboxylase (Lus10022077), a Laccase-13-related gene (Lus10026400), and two Rabgap/TBC domain containing proteins (Lus10036500 and Lus10040349). The two Rabgap/TBC domain containing proteins had 83.6% nucleotide identity, and bore the TE insertions in different regions (**Table 2.9**). Additionally, because of potential positional effects of the TE insertions, two TEs were inserted in exons and two in introns and

two TE-gene associations were antisense (**Figure 2.5**). Primers were designed downstream the insertion following the theoretical gene transcription orientation (**Figure 2.5**). Five flax accessions that were polymorphic for insertions in these four genes were selected to be assayed on three different tissues (leaves, root and stem) by qRT-PCR: Lutea (TE insertion in exon 8 of pyruvate carboxylase), Oleane (insertion in intron 6 of first Rabgap/TBC domain containing protein); Stormont Cirrus (TE insertion in exon 3 of the Laccase-13-related gene), and an insertion in a second Rabgap/TBC domain containing protein of intron 4, which was also present in Bethune) (**Figure 2.5**); and Oliver, which had no TE insertions in any these four genes. The results of the qRT-PCR analysis showed one relationship which was in agreement with the polymorphic pattern on insertion of a TE in a gene: the Laccase gene presented a significant decrease in root gene expression between Stormont Cirrus (which had the TE insertion), and the other four cultivars evaluated that did not bear the insertion (**Figure 2.6**). The three other genes did not show decrease in gene expression in the accession containing the TE insertion.

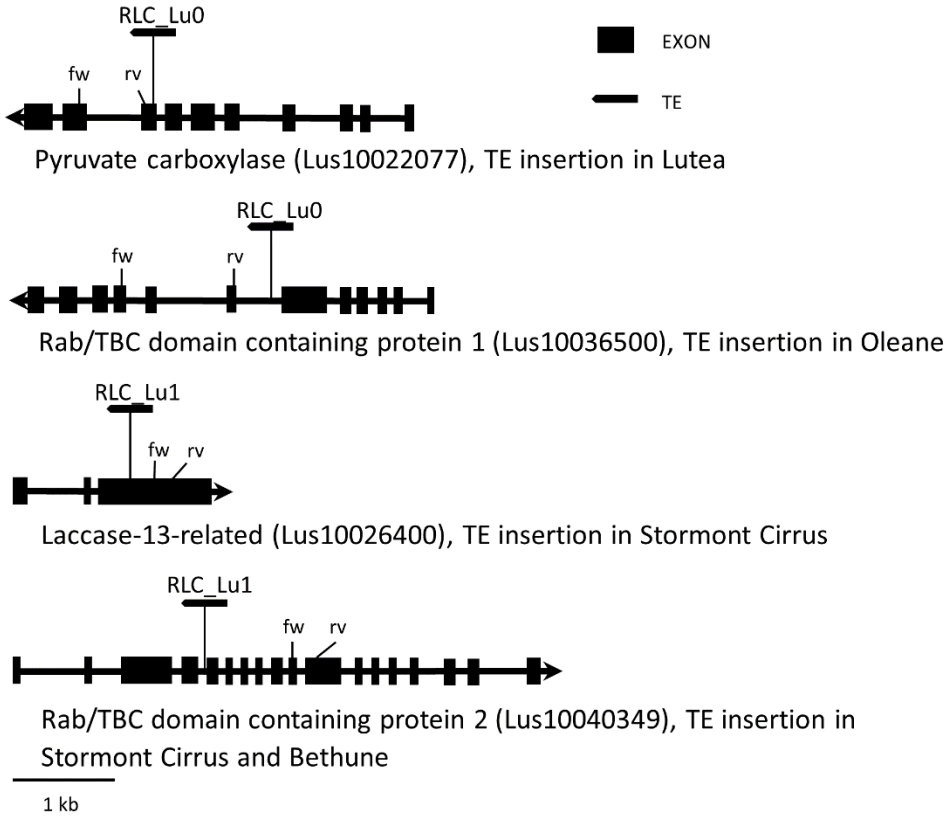


Figure 2.5 Diagrams of genes bearing Ty1-*copia* TE insertions. The location of the primers used to test for changes in gene expression is displayed. Gene expression was tested in five flax cultivars (Lutea, Oleane, S. Cirrus, Bethune and Oliver), which were polymorphic for the insertions. The name of the TE family is above each one of the represented TEs. Orientation of the genes and TEs is as depicted after mapping using phytozome. Genes are drawn according to scale while TEs (not to scale) are depicted only to show insertion sites.

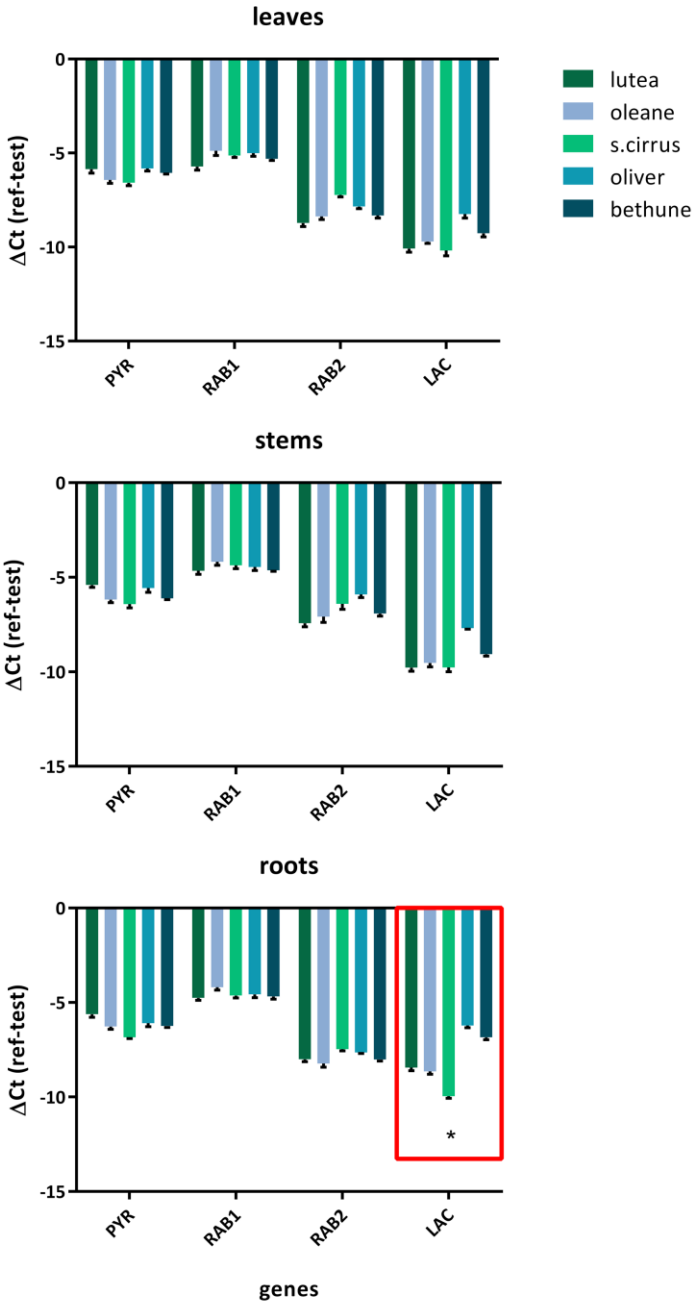

leaves
stems
roots
ΔCt (ref-test)
0
-5
-10
-15
PYR
RAB1
RAB2
LAC
*
genes
lutea
oleane
s.cirrus
oliver
bethune

Figure 2.6 Normalized gene expression of four genes bearing TE insertions in three different tissues. Each ΔCt corresponds to the average of four biological replicates, each with 3 technical replicates in the qRT-PCR. PYR (Pyruvate carboxylase – Lus10022077), RAB1 (Rabgap/TBC domain containing protein 1 – Lus10036500), RAB2 (Rabgap/TBC domain containing protein 2 – Lus10040349), LAC (Laccase-13-related – Lus10026400). All pairwise comparison in each gene and tissue, were performed using unpaired two-tailed *t*-tests, and significant differences were calculated after Bonferroni correction ($p < 0.005$). Only one primer in one tissue (red outline) showed a significant difference which agreed with the polymorphic insertion of a TE in a gene (*).

2.5 Discussion

2.5.1 TE activity and genomic copy number

Quantitative PCR using six Ty1-*copia* families in 14 flax cultivars demonstrated copy number variation between TE families, but also within each family between cultivars. Artificial selection through plant breeding involves subjecting plants to diverse stress conditions including testing plants in harsh conditions (e.g. drought, cold), and growing them under agricultural and laboratory practices which are not common in natural environments. Mobilization of TEs has accompanied the processes of breeding, as evidenced by the level of inter-varietal polymorphisms found using different transposon-derived markers [174,181,189,196,225].

A first clue that genomes diverge with respect to transposon history is a difference in abundance of specific TEs. Our approach to assess copy numbers of the selected TE families in flax followed a previous report that quantified the highly abundant *BARE*-1 retrotransposons from barley in several cultivars [173]. We found significant differences in copy numbers between cultivars for each TE family examined and differences when testing for multiple comparisons in five out of six TE families (**Figure 2.2**). One of the most extreme examples of genome diversification due to active TEs in plant genomes was demonstrated by amplification of the *mPing* MITE (Miniature Inverted-repeat TE) in different rice landraces where the element was actively transposing and ranged from 50 to more than 1000 copies [251]. An example of retrotransposon-related diversification includes variation in TE copy numbers among wild accessions and cultivars of sunflower, where a trend was found for cultivars to have a larger proportion of LTR-retrotransposons than wild accessions [252]. While the differences in abundance of the Ty1-*copia* families we tested in flax were not as large as reported for *mPing* in

rice, they were nevertheless indicative of recent activity of these TEs, and the potential for polymorphisms between cultivars.

In general, the absolute copy numbers we reported (**Figure 2.2**) probably underestimated the actual abundance of the Ty1-*copia* families, due to frequent recombination and high mutation rates expected among LTR elements [110]. This can result in modification of binding sites for qPCR primers or loss of internal retrotransposon domains, with concomitant creation of solo LTRs [109,144,240], which would not be accounted for by our method, based on amplification of the internal RT gene. Furthermore, the expected number of annealing sites from our BLAST analysis (**Table 2.3**) was almost always lower than the calculated copy number by qPCR (**Figure 2.2**) in CDC Bethune (this happens for 5 out of 6 families). This is probably a result of unassembled genome regions that are yet to be reported in the database (in general regions which are difficult to assemble are rich in repeats such as TEs). Nonetheless, the copy numbers we reported for each family for CDC Bethune (**Figure 2.2**) were correlated with the number of TEs identified by BLAST alignment of primer binding sites to the CDC Bethune genome assembly ($r = 0.85$), showing the validity of our approach. There was also a proportional high correlation between the TEs counted by BLAST alignment and the number of SSAP bands found for CDC Bethune ($r = 0.80$). In this case the number of SSAP bands for CDC Bethune (**Table 2.8**), was generally lower than the number of expected hits (**Table 2.3**), and calculated TE copy numbers (**Figure 2.2**) in each family, because this transposon display technique is only efficient for insertions located close to a restrictions site, and thus only reveals a subset of all insertions.

The highest estimated copy numbers and SSAP bands across all cultivars were found on family RLC_Lu6 (Compare **Figure 2.2** and **Table 2.8**). Interestingly, this family also had the lowest proportion of polymorphic bands (**Table 2.8**), and most of its flanking sequences were not related to gene regions (**Table 2.9**). An explanation for the insertion pattern and abundance of family RLC_Lu6 could be related to a lower level of negative selection, since its TEs inserted in regions of low gene abundance, may not be as detrimental for the genome, while the lower copy number found in other families could be related to purifying selection control by the host when TEs insert inside or close to genes, as has been seen in *Arabidopsis* [66]. Likewise, a lower level of polymorphism reflects inactivity, and it is therefore likely that family RLC_Lu6 expanded in the flax genome before breeding of these cultivars, and has been mostly quiescent since. In fact,

this family's representative sequence has the largest LTR divergence (**Table 2.4**), supporting this hypothesis.

Opposite to what happened with family RLC_Lu6, there are compelling clues that the low copy number families from our study were active in the recent past as demonstrated by the differences in TE copy numbers between cultivars (**Figure 2.2**), level of SSAP polymorphism (**Table 2.8**), their LTR similarity (**Table 2.4**), and that they relate more closely to genes (**Table 2.9**). This pattern is in agreement with low copy number TEs catalogued as being inserted closer to genes [253], and also being more active in recent past than high-copy number TEs in plants like maize [254,255]. Analysis of the maize genome suggests that the transition from low copy to high copy number TEs should be placed in the 10-100 copy range [253], which is in line with the difference between RLC_Lu6 and our low copy number families.

2.5.2 SSAP markers associate with flax types

SSAPs were performed with the same six families for which we measured TE copy number. Based on 140 polymorphic bands, we produced a neighbor-net, in which accessions showed associations which reflected the division between fiber and oil types, with the exception of Violin, a fiber winter type that was closer to the oil winter types than to its other fiber winter partner (Adelie) (**Figure 2.4**). Violin is a cultivar that behaves in the field like a dual purpose (oil/fiber) winter flax, and is genetically closer to an oil type, while Adelie has characteristics of a spring fiber (Jean-Paul Trouvé - personal communication). This would explain the close relationship of Violin to winter linseed types in the neighbor-net, as well as the close relationship of Adelie to the fiber spring types. Previously, the division of fiber and oil types has been supported by molecular studies looking at genes closely linked to the distinct phenotypes of these groups. For example, by using the *sad2* gene, which is involved in fatty acid metabolism, researchers were able to pinpoint a potential ancestral state of domestication of oil over fiber flax, and a relationship network where fiber flax was restricted to specific phylogenetic sections [256]. Additional candidate genes that can distinguish between fiber and linseed varieties were also found in a genetic diversity study by Soto-Cerda and collaborators [257]. Many of these genes were, as expected, closely related to fiber development (cell wall-related genes), or to the metabolism of oil production (fatty acid metabolism genes). The TE markers found in our study

are therefore a good source of molecular variation that can be linked to genes involved in the divergence of flax types (see more below).

2.5.3 Analysis of TE insertions and potential impact on genes

Characterization of insertion sites from polymorphic bands of the SSAPs in flax cultivars evidenced a high percentage of association of TEs to genes. Polymorphic TE insertions can result in genome divergence through genome restructuring, gene mutation and regulation changes (e.g. LTRs upstream of genes can change expression patterns), which at the same time depend on the TE's insertion site preference, and regulation of their transposition by host mechanisms (e.g. epigenetic control). Although disruption of gene function by transposition will often result in detrimental effects, some TE insertions can produce useful phenotypes in crops [26,258] or be co-opted by genomes to fulfill gene functions [259], and TE remnants can become important controlling factors in gene regulation [24,260]. While the mechanisms for insertion site selection are still not completely understood, insertional bias is evident for certain TE families. Young *Copia*-like retrotransposons have been shown to insert more randomly than *Gypsy*-like elements in *Arabidopsis*, and are associated with euchromatic gene-rich regions [66,261]. Similarly, we previously found that in flax, recently inserted Ty1-*copia* elements were non-randomly associated with gene regions and constituted the largest superfamily of TEs in the flax genome [218]. Our results here confirmed that numerous Ty1-*copia* TEs are biased towards insertion close to or inside genes. GO (Gene Ontology) classification (**Table 2.11**), however, showed no bias towards specific functional categories of genes.

TD has been often used to find intraspecies polymorphic markers to study genetic diversity [174,181,189,262,263], but these types of studies rarely characterize polymorphic insertion sites with detail at the sequence level. We successfully sequenced 66 non-redundant insertions in different genomic locations. Analysis of these insertion sites showed some interesting genes that could be related to agronomic traits, and represent potential candidates for future studies. For example, band 11 from LTR-RLC_Lu0-primer3 (**Table 2.9**) was characterized as a TE insertion on intron 19 of Pyruvate dehydrogenase E1. This gene is involved in fatty acid biosynthesis, and has been identified as potentially associated with divergent selection between flax types [257]. We also found a TE insertion on exon 2 of a Pinoresinol-lariciresinol reductase 3 gene (PLR3) for cultivars Bethune and the associated

mutant accession *rdf* (LTR-RLC_Lu2-primer1, band 5 - **Table 2.9**). It has been shown that PLR 1 is a key enzyme in flax lignan biosynthesis [264,265]. Flax seeds are rich in lignans, where they presumably act as antioxidants, as well as having beneficial effects on human health [266]. While the TE insertion found in our study matches PLR3, there is proof that different PLRs are needed for reduction of different pinoresinol enantiomers resulting in parallel lignan biosynthesis [267,268]. This marker is only 31kb from a region identified for divergent selection between linseed and fiber flax types in scaffold 86, containing five candidate genes related to this dichotomy (see Additional file 4 in [257]). Another interesting gene annotated as a Laccase-13-related, had a TE insertion on exon 3 of the gene (LTR-RLC_Lu1-primer1, band 18 - **Table 2.9**) that was present in six fiber cultivars. Laccases catalyze the last step in lignin biosynthesis [269], and downregulation of lignin biosynthetic genes (including laccases) has been associated with the hypolignification [270,271] that happens in bast fibers, and that makes peeling and harvesting of such fibers easier. A direct exon disruption by TE, as shown in our study, could prove relevant since it could be related to decreased transcript abundance of the disrupted gene, and contribute to this desirable characteristic of the fibers. Additionally, a recent study on the defense response of flax to *Fusarium oxysporum* f.sp *lini*, showed that several Laccase genes had increased transcript abundance in response to the pathogen (Galindo-Gonzalez & Deyholos, in preparation), and therefore these genes represent dual interest. Another gene with a TE insertion, was 1-aminocyclopropane-carboxylate synthase 2-related (LTR-RLC_Lu8-primer1, band 7 - **Table 2.9**), which was previously identified as Lu-ACS5 (1-aminocyclopropane-carboxylate synthase 5) [272]. ACS enzymes are involved in ethylene synthesis, and a previous study of ACS gene expression in flax roots showed that transcript abundance of ACS5 did not change in response to treatment with auxin antagonists, although transcripts of four other ACS genes did. Whether or not this might be related to the TE insertion is yet to be investigated. We also found a TE overlapping with a WRKY27 transcription factor (LTR-RLC_8-primer1, band 11 - **Table 2.9**). Mutants of this gene in *Arabidopsis* have delayed wilting upon infection with the bacterial pathogen *Ralstonia solanacearum*, showing that the gene might be a negative regulator of defense response [273], and multiple WRKY genes were found upregulated in response to *Fusarium oxysporum* f.sp. *lini* in flax (Galindo-González & Deyholos, in preparation). Therefore, this group of transcription factors are of interest in plant defense responses.

Our results showed that 83.3% of insertions disrupt exons or introns or are otherwise in close proximity to coding regions (**Table 2.9**). This means that these TEs could affect gene function by multiple mechanisms that could be tested in future studies. Fourteen of the characterized insertions in our study disrupted exons. While the most common result of an exon disruption is loss of gene function, this loss can result in a desirable agronomic trait. As an example, glutinous rice is the product of a retrotransposon disrupting an exon of the granule-bound starch synthase gene [126]. We also found 30 TEs mapped to introns. While many of the TEs in introns could be spliced out, previous studies have demonstrated their regulatory influence. An LTR-retrotransposon insertion in different introns of a MADS-box transcription factor of different apple varieties causes transcript suppression leading to seedless fruits [120] and waxy kernel phenotypes in maize result from alternative splicing patterns caused by retrotransposon insertions in introns of an amylose biosynthesis gene [112]. Regulation of expression can also result from TEs that do not disrupt the coding sequence; we found 10 insertions within 1kb of genes. Examples of the impact of extragenic insertions include: insertions of LTR retrotransposons adjacent to MYB genes involved in anthocyanin biosynthesis resulting in skin color variation in grape cultivars [274], and in the production of blood oranges [105]; insertion of a retrotransposon in the 5' UTR region of a vernalization gene (VRN1), which allows winter wheat to grow as a spring-type wheat [275]; and an increase in disease resistance to rice blast due to an insertion of an LTR retrotransposon in the promoter of the *Pit* resistance gene [108]. Finally, transposable element insertions can cause gene silencing, due to small RNA driven methylation of the mobile element [276]; this process is common with antisense transcripts [114,277], and in our case there were 28 insertions in opposite orientation to the associated gene. TE-mediated epigenetic gene silencing which extends to genes in the vicinity has been proven only when TEs are inside or very close to genes [131], and has been presented for *Arabidopsis* [116,117].

We found a particular example of a TE carrying an F-box domain protein between its two LTRs. This TE (band 8 from LTR-RLC_Lu6-primer3 in **Table 2.9**) has a recognizable RNAse H (ribonuclease H) domain near the 3' LTR and therefore represents a functional retrotransposon that has acquired a gene. This gene capture and capacity to mobilize the gene is known as transduplication, and has been widely seen with over 3000 Pack-MULEs (Mu-like Elements) in rice that have captured over 1000 genes [75]. Furthermore, retrotransposons in rice and sorghum

have also been shown to capture numerous genes [129], making these TEs key players in the process of gene evolution. In the rice study it was shown that the captured genes could either be under purifying or positive selection, so each case of gene capture could represent a different evolutionary pattern.

2.5.4 TE impact on flax gene expression

A preliminary assay was performed to test the effects of flax TEs on gene expression, using qRT-PCR on four genes with insertions in either exons or introns. Only the Laccase gene demonstrated significant decreased transcript abundance in roots, correlated to the presence of the TE insertion: the cultivar Stormont Cirrus with the TE insertion had lower relative transcript abundance than the other cultivars devoid of the insertion (**Figure 2.6**). Observation of the qRT-PCR bands on a gel (not shown), demonstrated that the expected band was present in all cultivars, which would mean that likely scenarios for repression would be: i) anti-sense gene transcripts generated from readout of the TE inserted in opposite orientation of the gene (**Figure 2.5**), which could potentially be used for the generation of small RNAs tagging the gene for inactivation via methylation [114,115], or ii) the generation of a different splice form, which conserves the exon tested by qRT-PCR, but has reduced transcript abundance as a consequence of the modification [112,126].

For the pyruvate carboxylase gene, we expected that the TE insertion in exon 8 would result in transcript alterations for Lutea but this was not the case, and inspection of the qRT-PCR products in a gel indicated the presence of the expected band in all tested cultivars (not shown). For both of the Rabgap/TBC domain genes which had insertions on introns, no impact on gene expression was found.

These results showed no common mechanisms by which these insertions may alter gene expression. Insertion in exons would be expected to be directly disruptive but only in one of two cases a change was noticed. Opposite orientation of TE-derived transcripts could create epigenetic-mediated gene silencing [114,115], but this should depend on actual transcriptional activation of our TEs which might not be happening under our conditions. And TEs inserted in introns could change gene expression or splice forms, but this does not always happen, and alternate transcripts can be created from one single gene with an insertion [112]. Finally, if the

TE insertion is present in just one allele of the gene, the TE effects can be masked by the other allele functioning normally.

In future studies, stress conditions or treatments which upregulate genes with inserted TEs might prove to be a better strategy to discern if gene expression levels are affected by the TE, for two reasons: i) the stress can generate a higher response of the host gene that can be more distinct than a low constitutive expression if the TE really alters gene expression, and ii) the stress may also upregulate the transcription of the TE, increasing the chance of readout transcripts that can be used for the production of small RNAs that can mediate silencing. These examinations should be coupled with experiments: i) to assess the production of small RNAs and methylation state of the gene, and ii) revise if TE insertions are homozygous and if different splice forms are produced from the host gene.

2.6 Conclusions

Based on our findings, we can conclude retrotransposition events have occurred since breeding for the tested cultivars began. The TE markers found using SSAPs were useful to distinguish the major flax types, and their level of polymorphism further showed that they have an impact on diversifying flax cultivars. While not all flax TEs examined fall in gene-rich regions [218], we now know that the transposition of most studied families here is biased towards these regions and their study constitutes a good source of novel mutations that can be used to find potential linkage to diversifying phenotypes, which is the basis for creating new cultivars. In fact, strong proof of TE-mediated diversification exists in closely related species of *Arabidopsis* [117] and rice [110] and in cultivars of rice [251] and maize [278]. No matter what the adaptive fate of these insertions may be, the mobilization of TEs among flax cultivars constitutes a powerful tool in diversity studies. However, understanding how these insertions influence genome restructuring and shape gene evolution requires studying related cultivars and species to determine what insertions may be under purifying selection and which ones are being positively selected as part of the normal functioning of the genome. The TE insertions found here, in different gene regions and in different orientations, open the door to study their potential influence on gene regulation on a case by case basis.

CHAPTER 3 – Activation of TEs by stress

3.1 Abstract

Transposable elements (TEs) can be activated by a plethora of elicitors. Their activation depends on motifs embedded in their promoter regions, and also on the level of accumulated mutations and their local genomic context and epigenetic repression.

In a comparison of flax cultivars, polymorphisms in Ty1-*copia* retrotransposon insertion sites were demonstrated (Chapter 2), indicating that specific Ty1-*copia* families are active in this species. Finding elicitors that activate specific TEs could explain the diversification processes, and point to the TEs that play a current role in genomic changes.

Here we evaluated the dependence of Ty1-*copia* transcription on fungal elicitors, wounding and tissue type. Neither wounding of leaves, nor treatment of either leaves or excised shoots with fungal extract, significantly increased activity of the TEs assayed, although constitutive expression of some TE families was detected. However, we observed differences in expression of Ty1-*copia* families between plant tissues in three cultivars. Finally, a large-scale transcriptome study indicated that Ty1-*copia* TEs were not differentially expressed upon *Fusarium oxysporum* inoculation, but numerous TEs were constitutively expressed.

While most TE families evaluated did not seem to respond to most stress factors applied, the variability in their regulatory motifs, expression in different organs, and constitutive expression give clues about which TEs might be playing a role in flax genomic modification and diversification.

3.2 Introduction

Diverse stress factors activate TEs (see section 1.2.2 from Chapter 1). For example, after tissue culture, a large amount of somaclonal variation can be detected, and many of the molecular changes that result in phenotypic variation are due to TE activation [56,225,279–281]. Although many elicitors activate various families of TEs in diverse species, no correlations have been established between specific stresses and the activation of specific TE families.

LTR-retrotransposons have transcription factor binding sites (TFBS') in their promoter regions, i.e. within the LTRs. TFBS motifs can vary even between closely related TE families giving rise to different expression patterns, depending on the stress [45]. The study of LTRs in retrotransposons has shown their TFBS' respond to numerous biotic and abiotic stresses [46,47,49,50,52,53,55,92].

The activation of TEs can also be related to changes in the methylation status of the genome during events like tissue culture, which alters methylation patterns of TEs, allowing them to be transcribed and transposed [57,282]. The same elicitor may not activate the same TE family in two different species under tissue culture because activation of TEs not only depends on the controlling regions embedded into the TE sequence, but also on the genomic context, and on the level of degeneration of both the promoter sequences (e.g. LTRs) and the internal protein-coding domains (the latter allow the TE to complete its transpositional cycle). For example if a TE falls into heterochromatic regions, as is usual with many *gypsy*-type retroelements [66], it is more likely that the genomic context of epigenetic silencing is also maintained in the TE, and that the rate of mutation is higher. And while TEs falling close to genes can also be silenced by epigenetic mechanisms, their integration in promoter regions, introns or even exons represents a possibility of directly altering phenotype through gene regulation or modification. For example, when methylation is lifted upon stress, TEs close to genes can be transcribed, resulting not only in possible transposition, but also in potentially TE-directed regulation of adjacent genes (see sections 1.2.3.1 and 1.2.3.2 from Chapter 1). This demonstrates that the TE insertion context can affect the transcription of the TE, but the TE also affects the transcription and regulation of its context. Because every species, and even populations and individuals within a species, have different histories of activation and movement of TEs, even when TE families are common, their activation may differ upon the same elicitor.

To identify the elicitors that trigger TE movement in a new plant system, stresses that activate different TE families in other species can be tested. In crops, the response of plants to biotic and abiotic stresses that can damage crops are of interest to breeders. Microbial elicitors and wounding can trigger a stress defense response in plants but can also trigger activation of TEs [45,55,62,64,65,77,79,92,283].

Here, we test a fungal extract (onozuka) which acts as a cellulase (microbial elicitor), scarification treatments (wounding), differential tissue response, and inoculation with *Fusarium oxysporum,* to study the transcriptional responses of flax Ty1-*copia* elements to these stresses; these treatments have been shown to activate the expression of retrotransposons [48,52,55,64,65,92]. The TE families selected include some of the families studied in parallel for cultivar insertional polymorphism in Chapter 2, and comprise good candidates for activation, due

to high similarity of their LTRs, which indicates recent insertion/activity, and conserved protein domains (**Table 2.4**).

Our results did not show a consistent pattern of activation in response to fungal extracts or wounding, but the transcript abundance of some of the TE families examined was constitutively high. Since LTRs flanking retrotransposons act as promoters and contain transcription factor binding sites, it is very possible that besides motifs for stress response, there are tissue-specific controlling regions. This led us to hypothesize that their expression might be tissue-specific. We measured differential expression of most TE families examined in a comparison of samples derived from three different parts of the plant: shot apices (including young leaves); stems; and roots. No differential expression was found in response to *Fusarium oxysporum* treatment.

The experiments presented on this Chapter confirm that specific Ty1-*copia* families are active in flax, and some can be activated in a tissue-specific manner. Therefore, their effects on genome remodelling could depend not only on environmental influence but also on developmental stages.

3.3 Materials and Methods

3.3.1 Prediction of transcription factor binding sites (TFBS') in LTRs

Long terminal repeats (LTRs) from selected TE families were entered into the plant promoter analysis navigator: PlantPAN [284,285], to find motifs with similarity to the database of TFBS' in *A. thaliana*. Matches with over 70% similarity were tabulated [286,287].

3.3.2 Experimental overview and plant material

Four experiments were conducted to assess TE gene expression: in the first experiment onozuka fungal extract was applied through cut stems to shoot segments; in the second experiment, onozuka and scarification were applied to detached leaves; in the third experiment young leaves (including shoot apices), stems and roots were compared for TE differential expression. Experiments 1 and 2 were evaluated with end-point RT-PCR and experiment 3 was evaluated with quantitative RT-PCR (qRT-PCR). The fourth experiment was a full-scale RNA-seq study with a methodology entirely detailed in Chapter 4.

For the first experiment with onozuka extract, flax plants from cultivar CDC Bethune were grown under greenhouse conditions (14 hours of light, 24°C day / 20°C night, 40% humidity) at the National Institute for Agronomic Research (INRA) in Versailles, France. CDC Bethune seeds were sown in pods with a 50/50 soil/sand mix, and were harvested after two weeks of growth and roots were taken of plants before aerial sections (stem and leaves) were immersed in 10 mL tubes bearing 8 mL of either a 1mg/mL onozuka solution (treatment) or water (control) (**Figure 3.1**). Three replicates were used per treatment and per control, and plants were harvested at 2, 4, 8 and 24 hours after being placed in solution. Samples were instantly frozen in liquid nitrogen. Extra replicate controls were taken at time 0 without immersion in any solution.

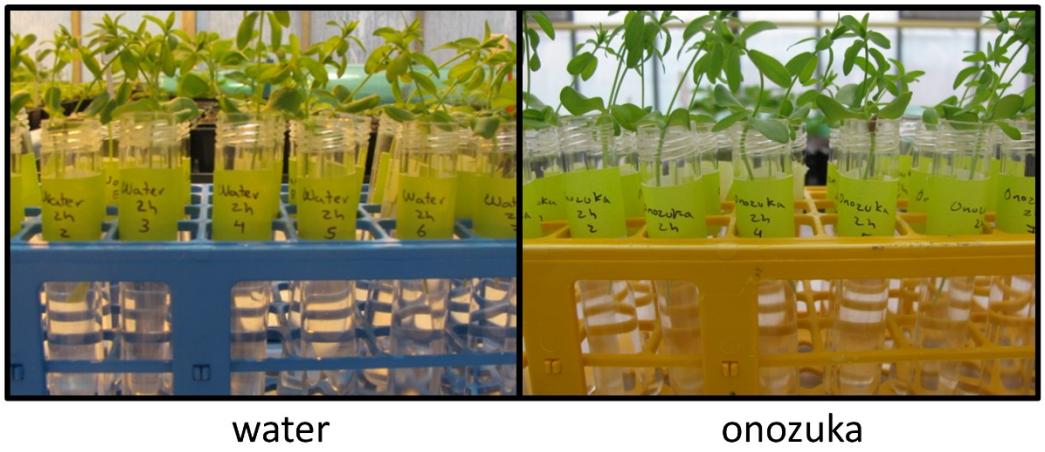


Figure 3.1 Experimental setup of aerial sections of flax plants in tubes containing a fungal extract (onozuka) or water.

For the second experiment, flax plants were grown as in experiment 1. Fifteen leaves were collected from the upper half of a one-month old plant and used for three treatments (5 for water control, 5 for onozuka and 5 to be scarified and placed in water). A total of four biological replicates (different plants) were used per treatment and time point (2, 4, 8 and 24 hours, plus control at time 0). Detached leaves were placed in 5 cm diameter petri dishes with 8 mL of either onozuka (1 mg/mL) solution, or water (for water control and scarified plants) (**Figure 3.2**); scarification was produced with a needle across all leaf surfaces. Four replicates were used per treatment and per control, and leaves were collected at 2, 4, 8 and 24 hours after being placed in solution. Samples were instantly frozen in liquid nitrogen.

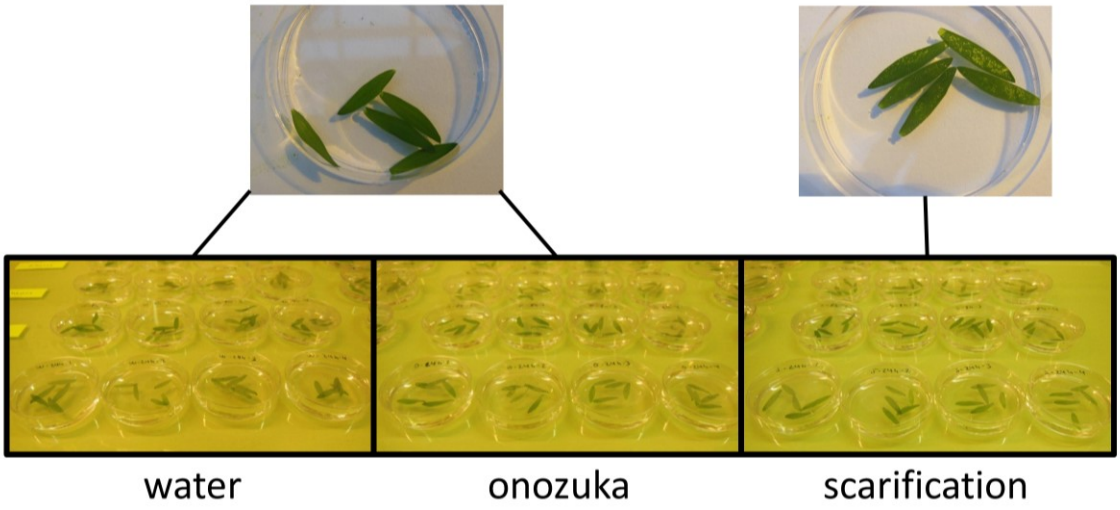


Figure 3.2 Experimental setup petri dishes containing leaves in a fungal extract (onozuka), or normal or scarified leaves in water. Pictures on top show how the scarified leaves differ from leaves placed directly on water or onozuka.

For the third experiment, plants from three different cultivars (CDC Bethune, Oliver and Stormont cirrus) were grown in a 50/50 soil-sand mixture in a growth chamber with 16 hours of light / 8 hours of dark, 22°C, 50% humidity, 0.132 µmol/m^2/s light). Four replicates of stems, 5-10 young leaves (including the apical meristem) and roots were harvested after 2-3 weeks of growth, and instantly frozen in liquid nitrogen. The cultivars were selected for the comparison of the TE families based on the distance reflected by SSAPs (Chapter 2), where both Oliver and S. Cirrus are separated from the elite cultivar CDC Bethune (Figure 2.4). The fourth experiment was carried out as described in Chapter 4.

3.3.3 Nucleic acid extraction and cDNA synthesis

Samples were either ground with plastic pestles in the sample tubes (experiments 1 and 2) maintaining the tubes in liquid nitrogen, or using an autoclaved 5.6 mm stainless steel bead placed in the sample tube, and using a Retsch MM301 mixer mill (Retsch, Haan, Germany) with two cycles of 1 minute at 20 Hz (experiments 3 and 4). RNA was extracted using an RNeasy Plant Mini Kit (QIAGEN, Venlo, The Netherlands). A DNAse treatment was performed for 30 minutes at 37°C with DNaseI (Thermo Scientific, Waltham, MA, USA). Quantity of obtained RNA was measured using a Nanodrop ND-1000 spectrophotometer (Thermo Scientific,

Waltham, MA, USA) and sample quality was assessed with a 2100 Bioanalyzer (Agilent, Mississauga, ON, Canada).

For experiments 1 and 2, equal amounts of RNA from individual replicates were pooled before cDNA synthesis, prior to end-point RT-PCR. Replicates from experiment 3 were left separated for qRT-PCR. One microgram of RNA (experiments 1 and 2) and 500 ng (experiment 3) were used with the RevertAid H Minus Reverse transcriptase under the manufacturer specifications and using oligo dT (18-21) (Thermo Scientific, Waltham, MA, USA). Experiment 4 details on nucleic acid extraction and pooling are displayed in Chapter 4.

3.3.4 Primers

Primers were designed either for end-point RT-PCR or for qRT-PCR. The former were designed for longer amplicons that could be easily resolved in standard agarose gels, while the latter where designed to produce smaller amplicons required for qRT-PCR. We used Primer3 (http://bioinfo.ut.ee/primer3/) (Untergasser et al., 2012) to design primers for end-point RT-PCR with the following parameters: primer size range between 18 and 24bp, temperature between 57 and 63 degrees, product size 200-300 bp, and GC content between 40 and 60% (the rest of the parameters were left by default). For qRT-PCR we changed the product size to 50-160 bp. Designed primers were initially tested on genomic DNA to assess specificity. Amplifications with a gradient PCR to determine the most suitable temperature for subsequent experiments yielded the expected amplicon size in all cases, and showed that 61°C could be used for end-point RT-PCR (Appendix 3.1).

For end-point RT-PCR of experiments 1 and 2, primers were designed from six chitinases (**Table 3.1**) from different classes within family 19 of glycosyl hydrolases, and one from family 18 (**Figure 3.3**). The chitinases were aligned to their closest cluster members using AlignX from the VectorNTI platform (Invitrogen, Carlsbad, CA, USA), and primers were designed to exclusively amplify the selected gene.

For end-point RT-PCR of experiments 1 and 2, three reference genes (ETIF3E, GAPDH and UBI2 – **Table 3.1**) were selected from a list of previously characterized genes suitable for quantitative gene expression normalization in flax [228], and primers were designed *de*-novo since amplicons for end-point RT-PCR were expected to be larger than for qRT-PCR. For qRT-PCR of experiment 3, seven primers were tested from the list of genes used for normalization

[228], and the three most stable primers: GAPDH, EF1A and ETIF5A (**Table 3.2**), were selected based on analysis carried out with GeNorm and BestKeeper [247,248].

For testing Ty1-*copia* elements in experiments 1 and 2 we selected five families of TEs (RLC_Lu0, RLC_Lu1, RLC_Lu2, RLC_Lu3 and RLC_Lu28) which had a representative sequence with 100% LTR similarity (the family and representative sequence designation are explained in section 2.3.3 of Chapter 2), which assumes recent activity of the element, and therefore increases the probability of transcriptional activation by an elicitor. Members from each cluster (family) were also aligned as was done for chitinase sequences, but primers were instead designed from conserved regions to test the expression of the family instead of an individual TE. We designed six primer pairs for end-point RT-PCR (two primers were designed for RLC_Lu0 – **Table 3.1**). All primers were designed in the conserved retrotranscriptase (RT) region of the TEs. For qRT-PCR from experiment 3 (organ comparison), we used primers (**Table 3.2**) from all families evaluated in Chapter 2.

The methodology to obtain the qRT-PCR primers for validation of RNA-seq from experiment 4 is detailed in Chapter 4.

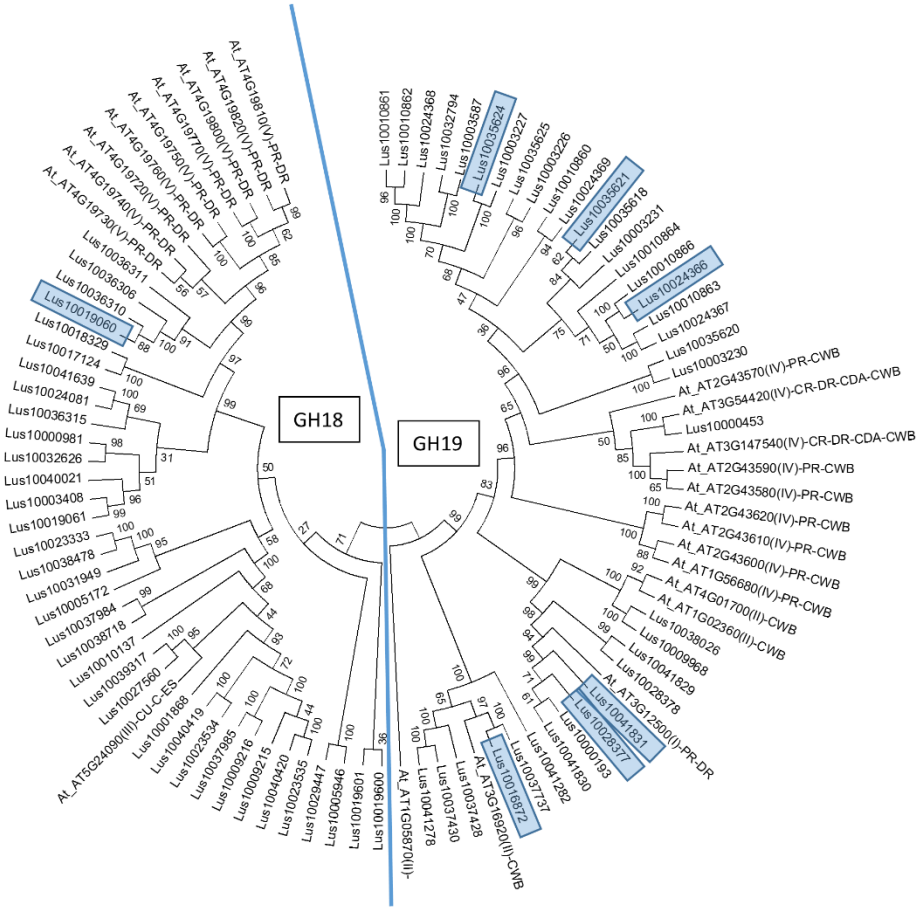


Figure 3.3 Relationship of flax chitinases with previously characterized *Arabidopsis* chitinases. The tree was done with amino acid sequences using a Muscle alignment (default parameters) and the dendrogram was built under the following parameters: neighbor joining (NJ), 1000 bootstrap replicates (each branch shows final support %), p-distance and pairwise deletion. Predicted function was taken from: *Arabidopsis thaliana*: A Genomic Survey [288]**: C – cytokinesis, CDA – cell death and aging, CR – cell rescue, CU – carbohydrate utilization, CWB – cell wall biogenesis, DR – defense-related, ES – extracellular secretion, PR – pathogenesis-related. Selected chitinases are outlined in rectangles and their respective primers were designed de-novo to test end-point RT-PCR (Table 3.1) and qRT-PCR (Table 3.2). Chitinase classes are in parentheses. GH = glycosyl hydrolase family.**

Table 3.1 End-point RT-PCR primers used in experiments 1 and 2.

Primer name	Sequence	Gene annotation
Lus10019060_ep_fw	ttttgtcgatcggaggcggt	chitinase
Lus10019060_ep_rv	taggagtagaggtggcttgc	chitinase
Lus10016872_ep_fw	gcgacgactactacaagctc	chitinase
Lus10016872_ep_rv	ccctttcgccagagtgtca	chitinase
Lus10028377_ep_fw	caatattacggcagaggaccc	chitinase
Lus10028377_ep_rv	gcgccacgtccacattca	chitinase
Lus10041831_ep_fw	aagcgggcagggcaattg	chitinase
Lus10041831_ep_rv	cccgacaatcacgctatgga	chitinase
Lus10035621_ep_fw	tctgttcgccgttcaaggtc	chitinase
Lus10035621_ep_rv	agccgtcgaagttgtttgcc	chitinase
Lus10024366_ep_fw	tctacacacgagaagccttca	chitinase
Lus10024366_ep_rv	ccgtagtagctcttcccgg	chitinase
Lus10035624_ep_fw	aggggcgcattcctcaac	chitinase
Lus10035624_ep_rv	gccactgtctctccgacct	chitinase
ETIF3E_ep_fw	aacaaaagaagacgtccccag	eukaryotic translation initiation factor 3E
ETIF3E_ep_rv	gaaaagcttccatcttcaactcg	eukaryotic translation initiation factor 3E
GAPDH_ep_fw	aaggttcttcccgctctcaat	glyceraldehyde 3-phosphate dehydrogenase
GAPDH_ep_rv	gttcaatgcgattccagccc	glyceraldehyde 3-phosphate dehydrogenase
UBI2_ep_fw	aaggtgttcaccaagccgaa	ubiquitin extension protein
UBI2_ep_rv	agagcccgccttgttgtaaa	ubiquitin extension protein
Cluster-RTs-0-a_ep_fw	ctatggagtggactacgaggag	RLC_Lu0
Cluster-RTs-0-a_ep_rv	cttcttcagtctgcacaccat	RLC_Lu0
Cluster-RTs-0-b_ep_fw	ggactacgaggagacatttgc	RLC_Lu0
Cluster-RTs-0-b_ep_rv	ttcatcatcccctgtgatgat	RLC_Lu0
Cluster-RTs-1-a_ep_fw	ggagagacacaaggctaggc	RLC_Lu1
Cluster-RTs-1-a_ep_rv	tccagcttgcacacctttct	RLC_Lu1
Cluster-RTs-2-a_ep_fw	aaatgcaagctctcgaggca	RLC_Lu2

Primer name	Sequence	Gene annotation
Cluster-RTs-2-a_ep_rv	tttgccactggtgagaacgt	RLC_Lu2
Cluster-RTs-3-a_ep_fw	tggcaaaaggctattcacaaca	RLC_Lu3
Cluster-RTs-3-a_ep_rv	ggtgcttgtttgagtccgtag	RLC_Lu3
Cluster-RTs-28-a_ep_fw	tggaggagttacgagctttgg	RLC_Lu28
Cluster-RTs-28-a_ep_rv	gccactggtgcaaaggtttc	RLC_Lu28

Table 3.2 qRT-PCR primers used in experiment 3.

Primer name	Sequence	Gene annotation/alias
GAPDH_qrt_fw	gaccatcaaacaaggactgga	glyceraldehyde 3-phosphate dehydrogenase
GAPDH_qrt_rv	tgctgctgggaatgatgtt	glyceraldehyde 3-phosphate dehydrogenase
EF1A_qrt_fw	gctgccaacttcacatctca	eukaryotic translation initiation Factor 1
EF1A_qrt_rv	gatcgcctgtcaatcttggt	eukaryotic translation initiation Factor 1
ETIF5A_qrt_fw	tgccacatgtgaaccgtact	eukaryotic translation factor 5A
ETIF5A_qrt_rv	ctttaccctcagcaaatccg	eukaryotic translation factor 5A
Cl-RTs-0-a-2_qrt_fw	ggcccctataccaattagatgtg	RLC_Lu0
Cl-RTs-0-a-2_qrt_rv	cttcttcagtctgcacaccat	RLC_Lu0
Cl-RTs-1-a-1_qrt_fw	ggagagacacaaggctaggc	RLC_Lu1
Cl-RTs-1-a-1_qrt_rv	gacgtccatttgatataggggc	RLC_Lu1
Cl-RTs-2-a-2_qrt_fw	ttctcaccagtggcaaagat	RLC_Lu2
Cl-RTs-2-a-2_qrt_rv	tcctcatccaagtctccatg	RLC_Lu2
Cl-RTs-6-a-1_qrt_fw	ttcagtcaaaggaagggcatc	RLC_Lu6
Cl-RTs-6-a-1_qrt_rv	tcttcctccaaatcgccatg	RLC_Lu6
Cl-RTs-8-a-1_qrt_fw	tggtgacctgcatgaagaagt	RLC_Lu8
Cl-RTs-8-a-1_qrt_rv	agtaccactgccttgatgct	RLC_Lu8
Cl-RTs-28-a-1_qrt_fw	tggaggagttacgagctttgg	RLC_Lu28
Cl-RTs-28-a-1_qrt_rv	ccgtctgctctatatttaatggtg	RLC_Lu28
PME-32_qrt_fw	catggtggtcggtttgtg	pectin methylesterase
PME-32_qrt_rv	gtcgatcgccatgaatcc	pectin methylesterase

3.3.5 Experiment 1 and 2 – response to fungal extract and scarification

In experiment 1 we tested the effect of onozuka fungal extract on aerial sections of flax plants by estimating differences in transcript accumulation of selected genes and TEs using end-point RT-PCR. To test correct primer design, we performed a gradient PCR on genomic DNA to see whether amplicons matched the expected size. The PCR was run with 1X buffer, 2 mM $MgCl_2$, 0.2 mM dNTPs, 0.4 µM of each primer, 10 ng of DNA and 1.5 units of Taq polymerase (Thermo Scientific, Waltham, MA, USA). Cycling conditions were 94°C for 5 minutes, followed by 35 cycles of 94°C for 30 seconds, (52 – 55 – 58 - 61°C) for 30 seconds and 72°C for 1 minute, finalizing with an extension at 72°C for 5 minutes.

The response of chitinases, Ty1-*copia* elements and reference genes during the time course as 0, 2, 4, 8 and 24 hours after treatment with onozuka fungal extract, was assessed with end-point RT-PCR on cDNA pooled samples with the same profile as described for testing the primers, but with 5 ng of cDNA, and keeping the extension temperature at 61°C. All agarose gels were run 1% agarose at 75V for 40 minutes.

In experiment 2, where we tested the effect of onozuka and scarification on detached flax leaves, end-point RT-PCR was conducted similarly as for experiment 1. Band quantification for both experiments was performed using GelAnalyzer (http://www.gelanalyzer.com/index.html), with the following procedure (**Figure 3.4**): i) Images from all tested genes under all treatments were merged for comparison among genes; ii) merged gel images were loaded in the program and lanes corresponding to each treatment were outlined manually; iii) bands in each lane were automatically detected but corrected according to intensity profiles (images are converted to 8-bit gray scale with maximum intensity value of 255); iv) three bands with low, medium and high intensity were used as calibration bands to calculate a linear fitting curve (y=mx+b) to raw volume values (based on band thickness and intensity). The curve was used to calculate the rest of the unknown values; v) values were exported to Excel where a geometric mean was calculated using the three reference genes in each treatment, and then each gene value was divided by this geometric mean for normalization. Additionally, each value was further compared to the control at time 0 to calculate linear fold changes between treatments and control. The procedure of band quantification for end-point RT-PCR was performed as a way to better visualize how end amplification products change. However, since these bands were the products of end-point RT-

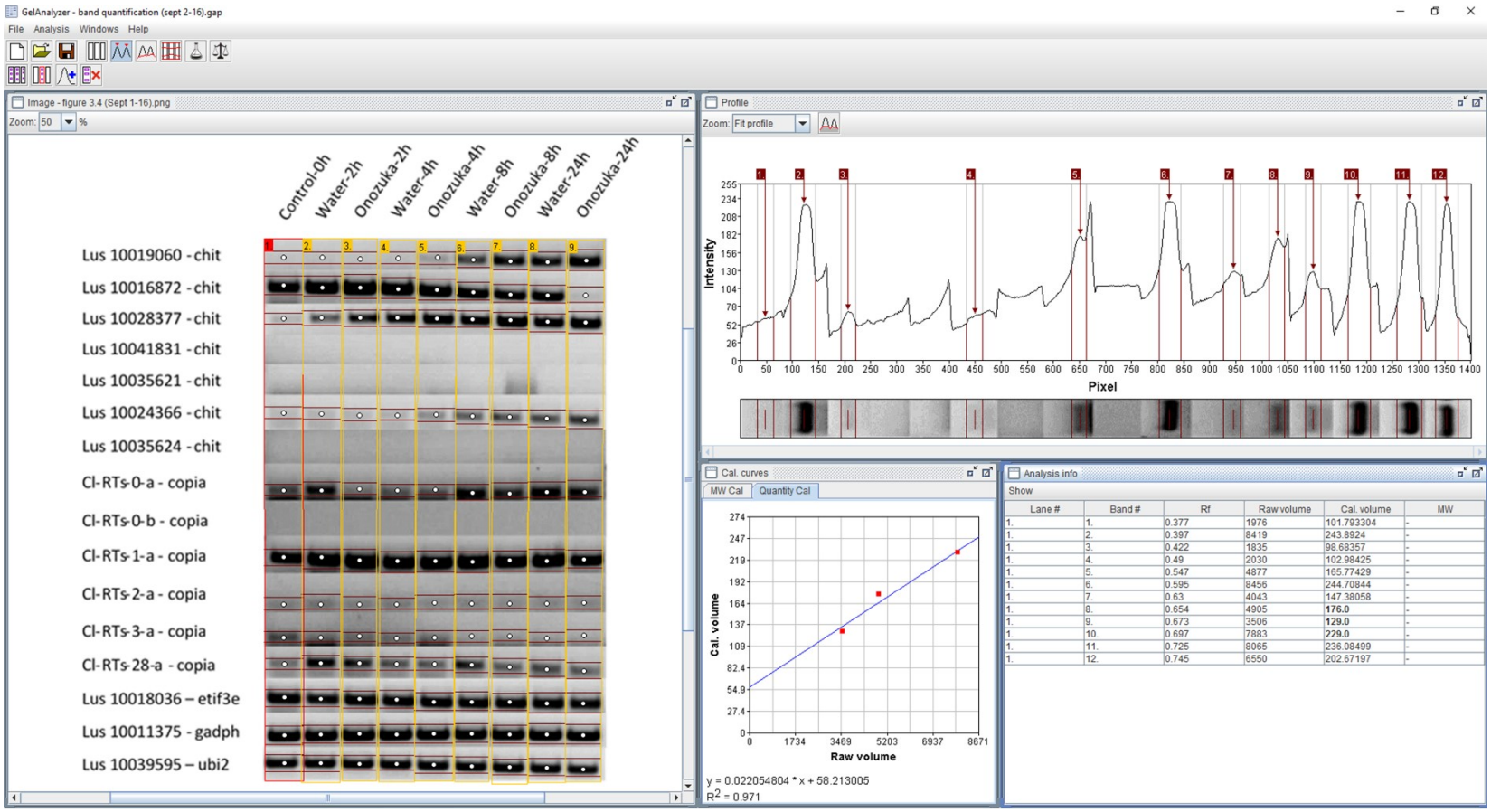


Figure 3.4 Quantification of end-point RT-PCR bands. A merged image was loaded to GelAnalyzer and lanes and bands were detected (image panel). The intensities of each band are displayed on the profile panel. Three bands are selected to generate a fitting curve (calibration curve panel), which allows interpolation of the unknown values, generating calibrated values (analysis info panel).

PCR, we do not know which, if any, reactions reached the plateau of amplification. Therefore, PCR results may not be fully quantifiable.

3.3.6 Experiment 3 – specific organ response

A test for quality of cDNA and absence of residual DNA contamination was performed using ETIF3E and PME gene primers (**Table 3.2**). The former produced 106 bp and 378 bp bands for cDNA and genomic DNA respectively, while the latter produces 123 bp and 848 bp bands. The PCR was run with 1X buffer, 2 mM $MgCl_2$, 0.2 mM dNTPs, 0.4 µM of each primer, 5 ng of DNA/cDNA (DNA is used as positive control) and 1.5 units of Taq polymerase (Thermo Scientific, Waltham, MA, USA). Cycling conditions were 94°C for 5 minutes, followed by 35 cycles of 94°C for 30 seconds, 60°C for 30 seconds and 72°C for 1 minute, ending with an extension at 72°C for 5 minutes.

To quantify the differences in Ty1-*copia* family expression across different organs, four individual replicates per time point and treatment were used for qRT-PCR. For testing differential expression, samples were aliquoted in 384-well plates (with three technical replicates per sample and organ combination) using a Biomek 3000 Laboratory Automation System (Beckman Coulter, Brea, CA, USA), and the qRT-PCR was run using a QuantStudio 6 Flex Real-Time PCR system (Applied Biosystems-Life Technologies, Carslbad, CA, USA). Sample reactions were done in 10 µL with 5 µL of SYBR-green (Molecular Probes – Thermo Fisher Scientific, Waltham, MA, USA), 2.5 µL of the mixed primer pair (3.2 µM) and 2.5 µL of a 1:50 dilution of the synthesized cDNA. Cycling conditions were: 95°C for 2 minutes followed by 40 cycles of 95°C for 30 seconds, 60°C for 1 minute. A melting curve stage was added: 95°C for 15 seconds, 60°C for 1 minute and 95°C for 15 seconds.

Relative quantification was performed using the geometric mean of the three selected reference genes with the ΔCt of the reference – the test gene. Statistical differences in each TE family were calculated using ANOVA followed by a Tukey test for multiple comparison with GraphPad Prism version 6.0 (GraphPad Software, La Jolla California USA); homogeneity of variances was tested with the Brown-Fosythe test (a variation of the Bartlett's test) before performing ANOVA. Fold changes were calculated using the $2^{(-\Delta\Delta Ct)}$ relative quantification method [289], to calculate the log2 value.

3.3.7 Experiment 4 – RNA-seq transcriptome response to *Fusarium oxysporum*

Details on RNA-seq setup and analysis are described in Chapter 4. We used the RNA-seq analysis tool from CLC Genomics Workbench 8.0.2 (https://www.qiagenbioinformatics.com/**)** to map the reads from the fastq files to a database of 699 non-redundant Ty1-*copia* elements from flax generated from a previous study [218]. This database has been uploaded as a dataverse file (all copia elements.fa – http://dx.doi.org/10.7939/DVN/10933). For read mapping we used the following parameters: mismatch cost (2), insertion cost (3), deletion cost (3), length fraction (0.8), similarity fraction (0.8), and maximum number of hits for a read 30. Statistical analysis was performed similarly as for cufflinks. RPKM (reads per kilobase of transcript per million mapped reads) was calculated from unique reads matching each retrotransposon and biological replicates were used to calculate unpaired t-tests. Significant differences (*p*-values) were corrected for multiple comparisons using the benjamini-hockberg correction [290]. Fold changes were calculated from the average of the two treatments (fungal inoculation / water control) and a log2-fold change was calculated from this value. Protein domains from TEs with largest RPKMs were previously identified with RepeatMasker **(**http://www.repeatmasker.org**)** [218]**.**

3.4 Results

3.4.1 *In-silico* prediction of transcription factor binding sites (TFBS') in retrotransposon LTRs

We searched LTRs (Appendix 3.2) for the most common TFBS' related to plant defense responses [286,287]. From the 10 tabulated TFBS', only Whirly was absent from all TE families evaluated, and a heat stress TFBS was only found in family RLC_Lu6-1 (**Table 3.3**). In the meantime, bZIP, AP/ERF and DOF TFBS were found in all families examined. The size of the LTRs was correlated with the total number of TFBS' identified (correlation = 0.97), and therefore in longer LTRs, the likelihood of finding more TFBS' was higher. For example, RLC_Lu6-1, which bears the longest LTR, had a total of 166 TFBS hits, and had the most hits in each TFBS analyzed with the exception of the WRKY TFBS, which was not found in this TE. Adding all hits in all TEs the two most abundant TFBS' were bZIP and AP2/ERF, being this latter one the most abundant (**Table 3.3**). Looking at each TE, most retrotransposons had the highest proportion of hits to the AP2/ERF TFBS', but for RLC_Lu0-1 and RLC_Lu8-1 the MYB and bZIP TFBS' had respectively the largest proportions. An additional search for a 13-bp

sequence (TGGTAGGTGAGAT), that has been shown to function as a *cis*-regulatory activated in response to tissue culture, jasmonate, wounding and fungal elicitors in LTRs of tobacco retrotransposons [92], was not found in any of the flax retrotransposons evaluated.

Table 3.3 Transcription factor binding sites (TFBS') present in flax Ty1-*copia* retrotransposon representative sequences from each family.

TFBS	Main responses/functions	RLC_Lu0-1	RLC_Lu1-1	RLC_Lu2-1	RLC_Lu3-1	RLC_Lu6-1	RLC_Lu8-1	RLC_Lu28-1	Total per TFBS
bZIP (basic-region leucine ZIPper Protein)	defense [1] - light signalling, abiotic stress response, pathogen defense, seed maturation, flower development [2]	3	4	3	15	42	36	2	**105**
bHLH (basic-Helix-Loop-Helix)	hormone signalling, flavonoid biosynthesis, seed and root differentiation, biotic and abiotic stress response light signalling [2]	1	0	2	6	27	10	2	**48**
MYB (MYeloBlastosis)	defense [1] - plant secondary metabolism, cell fate, biotic and abiotic stress response, cellular and organ morphogenesis and differentiation, cyrcadian rhythm [2]	28	16	0	14	25	10	5	**98**
HSF (Heat Stress Transcription Factors)	accumulation of heat-shock proteins, abiotic stress response, thermotolerance [2]	0	0	0	0	1	0	0	**1**
AP2/ERF (APETALA2/Ethylene Response Factor)	defense [1] - biotic and abiotic stress (drought-salt-cold) response, ethylene response [2]	18	24	8	22	49	13	12	**146**
WRKY (WRKY conserved domains)	defense [1] -biotic and abiotic stress response, development, hormone signalling, flavonoid biosynthesis [2]	0	2	0	1	0	1	0	**4**

TFBS	Main responses/functions	RLC_Lu0-1	RLC_Lu1-1	RLC_Lu2-1	RLC_Lu3-1	RLC_Lu6-1	RLC_Lu8-1	RLC_Lu28-1	Total per TFBS
NAC (NAM (No Apical Meristem) - ATAF1 - CUC2 (CUp-shaped Cotyledon))	defense [1] - biotic and abiotic stress (wounding) response, development [2]	0	1	0	0	6	0	0	**7**
TCP (Teosinte branched 1 - Cycloidea - Proliferating cell factor 1 (PCF))	plant growth and development (flower, leaf morphogenesis and senescence, embryo growth, plant architecture, cyrcadian rhythm) [2]	0	0	0	0	5	0	0	**5**
DOF (DNA-binding with One Finger protein)	defense [1] - light, phythormone and defense responses, seed development, germination [3]	2	7	5	2	11	7	4	**38**
Whirly (Whirly quaternary structure)	defense [1,4]	0	0	0	0	0	0	0	**0**
	Total TFBS' per TE family	**52**	**54**	**18**	**60**	**166**	**77**	**25**	**452**
	LTR size	**252**	**217**	**200**	**360**	**826**	**351**	**197**	

The numbers in the retrotransposon columns correspond hits per each TFBS. TFBS' with the largest amount in each TE are highlighted.
Correlation of total TFBS' with LTR size = 0.97
Minimum similarity score for the analysis is 0.7
[1] [287]
[2] [286]
[3] [291]
[4] [292]

3.4.2 Response to fungal extracts

TFBS' indicated that the selected TE families might respond to different stress treatments. We first tested a fungal extract (onozuka) to see its effect on the transcription of stress response genes (chitinases) and Ty1-*copia* families.

End-point RT-PCR showed the amplicons derived from cDNA had the expected sizes in all cases (not shown). A summary figure was built to compare the patterns of transcriptional response over the time course (**Figure 3.5**), and the bands were quantified and normalized to the geometric mean of the three reference genes (**Figure 3.7A**). Chitinase primers were designed as response markers for scarification and the fungal extract since chitinases usually respond to these elicitors [293]. To select diverse chitinases a dendrogram of the relationships of flax chitinases to previously characterized *Arabidopsis thaliana* chitinases was built (**Figure 3.3**). Chitinase Lus10019060 was induced after 4 h in response to onozuka, and an increasing induction was seen at 8 and 24 h for both the water control and the onozuka extract, although the end product in the onozuka treatment was higher. Chitinase Lus10016872 had high constitutive expression, demonstrated by its larger relative abundance in the control at 0 h. More abundant end-products were apparent at 2 h in both the water control and the fungal extract, but the expression decreased with time and was inhibited with onozuka at 24 h. Chitinase Lus10028377 showed low relative expression at time point 0 (**Figure 3.7A**), and increased induction over time with both the control and the fungal extract; however, the level of induction in each time point was higher with onozuka when compared to the control. Chitinases Lus10041831, Lus10035621 and Lus10035642 showed no induction or constitutive expression. Chitinase Lus10024366 had a low relative expression at 0 h and a difference in induction was evident at 4 h when more end-product was seen for onozuka than for control (**Figure 3.7A**). This pattern continued until 24 h. The patterns of expression found for the Ty1-*copia* families were more erratic. Primers Cl-RTs-0-a from family RLC_Lu0 had a low level of constitutive expression (control 0h), its product increased for water at 2 h and then decreased at 4h. The end products had higher relative quantities at 8 and 24 h (**Figure 3.7A**), but with water displaying more product. Primers CL-RTs-0-b displayed no induction, even though it belonged to the same RLC-Lu0 family. Primers Cl-RTs-1-a from family RLC-Lu1 had high constitutive expression across the time course and treatments. Primers from families RLC_Lu2 an RLC_Lu3 had some constitutive expression as demonstrated by their faint bands in the control at 0 h, and some expression was maintained in

the other time points with no clear distinction between onozuka and water. Primers from family RLC_Lu28 exhibited some constitutive expression (control at 0 h), and different levels of induction throughout the time course, with more relative end product on the controls of each time point, but no clear temporal pattern. Finally, the three reference genes showed constitutive high expression in all time points, demonstrating that the patterns of the query genes were reliable.

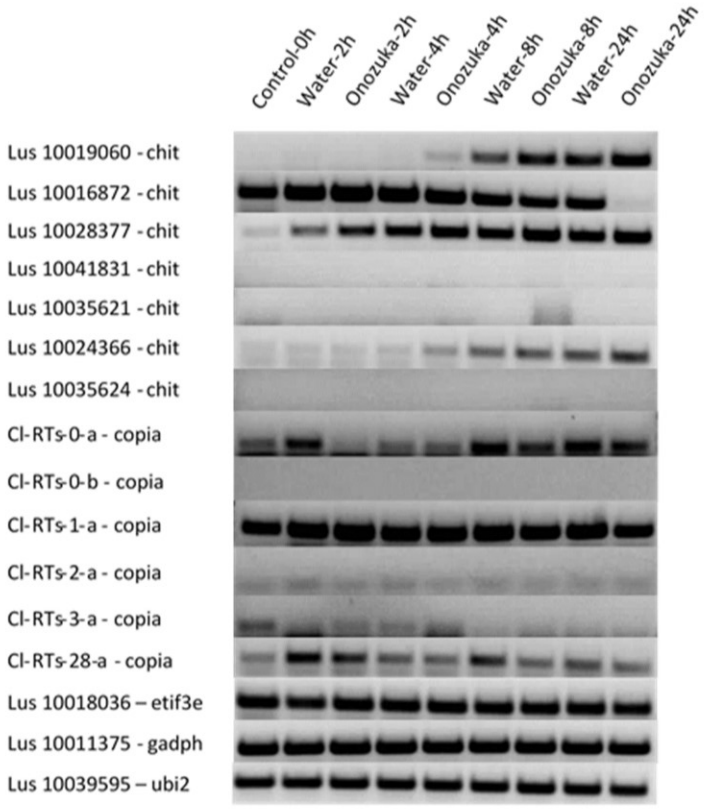


Figure 3.5 End-point RT-PCR summary for experiment 1. Aerial sections (stems+leaves) were placed in water or onozuka fungal extract and evaluated over a period of 24 hours to assess gene expression changes in chitinase, Ty1-*copia*, and reference genes.

3.4.3 Response to fungal extract and wounding

We then tested onozuka extract along with a wounding treatment (scarification) to assess the transcriptional response of the same chitinase genes and TE families used in the previous section. To assess gene expression responses in flax leaves, a similar procedure as for the onozuka assay on aerial sections, was followed. A summary figure was also generated to compare the responses to these two elicitors over the time course (**Figure 3.6**) and the bands

were quantified and normalized to the geometric mean of the three reference genes (**Figure 3.7B**). Chitinase Lus10019060 showed a trend for increased abundance with time; at 4 h, increased abundance was evident for the water control compared to the treatments, but at 8 and 24 h a thicker end-product band indicated higher transcription for the onozuka and scarification treatments when compared with the water control (especially for the former). Chitinase Lus10016872 showed constant expression, with the exception of a slight increase at 24 h specifically for the onozuka treatment (**Figure 3.7B**). The response of this gene seemed lower upon scarification (24 h), even when compared to the water control. Chitinase Lus10028377 had high constitutive expression (control-0h) with a peak of induction for both onozuka and scarification at 8 and 24 h when compared to the water control. Chitinase Lus10041831 showed larger end products by onozuka at 8 h, and by onozuka and scarification at 24 h. Chitinases Lus10035621 and Lus10035624 demonstrated no constitutive or induced expression. Chitinase Lus10024366 showed an intermediate level of constitutive expression and some level of induction of the onozuka and scarification treatments was seen at 8 and 24 h when compared to the water controls (the change is more obvious for the fungal extract treatment). Primers CL-RTs-0-a showed a very erratic pattern. The absence of a band for the water control and presence of the band for the onozuka and scarification treatments at 4 and 24 h, argues for an induction by the elicitors, but this pattern was not consistent on the remaining time points (**Figure 3.6**). The other primer pair from family RLC-Lu0 (CL-RTs-0-b) had low abundance in later time points, but it was difficult to discern a pattern. Primers from family RLC_Lu1 demonstrated high constitutive expression; after 4 h the stress treatments had larger relative end-products than their respective water control (**Figure 3.7B**). Primers from families RLC_Lu2 and RLC_Lu3 had low relative amounts as compared with most other genes in this experiment, and no pattern could be observed. Primers from family RLC_Lu28 showed one of the two largest levels of relative constitutive expression (control-0h) with the other time points and treatments showing slightly higher relative abundance of their end products, but with no clear pattern (**Figure 3.7B**). Similarly, as with experiment 1, all three reference genes used presented constant expression throughout treatments and time points.

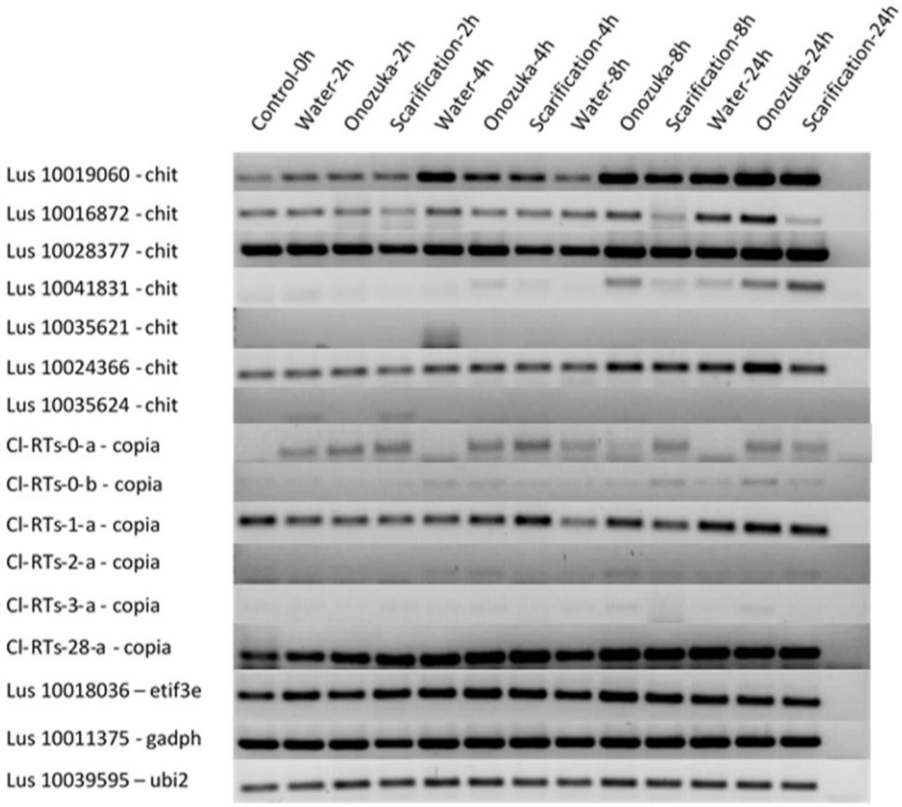


Figure 3.6 End-point RT-PCR summary for experiment 2. Detached leaves were immersed in water (with or without scarification) or onozuka fungal extract and evaluated over a period of 24 hours to assess gene expression changes in chitinase, Ty1-*copia*, and reference genes.

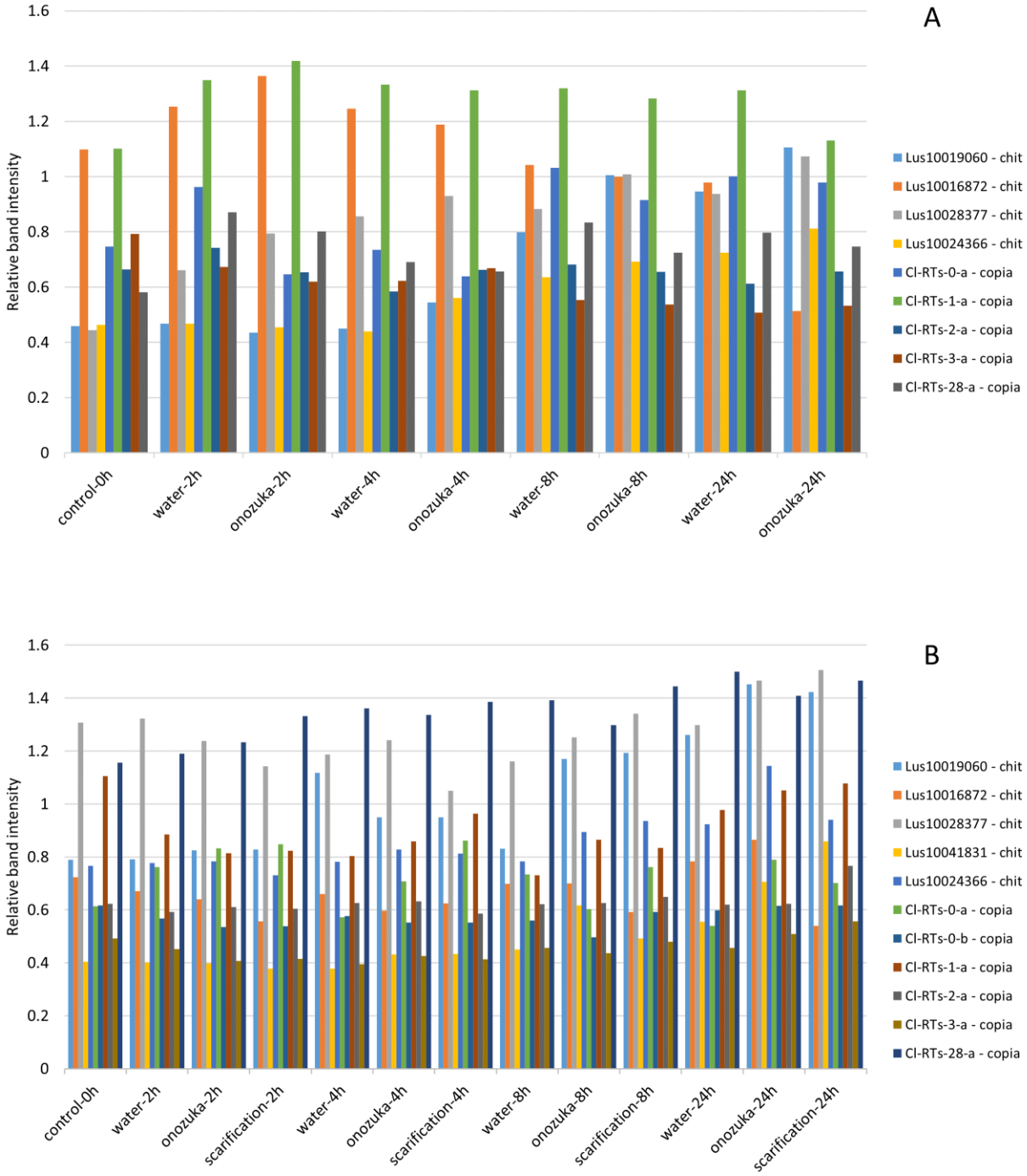


Figure 3.7 Relative calibrated amounts of end-point RT-PCR products. A. Experiment 1 (shoots in onozuka fungal extract). B. Experiment 2 (detached leaves placed in water with and without scarification, or in onozuka fungal extract).

3.4.4 Differential response of TEs in flax organs

While no Ty1-*copia* families presented clear evidence of increased activity in response to fungal extracts or wounding in the end-point RT-PCR experiments, several TEs showed constitutive expression. We tested if the level of expression of the families was constant among different plant organs made of distinct tissues, by using root as reference to compare with the leaves and stems (**Figure 3.8**). In the cultivar CDC Bethune, significant differences in the expression were seen between root and leaves in family RLC_Lu2, between roots and leaves and stem for family RLC-Lu8 and between root and stem in RLC_Lu28. For the cultivar Oliver there was only one significant difference between root and leaves for family RLC_Lu6. Finally, for Stormont Cirrus, the level of expression in roots was significantly different from leaves and stem in family RLC_Lu6, and root and stem were different from the level of expression in leaves for family RLC_Lu8.

We then compared the relative changes of each TE family in the same organ among the three cultivars and used CDC Bethune as reference (**Figure 3.9**). The differences found in this analysis were more substantial than for the previous analysis. In leaves CDC Bethune had levels of expression which were significantly different from Oliver and S. Cirrus for TE families RLC_Lu0 and RLC_Lu1, with log2-fold changes >2. For family RLC_Lu2 the log2-fold change of Oliver was >3 when compared to CDC Bethune. In family RLC_Lu6 the expression in Oliver was significantly less than for CDC Bethune and S. Cirrus, while S. Cirrus had significantly higher levels of expression than both of the other cultivars. For family RLC_28, the level of expression in S. Cirrus was significantly higher than in CDC Bethune. Analyses of roots showed that Oliver and S. Cirrus all had log2-fold changes close to or above 3 when compared with CDC Bethune for families RLC_Lu0, RLC_Lu1 and RLC_Lu28. Family RLC_Lu2 had the same pattern of expression as for leaves, while for family RLC_Lu6 the expression in Oliver had a negative log2-fold change of almost 5. Finally, when comparing expression in stems across cultivars almost the exact pattern as for leaves was replicated with the exception of family RLC_Lu28, where the level of expression in S. Cirrus was significantly less than in the other two cultivars.

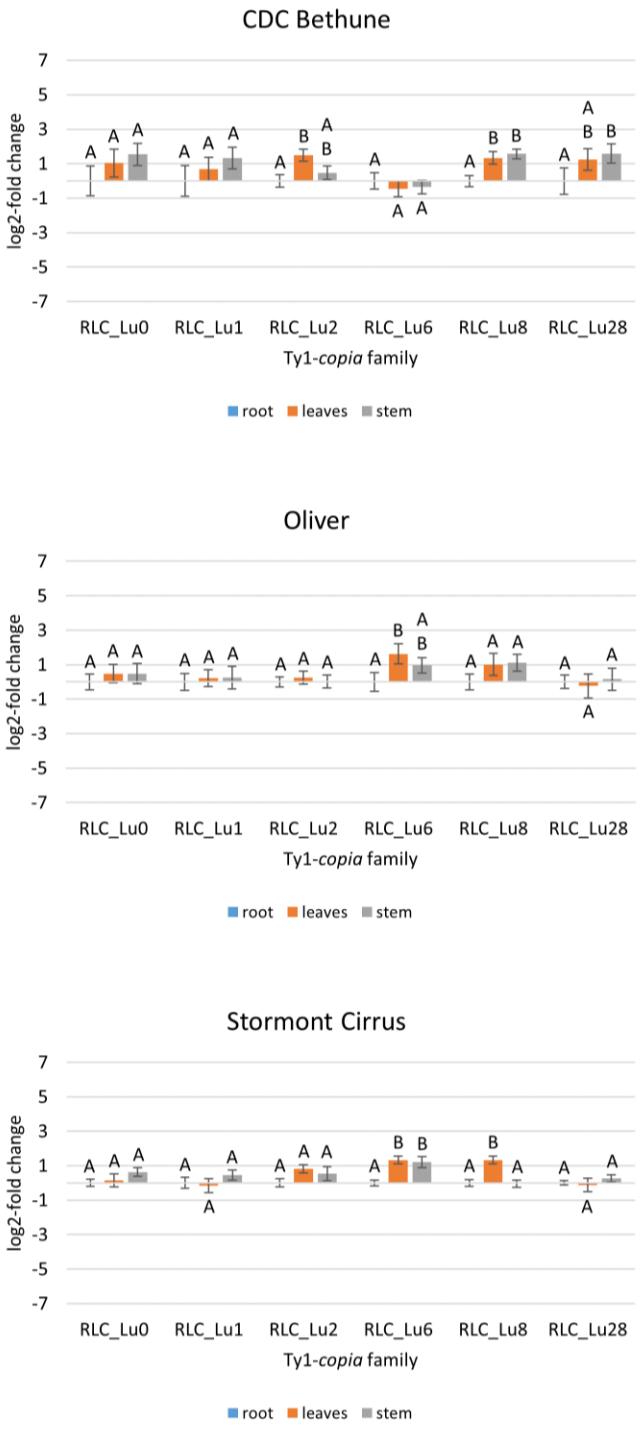


Figure 3.8 Log2-fold gene expression changes between tissues in different Ty1-*copia* families, for three different cultivars. Abundance in leaves or stem is shown relative to root. Different letters represent significant statistical differences after Tukey multiple comparisons ($p < 0.05$). Error bars = standard error of the mean (SEM).

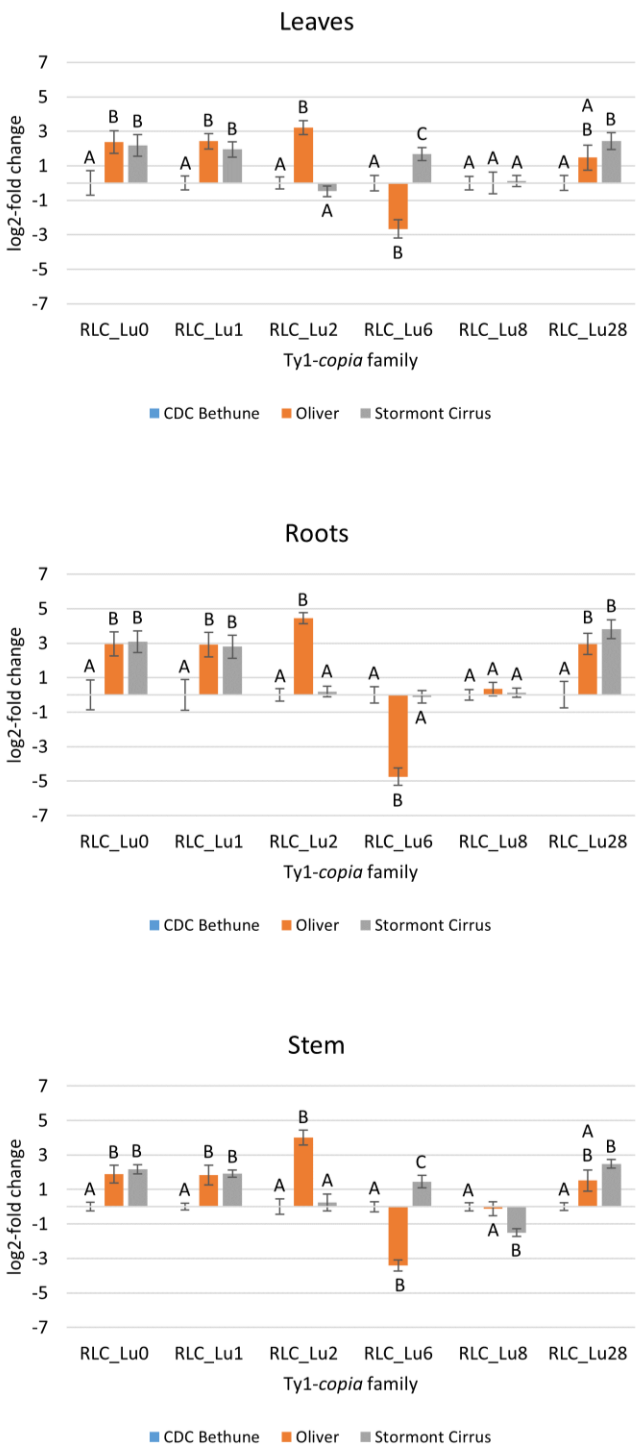


Figure 3.9 Log2-fold gene expression changes between cultivars in different Ty1-*copia* families, in three different tissues. Abundance in Oliver or Stormont Cirrus is shown relative to CDC Bethune. Different letters represent significant statistical differences after Tukey multiple comparisons ($p < 0.05$). Error bars = standard error of the mean (SEM).

3.4.5 Differential expression of TEs in flax plants inoculated with *Fusarium oxysporum*

Our RNA-seq experiment (see chapter 4) showed that after correction for multiple comparisons using the Benjamini-Hochberg statistic, none of the inividual Ty1-*copia* elements found in the genome displayed differential expression in either of the four days post inoculation (DPI) evaluated (dataverse file – TE differential expression in flax upon *Foln* infection.xlsx - http://dx.doi.org/10.7939/DVN/10933). Some TEs showed individual significant *p*-values (< 0.05): 20 TEs 2 DPI, 28 TEs 4 DPI, 24 TEs 8 DPI and 10 TEs 18 DPI. However, these values should be approached with caution since many of them have a value of 0 in one of the two conditions and a low RPKM (<10) which usually reflects just a few reads, and probably stochastic variation, which is not really indicative of a real expression difference.

As with the previous experiments many of these TEs had some level of constitutive expression, with many of them having high RPKMs (> 1000). Two days post-inoculation there were four TEs with RPKMs > 5000 for both the control and the treatment (dataverse file – TE differential expression in flax upon *Foln* infection.xlsx - http://dx.doi.org/10.7939/DVN/10933). The same four TEs also had RPKMs > 5000 at 4 DPI, and at 8 DPI, but on this latter day, there were two additional TEs with RPKMs > 5000. On the last day (18 DPI), six TEs had RPKMs > 5000, from which 5 out of 6 were common to the ones found 8 DPI. This constitutes a total of seven Ty1-*copia* elements with a RPKM > 5000 during the time course, representing hundreds of mapped reads to each TE. From these, only two belong to families RLC_Lu2 and RLC_Lu8, which were evaluated in other experiments in this Chapter (**Table 3.4**), while three others belong to families RLC_7, RLC_15 an RLC_26, and two could not be placed in a family due to the absence of a reverse transcriptase sequence (for family placement see methodology in Chapter 4 – section 2.3.3).

Finally, the presence of protein domains in each of these TEs was assessed as a way to evaluate conservation of the retrotransposons. This characteristic, along with LTR similarity (**Table 3.4**), are proxies for potential TE activity. It was found that in general, Ty1-*copia* elements with higher LTR similarity had more of their internal protein-coding domains intact.

Table 3.4 Ty1-*copia* elements with RPKM > 5000 in both water control and fungal treatment.

Identifiers			Days post inoculation				Domains[c]				
Scaffold identifier[a]	**TE size**	**Family[b]**	**2**	**4**	**8**	**18**	**GAG**	**PR**	**INT**	**RT**	**RH**
Copia/LF/S116_455741_459409_-/93.3	3668	not classified	yes	yes	yes	no	no	no	no	no	yes
Copia/LF/S139_77307_85315_-/99.8	8008	RLC_Lu8	no	no	yes	yes	no	yes	yes	yes	yes
Copia/LF/S196_897314_902511_-/100	5197	RLC_Lu26	no	no	yes	yes	yes	yes	yes	yes	yes
Copia/LF/S257_519492_536775_+/92.7	17283	RLC_Lu7	yes	yes	yes	yes	no	yes	yes	yes	no
Copia/LF/S280_703358_708311_+/99.1	4953	RLC_Lu2	yes	yes	yes	yes	yes	yes	yes	yes	yes
Copia/LF/S426_424_19200_-/93.2	18776	RLC_Lu15	no	no	no	yes	yes	yes	yes	yes	yes
Copia/LF/S480_180560_198307_+/76.7	17747	not classified	yes	yes	yes	yes	no	no	yes	no	yes

[a]Scaffold identifier are given as output from LTR finder: e.g. Copia/LF/S116_455741_459409_-/93.3 = *Copia* TE/identified by LTR finder/in scaffold 16 from position 455741 to 459409 in the minus strand/with similarity between its LTRs of 93.9%.
[b]Family refers to the classification of the respective TE according to its reverse transcriptase sequence (section 2.3.3 – Chapter 4). When element is not classified the reverse transcriptase was not found for this element.
[c]GAG (group-specific antigen), PR (protease), INT (integrase), RT (reverse transcriptase), RH (ribonuclease H).

3.5 Discussion

3.5.1 Response of chitinases to fungal extracts and wounding

In our study chitinases, but not TE families, demonstrated differential transcript abundance when challenged with fungal elicitors or by wounding (**Figure 3.5, Figure 3.6 and Figure 3.7**). Chitinase expression is well-documented in response to fungal attack, and a multitude of other stresses, including wounding, drought, cold, growth and development [293,294]. Chitinases inhibit the growth of the fungi by acting on the exposed fungal tips with nascent chitin chains being synthesized [293,295]. Chitinases have been shown to be triggered by different pathogens [157,296], or synergistically activated by both fungal and wounding responses. [297–299].

In flax, expression of chitinases upon treatment with fungal elicitors has been studied in the context of its interaction with rust (*Melampsora lini*) [300], where chitinases increased their expression in response to both virulent and avirulent strains, with lesser expression when challenged with the former strains. A chitinase was also used as a marker of *F. oxysporum* infection progression, in a study to determine changes in cell wall polymers in flax upon interaction with the fungus [222]. Finally, we detected the activation of multiple chitinases in our study following the gene expression changes of flax when challenged with *F. oxysporum* (see Chapter 4). Our results show that the treated plants had perceived an induced stress and were responding to it.

3.5.2 Response of TE families to fungal extracts and wounding

Our end-point RT-PCR analysis of the expression of selected Ty1-*copia* families did not identify a pattern of TE activation by either onozuka or wounding (Figures 3.7). Microbial elicitors as well as wounding can activate tobacco retrotransposons [55,62,92,301]. This response parallels the general plant defense response, due in part to common *cis* motifs shared by defense genes and LTRs in retrotransposons [50,55,64,77,92]. Although all of the examined TE families contained TFBS' that have been shown to be involved in defense responses and wounding (**Table 3.3**), we did not find any difference in expression of these TEs in treated samples as compared to controls. All of the retrotransposons evaluated had a high proportion of AP2/ERF, which are usually located in pathogen and hormone induced genes [101,287,302–304]. Another stress-modulated TF that is commonly found in plant defense genes (WRKY)

[305], had very few sites in our TE sequences, and a 13 bp sequence found in LTRs of tobacco, than controls induction by tissue culture, wounding, fungal elicitors and methyl jasmonate [92], was not found in any of the flax retrotransposons. In our experiment, the lack of responses of TEs to stress treatments, despite the presence of many predicted stress-responsive TFBS', demonstrates that the presence of these TFBS' is not sufficient to modulate the activity of these TEs in response to stress. Many TFBS' are short and therefore motifs can occur often without having a function. Addditionally, structural restrictions might impair the transcription factors from reaching the binding site. In the future, an enrichment analysis should be performed to see if retrotransposon promoters have specific motifs more often than expected by chance.

TE activation depends on numerous factors. A TE can fall in heterochromatic regions, which have higher rates of mutation than gene coding regions [306], resulting in a higher chance of degeneration of promoter and protein coding regions. Additionally, TEs have rates of evolution which are higher than those of genes [38], which further increase the likelihood of rapidly degenerating and becoming non-functional. It has also been found that many TEs target other TEs for insertion, generating a nested pattern [39] which disrupts transposon function. For those TEs that remain functional, the activation may depend on changes in their silencing status. A TE can be silenced due to epigenetic mediated methylation [116,117] and such methylation would have to be lifted to allow transcription and consequent transposition. Additionally, TEs can be activated developmentally or in a tissue specific manner [307–309], which further complicates trying to trigger their activation under laboratory settings. Under all these factors, detecting the activity of TEs is difficult, because even when the same family is detected in a closely related species or interspecifically, the history of TE-mediated evolution and insertion sites may vary.

The erratic patterns of activation of TEs in our experiment probably depended on many of these variables. Additional factors like a more localized TE activation (a few cells in a tissue), was suggested before [55,77]. Finally, amplification of sections of the conserved reverse transcriptase, as performed in our experiment, implies maximizing capturing transcripts derived from different family members, but also results in variability depending on activation of different elements. In fact, the use of different primers for the same family (CL-RTs-0-a and CL-RTs-0-b) resulted in different patterns of response, demonstrating that different sections of a TE may yield

different results depending on the accumulated mutations, which at the same time may vary among members of a same TE family.

While extensively studied retrotransposons like *Tnt1* can be used as markers of plant defense responses [77], the regular activity of other TEs for this type of studies seems elusive. However, some of our evaluated TE families showed expression even in controls at 0 h (constitutive expression), with family RLC_Lu1 an RLC_Lu28 demonstrating higher expression than the other families in experiments 1 and 2 respectively, throughout all treatments (**Figure 3.5 and Figure 3.6**). Using RT-PCR, a Ty3-*gypsy* retrotransposon (*Ogre*) was shown to be constitutively expressed in different plant tissues and upon wounding [310]. Likewise, *Rider*, a Ty1-*copia* element from tomato, was constitutively expressed across plant tissues using RT-PCR (which is the same as our end-point RT-PCR) [311]. *Rider* was further characterized as a mid copy number TE (around 100 copies) and therefore there seemed to be no correlation of transposition and the fact that transcription was not restricted. Similarly, we saw no correlation here of the transcriptional levels of families RLC_Lu1 and RLC_Lu28, with the copy numbers demonstrated in Chapter 2 (**Figure 2.2**); while family RLC_Lu1 has a low copy number, RLC_Lu28 has an intermediate copy number, and yet both seem to have high constitutive transcription levels (Figures 3.5 and 3.6). This lack of correlation was previously observed for TEs in maize [308] and for flax TEs [218], when comparing predicted TE copy numbers and associated ESTs by bioinformatics approaches.

Constitutive transcription indicates that at least some of the members of these families could escape epigenetic control under the conditions examined. Such escape mechanisms from epigenetic control could depend on TE location in the genome, like when TEs insert inside genes, and seem to have a lower level of methylation than intergenic TEs [136]; for example, the sequencing of TE insertions performed in Chapter 2, showed multiple transposons mapped inside genes (**Table 2.9**). Alternatively, changes in controlling mechanisms may be tissue specific (see next section).

3.5.3 Tissue-specific expression

The barley Ty1-*copia* retrotransposon *BARE-1* is one of the most studied and abundant TEs from a plant genome [173]. This TE was shown to have tissue-specific expression, especially in meristematic regions and reproductive tissue [309]. In fact, ovule related tissue

presented high concentration of TE-derived proteins, where cells are supposed to be demethylated, and it has been proposed that this demethylation in reproductive tissue can result in TE reactivation [312]. Furthermore, the LTR contains promoter elements that are associated with gibberellin and sugar metabolism, which are usually localized to specific tissues [309]; therefore, motifs in LTRs along with methylation changes could account for tissue specific expression.

The constitutive expression of some TE families from our experiments 1 and 2, is therefore indicative of possible tissue-dependent expression. However, we found only a few significant differences between roots, leaves, and stems when we examined six families of retrotransposons in three different flax cultivars (**Figure 3.8**). These differences usually indicated higher expression in leaves and stem as compared to roots. Strong expression of *BARE-1* was seen in barley axillary shoot apical meristems and vascular tissues of the stem, while in roots the expression was exclusively localized to the root tips [309].

Differential tissue regulation of TEs has also been shown in other plants. In *Quercus suber* the *gypsy* retrotransposon *corky* is active throughout plant development, but differential transcriptional abundance depends on tissue and potentially on environmental triggers [307]. Expression is high in reproductive tissue, and also in roots, which is attributed to the presence of meristematic tissues, and possible wounding caused by stress as the roots grow through the soil. In maize, *in silico* transcript analysis shows that TE families are expressed at higher levels in certain tissues, for example some retrotransposons are especially abundant in apical meristems and reproductive tissues [308]. In sugarcane, no specific families could be identified as having a bias for tissue-specific expression, but individual TEs had different levels of expression in different tissues [313].

When the relative expression of the TE families in each tissue was compared between cultivars the differences were more evident and larger than when comparing tissues within each cultivar (compare **Figure 3.8 and Figure 3.9**). Almost exact patterns of expression were achieved with the three tissues, demonstrating a clear difference of expression of most TE families between cultivars. In three TE families, cultivars Oliver and S. Cirrus had higher TE expression than CDC Bethune. In the meantime, in family RLC_Lu2 only Oliver was significantly upregulated when compared to the other two cultivars; in family RLC_Lu6 Oliver expression was significantly lower than in CDC Bethune and S. Cirrus, and in family RLC_Lu8

there was no change between cultivars. Similar to experiments 1 and 2, there does not seem to be any relationship between transcriptional differences between cultivars and their copy numbers: e.g. increased transcription in Oliver and S. Cirrus in three families of TEs (**Figure 3.9**) does not show any trend in correlation with differences in copy number for the same three families (**Figure 2.2**). Nevertheless, the differences in relative transcription among the three cultivars tested supports the level of polymorphisms in insertion sites as assessed by SSAP (**Figure 2.4**), and confirms that most of these TE families are still active.

A higher spatial resolution of TE expression could be achieved in the future by examining specific tissues within organs, and targeting for meristematic and reproductive tissues as a way to assess *de-novo* insertions that will potentially be inherited. While the families of TEs we examined here do not have as high copy numbers as families like *BARE-1* from barley, it is possible that TEs in flax could use a similar strategy of activation in reproductive tissue, which could be related to the polymorphisms we have detected here among cultivars (see Chapter 2). While tissue specific expression does not exclude response to stress, in the future a better strategy will comprise tagging specific members of a TE family, and testing a larger array of stress elicitors.

3.5.4 TE response to flax inoculation with *Fusarium oxysporum*

We were not able to detect differential expression of Ty1-*copia* retrotransposons under our experimental conditions after correcting for multiple comparisons. Contrary to the erratic activation of TE families, here we were able to evaluate individual TEs, and therefore confounding factors like conserved primers used for all members of a TE family, or TEs only falling in heterochromatic regions, are not a factor. Alternatively, regulation could happen at a different time points than the ones that were tested, or TEs could be activated only in specific tissues [307,309] and the response would be diluted in our samples which represent full plants. Nevertheless, hundreds of TEs showed some level of constitutive expression in each of the days evaluated (dataverse file – TE differential expression in flax upon *Foln* infection.xlsx - http://dx.doi.org/10.7939/DVN/10933), and a few had RPKMs far above background (**Table 3.4**), indicating that epigenetic repression is not a factor in them. It is possible that since the base control condition for this experiment are plants grown in closed tubes for over two weeks, this already constitutes an stress which lifts epigenetic repression of some TEs, resulting in what

seems to be a constitutive pattern of expression under our experimental conditions. Furthermore, these TEs are regulated through time (without differential expression between control and treatment), as evidenced by their differing RPKM in the different days sampled (dataverse file – TE differential expression in flax upon *Foln* infection.xlsx - http://dx.doi.org/10.7939/DVN/10933); for example, the TE with identifier Copia/LF/S196_897314_902511_-/100 started at day 2 with an RPKM over 3000 (for both control and treatment), which increased to >12000 18 DPI. In this sense, a shortcoming of our experimental design is not having an additional set of control plants growing in conditions that more closely approximate normal field conditions. If, for example water depletion is a factor that can activate TEs, the Ty1-*copia* elements displaying the highest relative expression (**Table 3.4**) are good candidates to explore this stress in the future for TE activation.

Lastly, similarity between LTRs of the same element, and domain conservation (Table 3.4) indicated that most recently inserted elements also had most of their domains conserved, and are the most likely to be active. However, some of the retrotransposons with RPKMs > 5000 did not have LTR similarities close to 100%, and were missing at least two of the five protein coding domains, which seems counterintuitive. The TE located in scaffold 116 was much shorter than the average (5.3 kb) found previously for flax *copia*-type retrotransposons [218], and therefore the absence of domains could be due to internal deletions that may not necessarily impair transcription. Three other TEs from scaffolds 257, 426 and 480 were extremely large (Table 3.4), and are probably carrying foreign regions that have been captured through processes like transduplication and recombination [75,129], and may therefore align to reads that also align to non-TE genes, elsewhere in the genome. In this case the RPKM calculated could be inaccurate, although it could be argued that a captured gene section becomes a structural part of a TE. A more in-depth bioinformatics analysis will have to be performed to find what is contained in the regions between the two LTRs on these long TEs. Therefore, the Ty1-*copia* elements from families RLC_Lu2, RLC_Lu8 and RLC_26 (Table 3.4) constitute the most likely candidates to study other stresses since they have high LTR similarity, domain conservation, a close to average size, and seem to be modulated at least over the time course.

A complete analysis that includes bioinformatics predictions along with a large-scale study can give clues of which TEs may be regulated, and become candidates to test new stress elicitors.

However, the difficulty in mapping reads and of analyzing TEs on a case-by-case basis remains, due to the highly repetitive and mutational nature of the elements.

3.6 Conclusions

Our expression analysis showed that it is better to study TEs individually (as opposed to TE family evaluation) when trying to assess potential elicitors of expression, but even when TEs are evaluated separately, there is a chance that neither the elicitor nor the experimental design are able to detect changes in transcription. Likewise, the prediction of TFBS' can give clues on potential regulators, but since so many can be detected in LTRs, prediction of the potential elicitor to factor into an experiment becomes difficult. The transcriptome-wide study could be the best approach from among the ones used here, to find candidates for a more in-depth analysis. We detected some TEs with high normalized transcription values (RPKMs) in this experiment. These were not regulated by the fungal stress, but a change through time and/or due to either water or nutrient depletion is suggested as a possible stress to be evaluated in future studies. On these candidates the prediction of TFBS' in the LTRs would be a first step to confirm potential elicitors to evaluate, but an experiment which includes methylation mutants would also aid in evaluating which TEs are under epigenetic control. If an elicitor can be directly associated with changes in TE expression, then promoter deletion analysis would allow us to tag the most important controlling factors.

While most of the elicitors assayed here were unsuccessful in regulating TE expression, there were some differences in tissue expression, indicating that this might be a much better approach to detecting transcriptional variation. Most studies which have detected tissue-specific expression of TEs evidenced meristematic and reproductive tissues as potential sites of TE activation [173,308,309,312], in part due to epigenetic reprogramming happening in these tissues. Future studies should focus on sampling specific meristematic tissues with parallel evaluation of methylation status.

CHAPTER 4 - RNA-seq transcriptome response of flax (*Linum usitatissimum* L.) to the pathogenic fungus *Fusarium oxysporum* f.sp. *lini*.

This chapter is based on a published article: Galindo-González L. & Deyholos M.K. 2016. RNA-seq transcriptome response of flax (*Linum usitatissimum* L.) to the pathogenic fungus *Fusarium oxysporum* f.sp. *lini*. Frontiers in Plant Science. 7:1766.

4.1 Abstract

Fusarium oxysporum f. sp. *lini is* a hemibiotrophic fungus that causes wilt in flax. Along with rust, fusarium wilt has become an important factor in flax production worldwide. Resistant flax cultivars have been used to manage the disease, but the resistance varies, depending on the interactions between specific cultivars and isolates of the pathogen. This interaction has a strong molecular component (resistance of the plant depends on the interaction of its gene products with pathogen elicitors), but no genomic information is available on how the plant responds to attempted infection to inform breeding programs on potential candidate genes to evaluate or improve resistance across cultivars. In the current study, disease progression in two flax cultivars (CDC Bethune and Lutea), showed earlier disease symptoms and higher susceptibility in the latter cultivar. Chitinase gene expression was also divergent and demonstrated an earlier molecular response in Lutea. The most resistant cultivar (CDC Bethune) was used for a full RNA-seq transcriptome study through a time-course at 2, 4, 8 and 18 days post-inoculation (DPI). While over 100 genes were significantly differentially expressed at both 4 and 8 DPI, the broadest deployment of plant defense responses was evident at 18 DPI with transcripts of more than 1,000 genes responding to the treatment. These genes provided evidence of a reception and transduction of pathogen signals, a large transcriptional reprogramming, induction of hormone signalling, activation of pathogenesis-related (PR) genes, and changes in secondary metabolism. Among these, several key genes that consistently appear in studies of plant-pathogen interactions had increased transcript abundance in our study, and constitute suitable candidates for resistance breeding programs. These included: an *R*PMI-induced protein kinase (*RIPK*); transcription factors *WRKY3*, *WRKY70*, *WRKY75*, *MYB113* and *MYB108*; the ethylene response factors ERF1 and ERF14; two genes involved in auxin/glucosinolate precursor synthesis (*CYP79B2* and *CYP79B3*); the flavonoid-related enzymes chalcone synthase, dihydroflavonol reductase and multiple anthocyanidin synthases; and a peroxidase implicated in lignin formation (*PRX52*). Additionally, regulation of some genes indicated potential pathogen manipulation to facilitate infection. These included: four disease resistance proteins that were repressed; indole acetic acid amido/amino hydrolases, which were upregulated; activated expansins and glucanases, amino acid transporters and aquaporins; and finally, repression of major latex proteins.

4.2 Introduction

Flax (*Linum usitatissimum*) is an important crop for the production of fiber, oil, and nutraceuticals [3]. Among flax diseases caused by fungal pathogens, fusarium wilt, caused by *Fusarium oxysporum* f. sp. *lini* (*Foln*) has been an important factor limiting yield of this plant. Fusarium wilt was identified as a major flax disease problem in North America at the beginning of the 20th century [14]. *Fusarium* is a genus of filamentous, seed and soil-borne ascomycetes with numerous pathogenic members that have been reported to cause disease in over 100 major crop species worldwide [314]. Besides wilt disease, it can also produce rots, blights and cankers through invasive growth and the production of mycotoxins, using mainly a hemibiotrophic infection strategy [314]. Infection occurs through the roots, invading the water-conducting tissues, which impairs water transport and results in wilting, necrosis and chlorosis of aerial parts [14,314]. *Fusarium oxysporum* can persist in the soil for 5-10 years [14], which allows recurrent infections if soil and residues are not treated and if no crop rotation is implemented. While the generation of fusarium-resistant cultivars worldwide has reduced the impact of the pathogen, there is a wide range of susceptibility among varieties, dependent in part on the specific fungal isolates/races involved in infection [315].

Previous studies of interactions between flax and fusarium have focused on disease symptomology [316], physiology and the fungal colonization process [223,224,317], and molecular and metabolic responses [220,221,224,318,319]. Techniques that have been applied to study the infection process include transformation [320–323], tissue culture [324], and QTL analysis [325]. To date there have been no transcriptome-scale studies of the response of flax to *F. oxysporum* f. sp. *lini*, which limits information that can be used to further breeding improvements.

RNA-seq studies of plant responses to fungal pathogens have been performed in numerous pathosystems including: lettuce infected by *Botrytis cinerea* [326], *Arabidopsis thaliana* after treatment with *Pseudomonas syringae* [327], banana roots in response to *Fusarium oxysporum* [328–330], chrysanthemum leaf after infection with *Alternaria tenuissima* [304], a wheat resistant variety affected by *Fusarium graminearum* [331], and the early infection of peach leaves by *Xanthomonas arboricola* [332].

The sequencing and annotation of the flax genome [15] has unlocked new genomic tools that can be used for whole genome scale studies. The flax genome sequence was based on the

cultivar CDC Bethune, which is a highly inbred, elite oilseed cultivar widely grown in Canada [333]. Furthermore, CDC Bethune has been classified as moderately resistant to fusarium wilt [333], although other studies have found higher levels of susceptibility [334], which supports the need to investigate the resistance mechanisms of this elite cultivar.

Here, we present a multi-level study of the progression of *Foln*-induced responses in CDC Bethune contrasted with Lutea, which is an exemplar of a less-resistant cultivar. The relative susceptibility of the two cultivars was demonstrated by monitoring disease symptoms following inoculation with *F. oxysporum* f. sp. *lini*, and by measuring changes in chitinase transcript expression as a marker of defense responses. Finally, we conducted RNA-seq analysis on CDC Bethune, following infection by *F. oxysporum* f. sp. *lini*. Besides the deployment of a full defense response from the plant at the end of the evaluated time course, several genes had unexpected patterns of regulation which supported cell growth, weakening of the cell wall and favored fungal penetration, and may be indicative of partial manipulation of host genes by the pathogen. The genes identified can be used to inform breeding programs and improve understanding of molecular mechanisms underlying fusarium resistance.

4.3 Materials and methods

4.3.1 Plant material

Seeds from flax cultivars CDC Bethune and Lutea were grown according to the protocol of Kroes (1998b) with some modifications: sterilized seeds from each cultivar were grown in sterile 25 x 200 mm glass tubes filled with 5 mL of 10% Murashige-Skoog solution (MS basal medium Sigma-Aldrich, MO, USA) pH 5.8 and 2 g of vermiculite (**Figure 4.1**). Tubes were placed in a growth chamber at 22°C with 16 h day / 8 h night (light intensity = 167 µMol).

4.3.2 Pathogen

Fusarium oxysporum f. sp. *lini* isolates (#65 and #81) were kindly provided in potato dextrose agar (PDA) by Khalid Rashid (Agriculture and Agri-Food Canada, Morden, Manitoba). Isolate #65 is from the Indian Head Saskatchewan flax nursery, and isolate #81 was obtained from a farmer's field in Treherne, Manitoba. We grew *F. oxysporum* isolates in PDA (39 g/L) plates at 21°C under 12 h dark / 12 h light cycles. Cultures were started on three consecutive days, and viability was assessed by counting the percentage of germinated spores at 1, 4, 8 and

24 hours when the initiated cultures reached 13, 14 and 15 days, to select the best culture for inoculation. Spores (a mix of macro and microconidia) from isolates were harvested after flooding the plate with 15 mL of sterile water and a sterile inoculation loop was used to detach the mycelium/spores from the surface of the media. Spore count was performed using a haemocytometer. Spore suspensions were diluted to 10^5 spores mL^{-1} to perform the inoculations.

4.3.3 Comparison of cultivar response

Plants grown in test tubes (described above) until cotyledon expansion, were either inoculated with 1 mL of 10^5 spores mL^{-1} of the fungal isolates or with 1 mL of sterile water (control) directly on the surface of the vermiculite under sterile conditions. Disease symptoms and shoot length were recorded at 1, 8 and 22 days post-inoculation (DPI) for 7-10 plants from each treatment (control, isolates #81 and #65) in each cultivar. Plants were removed from vermiculite and roots were cleaned with sterile water and dried. Sections of 3-5 cm from root tips were taken from four plants of each treatment for microscopy and fungal isolation from infected plants (see below). Entire seedlings were placed in 2 mL tubes and flash-frozen in liquid nitrogen for further processing.

4.3.4 Fungal isolation from infected plants

To confirm that symptoms were a result of the fungal infection, *F. oxysporum* f. sp. *lini* was reisolated from the plant roots. Collected root sections (3-5 cm) were surface sterilized in 10% sodium hypochlorite for 30 seconds and then rinsed three times in sterile distilled water. Root sections were further cut into 3-5 mm sections and air-dried on Whatmann paper. Four to six of these root sections were transferred to sterile Komada medium [335] and grown for seven days at 22°C under 8 h dark / 16 h light cycles. Plates were examined for growth, and colonies were subcultured in PDA for 14 additional days at 21°C under 12h dark / 12 h light cycles.

4.3.5 Microscopy

To examine fungal penetration of plant tissues, we collected root sections of 3-5 mm in length that were fixed in FAA (3.7% Formaldehyde, 5% Acetic acid- 50% Alcohol), then dehydrated in an ethanol series (50 and 70%) and embedded in paraffin blocks using the TISSUE TEK II embedding center (Sakura, Torrance, CA, USA). Sections of 8 and 12 µm were cut from

the blocks using a RM2125 microtome (Leica, Wetzlar, Germany), and stained with 0.5% (w/v) Toluidine Blue. Sections were observed with a Leica DMRXA microscope (Meyer Instruments, Houston, TX, USA), photographed with the incorporated QI Click digital camera and captured using the Q Capture Pro 7 software (Q Imaging, Surrey, BC, Canada).

4.3.6 RNA extraction and cDNA synthesis

Entire plants collected from the time course were used for RNA extraction and cDNA synthesis to evaluate gene expression. Tissue was ground in 2 mL collection tubes with a 5.5 mm stainless steel bead, using a Mixer Mill MM 301 (Retsch, Haan, Germany). RNA was extracted using the RNeasy Plant Mini Kit (QIAGEN, Venlo, Netherlands), followed by a DNAse I treatment (Ambion-Life Technologies, Carlsbad, CA, USA). RNA quality was checked with a 2100 Bioanalyzer (Agilent, Mississauga, ON, Canada), and cDNA was synthesized with 250 ng of RNA using the RevertAid H Minus Reverse transcriptase using oligo dT (18) (Thermo Scientific, Waltham, MA, USA). Presence of contaminating genomic DNA was tested by PCR analysis using pectinesterase gene primers (**Table 4.1**), which give two distinct bands of 123 bp and 848 bp for cDNA and genomic DNA respectively.

4.3.7 Quantitative reverse transcription PCR (qRT-PCR)

To test the defense response of the two cultivars, selected flax chitinases were chosen as orthologs of genes previously characterized in *Arabidopsis thaliana* [288]. Four chitinases (chitinase-like CTL2, 4, 10 and 11 – **Table 4.1**) from Glycosyl Hydrolase family 19 (GH19) were selected to test the response to the pathogen (Mokshina *et al.* 2014). To select reference genes, we tested primers from six genes for stability upon our treatments from a list of 13 genes previously published as normalizers in qRT-PCR experiments in flax [228]: Elongation factor 1-α (EF1A), Eukaryotic translation initiation factor 3E (ETIF3E), Eukaryotic translation factor 5A (ETIF5A), Glyceraldehyde 3-phosphate dehydrogenase (GAPDH), Ubiquitin (UBI) and Ubiquitin extension protein (UBI2) (**Table 4.1**). The most stable reference genes after performing the analysis with Bestkeeper [247] and GeNorm [248] were: GAPDH, ETIF3E and UBI2.

Table 4.1 Primers qRT-PCR.

Primer name	Sequence	Gene annotation	Gene alias
Lus10037737_F	GCTCTCAGCGATCCTACTGC	chitinase	LusCTL2
Lus10037737_R	TCGAAGACGATGCCGATT	chitinase	LusCTL2
Lus10037430_F	TTGGTGAACTTGTTGGCAGT	chitinase	LusCTL4
Lus10037430_R	CTTCCCCTTCACCTTCTTCA	chitinase	LusCTL4
Lus10028377_F	AACAGAGTTCCCGGCTACG	chitinase	LusCTL10
Lus10028377_R	GCCACGTCCACATTCAAGA	chitinase	LusCTL10
Lus10041831_F	CGTCCATCCATAGCGTGATT	chitinase	LusCTL11
Lus10041831_R	TACCCGGGAACTCTGTTGG	chitinase	LusCTL11
EF1-A-fw	GCTGCCAACTTCACATCTCA	Elongation factor 1-α	EF1A
EF1-A-rv	GATCGCCTGTCAATCTTGGT	Elongation factor 1-α	EF1A
ETIF3E-fw	TTACTGTCGCATCCATCAGC	Eukaryotic translation initiation factor 3E	ETIF3E
ETIF3E-rv	GGAGTTGCGGATGAGGTTTA	Eukaryotic translation initiation factor 3E	ETIF3E
ETIF5A-fw	TGCCACATGTGAACCGTACT	Eukaryotic translation factor 5A	ETIF5A
ETIF5A-rv	CTTTACCCTCAGCAAATCCG	Eukaryotic translation factor 5A	ETIF5A
GADPH-fw	GACCATCAAACAAGGACTGGA	Glyceraldehyde 3-phosphate dehydrogenase	GADPH
GADPH-rv	TGCTGCTGGGAATGATGTT	Glyceraldehyde 3-phosphate dehydrogenase	GADPH
UBI-fw	CTCCGTGGAGGTATGCAGAT	Ubiquitin	UBI
UBI-rv	TTCCTTGTCCTGGATCTTCG	Ubiquitin	UBI

Primer name	Sequence	Gene annotation	Gene alias
UBI2-fw	CCAAGATCCAGGACAAGGAA	Ubiquitin extension protein	UBI2
UBI2-rv	GAACCAGGTGGAGAGTCGAT	Ubiquitin extension protein	UBI2
PME-32F	CATGGTGGTCGGTTTGTG	Pectinesterase	PME32
PME-32R	GTCGATCGCCATGAATCC	Pectinesterase	PME32
Lus10015351-N_tr_fw	CAGGTCCTTGTCGCTGCTTC	nitrate transporter downregulated	Nit_trp
Lus10015351-N_tr_rv	GCTCTGTACTGATTGCTGCC	nitrate transporter downregulated	Nit_trp
Lus10021936-_7s_glob_fw	TTCGGAGATGGCCCTTATGTC	basic 7s globulin like downregulated	7s glob-like
Lus10021936-_7s_glob_rv	GCTGTGCTCACCTTGTTCAG	basic 7s globulin like downregulated	7s glob-like
Lus10016424-sim_AER92600_fw	TCATCATGGCGTCGATCTTGTA	conserved proteins similar to AER92600 downregulated	sim_AER92600
Lus10016424-sim_AER92600_rv	ACTCATCCCACCACCACCTT	conserved proteins similar to AER92600 downregulated	sim_AER92600
Lus10005393-POX_fw	TGCTCTAGTCGACCCATTCTCCA	polyphenol oxidase downregulated	POX
Lus10005393-POX_rv	TCTTGAACTCCCTCCCCGCA	polyphenol oxidase downregulated	POX
Lus1001069-mlp423_fw	CAAGGTGATGTGGAGAAGTTAGAA	mlp like protein 423 downregulated	mlp423
Lus1001069-mlp423_rv	CCTGTCGTGACGCTTCTTCT	mlp like protein 423 downregulated	mlp423
Lus10039487-iaa7_fw	TACTGCCCAAACGAACTCATCTA	auxin responsive protein iaa7 downregulated	iaa7
Lus10039487-iaa7_rv	TCCATTATTATCTTCATCGCCGG	auxin responsive protein iaa7 downregulated	iaa7
Lus10041830-chit_fw	CCGCTGCTAGGTCCTTCAAC	chitinase upregulated	chitA

Primer name	Sequence	Gene annotation	Gene alias
Lus10041830-chit_rv	TGTCTCCCTAATGAAACAATACCC	chitinase upregulated	chitA
Lus10028377-chit_fw	CTACTGCTGGGTCCTTCAAT	chitinase upregulated	chitB
Lus10028377-chit_rv	GCTCCTTCTTACGGGTGTC	chitinase upregulated	chitB
Lus10004808-leuc_diox_fw	GACTTCAAGTGCGGAAAGACA	leucoanthodyanidin dioxygenase upregulated	leuc_diox
Lus10004808-leuc_diox_rv	TCTTGTAGACCGCGTTGCTA	leucoanthodyanidin dioxygenase upregulated	leuc_diox
Lus10020826-peroxid_fw	GATGCCAAGACTCAGCTCGAAA	peroxidase upregulated	peroxid
Lus10020826-peroxid_rv	CCGTCTCTTCGTCCCGTG	peroxidase upregulated	peroxid
Lus10019060-GH_chit_fw	CAGTTTATGACCTTTACCCAGACA	glycosyl hydrolase family protein with chitinase insertion domain upregulated	GH_chit
Lus10019060-GH_chit_rv	ACGTTAGCTCCACCGCCT	glycosyl hydrolase family protein with chitinase insertion domain upregulated	GH_chit
Lus10016836-PR_sth2_fw	TGTGACCCGCGACATACAG	pathogenesis related protein sth 2 upregulated	PR_sth2
Lus10016836-PR_sth2_rv	TCGACCATTGTGTACTTGCATAC	pathogenesis related protein sth 2 upregulated	PR_sth2
Lus10008930-mlp_fw	TCCTTCCAATATTCAGGCTGTCA	major latex protein upregulated	mlp
Lus10008930-mlp_rv	ACATCTCCTTCTAAGCCGTTCAG	major latex protein upregulated	mlp
Lus10027702-ETIF3_sub_C_fw	ATCTGACGAGTCTACTGATGAGG	eukariotic translation initiation factor 3 subunit c non-regulated	ETIF3_sub_C
Lus10027702-ETIF3_sub_C_rv	ACGTCCCAGGTAATTTCGCT	eukariotic translation initiation factor 3 subunit c non-regulated	ETIF3_sub_C

Primer name	Sequence	Gene annotation	Gene alias
Lus10025438-TEF1_fw	CTTGATTGGTGAAGCTTCGTGT	transcription elongation factor 1 non-regulated	TEF1
Lus10025438-TEF1_rv	CCATACGCAGCAGAGCACTA	transcription elongation factor 1 non-regulated	TEF1
Lus10015458-RNA_pol_fw	GAGGGGAAAAGGTGTGTTTGG	dna directed rna polymerase i subunit rpa12 non-regulated	RNA_pol
Lus10015458-RNA_pol_rv	TGCTGCATTTCTCACACTGC	dna directed rna polymerase i subunit rpa12 non-regulated	RNA_pol
Lus10005425-treh6P_synt_fw	AGGCTGAGATTGAGGAGAGTTG	trehalose 6 phosphate synthase non-regulated	treh6P_synt
Lus10005425-treh6P_synt_rv	ACATTATAGTAAGCTGCTCGTTCG	trehalose 6 phosphate synthase non-regulated	treh6P_synt
Lus10038622-ETIF3_sub_l_fw	CTTGGAAAGCTTGCGAATTACTTG	eukaryotic translation factor 3 subunit l like non-regulated	ETIF3_sub_l
Lus10038622-ETIF3_sub_l_rv	ATGATCTTCCTGTCAGAATCAACC	eukaryotic translation factor 3 subunit l like non-regulated	ETIF3_sub_l

QRT-PCR experiments were run on a QuantStudio 6 Flex Real-Time PCR system (Applied Biosystems-Life Technologies, Carslbad, CA, USA) after samples were aliquoted using a Biomek 3000 Laboratory Automation System (Beckman Coulter, Brea, CA, USA). Reactions were performed in 10 µL with 5 µL of SYBR-green, 2.5 µL of the pair of mixed primers (3.2 µM) and 15 ng of cDNA (2.5 µL of a 1:40 dilution of the synthesized cDNA). Cycling conditions were: 95°C for 2 minutes followed by 40 cycles of 95°C for 30 seconds, 60°C for 1 minute. A melting curve stage was added: 95°C for 15 seconds, 60°C for 1 minute and 95°C for 15 seconds.

Experiments were performed with four biological replicates (one plant = one replicate) for each combination of treatment and time point and three technical replicates for each biological replicate. The geometric mean of the three reference genes selected was used to perform relative quantification of expression using the $2^{-\Delta\Delta CT}$ method [289]. Statistical analysis to find significant differential expression was performed using *t*-tests with an Excel macro.

4.3.8 CDC Bethune transcriptome response

4.3.8.1 Experimental design

Full transcriptome response and the progression of molecular events were assessed for CDC Bethune plants that were either inoculated with the most aggressive fungal isolate (#81) or with sterile water following the procedures outlined above. Harvesting was performed at 2, 4, 8 and 18 DPI and six biological replicates were collected for each treatment and time point combination. Disease symptoms were scored and plants were harvested and frozen in liquid nitrogen for RNA extractions as aforementioned. RNA samples were pooled in groups of three (to decrease variability), resulting in two pooled biological replicates per treatment and time point, which were used for RNA-seq.

4.3.8.2 RNA-seq

Twenty-seven micrograms of RNA per pooled sample were sent to the Beijing Genomics Institute (BGI) for sequencing. In brief: total RNA was enriched using oligo (dT) magnetic beads, and then fragmented into short fragments (200bp). The first strand was synthesized using random hexamers, prior to second strand synthesis. The double stranded cDNA was purified using the QiaQuick PCR purification kit (QIAGEN, Venlo, Netherlands), and washed with EB

buffer (from the kit) for end repair and addition of base A. Sequencing adapters were ligated to the fragments, before agarose gel electrophoresis purification and enrichment via PCR. The library products were sequenced using an Illumina HiSeq 2000 (Illumina, San Diego, CA, USA) as single-end reads. Raw reads in fastq format were filtered with an in-house pipeline to remove adaptors, remove reads with unknown bases (more than 5%), and remove low quality reads (reads with more than 50% of bases with a quality value equal or less than 10). Reads were deposited in the NCBI sequence read archive as accession PRJNA232613.

Reads from each filtered fastq file were mapped to the flax genome and the flax genome gene models produced previously by our group [15] using TopHat [337]. Mapped reads and the gene models file were used as input for cufflinks v2.2.1 to generate transcripts and quantify differential expression. Cufflinks was run with the GTF-guide option using the previously annotated gene models file; fragment bias correction and multi-read correction were also applied. The gtf files from transcripts from all treatment and replicates were combined using cuffmerge. Cuffquant was performed using the merged file and cuffdiff was performed comparing the water controls to the inoculated plants in each one of the four time points post-inoculation. The levels of expression were quantified using Fragments Per Kilobase of transcript per Million fragments mapped (FPKM) and significant differential expression was assessed using the Benjamini-Hockberg correction for multiple comparisons [290].

To validate differential expression of the genes, qRT-PCR was performed using primers for five mainly upregulated, five mainly downregulated and five genes with no change in expression (**Table 4.1**), using the same qRT-PCR conditions previously mentioned.

4.3.8.3 *In*-silico analyses

To find gene regulation changes between inoculated and control plants, a systematic process of transcript annotation and differential expression analysis was performed. Since many new unannotated transcripts were found after the RNA-seq analysis, we performed an annotation of close to 50,000 transcripts, which included unannotated and previously annotated flax genes. The merged gtf file from cuffmerge bearing all transcripts was used along the genome fasta file as input for Transdecoder (https://transdecoder.github.io/), which uses a Perl script to construct a fasta file of all transcripts. The fasta file was parsed to obtain only the longest isoform from each gene for further annotation.

We annotated all transcripts using 12 cores (physical memory=2000mb/core) on a server at Westgrid/Compute Canada (https://www.westgrid.ca/ - https://www.computecanada.ca/). We performed blastx against the non-redundant (nr) Genbank database, the two databases from Uniprot (Trembl and Swissprot) and the TAIR10 protein release. We restricted our search to a maximum of 20 hits with an e-value threshold of 10^{-10}. The XML output file was loaded into blast2go [338], where the description of the blast hits in each case was compiled as the most common term found in the 20 resulting top hits for each transcript.

The TAIR10 hit IDs from significant differentially expressed genes ($q < 0.05$, after Benajmini-Hockberg multiple testing correction) were used as input for gene ontology (GO) enrichment analysis using AgriGO [246]. The parameters were as follows: species – *Arabidopsis thaliana*, statistical test – hypergeometric distribution, multi-test adjustment – Yekutieli, significance level – 0.05, minimum number of mapping entries – 5. As background we used a compiled list of all the RNA-seq transcripts that had at least 10 read alignments. We also used plantGSEA [339] to find enriched pathways (PlantCyc gene sets and KEGG) using the same parameters as for AgriGO. To see the gene expression changes of all differentially expressed genes at any time point we used multi-experiment viewer MeV4.9 [340].

4.4 Results

4.4.1 Differential response of two flax cultivars to *F. oxysporum* f. sp. *lini*

CDC Bethune is an elite, brown-seeded linseed cultivar of flax that is widely grown in Canada and has been reported to have moderate resistance to fusarium wilt [333]. To confirm that CDC Bethune was relatively resistant, we conducted preliminary experiments with a panel of linseed varieties selected in consultation with a flax pathologist (Khalid Rashid*, personal communication*), and identified Lutea (a yellow-seeded variety) as a candidate cultivar that could differ in fusarium wilt resistance from CDC Bethune. We inoculated both cultivars with two *F. oxysporum* f. sp. *lini* isolates (#65 and #81) that demonstrated high spore viability/germination (not shown). CDC Bethune plants generally did not show any symptoms until 22 DPI, but wilting was evident in plants inoculated with isolate #81 (**Figure 4.1**). In Lutea plants, disease symptoms appeared earlier (8 DPI) than in CDC Bethune (22 DPI) and consequently the disease state was more advanced at 22 DPI, with some plants having undergone complete necrosis (**Figure 4.1F**). Disease symptoms recorded at 22 DPI included yellowing of leaves, brown spots

on leaves, wilting, necrosis and root browning. While most of these characteristics were variable and some infected plants presented little or no symptoms, root browning (represented as a general brown-ashy appearance indicative of rot) was a consistent symptom of disease in both cultivars and with both fungal isolates (**Figure 4.2A**). When using shoot length to assess the influence of the fungal inoculations on plant growth [316], there was a significant difference between the shoot lengths of control plants when compared with the lengths of isolate #81 inoculated Lutea plants (**Figure 4.2B**); nevertheless both cultivars had a 13% shoot length reduction when inoculated with isolate #81 at 22 DPI. Together, these results showed that isolate #81 was the most aggressive *F. oxysporum* f. sp. *lini* isolate, and that CDC Bethune was more resistant to *F. oxysporum* f. sp. *lini* than Lutea, under our experimental conditions. We were able to re-isolate the fungus from surface-sterilized roots of previously inoculated plants of both CDC Bethune and Lutea (**Figure 4.3**). Spore morphology was consistent with the original inocula (**Figure 4.4**). As further evidence of infection, we also stained sections of inoculated roots with toluidine blue. Hyphal development in root sections was advanced at 22 DPI, at which point hyphae had colonized the cortical cells and penetrated xylem vessels (**Figure 4.5**).

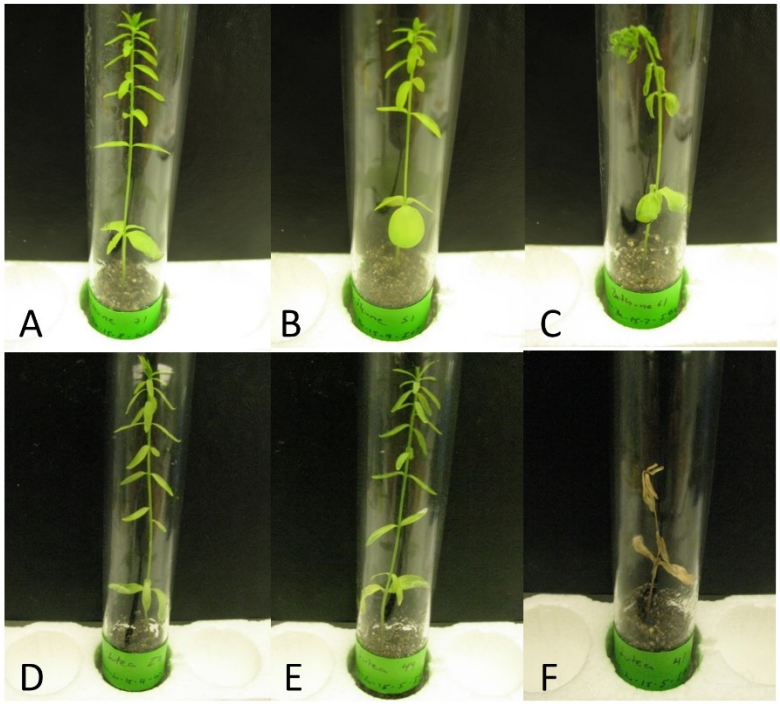


Figure 4.1 Disease symptoms 22 DPI in flax cultivars. CDC Bethune (A-B-C) and Lutea (D-E-F). A and D: Control plants treated with water. B and E: Plants inoculated with isolate #65. C and F: Plants inoculated with isolate #81.

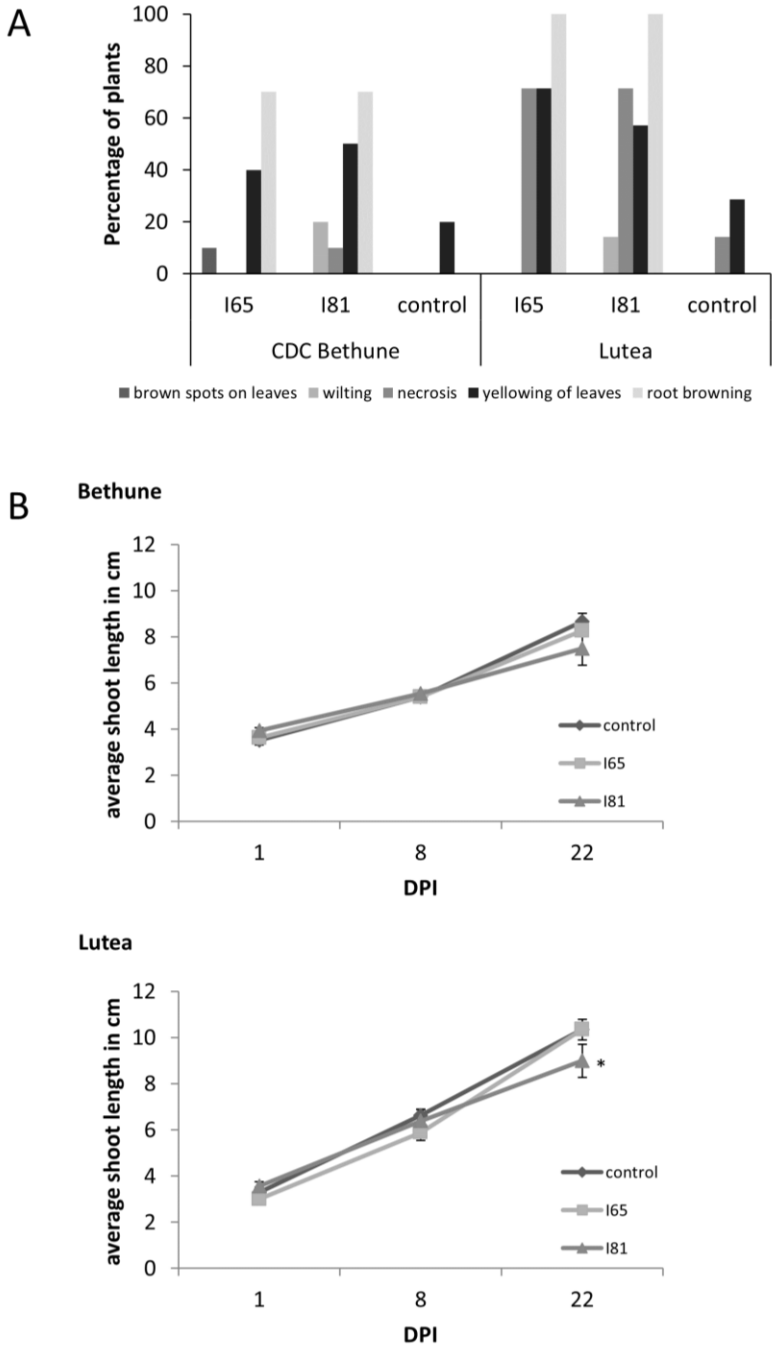
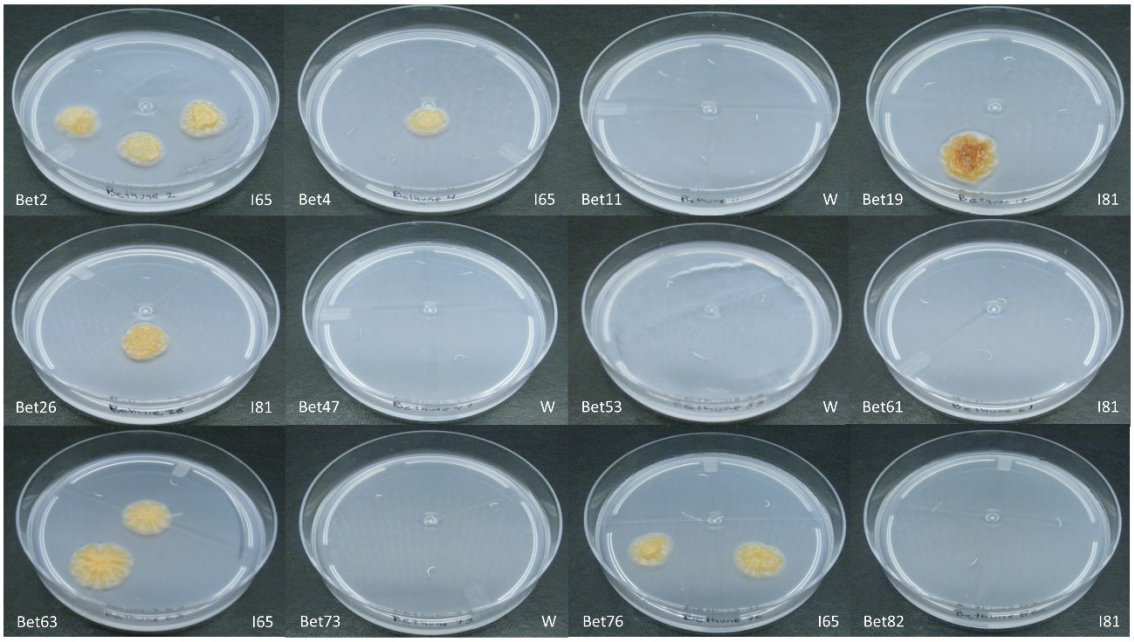
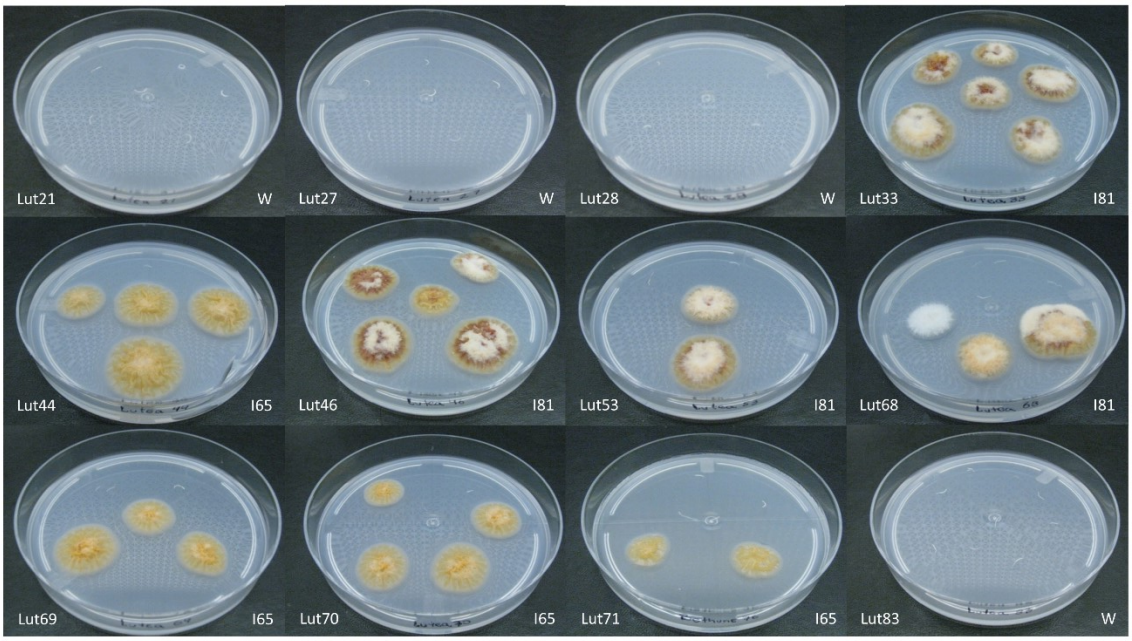


Figure 4.2 Disease symptoms and changes in shoot length. A. Percentage of plants presenting disease symptoms 22 DPI in the flax cultivars CDC Bethune and Lutea due to the infection with two isolates (#65 and #81) of *Fusarium oxysporum* f.sp. *lini*. B. Difference in average shoot length between control plants and plants inoculated with isolates #65 and #81 for cultivars CDC Bethune and Lutea. An asterisk (*) denotes a significant difference between the control plants and the isolate #81 Lutea plants at day 22 (one tail *t*-test, p = 0.04). Error bars = standard error, DPI = days post-inoculation.


Bet2 I65
Bet4 I65
Bet11 W
Bet19 I81
Bet26 I81
Bet47 W
Bet53 W
Bet61 I81
Bet63 I65
Bet73 W
Bet76 I65
Bet82 I81

A


Lut21 W
Lut27 W
Lut28 W
Lut33 I81
Lut44 I65
Lut46 I81
Lut53 I81
Lut68 I81
Lut69 I65
Lut70 I65
Lut71 I65
Lut83 W

B

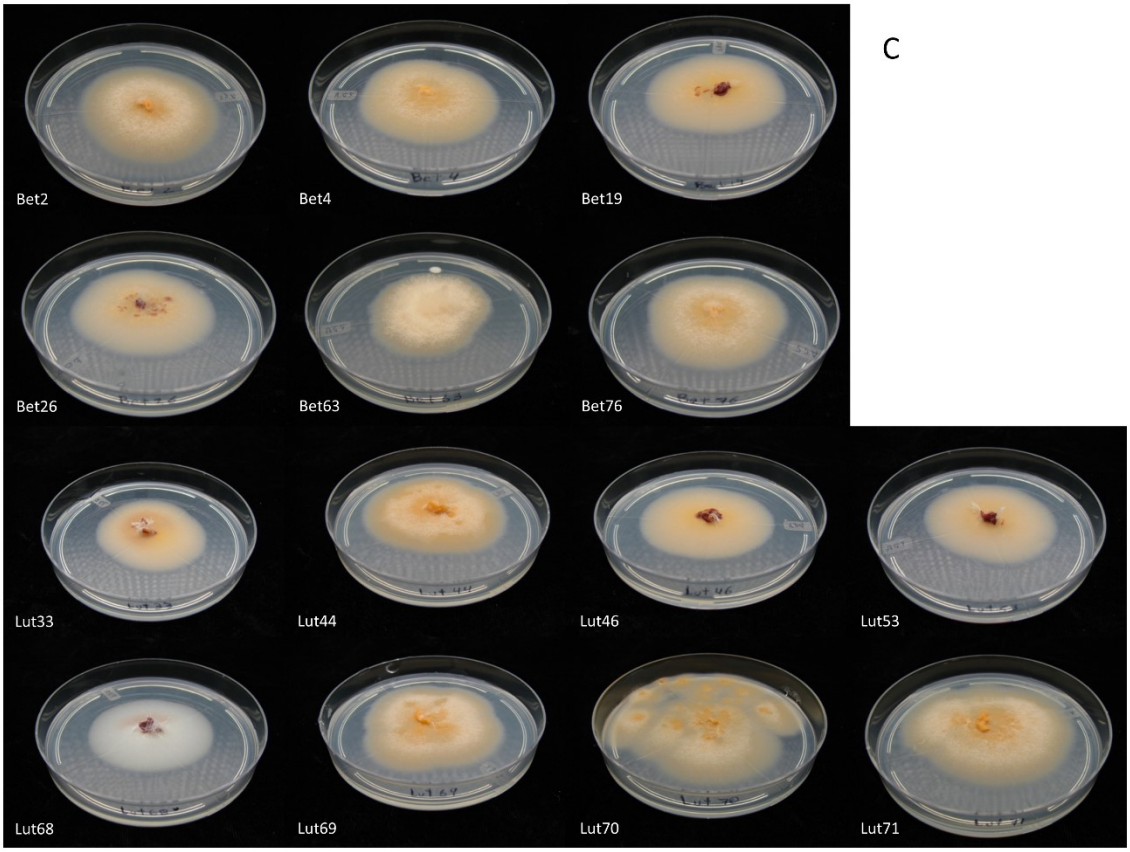


Figure 4.3 Reisolation of *F. oxysporum* from surface-sterilized roots. A. Growth in Fusarium-selective Komada medium for CDC Bethune plants. B. Growth in Fusarium-selective Komada medium for Lutea plants. C. Subculture of fungal isolates from Komada medium, grown in PDA. I65 and I81 = fungal isolates #65 and #81 respectively, W = water control.

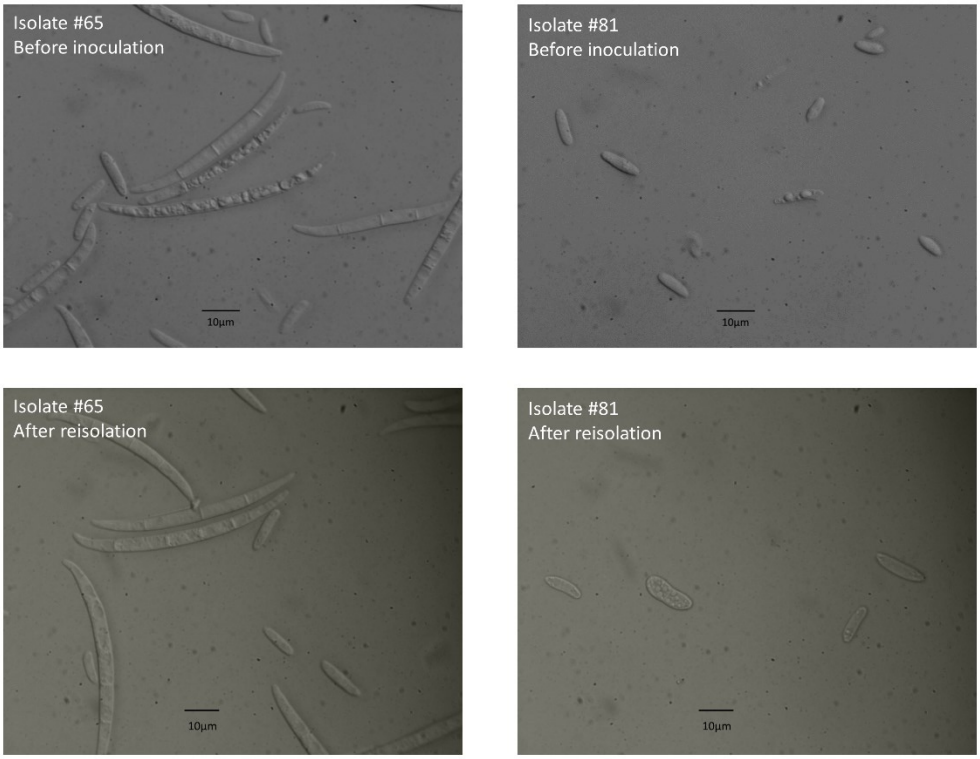


Figure 4.4 Comparison of spores used for inoculation. Before inoculation: upper pictures; spores from PDA subcultures after reisolation from infected roots: lower pictures. While both isolates had macro and microspores, isolate #65 cultures were dominated by macrospores while isolate #81 contained more microspores.

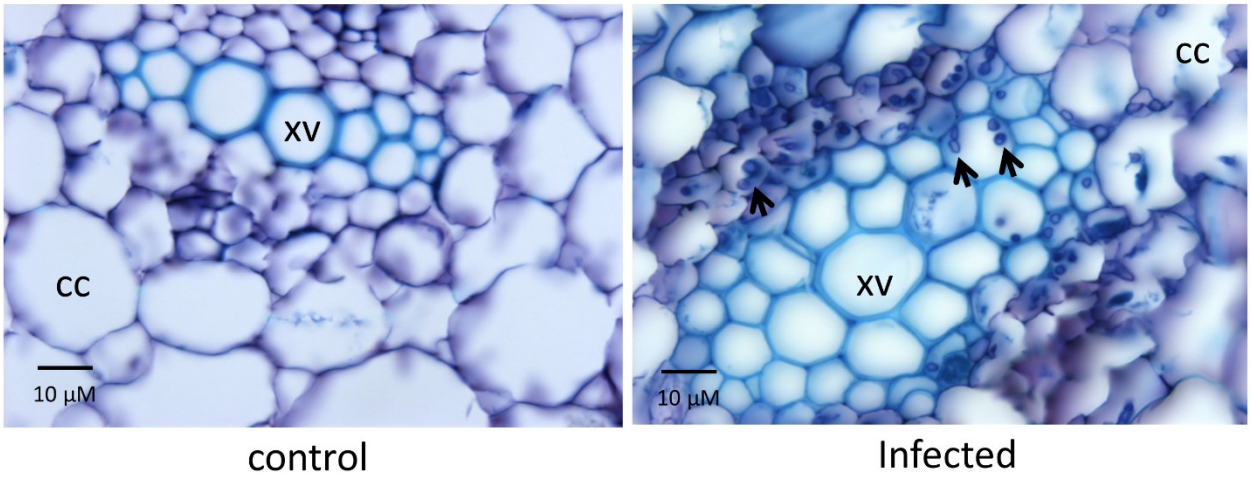


Figure 4.5 Root sections of Lutea plants 22 DPI. The control plant (inoculated with water) on the left shows no signs of infection while the treated plant (I81 inoculum) on the right has hyphae colonizing (arrowheads) the cortical cells (CC) and the xylem vessels (XV). Sections of 12 µm were stained with toluidine blue.

4.4.2 Chitinase differential expression

We next characterized a time course of molecular-scale responses to infection, in both CDC Bethune and Lutea. As markers of the response to fungal infection, we used quantitative PCR to measure transcript abundance of four Glycosyl Hydrolase family 19 (GH19) chitinase genes of flax [336]. These chitinases were selected based on homology to *A. thaliana* genes that had been previously characterized as responsive to pathogens or other related processes (**Figure 4.6**). Three of the four tested chitinases responded to the fungal inoculation (**Figure 4.7**). LusCTL4 in CDC Bethune showed a significant increase in transcript abundance at 8 DPI with both *F. oxysporum* f. sp. *lini* isolates, as compared to water controls. This chitinase also showed overexpression at 8 DPI in Lutea with isolate #65. The last two chitinases, LusCTL10 and LusCTL11, were the most responsive and over the time course appeared to increase in abundance in Lutea earlier that in CDC Bethune: chitinases peaked at 8 DPI for Lutea, while for CDC Bethune, the strongest chitinase responses to both fungal isolates occurred at 22 DPI (**Figure 4.7**).

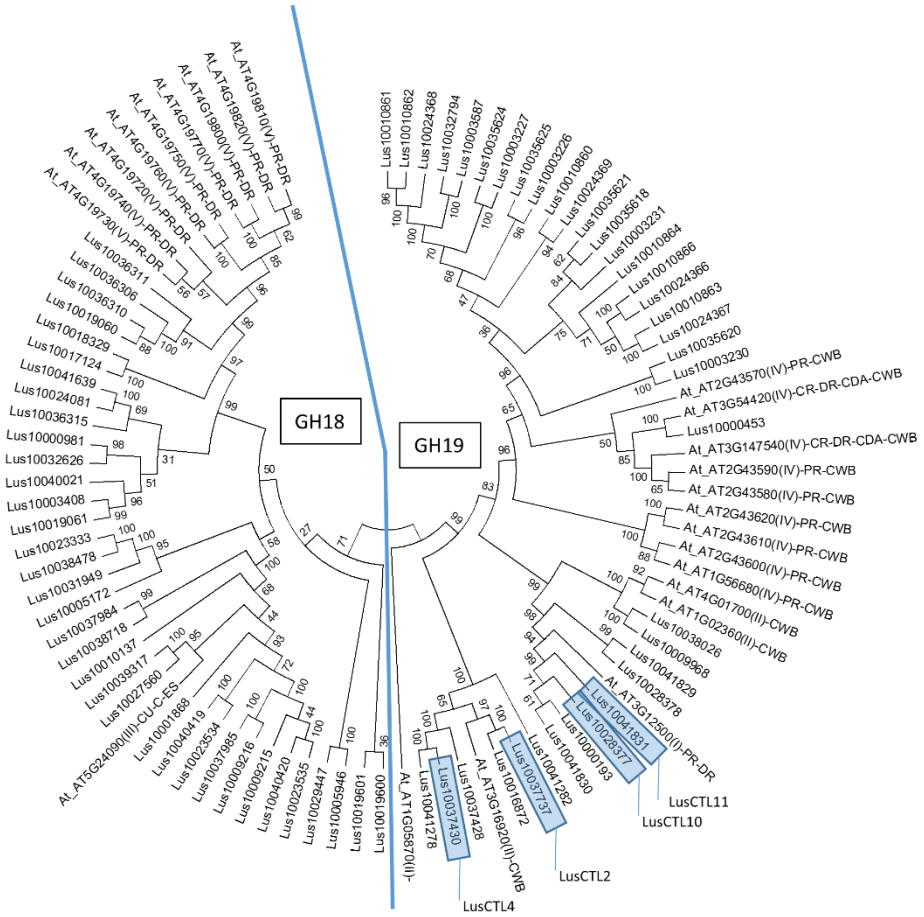


Figure 4.6 Relationship of flax chitinases with previously characterized *Arabidopsis* chitinases. The tree was done with amino acid sequences using a muscle alignment (default parameters) and the dendrogram was built under the following parameters: neighbor joining (NJ), 1000 bootstrap replicates (each branch shows final support %), p-distance and pairwise deletion. Predicted function was taken from: *Arabidopsis thaliana*: a Genomic Survey [288] **: C – cytokinesis, CDA – cell death and aging, CR – cell rescue, CU – carbohydrate utilization, CWB – cell wall biogenesis, DR – defense related, ES – extracellular secretion, PR – pathogenesis-related. Selected chitinases are outlined in rectangles and their respective labels (e.g. LuCTL14), correspond to a previous report** [336]**. Chitinase classes are in parentheses. GH = glycosyl hydrolase family.**

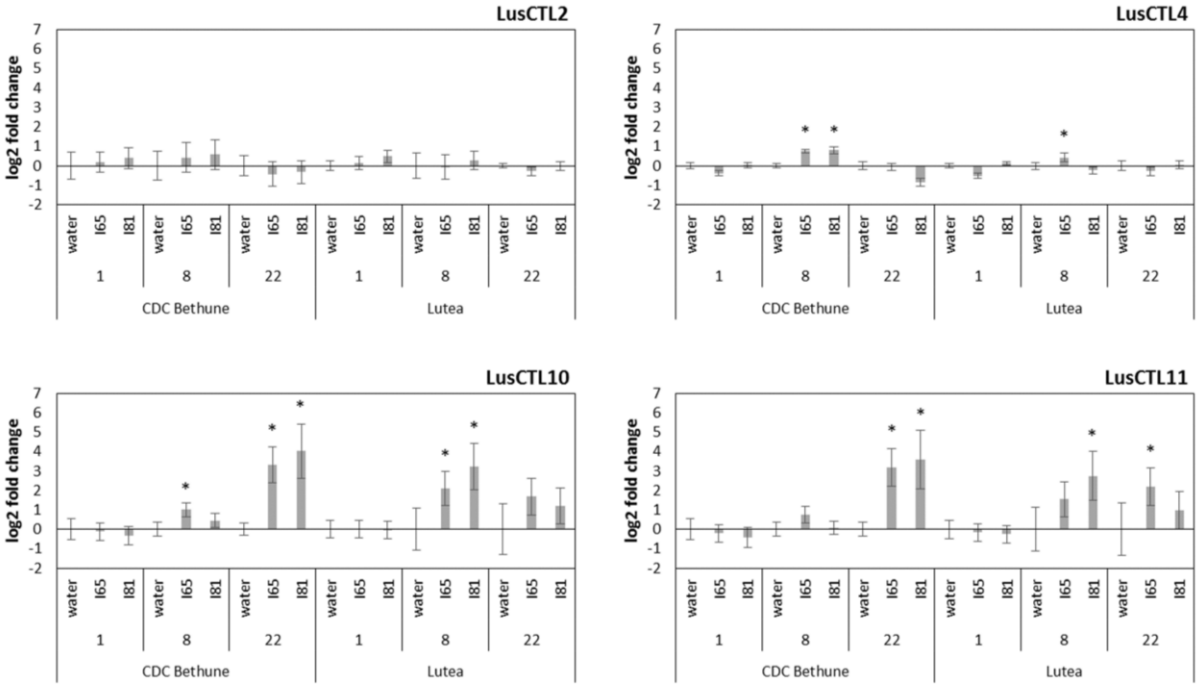


Figure 4.7 Expression changes in chitinase genes. QRT-PCR log2-fold expression changes were measured through a time course, of four chitinase genes in two flax cultivars (CDC Bethune and Lutea) upon *F. oxysporum* f. sp. *lini* inoculation. Water: control plants, I65: fungal isolate #65 inoculated plants, I81: fungal isolate #81 inoculated plants. Numbers below each treatment indicate days post inoculation. Error bars are the standard error of the ΔΔCt values (log2-fold changes) calculated as the square root of SEM^2(Δcontrol)+ SEM^2(Δtreatment). Asterisks denote significant differences between the inoculation treatment and the respective water control (one-tailed *t*-test, $p < 0.05$). Names of chitinases correspond to those referenced in Mokshina et al., 2014 [336]**.**

4.4.3 RNA-seq

Having demonstrated the relative resistance of CDC Bethune to *F. oxysporum* f. sp. *lini* inoculation, we conducted a RNA-Seq experiment to compare transcriptomes of control and inoculated plants at 2, 4, 8 and 18 DPI, following the same parameters as for our first experiment, but with additional sampling time points which could potentially capture molecular responses at higher temporal resolution. Two biological replicates of three pooled plants each were sequenced at each time point for each treatment (**Table 4.2**). Reads were mapped to a total of 49,998 transcripts, including published gene models and de-novo assembled fragments. Over 38,000 transcripts with detectable expression in each time point (transcripts had at least 10 reads aligned to each one of them – **Table 4.3**) made up a total of 40,042 non-redundant transcripts.

We used this set of transcripts for all subsequent analyses. No transcripts showed significant difference in abundance between control and treated plants at 2 DPI ($q < 0.05$), but over 100 transcripts were significantly different on each of days 4 and 8 (**Table 4.3**), and 1,043 were significant 18 DPI. While at 4 DPI there were a few more genes that decreased rather than increased in abundance, at both 8 and 18 DPI the majority of differentially expressed transcripts increased in abundance.

Table 4.2 RNA-seq statistics.

Treatment	Days post inoculation	replicate	Total number of reads	Total number of reads after filtering	TopHat mapped reads	Mapped reads %
Water	2	1	23,429,302	22,796,701	20,988,107	92.1
		2	21,010,496	20,459,410	18,891,971	92.3
	4	1	21,619,041	21,071,644	19,641,946	93.2
		2	16,789,921	16,305,931	149,99,487	92.0
	8	1	19,010,668	18,410,625	16,735,093	90.9
		2	19,515,487	19,047,923	17,806,731	93.5
	18	1	21,529,887	20,754,161	18,803,585	90.6
		2	21,542,682	20,901,401	19,238,289	92.0
Fol I81	2	1	23,080,270	22,511,052	20,812,618	92.5
		2	21,937,151	21,071,954	19,193,775	91.1
	4	1	17,545,450	16,998,485	15,443,102	90.8
		2	20,411,895	19,846,352	18,120,142	91.3
	8	1	18,541,028	18,098,407	16,694,252	92.2
		2	20,483,606	19,924,128	18,185,793	91.3
	18	1	19,615,184	18,916,943	17,036,447	90.1
		2	22,211,320	21,636,826	17,065,423	78.9
Total			**328,273,388**	**318,751,943**	**289,656,761**	**N/A**
Average			**20,517,086.7**	**19,921,996.4**	**18,103,547.6**	**90.9**

Table 4.3 Transcript comparison after gene expression analysis.

Days post inoculation	Number of transcripts with expression[a]	Number of differentially expressed transcripts (q<0.05)[b]	upregulated (q<0.05)[b]	downregulated (q<0.05)[b]
2	38,768	0	0	0
4	38,302	103	48	55
8	38,407	125	79	46
18	38,616	1043	1008	35

[a] Transcripts where the minimum number of read alignments (n=10) allowed significance testing between the two conditions.
[b] After correction for multiple testing (Benjamini-Hockberg).

Validation of RNA-seq results was performed by qRT-PCR of 15 gene primers (**Table 4.1**) with different expression patterns over the time course. Log_2-fold changes showed a correlation of 0.83 between the RNA-seq and the qRT-PCR results for all time points and treatment comparisons (**Table 4.4** and **Figure 4.8**).

Table 4.4 Log2-fold change (water vs. inoculum) agreement between RNAseq and qRT-PCR.

Description		Day 2		Day 4		Day 8		Day 18	
Flax gene ID	**Gene**	**RNAseq**	**qRT-PCR**	**RNAseq**	**qRT-PCR**	**RNAseq**	**qRT-PCR**	**RNAseq**	**qRT-PCR**
Lus10015351	nitrate transporter	0.0	0.1	-0.9	-0.6	-1.6	-2.4	-3.5	-2.7
Lus10021936	basic 7s globulin-like	-1.0	0.4	-1.8	-1.4	-2.1	-3.0	-1.2	-1.4
Lus10016424	conserved protein similar to AER92600	-2.4	-1.8	-2.9	-2.5	-2.1	-1.9	-0.5	-1.0
Lus10005393	polyphenol oxidase	1.9	2.9	-0.7	0.1	-1.8	-1.9	1.0	1.6
Lus10039487	auxin-responsive protein iaa7	-0.1	0.0	0.3	0.2	-0.3	-0.9	-1.3	-2.2
Lus10041830	chitinase	-0.1	0.4	0.3	0.7	0.3	0.8	3.2	2.0
Lus10004808	leucoanthocyanidin dioxygenase	0.1	0.9	-0.4	0.1	2.1	2.1	6.7	2.1
Lus10020826	peroxidase	0.4	0.8	-0.4	0.2	0.1	-0.4	2.6	1.0
Lus10019060	glycosyl hydrolase family protein with chitinase insertion domain	1.4	2.5	0.1	0.3	-0.5	0.2	4.3	1.5
Lus10008930	major latex protein	0.3	0.7	-0.3	-0.7	0.5	-0.2	1.7	1.6
Lus10027702	eukaryotic translation initiation factor 3 subunit c	0.0	0.0	0.0	0.0	0.0	-0.2	-0.2	0.0
Lus10025438	transcription elongation factor 1	-0.1	0.0	0.1	0.0	0.0	-0.1	0.1	0.2
Lus10015458	dna-directed rna polymerase i subunit rpa12	0.0	0.1	0.0	-0.1	0.0	-0.1	0.0	0.1
Lus10005425	trehalose-6-phosphate synthase	-0.1	-0.2	0.0	0.0	0.0	0.0	0.1	0.2
Lus10038622	eukaryotic translation initiation factor 3 subunit l-like	0.1	0.1	0.0	0.1	0.0	-0.1	-0.3	-0.1

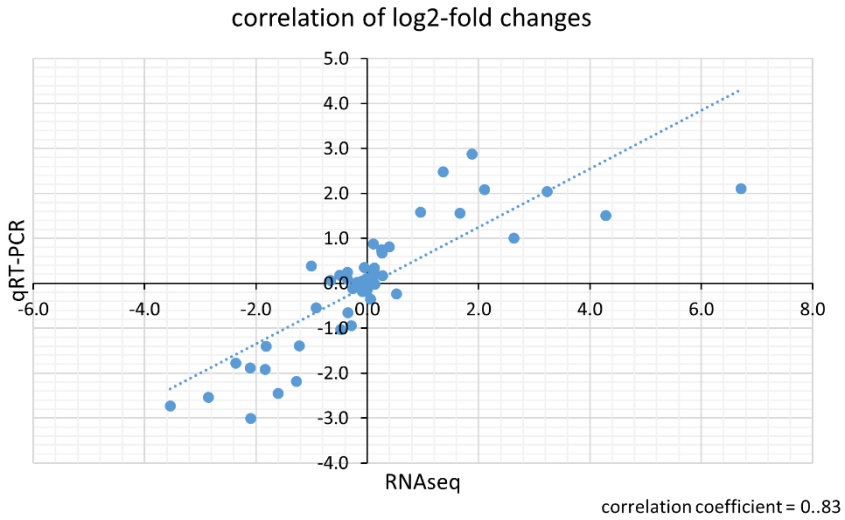


Figure 4.8 Correlation of all genes and time points of Table 4.4.

4.4.4 Functional categorization of differentially expressed transcripts.

We used complementary approaches to categorize the differentially expressed transcripts that we had identified by RNA-seq: Gene Ontology (GO) enrichment analysis using AgriGO [246]; metabolic pathway enrichment analysis using plantGSEA [339]; and a heatmap time course using MeV4.9 [340]. For Gene Ontology we defined a numerical level of hierarchy based on the acyclic graphs created by AgriGO, with more general terms having a lower number (e.g. biological process = 1) and more specific terms having higher numbers (**Table 4.5**).

Table 4.5 Biological process GO categories enriched from significantly different genes.

DPI	GO accession	Term[a]	*p*-value	FDR
4	GO:0045087	innate immune response (5)	7.40E-07	5.30E-05
	GO:0006955	immune response (3)	1.20E-06	5.30E-05
	GO:0002376	immune system process (2)	1.20E-06	5.30E-05
	GO:0006952	defense response (4)	2.60E-06	8.90E-05
	GO:0012501	programmed cell death (4)	1.60E-05	0.00043
	GO:0008219	cell death (3)	6.00E-05	0.0012
	GO:0016265	death (2)	6.00E-05	0.0012
	GO:0051707	response to other organism (4)	0.00018	0.0031
	GO:0009607	response to biotic stimulus (3)	0.00026	0.0039
	GO:0051704	multi-organism process (3)	0.00073	0.0098

DPI	GO accession	Term[a]	*p*-value	FDR
	GO:0006950	response to stress (3)	0.0017	0.021
	GO:0050896	response to stimulus (2)	0.0022	0.025
	GO:0009620	response to fungus (5)	8.90E-06	0.0021
	GO:0009607	response to biotic stimulus (3)	4.40E-05	0.0052
	GO:0019748	secondary metabolic process (3)	8.30E-05	0.0066
	GO:0051707	response to other organism (4)	0.00014	0.0084
	GO:0006955	immune response (3)	0.00023	0.0089
	GO:0002376	immune system process (2)	0.00023	0.0089
	GO:0006519	cellular amino acid and derivative metabolic process (4)	0.0003	0.01
	GO:0042398	cellular amino acid derivative biosynthetic process (6)	0.00074	0.016
	GO:0019752	carboxylic acid metabolic process (6)	0.00084	0.016
	GO:0006952	defense response (4)	0.00053	0.016
8	GO:0043436	oxoacid metabolic process (5)	0.00084	0.016
	GO:0051704	multi-organism process (2)	0.00076	0.016
	GO:0006082	organic acid metabolic process (4)	0.00086	0.016
	GO:0044283	small molecule biosynthetic process (4)	0.00095	0.016
	GO:0042180	cellular ketone metabolic process (4)	0.0011	0.017
	GO:0045087	innate immune response (5)	0.0012	0.018
	GO:0006629	lipid metabolic process (4)	0.0018	0.025
	GO:0050896	response to stimulus (2)	0.0022	0.029
	GO:0006520	cellular amino acid metabolic process (7)	0.0024	0.03
	GO:0006575	cellular amino acid derivative metabolic process (5)	0.0026	0.03
	GO:0044106	cellular amine metabolic process (5)	0.003	0.034
	GO:0042398	cellular amino acid derivative biosynthetic process (6)	2.20E-11	2.90E-08
	GO:0050896	response to stimulus (2)	6.80E-11	3.10E-08
	GO:0019748	secondary metabolic process (3)	4.90E-11	3.10E-08
	GO:0009699	phenylpropanoid biosynthetic process (7)	3.20E-10	1.10E-07
	GO:0009607	response to biotic stimulus (3)	3.90E-10	1.10E-07
18	GO:0006575	cellular amino acid derivative metabolic process (5)	5.20E-10	1.20E-07
	GO:0051707	response to other organism (4)	6.90E-10	1.40E-07
	GO:0044283	small molecule biosynthetic process (5)	9.40E-10	1.40E-07
	GO:0009611	response to wounding (4)	9.30E-10	1.40E-07
	GO:0006950	response to stress (3)	3.10E-09	4.30E-07
	GO:0009698	phenylpropanoid metabolic process (6)	1.20E-08	1.50E-06

DPI	GO accession	Term[a]	*p*-value	FDR
	GO:0051704	multi-organism process (2)	2.30E-08	2.70E-06
	GO:0006952	defense response (4)	2.60E-08	2.70E-06
	GO:0009605	response to external stimulus (3)	4.40E-08	4.30E-06
	GO:0019438	aromatic compound biosynthetic process (6)	5.50E-08	5.00E-06
	GO:0042221	response to chemical stimulus (3)	6.40E-08	5.50E-06
	GO:0009813	flavonoid biosynthetic process (8)	1.70E-07	1.30E-05
	GO:0009620	response to fungus (5)	1.60E-07	1.30E-05
	GO:0006519	cellular amino acid derivative metabolic process (5)	3.20E-07	2.30E-05
	GO:0009812	flavonoid metabolic process (7)	7.80E-07	5.40E-05
	GO:0006725	cellular aromatic compound metabolic process (5)	2.20E-06	0.00014
	GO:0009753	response to jasmonic acid stimulus (5)	3.00E-06	0.00018
	GO:0006979	response to oxidative stress (4)	3.00E-05	0.0018
	GO:0044281	small molecule metabolic process (3)	4.00E-05	0.0023
	GO:0009695	jasmonic acid biosynthetic process (11)	6.50E-05	0.0036
	GO:0031408	oxylipin biosynthetic process (10)	9.40E-05	0.0049
	GO:0032787	monocarboxylic acid metabolic process (7)	0.00013	0.0064
	GO:0009694	jasmonic acid metabolic process (10)	0.00013	0.0065
	GO:0031407	oxylipin metabolic process (9)	0.00018	0.0086
	GO:0009404	toxin metabolic process (4)	0.00026	0.011
	GO:0009407	toxin catabolic process (5)	0.00026	0.011
	GO:0009850	auxin metabolic process (6)	0.00037	0.016
	GO:0006629	lipid metabolic process (4)	0.00043	0.018
	GO:0006576	cellular biogenic amine metabolic process (6)	0.00068	0.028
	GO:0010260	organ senescence (6)	0.00071	0.028
	GO:0010033	response to organic substance (4)	0.00085	0.032

4.4.4.1 Day 2

No transcripts differed significantly in abundance between control and treated plants at day 2 (**Table 4.3**), therefore, no enriched functional categories were identified.

4.4.4.2 Day 4

At day 4 post-inoculation, 12 GO terms were significantly enriched (**Table 4.5**); the highest level categories (indicating the more general processes) were immune system process, death, and response to stimulus. The more specific categories pointed towards defense and

interaction responses with other organisms. Inspection of the transcripts corresponding to these specific categories included disease resistance proteins and pathogenesis-related (PR) thaumatin proteins (dataverse file: transcript differential expression in flax upon *Foln* infection http://dx.doi.org/10.7939/DVN/10933). From the plantGSEA analysis only the terpenoid backbone biosynthesis pathway and the metabolism of xenobiotics by cytochrome P450 showed significant enrichment (**Table 4.6**). Uncategorized transcripts represented by multiple hits included GDSL-like lipase acylhydrolase proteins, laccases, bifunctional inhibitor lipid-transfer proteins (LTPs), and major latex-like protein (MLP) 423 (dataverse file: transcript differential expression in flax upon *Foln* infection http://dx.doi.org/10.7939/DVN/10933).

Table 4.6 Enrichment analysis using plant GSEA.

DPI	pathway ID	Description	*p*-value	FDR
4	KEGG:ATH00980	Metabolism of xenobiotics by cytochrome P450	2.30E-04	0.0186
	KEGG:ATH00900	Terpenoid backbone biosynthesis	1.03E-03	0.0417
8	KEGG:ATH00945	Stilbenoid, diarylheptanoid and gingerol biosynthesis	8.73E-05	4.75E-03
	KEGG:ATH00903	Limonene and pinene degradation	1.31E-04	4.75E-03
	KEGG:ATH00260	Glycine, serine and threonine metabolism	1.57E-04	4.75E-03
	KEGG:ATH00680	Methane metabolism	5.28E-04	9.56E-03
	KEGG:ATH00360	Phenylalanine metabolism	4.69E-04	9.56E-03
	KEGG:ATH00908	Zeatin biosynthesis	7.15E-04	0.0108
	KEGG:ATH00940	Phenylpropanoid biosynthesis	8.50E-04	0.011
	KEGG:ATH01100	Metabolic pathways	1.12E-03	0.0126
	PlantCyc:2.4.1.91	quercetin glucoside biosynthesis (*Arabidopsis*)	2.02E-04	0.0134
	PlantCyc:1.1.1.219	leucopelargonidin and leucocyanidin biosynthesis	4.12E-04	0.0134
	PlantCyc:6.2.1.12	flavonoid biosynthesis	4.12E-04	0.0134
	PlantCyc:1.14.11.9	leucodelphinidin biosynthesis	4.12E-04	0.0134
18	KEGG:ATH01100	Metabolic pathways	4.47E-18	1.49E-15
	KEGG:ATH00360	Phenylalanine metabolism	1.93E-15	3.20E-13
	KEGG:ATH00940	Phenylpropanoid biosynthesis	4.08E-14	4.52E-12
	KEGG:ATH00680	Methane metabolism	5.37E-13	4.46E-11

KEGG:ATH01061	Biosynthesis of phenylpropanoids	1.46E-11	9.68E-10
KEGG:ATH00592	alpha-Linolenic acid metabolism	1.18E-08	6.56E-07
KEGG:ATH01070	Biosynthesis of plant hormones	2.36E-08	1.12E-06
KEGG:ATH00270	Cysteine and methionine metabolism	1.27E-07	5.27E-06
KEGG:ATH00350	Tyrosine metabolism	6.48E-07	2.39E-05
KEGG:ATH00945	Stilbenoid, diarylheptanoid and gingerol biosynthesis	5.40E-06	1.79E-04
KEGG:ATH00966	Glucosinolate biosynthesis	9.07E-06	2.74E-04
KEGG:ATH00980	Metabolism of xenobiotics by cytochrome P450	1.34E-05	3.70E-04
KEGG:ATH00941	Flavonoid biosynthesis	3.61E-05	9.23E-04
KEGG:ATH00950	Isoquinoline alkaloid biosynthesis	7.06E-05	1.68E-03
KEGG:ATH00903	Limonene and pinene degradation	1.11E-04	2.45E-03
PlantCyc:6.2.1.12	flavonoid biosynthesis	2.34E-05	3.92E-03
PlantCyc:5.3.99.6	jasmonic acid biosynthesis	1.95E-05	3.92E-03
PlantCyc:1.1.1.219	leucopelargonidin and leucocyanidin biosynthesis	2.34E-05	3.92E-03
PlantCyc:1.14.11.9	leucodelphinidin biosynthesis	2.34E-05	3.92E-03
KEGG:ATH00052	Galactose metabolism	4.81E-04	1.00E-02
KEGG:ATH00960	Tropane, piperidine and pyridine alkaloid biosynthesis	5.64E-04	0.011
PlantCyc:1.13.11.12	13-LOX and 13-HPL pathway	9.42E-05	0.0126
KEGG:ATH00010	Glycolysis / Gluconeogenesis	7.13E-04	0.0132
KEGG:ATH01064	Biosynthesis of alkaloids derived from ornithine, lysine and nicotinic acid	1.16E-03	0.0203
KEGG:ATH00330	Arginine and proline metabolism	2.84E-03	0.0472
KEGG:ATH00500	Starch and sucrose metabolism	3.21E-03	0.0485
KEGG:ATH01062	Biosynthesis of terpenoids and steroids	3.13E-03	0.0485

4.4.4.3 Day 8

On day 8 post-inoculation, the number of significantly enriched GO categories increased to 21 (**Table 4.5**). Higher-level categories included multi-organism processes, as well as some categories seen on day 4: immune system process, response to stimulus. Among immune system process transcripts there were PR thaumatin proteins, chitinases, and disease resistance proteins. The category with the highest number of hits (response to stimulus) included all genes from

immune system process, plus genes such as peroxidases and WRKY transcription factors. More specific categories indicated for the first time in this time series a direct interaction with another organism (e.g. response to fungus). Furthermore, GO categories associated with primary and secondary metabolism became enriched; these included metabolism of amino acid derivatives, organic acids, and lipids. The category of secondary metabolic process included cytochrome-related polypeptides and glutathione s-transferase (GST) family proteins. The category of lipid metabolic process contained mainly GDSL-like lipase acylhydrolases.

The plantGSEA categorization provided additional information about the metabolic pathways enriched at day 8, particularly pathways for the synthesis of phenylpropanoids and flavonoids (**Table 4.6**). Specific genes involved in these processes (dataverse file: transcript differential expression in flax upon *Foln* infection http://dx.doi.org/10.7939/DVN/10933) included: peroxidases, terpenoid synthases/cyclases, and 2-oxoglutarate and Fe-dependent oxygenase superfamily proteins.

Several uncategorized transcripts or gene transcripts with a common annotation but not placed in a specific category also showed distinct patterns of accumulation 8 DPI. These included 2-oxoglutarate and Fe-dependent oxygenase superfamily proteins, UDP-glycosyltransferases (UGTs), lacasses, LTPs and MLPs; and genes related to primary carbohydrate metabolism including some family 32 glycosyl hydrolases, sugar transporters, and several cell wall modifying enzymes (e.g. xyloglucan endotransglucosylase, pectinesterase, beta-d-xylosidase 1) (dataverse file: transcript differential expression in flax upon *Foln* infection http://dx.doi.org/10.7939/DVN/10933).

4.4.4.4 Day 18

The greatest number of enriched categories was found 18 DPI (**Table 4.5**). The majority of the genes belonging to enriched terms had increased transcript abundance in the inoculated plants as compared to controls. The more general enriched GO terms included those related to metabolism, response to stimulus and multi-organism processes (**Table 4.5**). More-specific categories indicated the activation of processes related to: organ senescence; metabolism of auxin, jasmonic acid, aromatics, flavonoids and toxins; and responses to wounding, oxidative stress, jasmonic acid and fungus. With reference to metabolism, transcripts annotated as 2-oxoglutarate and Fe-dependent oxygenases, NAD-binding rossmann-fold proteins and

cytochrome-related proteins were classified in the amino acid and aromatic-related processes, with the former genes also related to metabolism of flavonoids, phenylpropanoids, and terprenoids. Two other enriched pathways, lipid and monocarboxylic acid metabolic processes, are a source of fatty acids that can result in downstream synthesis of oxylipin and jasmonate derivatives. The GDSL-like lipase acylhydrolases and alpha beta hydrolases were abundant in the category of lipid metabolic process.

The categories of response to stimulus and multi-organism process were comprised of common genes of plant defense responses. Receptors of pathogen signals included leucine-rich repeat (LRR) protein kinases and receptor-like proteins (RLPs), while proteins that have a direct effect on the pathogens comprised chitinases and thaumatins. Inhibitors of pathogen disruptive enzymes were represented by diverse protease inhibitors (PIs). Genes related to the oxidative burst/lignification included peroxidases and laccases, and potential controllers of oxidative stress comprised GSTs, which belonged to both the response to stimulus and secondary metabolic process categories. Multidrug transporters that bind cytotoxic compounds for cell removal were part of the response to stimulus category and included ATP-binding cassette (ABC) transporters, transcripts classified as multidrug and toxic compound extrusion (MATE) efflux family proteins, and major facilitator superfamily (MFS) membrane proteins. Transcripts with similarity to aquaporins (major intrinsic proteins), and amino acid transporters were also activated.

Transcription factors (TFs) were found in multiple GO functional categories, and were one of the most numerous and diverse classes of genes activated 18 DPI (Dataverse file: transcript differential expression in flax upon *Foln* infection http://dx.doi.org/10.7939/DVN/10933), and included: basic helix-loop-helix (bHLH) DNA binding proteins, C2H2-type zinc finger proteins, WRKY DNA-binding proteins, MYB domain proteins, NAC domain transcriptional regulators and winged-helix-DNA-binding transcription factors.

Protein modification and degradation genes spread among multiple categories and were represented by increased abundance transcripts of cysteine and aspartyl proteases, ubiquitin-related proteins, and numerous protein kinases: lectin protein kinases, calcineurin B-like (CBL)-interacting protein kinases (CIPK), LRR protein kinases, and mitogen-activated (MAP) kinases.

The plantGSEA categorization provided further information about the metabolic processes that were enriched at 18 DPI, the most prominent of which were biosynthesis of

phenylpropanoids and plant hormones (**Table 4.6**). Other well represented categories included biosynthesis of flavonoids, glucosinolates, terpenoids and steroids, stilbenoid, diarylheptanoid and gingerol, tropane, piperidine and pyridine alkaloids, and isoquinoline alkaloids. Biosynthetic pathways for the amino acid precursors of many of these compounds were also enriched, including phenylalanine, tyrosine, cysteine, and methionine, arginine, and proline. Pathways for the synthesis of 13-LOX and 13-HPL, as well as alpha-linolenic acid were enriched as was the downstream jasmonic acid biosynthesis pathway. Finally, carbohydrate related pathways were also enriched, including glycolysis/gluconeogenesis, and metabolism of galactose, starch and sucrose metabolism. Carbohydrate metabolism and cell wall enzymes were highlighted by the presence of other glycosyl hydrolases (e.g beta glucosidases), expansins, pectin lyases (polygalactunorases), pectin methylesterase inhibitors (PMEIs) and xyloglucan endotransglucosylases.

Among enzymes that were not categorized by our analysis, UDP-glycosyltrasnferases were the most abundant with 19 transcripts, and calcium-binding or dependent proteins totalled 14 hits (Dataverse file: transcript differential expression in flax upon *Foln* infection http://dx.doi.org/10.7939/DVN/10933). Thirty-one transcripts with increased abundance were annotated as 2-oxoglutarate and Fe-dependent oxygenases, although most of these genes function in diverse pathways. Cytochrome-related proteins, NAD-binding Rossmann-fold proteins and peroxidases were among the transcripts with more hits 18 DPI (Dataverse file: transcript differential expression in flax upon *Foln* infection http://dx.doi.org/10.7939/DVN/10933).

4.4.5 Time course gene expression

Our enrichment analyses over the time course of infection evidenced relevant genes of the plant-pathogen interaction. These genes were retrieved from our list of all differentially expressed genes and used to build heatmaps of expression levels.

4.4.5.1 Pathogen elicitor perception and signalling

Thirty-three genes including TIR-NBS-LRR, receptor-like kinases (RLKs), receptor-like proteins (RLPs) and lectin protein kinases (LecRK), are part of the plant's pathogen signals perception and transduction. Most of these genes were activated 18 DPI, but seven were mainly downregulated throughout the time course (**Figure 4.9A**). Four of these seven are TIR-NBS-

LRR transcripts, and two of them correspond to the disease resistance proteins appearing at days 4 and 8 in the GO analyses (XLOC_041933 and XLOC_008811).

Downstream of signal reception, calcium acts as a secondary messenger and multiple modifying enzymes act to complete signal transduction. Calcium related genes included binding ef-hand family proteins, calcium transporters and calcium-dependent phosphodiesterases; in the meantime, 12 transcripts related to protein modification and signalling comprised MAPKs, CIPKs and phosphatases. All of these proteins showed increased transcript abundance 18 DPI (**Figure 4.9B-C**).

4.4.5.2 Transcription factors (TFs)

Seventy-six TFs belonging to 16 different TF types were regulated during pathogen infection (**Figure 4.9D**). Hierarchical clustering of the expression patterns did not group the TFs by type and most TFs had diverse patterns of regulation during the three initial sampling days, and increased transcript abundance mainly 18 DPI. Only five genes seemed consistently downregulated throughout the sampled days, with three of them corresponding to WRKY TFs: XLOC_030137, XLOC_027114 and XLOC_024419. These three transcripts correspond to the same annotation: *WRKY70* (Dataverse file: transcript differential expression in flax upon *Foln* infection http://dx.doi.org/10.7939/DVN/10933).

4.4.5.3 Hormones

Hormones only appeared as a GO enriched term on day 18. Genes for jasmonate, ethylene and auxin metabolism were abundant and presented increased transcript levels 18 DPI. Transcripts directly involved in synthesis of jasmonate included: allene oxide cyclase (*AOC*), allene oxide synthase (*AOS*), 12-oxophytodienoate reductase (*OPR*) and lipoxygenase (*LOX*) family proteins, but transcripts involved in the regulatory response to jasmonic acid were also present (jasmonate-zim-domain proteins - JAZ) (**Figure 4.9E**). Nine transcripts characterized as 2-oxoglutarate and fe-dependent oxygenases (Dataverse file: transcript differential expression in flax upon *Foln* infection http://dx.doi.org/10.7939/DVN/10933) were similar to 1-aminocyclopropane-1-carboxylate oxidase (ACO, a key enzyme in ethylene biosynthesis) (**Figure 4.9F**). Moreover, transcripts annotated as 1-aminocyclopropane-1-carboxylate synthases (ACS - also key in ethylene synthesis), and ethylene responsive factors (ERFs) were also present.

Auxin-related transcripts included indoleacetic acid (IAA) amido/amino hydrolases, auxin-induced transporters and auxin binding proteins (**Figure 4.9G**).

4.4.5.4 PR-proteins

PR-proteins included genes involved in deterring or avoiding pathogen attack (chitinases, thaumatins and protease inhibitors), as well as genes with additional functions like peroxidases (oxidative burst/lignification) and LTPs (non-specific lipid transfer).

Six chitinases from classes I, IV and V showed increased transcript abundance 18 DPI, while two class IV chitinases were mainly upregulated until day 8 (**Figure 4.9H**). Thaumatins presented a more homogeneous transcription pattern from beginning to end, with only one being repressed constantly (**Figure 4.9I**). In the meantime, 21 kunitz and serine PIs showed high transcript abundances 18 DPI (**Figure 4.9J**). Whereas most of the 29 peroxidases present were activated at 18 DPI, four clearly showed decreased transcript abundance 8 and 18 DPI (**Figure 4.9K**). Finally, LTPs showed two groups, one with members that are activated early, and are repressed either after day 4 or 8 (four transcripts showed high repression at 8 DPI), and a second group with LTPs with higher increased transcript abundance 18 DPI (**Figure 4.9L**).

4.4.5.5 Oxidative burst

Central to the oxidative burst that occurs upon pathogen attack are peroxidases, laccases and the glutathione-ascorbate complex in charge of regulating ROS. The main enzyme implicated in ROS generation (respiratory burst oxidase - RBO) had increased abundance 18 DPI and was represented by two transcripts: XLOC_008027 and XLOC_024171 (Dataverse file: transcript differential expression in flax upon *Foln* infection http://dx.doi.org/10.7939/DVN/10933). Peroxidases (discussed above) can both act in hydrogen peroxide production or in lignification, while laccases are involved in the latter process. From seven laccases, two were activated through 8 DPI while the rest had their activation peak 18 DPI (**Figure 4.9M**). Finally, the glutathione-related transcripts were dominated by the hydrogen peroxide scavenging proteins: GSTs (**Figure 4.9N**), with increased transcript abundance 8 and 18 DPI.

4.4.5.6 Secondary metabolism

The most visible changes through the infection cycle according to our GO analyses were evidenced in genes related to secondary metabolism. The phenylpropanoid genes encountered were mostly activated 18 DPI and comprised transcripts that are key in lignin formation: cinnamic acid 4-hydroxylase (*C4H*), cynnamoyl-CoA reductase (*4CL*), cinnamyl alcohol dehydrogenase (*CAD*) and shikimate quinate hydroxycynnamoyltransferase (*HCT*) (**Figure 4.9O**). Meanwhile, flavonoid-related genes were represented by anthocyanin biosynthetic genes but also by genes of the flavone and flavonoid synthesis pathways (**Figure 4.9P**). The flavonoid pathway transcripts were also mainly activated 18 DPI but some members demonstrated increased abundance at 8 DPI, while several anthocyanidin synthases (XLOC_005793, XLOC_001152, XLOC_002414) seemed to be on all the time. Finally, isoprenoid/carotenoid genes were also found with increased abundance 18 DPI. The main enzyme controlling carotenoid synthesis, phytoene synthase (PSY), had one of the highest increases from this group (**Figure 4.9Q**).

4.4.5.7 Transport

Three groups of transport proteins had high transcript abundances 18 DPI: multidrug transporters which comprise ABC transporters, MATEs and MFS' (**Figure 4.9R**), major intrinsic proteins (MIPs) also known as aquaporins (**Figure 4.9S**), and amino acid transporters (**Figure 4.9T**). Ten MIPs divided in six plasma membrane intrinsic proteins (PIPs) and four tonoplast intrinsic proteins (TIPs) were mainly repressed during days 4 and 8, but increased their transcripts 18 DPI. With one exception, all amino acid transporters increased their transcript abundance 18 DPI, while only three out of 37 multidrug transporters had decreased transcript abundance at the end of the sampled cycle.

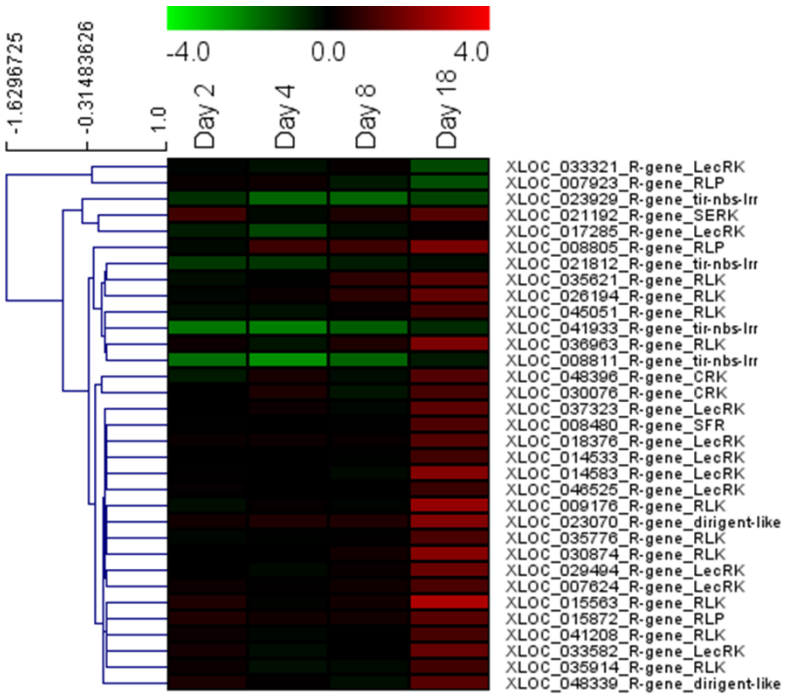
4.4.5.8 Cell wall

Extensive cell wall modification at 18 DPI was indicated by a diversity of genes that included expansins, endotransglycosylases and polygalactunorases (**Figure 4.9U**). However, a subset of the cell-wall related enzymes also indicated enzyme inhibition (e.g. PMEIs).

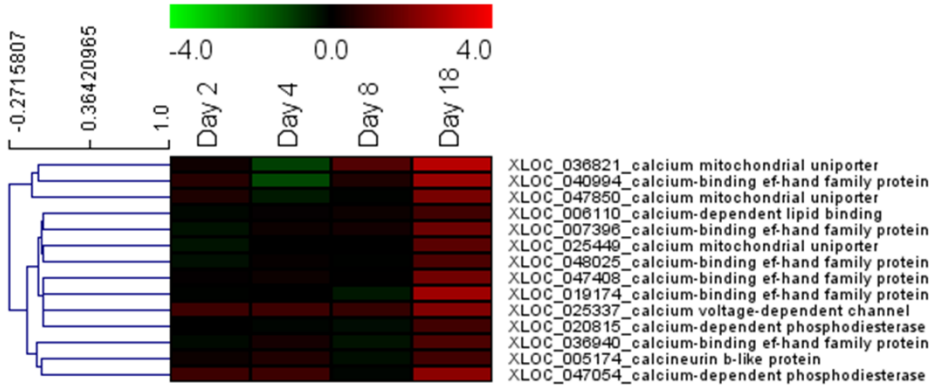
4.4.5.9 Major latex proteins

Major latex proteins comprise a group of genes initially isolated from opium [341], with a function which has not been completely elucidated. An interesting pattern emerged from MLPs, which were mostly repressed over the full time course, contrasting the behavior of most genes from this study (**Figure 4.9V**).

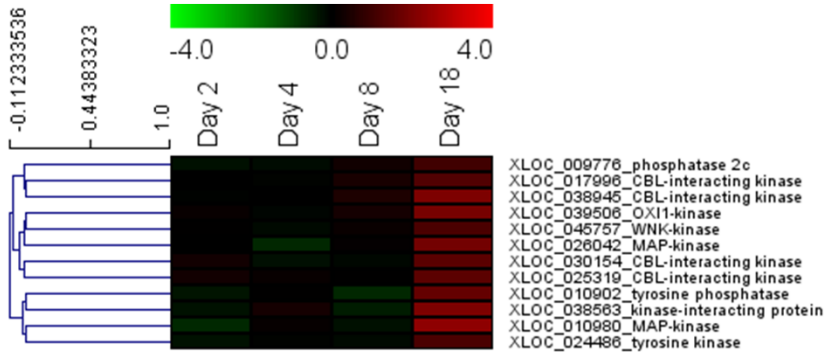
A


-4.0
0.0
4.0
-1.6296725
-0.31483626
1.0
Day 2
Day 4
Day 8
Day 18
XLOC_033321_R-gene_LecRK
XLOC_007923_R-gene_RLP
XLOC_023929_R-gene_tir-nbs-lrr
XLOC_021192_R-gene_SERK
XLOC_017285_R-gene_LecRK
XLOC_008805_R-gene_RLP
XLOC_021812_R-gene_tir-nbs-lrr
XLOC_035621_R-gene_RLK
XLOC_026194_R-gene_RLK
XLOC_045051_R-gene_RLK
XLOC_041933_R-gene_tir-nbs-lrr
XLOC_036963_R-gene_RLK
XLOC_008811_R-gene_tir-nbs-lrr
XLOC_048396_R-gene_CRK
XLOC_030076_R-gene_CRK
XLOC_037323_R-gene_LecRK
XLOC_008480_R-gene_SFR
XLOC_018376_R-gene_LecRK
XLOC_014533_R-gene_LecRK
XLOC_014583_R-gene_LecRK
XLOC_046525_R-gene_LecRK
XLOC_009176_R-gene_RLK
XLOC_023070_R-gene_dirigent-like
XLOC_035776_R-gene_RLK
XLOC_030874_R-gene_RLK
XLOC_029494_R-gene_LecRK
XLOC_007624_R-gene_LecRK
XLOC_015563_R-gene_RLK
XLOC_015872_R-gene_RLP
XLOC_041208_R-gene_RLK
XLOC_033582_R-gene_LecRK
XLOC_035914_R-gene_RLK
XLOC_048339_R-gene_dirigent-like

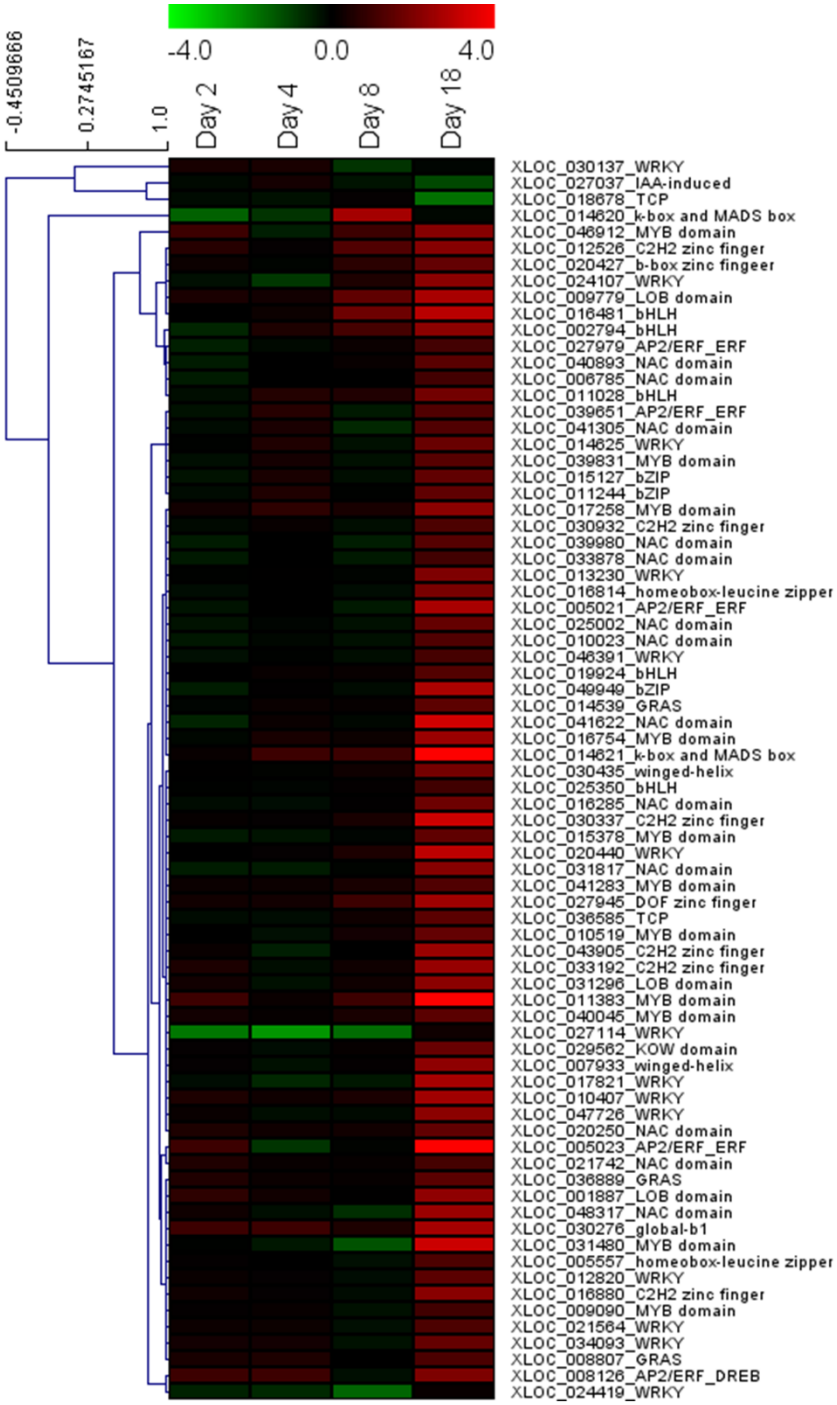
B


-4.0
0.0
4.0
-0.2715807
0.36420965
1.0
Day 2
Day 4
Day 8
Day 18
XLOC_036821_calcium mitochondrial uniporter
XLOC_040994_calcium-binding ef-hand family protein
XLOC_047850_calcium mitochondrial uniporter
XLOC_006110_calcium-dependent lipid binding
XLOC_007396_calcium-binding ef-hand family protein
XLOC_025449_calcium mitochondrial uniporter
XLOC_048025_calcium-binding ef-hand family protein
XLOC_047408_calcium-binding ef-hand family protein
XLOC_019174_calcium-binding ef-hand family protein
XLOC_025337_calcium voltage-dependent channel
XLOC_020815_calcium-dependent phosphodiesterase
XLOC_036940_calcium-binding ef-hand family protein
XLOC_005174_calcineurin b-like protein
XLOC_047054_calcium-dependent phosphodiesterase

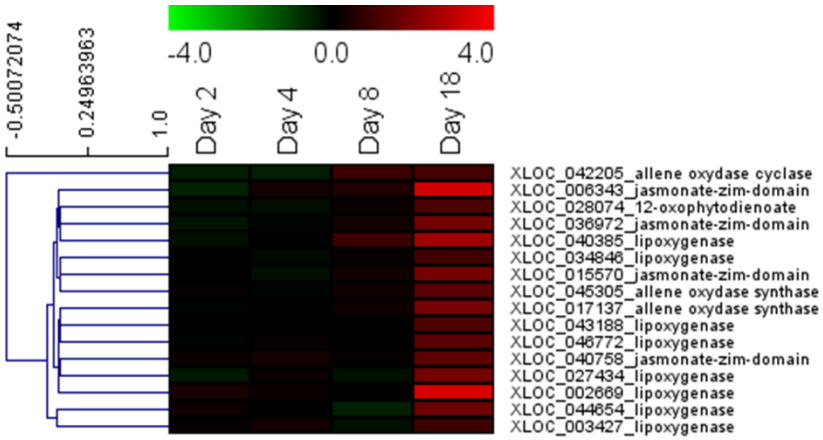
C


-4.0
0.0
4.0
-0.112333536
0.44383323
1.0
Day 2
Day 4
Day 8
Day 18
XLOC_009776_phosphatase 2c
XLOC_017996_CBL-interacting kinase
XLOC_038945_CBL-interacting kinase
XLOC_039506_OXI1-kinase
XLOC_045757_WNK-kinase
XLOC_026042_MAP-kinase
XLOC_030154_CBL-interacting kinase
XLOC_025319_CBL-interacting kinase
XLOC_010902_tyrosine phosphatase
XLOC_038563_kinase-interacting protein
XLOC_010980_MAP-kinase
XLOC_024486_tyrosine kinase

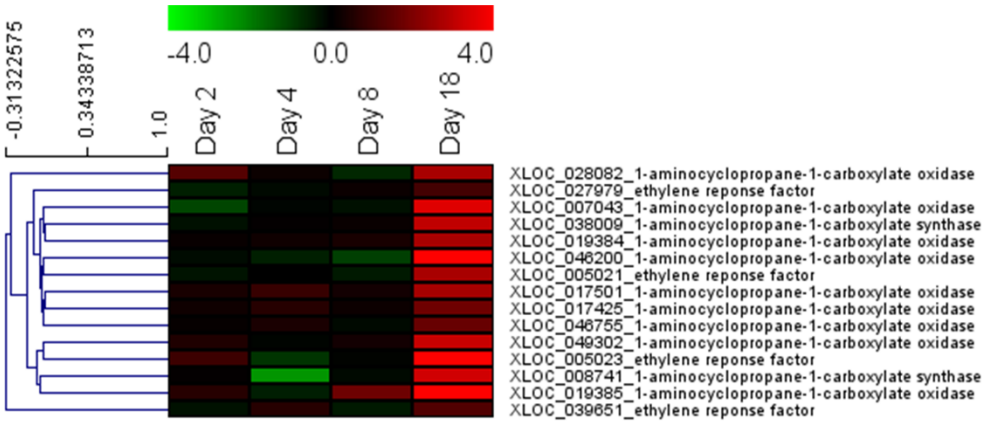
D


-4.0
0.0
4.0
-0.4509666
0.2745167
1.0
Day 2
Day 4
Day 8
Day 18
XLOC_030137_WRKY
XLOC_027037_IAA-induced
XLOC_018678_TCP
XLOC_014620_k-box and MADS box
XLOC_046912_MYB domain
XLOC_012526_C2H2 zinc finger
XLOC_020427_b-box zinc fingeer
XLOC_024107_WRKY
XLOC_009779_LOB domain
XLOC_016481_bHLH
XLOC_002794_bHLH
XLOC_027979_AP2/ERF_ERF
XLOC_040893_NAC domain
XLOC_006785_NAC domain
XLOC_011028_bHLH
XLOC_039651_AP2/ERF_ERF
XLOC_041305_NAC domain
XLOC_014625_WRKY
XLOC_039831_MYB domain
XLOC_015127_bZIP
XLOC_011244_bZIP
XLOC_017258_MYB domain
XLOC_030932_C2H2 zinc finger
XLOC_039980_NAC domain
XLOC_033878_NAC domain
XLOC_013230_WRKY
XLOC_016814_homeobox-leucine zipper
XLOC_005021_AP2/ERF_ERF
XLOC_025002_NAC domain
XLOC_010023_NAC domain
XLOC_046391_WRKY
XLOC_019924_bHLH
XLOC_049949_bZIP
XLOC_014539_GRAS
XLOC_041622_NAC domain
XLOC_016754_MYB domain
XLOC_014621_k-box and MADS box
XLOC_030435_winged-helix
XLOC_025350_bHLH
XLOC_016285_NAC domain
XLOC_030337_C2H2 zinc finger
XLOC_015378_MYB domain
XLOC_020440_WRKY
XLOC_031817_NAC domain
XLOC_041283_MYB domain
XLOC_027945_DOF zinc finger
XLOC_036585_TCP
XLOC_010519_MYB domain
XLOC_043905_C2H2 zinc finger
XLOC_033192_C2H2 zinc finger
XLOC_031296_LOB domain
XLOC_011383_MYB domain
XLOC_040045_MYB domain
XLOC_027114_WRKY
XLOC_029562_KOW domain
XLOC_007933_winged-helix
XLOC_017821_WRKY
XLOC_010407_WRKY
XLOC_047726_WRKY
XLOC_020250_NAC domain
XLOC_005023_AP2/ERF_ERF
XLOC_021742_NAC domain
XLOC_036889_GRAS
XLOC_001887_LOB domain
XLOC_048317_NAC domain
XLOC_030276_global-b1
XLOC_031480_MYB domain
XLOC_005557_homeobox-leucine zipper
XLOC_012820_WRKY
XLOC_016880_C2H2 zinc finger
XLOC_009090_MYB domain
XLOC_021564_WRKY
XLOC_034093_WRKY
XLOC_008807_GRAS
XLOC_008126_AP2/ERF_DREB
XLOC_024419_WRKY

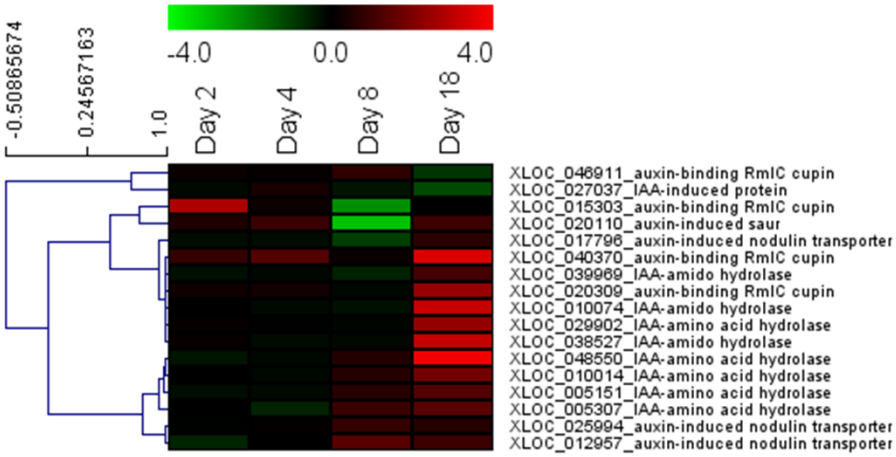
E


-0.50072074
0.24963963
1.0
-4.0
0.0
4.0
Day 2
Day 4
Day 8
Day 18
XLOC_042205_allene oxydase cyclase
XLOC_006343_jasmonate-zim-domain
XLOC_028074_12-oxophytodienoate
XLOC_036972_jasmonate-zim-domain
XLOC_040385_lipoxygenase
XLOC_034846_lipoxygenase
XLOC_015570_jasmonate-zim-domain
XLOC_045305_allene oxydase synthase
XLOC_017137_allene oxydase synthase
XLOC_043188_lipoxygenase
XLOC_046772_lipoxygenase
XLOC_040758_jasmonate-zim-domain
XLOC_027434_lipoxygenase
XLOC_002669_lipoxygenase
XLOC_044654_lipoxygenase
XLOC_003427_lipoxygenase

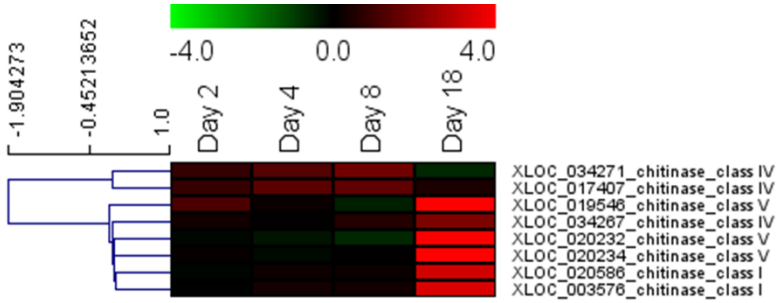
F


-0.31322575
0.34338713
1.0
-4.0
0.0
4.0
Day 2
Day 4
Day 8
Day 18
XLOC_028082_1-aminocyclopropane-1-carboxylate oxidase
XLOC_027979_ethylene reponse factor
XLOC_007043_1-aminocyclopropane-1-carboxylate oxidase
XLOC_038009_1-aminocyclopropane-1-carboxylate synthase
XLOC_019384_1-aminocyclopropane-1-carboxylate oxidase
XLOC_046200_1-aminocyclopropane-1-carboxylate oxidase
XLOC_005021_ethylene reponse factor
XLOC_017501_1-aminocyclopropane-1-carboxylate oxidase
XLOC_017425_1-aminocyclopropane-1-carboxylate oxidase
XLOC_046755_1-aminocyclopropane-1-carboxylate oxidase
XLOC_049302_1-aminocyclopropane-1-carboxylate oxidase
XLOC_005023_ethylene reponse factor
XLOC_008741_1-aminocyclopropane-1-carboxylate synthase
XLOC_019385_1-aminocyclopropane-1-carboxylate oxidase
XLOC_039651_ethylene reponse factor

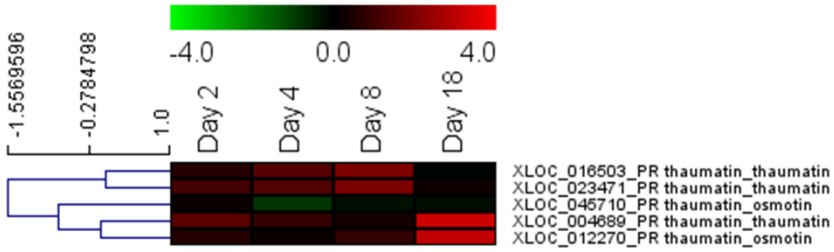
G


-0.50865674
0.24567163
1.0
-4.0
0.0
4.0
Day 2
Day 4
Day 8
Day 18
XLOC_046911_auxin-binding RmlC cupin
XLOC_027037_IAA-induced protein
XLOC_015303_auxin-binding RmlC cupin
XLOC_020110_auxin-induced saur
XLOC_017796_auxin-induced nodulin transporter
XLOC_040370_auxin-binding RmlC cupin
XLOC_039969_IAA-amido hydrolase
XLOC_020309_auxin-binding RmlC cupin
XLOC_010074_IAA-amido hydrolase
XLOC_029902_IAA-amino acid hydrolase
XLOC_038527_IAA-amido hydrolase
XLOC_048550_IAA-amino acid hydrolase
XLOC_010014_IAA-amino acid hydrolase
XLOC_005151_IAA-amino acid hydrolase
XLOC_005307_IAA-amino acid hydrolase
XLOC_025994_auxin-induced nodulin transporter
XLOC_012957_auxin-induced nodulin transporter

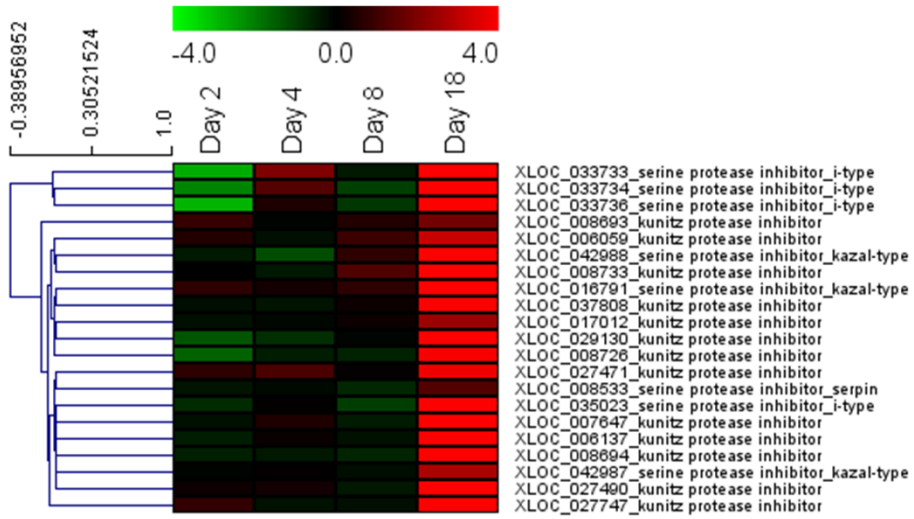
H


-4.0
0.0
4.0
-1.904273
-0.45213652
1.0
Day 2
Day 4
Day 8
Day 18
XLOC_034271_chitinase_class IV
XLOC_017407_chitinase_class IV
XLOC_019546_chitinase_class V
XLOC_034267_chitinase_class IV
XLOC_020232_chitinase_class V
XLOC_020234_chitinase_class V
XLOC_020586_chitinase_class I
XLOC_003576_chitinase_class I

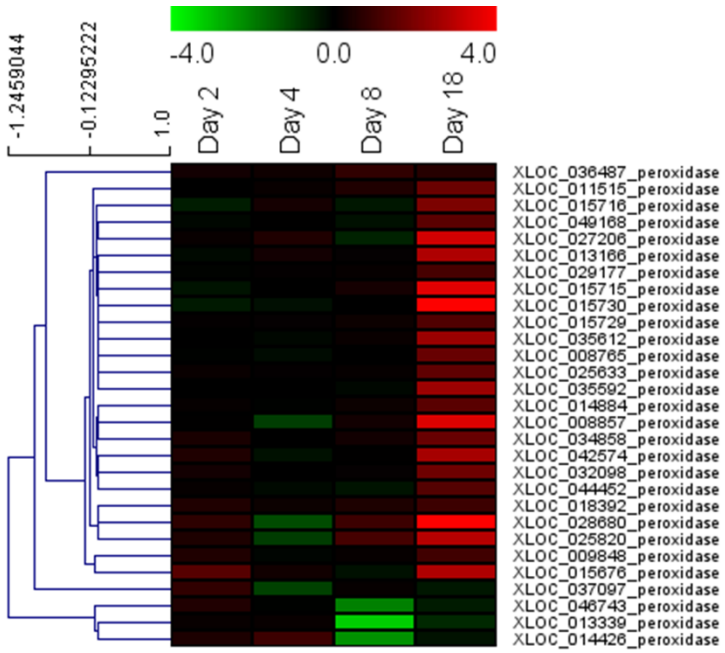
I


-4.0
0.0
4.0
-1.5569596
-0.2784798
1.0
Day 2
Day 4
Day 8
Day 18
XLOC_016503_PR thaumatin_thaumatin
XLOC_023471_PR thaumatin_thaumatin
XLOC_045710_PR thaumatin_osmotin
XLOC_004689_PR thaumatin_thaumatin
XLOC_012270_PR thaumatin_osmotin

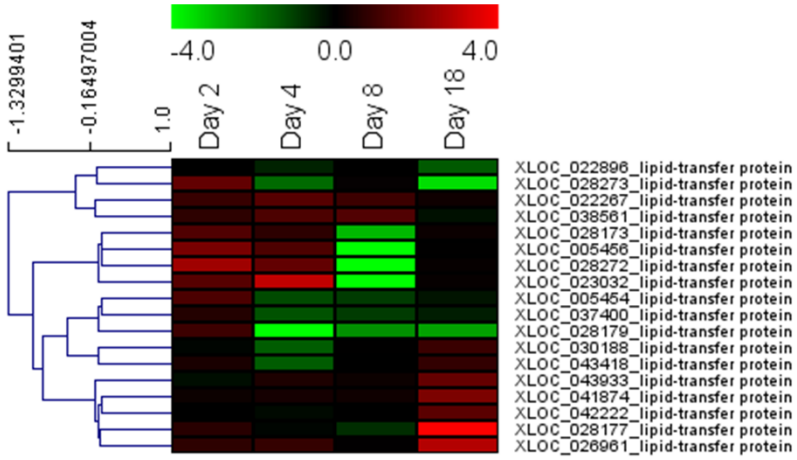
J


-4.0
0.0
4.0
-0.38956952
0.30521524
1.0
Day 2
Day 4
Day 8
Day 18
XLOC_033733_serine protease inhibitor_i-type
XLOC_033734_serine protease inhibitor_i-type
XLOC_033736_serine protease inhibitor_i-type
XLOC_008693_kunitz protease inhibitor
XLOC_006059_kunitz protease inhibitor
XLOC_042988_serine protease inhibitor_kazal-type
XLOC_008733_kunitz protease inhibitor
XLOC_016791_serine protease inhibitor_kazal-type
XLOC_037808_kunitz protease inhibitor
XLOC_017012_kunitz protease inhibitor
XLOC_029130_kunitz protease inhibitor
XLOC_008726_kunitz protease inhibitor
XLOC_027471_kunitz protease inhibitor
XLOC_008533_serine protease inhibitor_serpin
XLOC_035023_serine protease inhibitor_i-type
XLOC_007647_kunitz protease inhibitor
XLOC_006137_kunitz protease inhibitor
XLOC_008694_kunitz protease inhibitor
XLOC_042987_serine protease inhibitor_kazal-type
XLOC_027490_kunitz protease inhibitor
XLOC_027747_kunitz protease inhibitor

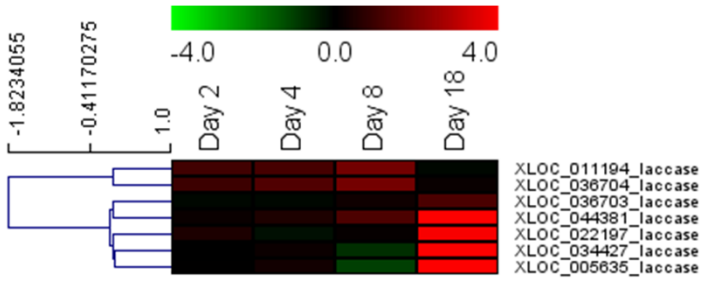
K


-1.2459044
-0.12295222
1.0
-4.0
0.0
4.0
Day 2
Day 4
Day 8
Day 18
XLOC_036487_peroxidase
XLOC_011515_peroxidase
XLOC_015716_peroxidase
XLOC_049168_peroxidase
XLOC_027206_peroxidase
XLOC_013166_peroxidase
XLOC_029177_peroxidase
XLOC_015715_peroxidase
XLOC_015730_peroxidase
XLOC_015729_peroxidase
XLOC_035612_peroxidase
XLOC_008765_peroxidase
XLOC_025633_peroxidase
XLOC_035592_peroxidase
XLOC_014884_peroxidase
XLOC_008857_peroxidase
XLOC_034858_peroxidase
XLOC_042574_peroxidase
XLOC_032098_peroxidase
XLOC_044452_peroxidase
XLOC_018392_peroxidase
XLOC_028680_peroxidase
XLOC_025820_peroxidase
XLOC_009848_peroxidase
XLOC_015676_peroxidase
XLOC_037097_peroxidase
XLOC_046743_peroxidase
XLOC_013339_peroxidase
XLOC_014426_peroxidase

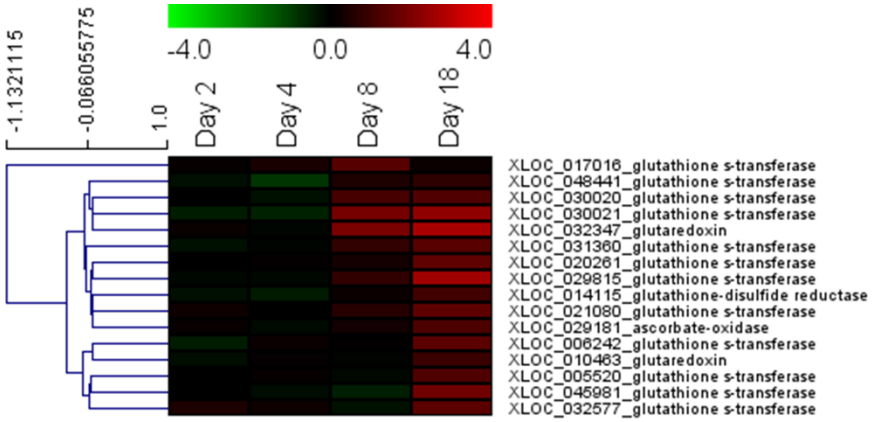
L


-1.3299401
-0.16497004
1.0
-4.0
0.0
4.0
Day 2
Day 4
Day 8
Day 18
XLOC_022896_lipid-transfer protein
XLOC_028273_lipid-transfer protein
XLOC_022267_lipid-transfer protein
XLOC_038561_lipid-transfer protein
XLOC_028173_lipid-transfer protein
XLOC_005456_lipid-transfer protein
XLOC_028272_lipid-transfer protein
XLOC_023032_lipid-transfer protein
XLOC_005454_lipid-transfer protein
XLOC_037400_lipid-transfer protein
XLOC_028179_lipid-transfer protein
XLOC_030188_lipid-transfer protein
XLOC_043418_lipid-transfer protein
XLOC_043933_lipid-transfer protein
XLOC_041874_lipid-transfer protein
XLOC_042222_lipid-transfer protein
XLOC_028177_lipid-transfer protein
XLOC_026961_lipid-transfer protein

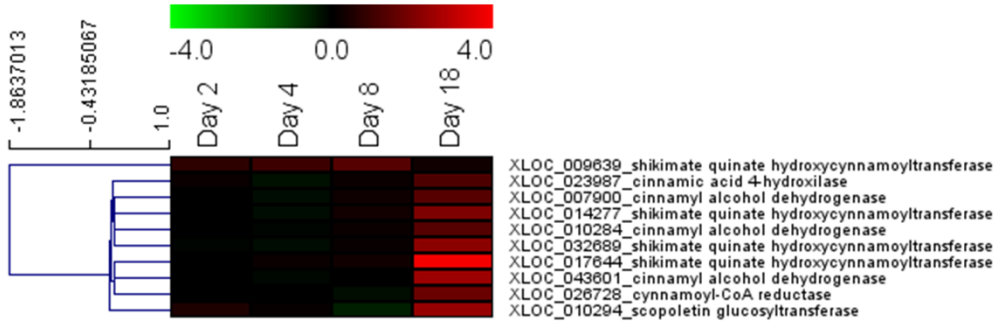
M


-1.8234055
-0.41170275
1.0
-4.0
0.0
4.0
Day 2
Day 4
Day 8
Day 18
XLOC_011194_laccase
XLOC_036704_laccase
XLOC_036703_laccase
XLOC_044381_laccase
XLOC_022197_laccase
XLOC_034427_laccase
XLOC_005635_laccase

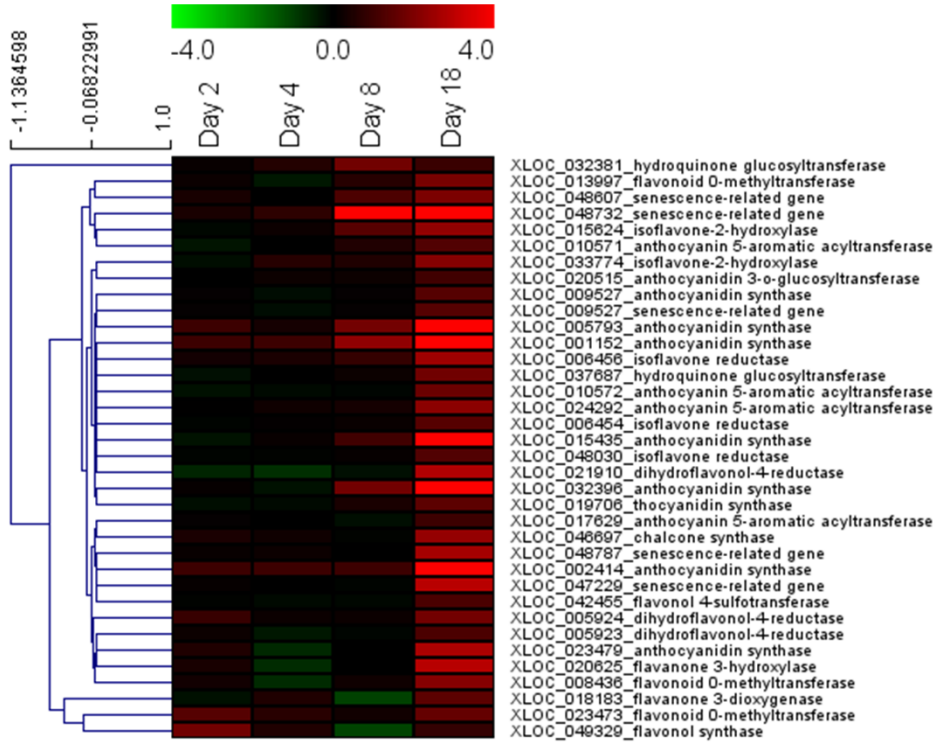
N


-1.1321115
-0.066055775
1.0
-4.0
0.0
4.0
Day 2
Day 4
Day 8
Day 18
XLOC_017016_glutathione s-transferase
XLOC_048441_glutathione s-transferase
XLOC_030020_glutathione s-transferase
XLOC_030021_glutathione s-transferase
XLOC_032347_glutaredoxin
XLOC_031360_glutathione s-transferase
XLOC_020261_glutathione s-transferase
XLOC_029815_glutathione s-transferase
XLOC_014115_glutathione-disulfide reductase
XLOC_021080_glutathione s-transferase
XLOC_029181_ascorbate-oxidase
XLOC_006242_glutathione s-transferase
XLOC_010463_glutaredoxin
XLOC_005520_glutathione s-transferase
XLOC_045981_glutathione s-transferase
XLOC_032577_glutathione s-transferase

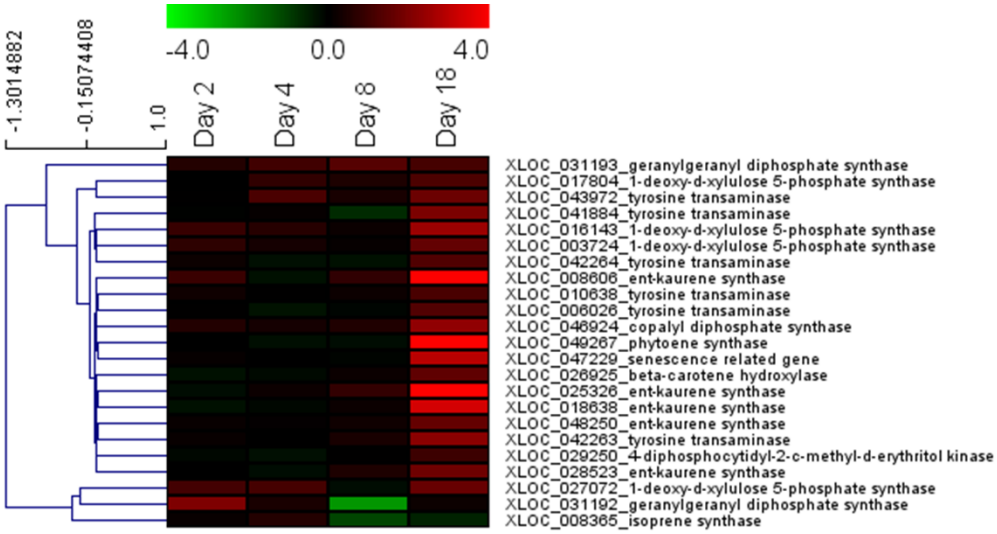
O


-1.8637013
-0.43185067
1.0
-4.0
0.0
4.0
Day 2
Day 4
Day 8
Day 18
XLOC_009639_shikimate quinate hydroxycynnamoyltransferase
XLOC_023987_cinnamic acid 4-hydroxilase
XLOC_007900_cinnamyl alcohol dehydrogenase
XLOC_014277_shikimate quinate hydroxycynnamoyltransferase
XLOC_010284_cinnamyl alcohol dehydrogenase
XLOC_032689_shikimate quinate hydroxycynnamoyltransferase
XLOC_017644_shikimate quinate hydroxycynnamoyltransferase
XLOC_043601_cinnamyl alcohol dehydrogenase
XLOC_026728_cynnamoyl-CoA reductase
XLOC_010294_scopoletin glucosyltransferase

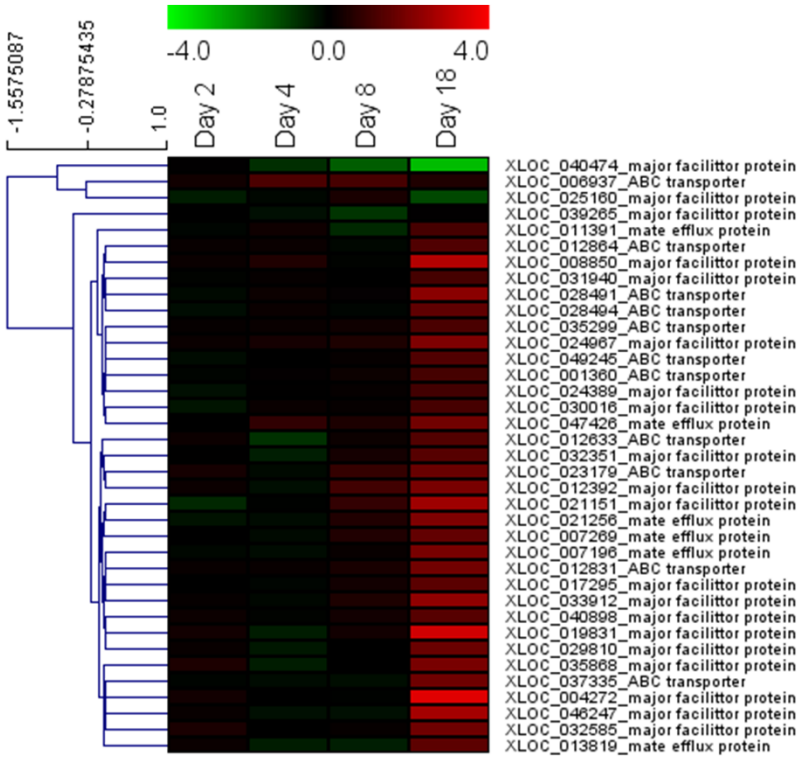
P


-1.1364598
-0.06822991
1.0
-4.0
0.0
4.0
Day 2
Day 4
Day 8
Day 18
XLOC_032381_hydroquinone glucosyltransferase
XLOC_013997_flavonoid 0-methyltransferase
XLOC_048607_senescence-related gene
XLOC_048732_senescence-related gene
XLOC_015624_isoflavone-2-hydroxylase
XLOC_010571_anthocyanin 5-aromatic acyltransferase
XLOC_033774_isoflavone-2-hydroxylase
XLOC_020515_anthocyanidin 3-o-glucosyltransferase
XLOC_009527_anthocyanidin synthase
XLOC_009527_senescence-related gene
XLOC_005793_anthocyanidin synthase
XLOC_001152_anthocyanidin synthase
XLOC_006456_isoflavone reductase
XLOC_037687_hydroquinone glucosyltransferase
XLOC_010572_anthocyanin 5-aromatic acyltransferase
XLOC_024292_anthocyanin 5-aromatic acyltransferase
XLOC_006454_isoflavone reductase
XLOC_015435_anthocyanidin synthase
XLOC_048030_isoflavone reductase
XLOC_021910_dihydroflavonol-4-reductase
XLOC_032396_anthocyanidin synthase
XLOC_019706_thocyanidin synthase
XLOC_017629_anthocyanin 5-aromatic acyltransferase
XLOC_046697_chalcone synthase
XLOC_048787_senescence-related gene
XLOC_002414_anthocyanidin synthase
XLOC_047229_senescence-related gene
XLOC_042455_flavonol 4-sulfotransferase
XLOC_005924_dihydroflavonol-4-reductase
XLOC_005923_dihydroflavonol-4-reductase
XLOC_023479_anthocyanidin synthase
XLOC_020625_flavanone 3-hydroxylase
XLOC_008436_flavonoid 0-methyltransferase
XLOC_018183_flavanone 3-dioxygenase
XLOC_023473_flavonoid 0-methyltransferase
XLOC_049329_flavonol synthase

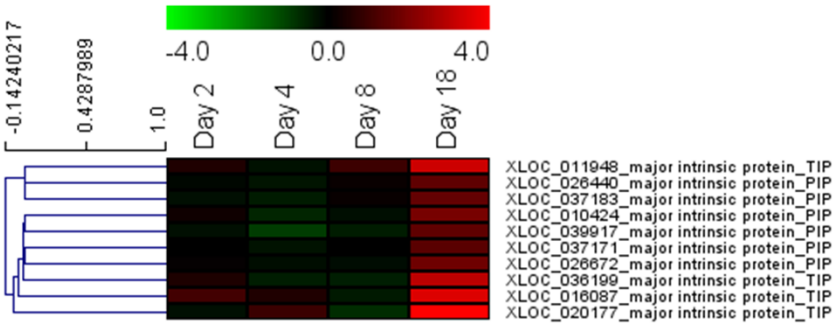
Q


-1.3014882
-0.15074408
1.0
-4.0
0.0
4.0
Day 2
Day 4
Day 8
Day 18
XLOC_031193_geranylgeranyl diphosphate synthase
XLOC_017804_1-deoxy-d-xylulose 5-phosphate synthase
XLOC_043972_tyrosine transaminase
XLOC_041884_tyrosine transaminase
XLOC_016143_1-deoxy-d-xylulose 5-phosphate synthase
XLOC_003724_1-deoxy-d-xylulose 5-phosphate synthase
XLOC_042264_tyrosine transaminase
XLOC_008606_ent-kaurene synthase
XLOC_010638_tyrosine transaminase
XLOC_006026_tyrosine transaminase
XLOC_046924_copalyl diphosphate synthase
XLOC_049267_phytoene synthase
XLOC_047229_senescence related gene
XLOC_026925_beta-carotene hydroxylase
XLOC_025326_ent-kaurene synthase
XLOC_018638_ent-kaurene synthase
XLOC_048250_ent-kaurene synthase
XLOC_042263_tyrosine transaminase
XLOC_029250_4-diphosphocytidyl-2-c-methyl-d-erythritol kinase
XLOC_028523_ent-kaurene synthase
XLOC_027072_1-deoxy-d-xylulose 5-phosphate synthase
XLOC_031192_geranylgeranyl diphosphate synthase
XLOC_008365_isoprene synthase

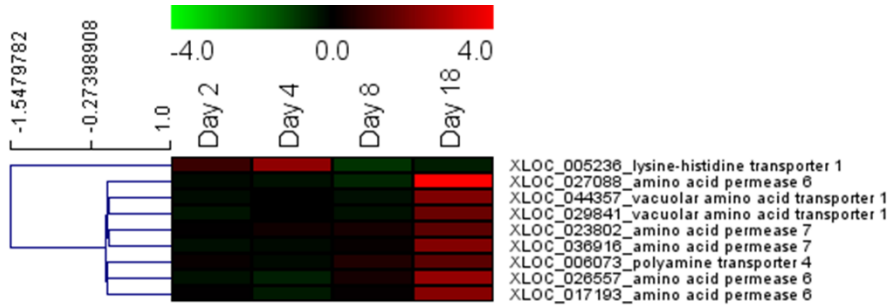
R


-1.5575087
-0.27875435
1.0
-4.0
0.0
4.0
Day 2
Day 4
Day 8
Day 18
XLOC_040474_major facilittor protein
XLOC_006937_ABC transporter
XLOC_025160_major facilittor protein
XLOC_039265_major facilittor protein
XLOC_011391_mate efflux protein
XLOC_012864_ABC transporter
XLOC_008850_major facilittor protein
XLOC_031940_major facilittor protein
XLOC_028491_ABC transporter
XLOC_028494_ABC transporter
XLOC_035299_ABC transporter
XLOC_024967_major facilittor protein
XLOC_049245_ABC transporter
XLOC_001360_ABC transporter
XLOC_024389_major facilittor protein
XLOC_030016_major facilittor protein
XLOC_047426_mate efflux protein
XLOC_012633_ABC transporter
XLOC_032351_major facilittor protein
XLOC_023179_ABC transporter
XLOC_012392_major facilittor protein
XLOC_021151_major facilittor protein
XLOC_021256_mate efflux protein
XLOC_007269_mate efflux protein
XLOC_007196_mate efflux protein
XLOC_012831_ABC transporter
XLOC_017295_major facilittor protein
XLOC_033912_major facilittor protein
XLOC_040898_major facilittor protein
XLOC_019831_major facilittor protein
XLOC_029810_major facilittor protein
XLOC_035868_major facilittor protein
XLOC_037335_ABC transporter
XLOC_004272_major facilittor protein
XLOC_046247_major facilittor protein
XLOC_032585_major facilittor protein
XLOC_013819_mate efflux protein

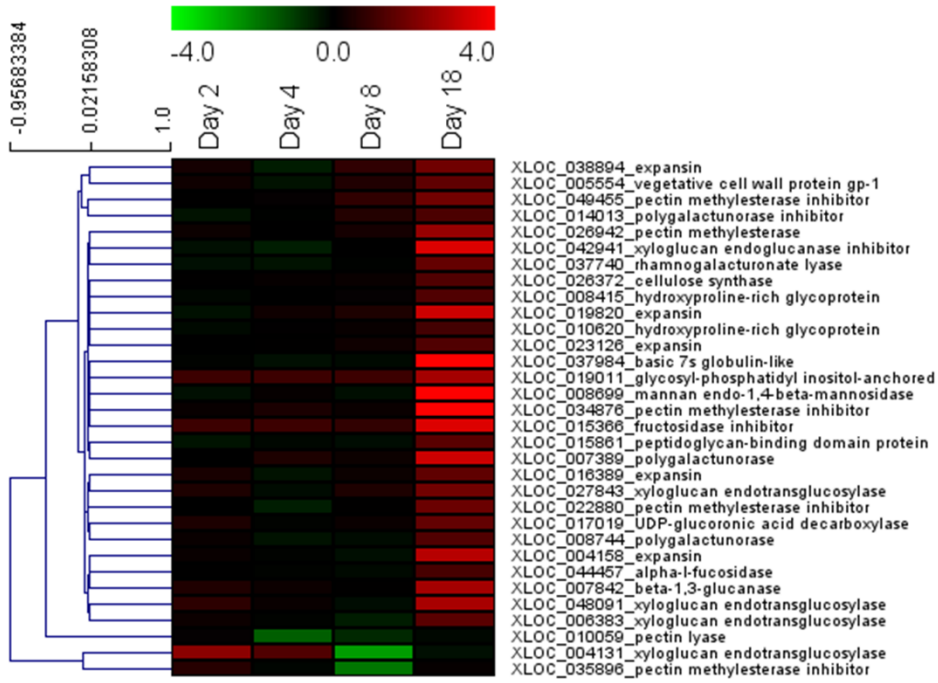
S


-0.14240217
0.4287989
1.0
-4.0
0.0
4.0
Day 2
Day 4
Day 8
Day 18
XLOC_011948_major intrinsic protein_TIP
XLOC_026440_major intrinsic protein_PIP
XLOC_037183_major intrinsic protein_PIP
XLOC_010424_major intrinsic protein_PIP
XLOC_039917_major intrinsic protein_PIP
XLOC_037171_major intrinsic protein_PIP
XLOC_026672_major intrinsic protein_PIP
XLOC_036199_major intrinsic protein_TIP
XLOC_016087_major intrinsic protein_TIP
XLOC_020177_major intrinsic protein_TIP

T


-4.0
0.0
4.0
-1.5479782
-0.27398908
1.0
Day 2
Day 4
Day 8
Day 18
XLOC_005236_lysine-histidine transporter 1
XLOC_027088_amino acid permease 6
XLOC_044357_vacuolar amino acid transporter 1
XLOC_029841_vacuolar amino acid transporter 1
XLOC_023802_amino acid permease 7
XLOC_036916_amino acid permease 7
XLOC_006073_polyamine transporter 4
XLOC_026557_amino acid permease 6
XLOC_017193_amino acid permease 6

U


-4.0
0.0
4.0
-0.95683384
0.02158308
1.0
Day 2
Day 4
Day 8
Day 18
XLOC_038894_expansin
XLOC_005554_vegetative cell wall protein gp-1
XLOC_049455_pectin methylesterase inhibitor
XLOC_014013_polygalactunorase inhibitor
XLOC_026942_pectin methylesterase
XLOC_042941_xyloglucan endoglucanase inhibitor
XLOC_037740_rhamnogalacturonate lyase
XLOC_026372_cellulose synthase
XLOC_008415_hydroxyproline-rich glycoprotein
XLOC_019820_expansin
XLOC_010620_hydroxyproline-rich glycoprotein
XLOC_023126_expansin
XLOC_037984_basic 7s globulin-like
XLOC_019011_glycosyl-phosphatidyl inositol-anchored
XLOC_008699_mannan endo-1,4-beta-mannosidase
XLOC_034876_pectin methylesterase inhibitor
XLOC_015366_fructosidase inhibitor
XLOC_015861_peptidoglycan-binding domain protein
XLOC_007389_polygalactunorase
XLOC_016389_expansin
XLOC_027843_xyloglucan endotransglucosylase
XLOC_022880_pectin methylesterase inhibitor
XLOC_017019_UDP-glucoronic acid decarboxylase
XLOC_008744_polygalactunorase
XLOC_004158_expansin
XLOC_044457_alpha-l-fucosidase
XLOC_007842_beta-1,3-glucanase
XLOC_048091_xyloglucan endotransglucosylase
XLOC_006383_xyloglucan endotransglucosylase
XLOC_010059_pectin lyase
XLOC_004131_xyloglucan endotransglucosylase
XLOC_035896_pectin methylesterase inhibitor

V

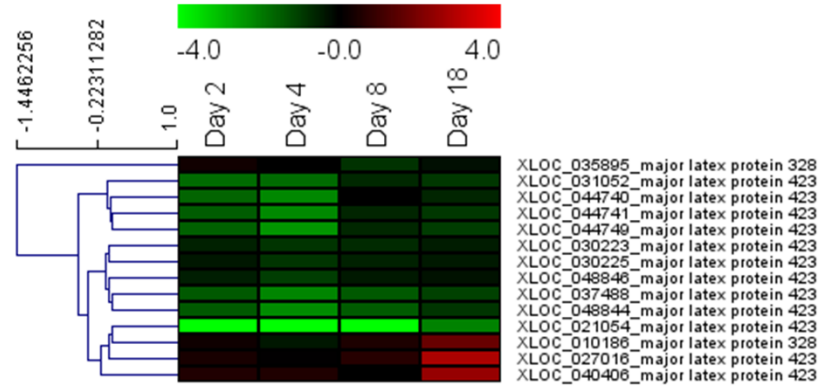

-4.0
-0.0
4.0
-1.4462256
-0.22311282
1.0
Day 2
Day 4
Day 8
Day 18
XLOC_035895_major latex protein 328
XLOC_031052_major latex protein 423
XLOC_044740_major latex protein 423
XLOC_044741_major latex protein 423
XLOC_044749_major latex protein 423
XLOC_030223_major latex protein 423
XLOC_030225_major latex protein 423
XLOC_048846_major latex protein 423
XLOC_037488_major latex protein 423
XLOC_048844_major latex protein 423
XLOC_021054_major latex protein 423
XLOC_010186_major latex protein 328
XLOC_027016_major latex protein 423
XLOC_040406_major latex protein 423

Figure 4.9 Expression patterns of major gene groups in flax through the time course upon inoculation with *F. oxysporum* f. sp. *lini*. Genes depicted are significantly differentially expressed at least at one time point (q < 0.05). A. Signal perception genes. B. Calcium-related genes. C. Kinases. D. Transcription factors. E. Jasmonate-related. F. Ethylene-related. G. Auxin-related. H. Chitinases. I. Thaumatins. J. Protease inhibitors. K. Peroxidases. L. Lipid transfer proteins. M. Laccases. N. Gluthatione-related. O. Phenylpropanoid metabolism. P. Flavonoid metabolism. Q. Isoprenoid metabolism. R. Transporters. S. Major intrinsic proteins. T. Amino acid permeases. U. Cell wall. V. Major latex proteins.

4.4.5.10 Other genes

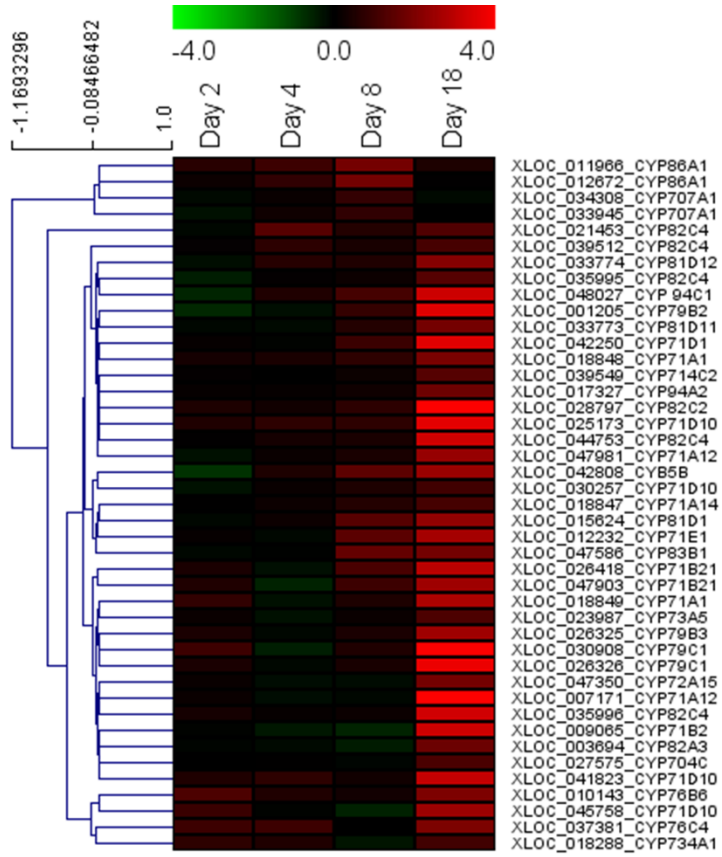
Many other groups of genes emerged during the time course of infection. Many of these genes are used in different metabolic processes and therefore are not easily placed in an exclusive category or group.

Cytochrome family proteins (CYP) P450 were the most numerous group of genes showing regulated transcripts after transcription factors. These genes are part of diverse processes in cells including primary and secondary metabolism. All but four of the transcripts bearing similarity to CYP genes showed increased transcript abundance 18 DPI (**Figure 4.10A**). Likewise, UGTs can transfer UDP-glucose to different molecules including hormones and secondary metabolites. The transcripts classified under this category were mainly repressed on day 2, with some genes showing increased transcripts at days 4 and 8, and most genes with higher activation 18 DPI (**Figure 4.10B**).

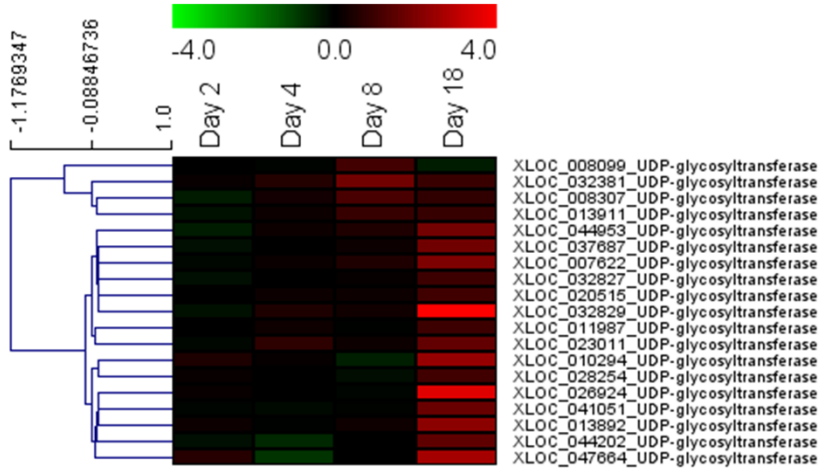
Other broad groups of genes are those that can modify or degrade other proteins or related compounds. These included enzymes like subtilases, aspartic and serine proteases and ubiquitin genes related to degradation via the 26S proteasome complex (**Figure 4.10C-D**). These enzymes had uniformly high transcript abundance 18 DPI.

Finally, two groups of enzymes related to the lipid metabolism (GDSL lipases and alpha beta hydrolases – **Figure 4.10E-F**) comprise genes with broad substrate specificity, and represent diverse functions. While the alpha beta hydrolases had a higher transcript number 18 DPI, the GDSL lipases had one group with increased transcript abundance from day 2 to 8, but with repression on day 18, and another group which mainly followed the opposite pattern.

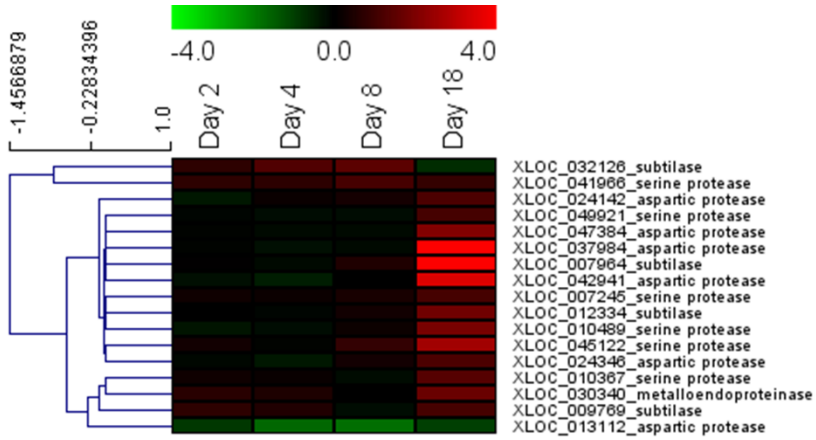
A


-1.1693296
-0.08466482
1.0
-4.0
0.0
4.0
Day 2
Day 4
Day 8
Day 18
XLOC_011966_CYP86A1
XLOC_012672_CYP86A1
XLOC_034308_CYP707A1
XLOC_033945_CYP707A1
XLOC_021453_CYP82C4
XLOC_039512_CYP82C4
XLOC_033774_CYP81D12
XLOC_035995_CYP82C4
XLOC_048027_CYP 94C1
XLOC_001205_CYP79B2
XLOC_033773_CYP81D11
XLOC_042250_CYP71D1
XLOC_018848_CYP71A1
XLOC_039549_CYP714C2
XLOC_017327_CYP94A2
XLOC_028797_CYP82C2
XLOC_025173_CYP71D10
XLOC_044753_CYP82C4
XLOC_047981_CYP71A12
XLOC_042808_CYB5B
XLOC_030257_CYP71D10
XLOC_018847_CYP71A14
XLOC_015624_CYP81D1
XLOC_012232_CYP71E1
XLOC_047586_CYP83B1
XLOC_026418_CYP71B21
XLOC_047903_CYP71B21
XLOC_018849_CYP71A1
XLOC_023987_CYP73A5
XLOC_026325_CYP79B3
XLOC_030908_CYP79C1
XLOC_026326_CYP79C1
XLOC_047350_CYP72A15
XLOC_007171_CYP71A12
XLOC_035996_CYP82C4
XLOC_009065_CYP71B2
XLOC_003694_CYP82A3
XLOC_027575_CYP704C
XLOC_041823_CYP71D10
XLOC_010143_CYP76B6
XLOC_045758_CYP71D10
XLOC_037381_CYP76C4
XLOC_018288_CYP734A1

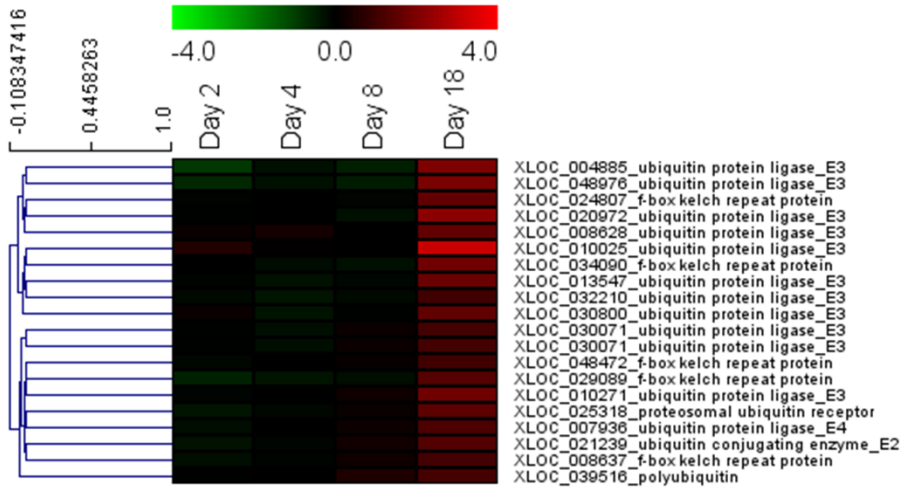
B


-1.1769347
-0.08846736
1.0
-4.0
0.0
4.0
Day 2
Day 4
Day 8
Day 18
XLOC_008099_UDP-glycosyltransferase
XLOC_032381_UDP-glycosyltransferase
XLOC_008307_UDP-glycosyltransferase
XLOC_013911_UDP-glycosyltransferase
XLOC_044953_UDP-glycosyltransferase
XLOC_037687_UDP-glycosyltransferase
XLOC_007622_UDP-glycosyltransferase
XLOC_032827_UDP-glycosyltransferase
XLOC_020515_UDP-glycosyltransferase
XLOC_032829_UDP-glycosyltransferase
XLOC_011987_UDP-glycosyltransferase
XLOC_023011_UDP-glycosyltransferase
XLOC_010294_UDP-glycosyltransferase
XLOC_028254_UDP-glycosyltransferase
XLOC_026924_UDP-glycosyltransferase
XLOC_041051_UDP-glycosyltransferase
XLOC_013892_UDP-glycosyltransferase
XLOC_044202_UDP-glycosyltransferase
XLOC_047664_UDP-glycosyltransferase

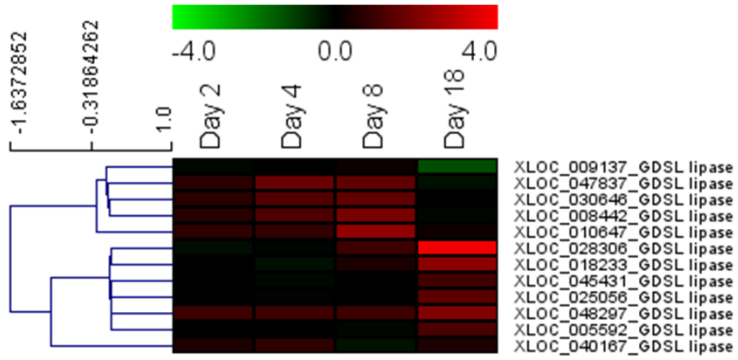
C


-1.4566879
-0.2283496
1.0
-4.0
0.0
4.0
Day 2
Day 4
Day 8
Day 18
XLOC_032126_subtilase
XLOC_041966_serine protease
XLOC_024142_aspartic protease
XLOC_049921_serine protease
XLOC_047384_aspartic protease
XLOC_037984_aspartic protease
XLOC_007964_subtilase
XLOC_042941_aspartic protease
XLOC_007245_serine protease
XLOC_012334_subtilase
XLOC_010489_serine protease
XLOC_045122_serine protease
XLOC_024346_aspartic protease
XLOC_010367_serine protease
XLOC_030340_metalloendoproteinase
XLOC_009769_subtilase
XLOC_013112_aspartic protease

D


-0.1083474
0.4458263
1.0
-4.0
0.0
4.0
Day 2
Day 4
Day 8
Day 18
XLOC_004885_ubiquitin protein ligase_E3
XLOC_048976_ubiquitin protein ligase_E3
XLOC_024807_f-box kelch repeat protein
XLOC_020972_ubiquitin protein ligase_E3
XLOC_008628_ubiquitin protein ligase_E3
XLOC_010025_ubiquitin protein ligase_E3
XLOC_034090_f-box kelch repeat protein
XLOC_013547_ubiquitin protein ligase_E3
XLOC_032210_ubiquitin protein ligase_E3
XLOC_030800_ubiquitin protein ligase_E3
XLOC_030071_ubiquitin protein ligase_E3
XLOC_030071_ubiquitin protein ligase_E3
XLOC_048472_f-box kelch repeat protein
XLOC_029089_f-box kelch repeat protein
XLOC_010271_ubiquitin protein ligase_E3
XLOC_025318_proteosomal ubiquitin receptor
XLOC_007936_ubiquitin protein ligase_E4
XLOC_021239_ubiquitin conjugating enzyme_E2
XLOC_008637_f-box kelch repeat protein
XLOC_039516_polyubiquitin

E


-1.6372852
-0.3186426
1.0
-4.0
0.0
4.0
Day 2
Day 4
Day 8
Day 18
XLOC_009137_GDSL lipase
XLOC_047837_GDSL lipase
XLOC_030646_GDSL lipase
XLOC_008442_GDSL lipase
XLOC_010647_GDSL lipase
XLOC_028306_GDSL lipase
XLOC_018233_GDSL lipase
XLOC_045431_GDSL lipase
XLOC_025056_GDSL lipase
XLOC_048297_GDSL lipase
XLOC_005592_GDSL lipase
XLOC_040167_GDSL lipase

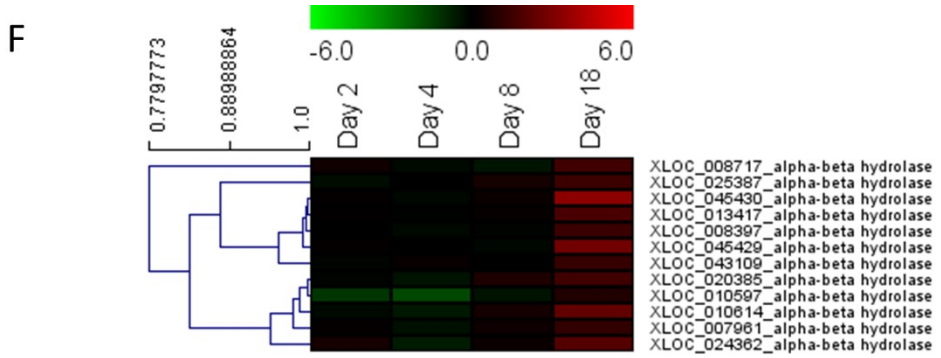


Figure 4.10 Expression patterns of other relevant gene groups in flax through the time course upon inoculation with *F. oxysporum* f. sp. *lini*. Genes depicted are significantly differentially expressed at least at one time point (q < 0.05). A. Cytochrome P genes. B. UDP glucosyltransferases. C. Protein degradation. D. Proteosome/ubiquitin-related. E. GDSL lipases. F. Alpha/beta hydrolases.

4.5 Discussion

4.5.1 Disease progression difference in two flax cultivars

The lack of molecular information about pathogenesis of *F. oxysporum* f. sp *lini* in flax limits efforts to increase resistance to this disease. In this study, Lutea developed more severe disease symptoms than CDC Bethune (**Figure 4.1** and **Figure 4.2**), and also had a significant shoot length reduction in response to one of the two *F. oxysporum* f. sp. *lini* isolates tested (**Figure 4.2B**). Differences between the cultivars were also evident at the molecular level: in two of the four chitinases evaluated by qRT-PCR, changes in transcript abundance were marked 8 DPI for Lutea, and not until 22 DPI for CDC Bethune, suggesting constitutive defenses or mechanisms to delay the pathogen interaction in the latter variety, as has been seen in other pathosystems [342]. These two chitinases are presumed orthologs of a class I basic chitinase (At3g12500) that was initially classified as a pathogen-induced and defense-related protein [288], and has shown upregulation in other systems upon pathogen incursion [343,344]. Therefore, the increase in the two orthologous chitinases represents a biomarker for *F. oxysporum* f. sp. *lini* infection, and can be linked to the defense response of flax.

4.5.2 Transcriptome regulation upon *F. oxysporum* f. sp. *lini* infection in CDC Bethune

We followed the transcriptome response of the more resistant cultivar, CDC Bethune, through a time course of 2, 4, 8, and 18 DPI. While earlier in the cycle, more downregulated than

upregulated genes were present (**Table 4.3**), the pattern was reversed after 8 DPI. This is in agreement with a pattern seen in wheat infected with *Zymoseptoria tritici*, where several genes involved in plant defense (pathogenesis-related proteins, resistance genes, WRKY TFs, oxylipin and lignin-related genes and detoxification proteins) where downregulated 1 DPI and the pattern was only reversed 8 DPI [345]. Interestingly both *Z. tritici* and *F. oxypsorum* are hemibitrophic fungi and it is possible that gene expression changes reflect transitions between the biotrophic stage, were perception takes place, and a second phase when the fungus becomes necrotrophic and the plant activates novel plant defenses.

4.5.2.1 Pathogen elicitor perception

Several genes are directly or indirectly associated with signal perception of pathogens [346]. Broadly, these genes can be divided in two groups: i) receptors that detect signals from general microbial products (related to pathogen fitness) including receptor-like kinases (RLKs) and receptor-like proteins (RLPs), which recognize pathogen-associated molecular patterns (PAMPs) and are involved in general immune response; and ii) resistance genes (*R*-genes) which recognize specific effectors (related to pathogen virulence) that are race specific and are related to triggering vertical resistance in plants. The two groups of genes are related to PAMP triggered immunity (PTI) and effector triggered immunity (ETI), although a clear cut division between these two classes of defense responses is not evident [148].

Among these, for example, a receptor-like cytoplasmatic kinase (RLCK) corresponding to transcripts XLOC_015563 and XLOC_030874, is an ortholog of *Arabidopsis* ATG05940.1. This gene encodes a RPMI-induced protein kinase (RIPK), and was induced >2- fold (log2 scale) by the treatments. The gene was identified as a key component of the phosphorylation of RIN4 in the presence of pathogen effectors, which is in turn recognized by the NB-LRR immune receptor RPM1, activating ETI [347].

Four disease resistance proteins (TIR-NBS-LRR), which would have been expected to increase their abundance upon pathogen attack, showed decreased transcript abundance throughout the cycle (**Figure 4.9A**). Downregulation of such genes has been found to be controlled by host miRNA in many plant-pathogen interactions, but the repression is usually lifted upon pathogen attack [348–351], which is contrary to our results. However, in wheat, rust

miRNAs that match NBS-LRR proteins could be an indication of how pathogens can interfere with plant defenses [352].

The remaining receptor-like kinases (RLKs) that showed increased transcript abundance 18 DPI (**Figure 4.9A**) could be related to recognition of pathogen proteins or sugars (e.g. chitobiose), or to cell death and immunity, as is common with some SERK1 family members [353,354].

4.5.2.2 Signal transduction

G-proteins have been shown to transduce the pathogen detection signals from RLKs [355]. A few G-protein-related transcripts differed significantly in abundance only at 18 DPI (e.g. XLOC_047108, XLOC_030817, XLOC_38580, XLOC_017620 - Dataverse file: transcript differential expression in flax upon *Foln* infection http://dx.doi.org/10.7939/DVN/10933). It is possible some of these proteins serve as transducers of fungal response as reported for necrotrophic fungi in *Arabidopsis thaliana* [356–358]. Calcium is a secondary messenger involved in environmental changes, and its influx from the apoplast and vacuoles into the cytoplasm is a classic response to pathogen infections [359–361]. Calcium-dependent genes increased significantly in transcript abundance only 18 DPI (**Figure 4.9B**). Changes in calcium-related gene response are in agreement with physiological measurements of calcium influx upon flax roots colonization by *F. oxysporum* [223], and are also represented in other fungal infections [331,362–364]. Finally, several kinases related to transduction signals (e.g. CIPKs, MAPKs, phosphatases) were mainly activated 18 DPI (**Figure 4.9C**). CIPKs have been implicated directly in processes like abscisic acid (ABA) perception and signalling; ABA itself, can act as messenger in response to pathogens [365].

Most signalling genes (*R*-genes, calcium and kinases) as well as TFs (see below), were strongly induced 18 DPI, however, several pathogen responsive genes were already active earlier in the cycle at 4 and 8 DPI (e.g. chitinases). A possible explanation is that some defense components may be constitutively activated and may be reinforced upon initial pathogen detection. For example, the chitin signalling process which can result in activation of several defense genes depends on chitinases degrading fungal cell walls, oligomer detection by the chitin elicitor binding protein (CEBiP), and signal transduction by a LysM domain-containing receptor-like kinase 1 (LysM RLK1) [366]. Examination of these genes showed that several chitinases

(FPKMs > 100) and *CEBiP* and *LysM* RLKs (FPKMs > 10) were constitutively expressed throughout the time course (not shown). This explains how pathogen signals could be detected and reinforce pathogen responsive genes.

4.5.2.3 Transcriptional regulation

TFs commonly reported to be involved in modulation of plant defenses include: WRKY, ethylene responsive factors (ERFs), basic-region leucine zipper protein (bZIP), MYB, DNA-binding with one finger protein (DOF), Whirly, MYC and NAC [287]. Most TFs in our study were differentially expressed at 18 DPI (**Figure 4.9K** and Dataverse file: transcript differential expression in flax upon *Foln* infection http://dx.doi.org/10.7939/DVN/10933).

The most prominent group of TFs in our study were the WRKYs (**Figure 4.9D**), which are key regulators of plant innate immunity [367]. A presumed flax ortholog of *WRKY70* was consistently less abundant at 2, 4 and 8 DPI, but not at 18 DPI. Overexpression of this gene positively regulates SA-mediated responses and supresses JA-mediated defenses [368,369]. However, expression of JA-related genes 18 DPI was not in agreement with this (see below). Other WRKY genes were responsive at 18 DPI in our study. For example, the ortholog transcripts of *WRKY75* (XLOC_020440 and XLOC_014625) which increased its transcript abundance in our infected flax plants, also increased in abundance in *Brassica napus* upon challenge with *Sclerotinia sclerotiorum* and *Alternaria brassicae* [369]. Finally, mutants of *AtWRKY3*, which also has a flax ortholog induced by infection in our study (XLOC_010407), have shown increased susceptibility to *B. cinerea* [370].

MYBs are a large family of TFs of which only a small number are directly involved in the response to pathogens [287,371]. *MYB113* has been previously reported as induced in *F. oxysporum* inoculations on *Arabidopsis* [372], and seems critical in the production of anthocyanins which comprise specific stages of phenylpropanoid metabolism [373]. The presumed ortholog of *MYB113* (XLOC_011383, Dataverse file: transcript differential expression in flax upon *Foln* infection http://dx.doi.org/10.7939/DVN/10933) was the MYB that showed the highest increase in transcript abundance from TFs in our study at 18 DPI (4.1 log2-fold change). Likewise, XLOC_015378 and XLOC_016754 show close similarity to *Arabidopsis MYB108*, which is necessary through the JA pathway for resistance to *Botrytis cinerea* infection [374] .

4.5.2.4 Hormone regulation

Besides their role in plant growth and development, plant hormones play a complex role in the signalling response to pathogen attack downstream of the initial plant-pathogen recognition [375]. Hormones like JA and ET control the expression of many PR-genes and have feedback loops of regulation to already expressed components.

Several genes related to JA biosynthesis increased in transcript abundance starting 4 DPI, but it was only 18 DPI when the majority of these transcripts presented increased abundance (**Figure 4.9E**). Transcripts induced included *LOXs*, *AOS*, *AOC* and *OPR*, and several JAZ proteins which negatively regulate JA transcriptional activity [376,377]. We also found three transcripts with increased abundance encoding a transcriptional activator of JA responses (*MYC2*) (Dataverse file: transcript differential expression in flax upon *Foln* infection http://dx.doi.org/10.7939/DVN/10933), which is repressed by JAZ proteins [377]. Simultaneous activation of both JA biosynthetic genes and their repressors (JAZ domain proteins) has also been reported in *Arabidopsis* [378,379]. This demonstrates there is a balance between JA production and inactivation, probably to impair excessive levels of JA [378,380].

Ethylene, another important hormone in signalling pathways, usually acts synergistically with JA in the defense against necrotrophic pathogens [375]. We observed increased transcript abundance of two key ethylene biosynthesis enzymes, corresponding to two transcripts of *ACS* and to nine transcripts of *ACO* at 18 DPI (**Figure 4.9F**), and of ethylene response factors *ERF1* (XLOC_005021, XLOC039651) and *ERF14* (XLOC_005023). The activation of these biosynthetic genes and of ERFs upon *Fusarium* species inoculation in plant hosts [328,331] and other fungal pathogens [326] has been documented in other species. *ERF1* is probably one of the most important markers involved in plant defense against fungal pathogens, linked to *Arabidopsis* resistance to both *F. oxysporum* sp. *conglutinans* and *F. oxysporum* sp. *lycopersici* [381]. *ERF14* has not only proven to be relevant in the defense against *F. oxysporum*, but also regulates the expression of other ERFs including *ERF1* [382]. *ERF14* was the most responsive ethylene-related gene in our study, with a 4.3 log2-fold change at 18 DPI (Dataverse file: transcript differential expression in flax upon *Foln* infection http://dx.doi.org/10.7939/DVN/10933).

Auxins are involved in most plant development processes and are known to also repress SA [375]. Two genes corresponding to an auxin-binding RmlC cupin and an auxin-induced saur

(small auxin-up RNA) were clearly repressed 8 DPI (**Figure 4.9G**). Saur genes are related to cell expansion [383] demonstrating that this process could be impaired at this time point, which is in agreement with the downregulation of this gene in *F. oxysporum*-infected *A. thaliana* roots [384]. At 18 DPI, two CYP450 family genes (*CYP79B2* and *CYP79B3* – **Figure 4.10A**) that transform tryptophan to indole-3-acetaldoxime (IAOx, a precursor of IAA and indole glucosinolates) [385,386] were present, indicating positive auxin regulation. Likewise, indole-3-acetic acid (IAA) amino acid/amido hydrolases increased in transcript abundance 8 DPI (**Figure 4.9G**), and became significantly regulated 18 DPI (Dataverse file: transcript differential expression in flax upon *Foln* infection http://dx.doi.org/10.7939/DVN/10933). The activation of these hydrolases should result in an increasing pool of the hormonally active IAA that is critical for plant germination and growth [387]. Favoring growth over defense could result in greater disease progression, but there can also be alternative functions for these enzymes. Auxin conjugate hydrolases genes *IAR3* which corresponds to transcripts XLOC_029902 - XLOC_048550, and *ILL6* which corresponds to transcripts XLOC_010014 – XLOC_005151 (**Figure 4.9G**), were expressed in *Arabidopsis thaliana* and in *Brassica rapa* upon challenge by microorganisms [388,389]. *IAR3* and *ILL6* not only control auxin metabolism but are involved in deconjugation of JA-Ile, which results in hormone turnover and repression of JA-responsive genes [380,390]. This pattern also supports the involvement of these genes in JA regulation (see above).

4.5.2.5 PR proteins

PR genes accumulate in plants in response to phytopathogens in the processes of hypersensitive response (HR) and systemic acquired resistance (SAR) [391,392]. Among them, chitinases (PR-3,4,8,11 families) [393], are a first line of defense to directly disrupt the fungal cell wall, which weakens the pathogen and produces oligomers that become elicitors of additional plant defenses. We observed increased significant transcript abundance of chitinases as early as 4 DPI (Dataverse file: transcript differential expression in flax upon *Foln* infection http://dx.doi.org/10.7939/DVN/10933), but the maximum number of activated chitinases was reached 18 DPI (**Figure 4.9H**). Chitinases have been shown to be key players in the response to *F. oxysporum* infection in cabbage [394], tomato [296] and cavendish banana [329], to *F. graminearum* in wheat [331,395], as well as in other plant-fungal interactions [396–398].

Other PR proteins (thaumatins, PIs and LTPs) were also regulated upon *F. oxysporum* f. sp. *lini* infection in flax; these genes are commonly regulated by plants in pathogenic attacks [328,331,396,397,399–402]. Thaumatins (PR-5) (**Figure 4.9I**) can cause increased permeabilization of the fungal cell wall inflicting direct damage to fungal hyphae [403,404]. A second group of genes which is also believed to cause membrane permeability in its action against pathogens are the lipid transfer proteins (LTPs) [405]. While LTPs usually transfer lipids across membranes [393], their induction may be related to cutin production stimulated by pathogen attack [405]. In our study LTPs showed mixed patterns of increased and decreased abundance during the time-course (**Figure 4.9L**), indicating functional diversification in the protein family.

PR proteins classified as PIs (PR-6) control pathogen proteases that work against plant cell wall components or in cell degradation to obtain nutrients for pathogen growth [393,406]. Numerous PIs were found with some of the highest transcript abundances in all gene groups 18 DPI (**Figure 4.9J**).

4.5.2.6 Reactive oxygen species (ROS)

Changes associated with the recognition of the pathogen resulting in signal transduction and changes in the calcium status of the cell are triggers of an oxidative burst. ROS can be involved in HR to produce cell death, in crosslinking with glycoproteins to reinforce the cell wall, or couple with other signalling factors to induce SAR [331,407,408]. The first step in ROS production is mediated by NADPH oxidase/RBO [407]. This gene was represented by two transcripts (XLOC_017620 and XLOC_008027) with increased abundance only 18 DPI (Dataverse file: transcript differential expression in flax upon *Foln* infection http://dx.doi.org/10.7939/DVN/10933).

Also central to the production of ROS are peroxidases that can catalyze the production of H_2O_2, resulting in activation of cell defenses and programmed cell death (PCD) to stop the spread of the disease, or to be used for the synthesis of lignin when H_2O_2 and phenolic substrates are available for cell wall strengthening [409]. While four peroxidases actually decreased in transcript abundance 4 and 8 DPI and maintained non-significant repression at 18 DPI (**Figure 4.9K**), most peroxidases were among the most abundant transcripts with increased abundance 18 DPI. Interestingly, the peroxidases with largest increase in abundance (> 3 log2-fold change) in

our study (XLOC_08857, XLOC_015715, XLOC_015730, XLOC_027206 and XLOC_028680) (dataverse file: transcript differential expression in flax upon *Foln* infection http://dx.doi.org/10.7939/DVN/10933) presented very low expression across undisturbed tissues (not shown), making them good candidate markers of the host response to the pathogen. Furthermore, XLOC_015730 and XLOC_027206 whose closest *A. thaliana* ortholog is At5g05340 (known as *AtPrx52*), has been directly implicated in lignin formation under normal development [410]. This gene increases more than 40-fold 21 dpi when *A. thaliana* is infected with the fungus *Verticillum longisporum* [411], which is in the same range of the non-logarithmic fold changes of our transcripts (16.9 and 10.0 respectively) (Dataverse file: transcript differential expression in flax upon *Foln* infection http://dx.doi.org/10.7939/DVN/10933), making it another good candidate for breeding studies.

To offset potentially negative effects of ROS, scavenging enzymes can help balance the localized HR response [407,409,412]. GSTs had decreased abundance at 2 and 4 DPI, but started increasing 8 DPI and were abundant and significantly upregulated 18 DPI along with other enzymes of the ascorbate-glutathione cycle (**Figure 4.9N**). This is similar to the situation in Chinese white poplar, where expression of multiple GSTs occurs upon stem blister canker infection [413]. We also found two transcripts (XLOC_00520 and XLOC_032577) orthologous to *Arabidopsis* GST At2g29420, which has been shown to be regulated by both *B. cinerea* and *Pseudomonas syringae* [414].

4.5.2.7 Secondary metabolism

Numerous functional categories and genes related to secondary metabolism were enriched in the treated samples. These compounds synthesized in response to pathogenic infections, can act by directly exerting an antimicrobial effect (pathogen membrane disruption and pathogen protein/enzyme alteration), or indirectly as in the case of cell wall reinforcement (e.g. lignification, callose deposition), or as signalling molecules leading to defense responses, HR or PCD [415].

Phenylpropanoid metabolism is central to secondary metabolite production of defense-related compounds including monolignols and flavonoids [416]. In a previous study, flax plants infected with *F. oxysporum* and *F. culmorum* showed regulation of phenylpropanoid genes and the derived metabolites showed increased abundance [220]. Furthermore, application of fungal

elicitors from mycelium, including *F. oxysporum* on flax cell suspensions, results in activation of monolignol gene expression [319].

Flavonoids were the most represented group of secondary compounds 18 DPI (**Figure 4.9P**). Flavonoids have high antioxidant capacity, which has been used to create increased resistance to *F. oxysporum* and *F. culmorum* through engineering of transgenic flax plants with a multi-construct including chalcone synthase (*CHS*), chalcone isomerase (*CHI*), and dihydroflavonol reductase (*DFR*) from petunia [322]. One of the three transcripts representing *DFR* (XLOC_021910) showed almost a 3 log2-fold increase, while *CHS* (XLOC_046697) had a 2.4 log2-fold change (Dataverse file: transcript differential expression in flax upon *Foln* infection http://dx.doi.org/10.7939/DVN/10933). Furthermore, the transgenic plants from the aforementioned study had increased levels of anthocyanins, and we found transcripts of both *DFR* and anthocyanidin synthases which are both implicated in anthocyanin biosynthesis (**Figure 4.9P**). From seven transcripts matching anthocyanidin synthases, five have log2-fold changes > 6 (XLOC_015435, XLOC_002414, XLOC_005793, XLOC_001152 and XLOC_032396), representing some of the largest transcript changes from the study. Additional antioxidant capacity in flavonoids can be achieved by glycosylation, which yields more stable flavonoids. The introduction of UGTs in flax plants resulted in increased resistance against *Fusarium* species through the generation of flavonoid glycosides and increased levels of proanthocyanin, lignans, phenolic acids and unsaturated fatty acids [323]. We found 19 UGTs in our study, and 18 of them showed increased transcript abundance 18 DPI (**Figure 4.10B**).

Isoprenoids or terpenoids are a group of chemicals employed in growth and development but also bearing specialized functions against different forms of stress [417]. Several genes related to the metabolism of terpenoids showed increased transcript abundance 18 DPI (**Figure 4.9Q**). PSY, the key controlling enzyme for carotenoid synthesis [418], had a 4.4 log2-fold change increase in our study. This enzyme is induced by diverse stresses including salt, drought and temperature [418], but we did not find literature on *in vivo* studies linking it to pathogen response. However, the introduction of a bacterial *PSY* gene under the 35S constitutive promoter in flax demonstrated increased resistance against *F. oxysporum* and *F. culmorum* in flax [320]. Since carotenoids can act as ROS scavengers too [419], the induction of key enzymes involved in their metabolism should be critical during oxidative stress triggered by pathogens.

Finally, other secondary metabolism genes with increased transcript abundance were related to glucosinolate synthesis. The two most-induced genes involved in glucosinolate synthesis corresponded to cytochrome P450 family genes *CYP79B2* and *CYP79B3* (also important for auxin precursor synthesis). Additional genes responsible for the synthesis of indole glucosinolates from IAOx (*CYP83B1* and *SUR1*) [420,421] had increased abundance in our study (XLOC_047586, XLOC_010638, Dataverse file: transcript differential expression in flax upon *Foln* infection http://dx.doi.org/10.7939/DVN/10933), but two additional enzymes for camalexin synthesis (*CYP71A3* and *CYP71B15*) [422] did not show regulation. Genes involved in indole glucosinolate and camalexin synthesis were highly induced in *A. thalaiana* upon *F. oxysporum* infection [372]. Camalexin also generated membrane permeabilization in *Alternaria brassicicola* [423]. The importance of these tryptophan-derived metabolites was highlighted by the study of the double mutant *cyp79b2/cyp79b3* in *A. thaliana*, which resulted in increased susceptibility to the fungal pathogen *Verticillum longisporum* [424].

4.5.2.8 Transport

Adjustments in molecule transport are usually made upon pathogen attack, but changes in plant transport mechanisms can also benefit the pathogen. For example, modifications of the relative flow of water and amino acid transport can have beneficial effects for the intruder.

Multidrug transporters constitute a large family of proteins that remove cytotoxic compounds from cells using ATP or a proton pump system [425]. Thirty-seven transcripts (most with increased abundance 18 DPI – **Figure 4.9R**) belonging to three multidrug transporter superfamilies comprised MATE efflux proteins, MFS proteins and ABC transporters. Plant ABC transporters are a large family of ATP-driven pumps that aid in secondary metabolite transport to deter pathogens, and can be used for detoxification of harmful compounds (fungal toxins) [426–428].

Major intrinsic proteins (aquaporins) showed increased transcript abundance 18 DPI (**Figure 4.9S**); it is possible that their regulation is controlled by the pathogen to improve invasion into plant tissues due to greater flow of water through these membrane pores which allows easier haustorial development [429]; however another study showed that these membrane water channels could facilitate the conduction of H_2O_2 [430], and support defense signalling.

In agreement with potential host manipulation by the pathogen, we also found increased transcript abundance for amino acid transporters/permeases (**Figure 4.9T**). In *A. thaliana*, mutants for the amino acid permeases *AAP3* and *AAP6*, presented reduced infestation by nematodes which otherwise benefit from increased amino acid transport to the site of infection [431]. Three transcripts with similarity to *AAP6* (XLOC_017193, XLOC_026557 and XLOC_027088) were represented in our study (**Figure 4.9T**).

4.5.2.9 Cell wall

While some cell wall-related genes indicate defense of the plant against *F. oxysporum* f. sp. *lini* (e.g. hydroxyproline-rich glycoproteins, polygalactunorase inhibiting proteins, pectin methyelesterase inhibitor proteins), the large majority of genes with increased transcript abundance 18 DPI suggest a modification that would favor the pathogen (**Figure 4.9U**). Expansins, endotransglycosylases, glucanases, and genes involved in the pectin metabolism, indicate a potential degradation of cell wall components which would provide sugar nutrients to the fungi and ease colonization by the pathogen [318,432,433].

4.5.2.10 Major latex proteins

An interesting trend throughout the time course was seen for MLPs. From 14 transcripts matching MLPs, 11 were consistently repressed (**Figure 4.9V**). MLPs are a group of genes initially discovered as abundant proteins in the latex of the opium poppy [341]. Their function has not been completely elucidated, but there is evidence of their regulation through hormone signalling [434–436], and activation in response to pathogens [434,437,438], plus a structural similarity to pathogenesis related proteins of group 10 (PR-10) [439]. However, flax response in the current study was opposite to these studies and we found only one study where 16 major latex proteins were repressed in response to oxidative stress in *Arabidopsis* [440].

4.5.3 A model for the deployment of flax defenses against fusarium.

Most plant defense responses were clearly activated 18 DPI, and the identification of the major groups of responsive genes allowed us to make general inferences about the interactions with the pathogen. We built a model (**Figure 4.11**) showing how upon interaction between *F. oxysporum* f. sp. *lini* and flax, the fungus potentially liberates PAMPs, which are detected by

membrane receptors like RLKs (pathogen recognition receptors - PRRs), triggering innate immune responses via PTI. At the same time, effectors that act as virulence factors are probably deployed and detected by *R*-genes (or the *R*-genes detect changes in the interaction of the effectors with other cell components), in a gene-for-gene resistance fashion (vertical resistance), resulting in ETI. A gene with similarity to *RIPK* from *Arabidopsis* was part of this response and may be an important component of resistance [347], while four disease resistance proteins downregulated throughout the time course represent interesting targets to study potential immune suppression by the pathogen, probably via small RNAs. The multitude of signal receptor genes indicate that flax (CDC Bethune) uses both non-specific and isolate-specific defenses to deter the pathogen, and therefore can activate both, a general immune response (e.g. PR-genes), as well as a HR.

The interactions between PAMPs/effectors and the plant proteins can promote changes in the phosphorylation status of both, the interacting plant proteins, and downstream proteins of signalling cascades like MAP kinases. Further modifications in calcium binding proteins also promote the signal transduction process. The result of the signalling processes is a transcriptional reprogramming via numerous transcription factors that activate hormonal control, PR genes, and secondary metabolism among other processes. From TFs, flax transcripts with similarity to *WRKY3*, *WRKY70* and *WRKY75* are responsive in other plant-pathogen interactions [369,370], and are likely general responders of plant defense against multiple pathogens. On the other hand a transcript with similarity to *MYB113*, which was the most upregulated TF, and is also responsive in the *A. thaliana* - *F. oxypsorum* interaction [372], indicates this could be a good marker gene to inform levels of resistance to *F. oxysporum* infection, if phenotypes can be associated with gene induction.

The regulation of key JA and ET biosynthetic and responsive genes found in this study indicates the importance of these hormones in both the reception and transduction of signals of the defense machinery in and outside the cell. The expression of many PR-genes is mediated by these hormones, which also act as signalling compounds in systemic resistance. For example, activation of many defense genes is done via ERFs, from which two regulated transcripts were found in our study (*ERF1* and *ERF14*), both involved in resistance against *F. oxysporum* [381,382].

Among PR genes regulated by transcriptional reprograming, chitinases and thaumatins may directly affect the fungal cell wall and its cell membrane integrity. The presence of LTPs could be related to transport of lipids for cutin deposition, or directly implicated in altering the fungal membrane [405]. The transcripts representing these genes showed different expression patterns in the time course with some early expressing genes which were repressed at 18 DPI, and others that were upregulated at this same time point. Additionally, among PR-genes were PIs, which had high transcript abundance at 18 DPI, demonstrating that the pathogen has probably entered a necrotrophic phase and has deployed its arsenal to break cell walls.

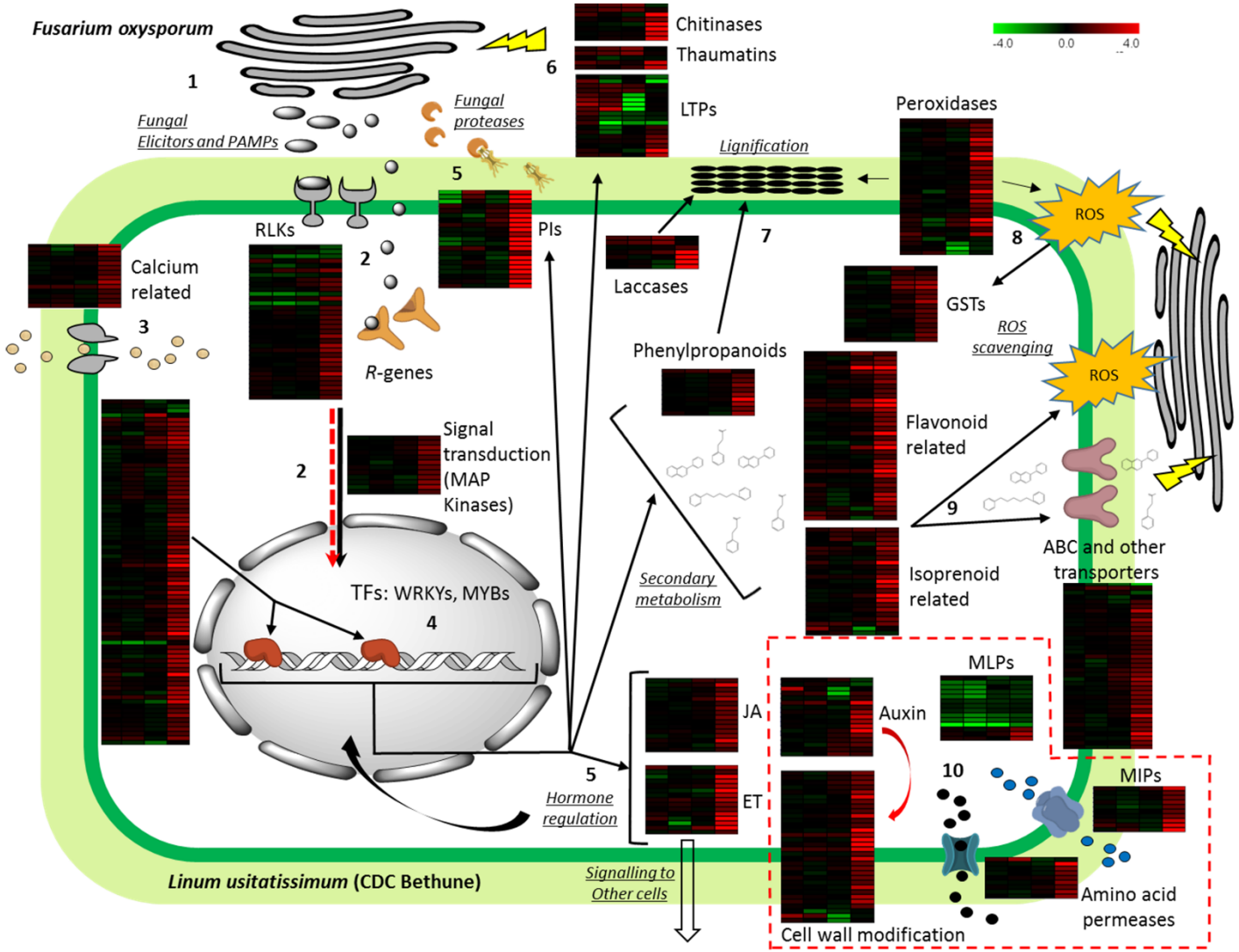

Fusarium oxysporum
1
Fungal Elicitors and PAMPs
Fungal proteases
6
Chitinases
Thaumatins
LTPs
-4.0
0.0
4.0
Peroxidases
Lignification
5
RLKs
PIs
2
Calcium related
3
R-genes
Laccases
7
ROS
8
GSTs
ROS scavenging
ROS
Phenylpropanoids
Flavonoid related
9
ABC and other transporters
Signal transduction (MAP Kinases)
2
TFs: WRKYs, MYBs
4
Secondary metabolism
Isoprenoid related
MLPs
JA
Auxin
10
MIPs
ET
5
Hormone regulation
Signalling to Other cells
Amino acid permeases
Cell wall modification
Linum usitatissimum (CDC Bethune)

Figure 4.11 Model depicting plant defense of flax upon *Foln* inoculation. Heatmaps for log2-fold gene expression changes at 2, 4, 8 and 18 DPI are shown besides each major gene group analyzed. The full deployment of plant defense is evidenced 18 DPI as seen in the forth column of the heatmaps. 1. During fungal attack, *Fusarium oxysporum* liberates elicitors (pathogen associated molecular patterns –PAMPs-), effectors and fungal proteases (which are also considered effectors) to facilitate infection. 2. Membrane receptors including receptor-like kinases (RLKs), an NBS-LRR (*R*-genes) interact with the PAMPs and effectors respectively causing downstream changes in phosphorylation of kinases (e.g. MAP kinases). 3. At the same time an influx of calcium causes changes in calcium-binding proteins that are also involved in signal transduction. 4. Activation of transcription factors results in activation of hormone-related, defense, and secondary metabolism genes. 5. Presence of jasmonate (JA), ethylene (ET) biosynthetic genes indicates further signalling to other cells and feedback loops to activate more defense genes. 6. Protease inhibitors (PIs) neutralize fungal proteases while chitinases, thaumatins and lipid transfer proteins (LTPs) act directly on the fungal cell wall or membrane. 7. Lignin precursors are created via phenylpropanoid metabolism and are polymerized into lignin by the action of laccases and peroxidases. 8. Peroxidases are also involved in the generation of reactive oxygen species (ROS) which are regulated by enzymes like glutathione S-transferases (GSTs). 9. Flavonoids and isoprenoids can act as antioxidants against ROS, or be directly translocated outside the cell by ABC transporters to impair fungal function and growth. 10. Some unexpected regulation was found in some specific transcripts of several gene groups: auxin-related genes, major latex proteins (MLPs), cell wall modification proteins, major intrinsic proteins (MIPs), and amino acid permeases; the potential manipulation of the host by the pathogen to regulate such genes is indicated by a red arrow that parallels signal transduction and by the gene groups surrounded by dashed red lines (see text for explanation).

Polymerization of monolignols generated via the phenylpropanoid metabolism occurs thanks to the presence of laccases and specific peroxidases (flax transcripts had similarity to *AtPrx52*, which is directly implicated in lignin formation) [410], while some of these peroxidases are also involved in the generation of ROS. The generation of these compounds is highlighted too by the overexpression of transcripts of *NADPH oxidase*. Because we were sampling a mixed population of cells undergoing HR and others getting the signals to deploy defenses but not undergoing cell death, we found indication of regulation of ROS by genes like GSTs and flavonoid biosynthetic genes, but it is likely that ROS is actively involved in damaging both host and pathogen cells at the points of pathogen invasion. Another alternative is that GSTs are hijacked by the pathogen to impair the HR by the plant. Nevertheless, the upregulation of GSTs coincides temporally with peroxidase action showing their importance in ROS regulation.

The other two large groups of regulated secondary metabolism transcripts were flavonoid and isoprenoid-related genes. The products of these metabolic pathways can act as antioxidants but also can be translocated (e.g. by ABC transporters) to directly change pathogen cell permeability and interact with membrane proteins, therefore impairing pathogen function [415]. Key enzymes in flavonoid, anthocyanin and carotenoid production included transcripts of chalcone synthase, dihydroflavonol reductase, anthocyanidin synthases and phytoene synthase. In fact, this latter group represents some of the transcripts with larger log2-fold changes in our study. Increased expression of some of these enzymes has demonstrated larger resistance against *F. culmorum* and *F. oxysporum* in flax [322], confirming these genes should be targets to identify natural variation among cultivars, or in gene modification efforts to increase their expression.

Lastly, several gene groups had members with unexpected regulation patterns, which we speculate could be caused by pathogen manipulation. For example, the presence of IAA amido/amino acid hydrolases, as part of auxin genes, could indicate a regulation of JA-conjugates but also results in active IAA which is used in growth. Furthermore, while some cell wall modification genes indicated reinforcement, others like expansins and glucanases weaken the cell wall and could provide an easier entry and nutrients for the pathogen. Under the same hypothesis, amino acid transporters would also provide increased nutrient input, while increased water exchange by aquaporins (MIPs) could facilitate haustorial development. Finally, MLPs were unexpectedly downregulated, and although a clear function upon pathogen response has not been established, these genes are usually upregulated in the plant-pathogen interaction.

4.6 Conclusions

This is the first transcriptome-wide study of the flax-fusarium interaction, and while confirmatory in many of the expected defense mechanisms of the plant, it also opens new possibilities for the exploration of specific genes. Several genes are candidate markers to explore the disease response across flax cultivars. Additionally, some of the top upregulated transcripts are still unannotated (Dataverse file: transcript differential expression in flax upon *Foln* infection http://dx.doi.org/10.7939/DVN/10933), which demands further investigation on their function. Both the candidate genes and the highest differentially expressed transcripts should be compared across cultivars to find variability that can be linked to resistance in the phenotypes. If gene

variability can be linked to phenotype, the resistance could be engineered back into susceptible cultivars having other desirable production characteristics using gene editing technology. Questions regarding which specific *Avr* genes interact with the plant cell components to supress immunity, and the cross-kingdom use of small RNAs for transcriptional control, should mark the avenues for new research.

CHAPTER 5 - Ion Torrent sequencing as a tool for mutation discovery in the flax (*Linum usitatissimum* L.) genome

This chapter is based on a published article: Galindo-González L.; Pinzon-Latorre D.; Bergen E.A.; Jensen D.C.; Deyholos M.K. 2015. Ion torrent sequencing as a tool for mutation discovery in the flax (*Linum usitatissimum* L.) genome. Plant Methods. (2015). 11:19.

5.1 Abstract

Detection of induced mutations is valuable for inferring gene function and for developing novel germplasm for crop improvement. Many reverse genetics approaches have been developed to identify mutations in genes of interest within a mutagenized population, including some approaches that rely on next-generation sequencing (e.g. exome capture, whole genome resequencing). As an alternative to these genome or exome-scale methods, we sought to develop a scalable and efficient method for detection of induced mutations that could be applied to a small number of target genes, using Ion Torrent™ technology. We developed this method in flax (*Linum usitatissimum*), to demonstrate its utility in a crop species.

We used an amplicon-based approach in which DNA samples from an ethyl methanesulfonate (EMS)-mutagenized population were pooled and used as template in PCR reactions to amplify a region of each gene of interest. Barcodes were incorporated during PCR, and the pooled amplicons were sequenced using an Ion Torrent™ PGM. A pilot experiment with known SNPs showed that they could be detected at a frequency of $> 0.3\%$ within the pools. We then selected eight genes for which we wanted to discover novel mutations, and applied our approach to screen 768 individuals from the EMS population, using either the Ion 314™ or Ion 316™ chips. Out of 29 potential mutations identified after processing the NGS reads, 16 mutations were confirmed using Sanger sequencing.

The methodology presented here demonstrates the utility of Ion Torrent™ technology in detecting mutation variants in specific genome regions for large populations of a species such as flax. The methodology could be scaled-up to test >100 genes using the higher capacity chips now available from Ion Torrent™.

5.2 Introduction

Flax (*Linum usitatissimum* L.) is cultivated as a source of either oil or fiber, both of which have distinct properties that make flax a valuable crop. The oil of flax seeds (i.e. linseed) is rich in polyunsaturated fatty acids including alpha-linolenic acid, which has purported health benefits and is also useful as a drying oil in manufacture of resins, finishes, and flooring. The stem phloem fibers (i.e. bast fibers) of flax are remarkably long and strong and are used for textiles and increasingly as substitutes for fiberglass in composite materials. The commercial potential of flax, as well as interesting aspects of its biology (including well-documented

phenotypic and genomic plasticity of some accessions [441]), have led to an increase in research activity in this species, highlighted by the release of an assembly of its whole genome sequence [15]. To accelerate the development of novel germplasm and to better exploit the available DNA sequence resources for flax, we sought to develop a mutant population and a reverse genetics platform for this crop.

Mutations can be induced by treating individuals with physical, biological or chemical mutagens [442]. Ethyl methanesulfonate (EMS) is widely used for inducing point mutations in plants [443–451], and results mostly in G/C to A/T transitions [443] that show a nearly random distribution throughout the genome. While one study showed that the frequency of EMS-induced mutations was estimated at about 1 mutation / 300kb screened [443], the density of mutations can vary for different plants and treatments [451]. Therefore, the frequency of SNVs for a given sequence length becomes an important factor in the probability of finding a phenotypic effect.

Two main approaches have been developed to relate genotype to phenotype in mutated populations. Forward genetics aims to evaluate the phenotype of hundreds or thousands of individuals to find abnormalities in characteristics like growth or development. Once a phenotypically abnormal individual is identified, map-based cloning or other molecular analyses must be used to identify the DNA sequence that was altered by mutation [452]. In reverse genetics, researchers start with a known DNA sequence of interest, and try to determine the effects of a mutation on the phenotype of the organism [442]. One advantage of reverse genetics is that it overcomes some of the limitations of forward genetics that are caused by functional redundancy [452]. In reverse genetics, mutations in a gene of interest can be obtained even in absence of a clear phenotypic effect, and therefore mutants of related genes can be combined to determine the impact of simultaneous loss-of-function or alteration of two or more genes.

Both forward and reverse genetics require researchers to screen a large number of individuals for the mutation of interest. Several methods have been developed to screen for mutations in a gene of interest within hundreds or thousands of individuals in parallel. TILLING (Targeting Induced Local Lesions in Genomes) was devised as one such methodology. In TILLING, the gene of interest is amplified by PCR of pools of DNA from members of an EMS-mutagenized population. Polymorphisms in the PCR amplicons are detected using denaturing high-performance liquid chromatography (DHPLC), or using a CEL I nuclease preparation, which cleaves the heteroduplexes that form between mutant and non-mutant DNA within the

amplicons from the pooled DNA; the activity of the nuclease is then detected by gel electrophoresis [453,454]. TILLING has been used in diverse species including *Arabidopsis* [443], rice [447], soybean [445], sorghum [448] and tomato [449]. Other alternatives to CEL I-based TILLING have also been described including high resolution DNA melting and conformation sensitive capillary electrophoresis [444,451,455]. However, with the advent of Next Generation Sequencing (NGS) technologies [456–458], the possibilities to improve the efficiency of reverse-genetic screening have increased. NGS provides direct information about the mutated sequence and does not require formation of heteroduplexes. While the cost of sequence is still too high to allow for whole-genome sequencing of every individual in a mutant population of a species such as flax with a genome size of 373Mb [15], the cost per-reaction may be reduced by incorporation of specific tags or barcodes in the primer sequences, allowing pooling of many samples in a single sequencing run by targeting specific regions of interest. An early approach using NGS to detect EMS mutations on tomato with GS FLX sequencing allowed screening of over 15000 plants [459]; GS FLX has also been used in the evaluation of Tef (*Eragrostis tef*) to examine genes related to lodging resistance [450]. Additional studies have used Illumina technology to perform TILLING by sequencing in rice and wheat [460], and in tobacco [461].

Here we present the first study using an Ion Torrent Personal Genome Machine ™ (PGM) to discover single nucleotide variants (SNVs) or rare variants (these two terms – along with "mutation" - are used interchangeably throughout the text) in an EMS mutant population of an elite linseed variety of flax. The Ion Torrent™ has one of the lowest instrument and per-run costs of the major NGS platforms [462,463], and its sequencing output is on a scale consistent with the expected requirements of this application. Ion Torrent™ works with chips bearing millions of microwells with transistor sensors that allow detection of changes in current produced by the release of hydrogen once new nucleotides are incorporated to the clonally amplified DNA strands attached to each one of the beads residing in each well [464]. Massive parallelism can be achieved with this technology and the sequencing capacity limit depends on the number of sensors in the array. During the development of the technology the increases in sequencing throughput have been achieved by growth of the chip size, closer packing of features (e.g. wells) and shrinking of features. In this way for example a 5.2 fold increase in sensor count has been achieved when moving from a 314 chip to a 316 chip [464].

We demonstrate the utility of this approach by identifying SNVs in eight genes of interest, after performing a pilot experiment in three genomic regions with known SNVs to validate the methodology. Our methodology allows identification of putative variants on the target genes and can be scaled up in the number of genes and individuals to screen large populations.

5.3 Materials and methods

5.3.1 Plant material

Seeds of *Linum usitatissimum* L. (var. CDC Bethune), an elite linseed cultivar, were obtained from Gordon Rowland (Crop Development Center, Saskatoon, SK). Seeds were soaked in 5 volumes (liquid volume/seed volume) of 0.5% ethyl methyl sulfonate (EMS) in 25 mM phosphate buffer (pH 7.6) for 4 h at room temperature and then were rinsed with distilled water (three times), and air dried prior to storage. These M_1 seeds were sown at the University of Alberta farm (Edmonton, AB), and their M_2 progeny were harvested as individual families. Approximately four seeds from each M_2 line were sown in rows at Kernen Farm (Saskatoon, SK) in summer 2010. Leaves were harvested and lyophilized for subsequent DNA extraction, and their progeny (i.e. M_3 families) were harvested from individual plants, then threshed and stored at ambient temperature in envelopes until screening.

5.3.2 DNA extraction and pooling

DNA extraction was performed for 96 samples at a time using CTAB [465] with some modifications. Lyophilized leaf samples (10-20 mg) were placed in 8 strip 1.2 mL collection tubes (QIAGEN, Hilden, Germany) containing a sterile 3 mm tungsten carbide beads. Tubes were capped and ground for 2 minutes at 25 Hz using a Retsch MM301 mixer mill (Retsch, Haan, Germany). Samples were centrifuged at 1,450 × g for 1 minute. CTAB extraction buffer was prepared with 2% CTAB (w/v), 2% PVP-40, 100 mM Tris-Cl pH 8.0, 25 mM EDTA pH 8.0, 1 M NaCl and 0.5 g L^{-1} of spermidine. The buffer was pre-warmed (60ºC) and supplemented with 10 µg mL^{-1} of RNAse A (Sigma-Aldrich, St. Louis, MO, USA), 100 µg mL^{-1} of proteinase K (Fermentas, Waltham, MA, USA) and 5% mercaptoethanol, before adding 500 µL to each sample. The tubes were re-capped with new caps and mixed by inversion (20 times) and incubated at 60ºC for 2 hours, mixing the tubes by inversion every 20 minutes. After incubation,

samples were centrifuged for 5 minutes at 5,800 × *g* and the supernatants were transferred to new tubes. Five hundred microliters of chloroform : isoamyl (24:1) were added to each sample and re-capped tubes were mixed by inversion (60 times) before centrifugation for 5 minutes at 5,800 × *g*. The supernatant was transferred to new tubes. Chloroform : isoamyl extraction was repeated once again. Three hundred microliters of ice-cold isopropanol were added to the tubes with the supernatant and the samples were mixed by inversion (20 times) before transferring to -20ºC for 2 hours. Incubation was followed by centrifugation at 5,800 × g for 15 minutes. Supernatants were decanted to waste and 500 µL of ice cold 95% ethanol were added to the pellets and samples were gently vortexed prior to centrifugation at 5,800 g for 5 minutes. The previous step was repeated with 500 µL of ice cold 70% ethanol. The ethanol was decanted and the samples were air-dried, and resuspended in 125 µL of TE 10:1 and stored at -20 ºC.

DNA samples were quantified using Picogreen (Invitrogen, Carlsbad, CA, USA) in a Fluorostar BMG plate reader (BMG labtech, Ortemberg, Germany), using a standard curve of flax DNA. Samples were diluted to 10 ng μL^{-1} and 1 µL of each pooled in groups of 64 or 96. According to the formula used by Tsai et al. [460] using 10 ng of DNA with a flax genome of 2C = 0.764 pg [15] would yield 204.5 copies per allele (for each individual) for the 1 in 64 dilution or 136.3 copies per allele for the 1 in 96 dilution. This is higher than the minimum number of recommended copies (40) to avoid absence or fluctuation of copies among individuals [460].

The pools were created using the following methodology: each sample was in an individual well of eight 96-well plates; all individuals from each plate were pooled creating the first eight pools containing 96 individuals each (designated pools A1 to A8). Then all the individuals from the same column in each plate were pooled creating 12 pools of 64 individuals each (designated pools B1 to B12). And finally all individuals from the same row in each plate were pooled creating the last eight pools of 96 individuals each (designated pools C1 to C8). In this way each individual was part of three different pools and therefore a mutation detected in three intersecting pools would allow us to pinpoint the source individual.

5.3.3 Primer design

The primers were designed using Primer 3 [237] with the following parameters: Two C's or G's in the last five nucleotides towards the 3` end; up to three nucleotides long homopolymers; a delta G lower than -9 kcal/mole, Tm between 59 and 61°C and size between 19

and 21 nucleotides (when conditions were not met parameters were relaxed). For the forward primer the universal primer tag 5'-CAGTCGGGCGTCATCA-3' was added (designed by Travis Glenn, Univ. of SC, http://www.gvsu.edu/dna/universal-primer-tag-6.htm) and for the primer Rv the adaptor trP1 was added (5'-CCTCTCTATGGGCAGTCGGTGAT–3') (**Table 5.1**).

Table 5.1 Primer sequences and adaptors used for pilot and test studies.

Gene region or ID	Annotation	Alias	Tag[b]+Fw (5'- 3')	trP1[c]+Rv (5'- 3')
5_scaffold20_536009	Genomic region	S20	ccggtgtcttcattgttgcgtctt	cagccaggttgcggaaagaacata
13_scaffold411_257117	Genomic region	S411	gttagagaaaggcaagaccaaccc	tagacggacgaacggaatcgtaga
23_scaffold900_176381	Genomic region	S900	aaagccgacctactgttcgtggta	ctcttctggagggcatcattgtca
Lus10004720	Pectinmethylesterase	LuPME10	gttacattcaatagcgaagag	atccaccgtgtgcaaccacgtc
Lus10031470	Pectinmethylesterase	LuPME79	tcccgatggcccaccagttcaac	gcgatgtacgaattcatcac
Lus10043035	Pectinmethylesterase	LuPME105	cgttcgtgggctgcagattc	gtgacacccctttcatttacg
G25305[a]	Pectinmethylesterase	LuPME73	gttgacgtttaggaacactg	gcagttctggaacacgacgg
G24175[a]	Cyclic peptide	CLE	accttgtctcctatttctgg	tcctgacaaccttccttg
Lus10029955	Acetolate synthase	ALS-2	atgatactgaacaaccagcat	acatcctctgaatcccctcc
Lus10016751	Acetolate synthase	ALS-1	gttcaagagctggcaactatt	gaattccacaaccttcagca
Lus10017825	UDP glucuronosyltransferase	UGT	accacagcggacttatatt	cgcatgactgagtgagaa

[a] Gene ID from first version of flax assembly.
[b] Tag added on 5' region: 5' cagtcgggcgtcatca 3'
[c] trP1 added on 5' region: 5' cctctctatgggcagtcggtgat 3'

5.3.4 Pilot experiment

A pilot experiment was performed using known SNVs showing polymorphisms between cultivars CDC Bethune and Macbeth. Since for the pilot experiment we were not trying to discover new mutations we did not have to pool DNA from different individuals as described in the DNA extraction and pooling section; instead Macbeth DNA was diluted 1:64 or 1:96 in Bethune DNA to simulate the presence of an individual with a mutation within the population. Three genomic regions with previously reported SNVs [466], from three flax scaffolds were chosen to test the methodology and were named S20, S411 and S900 (names were derived from the names of the scaffolds that contained them- see **Table 5.1**).

5.3.4.1 PCR amplification and barcoding

A two-step PCR strategy was adopted: the first step used a PCR with two sets of cycles at different temperatures (below) to amplify the target gene and the second step incorporated a bar code oligonucleotide to distinguish different pools (**Figure 5.1**). First step PCR was performed on each of 28 sample pools and three genes per pool (pilot experiment), with forward primers bearing a universal tag at their 5' end (**Table 5.1**), and reverse primers carrying an additional adaptor tail which binds the ionospheres used in the sequencing step.

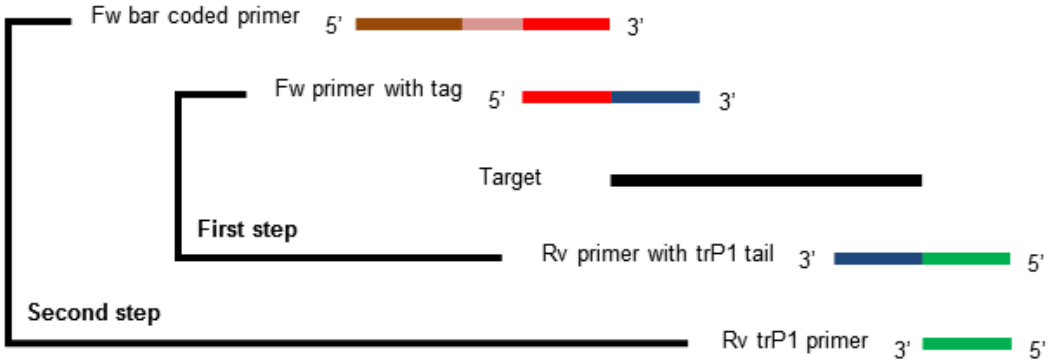


Figure 5.1 Two-step PCR strategy adopted for high throughput sequencing. On the first-step PCR the specific gene section (target) is amplified with a forward primer (blue) bearing a universal tag (red), and a reverse primer (blue) carrying a trP1 adapter (green). For the second-step PCR the amplicons of the first step are amplified with a reverse primer that matches the trP1 adapter (green) and a forward bar-coded (pink) primer (brown) for each desired pool of individuals and genes.

For the pilot experiment, first-step amplifications were performed on a template that either included DNA with known SNVs S20, S411 and S900 (Macbeth DNA in 1:64 or 1:96 dilutions in CDC Bethune DNA) or homogeneous template without the SNVs (i.e. only CDC Bethune DNA). The PCR products were diluted 1:100 and equal amounts (5 µL) of PCR product from each of the three target regions were mixed in each of 28 pools for barcoding; the changes of Macbeth genic regions SNVs where introduced in all pools for the gene of S20, and in 10 pools for genes S411 and S900. For the second step, a total of 28 PCRs were performed with forward bar-coded primers (**Figure 5.1** and **Appendix 5.1**), and a trP1 primer complementary to the tail from the reverse primer of the first-step PCR. The 28 bar codes allowed us to discriminate between the respective DNA pools after sequencing.

First step PCR was carried under the following conditions: 1X PCR buffer, 2 mM Mg, 0.2 mM dNTPs, 0.2 µM of each forward and reverse primers (**Table 5.1**), 1M ethylene glycol, 4% dimethyl sulfoxide (DMSO), 10 ng of DNA and 1.25 units of Platinum® Taq DNA polymerase (Invitrogen, Burlington, ON, Canada). The PCR protocol included an initial denaturation step at 94ºC for 3 minutes, followed by 5 cycles of 94ºC for 20 seconds, 50ºC for 30 seconds and 72ºC for 1 minute. Finally, an additional 25 cycles were done by changing the annealing temperature to 60ºC (touch-up), and a final extension step was performed at 72ºC for 10 minutes.

The second-step PCR was carried out as follows: 1X PCR buffer, 2 mM Mg, 0.2 mM dNTPs, 0.2 µM forward bar-coded primer and 0.2 µM reverse trP1 primer (**Table 5.1** and **Table 5.2**), 1 µL of the DNA amplicon dilution (three genes combined per pool) and 1.25 units of Platinum® Taq DNA polymerase (Invitrogen, Burlington, ON, Canada). The PCR protocol included an initial denaturation step at 94ºC followed by 30 cycles of 94ºC for 30 seconds, 62ºC for 30 seconds and 72ºC for 1 minute, and a final extension of 72ºC for 10 minutes.

Table 5.2 Primer sequences (with barcodes) used for second step-PCR.

primer	sequence
A_GLENN_IonXpress_1	ccatctcatccctgcgtgtctccgactcagctaaggtaaccagtcgggcgtcatca
A_GLENN_IonXpress_2	ccatctcatccctgcgtgtctccgactcagtaaggagaaccagtcgggcgtcatca
A_GLENN_IonXpress_3	ccatctcatccctgcgtgtctccgactcagaagaggattccagtcgggcgtcatca
A_GLENN_IonXpress_4	ccatctcatccctgcgtgtctccgactcagtaccaagatccagtcgggcgtcatca

A_GLENN_IonXpress_5	ccatctcatccctgcgtgtctccgactcagcagaaggaaccagtcgggcgtcatca
A_GLENN_IonXpress_6	ccatctcatccctgcgtgtctccgactcagctgcaagttccagtcgggcgtcatca
A_GLENN_IonXpress_7	ccatctcatccctgcgtgtctccgactcagttcgtgattccagtcgggcgtcatca
A_GLENN_IonXpress_8	ccatctcatccctgcgtgtctccgactcagttccgataaccagtcgggcgtcatca
A_GLENN_IonXpress_9	ccatctcatccctgcgtgtctccgactcagtgagcggaaccagtcgggcgtcatca
A_GLENN_IonXpress_10	ccatctcatccctgcgtgtctccgactcagctgaccgaaccagtcgggcgtcatca
A_GLENN_IonXpress_11	ccatctcatccctgcgtgtctccgactcagtcctcgaatccagtcgggcgtcatca
A_GLENN_IonXpress_12	ccatctcatccctgcgtgtctccgactcagtaggtggttccagtcgggcgtcatca
A_GLENN_IonXpress_13	ccatctcatccctgcgtgtctccgactcagtctaacggaccagtcgggcgtcatca
A_GLENN_IonXpress_14	ccatctcatccctgcgtgtctccgactcagttggagtgtccagtcgggcgtcatca
A_GLENN_IonXpress_15	ccatctcatccctgcgtgtctccgactcagtctagaggtccagtcgggcgtcatca
A_GLENN_IonXpress_16	ccatctcatccctgcgtgtctccgactcagtctggatgaccagtcgggcgtcatca
A_GLENN_IonXpress_17	ccatctcatccctgcgtgtctccgactcagtctattcgtccagtcgggcgtcatca
A_GLENN_IonXpress_18	ccatctcatccctgcgtgtctccgactcagaggcaattgccagtcgggcgtcatca
A_GLENN_IonXpress_19	ccatctcatccctgcgtgtctccgactcagttagtcggaccagtcgggcgtcatca
A_GLENN_IonXpress_20	ccatctcatccctgcgtgtctccgactcagcagatccatccagtcgggcgtcatca
A_GLENN_IonXpress_21	ccatctcatccctgcgtgtctccgactcagtcgcaattaccagtcgggcgtcatca
A_GLENN_IonXpress_22	ccatctcatccctgcgtgtctccgactcagttcgagacgccagtcgggcgtcatca
A_GLENN_IonXpress_23	ccatctcatccctgcgtgtctccgactcagtgccacgaaccagtcgggcgtcatca
A_GLENN_IonXpress_24	ccatctcatccctgcgtgtctccgactcagaacctcattccagtcgggcgtcatca
A_GLENN_IonXpress_25	ccatctcatccctgcgtgtctccgactcagcctgagataccagtcgggcgtcatca
A_GLENN_IonXpress_26	ccatctcatccctgcgtgtctccgactcagttacaacctccagtcgggcgtcatca
A_GLENN_IonXpress_27	ccatctcatccctgcgtgtctccgactcagaaccatccgccagtcgggcgtcatca
A_GLENN_IonXpress_28	ccatctcatccctgcgtgtctccgactcagatccggaatccagtcgggcgtcatca

5.3.4.2 Purification and quantification of PCR products

The products of PCR reactions of 28 pools and 3 genes per pool were run on 1.5% agarose gels. To eliminate primer dimers, a band of the expected size was gel purified using the Wizard®SV gel and PCR clean-up system (Promega, Madison, WI, U.S.A). Samples were quantified using a Qubit® 2.0 fluorometer using a dsDNA HS assay (Invitrogen, Burlington, ON, Canada). All quantified samples were diluted to 1 ng μL^{-1} and equal amount of the PCR products

were combined and were re-measured on the Qubit to confirm that concentration was still 1 ng μL^{-1}. The sample was diluted with low Tris EDTA buffer (TE 10:1) to obtain $15.5x10^6$ molecules per microliter (26 pM) which is the recommended concentration for template preparation using Ion Torrent™ technology.

5.3.4.3 Sequencing

All procedures for emulsion PCR and next-generation sequencing were performed with Ion Torrent™ equipment and Ion Torrent™ kits under the manufacturer's specifications (Life Technologies, Carlsbad, CA, U.S.A.): emulsion PCR was performed with the Ion OneTouch™ 200 template kit in an Ion OneTouch™. Enrichment of template positive Ionospheres (ISPs) was performed with an Ion OneTouch™ ES (Life Technologies, Carlsbad, CA, U.S.A.). Sequencing of enriched templates bound to ionospheres was done using the Ion PGM™ 200 sequencing kit in an Ion PGM™ Sequencer with either 314 or 316 chips. FASTQ files of each barcoded group of sequences were recovered from the Ion Torrent™ server for further analysis. Reads have been deposited in the Sequence Read Archive (SRA) from NCBI under study accession number: SRP052626.

5.3.5 Detection of induced mutations in a population of EMS mutagenized flax

Procedures were similar to the pilot experiment, unless stated otherwise. The experimental design was adapted from Tsai et al., 2011 [460]. A total of 28 pools of DNA from distinct individuals (768 lyophilized leaf samples) were created to facilitate detection of mutations as described in DNA extraction and pooling.

A total of eight primer pairs that amplified pectin methylesterases (PMEs) were designed to target conserved regions presumed to be essential for enzymatic function and tertiary structure stability of these genes [467,468]. Preliminary tests showed that four of the primers pairs (**Table 5.1**), gave stronger products, and these were used for second-step PCR as described above. Separately, 12 primer pairs from three different metabolism-related genes that constitute important breeding traits (cyclic peptides, acetolactate synthase and UDP - glucuronosyl/glucosyl transferases) were also designed and four primer pairs were selected after testing them by PCR (**Table 5.1**).

5.3.6 Analysis of Single Nucleotide Variants (SNVs)

Reads obtained from Ion Torrent PGM™ sequencing were uploaded to the CLC Genomics workbench platform (CLCbio, Aarhus N, Denmark). Reads were mapped to reference sequences previously confirmed by Sanger capillary sequencing of target amplicons (data not shown), using the following parameters: masking mode = no masking, mismatch cost = 2, insertion cost = 3, deletion cost = 3, length fraction = 0.8 and similarity fraction = 0.8, global alignment = no, non-specific match handling = map randomly, output mode = create stand-alone read mapping, create report = yes, collect unmapped reads = yes. Once the reads were mapped to the reference, the mapped reads files were used as input to discover rare variants using quality score with the following parameters: neighborhood radius = 5, maximum gap and mismatch count = 5, minimum neighborhood quality = 15, minimum central quality = 20, ignore non-specific matches = yes, ignore broken pairs = yes, minimum coverage = 100, minimum variant frequency (%) = 0.1 (selected according to a previous study [464]), maximum expected alleles = 4, advanced = no, require presence in both forward and reverse strands = no, filter454/ion homopolymer indels = yes, create track = yes, create annotated table = yes, genetic code = 1 standard. The whole process was automated by creating a CLC workbench workflow.

Tables were created by calculating mutation frequency per gene and per pool after filtering homopolymeric tracts and indel artifacts created by the sequencing technology. Graphs showing the frequency changes by position for the 28 pools in a specific base change (e.g. G to A) were made to visually identify outliers, which were indicative of a rare variant. SNV candidates were chosen if the mutation was present in three intersecting pools or in two intersecting pools with high frequency (>0.3% in pilot experiment and >0.5% in remaining experiments). DNAs from individuals (M_3 generation) with potential mutations were re-sequenced using Sanger sequencing to confirm the analysis performed with Ion Torrent™.

5.4 Results

We conducted three experiments to develop an Ion Torrent™-based method for discovery of single nucleotide variants (SNVs) in flax: (i) a pilot experiment with combinations of known SNVs (using an Ion 314™ chip); (ii) a proof of concept experiment with a mutagenized population of flax (also using an Ion 314™ chip); and (iii) a scale-up experiment using the higher capacity, Ion 316™ chip.

5.4.1 Experiment I: Pilot

To evaluate our ability to detect known variants in selected regions of DNA, we used DNA from two non-mutagenized cultivars of linseed flax: CDC Bethune and Macbeth [469]. We designed primers (**Table 5.1**) encompassing SNVs that had been previously identified in a comparison of CDC Bethune and Macbeth DNA sequences [466] and designated these regions as S20, S411 and S900 using their scaffold of origin (e.g. S20 = scaffold 20 of the published genome assembly [15]). We mixed DNA from CDC Bethune with DNA from Macbeth to simulate a total of 28 pools from either 64 or 96 individuals, in which one individual in the pool was polymorphic (i.e. carried a SNV not present in any other member of the pool). As a negative control, we also constructed simulated pools that consisted of only DNA from CDC Bethune.

We amplified the three target regions using a two-step PCR (**Figure 5.1** and **Appendix 5.1A-B**). The two-step PCR was used because it allowed us to incorporate specific barcodes for each pool (**Table 5.2**). After the second PCR step, we gel-purified the amplification products to eliminate primer dimers, which could otherwise be preferentially amplified during subsequent emulsion PCR. Gel-purified DNA was diluted to 1 ng μL^{-1}, and pooled before diluting all mixed products to 26 pM. This pooled sample was diluted one time to obtain a second pool of half the concentration (13 pM), which was used to perform a second emulsion PCR. We measured the percent of templated Ion Sphere^{TM} particles (ISPs), as 37.3% for the 26 pM sample and 27.6% for the 13 pM sample. The latter sample was selected for sequencing since the template ISPs fell in the acceptable range of 10 to 30% [470]. The loading of the 26 pM sample was deemed too high for sequencing.

A total of 119.38 Mbp of sequence were obtained which represented 678,532 library reads after filtering for polyclonal, dimer and low quality sequences (**Table 5.3**). While the modal read length for tested genes was > 200 bp, the mean read size was 176 bp due to a large number of reads in the 50 bp range. These short reads were comprised of incomplete sequence extensions, and sequence artifacts that were filtered out during the subsequent mapping step, leaving 47.35% of all reads to be mapped to the 28 pools (**Table 5.3** and **Table 5.4**), in each one of the three genomic regions evaluated (**Appendix 5.2**). Average read coverage was 4,103, 2,392 and 4,794 for sequences S20, S411 and S900 respectively (**Table 5.4**). While coverage did not seem to vary along the sequence, the coverage between pools did vary (**Figure 5.2**).

Table 5.3 Read statistics of the three experiments performed.

Experiment / replicate	Chip type	Percentage of wells with beads in chip	Total number of bases (Mbp)[a]	Total number of reads[a]	Percentage of mapped reads to all genes in experiment	Mean read length (bp)
Pilot	314	74%	119.38	678,532	47.35	176
Proof of concept-1	314	79%	71.80	459,888	60.31	156
Proof of concept-2	314	76%	85.40	543,659	63.02	157
Scale up	316	74%	649.00	3,403,220	92.04	190

[a] After filtering polyclonal wells, test fragments, adapter dimers sequences and low quality reads.

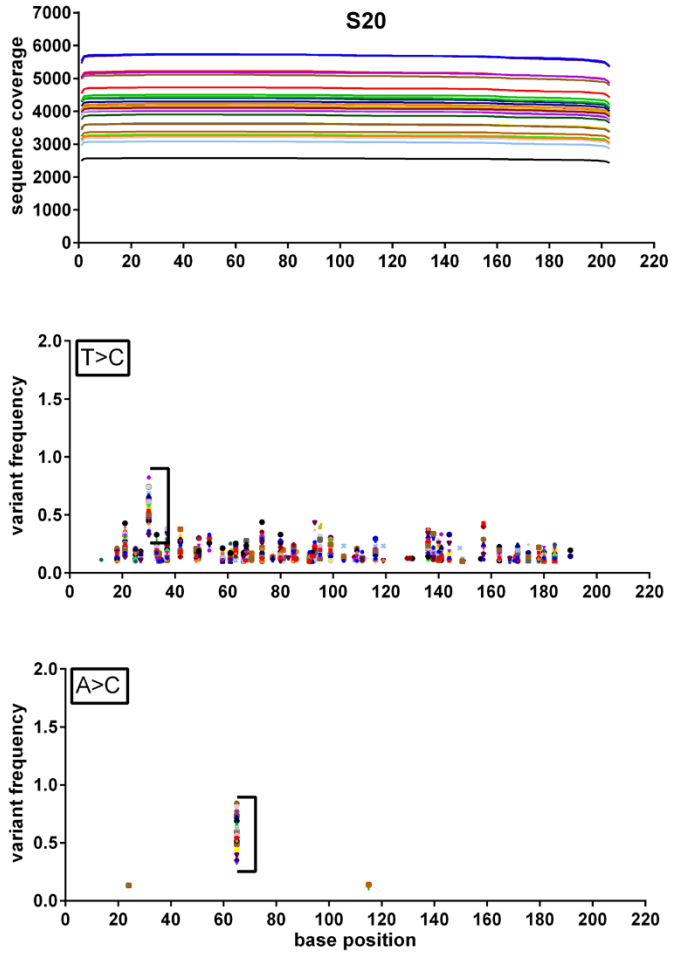
Table 5.4 Read statistics of the three experiments performed.

Experiment / replicate	Scaffold ID / gene ID	Gene name / annotation[b]	Number of reads mapped in all 28 pools	Percentage of mapped reads in all 28 pools[c]	Average read coverage per pool
Pilot	S20	N/A	115,998	17.09	4,103.69
	S411	N/A	67,971	10.02	2,392.08
	S900	N/A	137,314	20.23	4,794.04
Proof of concept-1	Lus10004720	*LuPME10*	91,755	19.95	2,975.50
	G25305[a]	*LuPME73*	94,242	20.49	2,942.75
	Lus10031470	*LuPME79*	38,154	8.30	1,059.81
	Lus10043035	*LuPME105*	53,216	11.57	1,621.20

Experiment / replicate	Scaffold ID / gene ID	Gene name / annotation[b]	Number of reads mapped in all 28 pools	Percentage of mapped reads in all 28 pools[c]	Average read coverage per pool
Proof of concept-2	Lus10004720	*LuPME10*	87,889	16.17	2,815.02
	G25305*	*LuPME73*	143,832	26.46	4,404.28
	Lus10031470	*LuPME79*	47,620	8.76	1,255.62
	Lus10043035	*LuPME105*	63,282	11.64	1,941.29
Scale up	Lus10016751	ALS-1	1,443,997	42.43	44,199.34
	Lus10029955	ALS-2	424,425	12.47	13,027.62
	G24175[a]	CLE	534,255	15.70	16,146.73
	Lus10017825	UGT	729,799	21.44	20,472.63

[a] Gene Id correspond to first draft assembly of flax (unpublished).
[b] PME = Pectinmethylesterase, ALS = acetolactate synthase, CLE = cyclic peptide, UGT = glucoronosyl/glucosyl transferase.
[c] Percentage from total number of reads in **Table 5.3**.

A


S20
sequence coverage
7000
6000
5000
4000
3000
2000
1000
0
0 20 40 60 80 100 120 140 160 180 200 220
T>C
variant frequency
2.0
1.5
1.0
0.5
0.0
A>C
base position

B

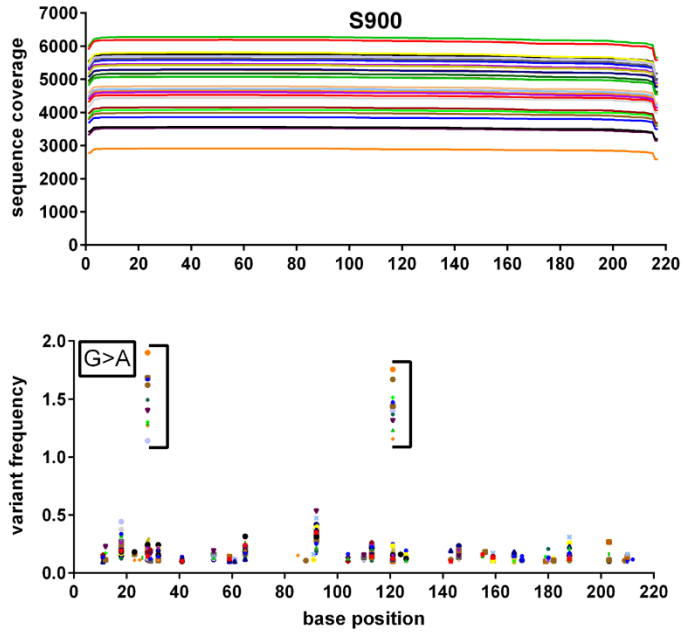
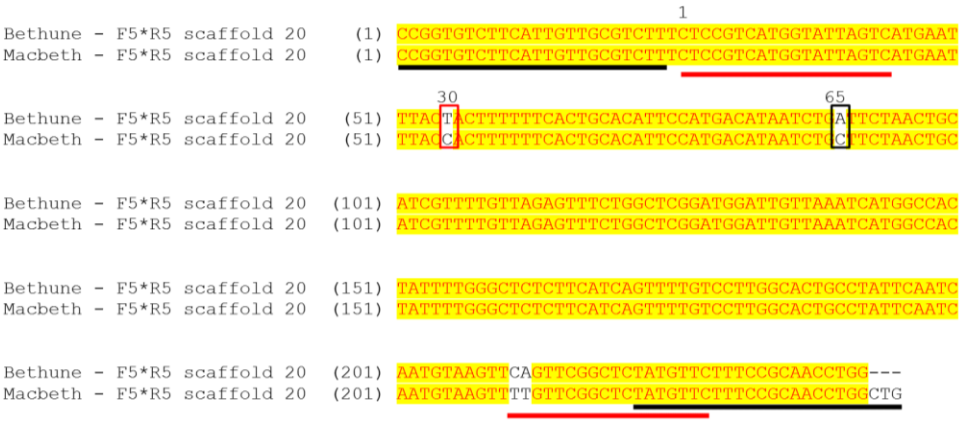

S900
sequence coverage
7000
6000
5000
4000
3000
2000
1000
0
0 20 40 60 80 100 120 140 160 180 200 220
G>A
variant frequency
2.0
1.5
1.0
0.5
0.0
base position

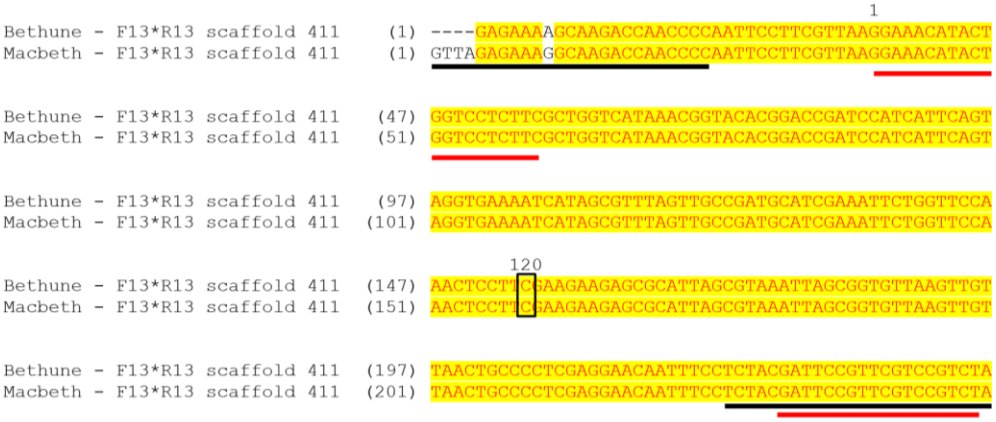
Figure 5.2 Sequence coverage and frequency of variants in gene sections of the pilot experiment. DNA from the cultivar Macbeth was diluted (1:64 or 1:96) in CDC Bethune DNA in several simulated pools as described in methods. Each line in the sequence coverage graphs represents one of 28 pools. The frequencies of the variants are plotted against the position in the respective reference sequence section. Each symbol in the frequency graphs represents a frequency of at least 0.1% for each of the 28 pools in each position. No graphs for S411 are shown since no variants were detected in that sequence.

Analysis of mapped reads from the simulated pools identified expected SNVs in two of the three targeted loci. For S20, a SNV was identified in base 65 (A->C) (**Figure 5.2**). For S900, a SNV was identified in base 28 (G>A), but the expected SNV at position 120 (C>T) of S411 that was previously reported was not found in any of the pools. However, novel SNVs (i.e. polymorphisms between Macbeth and CDC Bethune that had not been previously reported) were found in two of the targeted loci: we discovered an additional SNV at position 30 (T>C) in S20, and an additional SNV at position 121 (G>A) in S900 (**Figure 5.2**). All of these observations were confirmed by Sanger sequencing of targeted loci (**Figure 5.3**).

Alignment F5*R5 – scaffold 20



Alignment F13*R13 – scaffold 411



Alignment F23*R25 – scaffold 900

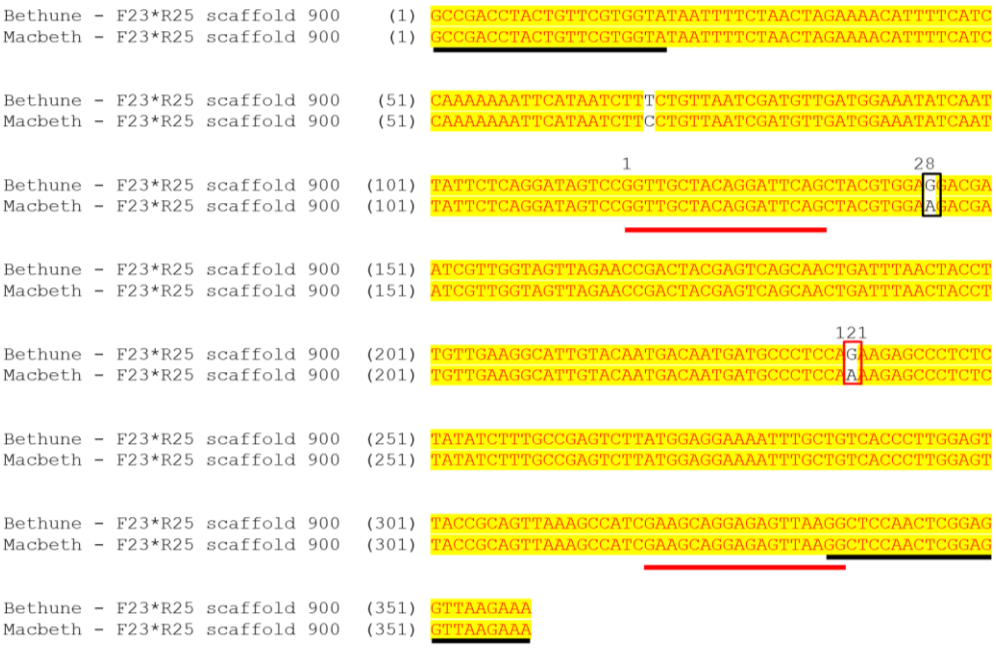


Figure 5.3 Alignment of sequenced fragments of the pilot experiment. The number 1 over the alignment indicates position 1 for reference of mutations found. Primers of the amplicons used for Ion torrent: black bar, primers used for sequencing: red bar, expected mutations: black outline, unreferenced mutations: red outline.

5.4.2 Discovery of EMS-induced mutations in PME genes

Having demonstrated in the pilot experiment that we could detect known SNVs within simulated pools of DNA, we next attempted to discover novel SNVs within pools of DNA obtained from an actual mutagenized population (proof of concept). We used 10 ng of DNA from each of 768 individuals and pooled the DNA as explained in materials and methods. Because of the way our experiment was designed, each one of the 768 individual DNA samples was present in three pools; when a SNV is found in three intersecting pools we could pinpoint the sample of origin. We targeted four genes of the pectin methylesterase (PME) family for discovery of SNVs (*LuPME10*, *LuPME73*, *LuPME79*, *LuPME105*, Table 2). These genes were selected because they are relevant to ongoing cell wall research in our laboratory [467]. To minimize the amplification of primer dimers, we tested the PME primers (**Table 5.1**), under a range of annealing temperatures and found that the optimal temperature range for the touch-up first-step PCR was 56-66°C (this was higher than the annealing temperature range 50-60°C in the pilot experiment), and the optimal second-step PCR annealing temperature was 68°C. This highlighted the importance of empirically testing PCR conditions for any new set of primers. Amplicons were analyzed and purified on agarose electrophoretic gels, eluted, and quantified (as in the pilot experiment) before Ion Torrent™ sequencing.

For sequencing, we diluted the pooled PME amplicon DNA to 13 pM. This DNA was sequenced in two replicate runs (to test for consistency). The percent of template ISPs for the two replicates was 23.87% and 20.13%, which made both samples suitable for sequencing. A total of 71.8 Mbp and 85.4 Mbp were obtained with an average read length of 156 bp and 157 bp for the two technical replicates (**Table 5.3**). The total number of usable reads after performing filtering of polyclonal, low quality and primer dimers were 459,888 and 543,659. However, we found that even after read filtering, there remained a fraction of short reads in the 50bp range, which presumably represented primer dimers and incomplete products (**Figure 5.4**).

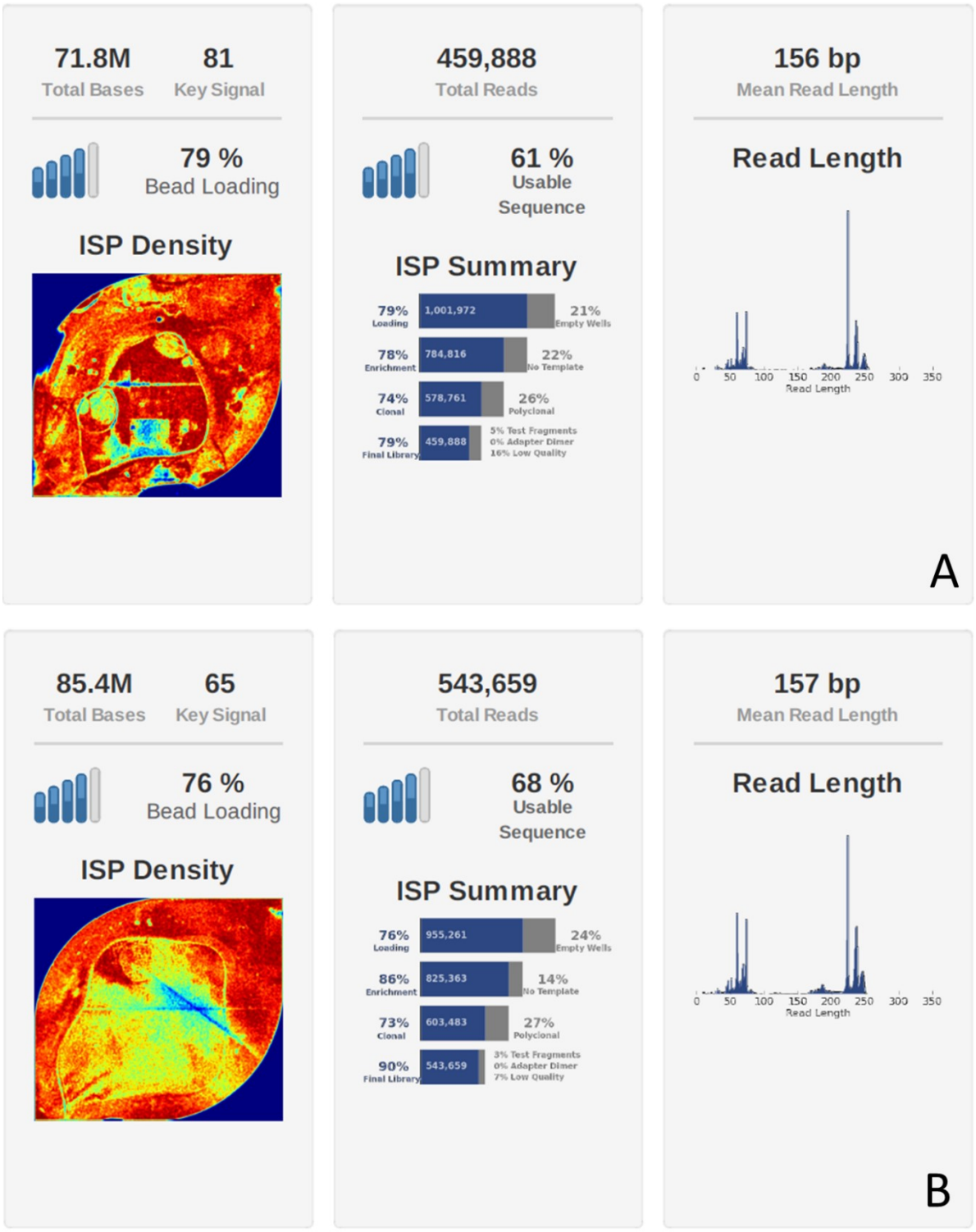


Figure 5.4 Ion Sphere Particles (ISPs) and read identification summary. The data is given for two technical replicates runs (A and B) from the proof of concept experiment using PMEs.

When comparing to the pilot experiment the percentage of mappable reads increased to over 60% for both replicates (**Table 5.3**), but the average read coverage per pool in each of the evaluated genes was proportionally lower than in the pilot since reads in this case were distributed among four genes (**Table 5.4**). Furthermore, the proportion of reads mapped was not equally distributed among the four genes in any of the pools. When using the mapped reads in all pools to calculate the coefficient of variation (CV) for each gene and replicate, the number ranged from 24.9 to 57.5 (**Table 5.5**), however the variation was constant among the two replicates for each gene.

Table 5.5 Average read count and dispersion among the mapped reads for two replicate sequencing runs of PME genes. Statistics are given for the 28 pools for all mapped reads or for each individual gene.

Reads	Replicate	Average	Mean deviation	Standard deviation	Coefficient of variation
Mapped reads	1	9906.0	1939.3	2461.8	24.9
	2	12236.5	2499.4	3175.1	25.9
Lus10031470 (*LuPME79*)	1	1362.6	464.2	693.8	50.9
	2	1700.7	485.2	715.0	42.0
Lus10004720 (*LuPME10*)	1	3277.0	1199.8	1534.1	46.8
	2	3138.9	1147.7	1478.6	47.1
G25305 (*LuPME73*)	1	3365.8	887.6	1279.2	38.0
	2	5136.9	1387.4	1977.0	38.5
Lus10043035 (*LuPME105*)	1	1900.6	820.7	1092.9	57.5
	2	2260.1	907.0	1227.5	54.3

The coverage per position for each gene was high throughout the sequence, with the exception of *LuPME79*, where a drastic decrease in coverage was observed after position 162 of the reads (**Figure 5.5**). Analysis of the sequence with Mfold [471] (not shown), did not predict a secondary structure that would explain this apparent hard stop in sequencing. Additionally, GC content of *LuPME79* (57.14%) was similar to *LuPME73* (57.34%), so a bias in GC content could not explain this difference either.

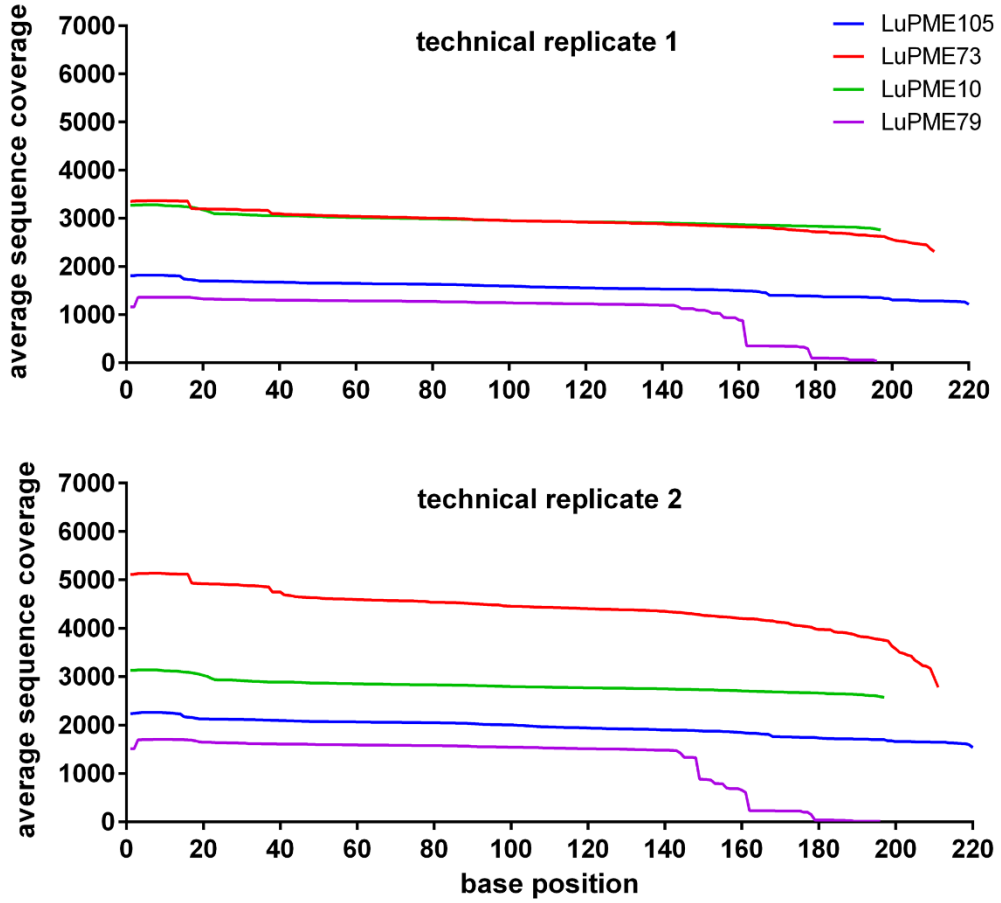


Figure 5.5 Coverage of four PME genes in two technical replicates. The average sequence coverage from 28 pools in each one of the base positions for the four PME genes is shown.

Based on our experience in the pilot experiment, we selected a minimum coverage per position of 500x, with a frequency of at least 0.5% in three intersecting pools, for defining putative mutations. When only two intersecting pools were found with the expected minimum frequency, all individuals from the intersection where sequenced. There was consistency between replicate runs for most SNVs but some of the SNVs were detected by complementary intersecting pools between both replicates. There was no correlation of false positives with the technically consistent SNVs or the ones found by complementarity.373

In our analysis of four targeted PMEs amplified from 768 individuals, we found a total of 13 putative SNVs. Sanger sequencing on the original DNA from the pooled individuals confirmed only five of the 13 putative SNVs (**Table 5.6**). When the sequenced sections from the original and mutated individuals were translated, it was found that neither of the two non-

synonymous changes found was within the predicted enzyme active sites [468] (Figure 4). Nevertheless, the methodology proved useful for finding mutations in pooled mutated populations when testing several genes at the same time.

Table 5.6 SNVs found in four PME genes.

Gene	Base No.	Change	Sanger confirmation	Nucleotide substitution	Amino acid substitution
LuPME79	33	G>A	No	N/A	N/A
LuPME79	96	G>A	Yes	Heterozygous	Non-synonymous
LuPME73	25	G>A	No	N/A	N/A
LuPME73	54	G>A	Yes	Heterozygous	Non-synonymous
LuPME73	81	T>A[a]	No	N/A	N/A
LuPME73	88	C>T	Yes	Heterozygous	Synonymous
LuPME73	97	A>G[a]	No	N/A	N/A
LuPME73	139	C>T	Yes	Homozygous	Synonymous
LuPME73	189	C>T	No	N/A	N/A
LuPME10	154	C>T	Yes	Homozygous	Synonymous
LuPME105	34	A>G[a]	N/A	N/A	N/A
LuPME105	57	A>G[a]	No	N/A	N/A
LuPME105	115	G>A	No	N/A	N/A

[a] Mutation not expected from EMS, but discovered using this methodology.

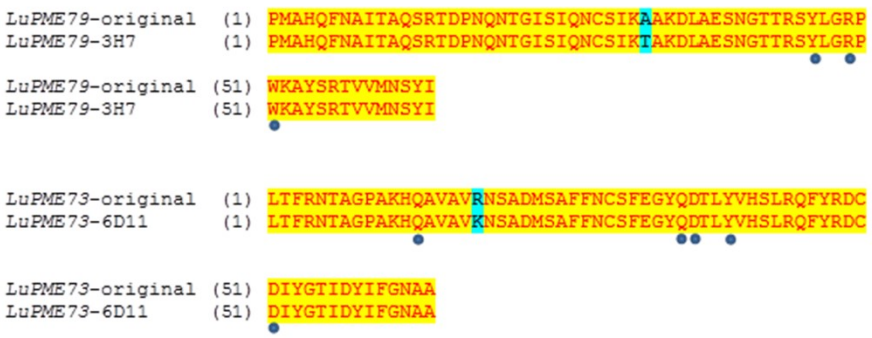


Figure 5.6 Alignment of amino acid sections from individuals bearing non-synonymous mutations (Table 5.6) to the original non-mutated sequences. Gene IDs are followed by an identifier given to the sequenced individuals. Circles below the alignment indicate enzyme active sites. Blue background indicates the amino acid change.

5.4.3 Increased read depth for discovery of EMS-induced mutations

Because the previous experiment showed a large variation in mapped reads between genes and read depth among pools in each gene (**Table 5.4** and **Table 5.5**) and less than half of the predicted SNVs could be confirmed by Sanger sequencing, we decided to increase read depth by switching from Ion 314™ chips to the higher capacity Ion 316™ chips (scale up). We used four genes related to flax metabolism (**Table 5.1** and **Table 5.4**). These genes are related to characteristics related to bitter taste in flax (cyclic peptides), targeting of group 2 herbicides (acetolactate syntases), or important as major components of cell wall formation (glucoronosyl/glucosyl transferases). We selected regions in these genes based on previous studies showing critical sections and/or amino acids for the function of these proteins [472–475]

We tested again two dilutions at 13 and 26 pM to assess which of these would give a better percentage of template ISPs. We obtained 12.15 and 17.35% of templated ISPs respectively and sequenced only the latter sample, which had the highest percent loading. A total of 649 Mbp were obtained with an average read length of 190bp (**Table 5.3**). The total number of usable reads after filtering was 3,403,220 which was an approximately 5-fold increase from the 314 chips used in the first two experiments. Although the coverage of pools among genes fell slightly towards the end of the sequences (result not shown), the average coverage for the four evaluated genes was 10 times higher than in the previous experiment (**Table 5.4**), and therefore the depth was sufficient to assess variants in any position throughout pools and genes.

Using similar parameters as for the PMEs, we found a total of 16 putative SNVs from which 11 were confirmed by Sanger sequencing (Table 4). From these, two were found to be homozygous. One of the heterozygous mutations resulted in the generation of a stop codon. We tested the heritability of the SNVs discovered in ALS1, ALS2, UGT and CLE by Sanger sequencing of the progeny of plants in which the mutations were initially identified. The presence of the mutation was confirmed in the progeny of all of the lines (Table 4).

Table 5.7 SNVs found in four genes of interest.

Gene	Base No.	Change	Sanger confirmation	Nucleotide substitution	Amino acid substitution	Sanger confirmation on M_4	Nucleotide substitution in progeny[b]
ALS-1	119	C>T	Yes	Heterozygous	A/V	Yes	2 homozygous, 1 non-mutant
ALS-1	140	C>T	No	N/A	P/L	N/A	N/A
CLE	89	G>A	Yes	Heterozygous	G/D	Yes	3 heterozygous
CLE	94	G>A	Yes	Heterozygous	E/K	Yes	1 homozygous, 1 heterozygous
CLE	134	G>A	Yes	Heterozygous	R/H	Yes	1 homozygous
ALS-2	26	G>A	No	N/A	G/E	N/A	N/A
ALS-2	43	G>A	Yes	Heterozygous	E/K	Yes	2 homozygous
ALS-2	100	G>A	No	N/A	E/K	N/A	N/A
ALS-2	161	C>T	Yes	Heterozygous	A/V	Yes	2 homozygous
ALS-2	161	C>T	Yes	Homozygous	A/V	Yes	2 homozygous
UGT	27	C>T	No	N/A	P/S	N/A	N/A

UGT	33	C>T	No	N/A	H/Y	N/A	N/A
UGT	81	C>T	Yes	Homozygous	L/F	Yes	3 homozygous
UGT	99	G>A[a]	Yes	Heterozygous	E/STOP	Yes	3 heterozygous
UGT	99	G>A[a]	Yes	Heterozygous	E/STOP	Yes	2 heterozygous
UGT	184	G>A	Yes	Heterozygous	G/E	Yes	1 homozygous, 1heterozygous

[a] Mutation was found by looking at intersecting pools with frequencies below the set threshold.
[b] Six individuals from progeny examined per mutation.

5.5 Discussion

5.5.1 Ion Torrent™ technology in SNV detection

The advent of next-generation sequencing technologies has opened new doors for genomic-scale analyses [457,458]. Among common sequencing platforms, IonTorrent™ offers potential advantages including low instrument cost, low cost per base, and fast output (up to 333 Mbp/h) [462,476]. Ion Torrent™ has also been reported to be superior for variant calling than Illumina™ [462], although other studies report similar or slightly higher sensitivity for MiSeq™ [477]. While Ion Torrent™ has a high rate of indels caused by homopolymeric runs, and long-range sequence quality can be lower than that of other instruments [462,476,478,479], this is not a problem for calling SNVs with high read depth.

We used Ion 314™ and Ion 316™ chips and were able to reach reads in the 200bp range with a total sequence throughput that guaranteed high depth. The size reached by these reads facilitated the evaluation of critical gene regions without the need for post-sequencing assembly. Likewise, we were able to achieve a high-throughput in the number of samples analyzed for several genes. A similar approach used Ion Torrent™ to map mutations in mice, but examined a large number of regions in a few samples [480].

In the pilot experiment, we detected the expected mutations and additional unreported changes in the tested genomic regions (**Figure 5.2** and **Figure 5.3**). Our experiment with PMEs had a larger variability in read depth among pools and genes (**Table 5.5**), which may have had an influence on the number of false positives. When we increased our read depth by using the larger Ion 316™ chip, the number of false positives decreased significantly.

Several variables were optimized during our experiment. Since sample pooling was used, high-quality DNA in equal amounts was needed to increase the probability of detecting one mutated individual among the population. We standardized a high-throughput CTAB protocol yielding high quality DNA and further quantified the samples by fluorescence to add equal amounts of DNA from each individual in each pool. Additionally, the first step PCR required addition of different PCR additives (ethylene glycol and DMSO), which decreased the formation of secondary structures since preliminary tests showed that a standard PCR resulted in a high proportion of primer dimers (or secondary structures) due to the length of the primers; nevertheless, some residual dimers were still unavoidable (**Appendix 5.1**). A PCR cleanup did not suffice to get rid of such dimers, which were still carried over to the emulsion PCR, resulting

in a preferential amplification of these smaller products in preliminary runs (result not shown). Eluting the specific products from agarose gels improved the detection of the larger specific products but still with some residual carryover. A better primer removal method like a solid-phase reversible immobilization (SPRI) technology was suggested for the process of PCR cleanup for Illumina [481], and could also be used in future experiments.

Differences in read coverage were detected in our different experiments, but this is not uncommon in NGS technologies [460]. As a general trend we found that shorter amplicons resulted in higher average coverage over all positions in all pools for the evaluated gene sections of the three experiments. This can be related especially to emulsion PCR, since this constitutes a step where all genes are mixed and there can be a preference for preferential amplification of shorter amplicons. Factors like shorter denaturation times and faster extension on smaller products may lead to this preferential amplification [482].

However, the exact same relationship was not found for the PMEs (**Figure 5.5**). The two PME amplicons corresponding to *LuPME79* and *LuPME105* had a lower average coverage, and while the latter does correspond to the larger amplicon of this gene set, the former is the smallest. Therefore, different factors may have had an influence on the variability in read count that we encountered. When we calculated GC content it was seen that *LuPME105* had the lowest GC content (43%) of the four PME sections evaluated. Ion Torrent^{TM} read coverage has been shown to decrease upon high or low GC content or under different levels of genome complexity [462,464,483]. Likewise, gel extraction has been shown also to have a bias for recovery of GC-rich double stranded templates that have higher affinity for kit columns, than AT-rich amplicons which become single stranded upon agarose melting conditions [481].

Neither length, nor GC, seemed correlated with the lower coverage of *LuPME79*, however, the primers from the two low coverage gene regions (**Table 5.1**) had a lower value of Gibbs energy – ΔG - (-9.28 and -10.24 kcal mol^{-1} respectively) compared with the primers of *LuPME10* and G25305 (-3.61 and -6.3 kcal mol^{-1}). Since a lower ΔG favors the formation of secondary structures, this could have had an effect of the PCR resulting in a differential amount of amplicons before pooling. Unforeseen changes like EMS mutations in priming sites can also contribute to differential amplification among samples. We also encountered a drop in coverage after position 162 in *LuPME79*; while we could not detect any evident secondary structure after the hard stop in the sequence reads, 16 out of the 20 previous nucleotides before the read

coverage fall are G or C and this could be related to the formation of a secondary structure that impairs the sequencing polymerase from continuing.

Length and GC content of the target locus are not entirely under control of the researcher. However, other factors can be better controlled to achieve near-homogeneous coverage when pooling samples for analysis. For example, an accurate quantification of the PCR products after the first round of PCR by comparison to a standard [484], would decrease biases in amounts of amplicons before pooling. Although previous studies have shown that non-normalized samples are suitable to detect high-frequency variants [484], our study comprised the detection of mutations in pooled samples, where a homozygous mutation could theoretically have a low frequency: aprox. 1% for a 1 in 96 dilution, and of 0.5% for an heterozygous allele. Additionally, very small amounts of DNA (10 ng) were used in the pools. Since differences in the amount of starting DNA can also result in differential amplification [482], it is important to guarantee close to equimolar amounts of starting DNA to avoid losing a variant due to a PCR deficiency.

Overall, the coverage on the 11 different gene regions that were tested allowed for the detection of SNVs in regions of up to 200bp. Although coverage varied slightly between some pools and genes, coverage along the length of amplicons was generally even but dropped only towards the ends of the sequences (**Figure 5.5**), which is common of sequencing-by-synthesis technologies [464]. With further optimization, the Ion 314^{TM} chips could easily accommodate the evaluation of 16 amplicons at an average of more than 500x coverage per pool under the 28-pool scheme we utilized. Theoretically, for the Ion 316^{TM} chip, 160 amplicons could be evaluated under the same conditions, but adjusting the technical parameters under the current technology to guarantee little variability from pooling to sequencing becomes harder unless a similar system to the Ampliseq^{TM}, used for human genes [464], can be rapidly implemented for plants.

Another factor that came into consideration was how to decrease the level of false positives in data. Whereas it is has been reported that the level of false positives in Ion Torrent^{TM} data is larger than Illumina^{TM}, it has also been shown that Ion Torrent^{TM} can detect more true positives given enough coverage [462]. From our data it was evident that the increase in depth upon using the Ion 316^{TM} chip was concomitant with a decrease in false positives (compare Tables 3 and 4), showing that read depth is key for separating real mutations from noise in SNV studies [460], and in studies where allele frequency needs to be resolved [485]. While our proof of concept and scale experiments differed in the amplicons used, read mapping statistics (**Table**

5.3 and **Table 5.4**) demonstrated that the experiment with the 316 chip had an increase in the number of reads by over 5 fold when compared to the 314 chip. Since other factors like GC content and small differences in amplicon size are still difficult to control for and do not have a clear correlation always with read depth, we believe that the technical increase on read number by selecting a higher capacity chip is the key factor in obtaining more rare variants. Unfortunately, most second generation NGS technologies still present high error rates (see below). While we tried to control for equimolar amounts of DNA when pooling samples, PCR steps result in uncontrollable dilutions of some samples before sequencing which will result in the loss of some SNVs among noise. Improvements to eliminate such technical variability will increase our ability for SNV detection.

There are additional elements to take into account. For example, it was noticed that the background frequency of base substitution varied between the type of substitutions and among genes. While an A>C change had little or no background over 0.1% for the S20 region (see **Figure 5.2**), a T>C change in the same region and a G>A in the S900 region had larger substitution noise. A differential rate of substitution has been linked to the PGM from Ion Torrent™ upon studying bacteria, with G>A and T>C transitions presenting the higher rates of substitution with the Ion OneTouch 200™ template kit [479].

Interestingly while homopolymer errors are the most common error type from this technology [462,476,479], an study with bacteria found that substitutions have the highest variation frequencies, with standard deviations ranging from 26%-56% [479]. This has implications in the detection of rare variants (including false negatives) which may come up at lower frequencies as we detected in our study due to sample dilution in pools. For example, a 0.3% frequency was found for S20 variants in the pilot experiment (**Figure 5.2**), and since some random error can reach this frequency, this can lower our detection ability. While there were a few false positives embedded into homopolymeric tracts which constitute the bigger source of error of Ion Torrent technology [479], no specific position or sequence-specific bias in the SNVs that were not confirmed by sanger could be inferred to make any generalization.

Compared to Roche/454 technologies where mutagenized populations are used to discover rare variants [450,459] the Ion Torrent technology offers a higher read depth in short times which results in a higher probability to find the mutated bases. While one of the main advantages of 454 sequencing over other technologies was their read length (>400bp), Ion

Torrent is quickly catching up to offer similar read lengths of high quality [464]. Our study achieved similar throughput (314 chips) in the number of reads obtained as the aforementioned 454 studies, but the time required for a run on an Ion Torrent PGM is just over 2 hours while the most basic 454 sequencers use at least 10 hours and with prices of equipment and cost per base which are above the ones of Ion Torrent [476,486]. Furthermore higher throughput was achieved (5-fold) when changing to a 316 chip without a change in runtime while for 454 technology the use of higher-end sequencers may require up to 23 hours for a run [486]. Finally, the rate of false positives seems to be on a similar range for these two technologies.

Our methodology was based on a previous Illumina study [460] and therefore some conclusions can be drawn by comparing the two. It was clear that because of the similar methodology similar results were obtained in several fields. For example, low coverage resulted in increased noise which impaired SNV detection in both studies. The throughput of Illumina is generally higher resulting in detection of many more mutations. However, this is achieved in longer runtimes (days) and with more expensive equipment, although cost per base is lower on Illumina [476,477,486]. Because of the higher error rates of the Ion Torrent technology the amount of false positives is usually higher than on Illumina [462].

5.6 Conclusion

We have demonstrated that the Ion Torrent™ can be used in a scalable, amplicon-based approach for efficient discovery of mutations in a small number of genes. The efficiency of the method is limited by the rate of false positives, which may be decreased by higher read-depth and further technical optimization. Ion Torrent™ technology has been demonstrated to introduce biases at errors at specific steps during sequencing [483], as have other sequencing technologies, especially when PCR step is used in sample preparation [487]. Nevertheless, the Ion Torrent PGM™ platform detected rare variants with as low as 0.3% frequency per pool according to our results, which is above the substitution error of 0.1% calculated for the technology [464], and we showed that 768 individuals could be easily pooled per run. The Ion Torrent™ is one of the first technologies that does need optical systems to detect nucleotide incorporation, and does not use modified nucleotides [464,488]; in addition it has a good combination of throughput, cost and time saving compared to other systems [462,476,484,489]. The use of chips with larger capacity [464], will allow increasing both the number of genes

and/or pooled samples. Additionally, the availability of 400bp kits now allows exploring larger regions of interest without the need of using paired ends.

CHAPTER 6 - General discussion and conclusions

6.1 General outcomes

Transposable elements play an essential role in the evolution and regulation of plant genomes [21,25,26]. Their initial characterization as junk DNA has changed as research has revealed their regulatory influence on genes, and their ability to generate genomic rearrangements.

Our previous analysis showed that at least 23% of the flax genome is covered by TE-derived sequences [218]. Furthermore, we believe that our analysis underestimated the TE content of the genome, due to degeneracy in some TE residual fragments, and because some TE-rich regions of the genome may not have been incorporated into the assembly. Thus, the actual proportion of TEs in the flax genome is probably closer to 30%, which is consistent with similar-sized genomes such as rice, which has an estimated transposon content of 35% [490].

Since most annotated TEs in the flax genome belonged to the Ty1-*copia* superfamily, a superfamily that showed increasing activity in the last 5 million years, and generally had close proximity to genes [218], we believed that Ty1-*copia* could have a strong influence on genome structure and gene regulation, which could potentially affect the phenotype. Moreover, our literature survey of the influence of *copia*-type elements on gene regulation and their potential influence on recombination of gene families, like disease resistance genes (Chapter 1), demonstrated how these elements have become important factors in many plant genomes.

In Chapter 2 we demonstrated that particular families of Ty1-*copia* retrotransposons must have been active in the past few hundred years, after selective breeding began. This observation confirms my hypothesis (overview in Chapter 1) that our bioinformatics prediction for selected families would likely render them active. The cultivation of flax with specific traits is therefore correlated with the differential activation of TEs, resulting in a large amount of retrotransposon-derived DNA polymorphisms. We cannot determine whether breeding practices per se resulted in TE activation, but the influence of trait selection itself (e.g. cross-breeding) for cultivation, can result in differential stresses that can elicit a differential TE response [40]. Additionally, the genomic context, epigenetic regulation and presence of regulatory motifs in the LTRs are factors that alter how TEs are differentially expressed among cultivars. From the Ty1-*copia* families studied, 66.7% of the sequenced polymorphic insertions fell within genes and an additional 16.7% were within 1 kb of genes. This pattern of insertion and the characteristics of the families studied in flax resembles what happens in *A. thaliana,* where *copia*-type elements insert more

randomly, have more recent activity and associate with euchromatic regions, while *gypsy* retrotransposons tend to insert in heterochomatic regions and were more active in the past [66]. The study of the insertional bias of *A. thaliana* retrotransposons showed that while *copia* elements can be located more proximally to genes, they are also subject to negative selection. If this holds true for flax, it is likely that many of the novel insertions among cultivars could be purified or degenerate rapidly, and would not affect regulation of the genes over the long term. However, even if residual sections of TEs were to remain in close proximity or inside genes, they could have a regulatory influence due to their conserved motifs in LTR sections [92,106]. The insertion of Ty1-*copia* elements in gene-rich regions of flax is an important source of variability among cultivars that can be explored to evaluate the impact of the insertion on gene expression. Our preliminary tests with TE insertion in four flax genes showed one with variable expression linked to the retrotransposon insertion (Chapter 2).

In Chapter 3 we analyzed different tissues, wounding, and treatment with a fungal extract or *F. oxysporum*, for evidence of changes in TE transcript abundance. None of the stress conditions elicited clear changes in retrotransposon regulation, which is contrary to what we expected with our hypothesis. Significant differential expression was only present when comparing tissues within a cultivar, or comparing the same tissue between cultivars. The search for transcription factor binding sites in LTRs showed numerous, conserved cis-elements that are putatively associated with stress-responses [286,287], which is consistent with the ability of retrotransposons to be activated by diverse elicitors; however, the context of the insertion (genomic region where the TE is inserted), and the state of epigenetic silencing are probably strong factors in determining how these TEs are activated. Based on reports in other species, it is likely that activation is most common in meristematic/reproductive tissues or during tissue culture [56,57,92,309,312], and this activation is probably linked to common epigenetic reprogramming in these tissues and conditions [491,492]. Pursuing a tissue or cell-specific characterization of TE transcriptional changes will potentially give better results in the search for flax retrotransposon activation. At this higher resolution, additional stresses can be tested in parallel with the study of the epigenetic changes to find how they relate to TE activation. The Ty1-*copia* families used in this thesis still make good candidates for research in specific tissues since they showed to be recently active in cultivars and constitutively expressed, but with potential to be regulated in our stress-response experiments.

In Chapter 4 we characterized two flax cultivars to follow disease progression upon inoculation with *F. oxysporum* and found that in the most susceptible cultivar (Lutea), chitinase genes presented an earlier response than in the moderately resistant cultivar (CDC Bethune). Furthermore, the study of the full transcriptional response in CDC Bethune demonstrated that defense responses were deployed mainly 18 DPI, which is contrary to my hypothesis that molecular responses would appear in the first two days post-inoculation. I believe that a later deployment of the molecular response could be related to several factors. Because CDC Bethune has moderate resistance to fusarium wilt [333], I speculate that this cultivar might have several constitutive defenses in place even before the interaction with the pathogen. For example, tomato breeding lines had higher constitutive expression of chitinases and glucanases than the susceptible lines to *Alternaria solani* [342,493], and this pattern of gene expression is related to the inheritance of resistance to the pathogen. This constitutive activation of chitinases and glucanases allows degradation of pathogen cell wall molecules that would act as elicitors and activate a cascade leading to a general stress response. For example, the chitin signalling process, which can result in activation of several defense genes, depends on chitinases degrading fungal cell walls, oligomer detection by the chitin elicitor binding protein (CEBiP), and signal transduction by a LysM domain-containing receptor-like kinase 1 (LysM RLK1) [366]. Examination of these genes showed that several chitinases and CEBiP and LysM RLKs had detectable non-differential expression levels throughout the time course of our study. Constitutive expression of defense-related genes has also been proposed as a mechanism to develop partial resistance (a broad-spectrum resistance that builds up with age and activates defense genes constitutively) in rice before infection with *Magnaporthe oryzae* [494]. Besides depending on age (with resistance changing week to week and even between young and older leafs), cultivars in rice present broad differences in the levels of constitutive expression of the defense related genes. This pattern would explain some of the differences in flax cultivars, but also agree with changes in constitutive transcript abundance that we found when examining Ty1-*copia* elements using the RNA-seq data (see chapter 3). Other examples showing high constitutive defense gene expression include: i) the maize rachis, where resistant inbred lines depend on high level of constitutive defenses while susceptible ones rely on induced gene changes, when confronted with *Aspergillus flavus* [495]; ii) a wheat resistant genotype which has higher levels of gene expression of several pathogen-defense genes when compared with the

susceptible genotype before infection with *Blumeria graminis* [496]; iii) an olive resistant cultivar where the basal expression of 17 genes was higher than in the susceptible cultivar before infection with the fungus *Spilocaea oleagina*; and iv) rice cultivars that were resistant or susceptible to *Xanthomonas oryzae* and *Pyricularia grisea* where 12 defense genes showed constitutive expression that increased after infection [497]. Finally, we cannot exclude the possibility that the greater resistance of CDC Bethune is due to preformed, broad-spectrum anatomical defenses like reinforced cell walls, or waxy cuticles.

Additionally, most transcriptional changes found with our RNA-seq analysis fit a typical model of activation of genes in response to pathogen attack: with genes perceiving pathogen elicitors (receptors), transductions of signals (kinases), a transcriptional reprogramming (transcription factors), and responses of defense genes and hormone signalling. However, several genes were regulated in unexpected ways that would seemingly favor pathogen entry. Many cell wall genes that are usually activated during growth were upregulated, and this could result in cell wall weakening; amino acid transporters and aquaporins activation could favor the establishment of nutrient sinks and haustorial development [429,431]. But what is still unknown is the role of the major latex proteins (MLPs) in this interaction, which were mostly downregulated in our study, while expression patterns in other pathosystems seem the opposite [434,437,498]. In cotton, MLP28 interacts with an ethylene response transcription factor and enhances its binding activity to target defense gene promoters [434] and was previously speculated to interact with flavonoids [499]. Interestingly, the upregulation of two MLP proteins against *Verticillium dahliae*, was inconsistent with their respective patterns of downregulation of transcription in cotton [500]. In other processes like cell wall maturation, MLP423 (which matches the annotation of most of the hits found in our study) is upregulated when comparing regions of fast directional growth in the stem with regions where elongative growth ceases [501], showing this protein could be involved in cell growth also.

In Chapter 5 we described a novel method to simultaneously analyze hundreds of plants from a mutagenize population to find rare variants in selected genes. This technology is best suited to finding single nucleotide mutations in a handful of target genes, but the rapid advancement of next generation sequencing technologies now allows sequencing full genomes or exomes for hundreds or thousands of individuals, which might be more effective (if funds are available) when mutations in a large number of genes are being sought. However, the underlying

methodology presented here still presents the opportunity of targeting genes of interest, after populations have been subjected to mutagenesis, to find altered phenotypes. In our case, this methodology could be used to target mutations in genes involved in TE methylation, since demethylation is related to re-activation of TEs [57,309,312]. For example, silencing of transposons in *A. thaliana* is lifted in mutants of DNA methyltransferase (*met1*), a chromatin remodelling ATPase (*ddm1*), and a histone modification gene (*sil1*) [502]; however, such silencing is not generalized and different TEs behave distinctly in different mutants, showing how regulation does not depend on a single mechanism for all TEs. The study of the different genes involved in epigenetic modification using mutants in flax would allow dissection of these mechanisms in greater detail.

6.2 Ongoing research and future perspectives

6.2.1 Analysis of full genomes for TE-derived polymorphisms

Based on the findings from Chapter 2, Ty1-*copia* families represent a reservoir of genetic variation. We currently have Illumina whole genome sequencing data for 16 flax cultivars (sequence read archive: SRA114122) which could be used to detect genome polymorphisms for selected TE families using bioinformatics tools like ITIS (Identification of Transposon Insertion sites) [217]. This approach can also be used to detect polymorphisms in whole-genome samples from four stages of regeneration of flax plants from hypocotyls (sequence read archive: SRA188726), that were prepared based on tissue culture being one of the common elicitors of TE activity [56,57]. In the case of cultivar comparison TE-derived variability can be used for assessing diversity but also as a source of potential mutations that can be related to phenotypic changes. Once mutations have been identified in genes of interest, transcriptional regulation of selected genes can be assessed using qRT-PCR. In parallel it will be necessary to determine if such TE modifications are homozygous or heterozygous because transcriptional changes in the host gene may vary accordingly. In the case of tissue culture, the progression of TE-derived somaclonal variation can be determined by comparing to normal flax plants.

6.2.2 Study of the flax-fusarium pathosystem

In continuing to dissect the interaction between flax and its fungal pathogen *F. oxysporum*, we have sequenced the *Fusarium oxysporum* f. sp. *lini* genome in collaboration with

Professor Lina Maria Quesada (North Carolina State University). The full assembly and annotation of the genome is underway and when finished, will be used to map RNA-seq reads found in our mixed sample utilized in Chapter 4 and reads from fungal controls, to determine the differentially expressed genes used by the fungi in the infection process.

The full genome annotation of *F. oxysporm* f. sp, *lini*, will reveal specific pathogenicity regions (e.g. chromosomes) [314,503,504] giving specificity to its interaction with flax, further contributing to the current knowledge of the evolution of variability among formae speciales, to produce specific host-pathogen interactions. Interestingly, the discovery of the lineage-specific pathogenicity genomic regions in *F. oxysporum* f. sp. *lycopersici* (which infects tomato) indicates that these regions may be rich in transposable elements. This represents an opportunity to study if TEs have an influence also in fungi for the evolution of regions that should be subject to rapid evolution to generate new virulence factors that can overcome plant defenses. This would be the matching side from the fungal perspective on what was proposed on Chapter 1 for regions rich in resistance genes in plants, where TEs could also be players of new gene variants in this arms race between plant and pathogen.

The identification of pathogenicity regions will also allow easier targeting of secretome/effector genes which are the base for the infection process and may determine resistance or susceptibility of the host [505]. The prediction of the secretome will allow easier identification and selection of the genes from the RNA-seq experiment as key factors for the pathosystem interaction. This knowledge can be then used to breed and select resistant varieties. Finally, in an attempt to explore the possibility of cross-kingdom manipulation of gene regulation [156,506] (one of the hypothesis we posed on Chapter 4), we started isolating small RNAs from our flax samples infected with *F. oxysporum*, in collaboration with Dr. Juan Antonio Jovel (Faculty of Medicine, University of Alberta). We have yet to achieve high quality in our small RNA isolation, but the few samples used for a pilot study showed no differential expression between water control and Fusarium-inoculated samples.

References

1. Hickey M, King C. 100 families of flowering plants. 2nd. editi. New York: Cambridge univerisity press.; 1991.

2. Mcdill J, Repplinger M, Simpson BB, Kadereit JW. The phylogeny of *Linum* and Linaceae subfamily Linoideae, with implications for their systematics, biogeography, and evolution of heterostyly. 2009;34:386–405.

3. Vaisey-Genser M, Morris DH. Introduction: History of the cultivation and uses of flaxseed. In: Muir AD, Westcott ND, editors. Flax, the genus *Linum*. London: Taylor & Francis; 2003. p. 1–21.

4. Kvavadze E, Bar-Yosef O, Belfer-Cohen A, Boaretto E, Jakeli N, Matskevich Z, et al. 30,000-Year-old wild flax fibers. Science (80-.). 2009;325:1359.

5. Fu Y-B, Diederichsen A, Allaby RG. Locus-specific view of flax domestication history. Ecol. Evol. 2012;2:139–52.

6. Diederichsen A, Richards K. Cultivated flax and the genus *Linum* L.: Taxonomy and germplasm conservation. Flax, the genus *Linum*. London: Taylor & Francis; 2003. p. 22–54.

7. Marchenkov A, Rozhmina T, Uschapovsky I, Muir AD. Cultivation of flax. Flax, the genus *Linum*. London: Taylor & Francis; 2003. p. 74–91.

8. Canada - a flax leader [Internet]. Flax Counc. Canada. 2016. Available from: http://flaxcouncil.ca/resources/about-flax/canada-a-flax-leader/

9. Flax: a healthy food [Internet]. Flax Counc. Canada. 2016. Available from: http://flaxcouncil.ca/resources/nutrition/general-nutrition-information/flax-a-healthy-food/

10. Westcott ND, Muir AD. Chemical studies on the constituents of *Linum* spp . Flax, the genus *Linum*. London: Taylor & Francis; 2003. p. 55–73.

11. Rajaram S. Health benefits of plant-derived alpha-linolenic acid. Am. J. Clin. Nutr. 2014;100:443–8.

12. Mason JK, Thompson LU. Flaxseed and its lignan and oil components: can they play a role in reducing the risk of and improving the treatment of breast cancer? Appl. Physiol. Nutr. Metab. 2014;39:663–78.

13. Peterson J, Dwyer J, Adlercreutz H, Scalbert A, Mccullough ML. NIH Public Access. Nutr. Rev. 2011;68:571–603.

14. Rashid KY. Principal diseases of flax. Flax, the genus *Linum*. London: Taylor & Francis; 2003. p. 92–123.
15. Wang Z, Hobson N, Galindo L, Zhu S, Shi D, Mcdill J, et al. The genome of flax (*Linum usitatissimum*) assembled de novo from short shotgun sequence reads. Plant J. 2012;72:461–73.
16. Sveinsson S, McDill J, Wong GKS, Li J, Li X, Deyholos MK, et al. Phylogenetic pinpointing of a paleopolyploidy event within the flax genus (*Linum*) using transcriptomics. Ann. Bot. 2014;113:753–61.
17. McClintock B. The origin and behavior of mutable loci in maize. Proc. Natl. Acad. Sci. USA. 1950;36:344–55.
18. McClintock B. Chromosome organization and genic expression. Cold Spring Harb. Symp. Quant. Biol. 1951;16:13–47.
19. Fedoroff N, Wessler S, Shure M. Isolation of the transposable maize controlling elements *Ac* and *Ds*. Cell. 1983;35:235–42.
20. McClintock B. The significance of responses of the genome to challenge. Science (80-.). 1984;226:792–801.
21. Bennetzen JL, Wang H. The contributions of transposable elements to the structure, function, and evolution of plant genomes. Annu. Rev. Plant Biol. 2014;65:505–30.
22. Doolittle WF, Sapienza C. Selfish genes, the phenotype paradigm and genome evolution. Nature. 1980;284:601–3.
23. Orgel LE, Crick FHC. Selfish DNA: the ultimate parasite. Nature. 1980;284:604–7.
24. Rebollo R, Romanish MT, Mager DL. Transposable elements: An abundant and natural source of regulatory sequences for host genes. Annu. Rev. Genet. 2012;46:21–42.
25. Grandbastien MA. LTR retrotransposons, handy hitchhikers of plant regulation and stress response. Biochim. Biophys. Acta. 2015;1849:403–16.
26. Vitte C, Fustier M-A, Alix K, Tenaillon MI. The bright side of transposons in crop evolution. Brief. Funct. Genomics. 2014;13:276–95.
27. Lisch D. How important are transposons for plant evolution? Nat Rev Genet. 2012;14:49–61.
28. Civan P, Svec M, Huptvogel P. On the coevolution of transposable elements and plant genomes. J. Bot. 2011;2011:1–9.
29. Tenaillon MI, Hollister JD, Gaut BS. A triptych of the evolution of plant transposable elements. Trends Plant Sci. 2010;15:471–8.

30. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 2000;408:796–815.
31. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. Science. 2009;326:1112–5.
32. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. 2007;8:973–82.
33. Witte CP, Le QH, Bureau T, Kumar a. Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. Proc. Natl. Acad. Sci. U. S. A. 2001;98:13778–83.
34. Vitte C, Panaud O. LTR retrotransposons and flowering plant genome size : emergence of the increase / decrease model. 2005;107:91–107.
35. Kalendar R, Vicient CM, Peleg O, Anamthawat-jonsson K, Bolshoy A, Schulman AH. Large retrotransposon derivatives: abundant , conserved but nonautonomous retroelements of barley and related genomes. Genetics. 2004;166:1437–50.
36. Feschotte C, Jiang N, Wessler SR. Plant transposable elements: where genetics meets genomics. Nat. Rev. Genet. 2002;3:329–41.
37. Kidwell MG. Transposable elements and the evolution of genome size in eukaryotes. Genetica. 2002;115:49–63.
38. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. The paleontology of intergene retrotransposons of maize. Nature. 1998;20:43–5.
39. SanMiguel P, Tikhonov A, Jin Y, Motchoulskaia N, Zakharov D, Melake-berhan A, et al. Nested retrotransposons in the intergenic regions of the maize genome. Science (80-.). 1996;274:765–8.
40. Kalendar R, Tanskanen J, Immonen S, Nevo E, Schulman AH. Genome evolution of wild barley (Hordeum spontaneum) by *BARE-1* retrotransposon dynamics in response to sharp microclimatic divergence. Proc. Natl. Acad. Sci. 2000;97:6603–7.
41. Hu TT, Pattyn P, Bakker EG, Cao J, Cheng J-F, Clark RM, et al. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat. Genet. 2011;43:476–81.
42. Lockton S, Gaut BS. The evolution of transposable elements in natural populations of self-fertilizing *Arabidopsis thaliana* and its outcrossing relative *Arabidopsis lyrata*. BMC Evol. Biol. 2010;10:10.

43. Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, et al. Doubling genome size without polyploidization : Dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. Genome Res. 2006;16:1262–9.
44. Kumar A, Bennetzen J. Plant retrotransposons. Annu. Rev. Genet. 1999;33:479–532.
45. Beguiristain T, Grandbastien M-A, Puigdomènech P, Casacuberta JM. Three *Tnt1* subfamilies show different stress-associated patterns of expression in tobacco. Consequences for retrotransposon control and evolution in plants. Plant Physiol. 2001;127:212–21.
46. Cavrak V V., Lettner N, Jamge S, Kosarewicz A, Bayer LM, Mittelsten Scheid O. How a retrotransposon exploits the plant's heat stress response for its activation. PLoS Genet. 2014;10:e1004115.
47. He P, Ma Y, Dai H, Li L, Liu Y, Li H, et al. Characterization of the hormone and stress-induced expression of *FaRE1* retrotransposon promoter in strawberry. J. Plant Biol. 2012;55:1–7.
48. Mhiri C, Vernhettes S, Casacuberta JM. The promoter of the tobacco Tnt1 retrotransposon is induced by wounding and by abiotic stress. Plant Mol. Biol. 1997;33:257–66.
49. Salazar M, González E, Casaretto JA, Casacuberta JM, Ruiz-Lara S. The promoter of the *TLC1.1* retrotransposon from *Solanum chilense* is activated by multiple stress-related signaling molecules. Plant Cell Rep. 2007;26:1861–8.
50. Sugimoto K, Takeda S, Hirochika H. MYB-related transcription factor NtMYB2 induced by wounding and elicitors is a regulator of the tobacco retrotransposon *Tto1* and defense-related genes. Plant Cell. 2000;12:2511–27.
51. Suoniemi A, Narvanto A, Schulman AH. The *BARE*-1 retrotransposon is transcribed in barley from an LTR promoter active in transient assays. Plant Mol. Biol. 1996;31:295–306.
52. Tapia G, Verdugo I, Poblete F, Gonza E. Involvement of ethylene in stress-induced expression of the *TLC1 . 1* retrotransposon from *Lycopersicon chilense* Dun. Plant Physiol. 2005;138:2075–86.
53. Woodrow P, Pontecorvo G, Ciarmiello LF, Fuggi A, Carillo P. *Ttd1a* promoter is involved in DNA-protein binding by salt and light stresses. Mol. Biol. Rep. 2011;38:3787–94.
54. Woodrow P, Pontecorvo G, Fantaccione S, Fuggi A, Kafantaris I, Parisi D, et al. Polymorphism of a new Ty1-*copia* retrotransposon in durum wheat under salt and light stresses. Theor. Appl. Genet. 2010;121:311–22.

55. Pouteau S, Grandbastien M-A, Boccara M. Microbial elicitors of plant defence responses activate transcription of a retrotransposon. Plant J. 1994;5:535–42.
56. Hirochika H, Sugimoto K, Otsuki Y, Tsugawa H, Kanda M. Retrotransposons of rice involved in mutations induced by tissue culture. Proc. Natl. Acad. Sci. USA. 1996;93:7783–8.
57. Liu ZL, Han FP, Tan M, Shan XH, Dong YZ, Wang XZ, et al. Activation of a rice endogenous retrotransposon *Tos17* in tissue culture is accompanied by cytosine demethylation and causes heritable alteration in methylation pattern of flanking genomic regions. Theor. Appl. Genet. 2004;109:200–9.
58. Grandbastien M, Audeon C, Bonnivard E, Casacuberta JM, Chalhoub B, Costa AP, et al. Stress activation and genomic impact of *Tnt1* retrotransposons in Solanaceae. Cytogenet. Genome Res. 2005;241:229–41.
59. Alzohairy A, Sabir JSM, Gyulai G, Younis R, Jansen R, Bahieldin A. Environmental stress activation of plant LTR-retrotransposons. Funct. Plant Biol. 2014;41:557–67.
60. Makarevitch I, Waters AJ, West PT, Stitzer M, Hirsch CN, Ross-Ibarra J, et al. Transposable elements contribute to activation of maize genes in response to abiotic stress. PLoS Genet. 2015;11:e1004915.
61. Petit M, Guidat C, Daniel J, Denis E, Montoriol E, Bui QT, et al. Mobilization of retrotransposons in synthetic allotetraploid tobacco. New Phytol. 2010;186:135–47.
62. Melayah D, Bonnivard E, Chalhoub B, Audeon C, Grandbastien M-A. The mobility of the tobacco *Tnt1* retrotransposon correlates with its transcriptional activation by fungal factors. Plant J. 2001;28:159–68.
63. Manninen O, Kalendar R, Robinson J, Schulman AH. Application of *BARE*-1 retrotransposon markers to the mapping of a major resistance gene for net blotch in barley. Mol. Gen. Genet. 2000;264:325–34.
64. Mhiri C, De Wit PJGM, Grandbastien M-A. Activation of the promoter of the *Tnt1* retrotransposon in tomato after inoculation with the fungal pathogen Cladosporium fulvum. Mol. Plant-Microbe Interact. 1999;12:592–603.
65. Takeda S, Sugimoto K, Otsuki H, Hirochika H. Transcriptional activation of the tobacco retrotransposon *Tto1* by wounding and methyl jasmonate. Plant Mol. Biol. 1998;36:365–76.
66. Pereira V. Insertion bias and purifying selection of retrotransposons in the *Arabidopsis thaliana* genome. Genome Biol. 2004;5:R79.

67. Schmidt T. LINEs , SINEs and repetitive DNA : non-LTR retrotransposons in plant genomes. Plant Mol. Biol. 1999;40:903–10.
68. Schwarz-Sommer Z, Leclercq L, Göbel E, Saedler H. *Cin4*, an insert altering the structure of the A1 gene in *Zea mays*, exhibits properties of nonviral retrotransposons. EMBO J. 1987;6:3873–80.
69. Kojima KK. Different integration site structures between *L1* protein-mediated retrotransposition in cis and retrotransposition in trans. 2010;1–9.
70. Kramerov DA, Vassetzky NS. Short retroposons in eukaryotic genomes. Int. Rev. Cytol. 2005;247:165–221.
71. Bennetzen JL. Transposable element contributions to plant gene and genome evolution. Plant Mol. Biol. 2000;42:251–69.
72. Han Y, Qin S, Wessler SR. Comparison of class 2 transposable elements at superfamily resolution reveals conserved and distinct features in cereal grass genomes. BMC Genomics. 2013;14:71.
73. Bureau TE, Wessler SR. *Stowaway* - a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. Plant Cell. 1994;6:907–16.
74. Bureau TE, Wessler SR. Mobile inverted-repeat elements of the Tourist family are associated with the genes of many cereal grasses. Proc. Natl. Acad. Sci. U. S. A. 1994;91:1411–5.
75. Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR. Pack-MULE transposable elements mediate gene evolution in plants. Nature. 2004;431:569–73.
76. Lisch D. Mutator transposons. Trends Plant Sci. 2002;7:498–504.
77. Grandbastien M-A, Lucas H, Morel J-B, Mhiri C, Vernhettes S, Casacuberta JM. The expression of the tobacco *Tnt1* retrotransposon is linked to plant defense responses. Genetica. 1997;100:241–52.
78. Okamoto H, Hirochika H. Efficient insertion mutagenesis of *Arabidopsis* by tissued culture-induced activation of the tobacco retrotransposon *Tto1*. Plant J. 2000;23:291–304.
79. Kimura Y, Tosa Y, Shimada S, Sogo R, Kusaba M, Sunaga T, et al. *OARE-1*, a Ty1-*copia* retrotransposon in oat activated by abiotic and biotic stresses. Plant Cell Physiol. 2001;42:1345–54.
80. Staton SE, Burke JM. Evolutionary transitions in the Asteraceae coincide with marked shifts in transposable element abundance. BMC Genomics. 2015;16:623.

81. Tam SM, Lefebvre V, Palloix A, Sage-Palloix A-M, Mhiri C, Grandbastien M-A. LTR-retrotransposons *Tnt1* and *T135* markers reveal genetic diversity and evolutionary relationships of domesticated peppers. Theor. Appl. Genet. 2009;119:973–89.
82. Senerchia N, Felber F, Parisod C. Contrasting evolutionary trajectories of multiple retrotransposons following independent allopolyploidy in wild wheats. New Phytol. 2014;202:975–85.
83. Parisod C, Mhiri C, Lim KY, Clarkson JJ, Chase MW, Leitch AR, et al. Differential dynamics of transposable elements during long-term diploidization of *Nicotiana* section Repandae (Solanaceae) allopolyploid genomes. PLoS One. 2012;7:e50352.
84. Piednoël M, Carrete-Vega G, Renner SS. Characterization of the LTR retrotransposon repertoire of a plant clade of six diploid and one tetraploid species. Plant J. 2013;75:699–709.
85. Kraitshtein Z, Yaakov B, Khasdan V, Kashkush K. Genetic and epigenetic dynamics of a retrotransposon after allopolyploidization of wheat. Genetics. 2010;186:801–12.
86. Parisod C, Salmon A, Zerjal T, Tenaillon M, Grandbastien M-A, Ainouche M. Rapid structural and epigenetic reorganization near transposable elements in hybrid and allopolyploid genomes in *Spartina*. New Phytol. 2009;184:1003–15.
87. Sarilar V, Palacios PM, Rousselet A, Ridel C, Falque M, Eber F, et al. Allopolyploidy has a moderate impact on restructuring at three contrasting transposable element insertion sites in resynthesized *Brassica napus* allotetraploids. New Phytol. 2013;198:593–604.
88. Paz RC, Rendina González AP, Ferrer MS, Masuelli RW. Short-term hybridisation activates *Tnt1* and *Tto1* Copia retrotransposons in wild tuber-bearing Solanum species. Plant Biol. 2015;17:860–9.
89. Parisod C, Alix K, Just J, Petit M, Sarilar V, Mhiri C, et al. Impact of transposable elements on the organization and function of allopolyploid genomes. New Phytol. 2010;186:37–45.
90. Grandbastien M-A, Spielmann A, Caboche M. Tnt1, a mobile retroviral-like transposable element of tobacco isolated by plant cell genetics. Nature. 1989;337:376–80.
91. Hirochika H, Otsuki H, Yoshikawa M, Otsuki Y, Sugimoto K, Takeda S. Autonomous transposition of the tobacco retrotransposon Tto1 in rice. Plant Cell. 1996;8:725–34.
92. Takeda S, Sugimoto K, Otsuki H, Hirochika H. A 13-bp cis-regulatory element in the LTR promoter of the tobacco retrotransposon *Tto1* is involved in responsiveness to tissue culture, wounding, methyl jasmonate and fungal elicitors. Plant J. 1999;18:383–93.

93. Ramallo E, Kalendar R, Schulman AH, Martínez-Izquierdo JA. *Reme1*, a *Copia* retrotransposon in melon, is transcriptionally induced by UV light. Plant Mol. Biol. 2008;66:137–50.
94. Ito H, Gaubert H, Bucher E, Mirouze M, Vaillant I, Paszkowski J. An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. Nature. 2011;472:115–9.
95. Ito H, Yoshida T, Tsukahara S, Kawabe A. Evolution of the *ONSEN* retrotransposon family activated upon heat stress in Brassicaceae. Gene. Elsevier B.V.; 2013;518:256–61.
96. Matsunaga W, Ohama N, Tanabe N, Masuta Y, Masuda S, Mitani N, et al. A small RNA mediated regulation of a stress-activated retrotransposon and the tissue specific transposition during the reproductive period in *Arabidopsis*. Front. Plant Sci. 2015;6:48.
97. Cao Y, Jiang Y, Ding M. Molecular characterization of a transcriptionally active Ty1/ *copia*-like retrotransposon in *Gossypium*. Plant Cell Rep. Springer Berlin Heidelberg; 2015;34:1037–47.
98. He P, Ma Y, Zhao G, Dai H, Li H, Chang L, et al. *FaRE*1: A transcriptionally active Ty1-*copia* retrotransposon in strawberry. J. Plant Res. 2010;123:707–14.
99. Casacuberta JM, Grandbastien M-A. Characterisation of LTR sequences involved in the protoplast specific expression of the tobacco *Tnt1* retrotransposon. Nucleic Acids Res. 1993;21:2087–93.
100. Cheng M-C, Liao P-M, Kuo W-W, Lin T-P. The *Arabidopsis* ETHYLENE RESPONSE FACTOR1 regulates abiotic stress-responsive gene expression by binding to different cis-acting elements in response to different stress signals. Plant Physiol. 2013;162:1566–82.
101. Lorenzo O, Piqueras R, Sanchez-Serrano JJ, Solano R. ETHYLENE RESPONSE FACTOR1 integrates signals from ethylene and jasmonate pathways in plant defense. Plant Cell. 2003;15:165–78.
102. Vukich M, Giordani T, Natali L, Cavallini A. *Copia* and *Gypsy* retrotransposons activity in sunflower (*Helianthus annuus* L.). BMC Plant Biol. 2009;12:1–13.
103. Bennetzen JL. Transposable element contributions to plant gene and genome evolution. Plant Mol. Biol. 2000;42:251–69.
104. Guo M, Liu J-H, Ma X, Luo D-X, Gong Z-H, Lu M-H. The plant heat stress transcription factors (HSFs): structure, regulation, and function in response to abiotic stresses. Front. Plant

Sci. 2016;7:114.
105. Butelli E, Licciardello C, Zhang Y, Liu J, Mackay S, Bailey P, et al. Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. Plant Cell. 2012;24:1242–55.
106. Warner SAJ, Scott R, Draper J. Characterisation of a wound-induced transcript from the monocot asparagus that shares similarity with a class of intracellular pathogenesis-related (PR) proteins. Plant Mol. Biol. 1992;19:555–61.
107. Studer A, Zhao Q, Ross-Ibarra J, Doebley J. Identification of a functional transposon insertion in the maize domestication gene tb1. Nat. Genet. 2011;43:1160–3.
108. Hayashi K, Yoshida H. Refunctionalization of the ancient rice blast disease resistance gene *Pit* by the recruitment of a retrotransposon as a promoter. Plant J. 2009;57:413–25.
109. Vitte C, Panaud O. Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice *Oryza sativa* L. Mol. Biol. Evol. 2003;20:528–40.
110. Ma J, Bennetzen JL. Rapid recent growth and divergence of rice nuclear genomes. Proc. Natl. Acad. Sci. USA. 2004;101:12404–10.
111. De Souza FSJ, Franchini LF, Rubinstein M. Exaptation of transposable elements into novel Cis-regulatory elements: Is the evidence always strong? Mol. Biol. Evol. 2013;30:1239–51.
112. Varagona MJ, Purugganan M, Wessler SR. Alternative splicing induced by insertion of retrotransposons into the maize *waxy* gene. Plant Cell. 1992;4:811–20.
113. Bureau TE, White SE, Wessler SR. Transduction of a cellular gene by a plant retroelement. Cell. 1994;77:479–80.
114. Kashkush K, Feldman M, Levy AA. Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. Nat. Genet. 2002;33:102–6.
115. Kashkush K, Khasdan V. Large-scale survey of cytosine methylation of retrotransposons and the impact of readout transcription from long terminal repeats on expression of adjacent rice genes. Genetics. 2007;177:1975–85.
116. Hollister JD, Gaut BS. Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res. 2009;19:1419–28.
117. Hollister JD, Smith LM, Guo Y, Ott F, Weigel D, Gaut BS. Transposable elements and

small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*. Proc. Natl. Acad. Sci. USA. 2011;108:2322–7.
118. Marillonnet S. Retrotransposon insertion into the maize *waxy* gene results in tissue-specific RNA processing. Plant Cell. 1997;9:967–78.
119. Tsuchiya T, Eulgem T. An alternative polyadenylation mechanism coopted to the *Arabidopsis RPP7* gene through intronic retrotransposon domestication. Proc Natl Acad Sci U S A. 2013;110:E3535-43.
120. Yao J-L, Dong Y-H, Morris BAM. Parthenocarpic apple fruit production conferred by transposon insertion mutations in a MADS-box transcription factor. Proc. Natl. Acad. Sci. USA. 2001;98:1306–11.
121. Costa JH, De Melo DF, Gouveia Z, Cardoso HG, Peixe A, Arnholdt-Schmitt B. The alternative oxidase family of *Vitis vinifera* reveals an attractive model to study the importance of genomic design. Physiol. Plant. 2009;137:553–65.
122. Li Q, Xiao G, Zhu YX. Single-nucleotide resolution mapping of the *Gossypium raimondii* transcriptome reveals a new mechanism for alternative splicing of introns. Mol. Plant. The Authors 2014.; 2014;7:829–40.
123. Pouteau S, Spielmann A, Meyer C, Grandbastien M-A, Caboche M. Effects of *Tnt1* tobacco retrotransposon insertion on target gene transcription. Mol. Gen. Genet. 2000;228:233–9.
124. Liu B, Kanazawa A, Matsumura H, Takahashi R, Harada K, Abe J. Genetic redundancy in soybean photoresponses associated with duplication of the phytochrome A gene. Genetics. 2008;180:995–1007.
125. Kanazawa A, Liu B, Kong F, Arase S, Abe J. Adaptive evolution involving gene duplication and insertion of a novel *Ty1/copia*-like retrotransposon in soybean. J. Mol. Evol. 2009;69:164–75.
126. Hori Y, Fujimoto R, Sato Y, Nishio T. A novel wx mutation caused by insertion of a retrotransposon-like sequence in a glutinous cultivar of rice (*Oryza sativa*). Theor. Appl. Genet. 2007;115:217–24.
127. Elrouby N, Bureau TE. *Bs1*, a new chimeric gene formed by retrotransposon-mediated exon shuffling in maize. Plant Physiol. 2010;153:1413–24.
128. Xiao H, Jiang N, Schaffner E, Stockinger EJ, Van Der Knaap E. A retrotransposon-mediated gene duplication underlies morphological variation in tomato fruit. Science (80-.).

2008;319:1527–31.
129. Jiang S-Y, Ramachandran S. Genome-wide survey and comparative analysis of LTR retrotransposons and their captured genes in rice and sorghum. PLoS One. 2013;8:e71118.
130. Hammond SM, Caudy AA, Hannon GJ. Post-transcriptional gene silencing by double-stranded RNA. Nat Rev Genet. 2001;2:110–9.
131. Lippman Z, Gendrel A-V, Black M, Vaughn MW, Dedhia N, Mccombie WR, et al. Role of transposable elements in heterochromatin and epigenetic control. Nature. 2004;430:471–6.
132. Wang X, Weigel D, Smith LM. Transposon variants and their effects on gene expression in *Arabidopsis*. PLoS Genet. 2013;9:e1003255.
133. McCue AD, Slotkin RK. Transposable element small RNAs as regulators of gene expression. Trends Genet. Elsevier Ltd; 2012;28:616–23.
134. Mccue AD, Nuthikattu S, Reeder SH, Slotkin RK. Gene expression and stress response mediated by the epigenetic regulation of a transposable element small RNA. PLoS Genet. 2012;8:e1002474.
135. McCue AD, Nuthikattu S, Slotkin RK. Genome-wide identification of genes regulated in trans by transposable element small interfering RNAs. RNA Biol. 2013;10:1379–95.
136. Le TN, Miyazaki Y, Takuno S, Saze H. Epigenetic regulation of intragenic transposable elements impacts gene transcription in *Arabidopsis thaliana*. Nucleic Acids Res. 2015;43:3911–21.
137. Hickey DA, Benkel B. Introns as relict retrotransposons: Implications for the evolutionary origin of eukaryotic mRNA splicing mechanisms. J. Theor. Biol. 1986;121:283–91.
138. Purugganan M, Wessler S. The splicing of transposable elements and its role in intron evolution. Genetica. 1992;86:295–303.
139. Catania F, Gao X, Scofield DG. Endogenous mechanisms for the origins of spliceosomal introns. J. Hered. 2009;100:591–6.
140. Rogozin IB, Carmel L, Csuros M, Koonin E V. Origin and evolution of spliceosomal introns. Biol. Direct. 2012;7:11.
141. Borchert GM, Holton NW, Williams JD, Hernan WL, Bishop IP, Dembosky JA, et al. Comprehensive analysis of microRNA genomic loci identifies pervasive repetitive-element origins. Mob. Genet. Elements. 2011;1:8–17.
142. Voinnet O. Origin, biogenesis, and activity of plant MicroRNAs. Cell. 2009;136:669–87.

143. Li Y, Li C, Xia J, Jin Y. Domestication of transposable elements into MicroRNA genes in plants. PLoS One. 2011;6:e19212.
144. Devos KM, Brown JKM, Bennetzen JL. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. Genome Res. 2002;12:1075–9.
145. Huang JT, Dooner HK. Macrotransposition and other complex chromosomal restructuring in maize by closely linked transposons in direct orientation. Plant Cell. 2008;20:2019–32.
146. Onozawa M, Zhang Z, Kim YJ, Goldberg L, Varga T, Bergsagel PL, et al. Repair of DNA double-strand breaks by templated nucleotide sequence insertions derived from distant regions of the genome. Proc. Natl. Acad. Sci. U. S. A. 2014;111:7729–34.
147. Flor HH. The complementary genic systems in flax and flax rust. Adv. Genet. 1956;8:29–54.
148. Thomma BPHJ, Nürnberger T, Joosten MHAJ. Of PAMPs and effectors: the blurred PTI-ETI dichotomy. Plant Cell. 2011;23:4–15.
149. Richter TE, Ronald PC. The evolution of disease resistance genes. Plant Mol. Biol. 2000;42:195–204.
150. Andersson G, Svensson AC, Setterblad N, Rask L. Retroelements in the human MHC class II region. Trends Genet. 1998;14:109–14.
151. Song W, Pi L, Wang G, Gardner J, Hoisten T, Pamela CR. Evolution of the rice *Xa21* disease resistance gene family. Plant Cell. 1997;9:1279–87.
152. Song W, Pi L, Bureau TE. Identification and characterization of 14 transposon-like elements in the noncoding regions of members of the *Xa21* family of disease resistance genes in rice. Mol. Gen. Genet. 1998;258:449–56.
153. Nagy ED, Bennetzen JL. Pathogen corruption and site-directed recombination at a plant disease resistance gene cluster. Genome Res. 2007;18:1918–23.
154. Ma J, Bennetzen JL. Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice. Proc. Natl. Acad. Sci. 2006;103:383–8.
155. Noël L, Moores TL, van der Biezen EA, Parniske M, Daniels MJ, Parker JE, et al. Pronounced intraspecific haplotype divergence at the *RPP5* complex disease resistance locus of *Arabidopsis*. Plant Cell. 1999;11:2099–111.
156. Weiberg A, Wang M, Bellinger M, Jin H. Small RNAs: a new paradigm in plant-microbe interactions. Annu. Rev. Phytopathol. 2014;52:495–516.

157. Zhu Q, Shan W, Ayliffe MA, Wang M. Epigenetic mechanisms: an emerging player in plant-microbe interactions. Mol. Plant-Microbe Interact. 2016;29:187–96.
158. Akimoto K, Katakami H, Kim H-J, Ogawa E, Sano CM, Wada Y, et al. Epigenetic inheritance in rice plants. Ann. Bot. 2007;100:205–17.
159. Tsukahara S, Kobayashi A, Kawabe A, Mathieu O, Miura A, Kakutani T. Bursts of retrotransposition reproduced in *Arabidopsis*. Nature. 2009;461:423–6.
160. Dowen RH, Pelizzola M, Schmitz RJ, Lister R, Dowen JM, Nery JR, et al. Widespread dynamic DNA methylation in response to biotic stress. Proc. Natl. Acad. Sci. U. S. A. 2012;109:E2183-91.
161. Zhang X. Dynamic differential methylation facilitates pathogen stress response in Arabidopsis. Proc. Natl. Acad. Sci. 2012;109:12842–3.
162. Le T-N, Schumann U, Smith NA, Tiwari S, Au P, Zhu Q-H, et al. DNA demethylases target promoter transposable elements to positively regulate stress responsive genes in *Arabidopsis*. Genome Biol. 2014;15:458.
163. Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN, Richardson AO, et al. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature. 2009;461:1130–4.
164. Waugh R, McLean K, Flavell AJ, Pearce SR, Kumar A, Thomas BBT, et al. Genetic distribution of *Bare-1*-like retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphisms (S-SAP). Mol. Gen. Genet. 1997;253:687–94.
165. Syed NH, Flavell AJ. Sequence-specific amplification polymorphisms (SSAPs): a multi-locus approach for analyzing transposon insertions. Nat. Protoc. 2006;1:2746–52.
166. Van Den Broeck D, Maes T, Sauer M, Zethof J, Keukeleire P De, Hauw D, et al. Transposon Display identifies individual transposable elements in high copy number lines. Plant J. 1998;13:121–9.
167. Kalendar R, Schulman AH. IRAP and REMAP for retrotransposon-based genotyping and fingerprinting. Nat. Protoc. 2006;1:2478–84.
168. Flavell AJ, Knox MR, Pearce SR, Ellis THN. Retrotransposon-based insertion polymorphisms (RBIP) for high throughput marker analysis. Plant J. 1998;16:643–50.
169. Kalendar R, Antonius K, Smýkal P, Schulman AH. iPBS: a universal method for DNA fingerprinting and retrotransposon isolation. Theor. Appl. Genet. 2010;121:1419–30.

170. Manninen I, Schulman AH. *BARE-1*, a *copia*-like retroelement in barley (*Hordeum vulgare* L.). Plant Mol. Biol. 1993;22:829–46.
171. Suoniemi A, Anamthawat-jnsson K, Arna T, Schulman AH. Retrotransposon *BARE*-1 is a major , dispersed component of the barley (*Hordeum vulgare* L.) genome. Plant Mol. Biol. 1996;30:1321–9.
172. Jääskeläinen M, Mykkänen A-H, Arna T, Vicient CM, Suoniemi A, Kalendar R, et al. Retrotransposon *BARE-1*: expression of encoded proteins and formation of virus-like particles in barley cells. Plant J. 1999;20:413–22.
173. Soleimani VD, Baum BR, Johnson DA. Quantification of the retrotransposon *BARE-1* reveals the dynamic nature of the barley genome. Genome. 2006;396:389–96.
174. Soleimani VD, Baum BR, Johnson DA. Genetic diversity among barley cultivars assessed by sequence-specific amplification polymorphism. Theor. Appl. Genet. 2005;110:1290–300.
175. Pagnotta MA, Mondini L, Porceddu E. Quantification and organization of *WIS2-1A* and *BARE-1* retrotransposons in different genomes of Triticum and Aegilops species. Mol. Genet. Genomics. 2009;282:245–55.
176. Petit M, Lim KY, Julio E, Poncet C, Dorlhac de Borne F, Kovarik A, et al. Differential impact of retrotransposon populations on the genome of allotetraploid tobacco (*Nicotiana tabacum*). Mol. Genet. Genomics. 2007;278:1–15.
177. Syed NH, Sureshsundar S, Wilkinson MJ, Bhau BS, Cavalcanti JJ V, Flavell AJ. Ty1-copia retrotransposon-based SSAP marker development in cashew (*Anacardium occidentale* L.). Theor. Appl. Genet. 2005;110:1195–202.
178. Tam SM, Mhiri C, Vogelaar A, Kerkveld M, Pearce SR, Grandbastien M-A. Comparative analyses of genetic diversities within tomato and pepper collections detected by retrotransposon-based SSAP, AFLP and SSR. Theor. Appl. Genet. 2005;110:819–31.
179. Acquadro A, Portis E, Moglia A, Magurno F, Lanteri S. Retrotransposon-based S-SAP as a platform for the analysis of genetic variation and linkage in globe artichoke. Genome. 2006;1159:1149–59.
180. Biswas MK, Chai L, Amar MH, Zhang X, Deng X xin. Comparative analysis of genetic diversity in *Citrus* germplasm collection using AFLP, SSAP, SAMPL and SSR markers. Sci. Hortic. (Amsterdam). 2011;129:798–803.
181. He P, Ma Y, Dai H, Li L, Liu Y, Li H, et al. Development of Ty1-copia retrotransposon-

based S-SAP markers in strawberry (*Fragaria x ananassa* Duch.). Sci. Hortic. (Amsterdam). 2012;137:43–8.
182. Sanz AM, Gonzalez SG, Syed NH, Suso MJ, Saldaña CC, Flavell AJ. Genetic diversity analysis in *Vicia* species using retrotransposon-based SSAP markers. Mol. Genet. Genomics. 2007;278:433–41.
183. V. Melnikova N, V. Kudryavtseva A, S. Speranskaya A, A. Krinitsina A, A. Dmitriev A, S. Belenikin M, et al. The *FaRE*1 LTR-retrotransposon based SSAP markers reveal genetic polymorphism of strawberry (*Fragaria x ananassa*) cultivars. J. Agric. Sci. 2012;4:111–8.
184. Bousios A, Saldana-Oyarzabal I, Valenzuela-Zapata AG, Wood C, Pearce SR. Isolation and characterization of Ty1-*copia* retrotransposon sequences in the blue agave (*Agave tequilana* Weber var. azul) and their development as SSAP markers for phylogenetic analysis. Plant Sci. 2007;172:291–8.
185. Jiang S, Zheng X, Yu P, Yue X, Ahmed M, Cai D, et al. Primitive genepools of Asian pears and their complex hybrid origins inferred from fluorescent sequence-specific amplification polymorphism (SSAP) markers based on LTR retrotransposons. Moriguchi T, editor. PLoS One. 2016;11:e0149192.
186. Chee HT, Siang HT, Chai LH, Faridah QZ, Othman YR, Heslop-Harrison JS, et al. Genome constitution and classification using retrotransposon-based markers in the orphan crop banana. J. Plant Biol. 2005;48:96–105.
187. Branco CJS, Vieira EA, Malone G, Kopp MM, Malone E, Bernardes A, et al. IRAP and REMAP assessments of genetic similarity in rice. J. Appl. Genet. 2007;48:107–13.
188. Vukich M, Schulman AH, Giordani T, Natali L, Kalendar R, Cavallini A. Genetic variability in sunflower (*Helianthus annuus* L.) and in the *Helianthus* genus as assessed by retrotransposon-based molecular markers. Theor. Appl. Genet. 2009;119:1027–38.
189. Smykal P, Bacova-Kerteszova N, Kalendar R, Corander J, Schulman AH, Pavelek M. Genetic diversity of cultivated flax (*Linum usitatissimum* L.) germplasm assessed by retrotransposon-based markers. Theor. Appl. Genet. 2011;122:1385–97.
190. Marcon HS, Domingues DS, Silva JC, Borges RJ, Matioli FF, Fontes MR de M, et al. Transcriptionally active LTR retrotransposons in *Eucalyptus* genus are differentially expressed and insertionally polymorphic. BMC Plant Biol. 2015;15:198.
191. Lee S Il, Kim JH, Park KC, Kim NS. LTR-retrotransposons and inter-retrotransposon

amplified polymorphism (IRAP) analysis in *Lilium* species. Genetica. 2015;143:343–52.
192. El Baidouri M, Carpentier MC, Cooke R, Gao D, Lasserre E, Llauro C, et al. Widespread and frequent horizontal transfers of transposable elements in plants. Genome Res. 2014;24:831–8.
193. Jing R, Knox MR, Lee JM, Vershinin A V., Ambrose M, Ellis THN, et al. Insertional polymorphism and antiquity of *PDR1* retrotransposon insertions in *Pisum* species. Genetics. 2005;171:741–52.
194. Jing R, Vershinin A, Grzebyta J, Shaw P, Smykal P, Marshall D, et al. The genetic diversity and evolution of field pea (*Pisum*) studied by high throughput retrotransposon based insertion polymorphism (RBIP) marker analysis. BMC Evol. Biol. 2010;10:44.
195. Vitte C, Ishii T, Lamy F, Brar D, Panaud O. Genomic paleontology provides evidence for two distinct origins of Asian rice (*Oryza sativa* L.). Mol. Genet. Genomics. 2004;272:504–11.
196. Huang X, Lu G, Zhao Q, Liu X, Han B. Genome-wide analysis of transposon insertion polymorphisms reveals intraspecific variation in cultivated rice. Plant Physiol. 2008;148:25–40.
197. Petit J, Bourgeois E, Stenger W, Bès M, Droc G, Meynard D, et al. Diversity of the Ty-1 copia retrotransposon Tos17 in rice (Oryza sativa L.) and the AA genome of the Oryza genus. Mol. Genet. Genomics. 2009;282:633–52.
198. Kim H, Terakami S, Nishitani C, Kurita K, Kanamori H, Katayose Y, et al. Development of cultivar-specific DNA markers based on retrotransposon-based insertional polymorphism in Japanese pear. Breed Sci. 2012;62:53–62.
199. Jiang S, Zong Y, Yue X, Postman J, Teng Y, Cai D. Prediction of retrotransposons and assessment of genetic variability based on developed retrotransposon-based insertion polymorphism (RBIP) markers in *Pyrus* L. Mol. Genet. Genomics. 2014;290:225–37.
200. Guo DL, Guo MX, Hou XG, Zhang GH. Molecular diversity analysis of grape varieties based on iPBS markers. Biochem. Syst. Ecol. 2014;52:27–32.
201. Baloch FS, Alsaleh A, de Miera LES, Hatipoğlu R, Çiftçi V, Karaköy T, et al. DNA based iPBS-retrotransposon markers for investigating the population structure of pea (*Pisum sativum*) germplasm from Turkey. Biochem. Syst. Ecol. 2015;61:244–52.
202. Bretó MP, Ruiz C, Pina J a, Asíns MJ. The diversification of *Citrus clementina* Hort. ex Tan., a vegetatively propagated crop species. Mol. Phylogenet. Evol. 2001;21:285–93.
203. Biswas MK, Xu Q, Deng X xin. Utility of RAPD, ISSR, IRAP and REMAP markers for the

genetic analysis of *Citrus* spp. Sci. Hortic. (Amsterdam). 2010;124:254–61.
204. Hamon P, Duroy P-O, Dubreuil-Tranchant C, Mafra D'Almeida Costa P, Duret C, Razafinarivo NJ, et al. Two novel Ty1-*copia* retrotransposons isolated from coffee trees can effectively reveal evolutionary relationships in the *Coffea* genus (Rubiaceae). Mol. Genet. Genomics. 2011;285:447–60.
205. Smýkal P, Horáček J, Dostálová R, Hýbl M. Variety discrimination in pea (*Pisum sativum* L.) by molecular, biochemical and morphological markers. J. Appl. Genet. 2008;49:155–66.
206. Smýkal P, Hýbl M, Corander J, Jarkovský J, Flavell AJ, Griga M. Genetic diversity and population structure of pea (*Pisum sativum* L.) varieties derived from combined retrotransposon, microsatellite and morphological marker analysis. Theor. Appl. Genet. 2008;117:413–24.
207. Burstin J, Salloignon P, Martinello M, Magnin-Robert J-B, Siol M, Jacquin F, et al. Genetic diversity and trait genomic prediction in a pea diversity panel. BMC Genomics. 2015;16:1–17.
208. Nakatsuka T, Yamada E, Saito M, Hikage T, Ushiku Y, Nishihara M. Construction of the first genetic linkage map of Japanese gentian (Gentianaceae). BMC Genomics. 2012;13:672.
209. Andeden EE, Baloch FS, Derya M, Kilian B, Özkan H. iPBS-Retrotransposons-based genetic diversity and relationship among wild annual *Cicer* species. J. Plant Biochem. Biotechnol. 2013;22:453–66.
210. Metzker ML. Sequencing technologies - the next generation. Nat. Rev. Genet. 2010;11:31–46.
211. Witherspoon DJ, Xing J, Zhang Y, Watkins WS, Batzer MA, Jorde LB. Mobile element scanning (ME-Scan) by targeted high-throughput sequencing. BMC Genomics. 2010;11:410.
212. Sabot F, Picault N, El-Baidouri M, Llauro C, Chaparro C, Piegu B, et al. Transpositional landscape of the rice genome revealed by paired-end mapping of high-throughput re-sequencing data. Plant J. 2011;66:241–6.
213. Hormozdiari F, Hajirasouliha I, Dao P, Hach F, Yorukoglu D, Alkan C, et al. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. Bioinformatics. 2010;26:i350–7.
214. Fiston-Lavier A-S, Carrigan M, Petrov DA, González J. T-lex: a program for fast and accurate assessment of transposable element presence using next-generation sequencing data. Nucleic Acids Res. 2011;39:e36.
215. Sveinbjörnsson JI, Halldórsson B V. PAIR: polymorphic Alu insertion recognition. BMC

Bioinformatics. 2012;13 Suppl 6:S7.
216. Robb SMC, Lu L, Valencia E, Burnette JM, Okumoto Y, Wessler SR, et al. The use of RelocaTE and unassembled short reads to produce high-resolution snapshots of transposable element generated diversity in rice. G3. 2013;3:949–57.
217. Jiang C, Chen C, Huang Z, Liu R, Verdier J. ITIS, a bioinformatics tool for accurate identification of transposon insertion sites using next-generation sequencing data. BMC Bioinformatics. 2015;16:72.
218. Galindo González L, Deyholos MK. Identification, characterization and distribution of transposable elements in the flax (*Linum usitatissimum* L.) genome. BMC Genomics. 2012;13:644.
219. Grandbastien M-A. Activation of plant retrotransposons under stress conditions. Trends Plant Sci. 1998;3:181–7.
220. Kostyn K, Czemplik M, Kulma A, Bortniczuk M, Skała J, Szopa J. Genes of phenylpropanoid pathway are activated in early response to *Fusarium* attack in flax plants. Plant Sci. Elsevier Ireland Ltd; 2012;190:103–15.
221. Wojtasik W, Kulma A, Namysł K, Preisner M, Szopa J. Polyamine metabolism in flax in response to treatment with pathogenic and non–pathogenic *Fusarium* strains. Front. Plant Sci. 2015;6:1–12.
222. Wojtasik W, Kulma A, Dymi L, Hanuza J, Czemplik M, Szopa J. Evaluation of the significance of cell wall polymers in flax infected with a pathogenic strain of *Fusarium oxysporum*. BMC Plant Biol. 2016;16:75.
223. Olivain C, Trouvelot S, Binet M, Cordier C, Pugin A, Alabouvette C. Colonization of flax roots and early physiological responses of flax cells inoculated with pathogenic and nonpathogenic strains of *Fusarium oxysporum*. Appl. Environ. Microbiol. 2003;69:5453–62.
224. Hano C, Addi M, Fliniaux O, Bensaddek L, Duverger E, Mesnard F, et al. Molecular characterization of cell death induced by a compatible interaction between *Fusarium oxysporum* f. sp. *linii* and flax (*Linum usitatissimum*) cells. Plant Physiol. Biochem. 2008;46:590–600.
225. Carrier G, Le Cunff L, Dereeper A, Legrand D, Sabot F, Bouchez O, et al. Transposable elements are a major cause of somatic polymorphism in *Vitis vinifera* L. PLoS One. 2012;7:e32973.
226. Muir A, Wescott N (Eds). Flax, the genus *Linum*. Muir A, Wescott N, editors. London:

Taylor & Francis group; 2003.
227. Kalendar R, Flavell AJ, Ellis THN, Sjakste T, Moisy C, Schulman AH. Analysis of plant diversity with retrotransposon-based molecular markers. Heredity (Edinb). 2010;106:520–30.
228. Huis R, Hawkins S, Neutelings G. Selection of reference genes for quantitative gene expression normalization in flax (*Linum usitatissimum* L.). BMC Plant Biol. 2010;10:71.
229. Novak P, Neumann P, Pech J, Steinhaisl J, Macas J. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics. 2013;29:792–3.
230. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics. 2010;26:680–2.
231. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22:4673–80.
232. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 1980;16:111–20.
233. Marchler-Bauer A, Derbyshire MK, Gonzalez NR, Lu S, Chitaz F, Geer JY, et al. CDD: NCBI's conserved domain database. Nucleic Acids Res. 2015;43:D222–6.
234. CDD: NCBI's conserved domain database [Internet]. Available from: http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi. Accessed 26 January 2016.
235. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. 2013;30:2725–9.
236. Primer3web [Internet]. Available from: http://bioinfo.ut.ee/primer3/. Accessed 10 December 2013.
237. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3--new capabilities and interfaces. Nucleic Acids Res. 2012;40:e115.
238. Phytozome [Internet]. Available from: http://phytozome.jgi.doe.gov/pz/portal.html. Accessed 13 June 2015.
239. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: A comparative platform for green plant genomics. Nucleic Acids Res. 2012;40:1178–86.
240. Ma J, Devos KM, Bennetzen JL. Analyses of LTR-Retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res. 2004;14:860–9.

241. SciencePrimer.com [Internet]. Available from: http://scienceprimer.com/copy-number-calculator-for-realtime-pcr. Accessed 5 August 2015.
242. GelAnalyzer.com [Internet]. Available from: http://www.gelanalyzer.com/. Accessed 21 August 2014.
243. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 2006;23:254–67.
244. Gene Ontology at TAIR [Internet]. Available from: https://www.arabidopsis.org/tools/bulk/go/index.jsp. Accessed 29 February 2016.
245. AGRIGO [Internet]. Available from: http://bioinfo.cau.edu.cn/agriGO/. Accessed 1 March 2016.
246. Du Z, Zhou X, Ling Y, Zhang Z, Su Z. agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Res. 2010;38:W64–70.
247. Pfaffl MW, Tichopad A, Prgomet C, Neuvians TP. Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper--Excel-based tool using pair-wise correlations. Biotechnol. Lett. 2004;26:509–15.
248. Vandesompele J, Preter K De, Poppe B, Roy N Van, Paepe A De. Accurate normalization of real-time quantitative RT -PCR data by geometric averaging of multiple internal control genes. Genome Biol. 2002;3:1–12.
249. ThaleMine [Internet]. Available from: https://apps.araport.org/thalemine/begin.do. Accessed 25 August 2015.
250. Krishnakumar V, Hanlon MR, Contrino S, Ferlanti ES, Karamycheva S, Kim M, et al. Araport: The Arabidopsis information portal. Nucleic Acids Res. 2015;43:D1003–9.
251. Naito K, Cho E, Yang G, Campbell MA, Yano K, Okumoto Y, et al. Dramatic amplification of a rice transposable element during recent domestication. Proc. Natl. Acad. Sci. USA. 2006;103:17620–5.
252. Mascagni F, Barghini E, Giordani T, Rieseberg LH, Cavallini A, Natali L. Repetitive DNA and plant domestication: variation in copy number and proximity to genes of LTR-retrotransposons among wild and cultivated sunflower (*Helianthus annuus*) genotypes. Genome Biol. Evol. 2015;7:3368–82.
253. Baucom RS, Estill JC, Chaparro C, Upshaw N, Jogi A, Westerman RP, et al. Exceptional diversity , non-random distribution , and rapid evolution of retroelements in the B73 maize

genome. PLoS Genet. 2009;5:e1000732.
254. Meyers BC, Tingey S V, Morgante M. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res. 2001;11:1660–76.
255. Diez CM, Meca E, Tenaillon MI, Gaut BS. Three groups of transposable elements with contrasting copy number dynamics and host responses in the maize (*Zea mays ssp. mays*) Genome. PLoS Genet. 2014;10:e1004298.
256. Allaby RG, Peterson GW, Merriwether DA, Fu Y-B. Evidence of the domestication history of flax (*Linum usitatissimum* L.) from genetic diversity of the sad2 locus. Theor. Appl. Genet. 2005;112:58–65.
257. Soto-Cerda BJ, Diederichsen A, Ragupathy R, Cloutier S. Genetic characterization of a core collection of flax (*Linum usitatissimum* L.) suitable for association mapping studies and evidence of divergent selection between fiber and linseed types. BMC Plant Biol. 2013;13:78.
258. Oliver KR, McComb JA, Greene WK. Transposable elements: Powerful contributors to angiosperm evolution and diversity. Genome Biol. Evol. 2013;5:1886–901.
259. Volff J-N. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. Bioessays. 2006;28:913–22.
260. White SE, Habera LF, Wessler SW. Retrotransposons in the flanking regions of normal plant genes: A role for *copia*-like elements in the evolution of gene structure and expression. Proc. Natl. Acad. Sci. USA. 1994;91:11792–6.
261. Lockton S, Gaut BS. The Contribution of transposable elements to expressed coding sequence in *Arabidopsis thaliana*. J. Mol. Evol. 2009;68:80–9.
262. Zerjal T, Rousselet A, Mhiri C, Combes V, Madur D, Grandbastien MA, et al. Maize genetic diversity and association mapping using transposable element insertion polymorphisms. Theor. Appl. Genet. 2012;124:1521–37.
263. Smýkal P. Development of an efficient retrotransposon-based fingerprinting method for rapid pea variety identification. J. Appl. Genet. 2006;47:221–30.
264. Hano C, Martin I, Fliniaux O, Legrand B, Gutierrez L, Arroo RRJ, et al. Pinoresinol–lariciresinol reductase gene expression and secoisolariciresinol diglucoside accumulation in developing flax (*Linum usitatissimum*) seeds. Planta. 2006;224:1291–301.
265. Renouard S, Tribalatc M-A, Lamblin F, Mongelard G, Fliniaux O, Corbin C, et al. RNAi-mediated pinoresinol lariciresinol reductase gene silencing in flax (*Linum usitatissimum* L.) seed

coat: Consequences on lignans and neolignans accumulation. J. Plant Physiol. 2014;171:1372–7.
266. Westcott ND, Muir AD. Flax seed lignan in disease prevention and health promotion. Phytochem. Rev. 2003;2:401–17.
267. Fujita M, Gang DR, Davin LB, Lewis NG. Recombinant pinoresinol-lariciresinol reductases from western red cedar (*Thuja plicata*) catalyze opposite enantiospecific conversions. J. Biol. Chem. 1999;274:618–27.
268. Hemmati S, Von Heimendahl CBI, Klaes M, Alfermann AW, Schmidt TJ, Fuss E. Pinoresinol-lariciresinol reductases with opposite enantiospecificity determine the enantiomeric composition of lignans in the different organs of *Linum usitatissimum* L. Planta Med. 2010;76:928–34.
269. Zhao Q, Nakashima J, Chen F, Yin Y, Fu C, Yun J, et al. LACCASE is necessary and nonredundant with PEROXIDASE for lignin polymerization during vascular development in *Arabidopsis*. Plant Cell. 2013;25:3976–87.
270. Huis R, Morreel K, Fliniaux O, Lucau-Danila A, Fenart S, Grec S, et al. Natural hypolignification is associated with extensive oligolignol accumulation in flax stems. Plant Physiol. 2012;158:1893–915.
271. Chantreau M, Portelette A, Dauwe R, Kiyoto S, Crônier D, Morreel K, et al. Ectopic lignification in the flax lignified bast fiber1 mutant stem is associated with tissue-specific modifications in gene expression and cell wall composition. Plant Cell. 2014;26:4462–82.
272. Zhao Y, Hasenstein KH. Primary root growth regulation: the role of auxin and ethylene antagonists. J. Plant Growth Regul. 2009;28:309–20.
273. Mukhtar MS, Deslandes L, Auriac M-C, Marco Y, Somssich IE. The Arabidopsis transcription factor WRKY27 influences wilt disease symptom development caused by *Ralstonia solanacearum*. Plant J. 2008;56:935–47.
274. Kobayashi S, Goto-Yamamoto N, Hirochika H. Retrotransposon-induced mutations in grape skin color. Science. 2004;304:982.
275. Chu C-G, Tan CT, Yu G-T, Zhong S, Xu SS, Yan L. A novel retrotransposon inserted in the dominant Vrn-B1 allele confers spring growth habit in tetraploid wheat (*Triticum turgidum* L.). G3. 2011;1:637–45.
276. Almeida R, Allshire RC. RNA silencing and genome regulation. Trends Cell Biol. 2005;15:251–8.

277. Saze H, Kakutani T. Heritable epigenetic mutation of a transposon-flanked *Arabidopsis* gene due to lack of the chromatin-remodeling factor DDM1. EMBO J. 2007;26:3641–52.
278. Wang Q, Dooner HK. Remarkable variation in maize genome structure inferred from haplotype diversity at the bz locus. Proc. Natl. Acad. Sci. USA. 2006;103:17644–9.
279. Barret P, Brinkman M, Beckert M. A sequence related to rice *Pong* transposable element displays transcriptional activation by in vitro culture and reveals somaclonal variations in maize. Genome. 2006;1407:1399–407.
280. Miyao A, Nakagome M, Ohnuma T, Yamagata H, Kanamori H, Katayose Y, et al. Molecular spectrum of somaclonal variation in regenerated rice revealed by whole-genome sequencing. Plant Cell Physiol. 2012;53:256–64.
281. Kikuchi K, Terauchi K, Wada M. The plant MITE *mPing* is mobilized in anther culture. Nature. 2003;421:167–70.
282. Kubis SE, Castilho AMMF, Vershinin A V, Seymour J, Heslop-harrison P. Retroelements , transposons and methylation status in the genome of oil palm (*Elaeis guineensis*) and the relationship to somaclonal variation. 2003;69–79.
283. Ansari KI, Walter S, Brennan JM, Lemmens M, Kessans S, McGahern A, et al. Retrotransposon and gene activation in wheat in response to mycotoxigenic and non-mycotoxigenic-associated *Fusarium* stress. Theor. Appl. Genet. 2007;114:927–37.
284. Chang W-C, Lee T-Y, Huang H-D, Huang H-Y, Pan R-L. PlantPAN: Plant promoter analysis navigator, for identifying combinatorial cis-regulatory elements with distance constraint in plant gene groups. BMC Genomics. 2008;9:561.
285. Chow C-N, Zheng H-Q, Wu N-Y, Chien C-H, Huang H-D, Lee T-Y, et al. PlantPAN 2.0: an update of plant promoter analysis navigator for reconstructing transcriptional regulatory networks in plants. Nucleic Acids Res. 2016;44:D1154-60.
286. Hong JC. General aspects fo plant transcription factor families. In: Gonzalez DH, editor. Plant Transcr. factors Evol. Struct. Funct. Asp. London: Elsevier Inc.; 2016. p. 35–56.
287. Jalali BL, Bhargava S, Kamble A. Signal transduction and transcriptional regulation of plant defence responses. J. Phytopathol. 2006;154:65–74.
288. Passarinho PA, de Vries SC. Arabidopsis chitinases: a genomic survey. Arab. B. The American Society of Plant Biologists; 2002. p. 1–25.
289. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time

quantitative PCR and the 2-ΔΔCT method. Methods. 2001;25:402–8.
290. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. 1995;57:289–300.
291. Yanagisawa S. The Dof family of plant transcription factors. Trends Plant Sci. 2002;7:555–60.
292. Desveaux D, Maréchal A, Brisson N. Whirly transcription factors: Defense gene regulation and beyond. Trends Plant Sci. 2005;10:95–102.
293. Grover A, Taylor P, Grover A. Plant chitinases: genetic diversity and physiological roles. CRC. Crit. Rev. Plant Sci. 2012;31:57–73.
294. Kasprzewska A. plant chitinases - regulation and function. Cell. Mol. Biol. Lett. 2003;8:809–24.
295. Collinge DB, Kragh KM, Mikkelsen JD, Nielsen KK, Rasmussen U, Vad K. Plant chitinases. Plant J. 1993;3:31–40.
296. Jordão do Amaral DO, Alves de Almeida, Clébia Maria Dos Santo Correia MT, de Menezes Lima VL, da Silva MV. Isolation and characterization of chitinase from tomato infected by *Fusarium oxysporum* f. sp. *lycopersici*. J. Phytopathol. 2012;160:741–4.
297. Wang J, Tian N, Huang X, Chen LY, Schläppi M, Xu ZQ. The tall fescue turf grass class I chitinase gene *FaChit1* is activated by fungal elicitors, dehydration, ethylene, and mechanical wounding. Plant Mol. Biol. Report. 2009;27:305–14.
298. Bravo JM, Campo S, Murillo I, Coca M, San Segundo B. Fungus- and wound-induced accumulation of mRNA containing a class II chitinase of the pathogenesis-related protein 4 (PR-4) family of maize. Plant Mol. Biol. 2003;52:745–59.
299. López RC, Gómez-Gómez L. Isolation of a new fungi and wound-induced chitinase class in corms of *Crocus sativus*. Plant Physiol. Biochem. 2009;47:426–34.
300. Mcfadden HG, Lawrence GJ, Dennis ES. Differential induction of chitinase activity in flax (*Linum usitatissimum*) in response to inoculation with virulent or avirulent strains of *Melampsora lini*, the cause of flax rust. Australas. plant Pathol. 2001;30:27–30.
301. Mhiri C, More J-B, Audeon C, Feraul M, Grandbastien M-A, Lucas H. Regulation of expression of the tobacco *Tnt1* retrotransposon in heterologous species following pathogen-related stresses. Plant J. 1996;9:409–19.
302. Van Verk MC, Gatz C, Linthorst HJM. Transcriptional Regulation of Plant Defense

Responses. Adv. Bot. Res. 2009;51:397–438.
303. Birkenbihl RP, Somssich IE. Transcriptional plant responses critical for resistance towards necrotrophic pathogens. Front. Plant Sci. 2011;2:76.
304. Li H, Chen S, Song A, Wang H, Fang W, Guan Z, et al. RNA-Seq derived identification of differential transcription in the chrysanthemum leaf following inoculation with *Alternaria tenuissima*. BMC Genomics. 2014;15:9.
305. Pandey SP, Somssich IE. The role of WRKY transcription factors in plant immunity. Plant Physiol. 2009;150:1648–55.
306. Makova KD, Hardison RC. The effects of chromatin organization on variation in mutation rates in the genome. Nat. Rev. Genet. 2015;16:213–23.
307. Rocheta M, Carvalho L, Viegas W, Morais-Cecílio L. *Corky*, a *gypsy*-like retrotransposon is differentially transcribed in Quercus suber tissues. BMC Res. Notes. 2012;5:432.
308. Vicient CM. Transcriptional activity of transposable elements in maize. BMC Genomics. 2010;11:601.
309. Jääskeläinen M, Chang W, Moisy C, Schulman AH. Retrotransposon *BARE* displays strong tissue-specific differences in expression. New Phytol. 2013;200:1000–8.
310. Neumann P, Požárková D, Macas J. Highly abundant pea LTR retrotransposon *Ogre* is constitutively transcribed and partially spliced. Plant Mol. Biol. 2003;53:399–410.
311. Cheng X, Zhang D, Cheng Z, Keller B, Ling H. A new family of Ty1-*copia*-like retrotransposons originated in the tomato genome by a recent horizontal transfer event. Genetics. 2009;181:1183–93.
312. Slotkin RK, Vaughn M, Tanurdžic M, Borges F, Becker JD, Feijó A, et al. Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. Cell. 2009;136:461–72.
313. de Araujo PG, Rossi M, de Jesus EM, Saccaro Jr NL, Kajihara D, Massa R, et al. Transcriptionally active transposable elements in recent hybrid sugarcane. Plant J. 2005;44:707–17.
314. Ma LJ, Geiser DM, Proctor RH, Rooney AP, O'Donnell K, Trail F, et al. *Fusarium* Pathogenomics. Annu. Rev. Microbiol. 2013;67:399–416.
315. Kroes GMLW, Loffler HJM, Parlevliet JE, Keizer LCP, Lange W. Interactions of *Fusarium oxysporum* f. sp. *lini*, the flax wilt pathogen, with flax and linseed. Plant Pathol. 1999;48:491–8.

316. Kroes GMLW, Sommers E, Lange W. Two in vitro assays to evaluate resistance in *Linum usitatissimum* to Fusarium wilt disease. Eur. J. Plant Pathol. 1998;104:561–8.
317. Kroes GMLW, Baayen RP, Lange W. Histology of root rot of flax seedlings (*Linum usitatissimum*) infected by *Fusarium oxysporum* f. sp. *lini*. Eur. J. Plant Pathol. 1998;104:725–36.
318. Wojtasik W, Kulma A, Kostyn K, Szopa J. The changes in pectin metabolism in flax infected with Fusarium. Plant Physiol. Biochem. 2011;49:862–72.
319. Hano C, Addi M, Bensaddek L, Crônier D, Baltora-Rosset S, Doussot J, et al. Differential accumulation of monolignol-derived compounds in elicited flax (*Linum usitatissimum*) cell suspension cultures. Planta. 2006;223:975–89.
320. Boba A, Kulma A, Kostyn K, Starzycki M, Starzycka E, Szopa J. The influence of carotenoid biosynthesis modification on the *Fusarium culmorum* and *Fusarium oxysporum* resistance in flax. Physiol. Mol. Plant Pathol. 2011;76:39–47.
321. Wróbel-Kwiatkowska M, Lorenc-Kukula K, Starzycki M, Oszmiański J, Kepczyńska E, Szopa J. Expression of β-1,3-glucanase in flax causes increased resistance to fungi. Physiol. Mol. Plant Pathol. 2004;65:245–56.
322. Lorenc-Kukuła K, Wróbel-Kwiatkowska M, Starzycki M, Szopa J. Engineering flax with increased flavonoid content and thus *Fusarium* resistance. Physiol. Mol. Plant Pathol. 2007;70:38–48.
323. Lorenc-Kukuła K, Zuk M, Kulma A, Czemplik M, Kostyn K, Skala J, et al. Engineering flax with the GT family 1 *Solanum sogarandinum* glycosyltransferase SsGT1 confers increased resistance to *Fusarium* infection. J. Agric. Food Chem. 2009;57:6698–705.
324. Rutkowska-Krause I, Mankowska G, Lukaszewicz M, Szopa J. Regeneration of flax (*Linum usitatissimum L.*) plants from anther culture and somatic tissue with increased resistance to *Fusarium oxysporum*. Plant Cell Rep. 2003;22:110–6.
325. Spielmeyer W, Green AG, Bittisnich D, Mendham N, Lagudah ES. Identification of quantitative trait loci contributing to Fusarium-wilt resistance on an AFLP linkage map of flax (*Linum usitatissimum*). Theor. Appl. Genet. 1998;97:633–41.
326. De Cremer K, Mathys J, Vos C, Froenicke L, Michelmore RW, Cammue BPA, et al. RNAseq-based transcriptome analysis of *Lactuca sativa* infected by the fungal necrotroph *Botrytis cinerea*. Plant. Cell Environ. 2013;36:1992–2007.

327. Howard BE, Hu Q, Babaoglu AC, Chandra M, Borghi M, Tan X, et al. High-throughput RNA sequencing of Pseudomonas-infected Arabidopsis reveals hidden transcriptome complexity and novel splice variants. Xing Y, editor. PLoS One. 2013;8:e74183.
328. Li C, Shao J, Wang Y, Li W, Guo D, Yan B, et al. Analysis of banana transcriptome and global gene expression profiles in banana roots in response to infection by race 1 and tropical race 4 of *Fusarium oxysporum* f. sp. *cubense*. BMC Genomics. 2013;14:851.
329. Li C, Deng G, Yang J, Viljoen A, Jin Y, Kuang R, et al. Transcriptome profiling of resistant and susceptible Cavendish banana roots following inoculation with *Fusarium oxysporum* f. sp. *cubense* tropical race 4. BMC Genomics. BMC Genomics; 2012;13:374.
330. Wang Z, Zhang J, Jia C, Liu J, Li Y, Yin X, et al. De novo characterization of the banana root transcriptome and analysis of gene expression under *Fusarium oxysporum* f. sp. *cubense* tropical race 4 infection. BMC Genomics. 2012;13:650.
331. Xiao J, Jin X, Jia X, Wang H, Cao A, Zhao W, et al. Transcriptome-based discovery of pathways and genes related to resistance against *Fusarium* head blight in wheat landrace Wangshuibai. BMC Genomics. BMC Genomics; 2013;14:197.
332. Socquet-Juglard D, Kamber T, Pothier JF, Christen D, Gessler C, Duffy B, et al. Comparative RNA-Seq analysis of early-infected peach leaves by the invasive phytopathogen *Xanthomonas arboricola* pv. *pruni*. PLoS One. 2013;8:e54196.
333. Rowland GG, Hormis YA, Rashid KY. CDC Bethune flax. Can. J. Plant Sci. 2002;82:101–2.
334. Diederichsen A, Rozhmina TA, Kudrjavceva LP. Variation patterns within 153 flax (*Linum usitatissimum* L.) genebank accessions based on evaluation for resistance to fusarium wilt, anthracnose and pasmo. Plant Genet. Resour. 2008;6:22–32.
335. Leslie JF, Summerell BA. The Fusarium laboratory manual. 1st editio. Ames, Iowa: Blackwell publishing; 2006.
336. Mokshina N, Gorshkova T, Deyholos MK. Chitinase-like (*CTL*) and cellulose synthase (*CESA*) gene expression in gelatinous-type cellulosic walls of flax (*Linum usitatissimum* L.) bast fibers. PLoS One. 2014;9:e97949.
337. Trapnell C, Pachter L, Salzberg SL. TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25:1105–11.
338. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. Blast2GO: a universal

tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005;21:3674–6.
339. Yi X, Du Z, Su Z. PlantGSEA: a gene set enrichment analysis toolkit for plant community. Nucleic Acids Res. 2013;41:W98–103.
340. Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, et al. TM4: a free, open-source system for microarray data management and analysis. Biotehcniques. 2003;34:374–8.
341. Nessler CL, Allen RD, Galewsky S. Identification and characterization of latex-specific proteins in opium poppy. Plant Physiol. 1985;79:499–504.
342. Lawrence CB, Singh NP, Qiu J, Gardner RG, Tuzun S. Constitutive hydrolytic enzymes are associated with polygenic resistance of tomato to *Alternaria solani* and may function as an elicitor release mechanism. Physiol. Mol. Plant Pathol. 2000;57:211–20.
343. Ascencio-Ibáñez JT, Sozzani R, Lee T-J, Chu T-M, Wolfinger RD, Cella R, et al. Global analysis of Arabidopsis gene expression uncovers a complex array of changes impacting pathogen response and cell cycle during geminivirus infection. Plant Physiol. 2008;148:436–54.
344. Nongbri PL, Johnson JM, Sherameti I, Glawischnig E, Halkier BA, Oelmüller R. Indole-3-acetaldoxime-derived compounds restrict root colonization in the beneficial interaction between *Arabidopsis* roots and the endophyte *Piriformospora indica*. Mol. Plant-Microbe Interact. 2012;25:1186–97.
345. Rudd J, Kanyuka K, Hassani-Pak K, Derbyshire M, Andongabo A, Devonshire J, et al. Transcriptome and metabolite profiling the infection cycle of Zymoseptoria tritici on wheat (Triticum aestivum) reveals a biphasic interaction with plant immunity involving differential pathogen chromosomal contributions and a variation on the hemibiotrop. Plant Physiol. 2015;167:1158–85.
346. Vidhyasekaran P. Disease resistance and susceptibility genes in signal perception and emission. Fungal Pathog. plants Crop. Mol. Biol. host Def. Mech. 2nd ed. Boca Raton, FL: CRC press, Taylor and Francis Group; 2008.
347. Liu J, Elmore JM, Lin ZJD, Coaker G. A receptor-like cytoplasmic kinase phosphorylates the host target RIN4, leading to the activation of a plant innate immune receptor. Cell Host Microbe. 2011;9:137–46.
348. Zhu QH, Fan L, Liu Y, Xu H, Llewellyn D, Wilson I. miR482 regulation of NBS-LRR defense genes during fungal pathogen infection in cotton. PLoS One. 2013;8.

349. Li F, Pignatta D, Bendix C, Brunkard JO, Cohn MM, Tung J, et al. MicroRNA regulation of plant innate immune receptors. Proc. Natl. Acad. Sci. U. S. A. 2011;109:1790–5.
350. Zhai J, Jeong DH, de Paoli E, Park S, Rosen BD, Li Y, et al. MicroRNAs as master regulators of the plant NB-LRR defense gene family via the production of phased, trans-acting siRNAs. Genes Dev. 2011;25:2540–53.
351. Shivaprasad P V, Chen H-M, Patel K, Bond DM, Santos BACM, Baulcombe DC. A MicroRNA superfamily regulates nucleotide binding site-leucine-rich repeats and other mRNAs. Plant Cell. 2012;24:859–74.
352. Mueth NA, Ramachandran SR, Hulbert SH. Small RNAs from the wheat stripe rust fungus (*Puccinia striiformis* f. sp. *tritici*). BMC Genomics. 2015;16:718.
353. Wu W, Wu Y, Gao Y, Li M, Yin H, Lv M, et al. Somatic embryogenesis receptor-like kinase 5 in the ecotype Landsberg erecta of *Arabidopsis* is a functional RD LRR-RLK in regulating brassinosteroid signaling and cell death control. Front. Plant Sci. 2015;6:1–16.
354. Gao M, Wang X, Wang D, Xu F, Ding X, Zhang Z, et al. Regulation of cell death and innate immunity by two receptor-like kinases in *Arabidopsis*. Cell Host Microbe. Elsevier Ltd; 2009;6:34–44.
355. Liu J, Ding P, Sun T, Nitta Y, Dong O, Huang X, et al. Heterotrimeric G proteins serve as a converging point in plant defense signaling activated by multiple receptor-like kinases. Plant Physiol. 2013;161:2146–58.
356. Llorente F, Alonso-Blanco C, Sánchez-Rodriguez C, Jorda L, Molina A. ERECTA receptor-like kinase and heterotrimeric G protein from *Arabidopsis* are required for resistance to the necrotrophic fungus *Plectosphaerella cucumerina*. Plant J. 2005;43:165–80.
357. Trusov Y, Rookes J, Chakravorty D, Armour D. Heterotrimeric G proteins facilitate *Arabidopsis* resistance to necrotrophic pathogens and are involved in jasmonate signalling. Plant Physiol. 2006;140:210–20.
358. Trusov Y, Sewelam N, Rookes JE, Kunkel M, Nowak E, Schenk PM, et al. Heterotrimeric G proteins-mediated resistance to necrotrophic pathogens includes mechanisms independent of salicylic acid-, jasmonic acid/ethylene- and abscisic acid-mediated defense signaling. Plant J. 2009;58:69–81.
359. Vidhyasekaran P. Perception and transduction of pathogen signals in plants. Fungal Pathog. Plants Crop. 2nd ed. Boca Raton, FL: CRC press, Taylor and Francis Group; 2008.

360. Dangl JL, Jones JDG. Plant pathogens and integrated defence responses to infection. Nature. 2001;411:826–33.
361. Lecourieux D, Ranjeva R, Pugin A. Calcium in plant defence-signalling pathways. New Phytol. 2006;171:249–69.
362. Zhuang X, McPhee KE, Coram TE, Peever TL, Chilvers MI. Rapid transcriptome characterization and parsing of sequences in a non-model host-pathogen interaction; pea-*Sclerotinia sclerotiorum*. BMC Genomics. 2012;13:668.
363. Serrazina S, Santos C, Machado H, Pesquita C, Vicentini R, Pais MS, et al. *Castanea* root transcriptome in response to *Phytophthora cinnamomi challenge*. Tree Genet. Genomes. 2015;11:6.
364. Alkan N, Friedlander G, Ment D, Prusky D, Fluhr R. Simultaneous transcriptome analysis of *Colletotrichum gloeosporioides* and tomato fruit pathosystem reveals novel fungal pathogenicity and fruit defense strategies. New Phytol. 2015;205:801–15.
365. Raghavendra AS, Gonugunta VK, Christmann A, Grill E. ABA perception and signalling. Trends Plant Sci. 2010;15:395–401.
366. Wan J, Zhang X, Stacey G. Chitin signaling and plant disease resistance. Mol. Plant Pathol. 2008;6:831–3.
367. Rushton PJ, Somssich IE, Ringler P, Shen QJ. WRKY transcription factors. Trends Plant Sci. 2010;15:247–58.
368. Li J, Brader G, Kariola T, Tapio Palva E. WRKY70 modulates the selection of signaling pathways in plant defense. Plant J. 2006;46:477–91.
369. Yang B, Jiang Y, Rahman MH, Deyholos MK, Kav NN V. Identification and expression analysis of *WRKY* transcription factor genes in canola (*Brassica napus* L.) in response to fungal pathogens and hormone treatments. BMC Plant Biol. 2009;9:68.
370. Lai Z, Vinod K, Zheng Z, Fan B, Chen Z. Roles of *Arabidopsis* WRKY3 and WRKY4 transcription factors in plant responses to pathogens. BMC Plant Biol. 2008;8:68.
371. Ambawat S, Sharma P, Yadav NR, Yadav RC. MYB transcription factor genes as regulators for plant responses: An overview. Physiol. Mol. Biol. Plants. 2013;19:307–21.
372. Zhu QH, Stephen S, Kazan K, Jin G, Fan L, Taylor J, et al. Characterization of the defense transcriptome responsive to *Fusarium oxysporum*-infection in *Arabidopsis* using RNA-seq. Gene. 2013;512:259–66.

373. Gonzalez A, Zhao M, Leavitt JM, Lloyd AM. Regulation of the anthocyanin biosynthetic pathway by the TTG1/bHLH/Myb transcriptional complex in Arabidopsis seedlings. Plant J. 2008;53:814–27.
374. Mengiste T, Chen X, Salmeron J, Dietrich R. The *BOTRYTIS SUSCEPTIBLE1* gene encodes an R2R3MYB transcription factor protein that is required for biotic and abiotic stress responses in *Arabidopsis*. Plant Cell. 2003;15:2551–65.
375. Pieterse CMJ, Van der Does D, Zamioudis C, Leon-Reyes A, Van Wees SCM. Hormonal modulation of plant immunity. Annu. Rev. Cell Dev. Biol. 2012;28:489–521.
376. Fonseca S, Chico JM, Solano R. The jasmonate pathway: the ligand, the receptor and the core signalling module. Curr. Opin. Plant Biol. 2009;12:539–47.
377. Chini A, Fonseca S, Fernández G, Adie B, Chico JM, Lorenzo O, et al. The JAZ family of repressors is the missing link in jasmonate signalling. Nature. 2007;448:666–71.
378. Chung HS, Koo AJK, Gao X, Jayanty S, Thines B, Jones AD, et al. Regulation and function of Arabidopsis *JASMONATE ZIM*-domain genes in response to wounding and herbivory. Plant Physiol. 2008;146:952–64.
379. Kidd BN, Kadoo NY, Dombrecht B, Tekeoglu M, Gardiner DM, Thatcher LF, et al. Auxin signaling and transport promote susceptibility to the root-infecting fungal pathogen *Fusarium oxysporum* in *Arabidopsis*. Mol. Plant-Microbe Interact. 2011;24:733–48.
380. Widemann E, Miesch L, Lugan R, Holder E, Heinrich C, Aubert Y, et al. The amidohydrolases IAR3 and ILL6 contribute to jasmonoyl-isoleucine hormone turnover and generate 12-hydroxyjasmonic acid upon wounding in *Arabidopsis* leaves. J. Biol. Chem. 2013;288:31701–14.
381. Berrocal-Lobo M, Molina A. Ethylene response factor 1 mediates Arabidopsis resistance to the soilborne fungus *Fusarium oxysporum*. Mol. Plant-Microbe Interact. 2004;17:763–70.
382. Oñate-Sánchez L, Anderson JP, Young J, Singh KB. AtERF14, a member of the ERF family of transcription factors, plays a nonredundant role in plant defense. Plant Physiol. 2006;143:400–9.
383. Spartz AK, Lee SH, Wenger JP, Gonzalez N, Itoh H, Inzé D, et al. The *SAUR19* subfamily of *SMALL AUXIN UP RNA* genes promote cell expansion. Plant J. 2012;70:978–90.
384. Chen YC, Wong CL, Muzzi F, Vlaardingerbroek I, Kidd BN, Schenk PM. Root defense analysis against *Fusarium oxysporum* reveals new regulators to confer resistance. Sci. Rep.

2014;4:5584.
385. González-Lamothe R, Mitchell G, Gattuso M, Diarra MS, Malouin F, Bouarab K. Plant antimicrobial agents and their effects on plant and human pathogens. Int. J. Mol. Sci. 2009;10:3400–19.
386. Glawischnig E, Hansen BG, Olsen CE, Halkier BA. Camalexin is synthesized from indole-3-acetaldoxime, a key branching point between primary and secondary metabolism in *Arabidopsis*. Proc. Natl. Acad. Sci. U. S. A. 2004;101:8245–50.
387. Rampey RA, LeClere S, Kowalczyk M, Ljung K, Sandberg G, Bartel B. A family of auxin-conjugate hydrolases that contributes to free indole-3-acetic acid levels during *Arabidopsis* germination. Plant Physiol. 2004;135:978–88.
388. Schuller A, Ludwig-Müller J. A family of auxin conjugate hydrolases from Brassica rapa: Characterization and expression during clubroot disease. New Phytol. 2006;171:145–58.
389. Truman WM, Bennett MH, Turnbull CGN, Grant MR. *Arabidopsis* auxin mutants are compromised in systemic acquired resistance and exhibit aberrant accumulation of various indolic compounds. Plant Physiol. 2010;152:1562–73.
390. Zhang T, Poudel AN, Jewell JB, Kitaoka N, Staswick P, Matsuura H, et al. Hormone crosstalk in wound stress response : wound- inducible amidohydrolases can simultaneously regulate jasmonate and auxin homeostasis in *Arabidopsis thaliana*. J. Exp. Bot. 2016;67:2107–20.
391. Linthorst HJM, Van Loon LC. Pathogenesis-related proteins of plants. CRC. Crit. Rev. Plant Sci. 1991;10:123–50.
392. Kitajima S, Sato F. Plant pathogenesis-related proteins: molecular mechanisms of gene expression and protein function. J. Biochem. 1999;125:1–8.
393. Sels J, Mathys J, De Coninck BMA, Cammue BPA, De Bolle MFC. Plant pathogenesis-related (PR) proteins: a focus on PR peptides. Plant Physiol. Biochem. 2008;46:941–50.
394. Ahmed NU, Park JI, Jung HJ, Kang KK, Hur Y, Lim YP, et al. Molecular characterization of stress resistance-related chitinase genes of *Brassica rapa*. Plant Physiol. Biochem. 2012;58:106–15.
395. Kong L, Anderson JM, Ohm HW. Induction of wheat defense and stress-related genes in response to *Fusarium graminearum*. Genome. 2005;48:29–40.
396. Teixeira PJPL, Thomazella DPDT, Reis O, do Prado PFV, do Rio MCS, Fiorin GL, et al.

High-resolution transcript profiling of the atypical biotrophic interaction between *Theobroma cacao* and the fungal pathogen *Moniliophthora perniciosa*. Plant Cell. 2014;26:4245–69.
397. Kawahara Y, Oono Y, Kanamori H, Matsumoto T, Itoh T, Minami E. Simultaneous RNA-seq analysis of a mixed transcriptome of rice and blast fungus interaction. PLoS One. 2012;7:e49423.
398. Windram O, Madhou P, McHattie S, Hill C, Hickman R, Cooke E, et al. *Arabidopsis* defense against *Botrytis cinerea*: chronology and regulation deciphered by high-resolution temporal transcriptomic analysis. Plant Cell. 2012;24:3530–57.
399. Wang X, Tang C, Deng L, Cai G, Liu X, Liu B, et al. Characterization of a pathogenesis-related thaumatin-like protein gene *TaPR5* from wheat induced by stripe rust fungus. Physiol. Plant. 2010;139:27–38.
400. Curto M, Krajinski F, Küster H, Rubiales D. Plant defense responses in *Medicago truncatula* unveiled by microarray analysis. Plant Mol. Biol. Report. 2015;33:569–83.
401. Lowe RGT, Cassin A, Grandaubert J, Clark BL, Van De Wouw AP, Rouxel T, et al. Genomes and transcriptomes of partners in plant-fungal- interactions between canola (*Brassica napus*) and two *Leptosphaeria* species. PLoS One. 2014;9:e103098.
402. Kirubakaran SI, Begum SM, Ulaganathan K, Sakthivel N. Characterization of a new antifungal lipid transfer protein from wheat. Plant Physiol. Biochem. 2008;46:918–27.
403. Vigers AJ. A new family of plant antifungal proteins. Mol. Plant-Microbe Interact. 1991;4:315.
404. Vigers AJ, Wiedemann S, Roberts WK, Legrand M, Selitrennikoff CP, Fritig B. Thaumatin-like pathogenesis-related proteins are antifungal. Plant Sci. 1992;83:155–61.
405. Kader J-C. Lipid transfer proteins in plants. Annu. Rev. Plant Physiol. Plant Mol. Biol. 1996;47:627–54.
406. Habib H, Fazili KM. Plant protease inhibitors : a defense strategy in plants. Biotechnol. Mol. Biol. Rev. 2007;2:68–85.
407. Torres MA, Jones JDG, Dangl JL. Reactive oxygen species signaling in response to pathogens. Plant Physiol. 2006;141:373–8.
408. Swarupa V, Ravishankar K V., Rekha A. Plant defense response against *Fusarium oxysporum* and strategies to develop tolerant genotypes in banana. Planta. 2014;239:735–51.
409. Apel K, Hirt H. REACTIVE OXYGEN SPECIES: metabolism, oxidative stress, and signal

transduction. Annu. Rev. Plant Biol. 2004;55:373–99.
410. Fernández-Pérez F, Pomar F, Pedreño MA, Novo-Uzal E. The suppression of *AtPrx52* affects fibers but not xylem lignification in *Arabidopsis* by altering the proportion of syringyl units. Physiol. Plant. 2015;154:395–406.
411. Floerl S, Majcherczyk A, Possienke M, Feussner K, Tappe H, Gatz C, et al. *Verticillium longisporum* infection affects the leaf apoplastic proteome, metabolome, and cell wall properties in *Arabidopsis thaliana*. PLoS One. 2012;7:e31435.
412. Edwards R, Dixon DP. Plant glutathione transferases. Methods Enzymol. 2005;401:169–86.
413. Liao W, Ji L, Wang J, Chen Z, Ye M, Ma H, et al. Identification of glutathione S-transferase genes responding to pathogen infestation in *Populus tomentosa*. Funct. Integr. Genomics. 2014;14:517–29.
414. Fode B, Siemsen T, Thurow C, Weigel R, Gatz C. The *Arabidopsis* GRAS protein SCL14 interacts with class II TGA transcription factors and is essential for the activation of stress-inducible promoters. Plant Cell. 2008;20:3122–35.
415. Pusztahelyi T, Holb IJ, Pócsi I. Secondary metabolites in fungus-plant interactions. Front. Plant Sci. 2015;6:1–23.
416. Ferrer JL, Austin MB, Stewart C, Noel JP. Structure and function of enzymes involved in the biosynthesis of phenylpropanoids. Plant Physiol. Biochem. 2008;46:356–70.
417. Throll D. Biosynthesis and biological functions of terpenoids in plants. Adv. Biochem. Eng. Biotechnol. 2015;148:63–106.
418. Nisar N, Li L, Lu S, Khin NC, Pogson BJ. Carotenoid metabolism in plants. Mol. Plant. 2015;8:68–82.
419. Stahl W, Sies H. Antioxidant activity of carotenoids. Mol. Aspects Med. 2003;24:345–51.
420. Mikkelsen MD, Naur P, Halkier BA. *Arabidopsis* mutants in the C-S lyase of glucosinolate biosynthesis establish a critical role for indole-3-acetaldoxime in auxin homeostasis. Plant J. 2004;37:770–7.
421. Bak S, Tax FE, Feldmann A, Galbraith DW, Feyereisen R. CYP83B1, a cytochrome P450 at the metabolic branch point in auxin and indole glucosinolate biosynthesis in Arabidopsis. Plant Cell. 2001;13:101–11.
422. Schuhegger R, Nafisi M, Mansourova M, Petersen BL, Olsen CE, Svatos A, et al. CYP71B15 (PAD3) catalyzes the final step in camalexin biosynthesis. Plant Physiol.

2006;141:1248–54.
423. Sellam A, Dongo A, Guillemette T, Hudhomme P, Simoneau P. Transcriptional responses to exposure to the brassicaceous defence metabolites camalexin and allyl-isothiocyanate in the necrotrophic fungus *Alternaria brassicicola*. Mol. Plant Pathol. 2007;8:195–208.
424. Iven T, König S, Singh S, Braus-Stromeyer SA, Bischoff M, Tietze LF, et al. Transcriptional activation and production of tryptophan-derived secondary metabolites in *Arabidopsis* roots contributes to the defense against the fungal vascular pathogen *Verticillium longisporum*. Mol. Plant. 2012;5:1389–402.
425. Eckardt NA. Move it on out with MATEs. Plant Cell. 2001;13:1477–80.
426. Kang J, Park J, Choi H, Burla B, Kretzschmar T, Lee Y, et al. Plant ABC transporters. Arab. B. The American Society of Plant Biologists; 2011. p. 1–25.
427. Yazaki K. ABC transporters involved in the transport of plant secondary metabolites. FEBS Lett. 2006;580:1183–91.
428. Kosaka A, Manickavelu A, Kajihara D, Nakagawa H, Ban T. Altered gene expression profiles of wheat genotypes against *Fusarium* head blight. Toxins (Basel). 2015;7:604–20.
429. Casassola A, Brammer SP, Chaves MS, Martinelli JA, Stefanato F, Boyd LA. Changes in gene expression profiles as they relate to the adult plant leaf rust resistance in the wheat cv. Toropi. Physiol. Mol. Plant Pathol. 2015;89:49–54.
430. Dynowski M, Schaaf G, Loque D, Moran O, Ludewig U. Plant plasma membrane water channels conduct the signalling molecule H2O2. Biochem. J. 2008;414:53–61.
431. Marella HH, Nielsen E, Schachtman DP, Taylor CG. The amino acid permeases AAP3 and AAP6 are involved in root-knot nematode parasitism of *Arabidopsis*. Mol. Plant-Microbe Interact. 2013;26:44–54.
432. Bellincampi D, Cervone F, Lionetti V. Plant cell wall dynamics and wall-related susceptibility in plant-pathogen interactions. Front. Plant Sci. 2014;5:1–8.
433. Malinovsky FG, Fangel JU, Willats WGT. The role of the cell wall in plant immunity. Front. Plant Sci. 2014;5:178.
434. Yang C-L, Liang S, Wang H-Y, Han L-B, Wang F-X, Cheng H-Q, et al. Cotton major latex protein 28 functions as a positive regulator of the ethylene responsive factor 6 in defense against *Verticillium dahliae*. Mol. Plant. 2015;8:399–411.
435. Wang Y, Yang L, Chen X, Ye T, Zhong B, Liu R, et al. *Major latex protein-like protein 43*

(*MLP43*) functions as a positive regulator during abscisic acid responses and confers drought tolerance in *Arabidopsis thaliana*. J. Exp. Bot. 2016;67:421–34.
436. Ruperti B, Bonghi C, Ziliotto F, Pagni S, Rasori A, Varotto S, et al. Characterization of a major latex protein (MLP) gene down-regulated by ethylene during peach fruitlet abscission. Plant Sci. 2002;163:265–72.
437. Kiirika LM, Schmitz U, Colditz F. The alternative *Medicago truncatula* defense proteome of ROS-defective transgenic roots during early microbial infection. Front. Plant Sci. 2014;5:341.
438. Schenk PM, Kazan K, Wilson I, Anderson JP, Richmond T, Somerville SC, et al. Coordinated plant defense responses in *Arabidopsis* revealed by microarray analysis. Proc. Natl. Acad. Sci. U. S. A. 2000;97:11655–60.
439. Osmark P, Boyle B, Brisson N. Sequential and structural homology between intracellular pathogenesis-related proteins and a group of latex proteins. Plant Mol. Biol. 1998;38:1243–6.
440. Stanley Kim H, Yu Y, Snesrud EC, Moy LP, Linford LD, Haas BJ, et al. Transcriptional divergence of the duplicated oxidative stress-responsive genes in the *Arabidopsis* genome. Plant J. 2004;41:212–20.
441. Cullis CA. Mechanisms and control of rapid genomic changes in flax. Ann. Bot. 2005;95:201–6.
442. Alonso JM, Ecker JR. Moving forward in reverse: genetic technologies to enable genome-wide phenomic screens in Arabidopsis. Nat. Rev. Genet. 2006;7:524–36.
443. Greene EA, Codomo CA, Taylor NE, Henikoff JG, Till BJ, Reynolds SH, et al. Spectrum of chemically induced mutations from a large-scale reverse-genetic screen in Arabidopsis. Genetics. 2003;740:731–40.
444. Botticella E, Sestili F, Hernandez-Lopez A, Phillips A, Lafiandra D. High resolution melting analysis for the detection of EMS induced mutations in wheat SBEIIa genes. BMC Plant Biol. 2011;11:156.
445. Cooper JL, Till BJ, Laport RG, Darlow MC, Kleffner JM, Jamai A, et al. TILLING to detect induced mutations in soybean. BMC Plant Biol. 2008;8:9.
446. Rowland GG. An EMS-induced low-linolenic-acid mutant in McGregor flax (*Linum usitatissimum* L .). Can. J. Plant Sci. 1991;71:393–6.
447. Till BJ, Cooper J, Tai TH, Colowit P, Greene EA, Henikoff S, et al. Discovery of chemically induced mutations in rice by TILLING. BMC Plant Biol. 2007;7:19.

448. Xin Z, Wang ML, Barkley NA, Burow G, Franks C, Pederson G, et al. Applying genotyping (TILLING) and phenotyping analyses to elucidate gene function in a chemically induced sorghum mutant population. BMC Plant Biol. 2008;8:103.
449. Okabe Y, Asamizu E, Saito T, Matsukura C, Ariizumi T, Brès C, et al. Tomato TILLING technology: development of a reverse genetics tool for the efficient isolation of mutants from Micro-Tom mutant libraries. Plant Cell Physiol. 2011;52:1994–2005.
450. Zhu Q, Smith SM, Ayele M, Yang L, Jogi A, Chaluvadi SR, et al. High-throughput discovery of mutations in tef semi-dwarfing genes by next-generation sequencing analysis. Genetics. 2012;192:819–29.
451. Kurowska M, Daszkowska-Golec A, Gruszka D, Marzec M, Szurman M, Szarejko I, et al. TILLING: a shortcut in functional genomics. J. Appl. Genet. 2011;52:371–90.
452. Østergaard L, Yanofsky MF. Establishing gene function by mutagenesis in *Arabidopsis thaliana*. Plant J. 2004;39:682–96.
453. McCallum CM, Comai L, Greene EA, Henikoff S. Targeting induced local lesions IN genomes (TILLING) for plant functional genomics. Plant Physiol. 2000;123:439–42.
454. Oleykowski CA, Bronson Mullins CR, Godwin AK, Yeung AT. Mutation detection using a novel plant endonuclease. Nucleic Acids Res. 1998;26:4597–602.
455. Gady ALF, Hermans FWK, Van de Wal MHBJ, van Loo EN, Visser RGF, Bachem CWB. Implementation of two high through-put techniques in a novel application: detecting point mutations in large EMS mutated plant populations. Plant Methods. 2009;5:13.
456. Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. J. Biomed. Biotechnol. 2012;2012:1–11.
457. Bräutigam A, Gowik U. What can next generation sequencing do for you? Next generation sequencing as a valuable tool in plant research. Plant Biol. 2010;12:831–41.
458. Deschamps S, Campbell MA. Utilization of next-generation sequencing platforms in plant genomics and genetic variant discovery. Mol. Breed. 2010;25:553–70.
459. Rigola D, van Oeveren J, Janssen A, Bonné A, Schneiders H, van der Poel HJA, et al. High-throughput detection of induced mutations and natural variation using KeyPoint technology. PLoS One. 2009;4:e4761.
460. Tsai H, Howell T, Nitcher R, Missirian V, Watson B, Ngo KJ, et al. Discovery of rare mutations in populations: TILLING by sequencing. Plant Physiol. 2011;156:1257–68.

461. Reddy TV, Dwivedi S, Sharma NK. Development of TILLING by sequencing platform towards enhanced leaf yield in tobacco. Ind. Crops Prod. Elsevier B.V.; 2012;40:324–35.
462. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 2012;13:341.
463. Hamilton JP, Buell CR. Advances in plant genome sequencing. Plant J. 2012;70:177–90.
464. Merriman B, Rothberg JM. Progress in ion torrent semiconductor chip based sequencing. Electrophoresis. 2012;33:3397–417.
465. Doyle JJ, Doyle JL. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem. Bull. 1987;19:11–5.
466. Kumar S, You FM, Cloutier S. Genome wide SNP discovery in flax through next generation sequencing of reduced representation libraries. BMC Genomics. 2012;13:684.
467. Pinzon-Latorre D, Deyholos MK. Pectinmethylesterases (PME) and pectinmethylesterase inhibitors (PMEI) enriched during phloem fiber development in flax (*Linum usitatissimum*). PLoS One. 2014;9:e105386.
468. Di Matteo A, Giovane A, Raiola A, Camardella L, Bonivento D, De Lorenzo G, et al. Structural basis for the interaction between pectin methylesterase and a specific inhibitor protein. Plant Cell. 2005;17:849–58.
469. Duguid SD, Kenaschuk EO, Rashid KY. Macbeth flax. Can. J. Plant Sci. 2003;83:803–5.
470. Life Technologies. Ion OneTouch System user guide. Carlsbad: Life technologies; 2012.
471. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 2003;31:3406–15.
472. Tranel PJ, Wright TR. Resistance of weeds to ALS-inhibiting herbicides : what have we learned ? Weed Sci. 2002;50:700–12.
473. Gui B, Shim YY, Datla RSS, Covello PS, Stone SL, Reaney MJT. Identification and quantification of cyclolinopeptides in five flaxseed cultivars. J. Agric. Food Chem. 2012;60:8571–9.
474. Bruhl L, Matthaus B, Fehling E, Wiege B, Lehmann B, Luftmann H, et al. Identification of bitter off-taste compounds in the stored cold pressed linseed oil. J. Agric. Food Chem. 2007;55:7864–8.
475. Walsh DT, Babiker EM, Burke IC, Hulbert SH. Camelina mutants resistant to acetolactate

synthase inhibitor herbicides. Mol. Breed. 2011;30:1053–63.
476. Loman NJ, Misra R V, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, et al. Performance comparison of benchtop high-throughput sequencing platforms. Nat. Biotechnol. 2012;30:434–9.
477. Li X, Buckton AJ, Wilkinson SL, John S, Walsh R, Novotny T, et al. Towards clinical molecular diagnosis of inherited cardiac conditions: a comparison of bench-top genome DNA sequencers. PLoS One. 2013;8:e67744.
478. Koshimizu E, Miyatake S, Okamoto N, Nakashima M, Tsurusaki Y, Miyake N, et al. Performance comparison of bench-top next generation sequencers using microdroplet PCR-based enrichment for targeted sequencing in patients with autism spectrum disorder. PLoS One. 2013;8:e74167.
479. Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW. Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. PLoS Comput. Biol. 2013;9:e1003031.
480. Damerla RR, Chatterjee B, Li Y, Francis RJB, Fatakia SN, Lo CW. Ion Torrent sequencing for conducting genome-wide scans for mutation mapping analysis. Mamm. Genome. 2014;25:120–8.
481. Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, et al. A large genome center's improvements to the Illumina sequencing system. Nat. Methods. 2008;5:1005–10.
482. Walsh PS, Erlich HA, Higuchi R. Preferential PCR amplification of alleles: mechanisms and solutions. Genome Res. 1992;1:241–50.
483. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, et al. Characterizing and measuring bias in sequence data. Genome Biol. 2013;14:R51.
484. Chan M, Ji SM, Yeo ZX, Gan L, Yap E, Yap YS, et al. Development of a next-generation sequencing method for BRCA mutation screening: a comparison between a high-throughput and a benchtop platform. J. Mol. Diagn. 2012;14:602–12.
485. Hartwig B, James GV, Konrad K, Schneeberger K, Turck F. Fast isogenic mapping-by-sequencing of ethyl methanesulfonate-induced mutant bulks. Plant Physiol. 2012;160:591–600.
486. Thudi M, Li Y, Jackson S a, May GD, Varshney RK. Current state-of-art of sequencing technologies for plant genomics research. Brief. Funct. Genomics. 2012;11:3–11.
487. Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, et al. Analyzing and

minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. 2011;12:R18.
488. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature. 2011;475:348–52.
489. Glenn TC. Field guide to next-generation DNA sequencers. Mol. Ecol. Resour. 2011;11:759–69.
490. Sequencing Project IRG. The map-based sequence of the rice genome. Nature. 2005;436:793–800.
491. Smulders MJM, Klerk GJ. Epigenetics in plant tissue culture. Plant Growth Regul. 2011;63:137–46.
492. Gutierrez-marcos JF, Dickinson HG. Epigenetic reprogramming in plant reproductive lineages special focus issue. Plant Cell Physiol. 2012;53:817–23.
493. Lawrence CB, Joosten AJ, Tuzun S. Differential induction of pathogenesis-related proteins in tomato by Alternaria solani and the association of a basic chitinase isozyme with resistance. Physiol. Mol. Plant Pathol. 1996;48:361–77.
494. Vergne E, Grand X, Ballini E, Chalvon V, Saindrenan P, Tharreau D, et al. Preformed expression of defense is a hallmark of partial resistance to rice blast fungal pathogen *Magnaporthe oryzae*. BMC Plant Biol. 2010;10:206.
495. Pechanova O, Pechan T, Williams WP, Luthe DS. Proteomic analysis of the maize rachis: Potential roles of constitutive and induced proteins in resistance to *Aspergillus flavus* infection and aflatoxin accumulation. Proteomics. 2011;11:114–27.
496. Xin M, Wang X, Peng H, Yao Y, Xie C, Han Y, et al. Transcriptome comparison of susceptible and resistant wheat in response to powdery mildew infection. Genomics. Proteomics Bioinformatics. 2012;10:94–106.
497. Wen N, Chu Z, Wang S. Three types of defense-responsive genes are involved in resistance to bacterial blight and fungal blast diseases in rice. Mol. Genet. Genomics. 2003;269:331–9.
498. Zhang WW, Jian GL, Jiang TF, Wang SZ, Qi FJ, Xu SC. Cotton gene expression profiles in resistant *Gossypium hirsutum* cv. Zhongzhimian KV1 responding to *Verticillium dahliae* strain V991 infection. Mol. Biol. Report. 2012;39:9765–74.
499. Chen J-Y, Dai X-F. Cloning and characterization of the *Gossypium hirsutum* major latex protein gene and functional analysis in *Arabidopsis thaliana*. Planta. 2010;231:861–73.
500. Wang F, Yang C, Zhao P, Yao Y, Jian G, Luo Y, et al. Proteomic analysis of the sea-island

cotton roots infected by wilt pathogen *Verticillium dahliae*. Proteomics. 2011;11:4296–309.
501. Hall H, Ellis B. Transcriptional programming during cell wall maturation in the expanding Arabidopsis stem. BMC Plant Biol. 2013;13:14.
502. Lippman Z, May B, Yordan C, Singer T, Martienssen R. Distinct mechanisms determine transposon inheritance and methylation via small interfering RNA and histone modification. PLoS Biol. 2003;1:420–8.
503. Ma L-J, van der Does HC, Borkovich K a, Coleman JJ, Daboussi M-J, Di Pietro A, et al. Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. Nature. 2010;464:367–73.
504. Williams AH, Sharma M, Thatcher LF, Azam S, Hane JK, Sperschneider J, et al. Comparative genomics and prediction of conditionally dispensable sequences in legume–infecting *Fusarium oxysporum formae speciales* facilitates identification of candidate effectors. BMC Genomics. BMC Genomics; 2016;17:191.
505. Lo Presti L, Lanver D, Schweizer G, Tanaka S, Liang L, Tollot M, et al. Fungal effectors and plant susceptibility. Annu. Rev. Plant Biol. 2015;66:513–45.
506. Weiberg A, Jin H. Small RNAs-the secret agents in the plant-pathogen interactions. Curr. Opin. Plant Biol. 2015;26:87–94.

Appendices

Appendix 2.1 Sequences from SSAP eluted bands. The LTR sequences from representative members of the six TE families investigated are presented along the sequenced sections of the SSAP bands (Genbank accession numbers: KX364308 to KX364373). Boxed sequences correspond to the LTR of representative elements of each TE family (Table 2.4), in direct and in reverse orientation; some LTR boundaries were adjusted from the original predictions of LTR finder (Table 2.4) after mapping analysis. The LTR primer region is shown in blue. The sequenced LTR region present in SSAP bands is shown in green. The polymorphic bands are named according to the original number given when eluting the band from the gel and match Table 2.9 descriptions. Sequences of the EcoRI adaptor primer and of the LTR primer have been trimmed from the sequences. Names of the TEs are explained in the methods section.

RLC_Lu0-primer3 * EcoRI

>RLC_Lu0-1 (LTR_sense)
TGTTGAATTACAGAAATAACTAGGATATTATTAGGCGTATAACTTGGAGTAATCTAG
CTATCCTATAGAATCAGTCTAGCCTGATTTAGGCAGTGTATTCATAGTAAGTGGTAG
GTAGGATAAATCCCTGATTTGTAGGGATTACTACGAGCTGGATCTGCATCCTAATGA
TGCAGACTGTGTATATATGTAAAGGAAACCACAGAAATAAGAATATCATACCAGAA
TCACTCAGATTTCTGGATTTCTGCA

>RLC_Lu0-1 (LTR_reverse)
TGCAGAAATCCAGAAATCTGAGTGATTCTGGTATGATATTCTTATTTCTGTGGTTTCC
TTTACATATATACACAGTCTGCATCATTAGGATGCAGATCCAGCTCGTAGTAATCCC
TACAAATCAGGGATTTATCCTACCTACCACTTACTATGAATACACTGCCTAAATCAG
GCTAGACTGATTCTATAGGATAGCTAGATTACTCCAAGTTATACGCCTAATAATATC
CTAGTTATTTCTGTAATTCAACA

>band_5
GATTAATCCTACCTACCACTTACTATGATTACACTGCCTAAATCAGGCTAGTCTGATT
CTATAGGATAGCTAGATTACACTAAGTGATACGCCTAATAATATCCTAGTTATTTCT
GTAATTCAACAAAAGCCATCAAAAACAGAAAACTGCAAGTGATAATAGCTGCTGCA
TCTTGGGAGATTTGAAAACCTCACCTCATTGTAAGGCCAAGTGTAGACCCGTGGAGG
CTTCTCTTCTACAGGTGGATTGGGAGCAGCTGCTGCATCAGTCCCGACTGCCTCCGA
AGTTACAAATCGAACATGGGAACTCGTGAAGCCAGAATGGAAGGTCCTAAATGGAG
TTTGAGCCGAGGAAAATGGGAACAATTTTCCCCTTGACATGCCTGATCAACGAAAA
ATGATTGCAATACTCAAGACATGCATCAGAAAGTAAAATAAGGCCATGGAAGCAAC
CAAAACGGAAAAGAACCGAGAAT

>band_6
GATTAATCCTACCTACCACTTACTATGATTACACTGCCTAAATCAGGCTAGTCTGATT
CTATAGGATAGCTAGATTACACTAAGTGATACGCCTAATAATATCCTAGTTATTTCT
GTAATTCAACACTACGCTGTTTCAACACAATTTCAAAGAGCCACGGTAGGCAACAA

ACAGTAGTGAATGAGAAGAAAAAGAGACAATTGATAATAGGTAAGAGGATTCGCA
GTGCCATAATCACCTGTAAACGCTTGTGTGGTTGGGAAGAGTTCTTGTCGAAACAAA
CATGTGGGCCGAACATTTCCACCAATTCCTTTGAAGTAGCCAAGTCATCTTCCGATT
CTAGAAAACTACCATCCACTAAAACATGGGCCATATCATTTCCTGAAGACTGAAGG
AGTCCAAGAAGCTCCTTCTTAGCCTCAGTAATAATATGGGAAAACTCTTCAAACTCA
TTANATCCGCTTGNNTTA

>band_7
GATTAATCCTACCTACCACTTACTATGATTACACTGCCTAAATCAGGCTAGTCTGATT
CTATAGGATAGCTAGATTACACTAAGTGATACGCCTAATAATATCCTAGTTATTTCT
GTAATTCAACACAGCCGAACCAACTTCTCCCAACTGTGTGATTCGCAAAATGTTGAT
TGAGCAACTTATTTGAAAATGTGTTAATTGAAGTACTAGAAGGAAATAAAATAGTCT
ATTACTATACCAGAGAAATACCTCGGTGAAATAGGACCCATCTTGCTGGTAAGTGGT
TGAAACTGTGAAGAGATTTGAGCTGCTAGTACCATATTCATTATCCCATTCAAATAC
CATGGCCCTTGTGGCATTATGATGAACTCTAAAAGGGGGATGTTTATACCATAACGA
GAGTGAACCTGTACTCCGTGGATATATTTCA

>band_8
GATTAATCCTACCTACCACTTACTATGATTACACTGCCTAAATCAGGCTAGTCTGATT
CTATAGGATAGCTAGATTACACTAAGTGATACGCCTAATAATATCCTAGTTATTTCT
GTAATTCAACACTATTTAATTAAAATAGTACCATTAATCTTGAACTGTTAACCTCAG
AGGTTCAACCTCTGAAGACAACATAGAAAAGTATATAAGCTACCTGTTCTTGAAGAT
CTGGAGCACTATCCATCTCTCCATTCTCTGTCAGACTAATAAGAACATCATCTGAATT
AGTAGATCCAGACTCTGAACAAGATAAGTCACCAGATGTTTGCATTCTTCCTTCTAC
CAACATTGATTTAGGATCATGCTTGAAATTATACAGCTTAGAGGCCAACCCTTTAGA
ATCTCTCCAAGCTT

>band_10
GATTTATTCTACCTACCACTTACTATGAATACACTGCCTAAATCAGGCTAGACTGATT
CTATAGGATAGCTAGATTACTCTAAGTTATACGCCTAATAATATCCTAGTTATTTCTG
TAATTCAACAAGTCTCATATAGCAAACATGATCAACCATGATCTCAACAACAAGCA
ACAGAAAACACTCCACGACTAACAAACTAACGAGGGACCAAAATGTAAACAGATG
GAAGTATTGAACTAAAACGACAATCACAACAGACTTGCACAATAAACCAAAACGCA
CAGTTTCATTAACTTATTAATTCCAACAAAAGTTAAGTAAATAAAATAGTAATACTT
ACTCTTTGCATCTGGCGTACATACTCGC

>band_11
GATTAATCCTACCTACCACTTACTATGATTACACTGCCTAAATCAGGCTAGTCTGATT
CTATAGGATAGCTAGATTACACTAAGTGATACGCCTAATAATATCCTAGTTATTTCT
GTAATTCAACACCTCTGCAGAGATATTATATTGTTGTTTATTTAGTATGCATCTATTG
TGGCATTCTGTTGTGGAGATTTATGCTTTAGGAAACCTCCTTTTTTATAGTGAATATC
TCTACTCACTCCATTCGTTGAATTTAGGCTGGATTTACTGGTATTGGAGTTGGCGCAG
CTTATCATGGGCTTAAGCCTATTATC

>band_12

GATTAATCCTACCTACCACTTACTATGATTACACTGCCTAAATCAGGCTAGTCTGATT
CTATAGGATAGCTAGATTACACTAAGTGATACGCCTAATAATATCCTAGTTATTTCT
GTAATTCAACAGTCTTCAATACCTGATTTTGCTTTTCCACTGCAGTCTACAATATCTG
ATTTTGCTTCACTAGCTGCTGAAGATGAAAATACAAGGGAAGTAGTTTCACGCAGAT
TCTTGCAGACAATGAAAATGATTTTGAAGGCTACGAAGCGTGCAGGACAAGCTGGA
AGGTCCAA

>band_14
GATTAATCCTACCTACCACTTACTATGATTACACTGCCTAAATCAGGCTAGTCTGATT
CTATAGGATAGCTAGATTCACTAAGTGATACGCCTAATAATATCCTAGTTATTTCTGT
AATTCAACAATTTAAAAGTAACAGTTTTATGTCTGCCTATCTTACATTTTAGGAGAA
CAAAAATAGATGCAGTTTGGACGACAAGCTGTGGCTTCTTCTGCTGTTGTTGGATTT
CATTTTTACTCAAATTGAAAAGATACTTCAACCTAGTAGA

>band_16
GATTAATCCTACCTACCACTTACTATGATTACACTGCCTAAATCAGGCTAGTCTGATT
CTATAGGATAGCTAGATTACACTAAGTGATACGCCTAATAATATCCTAGTTATTTCT
GTAATTCAACAGTTCTTACTTTATTTAGTGATCCAAGTTGTACAACTCCATAAGGAA
CAACAGCTACAACAGCAATTGTCTGCAGATCAATAAAAACAAGCAAAAAACATGAG
CATGATATCTCTGCTAAAACAATT

>band_17
GATTAATCCTACCTACCACTTACTATGATTACACTGCCTAAATCAGGCTAGTCTGATT
CTATAGGATAGCTAGATTACACTAAGTGATACGCCTAATAATATCCTAGTTATTTCT
GTAATTCAACATTTTTTTGTTGGAAATAGGAAACACAAGATAGAAGTAAAATAAAC
ATGACATAACTGTCCTTTTGTGGAGCGTAACT

RLC_Lu1-primer1 * EcoRI

>RLC_Lu1-1 (LTR_sense)
TGAGGAAATCCCGTTCCTTATTTGTATACAGTCAGATCTTATTAGTTTATTTGCTAGT
TAGATAGTTACCATATTAGTTCTAGTCCTATCCTAGTTTAGCTAGCAGATCCTATTAA
ATAGTTAGAGTAGATATTTTCTTACCTAGTCAGTAGGAGGAATGCTGTATATAATAA
ACCTACGAGACATGAATAAAAGTAATTCATTCTCTCAATCTTCA

> RLC_Lu1-1 (LTR_reverse)
TGAAGATTGAGAGAATGAATTACTTTTATTCATGTCTCGTAGGTTTATTATATACAGC
ATTCCTCCTACTGACTAGGTAAGAAAATATCTACTCTAACTATTTAATAGGATCTGCT
AGCTAAACTAGGATAGGACTAGAACTAATATGGTAACTATCTAACTAGCAAATAAA
CTAATAAGATCTGACTGTATACAAATAAGGAACGGGATTTCCTCA

>band_12

TAGGTAAGAAAATATCTACTCTAACTATTTAATAGGATCTGCTAGCTAAACTAGGAT
AGGACTAGAACTAACATGGTAACTATCTAACTAGCAAATAAACTAATAAGATCTGA
CTGTATACAAATAAGGAACGGGATTTCCTCAGAACTACATATGTATTACTGGATCTT
AGTGGACTCCTTTTTCTATATTTATATATATACCCTTCACGTCTTCCTTGTAGCTTTCA
TGGATTGGCTGAATTACATGTTAGCCAGGCATTATCCACTCCCATTGTAATTTGTAA
GTGGGTTCATCTTTGGAGCATTGTTATAGAGACTATAGCATGTGGTTGTGGTACATT
CATGAGAATTGTAGTAACTGCTTTTTTTTTTCACTACCTGTTTATTGGTGGCTTTTCTG
CCAGTGGCTTAATATCTAAGCAATGTGAAGTCTGCTTGCTTCATACGGACTTGCTTTC
GTTACTCTGCTGAAGATGGATTACTGTTATATTTGCGAGCATACAAGACAATCATGC
TCTGGTTTGATGTATTATGTCTGTCTGTATGCTCTTATCTCATAGTTGCGCACAAGGT
TTGAAGAATTAGAGAAGTTAAGTGAATGTTGACTGCTACAGTGCTACCTCATATGTT
GGTGTCAGTAGTAACTCCGTCCCAGTCCTATCTGTAAGTATGTTGAGCTGA

>band_13
TAGGTAAGAAAATATCTACTCTAACTAATTAATAGGATCTGCTAGTTAAACTAGGAT
AGGACTAGAACTAATATGGTAACTATCTAACTAGCAAATAAACTAATAAGATCTGA
CTGTATACACATAAGGAACGGGATTTCCTCACTATATAACCTTTAAGGCTAATTTTTT
GTTCAACATTTCTCTTCTTGAATTTGTGATGTTATTGTGGGTTTTTTTGTTGGGCATTC
ATTTGCTTCTCAATCATCTTGGATGGTTAAAAACTGTGATCGTTAATAATTCTCTAAC
CTATGATTTTGACGTATTTGCTTTCATACTGTTGAAGGTGTTATGAAGCTGAACAGA
AGCGTGAGGAGCGAGAAAGGCATCCTGTGAACCACCGGGAGAAACATGGACTCTAT
CCGGTAAGTTTGCAACAATTTGTTGTTGGTTGAAACTTTTCTCTCGTTACGATGAACT
GTAAACAGCATGACTGGAAATTTTTGAAGGTTGAAATCGGTGATGTGACTGTTGATA
CGAAGGACCAAGACGAAATTCTTGAGAGT

>band_14
ACTCTACTATTTATAGGATCTGCTAGCTAAACTAGGATAGGACTAGAACTAATATGG
TAACTATCTAACTAGCAAATAAACTAATAAGATCTGACTGTATACACATAAGGAAC
GGGATTTCCTCAATAACCTTTATGATGATCGATATATCCAATAAATACACATTGAGT
TTTAATGGTCAACTTAGTTTGATCTTTCTTGGGGAAGAACACAAAAAAGGTGCAACT
AAAGACCCTTAGTCAAGTGTCATCCGAAGGACATCCATTAAAACTTTAAAAGGTGA
CTGACTCTGTAGAACCAGCGTAGGTTGAAGATTCACAAGATAAACCATGGTATGAA
CCTTTTCGACCCTTAATTGTGAAGGACCCTGAGATTCGAGTAAGAGAGCTCGTGTAC
GATCCAAAACATGACGATGCTTACACTCTACAAACCATTTTGTTCAAACACACCCGG
ACAACAATCTTGGAACAGGATACCATTTTCTCAAAAATATAGATGAAGAGCATGCG
AGGT

>band_15
TAGGTAAGAAAATATCTACTCTAACTATTTAATAGGATCTGCTAGCTAAACTAGGAT
AGGACTAGAACTAATATGGTAACTATCTAACTAGCAAATAAACTAATAAGATCTGA
CTGTATACAAATAAGGAACGGGATTTCCTCAGGATCAAAAACCTCTCAACAGTGAT
ACATCATGACAAAGAACGAGTTACTCTTACGCAGATTATGCAACATGTTTTCCATCA
GATACACAAAATTCAATAACAAGGATGTTCAAGTTACCTGTTTCCATGGCAATAAAT
TACACAGGGCAGAGGCTTTTCTCCAGGACTGACGACGGGCAAATAATGGCTGCATT
GAAGAACATCCCCCCTGTCATTTGTTATCTGCATGATATTTGAAATGAAAATATTTA

GGAAGGCTAACTAAACTCATCATGCGGGATAAACATGAATATTGTGAATGGTTACCT
CCACATCCTTTCTCTGGTACAATTTCCCTCGTAACAT

>band_16
TAGGTAAGAAAATATCTACTCTAACTATTTAATAGGATCTGCTAGCTAAACTAGGAT
AGGACTAGAACTAATATGGTAACTATCTAACTAGCAAATAAACTAATAAGATCTGA
CTGTATACACATAAGGAACGGGATTTCCTCAGAAAACATCATCTATATGATTAAAAT
AGCATTTTTAGAGTCAGAGATACAGATCATGCACACAAAACTAAATAGTTCTGAAC
AGAAAGCATCCTGTGGCCTTACCTGCAACGCTCTTTTTTCTCCAGTAAAAACACGAG
CATCTAATTCATGTTGCATCTTTTCGGCAGCATGAATGGCTTGAAGAAAATCCATAA
CCAACTTTCCATCTTTTGTTATGTAACCTTCAGCAGATTCCTCGGTGCTACCCTGCAA
TTTCAACGAACCAAGGTTTATTGAACTCAAAGATTTATGATGTTGACAAAGTACTAA
AGTAGTCAAAATTT

>band18
TAGGTAAGAAAATATCTACTCTAACTATTTAATAGGATCTGCTAGCTAAACTAGGAT
AGGACTAGAACTAATATGGTAACTATCTAACTAGCAAATAAACTAATAAGATCTGA
CTGTATACAAATAAGGAACGGGATTTCCTCAACAACACCACCACCACTGCAATCCTC
GAATACAAATCTACAAAAGGAAAAACCTCTCCGGTTCTGCCTCAGCTTCCGGCCTTC
AACGACACTAACACTGCAAGGACATTCACTTCCCAAGTCAGGAGCCTTACATCCAAT
GTGAATGTACCCAAAAAGATCGATAAATCCTTGTTCTTCACCGTGGGGCTAGGGTTG
AACAATTGTACCAAACTGAACAGCCCTCGTTGCCAAGGTCCAAACGGCACCAGATT
CACCGCAAGCATCAACAATGTGTCGTTCGTGTTTCCCAGGAG

>band_19
TAGGTAAGAAAATATCTACTCTAACTATTTAATAGGATCTGCTAGCTAAACTAGGAT
AGGACTAGAACTAACATGGTAACTATCTAACTAGCAAATAAACTAATAAGATCTGA
CTGTATACAAATAAGGAACGGGATTTCCTCACCTATAGTAATAGGATCTCGTCATCT
TCTTGTCTGATCTAAATGCAGGATTTGCCTCGTACTTTTCCTGGCCATCCTGCCCTGG
ATATTGATGGCAGAAATGCTCTTAGGCGGATACTTACAGCCTATGCACGGCATAACC
CCTCAGTTGGATACTGCCAGGTACTTGACTTTAAACTTCATAGCAGGGGAAGATATT
TCTGTTATTCCTAGGCCAANNTGACTTTTAACATTGTTCTCATTATTGCTCTGCCATA
TATTCAGGTCAT

>band_21
TAGGTAAGAAAATATCTACTCTAACTATTTAATAGGATCTGCTAGCTAAACTAGGAT
AGGACTAGAACTAATATGGTAACTATCTAACTAGCAAATAAACTAATAAGATCTGA
CTGTATACACATAAGGAACGGGATTTCCTCAAATACGTTGGTCATTGCTGTTTTTATA
ACTCTTGCAGTATCCAAGCACTGCACAAGAAAGTGCTCCTTGGAGAAAATTATCAGG
TGGAAGGAAATGGTAAAGATATCATGGATGATAATTGGAATCATGATCAGGAAATT
CGTTGTGCAATCTGGAGGCTAGTTGGCATTTGTAGTTCAGATGACTCAGGAAGCATC
AGAGCCTTGGTTTCTGATTTTGTATCTCGGGTATTATTCTTGCT

>band_22
TAGGTAAGAAAATATCTACTCTAACTATTTAATAGGATCTGCTAGCTAAACTAGGAT
AGGACTAGAACTAATATGGTAACTATCTAACTAGCAAATAAACTAATAAGATCTGA

CTATATACAAATAAGGAACGGGATTTCCTCAGTAAATCACACATGTCAGGCTAGTTA
CATTCTTCTTTTTTAGTATGAAAATGAAAGCCAAAGACAGTATGTATCACTTACGGC
TGAGACAAGGTTATCATCCCACTTAACCATTGAAGAGAAAGAAGCAGAAGTTCATG
GAGCTCTTTTGCAGGAAGGTCTTTCTCCACTGCTTCAGCGTGCTTCAGGAAAAACAA
GCCGGCCTGTAAAAGGCAAACAAAGAGG

>band_23
TAGGTAAGAAAATATCTACTCTAACTATTTAATAGGATCTGCTAGCTAAACTAGGAT
AGGACTAGAACTAATATGGTAACTATCTAACTAGCAAATAAACTAATAAGATCTGA
CTGTATACAAATAAGGAACGGGATTTCCTCACCTCTCAAAGTGCATAAACTTCAGTA
AATTTTGACATTGAACCAGTAAGAAATTTACACAAGTTTCTTCACCACAAAAAACAC
TCTGAAGTCGCAGCAATCTGCAAGAAGACAAACACAGATTGAATAAGCATATGAAG
GAGTAATAACTAATATAGATAGAGAAATAGCACAACCAAATAACTTGTCAAACAAA
CAATGGG

>band_24
TAGGTAAGAAAATATCTACTCTAACTATTTAATAGGATCTGCTAGCTAAACTAGGAT
AGGACTAGAACTAACATGGTAACTATCTAACTAGCAAATAAACTAATAAGATCTGA
CTGTATACAAATAAGGAACGGGATTTCCTCAAAAATCGATATCCGCTGTGCCATCCA
AGGTGATGTTGTGCTTGAATGTGTCAGCATACGGGATGAAATGGAATCTGAGGAAA
TGATGTTTCGGGTAGTGTTCAATACAGCTTTCATCAGGTCAAACATCTTGATACTCA
ATCGAGATGAAATTGACATATTATGGGATGCTAAAGATCTATTCCCAAAG

>band_25
TAGGTAAGAAAATATCTACTCTAACTAATTAATAGGATCTGCTAGTTAAACTAGGAT
AGGACTAGAACTAATATGGTAACTATCTAACTAGCAAATAAACTAATAAGATCTGA
CTGTATACACATAAGGAACGGGATTTCCTCAGCATCTTTTTTCATCCTAAATGGTGTA
GATGGCATTCCCCTAGTCAAAGATAATGTTTCTACTTTCTACTGTCAGTAGTCTAGTG
AAA

>band_26
TAGGTAAGAAAATATCTACTCTAACTATTTAATAGGATCTGCTAGCTAAACTAGGAT
AGGACTAGAACTAACATGGTAACTATCTAACTAGCAAATAAACTAATAAGATCTGA
CTGTATACAAATAAGGAACGGGATTTCCTCAATATCTAATAACAGGAGCATAAAGT
ACATGTCCAAATTTATTTGAAATAAAAAATCCTTCTATC

RLC_Lu1-primer2 * EcoRI

>RLC_Lu1-1 (LTR_sense)
TGAGGAAATCCCGTTCCTTATTTGTATACAGTCAGATCTTATTAGTTTATTTGCTAGT
TAGATAGTTACCATATTAGTTCTAGTCCTATCCTAGTTTAGCTAGCAGATCCTATTAA
ATAGTTAGAGTAGATATTTTCTTACCTAGTCAGTAGGAGGAATGCTGTATATAATAA
ACCTACGAGACATGAATAAAAGTAATTCATTCTCTCAATCTTCA

>RLC_Lu1-1 (LTR_reverse)

TGAAGATTGAGAGAATGAATTACTTTTATTCATGTCTCGTAGGTTTATTATATACAGC
ATTCCTCCTACTGACTAGGTAAGAAAATATCTACTCTAACTATTTAATAGGATCTGCT
AGCTAAACTAGGATAGGACTAGAACTAATATGGTAACTATCTAACTAGCAAATAAA
CTAATAAGATCTGACTGTATACAAATAAGGAACGGGATTTCCTCA

>band_9
AACGGGATTTCCTCACTTTCTGTTTCTGTGCGGTGTTGACATATTGTTATATTTGTAA
ATTTTTCAAGGTCCAAAGTATGAAGAGAAATATTGGAAGATTTTCTGGCTTCGTTTG
GACTGGGAATGAGGTTTGGATCTGTTAAGTTATACATTTAATGTAGTTTTTTAGCTGA
AATTCTCCTATGAAGTGTATGTTGCAACTCCATGTGTGGACACAAGATTTACCTTAG
GACTTTAGTGACCATTTTGTGAAGCGTTTTTTGAATTGTAACTTTGATTCTGGTTGAC
AACTCTTGCTTGCTTGACAGGAAAAACAAAAGTCAAGAATGAAGGAAAAGCTTGAC
AAGTGTGTTAAGGAAAGTCTTCTAGACTTCTGTGATGTACTCAATATTCAAGTAACT
AAAGCCACCGTGAGAAAGGTCAGTTGTGACTTCTAATGTTAATATATTCGCCGGGTT
ATCCTTAAATTCATCTTTTAGCTCGATAATGTTTTCAGGAAGATCTCACTGTAAAAAT
CTTGGAGTTCTTGGAATCTCCTCATGCAACAACTGATGTTATGCTTGCTGACAAGGA
ACAGGTATTCAATACTTGAATGAGTTTGCTGCAGTTTGGGTAGTTGGGGTTTCTTTTT
CAATGCTATTACCTAGCTGATTTTGGATCTTATATTCGGTTTCAGAAAGTCAAGAGG
CGGAGGTCAATGACTGGGAAAAATTCAAGCCCTGGGGAAGCATCAGCTACACTAGC
TAAGGTTAGAATTACTGTAGTTAATGGACAACATAGATGTTCTGCAACTGTCAGAAA
TTTCCTCAATGCAGTGATCTCAATACAAGTAGCCCTGATAGTAGCAGTAAA

>band_12
AACGGGATTTCCTCACATGCTTTTATTTAACTAATTATACGGCATTCATTAAACTTCC
ATTCAGTTCTGCTTGCATGCTACTTGATAACATAAGATGTAAAACAGGAAACGGAGG
AATTGTTTTGAAGAACTGAAGTGAAAAACGAAGACAAGTGAAGACCATGGGCCACT
TACCAGAAAGGTGCTTCACATCAACATCCGTAATGCAGTTGCACCAATTGAGGTTAA
GAGATTCCAACTTTGACAAACCTACAAACAGCAGGACATTTTAATAGCCTAGAGAA
AACAATGCAAGATACCAGAAACTGAATGATGCAAGAGAAAATTAGATAAGCAAGT
AAGAGTTAGATAAAAGTGTGAAAATACAGAAATGGAAAGGTAAAGTGAGTACCGTT
GAGATGAACAAGAGCACCACCGCCAATGCCAGGACCCCTCTCCATGTCCAATTGAA
CTAGGTTTTCCAATGTGGCAACCGCCCTCAATCCCTCCGCACCGATGGCATTATTTC
GTCGGAAGCTCAATCTTACCAAGTTTGAGCAGCCTTCTGGAAGCAAACAATGGTATA
AGAAAATCTTCAACTCCCCAAAAAAATGTCACAATCATAACAAGCCAAATGGNTAA
TATCTATGAGCCGCAGA

>band_14
AACGGGATTTCCTCAGTATCGGATCCTAGTTGTAGAAGGAGAAAAAATGTTCACCCA
GTGTTTTTGTGGTCAAAACAGCAATTGTTCCCAAAGCCCTGAATTTGCCTCTGCTAGC
CAAAACGGTATGGCTGCGTGGATGGGTACTAGACAAAGGTCACTTCTAGTTGATGTA
GATAATTCTTTAATAATTTTTGAAAATGACCTTTATATGACTTCCGACTTCTGTGTTC
TAGCTTAATATAATGCAACATGTGATAAGCAGCTTAGTTTTTTTACCTACTTCCATTC
ACCAAAGTGTTCTACTGAAAGTCTGAAACAAGTTAGCAAACATCGTTTAAACTGAA
GCCAGTAAGAAATGTTATTATTCTCCTTTGTGCACGTGGGTTTGTTGACTGCACTGGC

TTTGTATGGCCTGCAGTTGAATGGGCTACAGGTAGGCTTATGGTTTCTCGGGTATTC
AGGAACCAGTTATTTCTTTCATTCGTTTTGGCTGAATGACGACCAGA

>band_15
AACGGGATTTCCTCAGTAGTATGCCTGTCCTTTGGCTGACAGACATAGGAACCATAA
ATAGAGCCTCACTACATCAGTGACGGCTATTTCTTCATCATTTGAGCCAGAGAAACA
AATTAGACTCATGAGAGTTCCTGTTTACAAATGCGGTTCTGTCCTTTGACTAACAGA
TGTATGAACCATTGAAGGTATTGTCACCATATCAATGACGGTACAAGCTCTTCAGCG
TGGGAGCTAGGTGAAACAAATTTGATCTATTCATTTTCCCTTTTTTCTTTTGTTTATTT
AAAAGCAGGATTGAAGCCTATGACAATCACACTTTCTTTAATTGCAGCGTTTAATAA
ATGTACTTTGTTTCACTTCTAAGCGATATTATATTCACTGTGTTTGGACATCTGAAAA
GGAGACCAATATTAGAGCGGGCAACAAAAGCAAGATACTACCGGGA

RLC_Lu2-primer1 * EcoRI

>RLC_Lu2-1 (LTR_sense)
TGTTAAATGGTGTTAAATTATAAGTATGTGAGTGAGTTGTAATAGATAATAGTAGTA
GTGGTATAAATAGAATTCTCCTTTCTGTTTGTACATTTAATTCATTCAGTACAGTAAT
AACAGAAAACGACTTTCCATTACAGCTTCTCCTTCTTCTTCGCTTTCTCTGTTCTTCTC
TTCTCTTCCATTCATCAAACTTCACA

> RLC_Lu2-1 (LTR_reverse)
TGTGAAGTTTGATGAATGGAAGAGAAGAGAAGAACAGAGAAAGCGAAGAAGAAGG
AGAAGCTGTAATGGAAAGTCGTTTTCTGTTATTACTGTACTGAATGAATTAAATGTA
CAAACAGAAAGGAGAATTCTATTTATACCACTACTACTATTATCTATTACAACTCAC
TCACATACTTATAATTTAACACCATTTAACA

>band_2
TTCTTCTCTTCTCTTCCATTCATCAAACTTCACACCTTGGCGTACTCGAGGACGAAGT
AGATCTTCGTTTTCGAGGCCAGGACTTCGTAGAGGTGCAACATGTTTTTGTGTTGGT
GAACGAGTCTCATGATGTGGATCTCTTGCTTGATGTTGGTGGTGGTTAGTCCTGCTTT
CTGAGCCTTTTCCTTGTCGATTACCTTGATTGCTACACTGTTTCCGGTTTGGAGGTCT
CGAGCGTAGTGGACTTTGGCGAATTGGCCTTGGCCTAGCAGCCTGCCGACTTCATAC
CTCTCCATCAAAATTTTGGTCCCATTGTTCTCCATTTCCAGTGAAGAGTTCTTTTTAA
CACACCGAATAGGTTCTACATGAGCTGGCAGGCTTCCATTACACGTTTCCTGCATGT
CCTGGTGTTGCAGCACTACAGATAGATCTGGATCAAAGAAGTAAAGAGTTAGAAAT
CCATTAGTATGTACTGTCAGCCAAAAGGATCATCAGATTACAGGCATTTTGGTTGGA
TTCTGGTATAAAAACCAATCTGATTAACCCTCTCCCATCAACTCAAGTTATTTGAAA
AACTCAAGTTATTTGAAAAAATTGAGATGCAATGNTTCTTAAGAGGTCTCCATTCT

>band_4
TTCTTCTCTTCTCTTCCATTCATCAAACTTCACATACTTAGCTGCAATCGTGAAAGAA
ACGTTGAGGCTACACCCGACGGGTCCTTTACTAATCCCCCACCGGTCCATGGGGACT

TCAGAAGTGATGAATTACACGATCCCCGAGGAGGCCTTGGTCGTCGTTAATATGTGG
GCGATCTCTCGAGATTGTTCGATCTGGGGGGATGATGCGTTGTCGTTTAGGCCGGAG
AGGTTCGTCGGTTCGAAGGTAGATTTTCGAGGACAAGATTTCGAGTTGTTGCCATTT
GGTGCAGGGAGGAGGATGTGTCCCGGGATGCCATTGGCGGCAAGGCAAATTCCTCT
GCTCTTGGCTAATTTGGTTTGGAACTTTGATTGGTGCTTGCCAGACGGAGGAGATCC
AGCGGTGGAGTTGGATATGAGCGAGAAGTTTGGGCTCATTTTACACAAGGAGCGGC
CTCTGGTTCTTGTTCCTCGTCCGAGTTCTATATTACAAGACTGATGAGTAAGTGTCTT
CTTCAAGTGTAGTAGAGTTATGTTGGTAATTTTAAGATATCCTATTT

>band_5
TTCTTCTCTTCTCTTCCATTCATCAAACTTCACACCATCCTCAACTGCTCTCACCCAA
CTCTCCTCCTCGTCAGAGAATCGGCTTTCCAAGACCCAATCAAAGCTCACAAGCTCC
AATCCCTCACCAATGCCGGCGCCTCCGTCATTAAGGTGTTATTCAATTGGGATAATG
CTTTTTCCAACTAGGGGAATTTTTGTTGTTCTTGATTCTTCTTCTTCTGCTGCTGTATT
TTTAGGTTTCATTGCAGGACGAGAGTAGCCTTGTGGAGGCAGTGAATCGAGTTGATG
TTGTGATCTGTTCGATTCCGTCTAAACAAGCTCATGATCAGAAGCTGCTTATCAATGT
CATCAAACAAGCCGGAAGCCGGATAAAGGTAACTGTTAGATTGTGATTGTCAAATT
GCTGTCTTCTTTGATTATCAAAAGCCGGATCAAGGCAACTGATTTTAGTGAAATTGA
AGCGATTTCTTTGATGGGATTTTGTACATTAGCACATTTAATGTTAACATGGGTTTCA
TTCATGATGATCATCATTTTCCAGAGGTTCATCCCGTCC

>band_6
TTCTTCTCTCTTCCATTCATCAAACTTCACACATTTGACCATTTCTCGGTCCTACTTCC
TATCAACATGTCTCGTTGGCGGTTCCCTCTACAGATTGTTTTGTGTTGTTGGTCGGCT
TTTTTGCCTGTTATGTCACCCCAAAAATTAGGGATCTGCAATTTTAAAACTGCTCTCA
TCCTTATCTCTTCACCCGTCACTAACTGCGATTCTTTTCCACTCGTGTTTATTTCCAAG
TTTGCATCACTATCCATCTTTCCCTTTTGCCCGATATCGCCTTCACCGATTCCAAATTC
TGTCTCTGACTATCTCATTCTTTACTTGGATTTCTTATTCCCTCCAGATTCACTTGATA
TGGATTCTCTGCCACCAGCCATTATCCCTCATGGATTCGATGTGGAGTTTACGACGA
AGGACGTGCGTGACGTTCCTCTAATTACACAACTCAGCCTTATTGGGTGATTTTTATG
GCCTCTTCCCAAACCGACACATACCCTTCTCCAAGGAATGGCTCGAGTTTGGA

>band_7
TTCTTCTCTTCTCTTCCATTCATCAAACTTCACAGAAACAACAACCCAACTGATTTCC
ATTTCGACGAAGTCGCAGACGAGCAGGAGGACCTCGTCGCCGTTCAGTTTCCGCCGC
GGTTTCGCGTCCTCGTCTAATTCCGCTTCTCTGCGGAGTTTACAGGTTTCCGATTCTC
CCGTTATCTTACTTGAGCTTAATCGGAGCTACCCGTCTCTGTATCTATGTGCTAAATA
GTATTATGCAACTTTTGTTTCTTGTGGGGGCAGTTTCAAATAACGGAGAAGATCTAA
TATAAGCGTGTGGTTGGAAAATTGATGAATTGAGGTACACGATGAGCAAGCAAAAG
TCGAGTAGCTGTGGAATCTGTGAGAATACGAATCGTGCTTCCGTTTGTGCAGTTTGT
GTCAATTACAGGTGAAATCTGCCTGACCCAACTTTTTTACGTTCTCTGGCTGGGAAA
GTTGAAATTTATCACTTACGCTGTTTGTTTCATTGCAGGCTGAATCGCA

>band_9
TTCTTCTCTTCTCTTCCATTCATCAAACTTCACATGAGTGTCCTCTAAAATTATGTAC
ACGTACTCTCCTTTCCCTTGCTTATGCAACTTGCAACATATATTTTAGCAAAGTTTTT

AGAAGCTAGTTCTTTTTTGTCTTACTTTAGTATTGGATTTAAGGTATTTTGATTCTATA
TCGGCTACCGCGTGTTTTCGTAGAGCAGTACAGTTTATATCAGGTTGAATTTAGTGC
ATTTTACATATGTATTTAAAAAATTAAAACTATGTTATTTCGATTGTCTGTATTGTGT
GGTTTATGAACAATGGCTACTTTTAGTAGAAATTCT

>band_10
TTCTTCTCTTCTCTTCCATTCATCAAACTTCACAATGCAATCAACTCTACGATTTTTCT
TTGAATTAAAAAAAAGCATAGGCCAAATGAGGTCATTATTCGGCACTATGTTAATCA
CAAATCAATTACAAATCTCCAATTTCCCGCAAACGACAATAAATTTATACTCATTTT
CTTTTCGAATAAATAAAAATACCTTATAATACATGTATTGCCGTACACTCACATATTT
AAAGGGACGATTGGCTTTGCAATTGTGACGAAAATTTGGCTTCCACCATAAA

>band_12
TCTTCTCTTCTCTTCCATTCTTCAAACTTCACAAAGGAATCAGAGTTCGTTAAGTGGC
AGAATGATGTAACCAAAATAGCCAAAATAGCCATCAGTATCTGGTTTACATGTATTC
GTCAATGCAGTCAGTGATGTTAGATTGTTTGTATGGACAGGTTAAAATGCAGTTCTT
GTTGAAAGGTGAGGTTTCGCCACAAGAGGTGATCGATGATTCACGAAGACAAAAGT
GTCGAGCAGGGTTAGGTACTGAATTGGNTCNNANTCTA

>band_13
TTCTTCTCTTCTCTTCCATTCATCAAACTTCACAATAGTTAGTTACCGTGTGTTGAATC
TGAAGCTATTCATGTTGTGATGGCAAGAAGAAGGAGAAATCCTACAGAATAAACCC
AACTTTTCAGGCTAATCTCCAGAAGCATATAATGACCGGGNGAGGCCACATGTTTCG
CAGCTCGACTATTTATCGTCACGTTAGTGAATGGATTCAGATGTTCATGTAGNATGT
TTTTGTGTCCT

RLC_Lu6-primer3 * EcoRI

>RLC_Lu6-1 (LTR_sense)
TGTTGGTCCCGGATAATTGATGTGCAACAAAAATATGTACCACCAATATTAAAGCAC
CACAAAATTAGTTGGATCTCTAACCAACTAATTGAATTCCCACCAATATTAACCCCC
ACCTCCAATTTAGTGCACATGGAAATTAGTTGGGAATTCAACTAATTGGATGCCACA
ACCAATTGCTTTCAATTGGTCTCCCCACATTTCTAGTATAAAAGGGAAGCTAGTGCA
TCCCATTTCAATCATCCCTCTCTTCTTCCTTTCTCACTTCTCTAAGTGTTGTAGTGTAG
CAATTTTCACTTGTTTAATAATTGAGATAAGTTATCTCAATTGGGTAGATAGGTGAG
CGGTAGAAAGTCCCGGTAAATGTTTTACCGTGGTAGGAATACTTTCTTGTGAGCGAT
AAAATAGTGAGTAGTTGTTTCGGGGTTGGGAAACACTTGCGAGACACTATTTTGGAT
CGGCTCGGATCACCTTGTAGCTACCTTGTTATAGTGAAGAAGTGCTCGTAGCTGTCG
CTGCTGCCGTAGATGTACTCTCCGCATTGGAGGGGAACTACGTAAATCCCGGTGTTA
TTTACTTACTGTTTTGTGCTTGGCAATTTCGAGAATATTCGTTGTATATTGCATTATTA
ATATTACCACAGTAAATTGGTCTAAGGAGGTTGGCTTAATTATCGTCATGATGGTAT
TGCGGTGGTAATCACCCATCCATAGTGATTTTAAGTGTGGCGGACTACCGCCACTTC
CAACTTATCTGGGAAATATTTACGGTGTGTGGTTTATTAGTGCAATATATTTACTCTA
TTCTCGTCCGCTGCGCCCCAACA

> RLC_Lu6-1 (LTR_reverse)
TGTTGGGGCGCAGCGGACGAGAATAGAGTAAATATATTGCACTAATAAACCACACA
CCGTAAATATTTCCCAGATAAGTTGGAAGTGGCGGTAGTCCGCCACACTTAAAATCA
CTATGGATGGGTGATTACCACCGCAATACCATCATGACGATAATTAAGCCAACCTCC
TTAGACCAATTTACTGTGGTAATATTAATAATGCAATATACAACGAATATTCTCGAA
ATTGCCAAGCACAAAACAGTAAGTAAATAACACCGGGATTTACGTAGTTCCCCTCCA
ATGCGGAGAGTACATCTACGGCAGCAGCGACAGCTACGAGCACTTCTTCACTATAA
CAAGGTAGCTACAAGGTGATCCGAGCCGATCCAAAATAGTGTCTCGCAAGTGTTTCC
CAACCCCGAAACAACTACTCACTATTTTATCGCTCACAAGAAAGTATTCCTACCACG
GTAAAACATTTACCGGGACTTTCTACCGCTCACCTATCTACCCAATTGAGATAACTT
ATCTCAATTATTAAACAAGTGAAAATTGCTACACTACAACACTTAGAGAAGTGAGA
AAGGAAGAAGAGAGGGATGATTGAAATGGGATGCACTAGCTTCCCTTTTATACTAG
AAATGTGGGGAGACCAATTGAAAGCAATTGGTTGTGGCATCCAATTAGTTGAATTCC
CAACTAATTTCCATGTGCACTAAATTGGAGGTGGGGGTTAATATTGGTGGGAATTCA
ATTAGTTGGTTAGAGATCCAACTAATTTTGTGGTGCTTTAATATTGGTGGTACATATT
TTTGTTGCACATCAATTATCCGGGACCAACA

>band_5
CCCATCTATAATGATTTTAAGTGTGGCGGACTACCGCCACTTTCAACCTATCTGAGA
TATATTTACGGTATATAGTTTATTAGTGCAATATATTTACTCTATTCTCGCCCGTTGC
GCCTGAAAGACGTATTATTGTATCTATCACTTTTANTAACANAAGCATACAAAGATG
CACATCAAAAGAGCGTCATGGTAATAAATGACACCTCCCTCTCAATCTTTGCTTGTT
GGAATCTACGTTCCCTCCATTGACAGGAGCTACTCCTCCCCTGCCGCTGCTGCTGCG
GCACATGTATCATCATCATCATGATCATAGTCACCCAGAAATAGTACATATGATGAC
GTTGACTTTCCATTN

>band_8
CCCATCCATAGTGATTTTAAGTGTGGCGGACTACCGCCACTTCCANCTTNCCTGGGA
AATATTTACGGTGTGTGGTTTATTAGTGCAATATATTTACTCTATTCTCGTCCGCTGC
GCCCCAACACATTGCTCTGTGTAATCTTCAACATAACTTCAATTCACATGTACTTGAG
TTTCAAAATTTAAGACGACATCATCTACATACACCATTAGCATTACGAGAACCC

>band_9
CCTATCCATAGTGATTTTAAGTGTGACTGACTTCCGTCACTTCCAACCTACCTGTGAA
ATATTTATGGTGTATGATTTATTAGTGCAATATATATACTCTATTCTCGTCCGCTGCG
CCCAACAGTAACCGATACATTTATCTTTGGTGGTGAGAGAACGACTTCAAGAAATTA
ACTAAGACCCGAATTCAACATTACCTTACCACGACAAAAATG

>band_13
CCCATCCATAGTGATTTTAATTGTGGCGGACTACCGCCACTTCCAACTTACCTGGGA
AATATTTACGGTGTGTGGTTTATTAGTGCAATATATTTACTCTATTCTCGTCCGCTGC
GCCCCAACAGGTTGGTCCCTTGAGNGNCGNGANNGCCTTGCTNNNGGCCGAGGTTG
TGTTAAGACT

>band_15

CCCATCCATAGTGATTTTAAGTGTGGCGGACTACCGCCACTTCCAACTTATCTGGGA
AATATTTACGGTGTGTGGTTTATTAGTGCAATATATTTACTCTATTCTCGTCCGCTGC
GCCCCAACAGACGGGCGCCACCATCCTCCTACTCAAACTCACANCCTCGCT

>band_16
CCCATCCATAGTGATTTTAAGTGTGGCGGACTACCGCCACTTCCAACTTACCTGGGA
AATATTTACGGTGTGTGGTTTATTAGTGCAATATATTTACTCTATTCTCGTCCGCTGC
GCCCCAACAGTAACCTCTTCTACGGNGATTCTGAGTCCTCTTN

RLC_Lu8-primer1 * EcoRI

>RLC_Lu8-1 (LTR_sense)
TATTGGAAATGATTTTTCATTTTCCCGCCAAACTTCACACTCTCCAAGCTTCAAGTGA
AACGGAGCGTTTCTTCCTCTCTACACCAACGACTAAATGAAACGGAGCGTTCTGTTA
AGTGATGAAGATAAGAACAAAACGTCATCGTTGCACTGATCAGAACAGCTCATCGT
CCTCTTCTCTGTTCCTCTGCATTTCTGCAAATTCCGTTAGAGCCTCAAGCTCACCTAC
TCTCTTTCAGCCAGCTACCAGCTGTGCACATTGTCTTTTAGCTTCTCATCACTTTGTAT
ATGTACCACCTTTCTATCAATGAGAACGTTGAGCCATTTCATTTGAACACAAACGAG
TTAATA

> RLC_Lu8-1 (LTR_reverse)
TATTAACTCGTTTGTGTTCAAATGAAATGGCTCAACGTTCTCATTGATAGAAAGGTG
GTACATATACAAAGTGATGAGAAGCTAAAAGACAATGTGCACAGCTGGTAGCTGGC
TGAAAGAGAGTAGGTGAGCTTGAGGCTCTAACGGAATTTGCAGAAATGCAGAGGAA
CAGAGAAGAGGACGATGAGCTGTTCTGATCAGTGCAACGATGACGTTTTGTTCTTAT
CTTCATCACTTAACAGAACGCTCCGTTTCATTTAGTCGTTGGTGTAGAGAGGAAGAA
ACGCTCCGTTTCACTTGAAGCTTGGAGAGTGTGAAGTTTGGCGGGAAAATGAAAAA
TCATTTCCAATA

>band_4
GCACTGATCAGAACAGCTCATCGTCCTCTTCTCTGTTCCTCTGCATTTCTGCAAATTC
CGTTAGAGCCTCAAGCTCACCTACTCTCTTTCAGCCAGCTACCAGCTGTGCACATTG
TCTTTTAGCTTCTCATCACTTTGTATATGTACCACCTTTCTATCAATGAGAACGTTGA
GCCATTTCATTTGAACACAAACGAGTTAATACAATCAAACTGGTTAAGATTACATAC
ATGAGCCTGCTTTTTTGTCCACCTATTCATTCGGACATGTTAATGAATTAAAGAGTTG
TGAAGAAATCAGCCCGTGATAAAAACTGAAAATGTGAATAATTGCGTAAGGCTCAG
ATCTGCCATTTGTTAATCAGACGGCGGTGGTCCAAGGTAGCAAGCCCGCATGATCAC
CGGCAGTTGATGGCTGCGAAATCGGAGCAGCTAGCTAGTGCTGCCTCGATCGCCAC
GCCCAACAGCCAGACCGAGTGAACTGGGACCCTCCACCTCAAAAATCGCCGTCCGA
GTAAAAACTCTCCGGACGTATCATCACAGCCTATGTTCCCAGAAGAAGTAAGAAGA
AGAAGCAGAAGAAGAAGTAGTATATATGATTAATTCCATGGAAACAGAGACAGACA
GAGATAGATCATAGATAGATAGATATATAGCTAGCTAGTAGTAGTAGTAGTGTTAG
CAGCAGCAGCAGTGAGGGTGGTGATGGTGGAATATAATGGATGGCGACGGCGGGTG
TTGGGTGCCGGAGAAATCATCGGCACTCACAGCCTCATCCTCGAACGAGAGGGTGG
CAGAGAACCTTTATATTAAATTAATTATACTTGGCTTTCTGTAGTTACAGGATTGCCC

TTTTGTTCTCGTCTTGTCTGTGCCAGTCCACAGGAGAAAAAAGAGGTGGTGGTAAAC
TAGTAATTAGAATGAACATGAGAGGGTATTTCCGTGTAAGCC

>band_5
TAAGAACAAAACGTCATCGTTGCACTGATCAGAACAGCTCATCGTCCTCTTCTCTGT
TCCTCTGCATTTCTGCAAATTCCGTTAGAGCCTCAAGCTCACCTACTCTCTTTCAGCC
AGCTACCAGCTGTGCACATTGTCTTTTAGCTTCTCATCACTTTGTATATGTACCACCT
TTCTATCAATGAGAACGTTGAGCCATTTCATTTGAACACAAACGAGTTAATAAGTGA
GCTGATTCAGGAAGCGGTTTCTGCTGCTAGGGGTGAGCCCTCTGATGAGAATTTGGT
GAGTTTGTTGAGTTTGATGTATGGGTACAGTTCGTTTAGGGATGGGCAACTTGAAGC
TATTAAAATGGTGCTTGATGGGAAATCGACCATGTTGATTTTGCCCACTGGAGCTGG
AAAATCACTTTGCTATCAAATTCCTGCCGTTATTTTGCCTGGGATACTTTAGTAGTAA
GCCCGTTAGTCGCATTGATGATTGATCAGCTTAAACGGTTGCCTCCAGAGATTCAGG
GTGGTCTTTTCTGTAGCAGTCAGGTAGTTTTCTCTCTTTATCTCTCTCTTCAATGGCCT
TTTGCGTGTACCATCATGCTGTTTTGTTTAATGTAGCTTCTTTTGAGTTCCATAGACG
CCTGAGGACGTTGCGGAACAATCAGGCAGCTTCAGCAAGGAGCCATTAGGGTAAGC
TAGGTTAAGGTTTAGTTATAAGAAAAATCTATCTGTTTCCTGTGTAGTTGGGATGTTT
GGTGAAAATTTATTGTTTGTTCCAGGTGCTATTTGTTTCGCCAGAGAGGTTCCTGAAC
GCAGATTTCTTGTCGCTTTTGTCTGAGATTCCTGTTTCCCTTCTGGTGGTCGATGAAG
CTCACTGTATCTCTGAATGGTGACTTTCCTGTTGCTATGGTCCTCGCATTTATAAATG
GGAGATGGGGTTTGTCATTG

>band_6
TAAGAACAAAACGTCATCGTTGCACTGATCAGAACAGCTCATCGTCCTCTTCTCTGT
TCCTCTGCATTTCTGCAAATTCCGTTAGAGCCTCAAGCTCACCTACTCTCTTTCAGCC
AGCTACCAGCTGTGCACATTGTCTTTTAGCTTCTCATCACTTTGTATATGTACCACCT
TTCTATCAATGAGAACGTTGAGCCATTTCATTTGAACACAAACGAGTTAATACACCG
CCGCCCGTCTGAAGGTGGTCGAGGGAAGCAGAGCAGGCCGTCGCGAGATGATAATG
ATGATGATGATGATGATCAGAAGGCTCCGTCCATGTCGATCGACGCTGAAGCAGAG
TCAGCCGTCGGCGAATCCAATAATGGTTTATTCGAGGATGATATCGACAGATGTCCT
ACATTGCATTGCAAAATTCCCGCAGGTAATGACTGATACATACTAGTTGCTTGTTTC
AGTGAGTTGTTTTCTGATTGTTGTTTCATATACAGGGAACGTTGTAAAAAATACAGG
GAACTGGCACTTTGCTGGAGGACAGGGTTGTGATGGTAATGAACAAAATCAATGAT
CATTTTTAATTATGATATTATTTCATTATCCATAAACAATAATTTTGAAGGTTAATAT
TCTTATTCTGGTTTTTTCCCGACGGAATCGCTTGTATTAGTTGTAATGCCAAAGGCAG
AGATCGAAGAGACGGAGAAGCTGGAGCGATTCAAAATCAAGCCCCGGAATTTCATG
AATGTCGGAACCACCAAGTAATGTGTCCACAGTATGTAACACTCCGATTTTTCCGAT
AAAATATTGATCGATTTGGCCATAATTAAGCTCTTATAGATTTGTTTTCGAGTGCATA
ATAATGACTTCTCAAGAGATCACCCATTTACACTGTTACTCTCTGAATNTGTTTAACT
TCANAT

>band_7
TAAGAACAAAACGTCATCGTTGCACTGATCAGAAAGCTCATCGTCCTCTTCTCTGTT
CCTCTGCATTTCTGCAAATTCCGTTAGAGCCTCAAGCTCACCTACTCTCTTTCAGCCA
GCTACCAGCTGTGCACATTGTCTTTTAGCTTCTCATCACTTTGTATATGTACCACCTT
TCTATCAATGAGAACGTTGAGCCATTTCATTTGAACACAAACGAGTTAATACACGTT

GACTTCCTAGTCTATGGTTACTCGTTACGTGTCACGGATTCAAAATGATCCAAAGTT
CTATATCGAAATCCTAACCGTTGATCAATTCTGACAGATTCGACCGGGATTTGAGAT
GGAGGACGGGAGTGCAGCTGGTACCAATCGAATGCCAGAGCTCCAACAACTTCAAG
CTGAGGATTGAAGATTTGGAATCTGCTTACGACATGGCTAAACTCAACAACGTTCGG
GTCAAAGGATTGCTCTTGACCAACCCATCCAACCCGCTGGGTACAATTCTCGACGGG
AACACTCTGAGAAGCATTGTCTCCTTCACCAACGACAACAACATTCACCTCATCTGC
GACGAGATCTACTCCGCCACCGTCTTCGACAAGCCTGATTACGTCAGCGTCGCCGAG
GTCGTCGACGAATACCTGAACAACGCTAACGACGATGGCGAGGGTGATGATGGTAA
TAGTAACGGACCCAGGCCCACTCTGAATCTGGACCTGATTCATATAGTGTACAGCCT
CTCAAAGGACATGGGCTTCCCGGGTTTCCGGGTCGGGATAATTTACTCATACAACGA
CGTCGTAGTCAGCTGCGCAAGGAAGATGTCGAGTTTCGGTTTGGTTTCGACCCAGAC
CCAGCACCTGATCGGGTCGATGCTCTCTGACNATGATTTCGTCGACTAT

>band_9
TAAGAACAAAACGTCATCGTTGCACTGATCAGAACAGCTCATCGTCCTCTTCTCTGT
TCCTCTGCATTTCTGCAAATTCCGTTAGAGCCTCAAGCTCACCTACTCTCTTTCAGCC
AGCTACCAGCTGTGCACATTGTCTTTTAGCTTCTCATCACTTTGTATATGTACCACCT
TTCTATCAATGAGAACGTTGAGCCATTTCATTTGAACACAAACGAGTTAATATGTGG
GTCAGTCAAATTACTCAAATTTATGAGTTCACTGTGCTTTGATTTTCTTTCATTCTGTT
AGTTTCTGGGTTTCTCACTGTTTGGTTGATTGAGTTCATCTCCAAAATCGTTGCTTGT
TAGATTCCCCCCCTGCATGATTTGTGCATTTTCACCAGTGGTGGTTGATTCATCACTT
TTTAACAAGTAGGTGCAATAGCATGAACAGTTTCATGCAGGCTCCAAAATATCCCCC
ATTTTTGCTGTGAGTTCATATGATCTTGTAGAGAGGGGTTTTCATCTCAGTGCCAAGT
TGCATGTCAGATTCCCTCGACTCCGCCAAAAAACAATTGTGTGTTGATAACTCGAAT
TGTTAGGAGAGAGTTAAAGATGGATTTTTATATTACTTAACAATACGCCTTTTGTTCC
CCTCAAACGGAAGGTTCTTAGAATAGTTGATTGTGTGCACAGTTTTGACAAGTCATG
TGCTCCGAATGTGCTCCGAATTAGGGCTACGG

>band_10
TAAGAACAAAACGTCATCGTTGCACTGATCAGAACAGCTCATCGTCCTCTTCTCTGT
TCCTCTGCATTTCTGCAAATTCCGTTAGAGCCTCAAGCTCACCTACTCTCTTTCAGCC
AGCTACCAGCTGTGCACATTGTCTTTTAGCTTCTCATCACTTTGTATATGTACCACCT
TTCTATCAATGAGAACGTTGAGCCATTTCATTTGAACACAAACGAGTTAATAGTATT
TGTCGGGTTGTTTCACGTTGACCCTCGGTGGGAAGTAGCAGAAGCGGTTCTCCTCCA
CGTCGGTCTTCATTCTTCTGATTTCGGAGTCGGCTTATTTCTGAGGGTCTGCAGCAGG
GCGATTTCCGGGTAGTGGCCGTTCGCGAATGGTGGTTTCGTTCATATAATAATCCCA
ACACTTCTTCAACAACTTTTCTTCGTTTTCGAACACATCTTTCTGTGAAAACGACCCA
CTAATTCCTGCAAATACCACGTCCTTACAGTACAGAAGATACGATGTCAAGTCTTCG
GAAAACTCACATGGTCCTCCAATCTTGCTATACTCTGTTTCAAGTTTGTGCAGAAATT
GCTCAGAGTCGAGCTCTGACGTATGAGTCGTCGGTGCACACTTGGATAAGCTCCTTT
CTCATGCACCAAAACCATGCAATTTATAGA

>band_11
TAAGAACAAAACGTCATCGTTGCACTGATCAGAACAGCTCATCGTCCTCTTCTCTGT
TCCTCTGCATTTCTGCAAATTCCGTTAGAGCCTCAAGCTCACCTACTCTCTTTCAGCC
AGCTACCAGCTGTGCACATTGTCTTTTAGCTTCTCATCACTTTGTATATGTACCACCT

TTCTATCAATGAGAACGTCGAGCCATTTCATTTGAACACAAACGAGTTAATACGGCT
ACTGCGGATCATGATATGATGACGAGGATCACGAGCGGTGATCACGTGGCGGACCT
GCAGGATCAGTTCGAGAGCGGTGACGATGATGATGCGAACGTGGGTGAGTTTGGTT
ATAGTCTTGATGATGACGAGGATGATGACGTGGACGACGGTAATTTTATTCGTGGTA
GTAATAATAATAATGATAATAATAACACTGGATTTGGGGATGACTTCTCTTCTTGGA
TATTCGGTAATAGCTCTTCCACCGGCGGCGGGGGGTGTTAGTGTAATTTTGCTTTTTG
GTTCCTGATTTACTTTGTAGGTTAGGTAATTAGGTGAAAAGTGTGTAGAATTTCATGT
AAAAATAATTCATTCCTCG

>band_12
TAAGAACAAAACGTCATCGTTGCACTGATCAGAACAGCTCATCGTCCTCTTCTCTGT
TCCTCTGCATTTCTGCAAATTCCGTTAGAGCCTCAAGCTCACCTACTCTCTTTCAGCC
AGCTACCAGCTGTGCACATTGTCTTTTAGCTTCTCATCACTTTGTATATGTACCACCT
TTCTATCAATGAGAACGTTGAGCCATTTCATTTGAACACAAACGAGTTAATACTCGG
CTTTTCACAAGCTTGATGGCTAGCTCTCCGCTCCTCGAAACCCTAGAAATCATCCGTT
ATGTTTACGGGATGAGGAACCTAAATTTTCCCAATCTGAAGACCCTCAAAATTTCAA
CAATCATCGACAGGGACAAGAGTACTGATGGATTGTTCATGGACGAGTTCATAGCC
CCTCAGCTGAATACTCTGGAAATTGATAATTGTTTTTATTTGAGATTGAGTGATGTAT
CTCGGGCAGTTTCTAAGCTCGAGAATCTGAAGTACTTGACCCTTACTCGATTCGATC
CACCAGAGAAGACACTGAAACTTTCGTGTCCCAAGCTCGAG

>band_13
CGTTGCACTGATCAGAACAGCTCATCGTCCTCTTCTCTGTTCCTCTGCATTTCTGCAA
ATTCCGTTAGAGCCTCAAGCTCACCTACTCTCTTTCAGCCAGCTACCAGCTGTGCAC
ATTGTCTTTTAGCTTCTCATCACTTTGTATATGTACCACCTTTCTATCAATGAGAACG
TTGAGCCATTTCATTTGAACACAAACGAGTTAATAGATCCTAAGCTGGATCCTGCCA
CCCTTTTATTCAACGACTTTTCTGTTTGATTCTTCTTCTTCCTCCTCCTCTCTATTTATG
GTTATGTCATCTCAATTTGTTTCGTAATTGAATTTGATCAGAGTCGTCAGTCTTTACA
CTCATTTCACTACTTTTTCAGTTTATTTCCCGTAATTTTTCCTTTTAAAAAAAACTGGA
TGAATGGAGGGTAATTTGGCCAATGCACAATTTACTCATTTTTTTCGGTGTTGCTATT
CATTTAATGAAATTTTGAAAGTTGTTACACGATGATTACAG

>band_14
TAAGAACAAAACGTCATCGTTGCACTGATCAGAACAGCTCATCGTCCTCTTCTCTGT
TCCTCTGCATTTCTGCAAATTCCGTTAGAGCCTCAAGCTCACCTACTCTCTTTCAGCC
AGCTACCAGCTGTGCACATTGTCTTTTAGCTTCTCATCACTTTGTATATGTACCACCT
TTCTATCAATGAGAACGTTGAGCCATTTCATTTGAACACAAACGAGTTAATACACAA
CATTCTTCGTAAGAAATTCCTAGCTTTGCAAAAGAAAACCAAAAAAAGTTTTAATGC
TGCTAGCTATTACTCCGTAGAAGCATTTCACATGTTAAAAAGGAACTTACCTGAAAG
GAAAGCAAGTGACGTGGAATTGAAGAGATAGGTAGTCCACACAGAAGCCTAGGAA
ATGAACAATTTAGCTGCTAAAGAATTGACAGAAAACTGAAGGCAGAGGCACGACA

>band_15
TAAGAACAAAACGTCATCGTTGCACTGATCAGAACAGCTCATCGTCCTCTTCTCTGT
TCCTCTGCATTTCTGCAAATTCCGTTAGAGCCTCAAGCTCACCTACTCTCTTTCAGCC
AGCTACCAGCTGTGCACATTGTCTTTTAGCTTCTCATCACTTTGTATATGTACCACCT

TTCTATCAATGAGAACGTTGAGCCATTTCATTTGAACACAAACGAGTTAATAATAGA
TACGGCGGTTTTGCTTCAAGGAAGTCTAATGTTAGGAATTGGGCTGATTCGGATGCT
AAGCCGGCAAAAGATTATTACTTTGACGGCCATGGTGATCGGGATAATTTAGCTTAT
GGCTCACTCTACAGGTTTTCGCTTTTGCTTTGTTTACACTGGAAATTATGCAGGATTT

>band_16
TAAGAACAAAACGTCATCGTTGCACTGATCAGAACAGCTCATCGTCCTCTTCTCTGT
TCCTCTGCATTTCTGCAAATTCCGTTAGAGCCTCAAGCTCACCTACTCTCTTTCAGCC
AGCTACCAGCTGTGCACATTGTCTTTTAGCTTCTCATCACTTTGTATATGTACCACCT
TTCTATCAATGAGAACGTTGAGCCATTTCATTTGAACACAAACGAGTTAATAACTAT
CAACCACAAGGACCAGGCCAAGTGTGTTGCGAATTGCGATGCGATGTCTATCTGAAT
CGAGATATTTTTCCTACACTAATTATAATTAATTAAGTAGATTTAAACAAGAATTTC
AGTTCCAATTATTAGAAGCATT

>band_17
TAAGAACAAAACGTCATCGTTGCACTGATCAGAACAGCTCATCGTCCTCTTCTCTGT
TCCTCTGCATTTCTGCAAATTCCGTTAGAGCCTCAAGCTCACCTACTCTCTTTCAGCC
AGCTACCAGCTGTGCACATTGTCTTTTAGCTTCTCATCACTTTGTATATGTACCACCT
TTCTATCAATGAGAACGTTGAGCCATTTCATTTGAACACAAACGAGTTAATACTTTTT
CCTGCCACTGGTAACCCGTAGAATGGCCTGGAAACGCGGCGAATGCATTTCCTTGGC
AAACGCCGCATGACGACTCTACAGATAATTGGACA

>band_18
TAAGAACAAAACGTCATCGTTGCACTGATCAGAACAGCTCATCGTCCTCTTCTCTGT
TCCTCTGCATTTCTGCAAATTCCGTTAGAGCCTCAAGCTCACCTACTCTCTTTCAGCC
AGCTACCAGCTGTGCACATTGTCTTTTAGCTTCTCATCACTTTGTATATGTACCACCT
TTCTATCAATGAGAACGTCGAGCCATTTCATTTGAACACAAACGAGTTAATAAACTG
CGAGAGCTTGGAGCATCTATTTTTGCAGGA

RLC_Lu28-primer1 * EcoRI

>RLC_Lu28-1 (LTR_sense)
TGTTGAAAAATATTATATTTTCCTTATTAGTATAGGAATAAGTTTCAATATTTTTCCT
TATTAGTATAGGAATAGTAATAAGAGTTTTCCTAGTTGAGGAAGGATTCTCCTATCC
TAACTCTATATAAACCCATGTACCCCTTATGTAATCTCATATATCATAATACCATTGA
AAACTTCCTCTCATAAATTCAATA

> RLC_Lu28-1 (LTR_reverse)
TATTGAATTTATGAGAGGAAGTTTTCAATGGTATTATGATATATGAGATTACATAAG
GGGTACATGGGTTTATATAGAGTTAGGATAGGAGAATCCTTCCTCAACTAGGAAAA
CTCTTATTACTATTCCTATACTAATAAGGAAAAATATTGAAACTTATTCCTATACTAA
TAAGGAAAATATAATATTTTTCAACA

>band_1

CCATATATCATAATACTATTGAAAAGCTTCCTCTCATAAATTCAATAAAGATCAACG
GCATACTCCCCAAAGTTTGAGTGACCAAGTAACCGGGCATACTTGTGCCGCAGTTCA
ACCTAAAAGAAGAAAGGAAAGATTCTCTTCAGGATACTACTTTAACAAAAGGTAAC
CAGGTCATATGCTTCATCATATCATTCCAAAGTTACATCAGTAAGTAGCAACACCAA
TCAAACACGGACTGCTCATGATAACTGACAACAACCTTGAGCTACTATTTCCTTGAA
TCAATGCTACTTATTTCTGAATCTTGATAGTTAGCTTTTTTATTTTTCTACTTTTCGGT
CCGTATGTTCTTTCTTTGTTAGGATAGGGAGTATATATAGAGCATAAACGACTGCAG
GTACCTTTAGCCAACAGCCGACTCGAATGTATAAGTAAACAAGGTTTAATAAAATCA
TCACTTGTTAGAAGTGTTCAAACAGTTTAGTAATTTCTTCAAAAGTTTAATACATAA
ACCCAAAATAGGAATAAGGTCACCGAAGGACGATATCATCCTAGGAATCTATTGAA
ACTTAAATGAATTGCTCTAACCGGTCAAACATTAAAGAATCTTACCAAACTTTCGAA
AACTGAAAGATTGATATGTCCACATCTCTTCCCGTATGCAACCGCAACTCGCTTTCT
GGTTTGTCCTACCTGTATACATTAAGAAACATAACACGGCCTTAACCCACCATTATC
CAAA

>band_4
CCATATATCATAATACTATTGAAAAGCTTCCTCTCATAAATTCAATATATACTTTACA
TATTGATTTGTTCTTCAAACGTTAAACACGTGAATGATGCTAGCTAATGTCTGCTGCT
GCTAACTCTGTTATGCAGTGGTTTGGAGATGCTGAGAAGTTGACAAAGGCCCTTTTC
TCCTTTGCCACCAAGCTTTCTCCTGTTATCATATTTATCGACGAGGTACCTCCCTAAA
CTCTCTTTTTCACTAGGGCTTTGTCAATCGACATGTTCGACTTAACATGTTGTGGTTG
GCGGGGGCGGAAAGAAAAGGTAGATAGTCTACTAGGGGCTCGAGGGGGTTCTCACG
AGCACGAGGCTNCGAGAAGAATGAGAAACGAGTTCATGGCAGCATGGGACGGATT
AAGATCGAAATACACTCAAAGGATCGTCATCCTTGGTGCCACGAATCGNCCGTATG
ATCTTGATGATGCTGTGANTCGNCGTTTNCCTAGGAGGTAAGGCGCGCNTGGAACNC
TCTCTTTGTCGTCCGANGANTTTTATAAATGCCNTCGGAAANNCANCTGTGCTCCTN
GTTTAGAATATNCGCG

>band_6
TCATATATCATAATACCATTGAAAACTTCCTCTCATAAATTCAATANATACCAAAAG
ATCAACTGNTNTGTTGACAAATTGCACAATCTCANNCAGCAATGTAGACCGTCCATA
AACAAAACCTNTGCAACAAACCACAAAGTAAACTCCACTTCATTAAACAAGCGAAT
CATCTTCACTTCGCTGGAACAACAAAAACCACCATCTCGCTCTTTCCTTTCTTCAGTG
CTGCCTTTTTCTTTCCAAGTTGATGATCATCAGTTCTTGAAACACCAACCACACTTTG
GTTGCTTAACTGCTCCAAAAACTTTCCTAACCCTTGAAATCTCCAGAAACAATTGGT
TTGGCCATCATATCTTGAAAGATTCACAATTGAATCATGTTCAAAACAGTATATAAT
GCATCGATACTATCCGATGGACCATGAGAAATCGATTATAACCTGAAGAAGAAACT
CAAAACTAATGTTTTCAGAACAATTGTACCTGAGATGGTGTTATGAGTTTCAAAACT
AATGTATTCTAACTACCAACTATGATGAATCCAAGCTCTCTAAACGTTTAACACAAC
AAAAATTGTGTGCATTCTATATCTATCTC

>band_7
CCATATATCATAATACTATTGAAAAGCTTCCTCTCATAAATTCAATACATACCTTAA
AACGAGCATGCAAGGTCAGCAAAATGTCATTCAGCAGAGCAGCTGGCCAAGCTGGA
TCAGAAAGCTCCGAAACAATGATTGATTCCAGCTCATCCAATGATTCATATGGAAGT

TGCATCCATATCAATGCTTTGATCACCATTGCTCGAACATAAACACATTCACAGGCA
ACAGTTGTTCGAACAACCTCCATCAGTGATGCTAATAAACTTGCAATTGTTTCCTTGC
TACCAGTTGAACTCGATAAAGAAGAAGCCTTTCGGGCACCTGAAATAACAATGAAT
CAATGAGAGGAAATGGGGGTACAAATGCGATGTAAAAAAAAATCAATACTGCAAA
AGCCAGAAACTCCTAATTTCTACCGAATCAAATTCAATGTAAGAAGGCGGAACCAT
AGAAACCCAAAAATCTTGAGACCATGTGAATAAACTAGGAAACTCCTTATACTGCA
TAAAAGGAGTAACAAGTGAGGTTTC

>band_9
CCATATATCATAATACTATTGAAAAGCTTCCTCTCATAAATTCAATACAACTCNNAC
CCCCGCAGGNCCACANAAGNAATTTCNGGAAAAAAAAAAATAGAATCTCTACCTCCT
TTCCCATTTTCTCAACTTTACCTTCTGTTTCATGCATTCCACCTTTTGGTGTAATCAAT
CATAGGGAACCTTAAATCACTAATTAATTAACATGCTATGTGAATGATGGCTCTTCT
TCAAGGCAAACATCCCATCTTTAGATGTTTCAAACACATTTAACCCATCGATGGTAA
TACCCTTTTAATGATAAGCTAGATTTTCAGAGTAAATAGGTATAATAGTAGCTCATA
AGTTAGATCAAACCTTGACGTACAATACCTCATCCAATCAGGTAACGCGAGTGAATA
GCCTAATGAGAGGCTAAGGAATGCTCGATCAAGGNGNAAAGANAAATTTTAGGATA
TTTTTATGTC

>band_11
CCATATATCATAATACTATTGAAAAGCTTCCTCTCATAAATTCAATAGTCTTGTTGCC
ATGCTTATTTGCATTTGACATCAAAAGCCAAATTTCCAGGCAAGGAGCTGTTGTTGA
CGAGGTTTTCCAAAATGTGTTGTCTCTTTTGAGCAAGAACTACACTTTAGTGACTGAT
GAGTTAGTTGGAGTTGATCATCATGTGGAAGAAGTGATGAAATTACTGAATTTGGGC
TCAGGATGTGTGACGACTGTTGGCATTTATGGAATGGGCGGAATTGGAAAAACCAC
CATCGCTACAGCTGTCTATAACAAAGTCTGCACGCTTTTTGACCGTTGTAGCTTTGTT
GATGATGTAAGAGAAACATTGTCATGGAGTGATGGTATTGTCACTTTGCAGAATAAG
CTCATCTATGGCATTACGAAAGATGGCTCTCCCATTGGTAGCACAAGTGAAG

>band_12
TCATATATCATAATACCATTGAAAACTTCCTCTCATAAATTCAATACTTCCACTCATG
CGACAGTAGCTGCTGAATCTTCTAAGTTATAACTGAAACCAATAGTGTATAACTTGC
ATAATAAACAGAAAAGTCTACGAGAAGGAAAAAATTAACCTACGTCAGTCTAAGGT
GTAGCCTGAACCTAGTTTATATATACAAACTAATATAGGACGATATATCATACAAGA
GATCTTAGCCCCTTCTCCACCAGCAATCTATATATCGAAAACTGTCCAGGCTGCATA
CCTTTGCCATTAATATTGTCAACTGAAATATCTTATGAGAAATATGTAAATCCCATCT
CTTCGAGAGCTGGAGAATAGATTTTGCAGCAGCTAATCGGATATAAGGCTTGTCACT
TGCACTGCCCAAAG

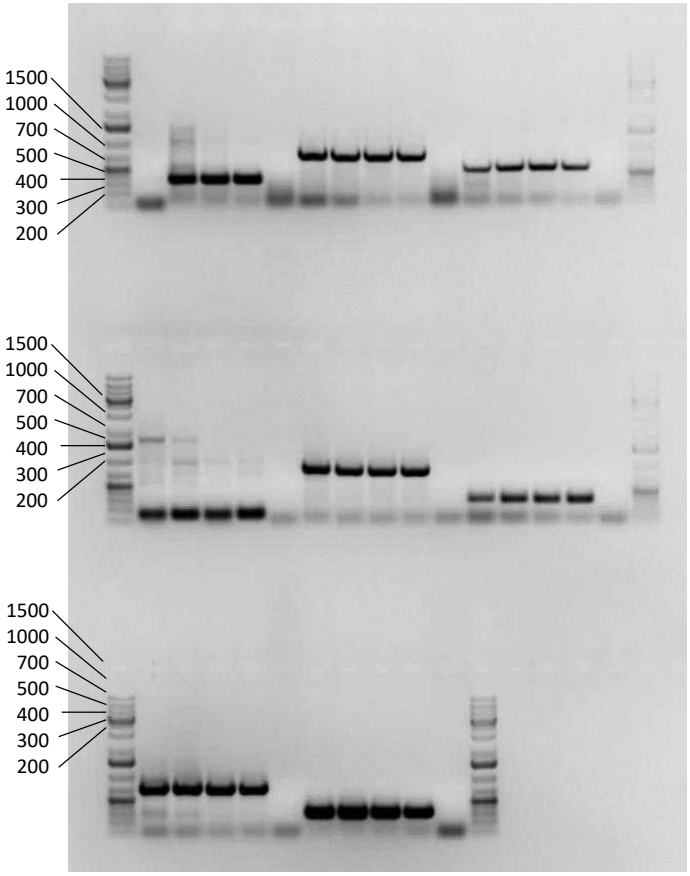
>band_14
TCATATATCATAATACCATTGAAAACTTCCTCTCATAAATTCAATAATTTATCATCTC
TAGTCAAAAGAACATTGAAATGTAATTCGTATAGGGAGTTCATATATGCCGTACTGA
AATACTTGTCGTAGTTAATCCGGGTACCGTAGTTACGTTGTCTTTCCAGTGAACAAG
AGTTTCCATTAACCGCCGCTATTAAAATCTGTCGACATCCCAGATGGTTCAAGTGTTT
TAAGAGTTT

>band_20
CCATATATCATAATACTATTGAAAAGCTTTCTCTCATAAATTCAATACATTCAAATGT
CGCTTAAGCTGAAATTTGAAGTATGATTGTTCCAGATGCACAATGTGCGGTGGAAA

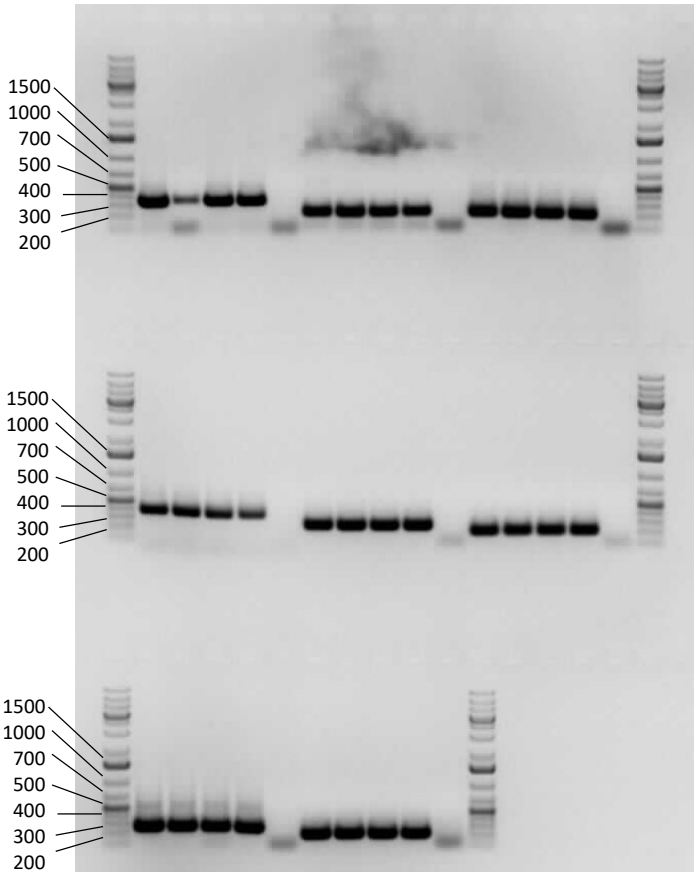
>band_22
CTATATATCATAATACTATTGAAAGCTTCCTCTCATAAATTCAATAGTTCCTTAAAAT
CCCCTGCAGCGTCTACTTTGGTGGCCCATTCAATTTTTTCAGCTGCACT

Appendix 3.1 Validation of primers for TE stress response. The two gels represent all primers evaluated in experiments 1 and 2 for end-point RT-PCR. A gradient PCR as performed with extension temperatures of 52, 55, 58 and 61°C which correspond to the four lanes of each primer evaluated (the fifth lane is the negative control). Expected sizes for amplicons are: 352 bp (Lus10019060), 789 bp (Lus10016872), 580 bp (Lus10028377), 143 bp (Lus10041831), 845 bp (Lus10035621), 311 bp (Lus10024366), 735 bp (Lus10035634). 326 bp (Lus10018035).

Sample	Primer	gene	amount
First (marker – 1kb)			7uL
1-5	Lus10019060	chitinase	25uL
6-10	Lus10016872	chitinase	25uL
11-15	Lus10028377	chitinase	25uL
16-20	Lus10041831	chitinase	25uL
21-25	Lus10035621	chitinase	25uL
26-30	Lus10024366	chitinase	25uL
31-35	Lus10035624	chitinase	25uL
36-40	Lus10018035	etif3e	25uL



Sample	Primer	gene	amount
First (marker – 1kb)			7uL
1-5	Lus10011375	gadph	25uL
6-10	Lus10039595	ubi2	25uL
11-15	Cl-RTs-0-a	copia	25uL
16-20	Cl-RTs-0-b	copia	25uL
21-25	Cl-RTs-1-a	copia	25uL
26-30	Cl-RTs-2-a	copia	25uL
31-35	Cl-RTs-3-a	copia	25uL
36-40	Cl-RTs-28-a	copia	25uL



Appendix 3.2 LTR sequences from representative members of each retrotransposon family evaluated.

>RLC_Lu0-1
TGTTGAATTACAGAAATAACTAGGATATTATTAGGCGTATAACTTGGAGTAATCTAG
CTATCCTATAGAATCAGTCTAGCCTGATTTAGGCAGTGTATTCATAGTAAGTGGTAG
GTAGGATAAATCCCTGATTTGTAGGGATTACTACGAGCTGGATCTGCATCCTAATGA
TGCAGACTGTGTATATATGTAAAGGAAACCACAGAAATAAGAATATCATACCAGAA
TCACTCAGATTTCTGGATTTCTGCA

>RLC_Lu1-1
TGAGGAAATCCCGTTCCTTATTTGTATACAGTCAGATCTTATTAGTTTATTTGCTAGT
TAGATAGTTACCATATTAGTTCTAGTCCTATCCTAGTTTAGCTAGCAGATCCTATTAA
ATAGTTAGAGTAGATATTTTCTTACCTAGTCAGTAGGAGGAATGCTGTATATAATAA
ACCTACGAGACATGAATAAAAGTAATTCATTCTCTCAATCTTCA

>RLC_Lu2-1
TGTTAAATGGTGTTAAATTATAAGTATGTGAGTGAGTTGTAATAGATAATAGTAGTA
GTGGTATAAATAGAATTCTCCTTTCTGTTTGTACATTTAATTCATTCAGTACAGTAAT
AACAGAAAACGACTTTCCATTACAGCTTCTCCTTCTTCTTCGCTTTCTCTGTTCTTCTC
TTCTCTTCCATTCATCAAACTTCACA

>RLC_Lu3-1
GTTGGGGCGCTAGCGGAAATGTAGCGGAATTTAGGATCGAAAAGATAGTATTTTGA
TAACTCACTCAATTGATCACTCAATTTTGTAAATATTACATTTTGGTCACTTAGACCT
TTACTCTCACCACTCACAAATGCCTCTCTACTTCTTCTTTCTTATTTCTTCAACTCACA
CTCCACTTTGTTACAAGTAAGGCTATTTATAATAGCCATGATACAAGTAGTAGGTAG
AATTTGTGGTTTGTAAAATTTTAATTATTTTAACATATCACATCTTAATTAATTTGGT
AGTTAGGAGTTGAGATATTTTTGTAAGAGTGGTAGAGACTAATCTAGAGAATTGTAG
ATTAATCTTCAACAC

>RLC_Lu6-1
TGTTGGTCCCGGATAATTGATGTGCAACAAAAATATGTACCACCAATATTAAAGCAC
CACAAAATTAGTTGGATCTCTAACCAACTAATTGAATTCCCACCAATATTAACCCCC
ACCTCCAATTTAGTGCACATGGAAATTAGTTGGGAATTCAACTAATTGGATGCCACA
ACCAATTGCTTTCAATTGGTCTCCCCACATTTCTAGTATAAAAGGGAAGCTAGTGCA
TCCCATTTCAATCATCCCTCTCTTCTTCCTTTCTCACTTCTCTAAGTGTTGTAGTGTAG
CAATTTTCACTTGTTTAATAATTGAGATAAGTTATCTCAATTGGGTAGATAGGTGAG
CGGTAGAAAGTCCCGGTAAATGTTTTACCGTGGTAGGAATACTTTCTTGTGAGCGAT
AAAATAGTGAGTAGTTGTTTCGGGGTTGGGAAACACTTGCGAGACACTATTTTGGAT
CGGCTCGGATCACCTTGTAGCTACCTTGTTATAGTGAAGAAGTGCTCGTAGCTGTCG
CTGCTGCCGTAGATGTACTCTCCGCATTGGAGGGGAACTACGTAAATCCCGGTGTTA
TTTACTTACTGTTTTGTGCTTGGCAATTTCGAGAATATTCGTTGTATATTGCATTATTA
ATATTACCACAGTAAATTGGTCTAAGGAGGTTGGCTTAATTATCGTCATGATGGTAT
TGCGGTGGTAATCACCCATCCATAGTGATTTTAAGTGTGGCGGACTACCGCCACTTC

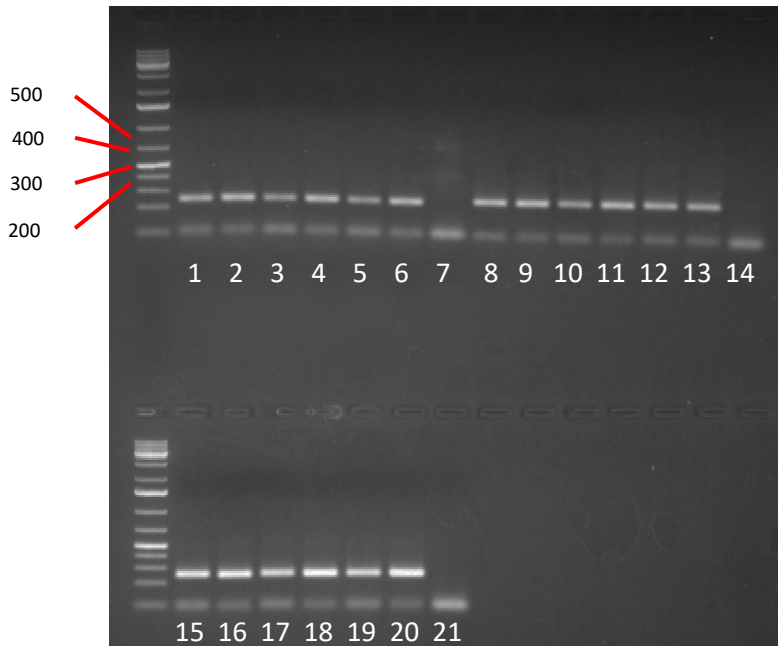
CAACTTATCTGGGAAATATTTACGGTGTGTGGTTTATTAGTGCAATATATTTACTCTA
TTCTCGTCCGCTGCGCCCCAACA

>RLC_Lu8-1
TATTGGAAATGATTTTTCATTTTCCCGCCAAACTTCACACTCTCCAAGCTTCAAGTGA
AACGGAGCGTTTCTTCCTCTCTACACCAACGACTAAATGAAACGGAGCGTTCTGTTA
AGTGATGAAGATAAGAACAAAACGTCATCGTTGCACTGATCAGAACAGCTCATCGT
CCTCTTCTCTGTTCCTCTGCATTTCTGCAAATTCCGTTAGAGCCTCAAGCTCACCTAC
TCTCTTTCAGCCAGCTACCAGCTGTGCACATTGTCTTTTAGCTTCTCATCACTTTGTAT
ATGTACCACCTTTCTATCAATGAGAACGTTGAGCCATTTCATTTGAACACAAACGAG
TTAATA

>RLC_Lu28-1
TGTTGAAAAATATTATATTTTCCTTATTAGTATAGGAATAAGTTTCAATATTTTTCCT
TATTAGTATAGGAATAGTAATAAGAGTTTTCCTAGTTGAGGAAGGATTCTCCTATCC
TAACTCTATATAAACCCATGTACCCCTTATGTAATCTCATATATCATAATACCATTGA
AAACTTCCTCTCATAAATTCAATA

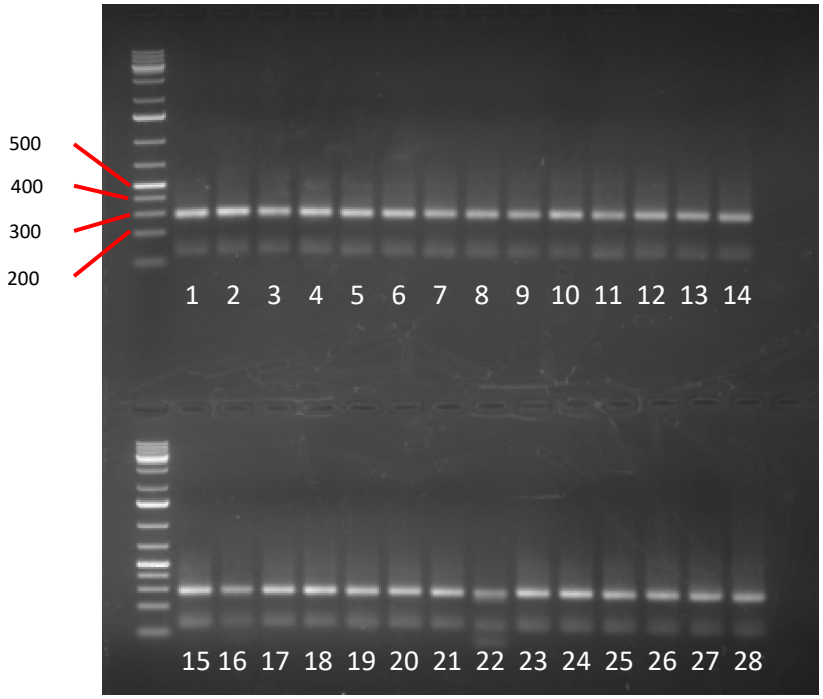
Appendix 5.1 Two-step PCR from pilot experiment.
A. Products of first step PCR from the pilot experiment. Amplifications were performed with 5 or 10 ng of pooled DNA in the different dilutions that simulated the inclusion of the mutated individual. Gel was run in 1.5% agarose in TAE 1X at 90V for 40 minutes. Size of marker bands is given in bp. Negative control refers to a PCR with no DNA template.

Samples	primer	DNA-dilution	proportion cultivars
First - marker			
1 - 2	S20	5ng-10ng	Macbeth:Bethune 1:96
3 - 4	S20	5ng-10ng	Macbeth:Bethune 1:64
5 - 6	S20	5ng-10ng	Bethune 100%
7	S20	negative	negative
8 - 9	S411	5ng-10ng	Macbeth:Bethune 1:96
10 - 11	S411	5ng-10ng	Macbeth:Bethune 1:64
12 - 13	S411	5ng-10ng	Bethune 100%
14	S411	negative	negative
15 - 16	S900	5ng-10ng	Macbeth:Bethune 1:96
17 - 18	S900	5ng-10ng	Macbeth:Bethune 1:64
19 - 20	S900	5ng-10ng	Bethune 100%
21	S900	negative	negative



B. Products of second step PCR from the pilot experiment. Amplifications were performed with 1:100 dilutions of the first-step PCR mixed products of the three genes (S20, S411, S900). Gel was run in 1.5% agarose in TAE 1X at 90V for 60 minutes. Size of marker bands is given in bp. Negative control refers to a PCR with no DNA template.

samples	tRP1 and barcoded primer
First - marker	
top 1-14	barcode primers 1-14
bottom 1-14	barcode primers 15-28



Appendix 5.2 Sanger-sequenced fragments of references genes used in the different experiments.

Pilot experiment

>scaffold20_(203bp)-S20
CTCCGTCATGGTATTAGTCATGAATTTACTACTTTTTTCACTGCACATTCCATGACAT
AATCTGATTCTAACTGCATCGTTTTGTTAGAGTTTCTGGCTCGGATGGATTGTTAAAT
CATGGCCACTATTTTGGGCTCTCTTCATCAGTTTTGTCCTTGGCACTGCCTATTCAAT
CAATGTAAGTTCAGTTCGGCTCTATGTTC

>scaffold411_(209bp)-S411
GGAAACATACTGGTCCTCTTCGCTGGTCATAAACGGTACACGGACCGATCCATCATT
CAGTAGGTGAAAATCATAGCGTTTAGTTGCCGATGCATCGAAATTCTGGTTCCAAAC
TCCTTCGAAGAAGAGCGCATTAGCGTAAATTAGCGGTGTTAAGTTGTTAACTGCCCC
TCGAGGAACAATTTCCTCTACGATTCCGTTCGTCCGTCT

>scaffold900_(220bp)-S900
GGTTGCTACAGGATTCAGCTACGTGGAGGACGAATCGTTGGTAGTTAGAACCGACT
ACGAGTCAGCAACTGATTTAACTACCTTGTTGAAGGCATTGTACAATGACAATGATG
CCCTCCAGAAGAGCCCTCTCTATATCTTTGCCGAGTCTTATGGAGGAAAATTTGCTG
TCACCCTTGGAGTTACCGCAGTTAAAGCCATCGAAGCAGGAGAGTTAAGG

PMEs

>Lus10031470_b3*b3_(196bp)-*LuPME79*
TCCCGATGGCCCACCAGTTCAACGCGATCACGGCTCAGAGCCGGACCGATCCGAAT
CAGAACACGGGGATATCGATCCAAAACTGTAGTATCAAGGCCGCGAAGGATCTTGC
GGAGAGCAACGGAACGACTAGGTCTTACCTCGGCCGGCCGTGGAAGGCGTATTCGA
GGACGGTGGTGATGAATTCGTACATCGC

>Lus10004720_b3*b2_(197bp)-*LuPME10*
GTTACATTCAATAGCGAAGAGATTTGGATCGATAGCGGCCCAAGACAGGAAATCGC
CAGACGAGAAGACGGGCTTTGCATTCTTGAACTGTAAGGTGACTGGAACGGGCCAG
CTCTACGTGGGCCGGGCCATGGGCCAGTACTCCAGGATTGTCTACTCACACACCTAC
TTTGATGACGTGGTTGCACACGGTGGAT

>G25305_a1*a1.2_(211bp)-*LuPME73*
GTTGACGTTTAGGAACACTGCCGGGCCGGCGAAGCACCAAGCAGTCGCCGTGAGAA
ACAGCGCCGACATGTCGGCGTTCTTCAACTGCAGCTTCGAAGGCTACCAGGATACAC
TATACGTACATTCCCTCCGCCAGTTCTACCGCGACTGTGACATCTACGGCACCATCG
ACTACATCTTCGGGAACGCGGCCGTCGTGTTCCAGAACTGC

>Lus10043035_a6*a6(220bp)-*LuPME105*
CGTTCGTGGGCTGCAGATTCCTGGGCGGACAGGACACTCTGTACGATCATTTCGGGA
GGCATTATTACAAAGGTTGCTACATTGAAGGATCTGTGGATTTCATCTTCGGGAACG

GCCTCTCCTACTTTGAGGTATGTATATTAATTTTTTGAATCAGAGAGAATATGTCGAA
TTGAAAGTGATGAATTTGGGGGTAAATCGTAAATGAAAGGGGTGTCAC

Metabolism genes

>Lus10016751_3.3F*3bR_(199bp)-acetolactate synthase-1
GTTCAAGAGCTGGCAACTATTCGGGTGGAGAATTTACCGGTGAAGATGATGCTGTTG
AATAATCAGCACTTGGGTATGGTGGTACAGTGGGAAGATCGTTTCTACAAAGCGAA
TAGAGCTCACACATATCTAGGGGATCCAGCAAGGGAATCGGAGATATTCCCGAACA
TGCTGAAGTTTGCTGAAGGTTGTGGAATTC

>G24175_4F*4R_(206bp)-cyclic peptide
ACCTTGTCTCCTATTTCTGGAAAGGATGGCGGCCTCCGCAACCAGGAGGAGAGCGAT
GGTATGTTGGTCTTCCCCTTATTTATATTCGGCAAGGAAGGTAGTCAGGACAAGTAT
AATGGAGCAGCTGCCCTCCGCGACCAGGAGGAGAGCGATGGTATGTTGATCCCCCC
CTTCTTTGTCATATTCGGCAAGGAAGGTTGTCAGGA

>Lus10029955_b.1F*bR_(208bp)-acetolactate synthase-2
ATGATACTGAACAACCAGCATTTGGGGATGGTGGTCCAGTGGGAGGACAGGTTTTA
CAAGGCGAACAGAGCGCATACGTTTCTGGGGAACCCGGCGGGGGAGGAAGGGGAG
ATTTTTCCGAACATGTTGAATTTCGCCGAGGCTTGTGGGATACCGGCGGCCAGGGTG
AGTAAGATCGGCGAGGTTAGGGAGGGGATTCAGAGGATGT

>Lus10017825_aF*aR_(210bp)-UDP glucoronosyl/glucosyl transferase
ACCACAGCGGACTTATATTGCAAGACCCAGAACACTATTCCCAAACCTTTCGGAGA
GTGGGTTCCGAAACCCTCACGGATCTCATCCGGAAACAGAGCGAATCNCGGCACCC
TGTCCACTGTATAATTTACGATGCAAGTATGCCCTGGTTCCTGGACGTCGCCAAGCG
GTTTGGGATTGTGGGAGCTGCATTTCTCACTCAGTCATGCG