Genomics and proteomics approaches to study flax seed

A thesis submitted to the
University of Pune

For the degree of
DOCTOR OF PHILOSOPHY

In
BIOTECHNOLOGY

By
VITTHAL BARVKAR

Research Guide
Dr. VIDYA S. GUPTA
Plant Molecular Biology Group
Division of Biochemical Sciences
CSIR-National Chemical Laboratory
Pune 411008 (INDIA)

April 2013

This work was performed at

CSIR-NATIONAL CHEMICAL LABORATORY, PUNE, INDIA

Scientists involved in this activity

NCL: Dr. Vidya S. Gupta (Research supervisor)

NCL: Dr. Narendra Y. Kadoo (Co-supervisor)

NCL: Dr. Ashok P. Giri (Co-supervisor)

NRC-PBI: Dr. Raju Datla

CERTIFICATE

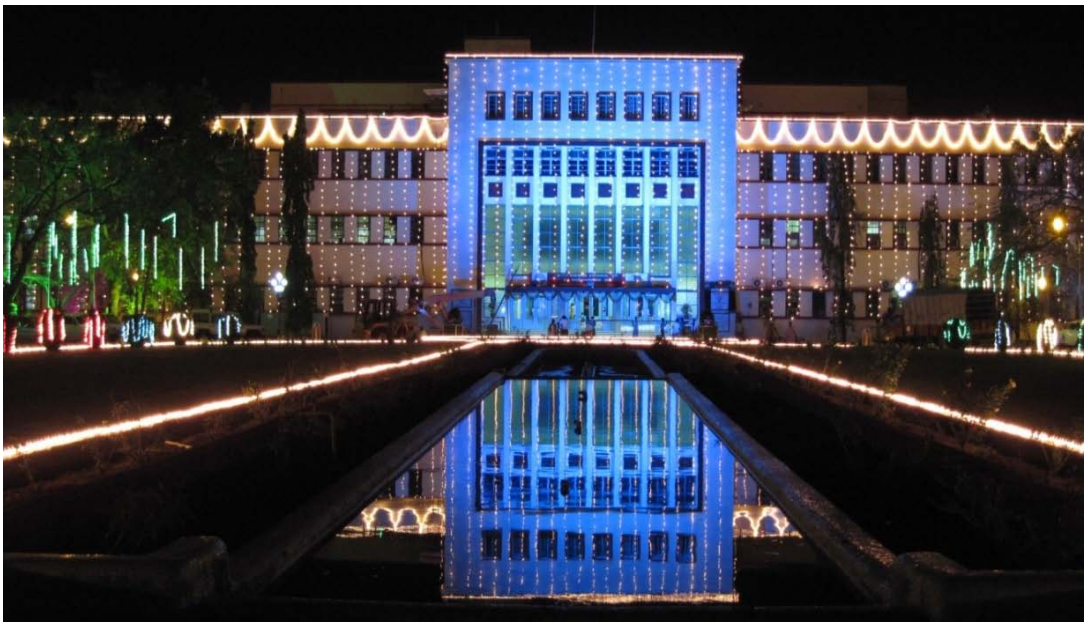
Certified that the work in the Ph.D. thesis entitled **"Genomics and proteomics approaches to study flax seed"** submitted by **Mr. Vitthal Barvkar** was carried out by the candidate under my supervision. The material obtained from other sources has been duly acknowledged in the thesis.

Date:
Place: Pune

Dr. Vidya S. Gupta
(Research Guide)

Dr. Narendra Y. Kadoo
(Research Co-guide)

Dr. Ashok P. Giri
(Research Co-guide)



CSIR-National Chemical Laboratory, Pune (India)

DECLARATION

I hereby declare that the thesis entitled **"Genomics and proteomics approaches to study flax seed"** submitted for Ph.D. degree to the University of Pune has been carried out at CSIR-National Chemical Laboratory, Pune 411008, India. This work is original and has not been submitted in part or full by me for any degree or diploma to any other university.

Date:

Vitthal T. Barvkar
Division of Biochemical Sciences,
CSIR-National Chemical Laboratory,
Pune 411008

Dedicated to my beloved parents and my wife Anupama

Contents

Title	Page
Acknowledgements	**i**
List of Abbreviations	**iii**
List of Figures	**vi**
List of Tables	**vii**
List of Supplementary Figures	**viii**
List of Supplementary Tables	**ix**
Thesis Abstract	**xi**
Chapter 1: Introduction and review of literature	
1.1 Origin and domestication	**1**
1.2 Flax as a crop plant	**1**
1.3 Flax seed as a functional food	**3**
1.3.1 Flax oil	**3**
1.3.2 Phenolic acids and lignan	**7**
1.3.4 Flax seed proteins	**9**
1.3.5 Dietary fibers	**10**
1.4 Economic importance of flax	**12**
1.5 Omics studies in flax	**13**
1.5.1 Genomics	**14**
1.5.2 Transcriptomics	**16**
1.5.3 Proteomics	**17**
1.5.4 Metabolomics	**18**
1.6 Genesis of thesis and its organization	**20**
Chapter 2: Proteome profiling of developing flax seed and characterization of functional metabolic pathways operating during seed development	
2.1 Introduction	**22**
2.2 Materials and Methods	**23**
2.2.1 Plant material	**23**
2.2.2 Fatty acid analysis of developing seeds	**23**
2.2.3 Optimization of protein pre-fractionation method	**23**

2.2.4 1D-SDS-PAGE and in-gel digestion **25**
2.2.5 Liquid chromatography mass spectrometry analysis **26**
2.2.6 Data processing and database searching **26**
2.2.7 Functional classification **27**
2.2.8 Digital expression analysis **27**
2.2.9 Reverse transcription quantitative real time PCR **28**
2.3 Results and Discussion **28**
2.3.1 Biophysical characterization of developing flax seed **29**
2.3.2 Fatty acid analysis of developing seeds **29**
2.3.3 1D-SDS-PAGE and LC-MS^E efficiently cover the flax seed proteome **31**
2.3.4 GO annotation, functional categorization and family relationship **31**
2.3.5 Transcriptional evidences correlate with proteome profiling **32**
2.3.6 Diversity of seed storage proteins in flax **32**
2.3.7 Flax seed maturation marker proteins **35**
2.3.8 Carbon entry and assimilation by glycolytic pathway **35**
2.3.9 Varied carbon flux pathways are functional during flax seed development **36**
2.3.10 Lipid accumulation is associated with oleosin and lipoxygenase expression **39**
2.3.11 High activity of β-oxidation and glyoxylate pathway might accumulate alpha-linolenic acid in flax seeds **39**
2.3.12 Flax seed development demands enhanced methionine metabolism **40**
2.3.13 Accumulation of health promoting compounds during flax seed filling **43**
I. Phenolic compounds **43**
II. Lignans **43**
III. Flavonoids **44**
2.3.14 Simultaneous biosynthesis and degradation of cyanogenic glycosides occurs in developing flax seed **45**

Chapter 3: Genome-wide identification and characterization of microRNA genes and their targets in flax and their role in oil accumulation

3.1 Introduction **47**

3.2 Materials and Methods **48**

3.2.1 Prediction of miRNAs from flax genome sequence **48**

3.2.2 Characterization of miRNA genes **49**

3.2.3 Digital expression analysis **50**

3.2.4 Computational prediction of miRNA protein coding target transcript **50**

3.2.5 RT-qPCR analysis of miRNA and target genes **51**

I. Plant material **51**

II. Isolation of total RNA and RT-qPCR analysis **51**

3.3 Results **52**

3.3.1 Identification of conserved miRNAs in the flax genome **52**

3.3.2 Characterization of miRNA genes **55**

3.3.3 Expression analysis **56**

I. Digital expression analysis of pre-miRNA **56**

II. RT-qPCR of pre-miRNA **56**

3.3.4 miRNA target identification and confirmation using RT-qPCR **57**

3.4 Discussion **59**

3.4.1 Flax miRNAs exhibit the typical features of plant miRNAs **61**

3.4.2 Diversity of expression of the flax miRNAs **62**

3.4.3 Identification and expression of target genes **64**

3.4.4 Role of miRNA genes and their targets in oil biosynthesis and accumulation **66**

Chapter 4: Phylogenomic analysis of UDP glycosyltransferase 1 multigene family in flax regulating secondary metabolite accumulation

4.1 Introduction **68**

4.2 Materials and Methods **69**

4.2.1 Probing the flax genome for UGT genes **69**

4.2.2 PCR amplification, cloning and sequencing **69**

4.2.3 Sequence alignment and phylogenetic analysis **70**
4.2.4 Intron mapping and organization **70**
4.2.5 Detection of orthologs of flax UGTs in four sequenced dicots **70**
4.2.6 Digital expression analysis **70**
4.2.7 Reverse transcription quantitative real time PCR **71**
4.3 Results **72**
4.3.1 Identification of flax UGT genes **72**
4.3.2 Phylogenetic analysis **72**
4.3.3 Detection of orthologs and duplicated genes **73**
4.3.4 Analysis of intron gain/loss events **73**
4.3.5 Expression analysis of flax UGT genes using EST data **77**
4.3.6 Expression analysis of flax UGT genes using microarray data **77**
4.3.7 Expression profiling using RT-qPCR **78**
4.4 Discussion **80**
4.4.1 Flax UGT family resembles the phylogenetic group structure of Arabidopsis UGTs **82**
4.4.2 Detection of orthologs and functional divergence of unique flax UGTs **83**
4.4.3 Intron mapping to understand the evolution of UGT family **84**
4.4.4 Expressed flax UGTs: identified by digital expression analysis and supported by RT-qPCR **85**

Chapter 5: Summary and future directions **88**
5.1 Developing seed proteome analysis **88**
5.2 Identification and characterization of miRNAs and their target genes **89**
5.3 Phylogenomic analysis of UGT multigene family **90**
5.4 Future directions **90**
Bibliography **92**
Curriculum vitae **117**

Acknowledgements

This thesis was made possible by the immense and much appreciated help and support of a number of people and I take this opportunity to thank them. First of all, I would like to express my sincere gratitude to my supervisor, Dr. Vidya Gupta for her continuous support, encouragement, guidance and advice throughout the course of my doctoral candidature. She herself by her punctuality, enthusiasm, and devotion to work sets an ideal example before me, which has always given me a boost for carrying out the Ph.D. work.

I gratefully acknowledge Dr. Narendra Kadoo for his crucial contribution, which made him a backbone of this thesis. His great efforts in computational analysis, improvement in manuscripts need special thanks. I am especially indebted to Dr. Ashok Giri who has been a adviser throughout my candidature. His enthusiasm has always inspired me and I thank him for helping me in developing a good research plan and its efficient execution. I feel happy to express my thanks to Dr. Mahesh Kulkarni for his help in proteomic analysis and generosity in exchange of consumables required for proteomic analysis. I highly appreciate efforts taken by Dr. Rebecca Zwart for critical reading of draft manuscripts and useful comments.

I would also like to acknowledge Dr. Raju Datla, NRC-PBI Canada for his help in microRNA expression analysis and regular interaction through mail. His contribution in my thesis is highly appreciated. I thank Dr. Peter Mackenzie, Australia for naming the glycosyltransferase genes and his untiring efforts for the same. I sincerely thank Dr. Martin Lott, University of East Anglia, UK for guiding me during microRNA analysis. Dr. Sam Griffiths-Jones, miRBase administrator is highly acknowledged for accepting flax microRNA sequences and naming them with international standards.

This work would never be complete without expressing my gratitude to Dr. Varsha Pardeshi for her contribution in developing my PhD objectives and for her suggestions and constant encouragement throughout the work. I am extremely grateful to Sandip Kale, my best friend for his helping hands in work and support all the way throughout. I would also like to thank Ashwini for her support in initial experiments, her helping nature and setting up ideal example for disciplined and systematic work. I would like to thank my friends Dr. Vishal Dawkar, Dr. Ashok

Chaugule, Rakesh and Sneha for being always with me through good times and rough patches. I sincerely thank Yashwant for his help in data analysis and making me updated about various proteomics data analysis softwares.

I would like to thank Dr. Bhushan Dholakia for always being helpful in giving suggestions for improvement in experimentation. I thank Ajit, Amey, Ashish, Atul, Hemangi, Gouri, Gayatri, Hemlata, Medha, Priyanka, Radhika, Ram, Rasika, Ramya, Reema, Rahul, Neha K., Neha M., Sheon, Sachin, Yojana, Purushottam, Ramesha, Tejas, Smuriti, Amol, Sucheta, Yamini and all other labmates for their timely help. I reserve special thanks to Charu, Sheetal and Yashashree who continue to manage the scientific requirements of the group.

I express my sincere thanks to Dr. Sivaram and Dr. Sourav Pal, the former and present Directors of NCL, for providing the research facilities at this institute. I also acknowledge the Council of Scientific and Industrial Research, Government of India for the research fellowship that enabled me to pursue PhD at NCL.

I find no words for my parents who have been a constant support and have always encouraged me to achieve my dreams. I deeply acknowledge my wife, Anu for her constant support and trust in me that inspired me to work hard. Special thanks to my brother, Ramdas and sisters Pallavi and Akki for providing encouragement and moral support on the bumpy ride of my PhD candidature. I thank Mama, from the bottom of my heart, for being with me from childhood and supporting me financially as well as morally.

Last but not the least, I thank the Almighty God for giving me the strength and courage at every step of life.

Vitthal

List of Abbreviations

2D-GE	Two dimensional gel electrophoresis
AA	Arachidonic acid
ACCase	Acetyl-CoA carboxylase
Acetyl-CoA	Acetyl-coenzyme A
ACN	Acetonitrile
ACP	Acyl carrier protein
ALA	α-linolenic acid
BAC	Bacterial artificial chromosome
BC	Before christ
bp, kb, Mb	Base pair, kilo-base pair, megabase pair
CBB-R250	Coomasssie Brilliant Blue-R250
cDNA	complementary deoxyribonucleic acid
Ct	Thresh hold cycles
CTAB	Cetyltrimethylammonium bromide
DAA	Days after anthesis
DHA	Docosahexaenoic acid
DiGE	Differential gel electrophoresis
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide triphosphate
DTT	Dithiothreitol
ED	Enterodiol
EDTA	Ethylenediamine tetra acetate
EFA	Essential fatty acid
EL	Enterolactone
EPA	Eicosapentaenoic acid
ER	Endoplasmic reticulum
ESI	Electrospray ionization
EST	Expressed sequence tags
FA	Fatty acid
FAD2	Fatty acid desaturase 2
FAD3	Fatty acid desaturase 3
FAME	Fatty acid methyl esters

FAO	Food and Agriculture Organization
FID	Flame ionization detector
g, mg, µg, ng	Gram, milligram, microgram, nanogram
GC	Gas chromatography
GO	Gene ontology
GT	Glycosyltransferases
ha	Hectare
HP	Hairpin
h, min, s	Hour, minute, second
IEF	Iso-electric Focussing
IPG	Immobilized pH gradient
IST	Indian standard time
KAS	3-ketoacyl-ACP synthases
kDa/kD	kilo Dalton
L, mL, µL	Liter, milliliter, microliter
LA	Linoleic acid
LC	Liquid chromatography
LC-PUFA	Long chain polyunsaturated fatty acids
LOX	Lipoxygenase
M, mM, µM, N	Molar, millimolar, micromolar, normal
MFEI	Minimum folding free energy index
Met	Methionine
miRNA	microRNA
mRNA	messenger RNA
MUFA	Monounsaturated fatty acid
NCBI	National Center for Biotechnology Information
NJ	Neighbour joining
nt	Nucleotides
OA	Oleic acid
^{o}C	Degree celsius
PA	Palmitic acid
PAGE	Poly-acrylamide gel electrophoresis
PCR	Polymerase chain reaction

PLACE	Plant cis-acting regulatory DNA elements
PLGS	Protein Lynx Global Server
PSPG	Plant secondary product glycosyltransferase
RNA	Ribonucleic acid
rpm	Revolutions per minute
RT-qPCR	Reverse transcription-quantitative polymerase chain reaction
SA	Stearic acid
SDG	Secoisolariciresinol diglucoside
SDS	Sodium dodecyl sulphate
SEA	Singular enrichment analysis
SECO	Secoisolariciresinol
SSR	Simple sequence repeat
SYBR	SYBR Green I
TF	Transcription factor
TFA	Trifluoroacetic acid
TSS	Transcription start site
TUFGEN	Total Utilization Flax GENomics
UDP	Uridine diphosphate
UGT	Uridine diphosphate glycosyltransferase
UPLC	Ultra performance liquid chromatography
UTR	Untranslated region
WGS	Whole genome sequence
WHO	World Health Organisation

List of Figures

Figure 1.1 Comparison of saturated and unsaturated fatty acids in dietary fats and oils

Figure 1.2 Fatty acid biosynthesis pathway in plants

Figure 1.3 Proposed major biosynthetic pathway to the SDG hydroxymethylglutaryl (HMG) ester linked biopolymer in developing flax seed

Figure 1.4 Flax tissues used for cDNA library construction

Figure 1.5 Tabular representation of omics studies in flax

Figure 2.1 Schematic representation of experimental design for proteome analysis of developing flax seeds

Figure 2.2 Biophysical and biochemical characterization of flax seed at seven seed developmental stages

Figure 2.3 GO analysis of proteins identified in developing flax seeds

Figure 2.4 RT-qPCR expression profiles of 19 candidate genes

Figure 2.5 Schematic view of carbohydrate metabolism during flax seed development

Figure 2.6 Schematic view of the methionine metabolic pathway operating during flax seed development

Figure 2.7 Schematic view of secondary metabolite biosynthesis during flax seed development

Figure 3.1 Phylogenetic distribution of the miRNA genes reported from five plant species

Figure 3.2 Length distribution of TSS-HP and HP-PolyA regions of pri-miRNA

Figure 3.3 Sequence logo of nucleotide frequencies surrounding the transcriptional start sites

Figure 3.4 Expression of selected pre-miRNAs in various flax tissues

Figure 3.5 Gene ontology (GO) analysis of flax miRNA target transcripts

Figure 3.6 The expression of nine miRNA target transcripts in various flax tissues

Figure 4.1 Phylogenetic analysis of the flax UGT family genes

Figure 4.2 Distribution of introns among 82 UGT genes of flax

Figure 4.3 Expression levels for flax UGT genes in various tissues by microarray analysis

Figure 4.4 RT-qPCR expression profile of 10 selected flax UGT genes in 12 different tissue types

List of Tables

Table 1.1 Components of extractable flax seed protein fractions

Table 1.2 Amino acid composition of proteins from flaxseed, soybean and canola

Table 1.3 Composition of relative neutral sugars in flaxseed and commercial gums

Table 3.1 Details of the four gene clusters detected in the microRNA study

Note: Those figures and tables which could not be accommodated in A4 size paper are provided as supplementary material in the compact disk (CD) attached to the inside of the back cover of the thesis.

List of Supplementary Figures

Figure S 2.1 Venn diagram of flax seed proteins annotated by one, a combination of two and all three GO vocabularies. (CC: cellular component, BP: biological process, MF: molecular function)

Figure S 2.2 Digital expression analysis of identified proteins using seed specific EST libraries. Identified 1716 proteins were TBlastN queried against developmental seed specific EST database and EST clone frequencies were represented as heat map. The left and upper side of the heat map shows hierarchical clustering based on Pearson correlation matrix.

Figure S 3.1 Predicted secondary hairpin structures for 116 flax miRNAs identified in this study. Mature miRNA sequences are shaded in green and miRNA* are shaded in magenta.

Figure S 3.2 Hierarchical representation of GO terms overrepresented in biological process (a) cellular component (b) and molecular function (c) categories generated by agriGO analysis. The boxes represent GO terms, GO Id, term definition and statistical information and are color coded as per the level of significance (0.05). Non-significant terms are shown as white boxes. The degree of color saturation of a box is positively correlated to the enrichment level of the term. Solid, dashed, and dotted lines represent two, one and zero enriched terms at both ends connected by the line, respectively. The rank direction of the graph is set to from left to right.

Figure S 4.1 Complete amino acid alignment of 137 Flax, 19 Arabidopsis and 1 Sesame UGTs

Figure S 4.2 Distribution of intron sizes in the flax UGTs

List of Supplementary Tables

Table S 2.1	Sequences of primers used for RT-qPCR in proteomics study
Table S 2.2	Summary of identified proteins. Functional classification, PLGS score and mass spectrometric data were represented. Sub-cellular localization of proteins was predicted by WoLF PSORT web server and values indicate scores. Protein homologue was identified by using BlastP suite from NCBI.
Table S 2.3	Digital expression analysis of identified proteins using seed specific EST libraries A. Details of TBlastN parameters and NCBI accession numbers of mapped ESTs B. Frequency of EST clones mapped onto the proteins
Table S 2.4	Digital expression analysis of identified proteins using complete EST database
Table S 3.1A	Information of 14 pre-miRNA primers for expression analysis
Table S 3.1B	Information of nine target gene primers for expression analysis
Table S 3.2A	Summary of the identified 116 flax miRNAs
Table S 3.2B	Mature miRNA sequences
Table S 3.3	Sequences of predicted transcription start sites (TSS) of 112 flax miRNAs indicating the TATA box in bold
Table S 3.4A	Count of promoters identified in 112 miRNA genes in 2kb upstream sequence of TSS
Table S 3.4B	Count of 17 promoters prevalent in the miRNA genes
Table S 3.5	Digital expression analysis of identified miRNAs using BlastN against flax EST database
Table S 3.6	Flax miRNA targets identified by psRNATarget web server. Target ESTs were further BlastX searched against NCBI protein database to predict their putative functions
Table S 4.1A & B	Information about primers used to clone and sequence full length UGTs and RT-qPCR
Table S 4.2	Accession numbers of proteins sequences encoded by genes included in the phylogenetic analysis

Table S 4.3 137 flax UGT CDS sequences submitted to Genbank with accession numbers (JN088282- JN088418)

Table S 4.4 Summary of 137 flax UGTs: information of genes and intron positions

Table S 4.5 Sequence similarity of the phylogenetic groups and families of 137 flax UGTs

Table S 4.6 Orthologues of flax UGTs identified from four sequenced dicots

Table S 4.7 Information about duplicated genes identified and their differential expression patterns

Table S 4.8 Summary of digital expression analysis with EST and microarray data

Thesis Abstract

Flax or linseed (*Linum usitatissimum* L., 2n = 30) is an annual dicot plant and belongs to the family Linaceae. Its seed has the highest content of the essential omega-3 fatty acid, i.e. alpha-linolenic acid (ALA) and other bioactive compounds such as lignans, predominantly secoisolariciresinol diglucoside (SDG), phenolic acids and flavonoids. ALA serves as a precursor for biologically active longer chain poly-unsaturated fatty acids (PUFA), mainly eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA). The presence of high levels of ALA in flax seed oil has led to its use in both, industrial as well as food and feed products including a wide variety of nutraceuticals and health foods. It is clinically proven that consumption of flax seeds reduces the risks of heart attack, inflammatory disorders and atherosclerosis and inhibits growth of prostate and breast cancers. The flax oil also has multiple industrial applications such as in the manufacture of linoleum and paints and in preserving wood and concrete. Due to these properties, flax is regarded as the cash crop of tomorrow and has enormous potential for significant economic impact in India.

Given the economic and health benefits of the nutraceutically important compounds accumulated in the seed during the seed developmental stages, understanding their biosynthesis and metabolic network at particular stages of seed development has become an important challenge. Identification and characterization of the expressed genes and proteins and their dynamic expression profiles during various seed development stages will help to understand metabolic regulation at the transcriptional and post-transcriptional levels. The present study systematically analyses seed proteome at various seed developmental stages and provides significant information with respect to the regulation and accumulation of economically important metabolites such as oil, lignan and storage proteins. Further, we performed dedicated characterization of the upstream and terminal regulatory genes such as microRNAs and UDP-glycosyltransferase multigene families and analyzed their roles in oil and lignan biosynthesis, respectively.

SDS-PAGE prefractionation coupled with MS^E mass spectroscopy for proteome profiling of developing flax seed was the best for greater proteome coverage

Seven seed developmental stages (4, 8, 12, 16, 22, 30 and 48 days after anthesis) in a flax variety, NL-97 were selected for proteome analysis. To ensure a systematic

coverage of the flax seed proteome, three protein pre-fractionation methods *viz*., 1D-SDS-PAGE, 2D-SDS-PAGE and in-solution digestion were evaluated. The peptides extracted from all the three pre-fractionation methods were analyzed using LC-MS^E. In 2D-SDS PAGE, 16 DAF stage showed the maximum number of protein spots. Hence, this stage was selected for further protein identification using the three pre-fractionation methods. All the protein spots from 2D-SDS-PAGE (16 DAF stage) were picked, extracted and subjected to LC-MS^E analysis for protein identification. This resulted in identification of 203 non-redundant proteins, whereas only 40 non-redundant proteins were identified by the in-solution digestion method. Total proteins (80 µg) were also subjected to 1D-SDS-PAGE gel that was cut into 13 slices and LC-MS^E analysis was performed, which identified 632 non-redundant proteins. Hence, 1D-SDS-PAGE was considered as an ideal pre-fractionation method for flax seed proteomic analysis and was further used for proteomic analysis of the seven seed developmental stages.

Developing seed proteome analysis revealed functional metabolic pathways for oil and secondary metabolite biosynthesis

A total of 1716 proteins were identified and their functional annotation revealed that a majority of them were involved in primary metabolism, protein destination, storage and energy. Further, the identified proteins were mapped on to developmental seed specific EST libraries of flax to obtain transcriptional evidence and 81% of them had detectable expression at mRNA level. Almost all the enzymes involved in developing seed specific pathways like cytosolic and plastidial glycolysis, fatty acid, secondary metabolite and methionine biosynthesis were identified along with their isoforms. The present study proposed three carbon flux routes supplying carbon for fatty acid synthesis from maternal sucrose and all the routes were supported with multiple isoforms of enzymes involved in it and RT-qPCR of selected genes. Lipid profile of developing seed was analysed by using GC-FID and it was observed that, fatty acid accumulation was associated with oleosin and lipoxygenase expression as also confirmed by RT-qPCR. Developing seed also showed the existence of lipid degradation pathways like β-oxidation and glyoxylate pathway, which might be responsible for preferential accumulation of ALA in flax seeds. This observation was also supported by studies conducted in developing rapeseed and Arabidopsis, which proposed that enhanced activity of enzymes involved in these pathways leads to

accumulation of specific fatty acids. Enzymes of metabolic pathway for phenylpropanoid and flavonoids class of compounds were also detected in this study and their importance with respect to accumulation has also been discussed. Cyanogenic glycosides are the main undesired compounds accumulated in flax seed and the enzymes involved in biosynthesis and degradation were expressed throughout seed development but abundance of transcript expression for anabolic enzymes was very high, resulting in accumulation at the end of development process. Metabolism of methionine has great importance in seed development and germination as it is one of the key amino acids in seed storage proteins and one carbon metabolism. Complete biosynthesis pathway for methionine turnover has been identified in this study and RT-qPCR revealed that methionine synthase and S-adenosylmethionine synthetase, key enzymes of S-methylmethionine cycle, were highly expressed enzymes besides seed storage proteins. More in-depth studies of the identified proteins will be useful for better understanding of the complexities of flax seed development.

Large numbers of conserved flax specific microRNAs identified and the role of miRNA targets in oil accumulation has been proposed

The miRNAs are small regulatory RNAs. Mature miRNA (21 nt) sequences are almost conserved across the plant species, however, the precursor-miRNA sequence may vary. The conserved nature of miRNA genes was utilized to identify these genes from flax. In this study, 116 conserved miRNAs belonging to 23 families were identified from the flax genome using a computational approach. These genes were also characterised with respect to their transcription start site (TSS), promoter region and polyadenylation site. The precursor miRNAs varied in length; while most of the mature miRNAs were 21 nt long, intergenic and showed conserved signatures of RNA polymerase II transcripts in their upstream regions. Promoter region analysis of the flax miRNA genes indicated prevalence of MYB transcription factor binding sites. Four miRNA gene clusters containing members of three phylogenetic groups were identified. Further, 142 target genes were predicted for these miRNAs and most of these represent transcriptional regulators. The miRNA encoding genes were expressed in diverse tissues as determined by digital expression analysis as well as real time PCR. Gene ontology analysis suggested that 40% of the putative target transcript were involved in cellular process, metabolic process, response to stimuli and biological regulations.

The expression of fourteen miRNAs and nine target genes was independently validated using the quantitative reverse transcription PCR (RT-qPCR) in eight different tissue types. There was inverse correlation in the expression of miRNA gene and its predicted target gene in a particular tissue indicating the miRNA mediated degradation of target transcripts. MicroRNA profiling in two *Brassica napus* cultivars substantiated the role of miRNA in oil biosynthesis and accumulation. Most of the identified target transcripts coded for transcription factors controlling various metabolic pathways including oil biosynthesis. Overexpression of endogenous miRNA in oil producing green algae *Chlamydomonas reinhardtii* gave enhanced oil yield and further supporting the role of miRNA in oil metabolism. Characterized flax miRNAs can be further utilized for enhancing oil yield by genetic manipulation.

Phylogenomic analysis of UDP glycosyltransferase 1 multigene family identified important regulatory genes for secondary metabolite accumulation

The glycosylation process, catalyzed by ubiquitous glycosyltransferase (GT) family enzymes, is a prevalent modification of plant secondary metabolites that regulates various functions such as hormone homeostasis, detoxification of xenobiotics and biosynthesis and storage of secondary metabolites. Further, these glycosyltransferases regulate the accumulation of specific secondary metabolites such as lignans. The plant UGT enzymes are characterized by a unique, well-conserved sequence of 44 amino acid residues designated as the plant secondary product glycosyltransferases (PSPG) box. This PSPG box sequence was employed to search the whole genome of flax, which identified 137 UGT genes. Phylogenetic analysis of these protein sequences clustered them into 14 major groups (A-N). Expression patterns of these genes were investigated using publicly available expressed sequence tag (EST), microarray data and RT-qPCR. Seventy-three per cent of these genes (100 out of 137) showed expression evidence in 15 tissues examined and indicated varied expression profiles. The RT-qPCR results of 10 selected genes were also coherent with the digital expression analysis. Further, expression analysis suggested that the gene *LuUGT74S1* highly expressed in developing seed tissue and might glycosylate secoisolariciresinol. This gene was heterologously expressed by Canadian group showing glycosylation activity towards secoisolariciresinol (From personal communication with TUFGEN coordinator; unpublished data). Interestingly, five duplicated UGT genes were identified, which showed differential expression in various tissues. Of the seven

intron loss/gain positions detected, two intron positions were conserved among most of the UGTs, although a clear relationship about the evolution of these genes could not be established. Comparison of the flax UGTs with orthologs from four other sequenced dicot genomes indicated that seven UGTs were flax diverged. Flax has a large number of UGT genes including few flax diverged ones. Phylogenetic analysis and expression profiles of these genes identified tissue and condition specific repertoire of UGT genes from this crop. This study facilitated precise selection of candidate genes (for e.g. *LuUGT74S1*) and their further characterization of substrate specificities and *in planta* functions.

The present study comprehensively characterized developing seed proteome and regulatory genes involved in oil and secondary metabolite biosynthesis. Moreover, the study also suggested candidate genes for genetic manipulation to enhance desirable traits in flax.

CHAPTER 1

Introduction and Review of Literature

1.1 Origin and domestication

Flax or linseed (*Linum usitatissimum* L.) is an important plant for its oil and fiber and is, therefore, vital for human culture and development for more than 8000 years. The history of flax can be traced by botanical findings in archaeological excavations, such as seeds, capsule fragments, stems and textiles and by help of ancient writings. The oldest botanical records are from the Near East and date to the 9th millennium BC (Helbæk, 1959; Zeist van and Bakker-Heeres, 1975), although they are frequently found in archaeological excavations all over the Europe.

Morphological (Diederichsen and Hammer, 1995), cytogenetic (Gill and Yermanos, 1967) and molecular (Fu et al., 2002; Allaby et al., 2005; Fu and Allaby, 2010) evidences suggest that the wild progenitor of cultivated flax is pale flax [*L. bienne* Mill. (Hammer, 1986)]. Cultivated flax when crossed with pale flax, produces progeny, which is observed to be fertile in nature (Tammes, 1928). Pale flax has a very wide biogeographical range spanning Western Europe and the Mediterranean, North Africa, Western and Southern Asia, and the Caucasus regions. It was earlier thought that these diversity regions of flax might represent independent domestication events within the wide biogeographical range of pale flax (Harlan, 1971). However, recent molecular and phylogenetic evidences (Allaby et al., 2005; Fu, 2005; Fu and Allaby, 2010; Fu, 2011), suggest that domesticated flax is of monophyletic origin and has probably evolved from a single domestication event of pale flax. However, the geographic location of this event is still obscure. In addition, these data also indicate that the flax was first domesticated for its oil, rather than for fiber use (Allaby et al., 2005; Fu and Allaby, 2010; Fu, 2011). The pale flax has strong branches, small seeds and dehiscent capsules. On the other hand, the main changes that were observed because of domestication were the shift to non-dehiscent capsules, increase in seed size, selection for high oil yielding, or longer stem with high amount of long fiber varieties.

1.2 Flax as a crop plant

The dual purpose of flax is known since ancient times. In ancient Egypt, linen (derived from the fiber) was used for wrapping the royal mummies and additionally linseed oil was used to embalm the bodies of deceased Pharaohs. Although, for a long time, flax has been cultivated as a dual-purpose crop, nowadays fiber flax and oilseed

flax (linseed) represent different gene pools. The application of flax is not restricted to the production of linen yarn. In fact almost the whole plant is used, justifying the species name given by Linnaeus, *L. usitatissimum*, which means useful flax. It is a cool season crop. Temperate and cool climatic conditions are best suited for its growth and the temperature required for its cultivation ranges from minimum 10^{o}C to maximum 38^{o}C. Flax can be raised in almost all types of soils, where sufficient moisture is available, but does better on heavier soils having greater water-retention capacity. It grows the best on well drained loam to clay loam soils rich in humus. It is also tolerant to wide range of soil pH (5.0-7.0). Annual rainfall of 700 to 750 mm is optimum; while under dry climates, irrigation is essential. The crop takes about 120 to 140 days to mature depending on the cultivar; where, the vegetative period is of 45 to 60 days, flowering period is 20 to 30 days and the maturation period consists of 30 to 40 days (Vittal et al., 2005).

Fiber flax varieties usually have longer stem, 80–120 cm tall, with fewer branches, fewer seed capsules, and smaller seeds. Whereas, oil type flax has shorter and heavily branched stems, 60–80 cm tall, with a higher number of seed capsules and larger seeds. The short fibers are used in paper, isolation material, matrix composites and linen painting textile. Total worldwide area under cultivation for flax is about 20 million hectares. Canada, China, the Russian federation and the United States of America are responsible for more than 65% of the worldwide production of flax, while India ranks 5^{th} in flax production. India ranks 3^{rd} for land under cultivation (0.33 million hectares) for flax crop. However, if the production in terms of yield is compared, it is only 4339 hectograms per hectare (Hg/Ha) as compared to Canada 13481 Hg/HA (FAOSTAT, 2010). The loss of yield in India is mostly because of pathogen attack, drought condition, prevalence of low yielding varieties, soil composition, low inputs (fertilizers, irrigation etc) and environmental changes such as temperature. From ancient times, flax was a major crop in India; however, during the course of farming practices, it was taken over by cash crops such as chickpea, soybean, canola etc. To reintroduce flax in farming practices the Government of India has initiated all India coordinated research project for linseed mainly at Kanpur and Nagpur. Their objectives are to develop new flax varieties by plant breeding and using molecular tools. As flax oil contains high proportion of polyunsaturated fatty acids (PUFAs) it becomes rancid quickly making it non-consumable. Recent breeding programmes, therefore developed a new flax type called ‘Solin’ with low (<5%) alfa

linolenic acid (ALA) content in the oil (Dribnenki and Green, 1995). It is agronomically not different from regular flax and developed for the edible oil market. The present study also attempt to generate genomic and proteomic resources to enhance yield and quality of this economically important crop.

1.3 Flax seed as a functional food

1.3.1 Flax oil

Fatty acids (FAs) are carboxylic acids with a long unbranched aliphatic chain, which are either saturated or unsaturated. Flax seed oil, compared to many other sources of dietary fats, has a very healthy FA profile, with low levels (approximately 9%) of saturated fat, moderate levels (18%) of monounsaturated fat and high concentrations (73%) of PUFAs **(Figure 1.1)**. The PUFA content comprises about 16% ω-6 fatty acids, primarily as linoleic acid (LA), and 57% ω-3 fatty acid, ALA (Bhatty and Cherdkiatgumchai, 1990; Daun and DeClercq, 1994).

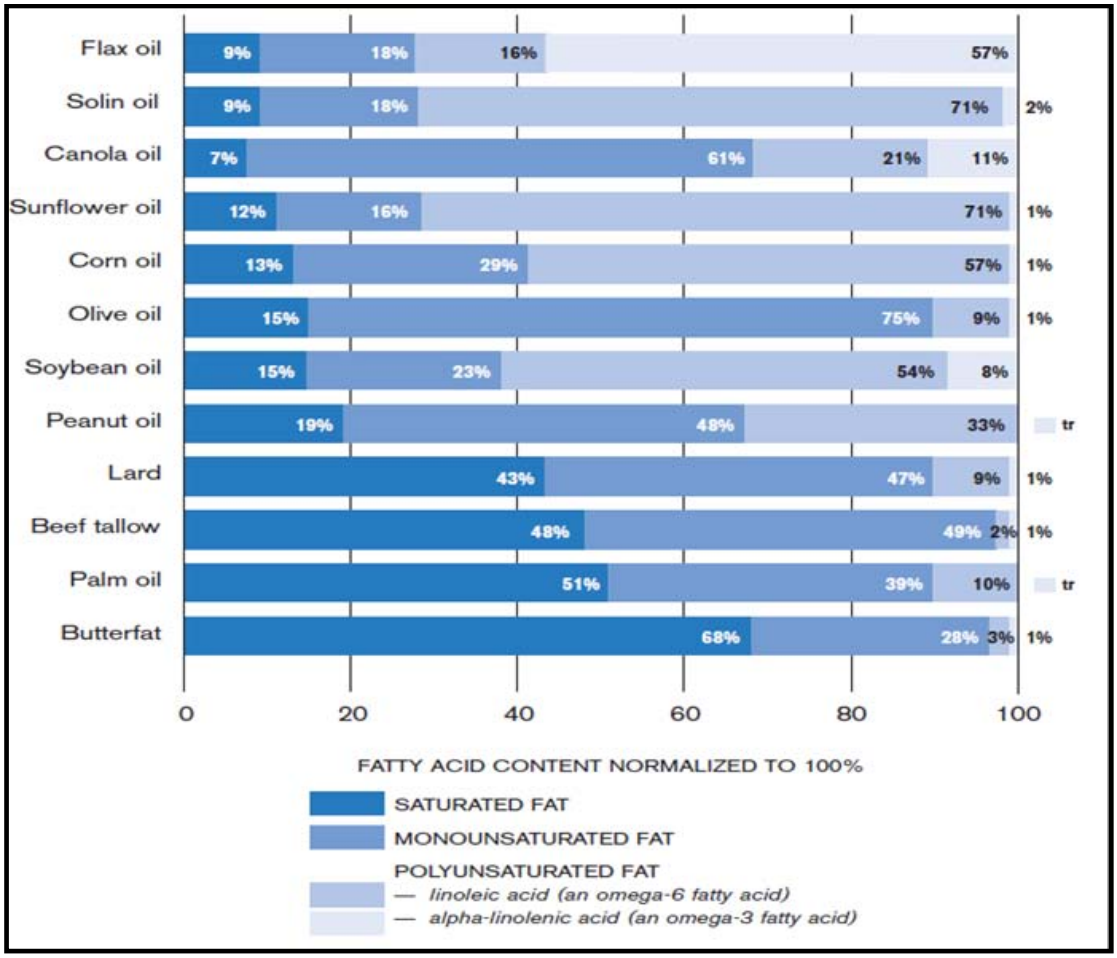


Figure 1.1 Comparison of saturated and unsaturated fatty acids in dietary fats and oils. [Adapted from Morris (2007)]

Both the fatty acids (LA and ALA) are essential fatty acids (EFA), since the human body cannot synthesize them and hence they must be obtained from external food sources or diet. They are further converted into arachidonic acid (AA) and eicosapentaenoic acid (EPA) in the human body and are metabolized to produce hormone like substances known as eicosanoids that affect physiological functions such as cell growth and division, inflammatory responses, muscle activity, blood pressure and immune function. Since LA and ALA compete with one another for the enzymes responsible for their conversion to AA and EPA, respectively, it is important to have a proper balance of ω-6 and ω-3 fatty acids in the diet. Flax has a lower ratio of ω-6/ω-3 and hence helps to improve this FA balance in our body. According to the joint recommendation by Food and Agriculture Organization and World Health Organization (FAO/WHO) committee, the ideal ω-6/ω-3 ratio should be between 5:1 and 10:1. In today's lifestyle, this ratio has considerably increased (20-25:1), indicating that modern diets are severely deficient in ω-3 FAs compared to the diet on which humans evolved and their genetic patterns were established (Eaton and Konner, 1985; Simopoulos, 1995). The higher ω-6/ω-3 ratio is a consequence of an increase in the intake of *trans* FAs, found in products made with hydrogenated vegetable oils, and ω-6 FAs, found in vegetable oils and animal products derived from grain-fed livestock.

High amount of ω-6 FAs in the body leads to the production of pro-inflammatory eicosanoids, which may contribute in the pathogenesis of many diseases, including cardiovascular diseases, cancer, osteoporosis and inflammatory and autoimmune diseases; whereas, increased levels of ω-3 PUFA (a lower ω-6/ω-3 ratio), exert suppressive effects by its anti-inflammatory eicosanoid production (Simopoulos, 2002; Mozaffarian, 2005; Mori, 2006). ALA reduces inflammatory reactions by blocking the formation of compounds that promote inflammation (anti-inflammatory eicosanoid production) and hence has an important role in prevention of such diseases. It also helps in lowering high blood pressure and cholesterol (Pellizzon et al., 2007; Pan et al., 2009) and thus used in treating cardiovascular disease. Further, ALA also has a role in prevention of certain types of cancer (Narisawa et al., 1994; Williams et al., 2007), neurodegenerative diseases (Joshi et al., 2006), and also in lowering blood glucose levels. ALA is also important in the growth and development of infants (especially brain and retina), as being the precursor of decosa hexanoic acid (DHA), which has a functional role in the membrane intensive organs like brain.

The FA biosynthesis pathway is a primary metabolic pathway, as it is found in every cell and is essential for growth. In plants, the major site of the FA synthesis is within plastid, whereas in animals and fungi, it is primarily synthesized in the cytosol. The carbon source for FA synthesis comes from pool of acetyl-coenzyme A (Acetyl-CoA), which is accumulated in the plastid (Post-Beittenmiller et al., 1991; Post-Beittenmiller et al., 1992). For *de novo* FA synthesis, malonyl-CoA is formed from acetyl-CoA by addition of CO_2 using the biotin prosthetic group of the enzyme acetyl-CoA carboxylase (ACCase) in the first committed step and serves as the carbon donor for subsequent elongation reactions of fatty acyl chain (Harwood, 1988). Further, FA synthesis in the plastids occurs through a repeated series of condensation, reduction and dehydration reactions that add two carbon units derived from malonyl-ACP to the elongating FA chain. Overall, nearly 30 enzymatic reactions are required to produce a 16- or 18-carbon FA from acetyl-CoA and malonyl-CoA (Ohlrogge and Browse, 1995). Before entering the FA synthesis pathway, the malonyl group of malonyl-CoA is transferred from CoA to a protein cofactor, acyl carrier protein (ACP) with the help of enzyme malonyl-CoA: ACP transacylase (Shimakata and Stumpf, 1983). All the further reactions of the pathway involve ACP until the 16- or 18-carbon product is ready for transfer to glycerolipids or for export from the plastid. Malonyl-ACP undergoes a number of condensation reactions mediated by a set of at least three separate condensing enzymes, known as 3-ketoacyl-ACP synthases (KAS) to produce an 18-carbon fatty acid. The first condensation of acetyl-CoA and malonyl-ACP to form a four-carbon product is catalyzed by KAS III. A second condensing enzyme, KAS I, is responsible for producing chain lengths from 6 to 16 carbons. Finally, KAS II is the enzyme required for the elongation of the 16 carbon palmitoyl-ACP to 18 carbon stearoyl-ACP (Shimakata and Stumpf, 1983; Jaworski et al., 1989). In addition to these three enzyme catalyzed reactions, after each condensation step, the 3-ketoacyl-ACP product is reduced, dehydrated, and reduced again, by 3-ketoacyl-ACP reductase, 3-hydroxyacyl-ACP dehydrase, and enoyl-ACP reductase, respectively. As a result, saturated FA that is two carbons longer than at the start of the cycle is obtained **(Figure 1.2)**.

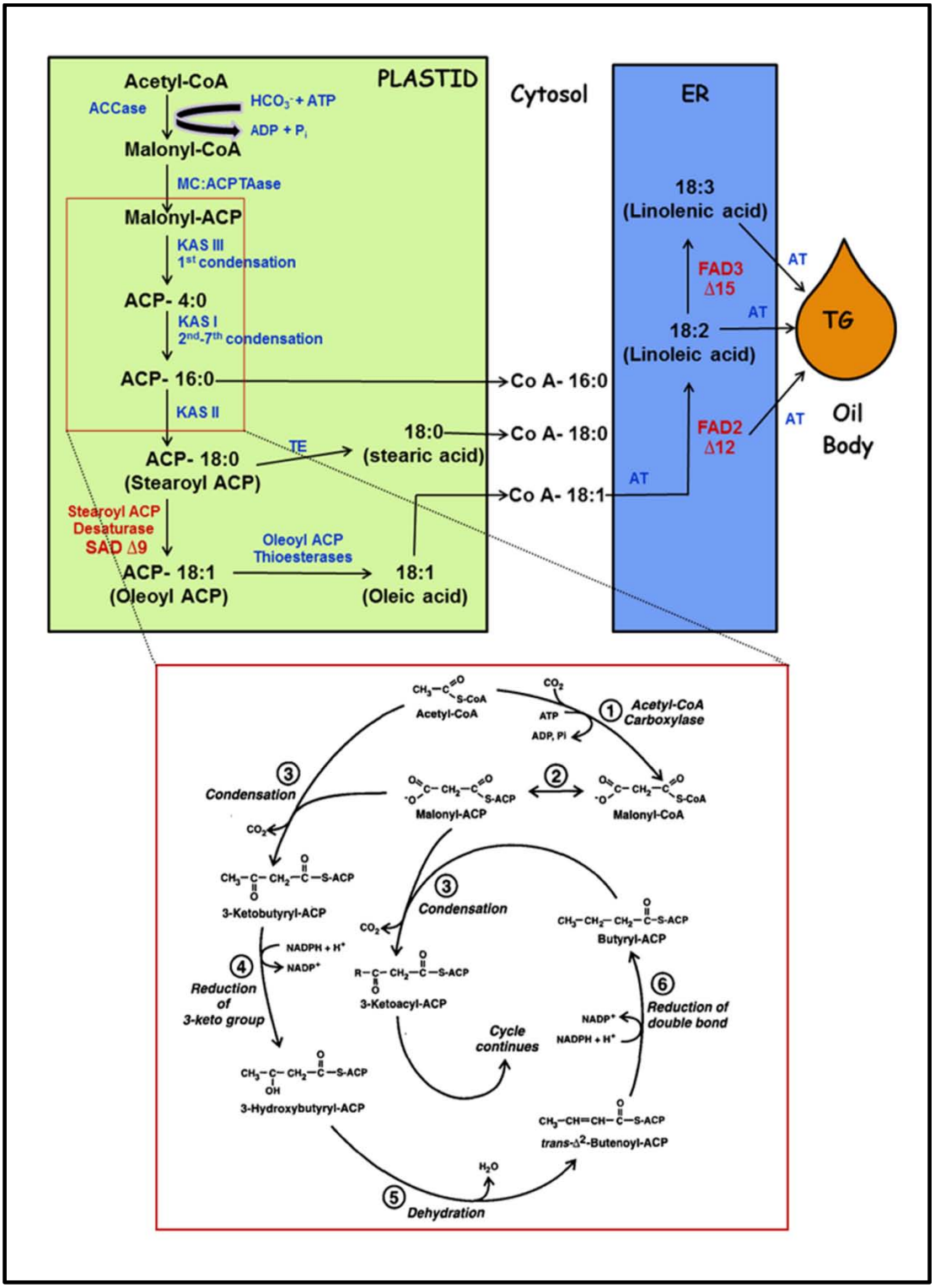


Figure 1.2 Fatty acid biosynthesis pathway in plants. After each condensation with 3-ketoacyl-ACP synthase (KAS), the 3-ketoacyl-ACP product is reduced (reaction 4), dehydrated (reaction 5), and reduced again (reaction 6), by 3-ketoacyl-ACP reductase, 3-hydroxyacyl-ACP dehydrase, and enoyl-ACP reductase, respectively [Adapted from Rajwade (2012)]. (AT- acyletransferases, TE- acyl-ACP thioesterase and TG- Triglycerides)

The elongation of FAs in the plastids is terminated when the acyl group is removed from ACP by either of the two enzyme systems, an acyl-ACP thioesterase or acyltransferases in the plastid. Acyl-ACP thioesterase hydrolyzes the acyl-ACP and releases free FA, whereas acyltransferases transfer the FA from ACP to glycerol-3-phosphate or to monoacylglycerol-3-phosphate. The transport of FA outside plastid is determined on the basis of their release from ACP by either a thioesterase or an acyltransferase. Only when thioesterase acts on acyl-ACP, free FAs can leave the plastid. On the outer membrane of the chloroplast envelope, an acyl-CoA synthetase assembles an acyl-CoA thioester that is then available for acyltransferase reactions to form glycerolipids in the endoplasmic reticulum (ER) (Harwood, 1988; Ohlrogge and Browse, 1995).

The above mentioned FAs are saturated in nature, but it has been observed that over 75% of the FAs found in most plant tissues, are unsaturated (Ohlrogge and Browse, 1995). Hence, these saturated FAs are further converted to unsaturated FAs with the help of desaturase enzymes by addition of double bonds at specific locations in their fatty acyl chains. The first double bond is introduced in plastid by the soluble enzyme stearoyl-ACP desaturase (SAD, Δ9) converting saturated stearic to monounsaturated oleic. Subsequent desaturation of the FAs to the highly unsaturated forms is carried out by membrane bound desaturases of chloroplast and endoplasmic reticulum (ER), where fatty acid desaturase 2 (FAD2 or Δ12*)* is the next in the series of desaturases, which converts oleic acid to linoleic acid by introduction of second double bond. Similarly, the third double bond is added to linoleic acid by FAD3 or Δ15 desaturase to form α-linolenic acid (Somerville and Browse, 1991; Heinz, 1993).

1.3.2 Phenolic acids and lignan

Flax seed is a natural source of major plant food phytochemicals such as flavonoids, coumarins, lignans and phenolic acids (Thompson et al., 1991). Dabrowski and Sosulski (1984) reported that the major phenolic acids in flax seed are; trans-ferulic (46%), trans-sinpaic (36%), *p*-coumaric (7.5%), and trans-caffeic (6.5%). These phenolic compounds function as antioxidants (Amarowicz et al., 1994).

Lignans are found in a variety of plant materials including flax seed, pumpkin seed, sesame seed, soybean, broccoli and some berries. Flax seed lignans are a good source of useful biologically active compounds and therefore, are considered as a functional food. The major lignan in flax seed is called secoisolariciresinol

diglucoside (SDG). Once ingested, SDG is converted in the colon into active mammalian lignans, enterodiol and entero-lactone, which have been shown to reduce growth of cancerous tumors, especially hormone-sensitive ones such as those of the breast, endometrium and prostate. The health benefits of flax seed lignans are thought to be due to antioxidant activity, primarily as hydroxyl radical scavengers and ability to complex divalent transition metal cations (Kitts et al., 1999; Toure and Xu, 2010). It also shows structural similarity to 17-b-estradiol for estrogenic and antiestrogenic compounds (Adlercreutz et al., 1992). The behavior of the lignans depends on the biological levels of estradiol. At normal estradiol levels, the lignans act as estrogen antagonists, but in postmenopausal women (at low estradiol levels) they can act as weak estrogens therefore, inhibit hormone dependent cancers (Hutchins and Slavin, 2003). Biosynthetic pathway of flax seed lignan was elucidated by stable radio isotope labeling studies (Ford et al., 2001) as shown in **Figure 1.3.** However, the terminal enzyme involved in the synthesis of SDG *viz.,* UDP-glycosyltransferase still remains unknown.

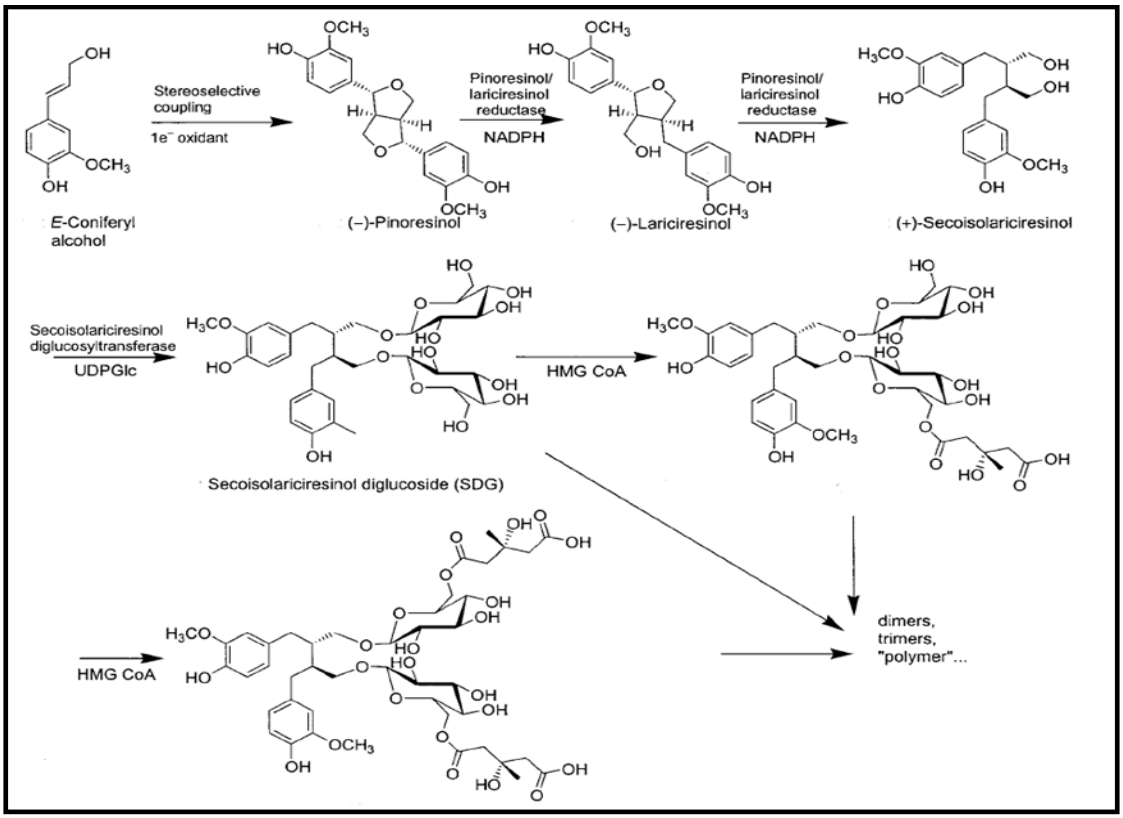


Figure 1.3 Proposed major biosynthetic pathway to the SDG-hydroxymethyl glutaryl (HMG) ester-linked biopolymer in developing flax seed. [Adapted from Ford et al., (2001)]

1.3.3 Flax seed proteins

The total proteins in flaxseed represent about 20-30% of the seed meal making it a good protein source (Sammour, 1999). Flaxseed proteins have similar nitrogen extractability at varying pH and ionic strength as observed in other oilseed proteins (Oomah and Mazza, 1993). Earlier research on flaxseed storage proteins indicated that it contains mixed or heterogeneous proteins (Sammour et al., 1994). Flaxseed proteins have been classified based on their solubility in a series of aqueous and non-aqueous solvents into different protein classes as per Osborne classification of proteins (Osborne, 1924). These classes include globulins or linins, albumins or conlinins, glutelins and prolamins **(Table 1.1)** (Sammour et al., 1994).

Flaxseed proteins comprise 20% albumins of low molecular weight proteins (1.6S and 2S) and 80% globulins of high molecular weight proteins (11S and 12S). They are structurally more lipophilic than soybean proteins due to the influence of their polysaccharide composition (Oomah and Mazza, 1993; Sammour, 1999). The amino acid composition is similar to soybean proteins (Oomah, 2001). They contain relatively higher levels of aspartic acid, glutamic acid and arginine indicating the high content of amides. Flaxseed proteins have less lipidimic and atherogenic effects when consumed compared to soybean proteins due to their low lysine/arginine ratio compared with soybean **(Table 1.2)** (Oomah, 2001).

When molecular and structural properties of flaxseed proteins were compared with those of other oilseed proteins, such as soybean, they showed similarity in terms of hydrophobicity, secondary structure and surface properties (Prakash and Rao, 1986). Addition of high polysaccharide containing flaxseed protein, enhanced the emulsion stability of the canned fish sauce (Dev and Quensel, 1989). Moreover, it is also found to reduce the fat losses during cooking when added to meat emulsions and improved the cooking emulsion and meaty flavor (Dev and Quensel, 1989). Addition of flaxseed protein with high content of mucilage to dough improved dough hardness and bread shelf life (Dev and Quensel, 1989). Recently, flaxseeds are being used for decorative as well as health benefits in bakery products (Oomah, 2001).

Table 1.1 Components of extractable flax seed protein fractions

Protein component	Total proteins %	Sedimentation coefficient (S20, w)	MW (proteins or subunits) (kDa)
Globulin (Linin)	40-80	11-12	14400 24.6 (basic) 30.0 (acidic) 35.2 (acidic) 50.9
Albumin (Conlinin)	20-40	1.6-2.0	16-18
Glutelin	13.5	-	-
Prolamin	6.5	-	-
Oleosin	7.2	-	16-24
Cd-binding protein	7.0	-	1.5
Linusitin	<1.0	-	25

[Adapted from Ayad (2010)]

1.3.4 Dietary fibers

Flaxseed is different from other oilseeds by having a relatively high content of a mucilaginous material composed of acidic and neutral polysaccharides **(Table 1.3)** (BeMiller, 1973; Mazza and Biliaderis, 1989). Fiber occurs as structural material in the cell walls of plants. Dietary fiber consists of non-digestible plant carbohydrates and other materials. They are not digested and absorbed by the human small intestine and therefore, passes relatively intact into the large intestine. Total fiber accounts for about 28% of the weight of flax seeds. The major fiber fractions in flax consist of cellulose, which is the main structural material of plant cell walls; mucilage gums, a type of polysaccharide that becomes viscous when mixed with water or other fluids; and lignin, a highly-branched fiber found within the cell walls of woody plants. They contribute to the strength and rigidity of the cell walls.

Table 1.2 Amino acid composition of proteins from flaxseed, soybean and canola

Amino acids	Flax seed			Soybean glycinin (g/100g)	Canola Globulin (g/100g)
	Total proteins (g/100g)	Globulins (g/16gN)	Albumins (g/16gN)		
Aspartic acid	8.3	11.3	5.5	12.7	8.5
Glutamic acid	22.8	19.8	35	15.5	19
Serine	4.1	5.1	3.9	5.3	5
Glycine	4.9	4.8	8.3	7.7	6
Histidine	2.7	2.5	1.6	1.8	3.3
Arginine	10.4	11.5	13.1	5.5	8.3
Threonine	3.4	3.9	2.1	3.7	3.4
Alanine	4.3	7.9	1.9	5.6	4.2
Proline	3.6	4.5	3	6.2	8.3
Tyrosine	2.2	2.3	1.4	2.8	2.9
Valine	5.7	5.6	2.6	5.7	3.9
Methionine	1.5	1.7	0.8	1.6	1.8
Cysteine	3.3	1.4	3.5	0.7	1.8
Isoleucine	4.8	4.6	2.8	4.6	3.4
Leucine	6.7	5.8	5.4	7	8.2
Phenylalanine	5.1	5.9	2.4	4.3	5.2
Lysine	4.4	3.1	4.9	4.2	4.7

[Adapted from Ayad (2010)]

Ground flaxseed consists of 30% of dietary fiber, 2/3 of which is insoluble (cellulose, hemicellulose and lignin) and 1/3 is soluble fiber. Insoluble fiber binds water and thus, increases the bulk in colon. Soluble fiber has similar effects as of guar gum or ispaghula, e.g. delay in gastric emptying, improvement in glycemic control and alleviation of constipation. Roughly ground flaxseed seems to have better water-binding capacity than the finely ground flaxseed meal. The mean dietary fiber intake in western countries is approximately 20 g/day. The recommended amount of daily dietary fiber of >25g is useful in the treatment of constipation, irritable bowel

syndrome and diverticular disease. Flaxseed supplement of 10-20 g/day would increase the intake to the recommended level of 25-30 g/day. However, the use of flaxseed has to be long term as the full effects are only observed after two months.

Table 1.3 Composition of relative neutral sugars in flaxseed and commercial gums (Cui and Mazza, 1996)

	Flaxseed gums			**Commercial gums**		
	Flax variety					
	NorMan	**Omega**	**Foster (%)**	**Arabic**	**Guar**	**Xanthan**
Rhamonse	21.2	27.2	25.6	34.0	0.0	0.0
Fucose	5.0	7.1	5.8	0.0	0.0	0.0
Arabinose	13.5	9.2	11.0	24.0	24.0	0.0
Xylose	37.4	28.2	21.1	0.0	0.0	0.0
Galactose	20.0	24.4	28.4	45.0	33.0	0.0
Glucose	2.1	3.6	8.2	0.0	0.0	50.7
Mannose	0.0	0.0	0.0	0.0	67.0	49.3

1.4 Economic importance of flax

Flax is one of the most important oilseed crop for industrial as well as food, feed and fiber purposes. Almost every part of the plant is utilized commercially, either directly or after processing. The stem yields good quality fiber having high strength and durability. The seed provides oil rich in omega-3 FAs, digestible proteins and lignans. Seed is primarily used for decoration and texture in baked products. It has also been incorporated into baked foods (Pohjanheimo et al., 2006), dairy products (Kangas et al., 2006) and dry pasta products (Lee et al., 2004). Omega-3 enriched eggs are produced by hens when fed on a diet containing 10% - 20% milled flax. As flax has higher proportion of ALA, and lower proportion of saturated fatty acids than other grains, the eggs produced are higher in total Omega-3 FAs. Omega-3 eggs look, cook and taste just like regular eggs, and the total fat content and cholesterol levels remain similar (Sim and Cherian, 1994). Similarly ‘Omega 3’ chicken meat and pork have also been produced to integrate ALA into the food chain (Romans et al., 1995; Kratzer and Vohra, 1996). A new type of dressing developed from flavonoid enriched genetically modified flax seeds called FlaxAid has shown to decrease the size of

diabetic foot ulcer (Skorkowska-Telichowska et al., 2012). Further, overexpression of the glycosyltransferase gene derived from *Solanum sogarandinum* resulted in significant increase in the content of ferulic acid, p-coumaric acid, caffeic acid, secoisolariciresinol and their glucoside derivatives. Such genetically engineered flax seed has shown to promote the proliferation of normal human dermal fibroblasts and the migration of fibroblasts in the wound scratch assay (Czemplik et al., 2012). These engineered flax-derived products are good candidates for application in the repair and regeneration of human skin and might also be an alternative to antibiotic therapy for infected wounds.

Besides health benefits, flax seed oil also has industrial applications. From the ancient times, flax seed oil has been used in paints, coatings and other industrial products. Flaxseed oil is a drying oil, which undergoes oxidation and forms a natural, plastic-like film. Due to these properties, paints and coatings containing flaxseed oil are still of the highest quality and the most durable products. Nowadays, the use of other modified vegetable oils and petroleum products in place of flax seed oil led to the reduced flax crop area.

Another bioactive compound of flax is lignan mainly secoisolariciresinol diglucoside (SDG). Epidemiological and experimental studies strongly suggest that lignans have a potential role in the prevention of menopausal symptoms, hormone-dependent cancers (e.g. breast and prostate cancer), cardiovascular disease, and possibly osteoporosis. The beneficial effects of flax seed are mediated mainly by its mammalian lignans, the precursor SDG, which upon the action of colonic microflora is converted to mammalian lignans, enterolactone (EL) and enterodiol (ED), which are subsequently absorbed and undergo enterohepatic circulation (Borriello et al., 1985). Mammalian lignans are positively linked to several bioactivities including antiestrogenic, anticarcinogenic and antioxidant activities. Antioxidant activities of SDG and its mammalian lignan metabolites are stronger than vitamin E (Prasad, 2000). Therefore SDG content can be one of the most important indicators of potential health benefit of flax in addition to ω-3 fatty acids.

1.5 Omics studies in flax

Omics studies involve systematic and multi-parallel approaches to investigate various facets of biological events. Due to the recent advances in sequencing techniques such

as next generation sequencing, it is now possible to sequence whole genome and transcriptome within few days to months. New generation mass spectrometers like triple TOF (time of flight), orbitrap, high resolution mass spectrometers analyse particular metabolite or peptide at a resolution of up to 50000 and provides accurate data to identify and characterise metabolome and proteome. These studies provide comprehensive idea about the protein and metabolite abundances based on relative and absolute quantification. Using these approaches one can study important processes such as seed development, oil and secondary metabolite biosynthesis and their regulation.

1.5.1 Genomics

Biochemistry and genetics generally focus on one gene and its encoded protein at a time. While powerful, these traditional approaches do not give a comprehensive view of the structure and activity of an organism's genome, its entire set of genes. The genomics tools encompass the molecular characterization of whole genomes and the determination of global patterns of gene expression. Recently, more than 50 species of plants genomes have been completely sequenced and this would permit comparisons of entire genomes from different species (Goodstein et al., 2012). Realizing the economic importance of flax, Genome Canada initiated project entitled Total Utilization of Flax GENomics (TUFGEN), which aims at providing genetic knowledge for the improvement of seed and fiber traits of oilseed flax. As a part of the TUFGEN project, bacterial artificial chromosome (BAC) library construction for physical mapping and whole-genome shotgun sequencing (WGS) of nuclear genome of the CDC Bethune variety of flax was conducted. The genome-wide physical map of flax constructed with BAC clones provided a framework for accessing target loci for marker development and positional cloning. The physical map consists of 416 contigs spanning ~368 Mb, assembled from 32,025 fingerprints, representing roughly 54.5% to 99.4% of the estimated haploid genome (370-675 Mb). The physical map and paired-end reads from BAC clones will also serve as scaffolds to build and validate the whole genome shotgun assembly (Ragupathy et al., 2011). For genome sequencing, seven paired-end libraries ranging from 300 bp to 10 kb in size, were sequenced using an Illumina genome analyzer. A *de novo* assembly, comprised exclusively of deep-coverage (approximately 94× raw, approximately 69× filtered) short-sequence reads (44–100 bp), produced a set of scaffolds with *N*50 = 694 kb,

including contigs with *N*50 = 20.1 kb. The contig assembly contained 302 Mb of non-redundant sequence representing 81% genome coverage. Up to 96% of published flax ESTs were aligned to the whole-genome shotgun scaffolds. A total of 43,384 protein-coding genes were predicted in the whole-genome shotgun assembly, and up to 93% of published flax ESTs, and 86% of *A. thaliana* genes aligned to these predicted genes, indicating excellent coverage and accuracy at the gene level (Wang et al., 2012).

After flax whole genome sequencing (WGS), large number of multiple genes have been identified and characterized by bioinformatics analysis such as transposable elements (Gonzalez and Deyholos, 2012), oleosin and coleosin genes (Hyun et al., 2013), NBS-LRR genes (Kale et al., 2013), cytochrome P450 genes (Babu et al., 2013) etc. The present study also identified microRNA genes and UDP-glycosyltransferase 1 family genes from whole genome and this will be described in detail in chapters 3 and 4, respectively. Similar kind of analysis for microRNA identification from EST and WGS has been performed earlier but they could identify very small number of microRNA genes from flax (Moss and Cullis, 2012; Neutelings et al., 2012).

Recently, a large number of SSR markers were developed from publically available flax EST libraries (Cloutier et al., 2009; Soto-Cerda et al., 2011) and from SSR-enriched genomic libraries or other genomic sequences (Roose-Amsaleg et al., 2006; Bickel et al., 2011; Deng et al., 2011; Rachinskaia et al., 2011; Kale et al., 2012). Ragupathy et al. (2011) identified 4,064 putative SSRs from bacterial artificial chromosome (BAC) end sequences (BES). There are currently 1,326 SSR markers published in flax (Cloutier et al., 2012). These markers were further used for constructing genetic maps and genetic diversity assessment (Fu and Peterson, 2010; Cloutier et al., 2011; Soto-Cerda et al., 2011). The first integrated consensus genetic and physical map of flax has been published by Cloutier et al. (2012), where they have constructed three linkage maps using three different populations containing about 385-469 mapped markers each. The consensus map of flax incorporated 770 markers based on 371 shared markers including 114 that were shared by all the three populations and 257 shared between any two populations. The map with 15 linkage groups corresponds to the haploid number of chromosomes of this species. The total length of the consensus genetic map is 1,551 cM with a mean marker density of 2.0 cM. A total of 670 markers were anchored to 204 of the 416 fingerprinted contigs of

the physical map corresponding to ~274 Mb or 74% of the estimated flax genome size of 370 Mb. Further, genome-wide single nucleotides polymorphism (SNP) discovery using next generation sequencing of reduced representation libraries in eight flax genotypes discovered 55,465 SNPs and quarter of these SNPs were found in genic regions (Kumar et al., 2012). These genomics studies will assist in generating high density maps of flax and facilitate QTL discovery, marker-assisted selection, phylogenetic analyses, association mapping and anchoring of the whole genome shotgun sequence.

1.5.2 Transcriptomics

Transcriptomics is the branch of molecular biology that deals with the study of messenger RNA molecules produced in an individual or population of a particular cell type. These studies involve characterization of the functional RNA transcript population of a cell/s or organism at a specific time point. Recently, extensive EST sequencing was carried out by Venglat et al. (2011) from various tissues as shown in **Figure 1.4**. The EST sequences were generated by constructing cDNA libraries and sequencing by modified Sanger method (Smith et al., 1986). A total of 261,272 ESTs were generated and assembly of these ESTs resulted in 30,640 unigenes and 82% of ESTs were annotated on the basis of homology to known and hypothetical genes from other plants. From the same group, there had been efforts to sequence the RNAs from various developing seed stages and vegetative tissue types using next generation sequencing technology. These studies will help to identify low expressed genes, alternately spliced transcripts, assist gene annotation, develop digital gene expression map and also discover new genes (Unpublished data and personal communication with Dr. Raju Datla NRC-PBI, Canada).

In addition, there have been efforts to generate high density oligo microarray platform to analyse the synchronized and large scale gene expression in various tissue types of flax. Fenart et al. (2010) developed and validated such oligo microarray platform for flax. They obtained RNA samples from nine different flax tissues such as inner- and outer-stems, developing seeds, leaves and roots. Further, RNA was used to generate a collection of 1,066,481 ESTs by massive parallel pyrosequencing. Sequences were assembled into 59,626 unigenes of which, only 48,021 sequences were selected for oligo design and high-density microarray (Nimblegen 385K) fabrication with eight, non-overlapping 25-mers oligos per unigene. Moreover,

validation experiments illustrated the capacity of this array to detect differential gene expression in a variety of flax tissues as well as between two contrasted flax varieties. The independent study of development and validation of microarray platform carried out on fiber flax varieties indicated that genes involved in the elongation phase of phloem fiber differentiation included transcripts related to cell-wall modification or primary-wall deposition (Roach and Deyholos, 2007, 2008).

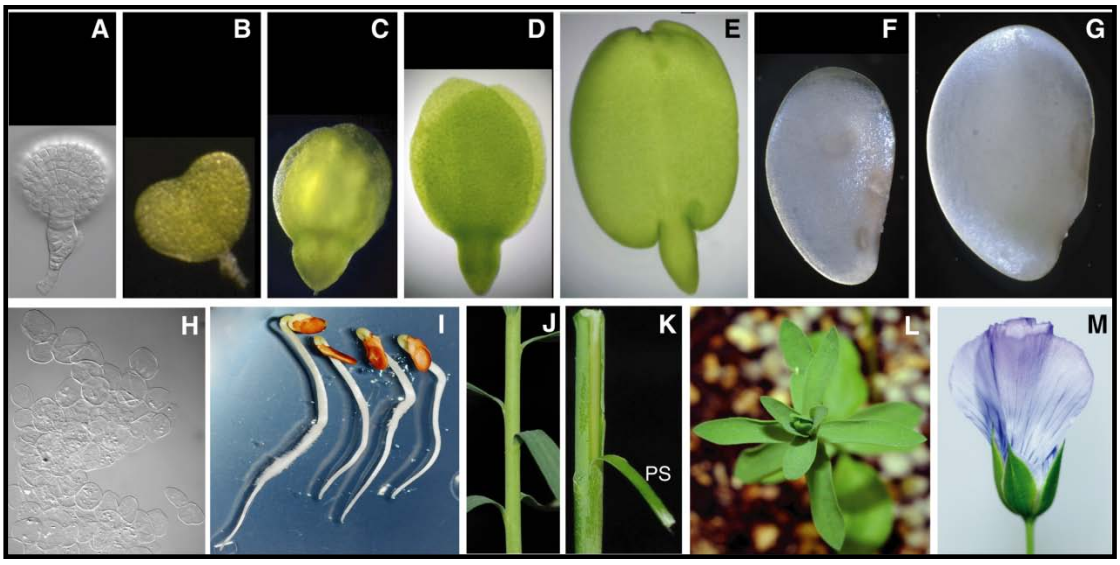


Figure 1.4 Flax tissues used for cDNA library construction (A) globular embryo; (B) heart embryo; (C) torpedo embryo; (D) cotyledon embryo; (E) mature embryo; (F) globular stage seed coat; (G) torpedo stage seed coat; (H) pooled endosperm from globular to torpedo stage seed; (I) etiolated seedlings; (J) stem; (K) stem peel "PS"; (L) leaves; and (M) mature flower. [Adapted from Venglat et al. (2011)]

1.5.3 Proteomics

Proteomics is the large-scale study of proteins, particularly their structures and functions. The proteome is an entire complement of proteins, including the modifications made to a particular set of proteins, produced by an organism or system. It varies with time and distinct requirements of cell or organism. Proteomic studies in flax were carried out mainly on fiber and seed. Early attempts to identify proteins involved in flax fiber development was performed by differential analysis of proteins enriched in bast fibers from individually dissected phloem fibers and non-fiber cells of the cortex using fluorescent (DiGE) labels and 2D-gel electrophoresis. The study showed that a K^+ channel subunit, annexins, porins, secretory pathway components, β-amylase, β-galactosidase and pectin and galactan biosynthetic

enzymes were among the most highly enriched proteins detected in developing flax fibers (Hotte and Deyholos, 2008). Recent study conducted to identify cell wall proteins in flax stem used sequential salt ($CaCl_2$, LiCl) extractions to obtain fractions enriched in cell wall proteins from the flax stem. High-resolution mass spectrometry analysis and use of predicted gene models from WGS as a database allowed the identification of 152 putative flax cell wall proteins. Further, transcript evidence from similar types of tissues indicated their role in flax fiber development (Day et al., 2013).

Flax has significant phytoremediation potential and represents effective low-cost approach for removing pollutants from contaminated soils. Proteomics studies were conducted to identify and characterize the proteins involved in detoxification mechanism of heavy metal such as cadmium (Cd) using two contrasting cultivars of flax (Hradilova et al., 2010). They observed significant changes in the expression of 14 proteins related to disease/defense, metabolism, protein destination and storage, signal transduction energy and cell structure from two dimensional electrophoresis and image analysis. The role of flax for adaptation in soil contaminated with radioisotopes has been very well established by proteomic analysis of mature as well as developing seeds. The first attempt of such study from mature flax seed identified only 28 proteins using 2D and MS^E approach (Klubicova et al., 2010). Two independent studies for systematic proteomic characterization of the mature and developing flax seed from the remediated Chernobyl area were carried out, resulting in quantitative information for 82 and 102 proteins, respectively. Further, these proteins were identified and categorized into various metabolic functional classes. However, due to the lack of genomic resources for flax at that time, proteins of unknown function comprised the largest group (Klubicova et al., 2011a; Klubicova et al., 2011b). In the present study, the proteome for various seed developmental stages has been analysed and explained in detail in chapter 2.

1.5.4 Metabolomics

The metabolome represents the collection of all metabolites in a living system. Metabolomics involves systematic study of the unique chemical fingerprints of small metabolite those are end products of specific cellular processes. Flax seed is a rich source of phenolic acids and lignans (Thompson, 1995). Most of these secondary metabolites are glycosylated and polymerised to form a compact structure. Organic

extraction of these polymeric fraction of flax seed and its base hydrolysis when separated by HPLC showed that they contain the same UV-absorbing components, mostly a lignan, secoisolariciresinol diglucoside (SDG). Further, HPLC purified hydroxycinnamic acid derivative structures were confirmed by nuclear magnetic resonance spectroscopy (NMR). The results of this study suggest that the glucosylated phenolic compounds of flax seed exist in polymeric structure(s) containing ester linkages (Johnsson et al., 2002).

Another study on alkaline hydrolysis of lignan macromolecule separated by reverse-phase chromatography and confirmation using LC-ESI-MS suggested that one molecule of caffeic acid corresponded with five molecules of p-coumaric acid and two molecules of ferulic acid to constitute a lignan macromolecule (Kosinska et al., 2011). There has been an attempt to investigate the dynamics of developing flax seed embryo metabolism using 13C-labelling experiments where the real-time kinetics of label incorporation into metabolites was monitored *in situ* using *in vivo* NMR. This approach was suitable to get a direct assessment of metabolic time-scales within living plant tissues and provides a valuable complement to steady state flux determinations (Troufflard et al., 2007). Moreover, concentration kinetics of important lignan SDG and its precursor coniferin was studied in developing flax seed using reversed-phase high-performance liquid chromatography–diode array detection (HPLC-DAD). It was found that, the biosynthesis of coniferin is not synchronous with that of SDG. Coniferin is synthesized and pooled in flax seed at a high concentration immediately at 0 days after flowering (DAF), then the concentration decreases dramatically and coniferin becomes undetectable 20 DAF, when embryos in flaxseed stop growing (Fang et al., 2013). Compositional analysis of metabolites from ten different flax lines using fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS) resulted in identification of 2606 spectral peaks, each representing a unique m/z (mass-to-charge) ratio. Further, principal component analysis (PCA) and hierarchical clustering (HCA) were able to separate the ten different flax lines/varieties based on the 2606 independent m/z detected (Daskalchuk et al., 2006). Metabolomic studies are very useful to determine the spatial and temporal biosynthesis and accumulation of specific metabolites and also help in determining the metabolic flux of particular metabolic pathway.

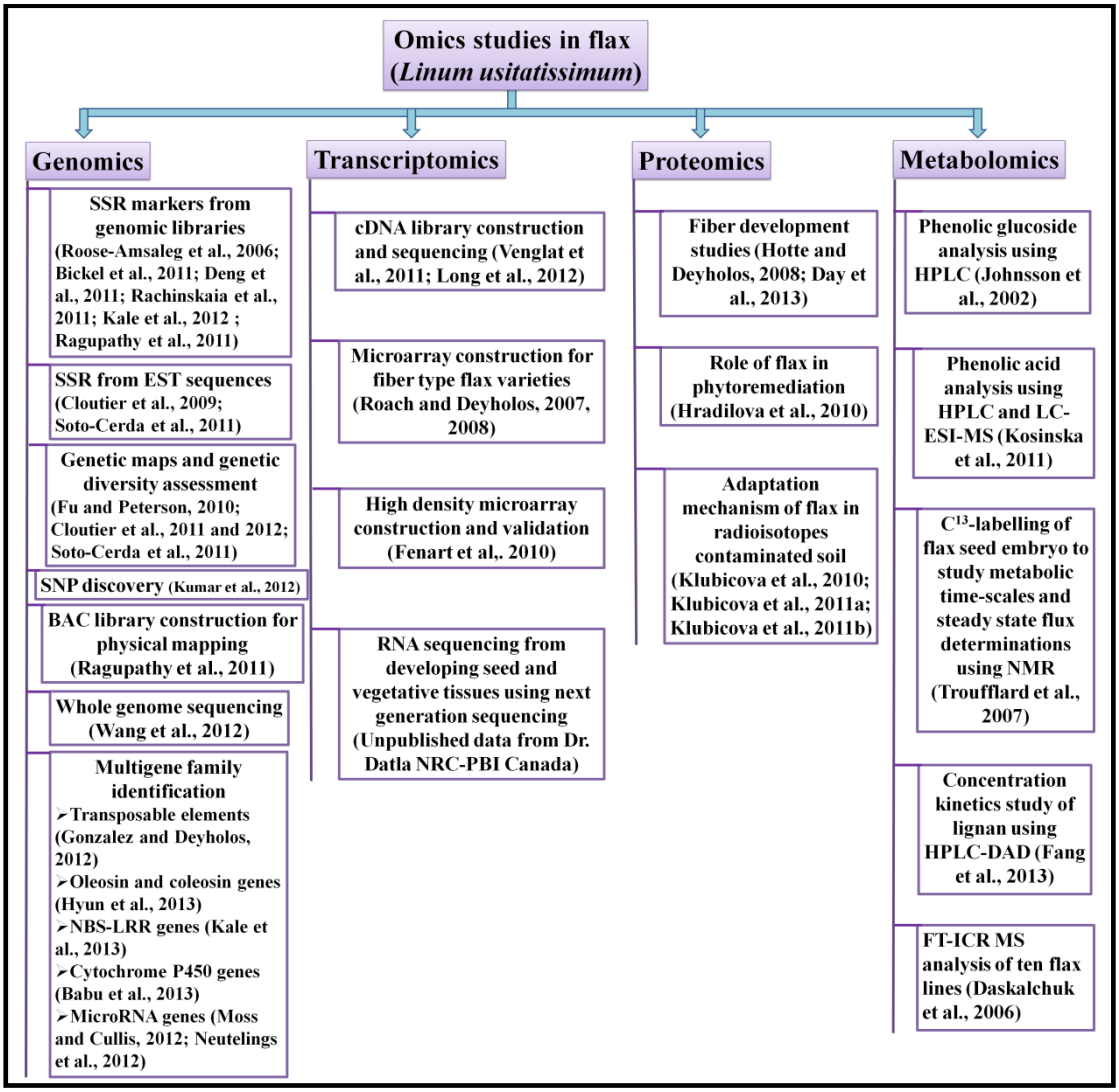


Figure 1.5 Tabular representation of omics studies in flax.

1.6 Genesis of thesis and its organization

Flax is one of the economically important crops for its nutraceutical, medicinal and industrial attributes. Recently, extensive genomic resources for flax have been developed; however, characterization and annotation of these identified genes is still a challenging task. The biosynthetic pathways for commercially important components of flax seed such as oil, lignan and dietary fibers were known for a long time but regulation mediated at post translational level (i.e. at proteome level) still remains unknown. Role of enzymes involved in various pathways, their regulatory post translational modifications and flux which determine abundance of particular metabolite is still a mystery in case of flax. The present study attempts to address these questions by profiling developing seed proteome comprehensively and characterizing various functional metabolic pathways along with quantitative

expression of enzyme transcripts involved in it. Further, to identify regulators of these metabolic pathways, the present study identified microRNAs from whole genome sequence. The microRNAs are well known upstream master regulators of the biological processes in plants. Target transcripts of identified microRNAs code for transcription factors and regulate metabolic pathways such as oil biosynthesis. Another, important aspect of flax is lignans. Their biosynthetic pathway is known to be regulated by terminal glycosyltransferase enzymes. This particular class of enzymes was not reported from flax. In the present study, multigene family of UDP-glycosyltransferase 1 was identified from flax using bioinformatics approach. Phylogentic and transcript expression analysis of various genes suggested candidate genes for lignan biosynthesis. This study provides new insights into the complex seed developmental processes operating in flax and various regulators of these processes.

The thesis is presented in five chapters and the content of each chapter is as follows:

Chapter 1: Introduction and review of literature (the current chapter)

Chapter 2: Proteome profiling of developing flax seed and characterization of functional metabolic pathways operating during seed development

This chapter describes various methods for protein prefractionation of complex seed proteome and further explains the functional existence of carbon flux pathways, fatty acid biosynthesis, methionine metabolism and secondary metabolite biosynthesis pathways in the developing flax seed.

Chapter 3: Genome-wide identification and characterization of microRNA genes and their targets in flax and their role in oil accumulation

This chapter describes bioinformatics based workflow for identification and characterization of microRNA genes and their respective protein coding target transcripts. Further role of microRNAs and their targets in enhancing oil yield has also been discussed.

Chapter 4: Phylogenomic analysis of UDP glycosyltransferase 1 multigene family in flax regulating secondary metabolite accumulation

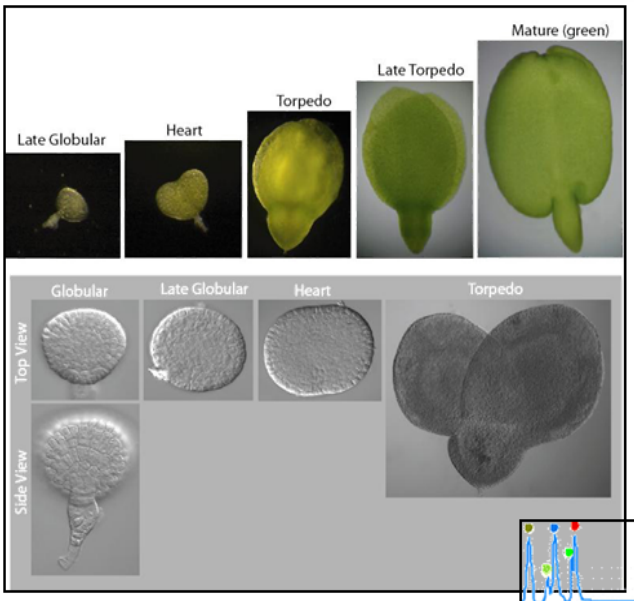
This chapter describes flax UDP glycosyltransferase 1 multigene family evolution, phylogenetic tree and intron architecture and also compares these data with closely related plants. Expression analysis using various techniques suggested candidate genes as lignan glycosyltransferase and this has been discussed in detail.

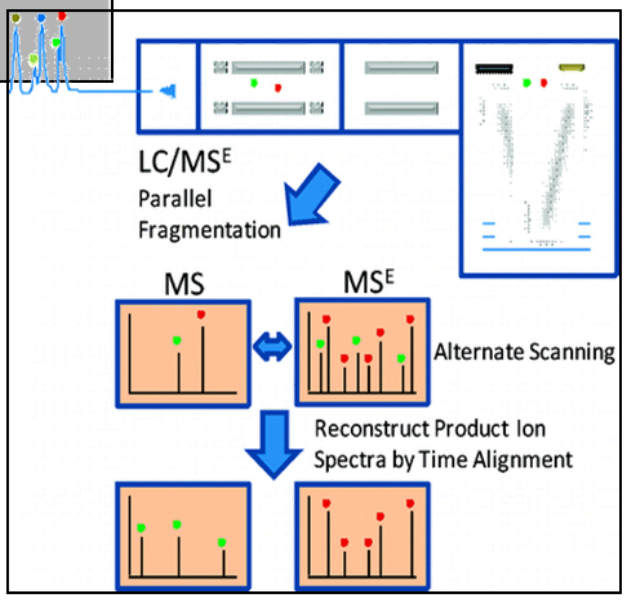
Chapter 5: Summary and future directions

Bibliography

CHAPTER 2

Proteome profiling of developing flax seed and characterization of functional metabolic pathways operating during seed development





A part of this work has been published in the *Journal of Proteome Research* (2012) 11, 6264–6276

2.1 Introduction

Seed development is a highly complex process and many genes involved in various pathways are under precise regulation for proper development. Therefore, it is necessary to apply systematic and parallel approaches on a global scale to elucidate the relationships among various metabolic networks operating during seed development. Recent advances in various high-throughput technologies provide a unique opportunity to analyze biological systems on a genome-wide scale. The relationship between protein and oil composition in Arabidopsis seed was revealed by the proteomics approach (Chen et al., 2009). Proteomes of oilseeds, such as soybean (Hajduch et al., 2005; Agrawal et al., 2008), rapeseed (Hajduch et al., 2006), Castor (Houston et al., 2009) and Arabidopsis (Hajduch et al., 2010) were investigated to understand the metabolic pathways functional during the respective seed development. Biochemical and genetic studies of flax seeds have been carried out to identify, characterize and understand the accumulation of fatty acids; cyanogenic compounds (Frehner et al., 1990; Niedzwiedz-Siegien, 1998), ALA (Daun and DeClercq, 1994), lignan (Ford et al., 2001; Eliasson et al., 2003) and seed storage proteins (DeClercq and Daun, 2002). Recently, extensive transcriptomic analyses of developing seeds have been performed to describe biosynthetic pathways leading to the accumulation of these storage compounds in flax seed (Fenart et al., 2010; Venglat et al., 2011). These studies catalogued temporal and spatial changes in gene expression specific to metabolic pathways related to the accumulation of oil, protein and carbohydrate (mucilage) reserves during seed development. Despite this, little is known about the translational and post-translational regulation of proteins during seed development. For example, although biosynthetic pathway of fatty acid production in flax is known, the biological mechanisms controlling linolenic acid level remains unknown (Rao et al., 2008). Thus, understanding the seed development process in flax is vital since they accumulate diverse seed storage compounds of nutritional and economic importance.

The goal of the present study was to develop a comprehensive proteomics based dataset for flax to better understand its seed development. By comparison of the results obtained in this study with those from other model plants, universal and species specific biological mechanisms involved in seed development can be identified.

2.2 Materials and Methods

2.2.1 Plant material

The flax variety NL-97 was grown at College of Agriculture, Nagpur (2010-2011, November-March). Flax flowers were tagged after opening (between 7.30 and 8.30 AM IST), and developing bolls were collected at 4, 8, 12, 16, 22, 30 and 48 days after anthesis (DAA). Developing seeds were harvested from bolls at 4 ºC to prevent dehydration and stored at -80 ºC until further use. To determine seed fresh weight and dry weight, three pools of 10 randomly selected seeds were weighed just after harvest and after drying at 50 ºC for 24 h in hot air oven. The seed length, thickness and width were measured using a scale.

2.2.2 Fatty acid analysis of developing seeds

Fatty acid methyl esters (FAMES) were extracted from seven seed developmental stages as described by Rajwade et al. (2010) with some modifications. 1µL of chloroform reconstituted extracts were injected in AutoSystem XL GC (PerkinElmer, USA) with SP-2330 Supelco capillary column, 30 m long and 0.32 mm diameter. Fatty acid (FA) peaks were identified by comparing them with the standard FA profiles (Sigma–Aldrich, USA). The area under the peak was expressed as percentage fatty acid content. FA profiling for each stage was repeated twice. Mean of three replicates was averaged to get the percent FA content of each developmental stage.

2.2.3 Optimization of protein pre-fractionation method

Total proteins from each seed developmental stage were extracted according to Wang et al. (2006). Three independent protein extractions were performed with the tissue. Total protein content was measured at each stage using a dye-binding protein assay (Bradford, 1976). To ensure a systematic coverage of the flax seed proteome, three protein pre-fractionation methods *viz*., 1D-SDS PAGE, 2D-SDS PAGE and in-solution digestion were evaluated. The peptides extracted from all the three pre-fractionation methods were identified using LC-MS^E (**Figure 2.1**). In 2D-SDS PAGE, 16 DAA stage showed the maximum number of protein spots (**Figure 2.1**). Hence, this stage was selected for further protein identification using the three pre-fractionation methods. All the protein spots from 2D-SDS-PAGE (16 DAA stage) were picked, extracted and subjected to LC-MS^E analysis for protein identification.

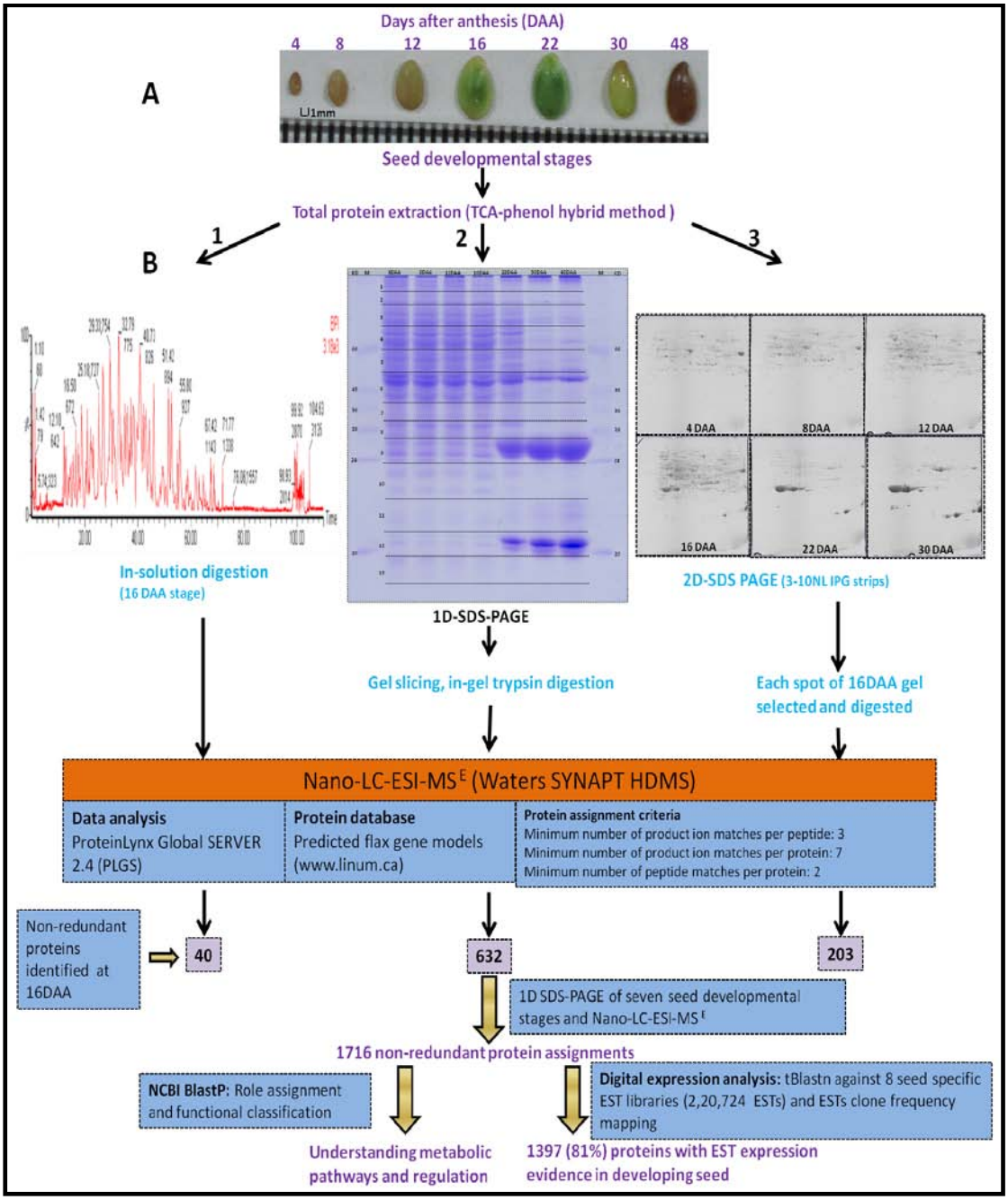


Figure 2.1 Schematic representation of experimental design for proteome analysis of developing flax seeds: A) The flax seed developmental stages (7 stages; 4, 8, 12, 16, 22, 30 and 48 DAA) subjected to proteome analysis. Total proteins were extracted using the TCA-phenol hybrid method. B) Three protein pre-fractionation methods were evaluated: 1) In-solution trypsin digestion followed by nano LC-ESI-MS^E, 2) 1D-SDS-PAGE (12%), gel slicing (13 pieces), in-gel trypsin digestion and nano LC-ESI-MS^E 3) 2D-SDS PAGE (pH 3-10 non-linear), protein spot picking followed by in-gel trypsin digestion and nano LC-ESI-MS^E to identify the best method giving maximum proteome coverage. The LC-MS/MS spectra were searched against the predicted flax gene models database (http://www.linum.ca) using

ProteinLynx Global Server 2.4 (PLGS) with criteria as described in section 2.2.6. The identified non-redundant proteins were BlastP searched against the NCBI-nr protein database along with Blast2Go software to assign the functional role to the identified proteins. The identified proteins were TBlastN searched against the seed specific EST libraries to study the digital expression analysis.

This resulted in identification of 203 non-redundant proteins, whereas only 40 non-redundant proteins were identified by the in-solution digestion method. Total proteins (80 µg) were also subjected to 1D-SDS-PAGE gel that was cut into 13 slices and LC-MS^E analysis was performed, which identified 632 non-redundant proteins. Hence, 1D-SDS-PAGE was considered as an ideal pre-fractionation method for flax seed proteomic analysis and was used further for all the seven seed developmental stages.

2.2.4 1D-SDS-PAGE and in-gel digestion

Protein samples (80 µg each) were loaded on 12% SDS polyacrylamide constant separation gel with a 4% stacking gel and electrophoresed at 20 °C using a vertical PROTEAN II xicell (BioRad, USA) at constant current of 50 A/gel. The gels were visualized with coomassie brilliant blue staining (CBB) and scanned at 300 dpi using a high resolution image scanner (BioRad GS 800, USA). Each lane (representing a seed developmental stage) of the gel was sliced with a sterile scalpel into 13 pieces based on protein abundance as outlined in **Figure 2.1**. Each slice was transferred into a 1.5 mL tube and in-gel digestion using trypsin was performed as described by Haynes et al. (1998) with slight modifications. In brief, protein bands were excised from the CBB-stained gels, washed twice with milli-Q water, destained with a 1:1 (v/v) solution of 50% acetonitrile and 50 mM NH_4HCO_3 and then dehydrated in 100% acetonitrile (ACN) until the gel pieces were shrunken. The dried gel pieces were reduced with 10mM dithiothreitol for 45 min at 56 °C and alkylated with 55 mM iodoacetamide in dark at RT for 40 min. Gel pieces were dehydrated and then digested with trypsin (0.02 IU/µL) (Sigma-Aldrich, USA) at 37 °C overnight. The resulting peptides were extracted using sonicator (Branson, USA) twice by adding 200 µL of solution containing 0.1% formic acid and 5% acetonitrile for 15 min, respectively. The peptides were dried in a freeze dryer (Labconco, USA) and then reconstituted in 10 µL of 5% aqueous ACN containing 0.1% formic acid for subsequent analysis.

2.2.5 Liquid chromatography-mass spectrometry analysis

All the samples were analyzed by LC-MS^E using a NanoAcquity ultra performance liquid chromatography (UPLC) system (Waters, USA) coupled to a SYNAPT High Definition Mass Spectrometer (Waters, USA). The nano-LC separation was performed using a bridged-ethyl hybrid (BEH) C18 reversed phase column (1.7 µm particle size) with an internal diameter of 75 µm length of 150 mm (Waters, USA). The binary solvent system that was used, comprised 99.9% water and 0.1% formic acid (mobile phase A) and 99.9% acetonitrile and 0.1% formic acid (mobile phase B). The samples were initially pre-concentrated and desalted online at a flow rate of 5 µL/min using a Symmetry C18 trapping column (internal diameter 180 µm, length 20 mm) (Waters, USA) with a 0.1% mobile phase B. Each sample (total digested protein) was applied to the trapping column and flushed with 0.1% solvent B for 3 min at a flow rate of 15µL/min. After each injection, peptides were eluted into the NanoLockSpray ion source at a flow rate of 300 nL/min using a gradient of 2-40% B over 50 min. The lock mass calibrant peptide standard, 600 fmol/µL Glu-fibrinopeptide B (Sigma-Aldrich, USA), was infused into the NanoLockSpray ion source at a flow rate of 300 nL/min and was sampled during the acquisition at 30s intervals. The mass spectrometer was operated in V-mode at a resolution of at least 9000 full width at half height (fwhh). For LC-MS^E, full scan (m/z 50-2000) data were collected using the "expression" mode of acquisition, which acquires alternating 1s scans of normal and elevated collision energy. Data were collected at a constant collision energy setting of low (4V) and high (ramp from 20 to 40 V) energy mode MS^E scans.

2.2.6 Data processing and database searching

The continuum LC-MS^E data were processed and searched using ProteinLynx Global Server 2.4 (PLGS; Waters, USA) software. Protein identifications were obtained by searching the 47,912 predicted gene models from flax whole genome sequence (November, 2011; http://www.linum.ca). LC-MS^E data were searched with a fixed carbamidomethyl modification for cysteine residues, along with a variable modification for oxidation of methionine, N-terminal acetylation, deamidation of asparagine and glutamine and phosphorylation of serine, threonine and tyrosine. The ion accounting search algorithm within PLGS was developed specifically for

searching data-independent MS^E data sets, and the algorithm has been described in detail by Li et al. (2009). The ion accounting search parameters were; precursor and product ion tolerance: automatic setting, minimum number of product ion matches per peptide: 3, minimum number of product ion matches per protein: 7, minimum number of peptide matches per protein: 2, and missed tryptic cleavage sites: 2. The false positive rate was 4%. Search results of the proteins and the individual MS/MS spectra with confidence level at or above >95% were accepted.

2.2.7 Functional classification

Gene Ontology (GO) annotation was performed with Blast2GO software (Conesa et al., 2005) based on sequence similarity. For annotation, the default configuration settings were used and the proteins were searched against the NCBI-nr protein database (December, 2011). Plant related GO terms were then retrieved using the GOslim viewer from the AgBase web server (http://www.agbase.msstate.edu). Proteins with unknown function or without specific homology or similarity descriptions were BlastP searched against the NCBI-nr database to confirm their functions. Sub-cellular localizations of proteins were predicted using WoLF PSORT (http://www.wolfpsort.org/) web server (Horton et al., 2007). The information about subcellular localization was incorporated into protein description. The identified proteins were then classified into 15 functional classes according to Hajduch et al. (2006).

2.2.8 Digital expression analysis

Developing flax seed specific ESTs were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/nucest/?term=linum%20usitatissimum) [eight libraries; Globular embryo (GE), heart shaped embryo (HE), torpedo embryo (TE), cotyledon embryo (CE), mature embryo (ME), globular stage seed coat (GC) and torpedo stage seed coat (TC) and pooled endosperm (EN), total 2,20,724 EST sequences; Jan, 2012)]. Makeblastdb (ftp://ftp.ncbi.nlm.nih.gov/blast/) was used to create developing flax seed EST database on an in-house server. Viroblast (v. 2.2+) (Deng et al., 2007) was configured to use the standalone suite of Blast programs (v. 2.2.24+). The identified proteins were TBlastN searched against the EST database with an E-value threshold of $\leq e^{-10}$ and $\geq$95% sequence identity criteria to map the ESTs onto gene models to obtain transcriptional evidence for individual proteins.

2.2.9 Reverse transcription quantitative real time PCR

RNA was extracted from seven seed developmental stages using Spectrum^{TM} plant total RNA kit (Sigma-Aldrich, USA). DNaseI treated total RNA was reverse transcribed using oligo(dT) primer and MultiScribe^{TM} reverse transcriptase (Applied Biosystems, USA). Gene specific primers for 19 genes were designed using Primer3 (**listed in Table S 2.1**). PCR conditions were optimized for annealing temperature and primer concentration. Real-time PCR was carried out in 7900HT Fast real-time PCR system (Applied Biosystems, USA) using FastStart universal SYBR green master mix (Roche, USA). Real-time PCR amplification reactions were performed as detailed below. Each 10 µL real-time PCR cocktail contained both forward and reverse gene-specific primers, 4 µL of 1:16 diluted first strand cDNA, ⅓ SYBR green master mix and sterile milliQ water to make up the reaction volume. Real-time PCR amplification reactions were performed with following conditions: 95 °C denaturation for 10 min, followed by 40 cycles of 95 °C for 3s, with primer annealing and extension at 60 °C for 30 s. Following amplification, a melting dissociation curve was generated using a 62–95 °C ramp with 0.4 °C increment per cycle in order to monitor the specificity of each primer pair. Eukaryotic translation initiation factor 5A (*ETIF5A*) gene from flax was used as a housekeeping or reference gene for all the real time PCRs (Huis et al., 2010). The housekeeping gene was selected after confirming the stability of this gene across all the tissue types used in the study. For each of the three biological replicates, two independent technical replications were performed and averaged for further calculations. PCR conditions were optimized such that PCR efficiencies of housekeeping gene and gene of interest will be closer to two. PCR efficiencies were calculated using LinRegPCR v12.x (Ramakers et al., 2003). Relative transcript abundance calculations were performed using comparative C_T (ΔC_T) method as described by Schmittgen and Livak (2008).

2.3 Results and Discussion

The present study of flax seed proteome describes a detailed analysis of proteins identified in developing flax seeds and characterizes the metabolic pathways operating during seed development. The major economic value of flax lies in its contents of oil and secondary metabolites and a thorough knowledge of the regulation

of their accumulation during seed development could be exploited for various applications.

2.3.1 Biochemical characterization of developing flax seed

The biochemical characterization of developing flax seed was performed to define various seed developmental events. Whole flax seeds were analyzed at 4, 8, 12, 16, 22, 30 and 48 days after anthesis (DAA). This time interval covered the major seed developmental events such as embryogenesis, seed filling and maturation (Gutierrez et al., 2006). Seed size and color changed during the course of seed development (**Figure 2.1**). The fresh seed weight increased till 22 DAA and declined later, indicating that the seeds entered the pre-desiccation phase; whereas, dry seed weight started increasing after 12 DAA (**Figure 2.2 A**) and continued to increase till maturity. This observation was also supported by shrinkage of seeds at the maturation phase indicated by the decrease in size (**Figure 2.2 B**). During embryogenesis (from 4 to 16 DAA), water content of the seed remained high (~90%); whereas, the protein content increased during seed filling, which coincided with increased fresh as well as dry weight of the seed revealing high metabolic activity in flax seeds during the 4 to 16 DAA (**Figure 2.2 C**).

2.3.2 Fatty acid analysis of developing seed

At the early developmental stages (4-8 DAA), palmitic acid (PA; C16:0), linoleic acid (LA; C18:2), and ALA (C18:3) were the most abundant FAs. The PA content was high (30.74%) at 4 DAA stage, further gradually decreased throughout seed development till maturity. The proportion of stearic acid (SA; C18:0) remained nearly constant. The oleic acid (OA; C18:1) content showed steady increase from 4 (6.19%) to 12 DAA (28%) and then plateaued. The accumulation pattern of LA and ALA showed a correlation between product and precursor ratio. LA content was high (34.28%) at the 4 DAA stage and gradually declined from 12 to 48 DAA. The ALA content was low at 4 DAA (22.66%) and from 12 DAA onwards showed a steady increase till maturity. At maturity the total ALA accumulation was 46% whereas, LA content was 11.84% of the total fatty acids quantified (**Figure 2.2 D**). Thus, biochemical analysis of flax seed indicated that the period from 4-16 DAA constitutes the embryogenesis phase while the period from 17-30 DAA and 31 onwards represents seed filling and maturation phases, respectively.

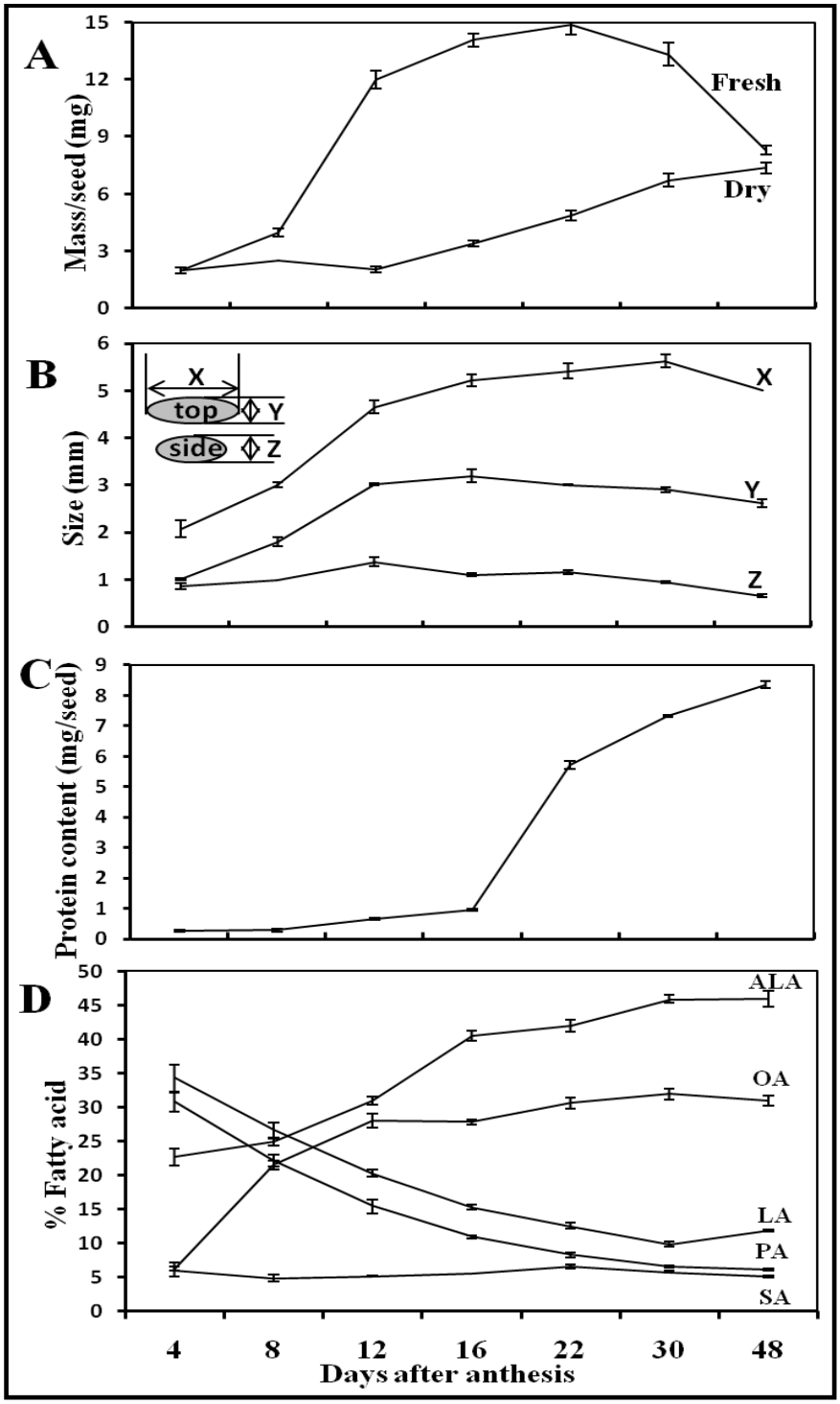


Figure 2.2 Biophysical and biochemical characterization of flax seed at seven seed developmental stages: A) Seed fresh and dry mass during the experimental period expressed as mass per seed. Values are the average of 10 determinations; SD is denoted as error bars. B) Individual seed size characteristics, including length (x), thickness (y), and width (z), were determined using scale. Each value is an average of 10 seeds; SD is denoted by error bars. C) Total protein content per seed during the seven seed developmental stages. Values are the average of 3 replicates; SD is

denoted as error bars. D) Percent fatty acid content per seed analyzed during the seven seed developmental stages. Accumulation of five main fatty acids, namely, palmitic (PA, C16:0), stearic (SA, C18:0), oleic (OA, C18:1), linoleic (LA, C18:2), and linolenic (ALA, C18:3) acids is illustrated. Values are the average of three biological replicates; SE is shown as error bars.

2.3.3 1D-SDS-PAGE and LC-MS^E efficiently cover the flax seed proteome

The developing seed proteome is a complex mixture and seed storage proteins constitute ~23% of total flax seed protein (DeClercq and Daun, 2002). However, this high abundance of storage proteins masks the identification of low abundant proteins and requires efficient pre-fractionation. To reduce and fractionate the protein complexity, 1D-SDS-PAGE was used and the fractionated proteins were identified using the LC-MS^E. A total of 1716 non-redundant proteins were identified (**for summary of identified proteins see Table S 2.2**). Good coverage of the flax proteome was obtained using 1D-SDS-PAGE. Earlier studies also reported 1D-SDS-PAGE as the most effective method for maximum proteome coverage and to fractionate complex protein mixtures (Fang et al., 2010). Previously, such a combination of 1D-SDS-PAGE protein separation and LC-MS/MS analysis was employed in shotgun proteomic analysis of Arabidopsis leaves (Lee et al., 2007), to analyze the roles of chloroplast proteases (ClpPR) in plastid biogenesis and homeostasis (Zybailov et al., 2009) and to understand the kiwi fruit ripening (Minas et al., 2012).

2.3.4 GO annotation, functional categorization and family relationship

To understand the biological functions of the expressed proteins, Gene Ontology (GO) annotation was performed. The identified 1716 proteins were analyzed using BlastP against the NCBI-nr plant database, which yielded 14,801 GO annotations. The Venn diagram (**Figure S 2.1**) shows the distribution of GO annotation in the three categories.

According to GO annotation, the cellular component assignments were mostly to the cellular components (15%), cell, intracellular and cytoplasm (13% each), plastid (10%), membrane (8%) and cytosol (6%) (**Figure 2.3 A**). As per the biological processes of the flax seed proteome, 18% proteins with metabolic processes, 17% with cellular processes, 15% with biological processes and 6% with biosynthetic

processes were observed in the flax proteome (**Figure 2.3 B**). Furthermore, the highly enriched molecular functions of the flax seeds were revealed as proteins with catalytic activity (25%) followed by binding proteins (18%) and hydrolase activity (13%), respectively (**Figure 2.3 C**).

Most of the identified proteins (19%) were involved in primary metabolism, followed by protein destination and storage (14%) and energy (10%). Similar observation was reported in other oilseeds such as rapeseed (Hajduch et al., 2006), soybean (Agrawal et al., 2008) and castor (Houston et al., 2009). More than half of the flax proteins showed high similarity to castor bean proteins, supporting the taxonomic grouping of flax and castor bean within the order Malpighiales. The flax transcriptomic data also showed similar results (Venglat et al., 2011). Such information will be helpful for molecular taxonomy of this order as the taxonomic relationship of families within this order is still poorly resolved.

2.3.5 Transcriptional evidences correlate with proteome profiling

To obtain preliminary transcriptional evidence, the frequency of the ESTs mapping per protein was determined using the publicly available flax EST datasets from the developing seeds. A total of 1397 proteins (81%) had corresponding transcripts expressed in previously mentioned libraries. This approach confirms gene expression at transcript level and would be helpful for candidate gene isolation and characterization (**Figure S 2.2 and Table S 2.3**). However, when the whole EST database of flax was used for the analysis, 95% proteins had transcriptional evidence **(Table S 2.4)**. The number of mapped transcripts onto identified proteins varied from 1-10 in various tissue types.

2.3.6 Diversity of seed storage proteins in flax

A majority of the flax seed proteins are storage proteins and constitute ~23% of the whole flax seed (DeClercq and Daun, 2002). Usually, seeds of many plant species accumulate two of the three (2S, 7S and 11S) types of storage proteins. Previous studies reported that flax seed contains an unusual diversity of storage proteins in the form of cupin, conlinin and cruciferin (Venglat et al., 2011). In the present study, 2S albumin (conlinin) and 11S globulin (legumin, glutenin type A, cupin) proteins were identified, confirming previous reports. In addition, 7S globulin (48kDa glycoprotein precursor) protein was also detected.

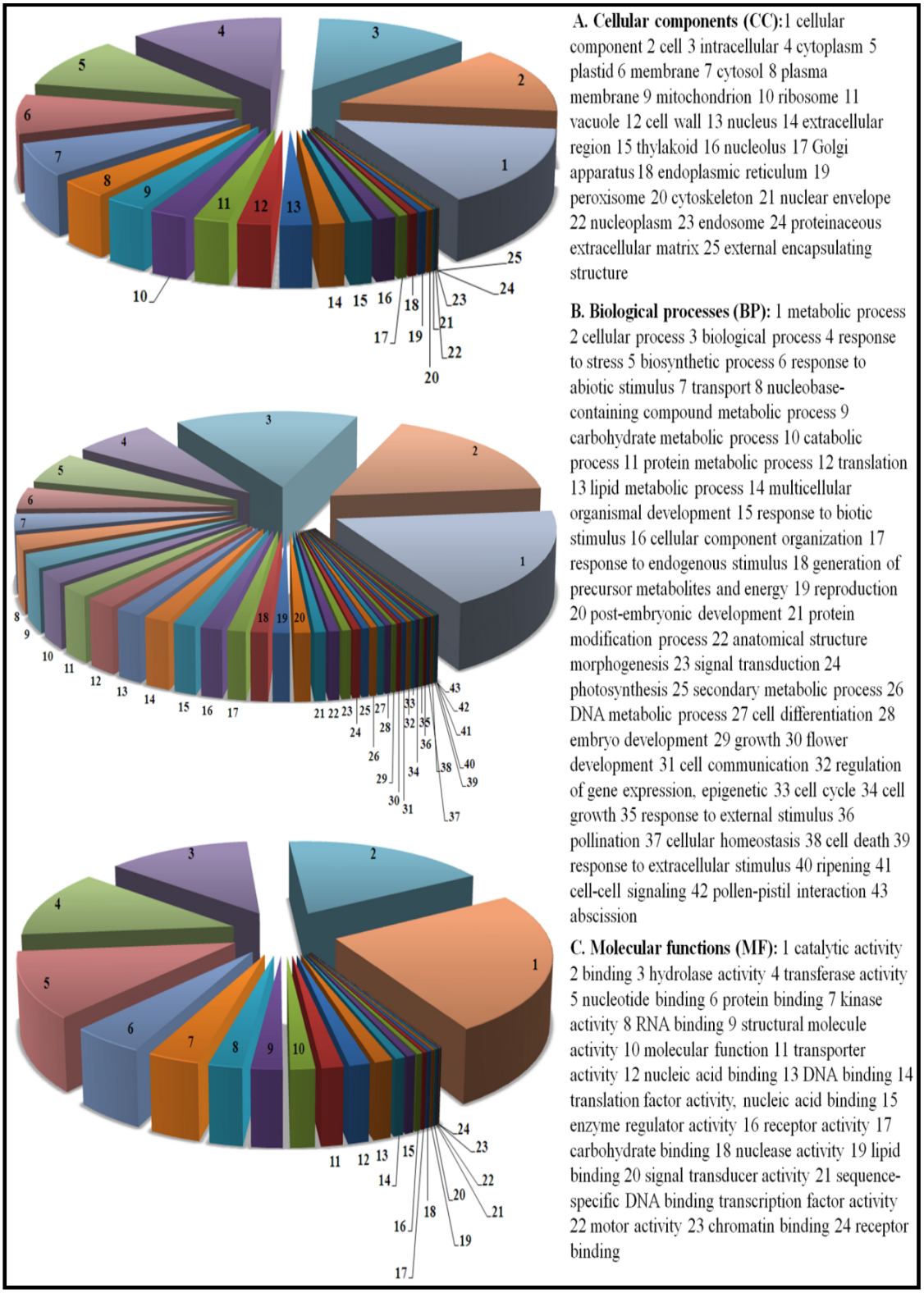


Figure 2.3 Gene ontology analysis of proteins identified in developing flax seeds: A total of 1716 unique proteins were analyzed with the Blast2GO program. The GO categories: (A) 25 cellular components (CC), (B) 43 biological processes (BP), and (C) 24 molecular functions (MF) are shown as pie charts.

This heterogeneity was consistent with the fact that seed reserve proteins are encoded by multi-gene families (Shewry et al., 1995). Expression of legumin B protein was much more abundant compared to other seed storage proteins analyzed (**Figure 2.4 A-D**). These storage proteins concentrated in a specific temporal order with 7S and 11S (except cupin) accumulating during the maturation phase, while 2S was identified throughout seed development. Transcripts encoding the abundant storage proteins were expressed in similar developmental stages (Venglat et al., 2011), suggesting that the temporal accumulation of storage proteins is transcriptionally controlled.

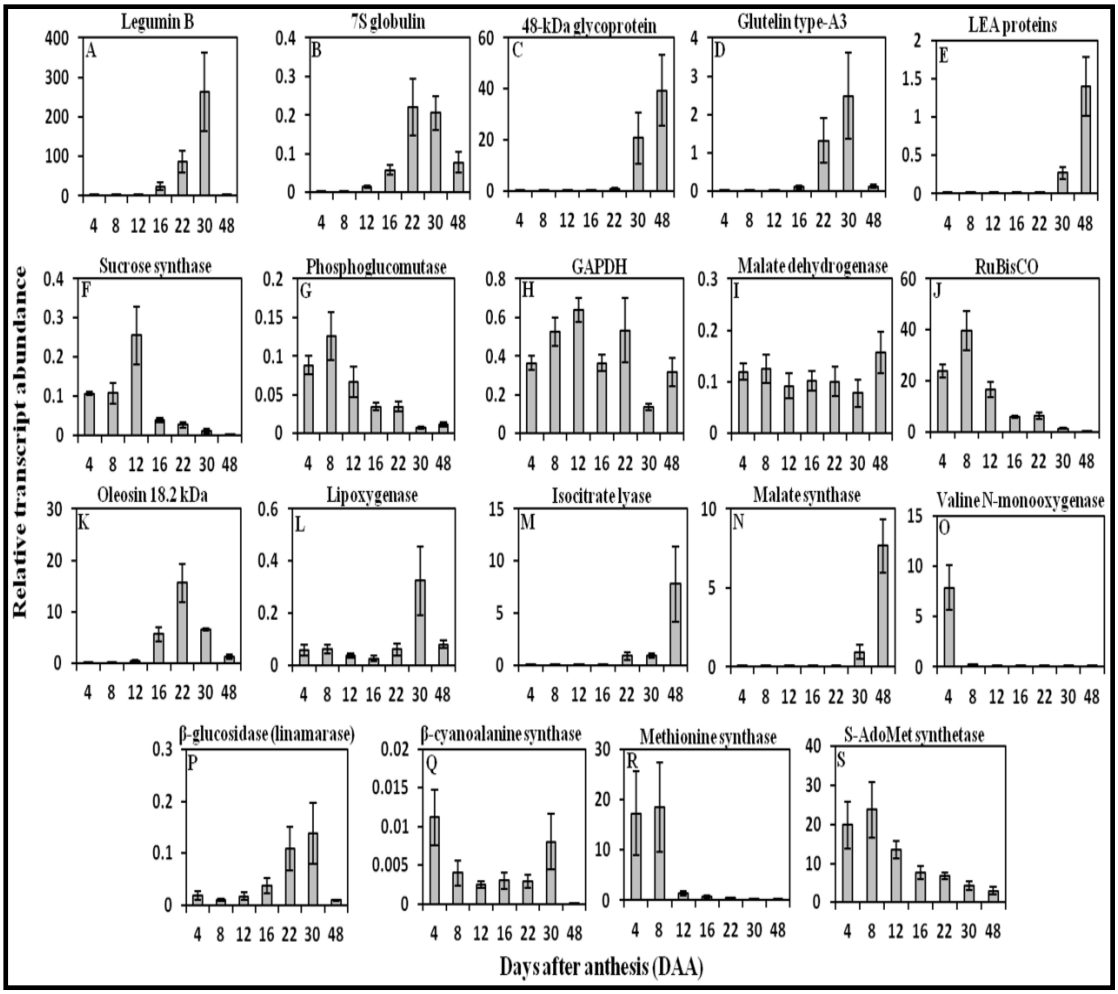


Figure 2.4 RT-qPCR expression profiles of 19 candidate genes. Seven seed developmental stages (4, 8, 12, 16, 22, 30, and 48 DAA) were used in the study. These graphs show the relative transcript abundance of each gene in comparison with the reference gene, *Linum usitatissimum ETIF5A* (GR508912). Expression values are reported as the average of three biological and two technical replicates for each. Values correspond to the mean and standard error of biological triplicates.

2.3.7 Flax seed maturation marker proteins

Several proteins known to play important roles in seed maturation were also identified. Gutierrez et al. (2006) reported the expression of cysteine protease gene, which marks the embryo growth phase, at 10 days after flowering (DAF) in the Barbara ecotype of flax. Whereas, in NL-97, it was identified at 8 DAA indicating that the embryo was still in growth phase. The LEA proteins were identified during the desiccation phase of seed maturation (Wise and Tunnacliffe, 2004). In NL-97, the LEA protein was identified at 30-48 DAA, and this was also supported by the RT-qPCR expression indicating that this period represents the late maturation phase of flax seed (**Figure 2.4 E**). Moreover, at 30 DAA, there was a sharp decrease in seed fresh weight indicating that the seed entered the desiccation or maturation phase at this time point.

In general, storage proteins are expressed abundantly during the cell expansion phase, after the embryo has been completely differentiated. Gutierrez et al. (2006) reported that flax seed filling phase lasts from 20-30 DAF; however, in NL-97, the 16 DAA stage indicated the start of seed storage phase and a majority of the flax storage proteins were identified at this stage. Therefore, 16 to 30 DAA might represent the seed filling stage in NL-97. The transition between the embryogenesis and seed filling phases probably occurred at around 16 DAA, which marked a key change in several aspects of seed development; most importantly, the beginning of storage product accumulation.

2.3.8 Carbon entry and assimilation by glycolytic pathway

It has been suggested that carbon assimilation during seed development begins with the transport of photosynthetically assimilated carbon in the form of sucrose into seed and cleavage of sucrose plays a key role in inducing the seed maturation process. Sucrose synthase (SuSy) is the key enzyme implicated in the first step of sucrose metabolism. Four isoforms of SuSy were identified mainly during embryogenesis and seed filling stages, suggesting the presence of two types of SuSy and active sugar metabolism during these stages. We analyzed the expression of one of the isoforms; g4346 using RT-qPCR and transcript abundance was higher during embryogenesis with peak at 12 DAA (**Figure 2.4 F**). The hexose phosphates generated via the action of SuSy are metabolized through glycolysis or the oxidative pentose phosphate

pathway (OPPP). Proteomic, biochemical and transcriptional studies suggest that glycolytic enzymes play a critical role in carbon metabolism during seed development (Dennis and Miernyk, 1982; Gallardo et al., 2007; Hajduch et al., 2011). In our study of flax seed development, both cytosolic and plastidial isoforms of all the glycolytic enzymes were identified suggesting both were important for flax during seed development (**Figure 2.5**). The presence of such complete glycolytic pathway has been reported in rapeseed embryos, cauliflower buds, castor seed endosperm and non-photosynthetic plastids from pea (Simcox et al., 1977; Journet and Douce, 1985; Denyer and Smith, 1988; Eastmond and Rawsthorne, 2000).

For almost all the enzymes, both cytosolic and plastidial isoforms were detected, except for the cytosolic phosphoglucomutase (PGM) and plastidial 2,3-biphosphoglycerate-independent phosphoglyceratemutase (iPGAM) isoforms. A possible explanation for this could be the lower expression of these proteins during seed development. Likewise, the plastidial iPGAM could not be detected even in rapeseed (Hajduch et al., 2006). A notable feature of our study was the detection of only plastidial isoforms of the PGM expressed especially during the embryogenesis stages (**Figure 2.4 G**). The plastidial PGM is reported to play an essential role in starch synthesis (Dietz, 1987) as well as degradation of assimilatory starch (Hattenbach and Heineke, 1999). The starch accumulated in young embryos of oilseeds provides carbon resources for lipid biosynthesis and thus, the plastidic PGM is a crucial factor affecting seed oil content (Periappuram et al., 2000). The studies in plants lacking the plastidial PGM indicate that it has a significant impact on the deposition of other storage products in seed (Casey et al., 1998). Therefore, in flax, plastidial PGM might have a crucial role to play in oil accumulation. In rapeseed, the seed specific expression of cytosolic glyceraldehyde 3-phosphate dehydrogenase (GAPDH) increased glyceraldehyde 3-phosphate levels by three to four fold in developing seeds, resulting in 40% increase in the oil content (Vigeolas et al., 2007). Therefore expression pattern of GAPDH was studied. RT-qPCR of GAPDH revealed that it is expressed throughout seed development indicating GAPDH in flax might play a key role in seed development (**Figure 2.4 H**).

2.3.9 Varied carbon flux pathways are functional during flax seed development

Oil synthesis studies in plant seeds propose two major routes of carbon flow for *de novo* fatty acid synthesis (Schwender et al., 2004). In the first route, cytosolic

glycolysis takes place till phosphoenolpyruvate (PEP) synthesis followed by its transport into plastid and conversion into acetyl-CoA, which is further used for FA synthesis. The functional existence of this route of carbon flux is supported by transcriptomics, stable isotopic labeling and proteomic analyses of developing seed from Arabidopsis and rapeseed (Ruuska et al., 2002; Hajduch et al., 2006; Plaxton and Podesta, 2006) although a plastid pyruvate translocator is yet to be identified. In the present study, six of the seven isoforms of enolase identified were cytosolic in origin, indicating that PEP was transported across plastid as a carbon source for FA production. Therefore, this route of carbon flux seems to be operating in flax seeds.

In the second route of carbon flow, phosphoenolpyruvate carboxylase (PEPC) converts cytosolic PEP to oxaloacetic acid, followed by its malate dehydrogenase (MDH) mediated conversion to malate. The malate is imported in plastids and plastidial NADP malic enzyme decarboxylates it to pyruvate. Further, the pyruvate is converted to acetyl-CoA (by plastidial pyruvate dehydrogenase) for *de novo* FA synthesis. Biochemical studies of developing seed and isolated leucoplasts from castor bean (Plaxton and Podesta, 2006) support this second route of carbon flux. In the present study, MDH and plastidial NADP malic enzyme decarboxylates were identified. The MDH was expressed throughout the seed developmental stages suggesting that this route might also be functional in flax seeds (**Figure 2.4 I**). Collectively, these data indicate that both the routes of carbon flow might be functional in flax seeds. Further studies of these pathway enzymes might help in understanding the major routes of carbon flow for *de novo* fatty acid synthesis in flax seeds.

A third pathway of carbon assimilation is via Ribulose-1,5-bisphosphate carboxylase oxygenase (RuBisCO) bypass to fulfill the high demand of carbon for fatty acid synthesis (Schwender et al., 2006). In green photoheterotrophic plastids of rapeseed embryos, RuBisCO is able to fix CO_2 apart from the Calvin cycle. The identification of plastidial enolase and RuBisCO bypass enzymes (RuBisCO and phosphoribulo kinase) suggested that the 3-phosphoglycerate produced during the plastidial glycolysis needs to be converted to PEP using cytosolic enzymes and then transported back into plastids for FA synthesis. The RuBisCO was expressed during embryogenesis stage and was one of the highly expressed enzymes among the selected 19 genes used for RT-qPCR analysis, indicating that this route is functional during early seed developmental stages (**Figure 2.4 J**).

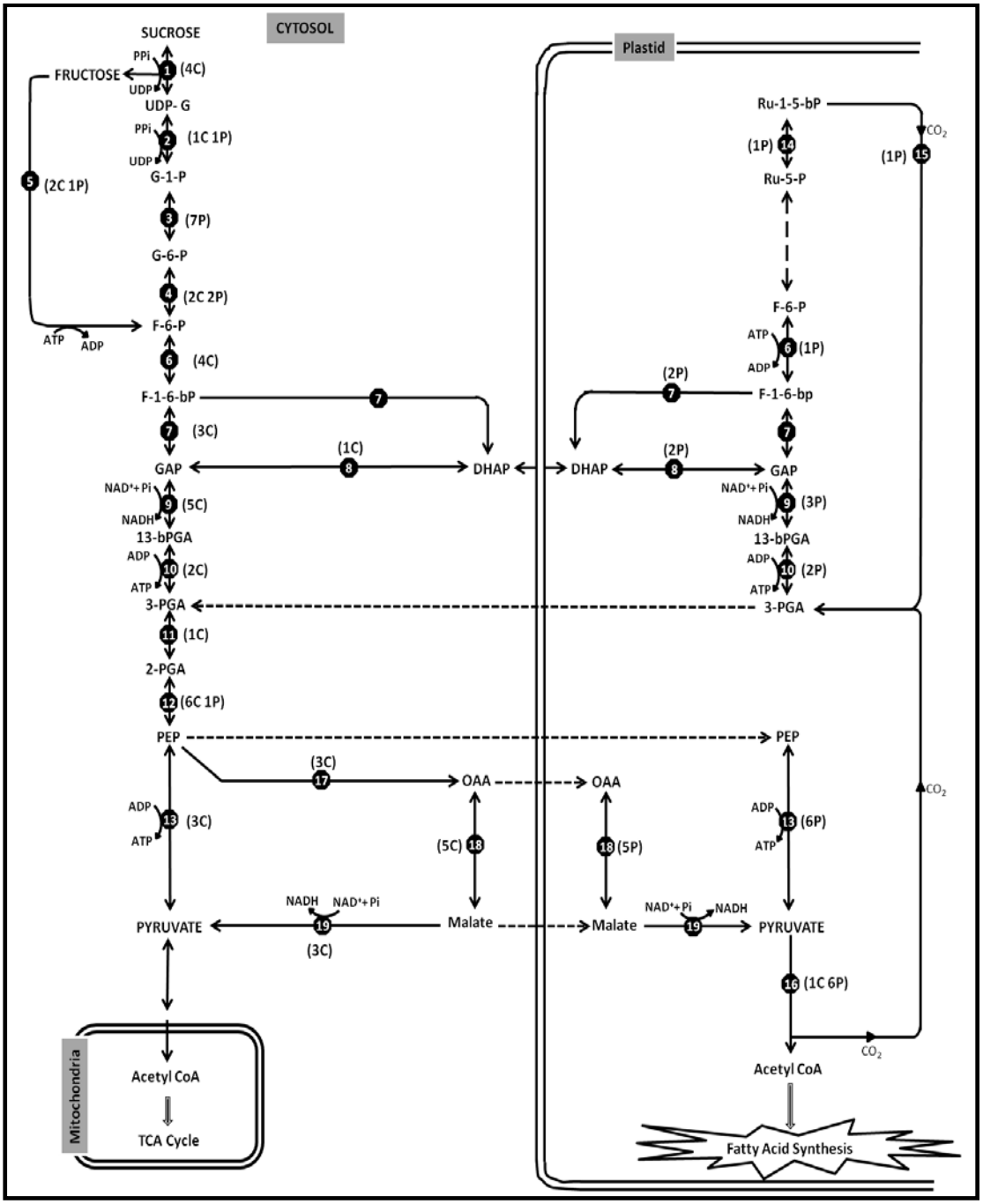


Figure 2.5 Schematic view of carbohydrate metabolism during flax seed development: Values in the parentheses represent the number of isoforms identified for that protein. C and P denote the cytosolic and plastidial locations of enzymes. Abbreviations for metabolites: UDP-G: uridine diphosphoglucose; G-1-P: glucose 1 phosphate; G-6-P: glucose 6 phosphate; F-6-P: fructose 6 phosphate; F-1,6bP: fructose 1,6-bisphosphate; GAP: glyceraldehyde 3-phosphate; DHAP: dihydroxyacetone phosphate; 1,3-bis PGA: 1,3-bisphosphoglyceric acid; 3-PGA: 3-phosphoglyceric acid; 2-PGA: 2-phosphoglyceric acid; PEP: phosphoenolpyruvate. The enzymes are (1) sucrose synthase, (2) UDP-glucose pyrophosphorylase, (3)

phosphoglucomutase, (4) glucose-6-phosphate isomerase, (5) fructokinase, (6) phosphofructose kinase, (7) fructose-bisphosphate aldolase, (8) triose-phosphate isomerase, (9) glyceraldehyde 3-phosphate dehydrogenase, (10) phosphoglycerate kinase, (11) 2,3-bisphosphoglycerate independent PGAM, (12) enolase, (13) pyruvate kinase, (14) phosphoribulokinase, (15) ribulose-1, 5-bisphosphate carboxylase oxygenase (RuBisCO), (16) pyruvate dehydrogenase, (17) PEP carboxylase, (18) malate dehydrogenase, and (19) NADP-dependent malic enzyme.

2.3.10 Lipid accumulation is associated with oleosin and lipoxygenase expression

Oil storage proteins such as oleosins (9 isoforms) and lipoxygenase (6 isoforms) were identified during seed maturation phase. Oleosins are structural proteins found in vascular plant oil bodies and help to stabilize oil bodies in seeds and account for about 2-8% of the total seed proteins (Huang, 1992). Nine isoforms of oleosin were detected during seed filling and mature stages. Expression pattern of oleosin isoform g44859 matched the pattern of flax FA desaturases (Banik et al., 2011) that are involved in the formation of the omega-3 fatty acids, indicating their role in the accumulation of oil (**Figure 2.4 K**).

Lipoxygenases (LOX) are widely distributed in plants and are critically important for plant growth and development (Porta and Rocha-Sosa, 2002). They are also involved in mobilizing storage lipids during germination (Feussner et al. 2001). Flax seed proteome as well as transcript analysis of isoform g14677 showed LOX proteins detected mainly during the maturation stage and therefore, most probably served as nutrient reservoir (**Figure 2.4 L**). High expression of LOX was also observed in developing seeds of Medicago (Gallardo et al., 2003) and soybean (Agrawal et al., 2008), thereby strengthening their role in oil accumulation in seeds.

2.3.11 High activity of β-oxidation and glyoxylate pathway might accumulate alpha-linolenic acid in flax seed

Fatty acids are widely found as major carbon and chemical energy reserves in seeds. Acetyl-CoA carboxylase (ACC), which catalyzes the conversion of acetyl-CoA into malonyl-CoA, is the key enzyme in FA biosynthesis. This plastid complex comprises four subunits, the biotin carboxylase, biotin carboxyl carrier protein and carboxyltransferase subunits (α and β). All the enzyme subunits (except the β subunit of the carboxyltransferase), were identified. Flax seed is enriched in ALA (C18:3n-3)

and fatty acid desaturases (SAD, FAD2 and FAD3) play important roles in synthesis of this fatty acid. The SAD (two isoforms), FAD2 and FAD3A proteins were detected at 22 DAA. Banik et al. (2011) detected peak in transcript activity of these enzymes during the similar seed developmental stages. These studies indicate that the expression of these proteins is transcriptionally controlled and these stages are important for ALA accumulation in flax seed.

Interestingly, the enzymes involved in β-oxidation of fatty acids during the embryogenesis and seed filling stages were also identified. β-oxidation occurs widely in plant tissues and also during seed development, implying a more general role during germination and seedling establishment than in lipid degradation (Moire et al., 2004). As a matter of fact, the 3-ketoacyl-CoA thiolase-2 (kat2) mutation in Arabidopsis resulted in loss of activity and decrease in lipids up to 30% in the seeds revealing the importance of β-oxidation during seed development (Germain et al., 2001). Such loss of lipids in embryos is also evident in other oilseed species such as castor bean (Hutton and Stumpf, 1969), cotton (Miernyk and Trelease, 1981), Arabidopsis (Baud et al., 2002) and rapeseed (Chia et al., 2005). Hence, degradation of stored oil is likely to be a normal maturation process in flax seed development. Moreover, key enzymes involved in glyoxylate pathway, isocitrate lyase (ICL) and malate synthase (MS) were also identified at later stage of seed filling and maturation also supported by their transcript expression pattern (**Figure 2.4 M-N**). Studies conducted in developing rapeseed proposed that enhanced activity of enzymes involved in β-oxidation and glyoxylate pathway leads to accumulation of specific fatty acids (Chia et al., 2005). This could be the possible reason for high accumulation of ALA in flax and the present study supports this observation.

2.3.12 Flax seed development demands enhanced methionine metabolism

Methionine (Met) is one of the important amino acids and a fundamental metabolite synthesized by plants. The proteomic analyses of Arabidopsis seed germination and priming (Gallardo et al., 2001), developing Medicago and Chinese fir seeds (Gallardo et al., 2003; Gallardo et al., 2007; Shi et al., 2010) suggested that Met metabolism has an important role in seed development and germination. In plants, Met can be synthesized through two pathways, *de novo* and S-methylmethionine (SMM) cycle. In the present study, two important enzymes involved in the *de novo* biosynthetic pathway of Met, cystathionine γ-synthase and methionine synthase were detected. The

expression pattern of methionine synthase was consistent with the high demand for protein synthesis during embryogenesis (**Figure 2.4 R**). Similarly, the enzyme involved in the Met recycling pathway, AdoMet:Met*S*-methyltransferase, (Hanson and Roje, 2001) was also detected. Additionally, two more enzymes, AdoMet synthetase and S-Adenosyl-L-homocysteine (AdoHcy) hydrolase, which constitute the SMM cycle and consume about half of the AdoMet produced (Ranocha et al., 2001) were also identified. These enzymes help in the short term control of AdoMet level, which is the regulator of methionine biosynthesis.

Among them, the AdoMet synthetase is a key enzyme in plant metabolism catalyzing the biosynthesis of AdoMet, which is one of the key regulators of seed metabolism and has an important influence on cell growth and development. It is the primary methyl-group donor for methylation of varied compounds and involved in biosynthesis of ethylene, biotin and polyamine (Ravanel et al., 1998). In the present study as well as in Medicago, a conspicuous decrease in the level of AdoMet synthetase was observed during seed development, indicating a switch from a period of highly active metabolism to a quiescence period (**Figure 2.4 S**). On the contrary, during germination in Arabidopsis, AdoMet synthetase accumulated in the transition from a quiescent to a highly active state (Gallardo et al., 2001).

AdoHcy hydrolase catalyzes the production of AdoHcy, which is a potent competitive inhibitor of methyltransferases crucial for cell growth and development (Turner et al., 2000) and is an important intermediate in maintaining proper AdoMet levels. AdoHcy hydrolase was active during flax seed development, supporting the hypothesis that methionine recycling via the AdoMet/AdoHcy and the SMM cycle is limiting in mature seeds and is unable to maintain appropriate methionine pool for germination and seedling establishment, as also observed in Medicago (Gallardo et al., 2002).

Besides the enzymes involved in the Met biosynthesis and SMM cycle, other enzymes involved in Met metabolism were also detected. Two isoforms of cysteine synthase, catalyzing the production of cysteine, which serves as the sulphur donor for Met from O-acetyl-serine and hydrogen sulphide, were detected. ACC oxidase, an enzyme involved in the synthesis of the plant ripening hormone, ethylene was also identified. During embryo development in rapeseed, this enzyme controls cotyledon expansion (Hays et al., 2000). Based on the above results, a functional model of methionine metabolism (**Figure 2.6**) has been proposed in the present study, which suggests that the metabolism of Met is critical for storage protein synthesis in flax.

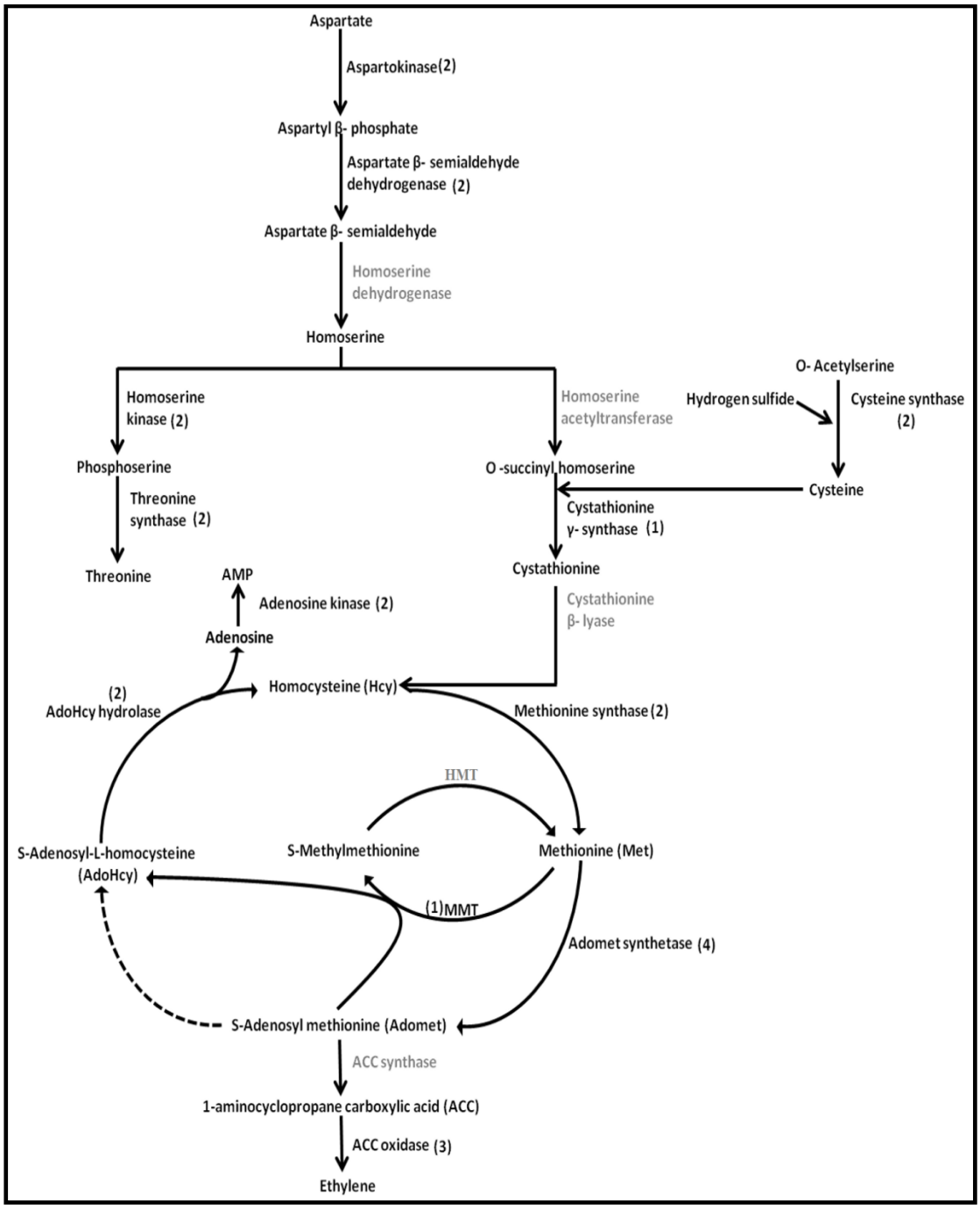


Figure 2.6 Schematic view of the methionine metabolic pathway operating during flax seed development. Values in the parentheses represent the number of isoforms identified for that protein. Enzymes shown in gray color were not identified in the present study. Abbreviations for enzymes and metabolites: MMT: Ado-Met: methionine S-methyltransferase; HMT: SMM: homocysteine S-methyltransferase; AMP: adenosine monophosphate.

2.3.13 Accumulation of health promoting compounds during flax seed filling

Metabolic pathways involved in the synthesis of health promoting compounds were the targets of this study, as flax seed is a rich source of such compounds *viz*. phenolics, lignan and flavonoids (Hall et al., 2006). These compounds also function as antioxidants and protect flax lipids against oxidation (Kitts et al., 1999).

I. Phenolic compounds

Phenolic compounds are plant secondary metabolites that constitute one of the most common and widespread groups of compounds in plants. These compounds are essential for pigmentation, growth, reproduction, resistance to pathogens etc. Phenolic acids such as coumaric (important for the synthesis of flavonoids and monolignol), ferulic and synapic are found in flax seed (Johnsson et al., 2002). Three enzymes *viz*. phenylalanine ammonia-lyase (PAL), cinnamic acid 4-hydroxylase (C4H), and hydroxycinnamate: coenzyme A ligase (4CL) catalyzing initial steps in general phenylpropanoid metabolism are necessary for the biosynthesis of phenolic compounds (Strack, 1997), among which, PAL plays a pivotal role. All these enzymes were detected in the present study. Besides these, cinnamoyl-CoA reductase (CCR), cinnamyl alcohol dehydrogenase (CAD) and o-methyltransferase (OMT) were also detected in the present study (**Figure 2.7**).

II. Lignans

Lignans are dimers of monolignol alcohol. Flax seed is the richest source of secoisolariciresinol diglucoside (SDG), providing 75 to 800 times higher levels than other plant sources (Eliasson et al., 2003). SDG is a phytoestrogen with antioxidant and anticancer activities. The primary pathway of lignan biosynthesis involves a dirigent protein and oxidases mediated coupling of coniferyl alcohol to yield pinoresinol (Davin and Lewis, 2000). Its sequential reduction by pinoresinol lariciresinol reductase (PLR) results in the formation of secoisolariciresinol followed by glucosylation by an UDP glucosyltransferase to yield SDG (Ford et al., 2001). Analysis of our flax seed proteome data identified three isoforms of PLR and nine UDP glycosyltransferases that are predominantly identified during seed filling stage (**Figure 2.7 and Table S 2.2**).

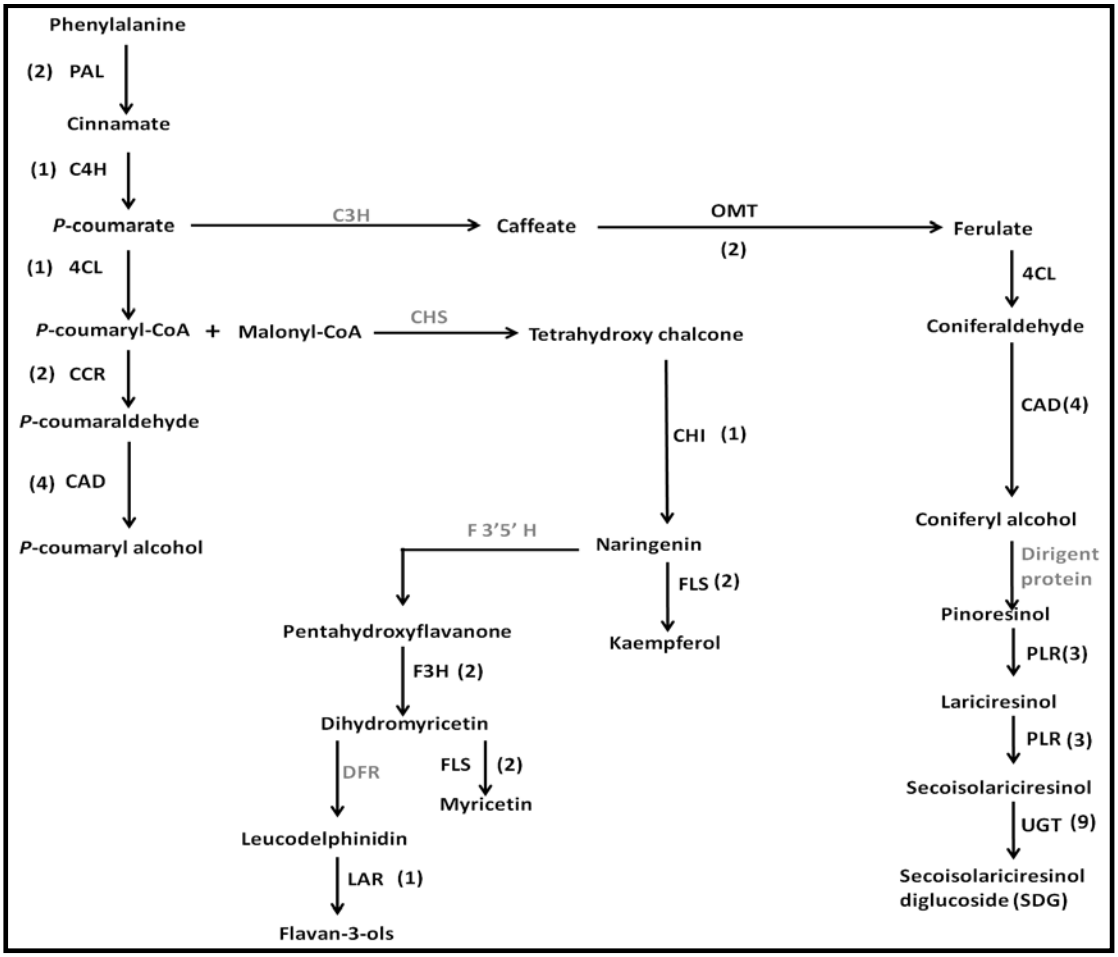


Figure 2.7 Schematic view of secondary metabolite biosynthesis during flax seed development: Values in the parentheses represent the number of isoforms identified for that protein. Enzymes shown in grey color were not identified in the present study. Abbreviations: PAL: phenylalanine ammonia lyase; C4H: Cinnamic acid 4-hydroxylase; 4CL: 4-coumarate:CoA ligase; CCR: cinnamoyl-CoA reductase; CAD: cinnamyl alcohol dehydrogenase; C3H: OMT, o-methyltransferase; CHS: Chalcone synthase; F3′5′H: flavonoid 3′5′ hydroxylase; CHI: chalcone isomerase; FLS: flavonol synthase; F3H: flavanone-3-hydroxylase; PLR: pinoresinol-lariciresinol reductase; UGT: UDP-glycosyltransferase; DFR: dihydroflavonol 4-reductase; LAR: Leucoanthocyanidin reductase.

III. Flavonoids

Flavonoids are ubiquitous in the plant kingdom and exert multiple biological effects including antioxidant activity. Hudson and Lewis (1983) reported that flavonoids have the potential to function as antioxidants in food lipid systems. Besides, they have beneficial effects on human health and as a result, increase commercial value of food

products. Flax seeds are a rich source of flavonoids such as flavonols and anthocyanidins (Toure and Xu, 2010). Chalcone isomerase (CHI) is one of the key enzymes involved in the synthesis of colorless flavanone naringenin, an important intermediate substrate for flavonoid synthesis. Flavanone 3 hydroxylase (F3H) converts flavanone naringenin into 3-OH-flavanones (dihydroflavonols), which are further converted into flavonols by flavonol synthase (FLS). Leucoanthocyanidin reductase (LAR) is involved in the synthesis of flavon-3-ols (e.g. catechin). All these enzymes were identified in the present study mainly during the embryogenesis and seed filling stages (**Figure 2.7**). However, their expression levels were low, which is also supported by the low transcript abundance of these genes in the flax seed EST library (Venglat et al., 2011).

2.3.14 Simultaneous biosynthesis and degradation of cyanogenic glycosides occurs in developing flax seed

Cyanogenic glycosides (CGs) are the biggest and extensively studied class of plant secondary metabolites found in more than 2,650 plant species and play important roles in primary metabolism especially in plant insect interactions (Zagrobelny et al., 2004; Ganjewala et al., 2010). After intake by humans, cyanogenic glycosides are degraded in intestines by the intestinal microorganisms' β-glycosidase enzymes, which produce thiocyanates; causing iodine-deficiency disorders such as goiter and cretinism (Delange et al., 1994). Flax seed contains a considerable amount of CGs such as linamarin, linustatin, lotaustralin and neolinustatin, (Niedzwiedz-Siegien, 1998) which limits the use of flax seed meal as food. Therefore, it is essential to understand the accumulation pattern of such toxic CGs during flax seed development. CGs biosynthetic pathway can be unanimously divided in three steps, wherein the first two steps are catalyzed by the enzymes of cytochrome-P450 family that produce cyanohydrins, which are glycosylated by UDP-glucosyltransferase.

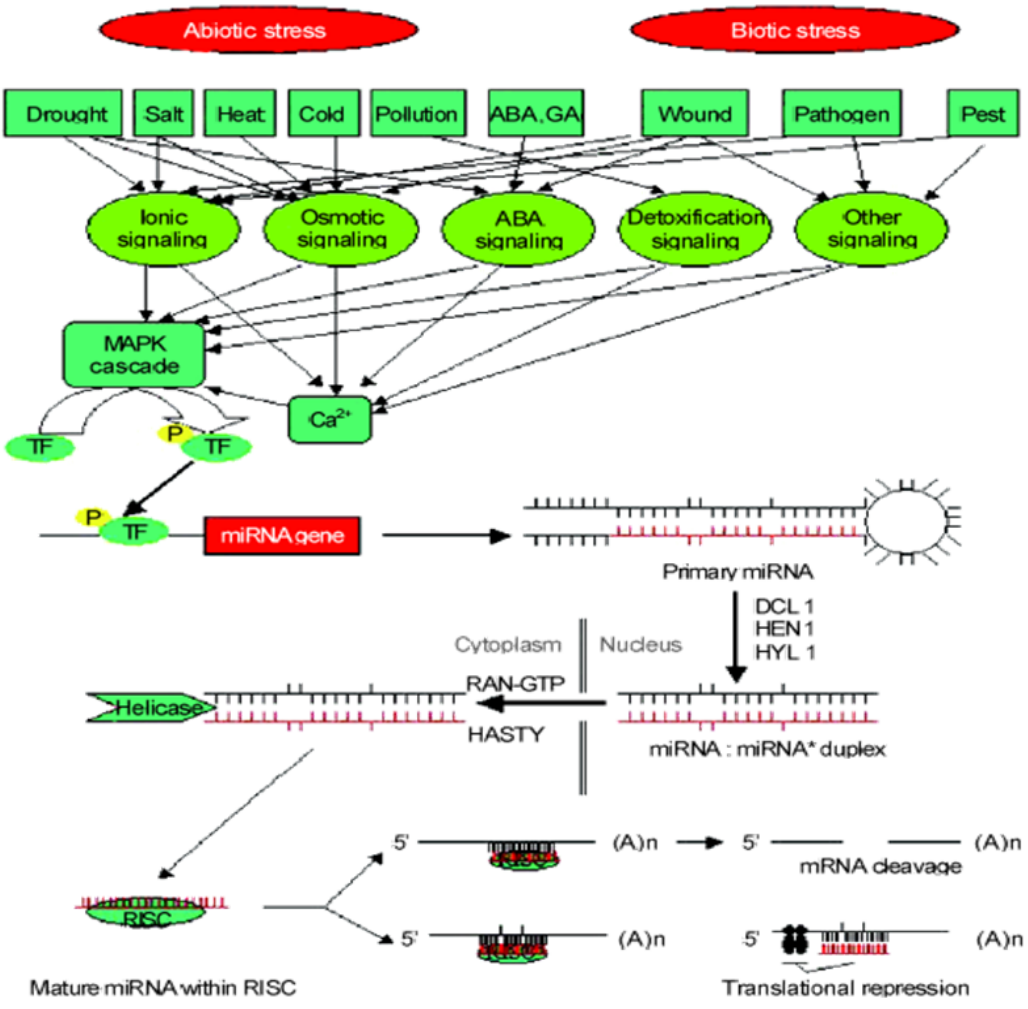
In the present study, three isoforms of cytochrome-P450 or valine N-monooxygenase (g28397, g29398 and g29400) were detected during embryogenesis, which were homologues to CYP79D1 and CYP71E1, respectively. The CYP79D1 catalyzes the first two committed steps in the biosynthesis of linamarin in cassava (Andersen et al., 2000), whereas the CYP71E7 exhibits broader substrate specificity and catalyzes the conversion of Z-p-hydroxyphenylacetaldoxime into cyanohydrin, p-hydroxy-mandelonitrile (Bak et al., 1998). Niedzwiedz-Siegien (1998) and Frehner et

al. (1990) reported that monoglucosides (linamarin and lotaustralin) and diglucosides (linustatin and neolinustatin) appeared in developing flax embryos soon after anthesis; however, mature seeds accumulated only diglucosides. In the present study, the enzymes responsible for the monoglucosides synthesis were identified during embryogenesis. Expression of valine N-monooxygenase at 4DAA confirms the earlier reports (**Figure 2.4 O**).

CGs undergo catabolic processes eventually leading to hydrogen cyanide and the enzymes, β-glucosidases (linamarase) and R-hydroxynitrile lyases are involved in this degradation. Both these enzymes were detected at the late seed filling and maturation stages. The expression pattern of β-glucosidases showed similar pattern and confirms the proteomic findings (**Figure 2.4 P**). Another catabolic enzyme, β-cyanoalanine synthase was also detected and expressed during the embryogenesis and seed filling stages of flax seed (**Figure 2.4 Q**).

In summary, an in-depth analysis of the proteome of developing flax seed is expected to provide significant information with respect to the regulation and accumulation of important storage compounds. A total of 1716 proteins were identified and their functional annotation revealed that a majority of them were involved in primary metabolism, protein destination, storage and energy. Three carbon assimilatory pathways appeared to operate in flax seeds. Reverse transcription quantitative PCR of selected 19 genes was carried out to understand their roles during seed development. Besides storage proteins, methionine synthase, RuBisCO and S-adenosylmethionine synthetase were highly expressed transcripts, highlighting their importance in flax seed development. Further, the identified proteins were mapped onto developmental seed specific expressed sequence tag (EST) libraries of flax to obtain transcriptional evidence and 81% of them had detectable expression at the mRNA level. This study provides new insights into the complex seed developmental processes operating in flax.

CHAPTER 3
Genome-wide identification and characterization of microRNA genes and their targets in flax and their role in oil accumulation



A part of this work has been published in *Planta* (2013) DOI 10.1007/s00425-012-1833-5

3.1 Introduction

MicroRNAs (miRNAs) are endogenous 20-24 nucleotide (nt) long non-coding RNAs derived from single-stranded RNA precursors that can form stem-loop structures. They regulate gene expression at the post-transcriptional level by repressing gene translation or targeted degradation of specific mRNAs, depending on the extent of complementarity between the miRNA and its target mRNAs (Bentwich, 2005). MiRNAs were first identified in *Caenorhabditis elegans* and subsequently a large number of miRNAs have been identified and characterized experimentally in both, animals and plants (Zhang et al., 2006; Meyers et al., 2008). The plant miRNAs have been shown to play important functions in diverse processes associated with plant growth and development, including developmental transitions, organ polarity, auxin signaling, boundary formation or organ separation, leaf growth, stem growth, floral organ identity and reproductive development (Sun, 2012).

Four approaches have been reported for discovering the miRNAs: i) genetic screening (Lee et al., 1993; Wightman et al., 1993), ii) isolation and sequencing of small RNAs (Lu et al., 2007), iii) computational prediction from genomic sequences (Mathews et al., 1999), and iv) analysis of expressed sequence tags (ESTs) (Song et al., 2010). In the first two approaches, miRNAs are discovered experimentally; while the latter two approaches are prediction based and require validation using experimental approaches such as cloning and expression analysis. Till date, the bioinformatics approaches have been widely applied and successful in the discovery and identification of conserved miRNA in various plant species such as cotton (Ruan et al., 2009), maize (Zhang et al., 2009), caster (Zeng et al., 2010), sorghum (Calvino et al., 2011), soybean (Turner et al., 2012), etc. So far, more than 5000 miRNA genes have been identified and annotated from 61 plants and listed in miRBase (ver. 19), the microRNA database (Griffiths-Jones et al., 2008). Moreover, the availability of complete genome sequence for a number of these miRNA encoding genes from diverse plant species has allowed a better understanding of miRNA evolution in plants (Zhu et al., 2011). For the bioinformatic based analysis and discovery, three major characteristics are primarily used: i) length of 20–22 nucleotides in case of animals and 20–24 nucleotides in case of plants (Ambros et al., 2003; Bartel, 2004); ii) presence of a well-defined stem-loop hairpin structure with low free energy (− 32 to − 57 kcal/mol) (Bonnet et al., 2004); and iii) their evolutionarily conserved nature, from

worm to human (Altuvia et al., 2005; Weber, 2005), or from ferns to higher plants (Zhang et al., 2006). Interestingly, mature miRNAs, rather than their miRNA precursors, are more conserved in plants; whereas the latter are more conserved in animals (Bartel, 2004).

Moss and Cullis (2012) identified 12 conserved and 649 potentially novel miRNA genes by analyzing the draft whole genome sequence and the ESTs of flax. Similarly, computational analysis of flax ESTs alone resulted in the identification of 20 conserved miRNAs (Neutelings et al., 2012). However, in both these studies, only a few conserved miRNA genes were identified compared to more than 100 in all the other plant species analyzed to date. The present study was undertaken with the following objectives: (i) to identify conserved miRNAs from the flax draft genome sequence (http://www.phytozome.net/flax ver. 1.0) using bioinformatic approaches, (ii) analyzing their genomic organization and genetic structure, (iii) confirming the expression of the identified miRNAs, and (iv) identifying their protein coding target genes. Further, role of miRNA genes and their target transcript has been proposed to understand miRNA mediated regulation of oil biosynthesis and accumulation in flax.

3.2 Materials and Methods

3.2.1 Prediction of miRNAs from flax genome sequence

The draft whole genome sequence of flax was accessed (downloaded) from the phytozome database (ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v8.0/Lusitatissimum/assembly/) (Wang et al., 2012). A total of 3430 mature miRNA sequences from 34 flowering plants (Magnoliophyta) were downloaded from miRBase (http://www.mirbase.org, August 2011) (Griffiths-Jones et al., 2008) and analyzed using the CD-HIT-EST (Huang et al., 2010) web server (http://weizhong-lab.ucsd.edu/_cdhit_suite/cgi-bin/index.cgi?cmd=cd-hit-est) with identity cut-off of 0.9, minimum and maximum alignment coverage as zero and unlimited respectively, to identify 1548 non-redundant miRNA sequences. These non-redundant mature plant miRNA sequences were queried against the whole genome sequence of flax using the PatMaN (Pattern Matching in Nucleotide databases) tool (Prufer et al., 2008) to identify homologous miRNA sequences from flax. PatMaN searches for short patterns in large DNA databases, allowing for approximate matches. The nucleotide sequence (250 bp) flanking the exact match of miRNA was extracted from the genome in two

ways; one sequence extending 180 nucleotides upstream and 70 nucleotides downstream from the hit, and the other covering 70 nucleotides upstream and 180 nucleotides downstream of the hit. Each sequence window was folded using RNAfold (Hofacker, 2003) and further trimmed and analyzed for resulting secondary structure to identify the possible miRNA hairpin. Additional checks were also performed such as: i) presence of no more than three consecutive mismatches between miRNA and miRNA*, ii) base-pairing of at least 17 of the 25 nucleotides centered around the miRNA, iii) hairpin length of at least 75 nucleotides, and iv) pairing of at least 50% bases in the hairpin. The most stable and valid hairpin from each sequence window was then chosen as the precursor miRNA candidate and tested using Randfold (Bonnet et al., 2004) with a cutoff of 0.1, which suggests that the miRNA precursors have lower folding energies than any random sequence or non-coding RNAs. The above-mentioned analysis was performed using the miRCat pipeline (Moxon et al., 2008). Also, minimal folding free energy index (MFEI) designate miRNA characteristics was calculated as suggested by Zhang et al. (2006).

3.2.2 Characterization of miRNA genes

The sequences of the identified miRNAs were submitted to Dr. Sam Griffiths-Jones, who manages the miRBase database (Griffiths-Jones et al., 2008). He performed phylogenetic grouping of the miRNAs and assigned universal names to them. Nucleotide sequence of 1 kb was extracted from the upstream region of the precursor miRNA (pre-miRNA) from the flax whole genome sequence. This sequence was used for prediction of transcription start sites (TSS) using the TSSP (http://linux1.softberry.com/berry.phtml?topic=tssp&group=programs&subgroup=promoter) program from the softberry package (Solovyev and Salamov, 1997). Transcription end sites marked by polyadenylation signal were predicted using POLYAH (http://linux1.softberry.com/berry.phtml?topic=polyah&group=help&subgroup=promoter). The upstream 2 kb nucleotide sequences of TSS were extracted and analyzed for the presence of promoters using the PLACE web server (http://www.dna.affrc.go.jp/PLACE/) (Higo et al., 1999). To identify intergenic and intragenic location of miRNA genes, the genomic coordinates of pre-miRNAs were employed using flax genome browser at phytozome web site (ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v8.0/Lusitatissimum/assembly/) and the presence of overlapping transcripts at that genomic region were searched.

3.2.3 Digital expression analysis

To obtain transcriptional evidence for the miRNA genes, publicly available flax EST sequences were downloaded. This includes the ESTs generated by Venglat et al. (2011), Fenart et al. (2010) and those available at NCBI (http://www.ncbi.nlm.nih.gov/dbEST/). The EST sequences were assembled using EGassembler (Masoudi-Nejad et al., 2006) web server (http://egassembler.hgc.jp/) and a non-redundant EST set was developed using the CD-HIT-EST (Huang et al., 2010) web server (http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi?cmd=cd-hit-est). Makeblastdb (ftp://ftp.ncbi. nlm.nih.gov/blast/) was used to create a flax EST database on an in-house server. Viroblast (v. 2.2+) (Deng et al., 2007) was configured to use the stand-alone suite of Blast programs (v. 2.2.24+). The pre-miRNA sequences were BlastN searched against the assembled EST database with an E-value threshold of $\leq e^{-25}$ and ≥88% sequence identity criteria to map the ESTs onto pre-miRNA sequences.

3.2.4 Computational prediction of miRNA protein coding target transcripts

The Plant Small RNA Target Analysis Server (psRNATarget; http://plantgrn.noble.org/psRNATarget/) was used to identify the miRNA targets. psRNATarget provides reverse complementary matching between a miRNA and its target transcript and determines target site accessibility by calculating unpaired energy (UPE) necessary for opening the secondary structure around the miRNAs target site, while distinguishing between translational and post-transcriptional inhibition (Dai and Zhao, 2011). The assembled EST data were used for prediction of protein coding target transcripts of the miRNAs. The "user-submitted small RNAs / user-submitted transcripts" option from psRNATarget website (http://plantgrn.noble.org/psRNATarget/?function=3) was used with default parameters. Identity of the target ESTs revealed by psRNATarget was confirmed by performing BlastX search against the NCBI protein database. Gene Ontology (GO) annotation of the identified targets was performed using GOanna tool from AgBase web server (http://www.agbase.msstate.edu) with plant specific database and expectation value of $\leq e^{-20}$ (McCarthy et al., 2006). The GO terms were obtained using Plant GOSlim and the GO terms enrichment analysis was performed with the agriGO web server [http://bioinfo.cau.edu.cn/agriGO/analysis.php (Du et al., 2010)] using

singular enrichment analysis (SEA) tool and the suggested background of Arabidopsis gene models (TAIR9). Chi-square test with Bonferroni multi-test adjustments was used at 0.05 significance level with a minimum of five mapping entries for the enrichment analysis.

3.2.5 RT-qPCR analysis of miRNA and target genes

I. Plant material

For validation of expression of the miRNAs and their target genes, the CDC Bethune flax genotype was used because the whole genome sequence of this genotype is determined and available. The seeds were germinated and grown in a growth chamber using a daily cycle of 16 h of light (23 °C) and 8 h of dark (16 °C). Tissue samples were collected and frozen immediately in liquid nitrogen. The leaf (LE) samples were collected from approximately two week old plants. The stem (ST) was harvested from one month old plants. The flower (FL) samples were a mixture of various stages. The anthers (AN) and ovaries (OV) were collected from unopened flowers. The seed (SD) samples represent mixture of various stages of seed development. The roots (RT) were collected from one month old plants and washed with water to remove the soil. Etiolated seedling (ES) samples were derived from seeds germinated on MS medium plates in the dark for four days followed by removal of the seed coats prior to harvesting.

II. Isolation of total RNA and RT-qPCR analysis

Total RNA was isolated from eight tissue types as outlined above using the RNeasy Plant Mini Kit (Qiagen, Cat. No. 74904). On-column DNase digestion was performed using the RNase-free DNase (Qiagen, Cat. No. 79254). Reverse transcription was performed using the QuantiTect Reverse Transcription Kit (Qiagen, Cat. No. 205311). The reactions were performed following the supplier's recommended conditions. For cDNA synthesis, 1µg of total RNA template was incubated at 42 °C for 45 min. After reverse transcription, 100 µl water was added to each sample, and 2 µl of the sample was used to perform 20 µl RT-qPCR. The hairpin specific primers were designed for 14 miRNA genes and also nine miRNA target specific primers were designed (**Table S 3.1A and 3.1B**). The Power SYBR Green PCR Master Mix (Applied Biosystems, USA) was used in the reaction following the suppliers' recommendations. The flax elongation factor 1-alpha (*EF1α*; Lus10023174) gene was used as a housekeeping or

reference gene for all the RT-qPCRs. The housekeeping gene (*EF1α*) was selected as reference internal control after confirming the consistent expression of this gene across all the tissue types used in the study. Three biological replicates were performed using "Step-one" Real-time PCR System (Applied Biosystems, USA).

3.3 Results

3.3.1 Identification of conserved miRNAs in the flax genome

A computational prediction method was employed to identify conserved miRNAs in the flax genome. As mature miRNAs have shown to be more conserved in plants rather than their precursor sequences (Bartel, 2004), mature miRNA sequences were used as queries for searching the flax draft genome sequence. Initially, mature miRNA sequences from 34 flowering plants (3430) were downloaded from miRBase and redundant sequences were removed using CD-HIT-EST (Huang et al., 2010). This set of 1548 non-redundant mature miRNA sequences was used to identify orthologous miRNA sequences from flax using the miRCat pipeline; this analysis initially identified 137 flax miRNA encoding genes. However, 21 of these miRNA sequences were eliminated, as 19 were redundant and for two miRNAs, the mature sequence was located in opposite arm of hairpin than the homologous miRNA hairpin from miRBase. The remaining 116 flax miRNAs were used for further analysis. These 116 miRNAs were also submitted to the miRBase database and they were assigned universal names by the miRBase administrator, Dr. Sam Griffiths-Jones. Detailed information about each pre-miRNA, as well as the mature miRNA, is provided in **Table S 3.2A and 3.2B**. The stem-loop hairpin structures of pre-miRNA sequences are provided in **Figure S 3.1**.

The 116 flax miRNAs identified in this study represent highly conserved plant miRNAs and they were classified into 23 phylogenetic families. The number of individual representative members per family varied with maximum in miR169 (12) and minimum in miR408 and miR828 (one each). The phylogenetic distribution of these miRNA family genes was compared in four top ranked flax related plant species (Rco: *Ricinus communis*, Ptc: *Populus trichocarpa*, Vvi: *Vitis vinifera* and Ath: *Arabidopsis thaliana*) and the distribution was similar in all these species (**Figure 3.1**).

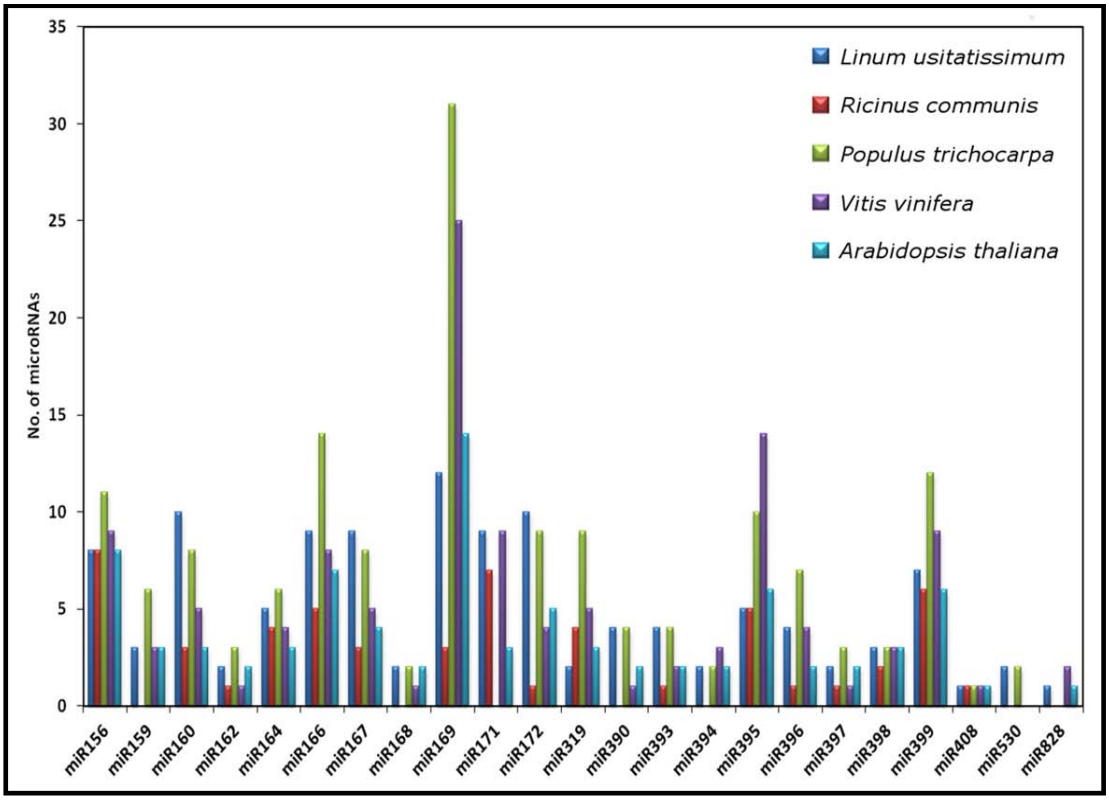


Figure 3.1 Phylogenetic distribution of the miRNA genes reported from five plant species. Lus: *Linum usitatissimum*, Rco: *Ricinus communis*, Ptc: *Populus trichocarpa*, Vvi: *Vitis vinifera* and Ath: *Arabidopsis thaliana*.

The identified flax miRNA sequences had high negative minimal fold energies (MFEs) with an average of -62.41 (-103.00 to -37.70, **Table S 3.2A**). The minimal folding free energy index (MFEI) was calculated for all the 116 flax miRNA sequences, to differentiate them from other non-coding RNAs like tRNA, rRNA and mRNA. The average MFEI for the miRNAs was -0.94, much lower than the recommended value of -0.85 (Zhang et al., 2006). Most of the conserved families had 21 nt long mature miRNAs, except for the families of miR394, miR398 and miR530, which had 20 nt miRNA. The average A+U content of the miRNA genes was 54% and about 73% of the mature miRNAs had U at their 5' end.

In the stem-loop hairpin pre-miRNA sequences, the mature miRNAs were located at 5'arm in 13 families and 3'arm in10 families (**Table S 3.2A**). The location of the mature miRNAs in the stem-loop hairpin was conserved according to families, with few exceptions [lus-MIR169b (3'arm), lus-MIR171d (5'arm), lus-MIR172j (5'arm), lus-MIR395d (5'arm) and lus-MIR398b (5'arm)].

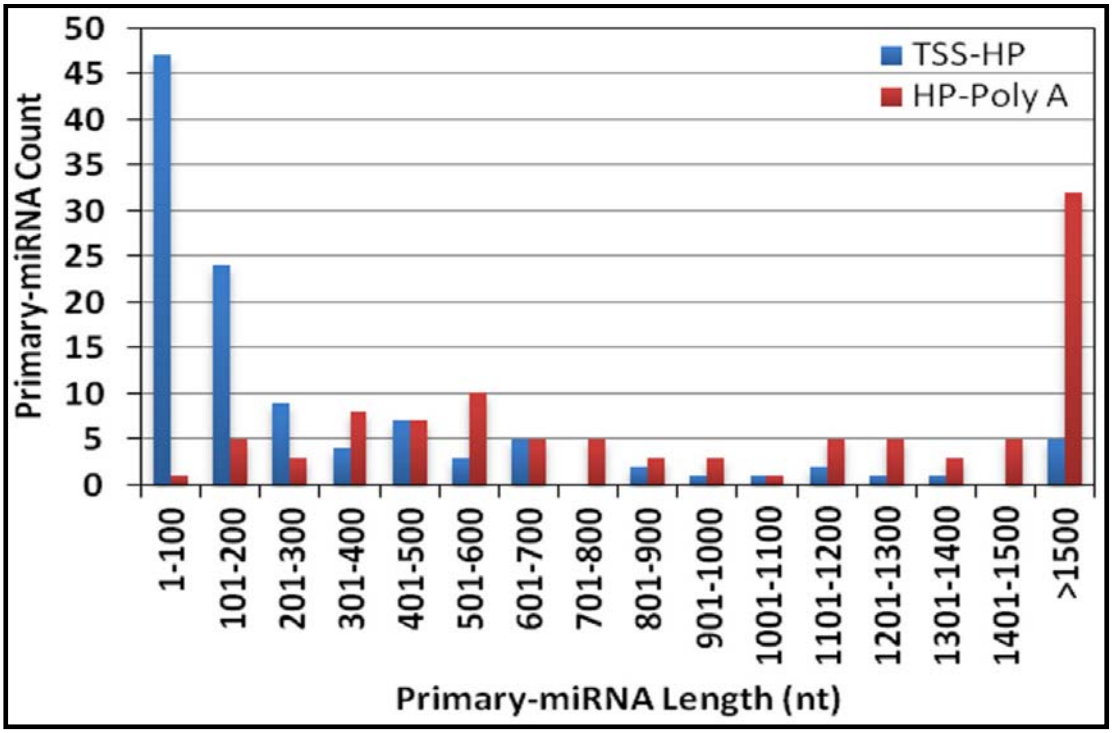


Figure 3.2 Length distribution of TSS-HP (blue bars) and HP-PolyA (red bars) regions of pri-mRNA. The TSS-HP region is defined as the segment from the transcription start site (TSS) to the hairpin (HP) structure; while the HP-PolyA region is the segment between the hairpin and the transcription stop site.

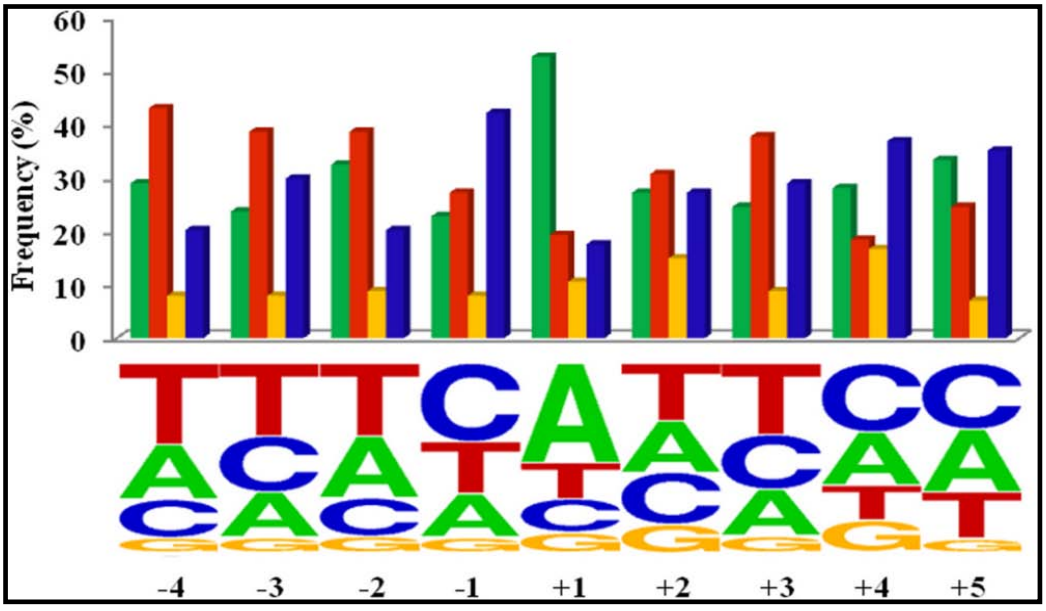


Figure 3.3 Sequence logo of nucleotide frequencies surrounding the transcriptional start sites (n=112). TSSs are predominately marked by an adenine at the +1 position and a cytosine at -1.

3.3.2 Characterization of miRNA genes

The miRNA sequences were characterized for the presence of characteristic features of miRNA transcripts, such as transcription start site (TSS), TATA box, promoter and polyadenylation site. The overall lengths for pre-miRNA transcripts ranged from 90-247 nt, with an average of 147 nt. The pri-miRNA length distribution (**Figure 3.2**) shows that the 3' region (measured from the hairpin to polyadenylation site) was longer than the 5' region (measured from the transcriptional start site (TSS) to the hairpin) for majority of the miRNAs genes. Out of 116 miRNA genes, the TSS site was identified for 112 genes and was similar to that of Arabidopsis (Xie et al., 2005) (**Figure 3.3**). To identify conserved motifs flanking the TSS at each mapped locus, a 60 bp genomic segment (50 bp upstream and 10 bp downstream of the locus) was computationally analyzed using the MEME webserver (http://meme.ebi.edu.au/meme/cgi-bin/meme.cgi). An 8 nt TATA box-like sequence was found in 99 (85.34%) of the miRNA genes as a conserved motif in this region (**Table S 3.3**).

To determine the putative promoter motifs, 2 kb sequence upstream to TSS of 112 miRNA genes were analyzed using the PLACE web server (http://www.dna.affrc.go.jp/PLACE/) (Higo et al., 1999). The sequences of four miRNAs could not be used for the promoter analysis because three of them lacked the TSS (lus-MIR319b, lus-MIR396b and lus-MIR828a), while one was polycistronic (lus-MIR169c and lus-MIR169d) and hence shared the same upstream region. A total of 275 promoter motifs were identified in the 112 miRNA genes (**Table S 3.4A**). The motif CACTFTPPCA1 (YACT) was the most abundant and occurred 3106 times among the 112 sequences. Likewise, as many as 796 motifs were present in the lus-MIR397b sequence. Secondly, promoter sequences occurring preferentially in plant miRNA genes were also searched (Megraw et al., 2006) and found that MYB binding domain was the most abundant among the 17 motifs (**Table S 3.4B**). At present it is difficult to predict the role of MYB in flax miRNA genes. It is quite possible that MYB is also most abundant in protein coding genes of flax.

The organization and position of these miRNA genes in the flax genome was also examined (www.phytozome.net/flax). Physical position of the miRNA genes with respect to protein coding genes was determined. Only two miRNAs were intragenic; lus-MIR172c was within the coding region of Lus10016514, while lus-

MIR530b was in the 5'UTR of Lus10016134. The miRNA genes were distributed in 85 scaffolds and four gene clusters were identified with gene number ranging from 2-4 per cluster (**Table 3.1**).

3.3.3 Expression analysis

I. Digital expression analysis of pre-miRNA

To obtain expression evidence for the 116 miRNAs, publically available flax ESTs were analyzed [NCBI and genolin (Fenart et al., 2010) (http://urgi.versailles.inra.fr/Species/Flax/Download-sequences)]. The assembly of the available ESTs resulted in 31,929 contigs and 19,400 singletons. When the 116 pre-miRNAs were BlastN searched against these assembled ESTs, 22 miRNA genes showed expression evidence (**Table S 3.5**) with majority of them in the developing seed tissues and stem.

II. RT-qPCR of pre-miRNA

For expression analysis using RT-qPCR, 22 miRNAs were initially selected for which the EST expression evidence was available. However, only 14 of these 22 miRNA genes produced specific amplifications in eight different tissue types. Hence, they were further selected for quantitative expression analysis. The results revealed distinct tissue specific expression profiles for these miRNAs. Seven miRNAs (lus-MIR-; 162a, 162b, 166b, 166d, 167a, 168b and 408a) were highly expressed in anthers; whereas lus-MIR172c and lus-MIR172i were highly expressed in etiolated seedlings and stem tissues, respectively. The miRNA lus-MIR156a showed the highest and similar expression in etiolated seedling and leaf. On the contrary, lus-MIR-; 159c, 167h, 172e and 396c miRNAs expressed in diverse tissue types and did not exhibit tissue specificity (**Figure 3.4**).

Table 3.1: **Details of the four gene clusters detected in the present study**

Name	Phylogenetic group	Chromosome	HP start	HP end	Orientation	Sequence	Distance in bp
lus-MIR167c	167	scaffold461	524864	524990	+	UGAAGCUGCCAGCAUGAUCUA	5112
lus-MIR167d	167	scaffold461	530102	530256	+	UGAAGCUGCCAGCAUGAUCUA	
lus-MIR169b	169	scaffold587	194841	195014	+	UAGCCAAGGAUGACUUGCCUG	2045
lus-MIR169c	169	scaffold587	197059	197242	+	UAGCCAAGGAUGACUUGCCUG	
lus-MIR169a	169	scaffold587	197291	197452	+	UAGCCAAGGAUGACUUGCCCA	49
lus-MIR399f	399	scaffold859	138722	138539	-	UGCCAAAGGAGAUUUGCCCAG	1679
lus-MIR399e	399	scaffold859	140542	140401	-	UGCCAAAGGAGAUUUGCCCGG	
lus-MIR399c	399	scaffold859	148520	148382	-	UGCCAAAGGAGAUUUGCCCGG	7840
lus-MIR399a	399	scaffold859	151882	151719	-	UGCCAAAGGAGAUUUGCCCGG	3199
lus-MIR399d	399	scaffold859	141913	142054	+	UGCCAAAGGAGAAUUGCCCUG	7400
lus-MIR399b	399	scaffold859	149454	149566	+	UGCCAAAGGAGAAUUGCCCUG	

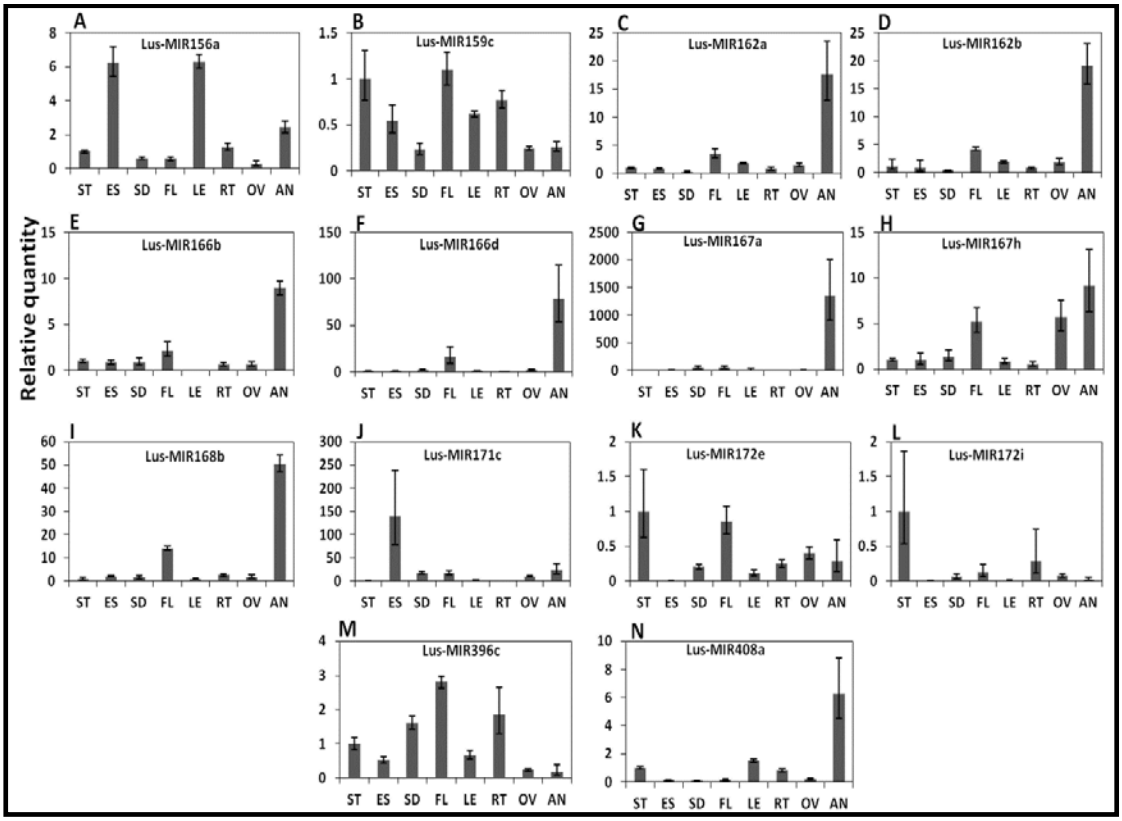


Figure 3.4 Expression of selected miRNAs in various flax tissues. Bar graph indicates the relative quantity of pre-miRNA transcript with respective expression in stem tissue. For all the expression analyses, normalization was performed using the housekeeping gene, *EF1α*. Error bars indicate maximum and minimum relative quantity in three biological replicates. Abbreviations:- ST: stem, ES: etiolated seedling, SD: pooled developing seed stages, FL: flower, LE: leaf, RT: root, OV: ovary and AN: anther.

3.3.4 miRNA target identification and confirmation using RT-qPCR

The psRNATarget web server (Dai and Zhao, 2011) was used to identify potential miRNA gene targets by searching the assembled flax EST database. Targets were identified for 105 (90.51%) miRNAs. A total of 479 (142 non-redundant) potential target transcripts were identified with almost 52% of these representing the unknown category as revealed by BlastX search against the NCBI protein database (**Table S 3.6**). The number of potential genes targeted by each miRNA family varied from 2 (lus-miR408) to 110 (lus-miR172). To understand the biological functions of these miRNA target transcripts, GO annotations were performed for 142 non-redundant target transcripts, which yielded 878 GO annotations for 89 targets, showing alignments of individual transcripts representing diverse functions, processes or

components. The targets regulated by flax miRNAs were subjected to agriGO analysis to investigate gene ontologies. The Arabidopsis gene models (TAIR9; 37,761) were used as a reference in order to retrieve the GO terms as flax gene models are not yet included in the agriGO webserver. Of the 89 targets, significant GO terms were obtained for 82 targets and these were used for further analysis. More than 40% of these genes were involved in cellular process, metabolic process, response to stimuli, biological regulation, etc. and were enriched in flax miRNA targets than in Arabidopsis genes (**Figure 3.5**). The graphical representation of GO ontology distribution in the three categories *viz*., biological processes, cellular component and molecular function is presented in **Figure S 3.2**.

To obtain further insights into the possible roles of the miRNAs, the expression of their putative target genes (with lowest unpairing energy and annotated) was analyzed using RT-qPCR. For this purpose, nine target genes were selected for profiling their expression patterns in relation to the expression profiles of 14 miRNAs whose RT-qPCR was also performed in the present study. As expected, an inverse correlation between the expression of the miRNAs and their putative target genes for seven pairs among different flax tissues was observed (lus-miR156a:squamosa promoter-binding protein, lus-miR166b:leucine zipper protein HOX32, lus-miR166d:homeodomain transcription factor, lus-miR168b:cytochrome P450, lus-miR172e:AP2 domain-containing transcription factor, lus-miR172i:AP2 domain-containing transcription factor andlus-miR396c:GREEN-RIPE LIKE 1), suggesting miRNA mediated regulation of their potential targets. For example, the potential target (squamosa promoter-binding protein) of lus-miR156a had the highest expression in the flower and ovary, while lower in etiolated seedling and leaf (**Figure 3.6**), in contrast to high expression of lus-miR156a in these tissues. On the contrary, two miRNAs (lus-miR167a and lus-miR167h) and their potential target genes (Corticosteroid 11-beta-dehydrogenase and LIM1) did not follow the miRNA mediated regulation, as all of them were highly expressed in the anthers, suggesting the possibility of spatially separated localization for miRNAs and their target genes (**Figure 3.6**).

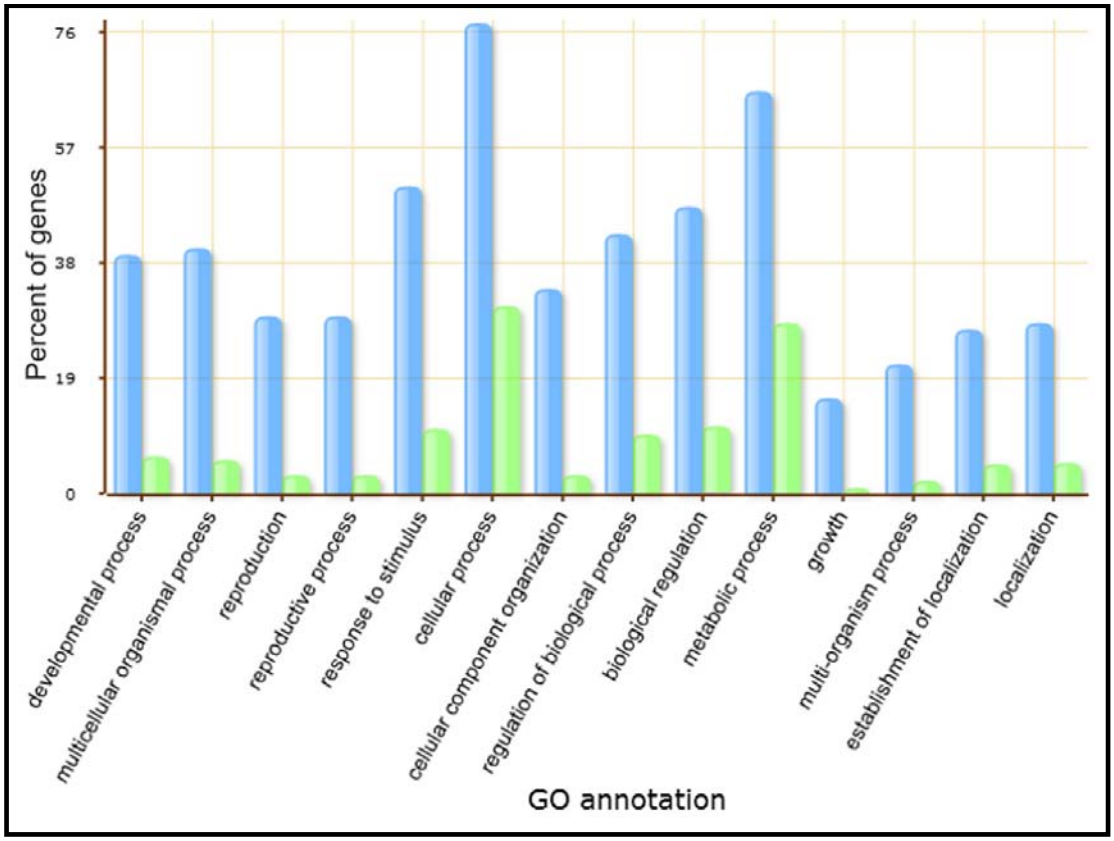


Figure 3.5 Gene ontology (GO) analysis of flax miRNA target transcripts. Blue bars indicate the enrichment of flax miRNA targets in GO terms. Green bars indicate the percentage of total annotated Arabidopsis genes mapping to GO terms.

3.4 Discussion

Identification of a comprehensive set of miRNAs in various plant species is a critical step to facilitate our in depth understanding of regulatory mechanisms operational in the respective plant species. A majority of known miRNAs in the plant kingdom are evolutionarily conserved allowing identification of new miRNAs using computational approaches in other plant species (Zhang et al., 2006). Using this approach, 116 conserved miRNAs belonging to 23 families from flax were identified. These represent significantly larger number than the recently reported miRNA genes in flax. A total of 32 conserved flax miRNAs have been reported previously (Moss and Cullis, 2012; Neutelings et al., 2012). Neutelings et al. (2012) performed miRNA identification from the flax EST dataset, while Moss and Cullis (2012) used both the genome and EST datasets for miRNA identification. They used the novoMIR algorithm with *ab initio* approach, wherein the miRNAs are identified by sequence analysis without any homology search, which identifies a large number of putative

miRNAs. In this study, miRCat pipeline was employed with homology based approach and stringent criteria were used for determining the 116 conserved miRNAs.

As neither the universal names nor the pre-miRNA sequences of the miRNAs identified by Moss and Cullis (2012) are available for comparison with the miRNAs detected in the present study, it is not possible to compare and determine if all the miRNAs detected in their study are also represented in this study. Neutelings et al. (2012) reported 20 conserved flax miRNAs, of which 14 miRNAs were identified in the present study. Some of the miRNAs detected by them might have been missed in this study due to the different softwares used for detection and the stringent criteria that were applied to detect conserved miRNAs. The results obtained in this study are comparable to the studies conducted in poplar [123 miRNAs from 32 families (Barakat et al., 2007)], maize [150 miRNAs from 26 families (Zhang et al., 2009)] and soybean [120 miRNAs from 31 families (Turner et al., 2012)].

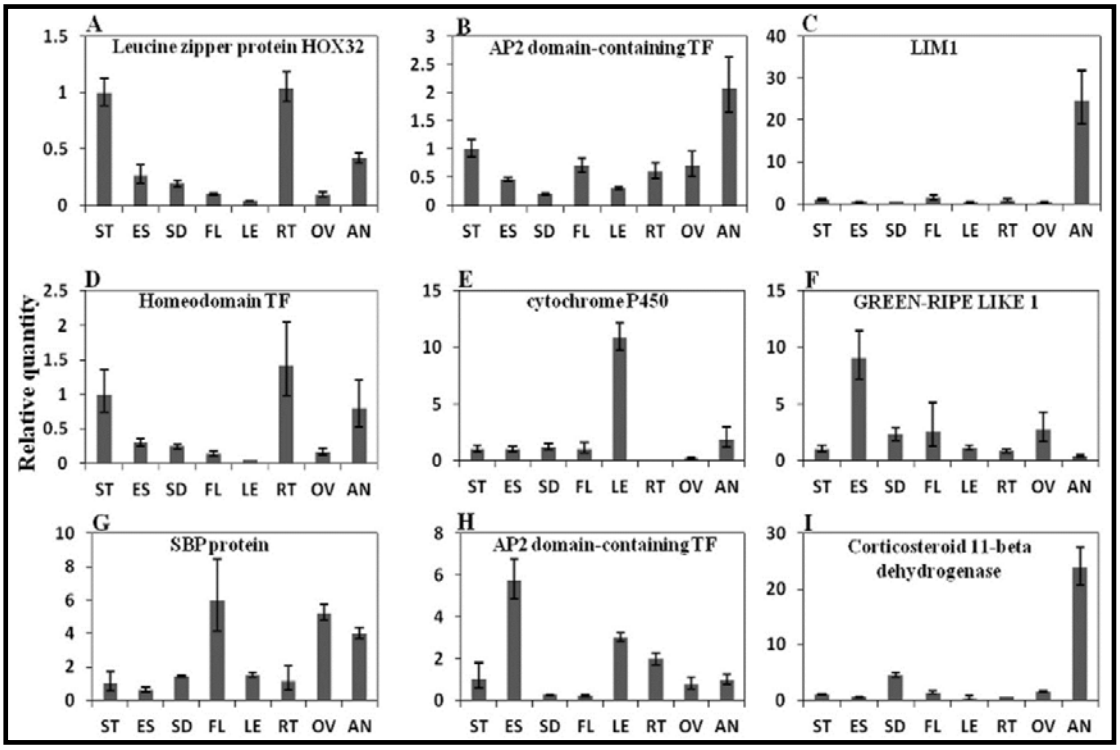


Figure 3.6 The expression of nine miRNA target transcripts in various flax tissues. Bar graph indicates the relative quantity of target transcript with respective expression in stem tissue. For all the expression analyses, normalization was done with the housekeeping gene, *EF1α*. Error bars indicate maximum and minimum relative quantity in three biological replicates. Abbreviations:- ST: stem, ES: etiolated seedling, SD: pooled developing seed stages, FL: flower, LE: leaf, RT: root, OV: ovary, AN: anther, SBP: squamosa promoter-binding protein, TF: transcription factor.

3.4.1 Flax miRNAs exhibit the typical features of plant miRNAs

A majority of the identified mature flax miRNAs (90%) were 21 nt in length, similar to those identified in maize (Zhang et al., 2006), soybean (Zhang et al., 2008), *Brachypodium distachyon* (Unver and Budak, 2009) and tobacco (Frazier et al., 2010). The flax pre-miRNAs also showed higher diversity in length (Zhang et al., 2006) ranging from 90-247 nt, with an average size of 147 nt, consistent with the length of plant pre-miRNAs (Thakur et al., 2011). More than half of the mature miRNAs (~59%) belonging to the same family were located in the 5' arm of the stem-loop structure. On average, the 5' region of pre-miRNA was shorter than the 3' region as reported in maize (Zhang et al., 2009). The minimal folding free energy index (MFEI) was significantly lower (-0.94) than the recommended MFEI value (-0.85) for plant miRNAs (Zhang et al., 2006), further supporting the accurate identification of flax miRNAs in this study. Conserved motifs flanking the TSS revealed the prevalence of adenine at TSS and cytosine at the -1 position similar to that observed in rice and Arabidopsis (Alexandrov et al., 2006). In 86% of the miRNAs, the TATA box-like motifs were found at -21 to -30 positions relative to TSS, suggesting that a majority of flax miRNA genes are transcribed by RNA polymerase II, and have the same promoters as the protein coding genes. The frequency of TATA box (86%) in flax miRNAs was similar to that observed in Arabidopsis and maize miRNA transcripts (Xie et al., 2005). The non-TATA box-containing miRNA genes might fall into the class termed TATA-less promoters usually reported for housekeeping and developmentally regulated genes (Smale, 2001) and also observed in Arabidopsis and rice (Xie et al., 2005; Cui et al., 2009). A total of 275 promoter sequences were identified, among which the MYB motif occurred in high proportion as also seen in other plant miRNA genes; suggesting that these conserved motifs may be also involved in the flax miRNA transcription.

Most of the annotated plant miRNA genes are located in the intergenic region (Rajagopalan et al., 2006; Tang, 2010). Likewise, almost all the identified flax miRNAs were intergenic. Only two miRNA genes were intragenic, lus-MIR172c was present within the exonic region of Lus10016514, while lus-MIR530b was located in the 5'UTR region of Lus10016134. It has been reported that sometimes, the genes in which the miRNAs are embedded, are also their putative targets for regulation and in such cases the expression of the miRNA and the putative target gene is tightly co-

regulated (Zhang et al., 2009). Hence, it would be interesting to further study the expression of these two pairs. However, these genes with embedded miRNAs correspond to predicted protein coding genes with no function assigned to them, as revealed by BlastN search against the NCBI database. Alternatively, it is also possible that these predicted genes could be the result of misannotation.

Several genome wide studies in plants suggest the presence of miRNA clusters (Zhou et al., 2011) and most of these clusters contain conserved miRNAs belonging to the same family. Four clusters involving three families from these 116 flax miRNA genes were identified. For searching the pre-miRNA structure, only 250 nt flanking the mature miRNA was included and hence families that have introns in their precursors might be missing in the present analysis. However, this number is more than that identified by Neutelings et al. (2012), wherein one flax pre-miRNA carrying two miRNA genes; lus-miR398a and lus-miR398b, was identified. The length of miRNA clusters varied, suggesting that they may be differentially transcribed and may perform unique functions for regulating expression of miRNAs. Further, the clustered genes also showed diversity in the amount of promoters they contain.

Two members of the miRNA169 family (miR169a and miR169c) on scaffold587 were separated by less than 50 nt, and also shared the same TSS. Such compact clustering suggests that these two genes are processed as a single polycistronic unit. Such co-location and polycistronic nature of miR169 family members was also observed in several plant species such as cotton (Zhang et al., 2007), soybean (Zhang et al., 2008), rice (Zhao et al., 2009) and tobacco (Frazier et al., 2010). The miR169 family members have been shown to play a role in environmental stress tolerance (Zhao et al., 2009). Therefore, it is possible that the identified flax miR169 genes might have similar evolutionary origins and might have similar functions.

3.4.2 Diversity of expression of the flax miRNAs

MicroRNA profiling is essential to analyze the functions of the miRNAs according to their expression patterns. In this study, miRNA expression was carried out using the RT-qPCR to validate and estimate the expression of selected mature miRNAs and the digital expression analysis using an in-house flax EST database. RT-qPCR based analysis of 14 miRNAs showed differential expression patterns, which suggests potentially different regulatory roles for these miRNAs in different tissues and

developmental stages of flax. Interestingly, of the 14 miRNAs, 8 genes (miR162-a,b; miR166-b,d; miR167-a,h; miR168b, miR408a) exhibited higher levels of expression in anther, indicating that these miRNAs might play an important role in anther development in flax and might be the potential candidates to develop genetic male-sterile lines in flax or other crops for development of hybrids. Among these genes, miR167 is known to control the expression of auxin response factors, thereby potentially controlling the fertility aspects of both ovules and anthers (Wu et al., 2006); thus implicating its involvement in flax anther development. Three genes (miR159c, miR172e and miR396c) showed variable expression patterns in almost all the tissues analyzed and hence, they most probably might be involved in the regulation and fine tuning the expression of some housekeeping genes. However, more studies are required to understand the functions of these genes in flax development.

The miR172 family genes target the AP2 transcription factor and are involved in specification of flower organ identity, flowering time and phase change (juvenile/adult vegetative) in plants (Jung et al., 2011). Earlier studies in flax revealed that the miR172 family was highly expressed in stem apex suggesting its role in phase change (Neutelings et al., 2012). Interestingly, this family also showed relatively higher expression in stem in the present study. miR159 is a highly conserved miRNA family, essential for plant development and fertility (Jones-Rhoades et al., 2006), and in Arabidopsis, miR159 genes have shown to be involved in anther development (Xie et al., 2011). Previous studies indicated abundant and widespread expression of miR159 in plants (Axtell and Bartel, 2005) and target mostly the MYB transcription factors. In the present study, a miR159 gene expressed in almost all the tissues analyzed, with the highest expression in flower. Similarly, the flax miR396 genes were expressed in all the tissues with the highest expression in flower. Thus, it is likely that these genes play diverse roles in flax development. The miR396 genes regulate conserved targets belonging to the growth regulating factor (GRF) family of transcription factors, which are known to control cell proliferation in Arabidopsis leaves. The miR156a showed the highest expression in etiolated seedlings and leaf. This miRNA family targets Squamosa promoter-binding protein (SBP), which is involved in controlling flowering time, phase change (Wang et al., 2009) and regulate anthocyanin accumulation in the stem in Arabidopsis (Gou et al., 2011). However, its expression in the flax leaf tissue is difficult to explain at this stage and possibly be

involved in the leaf development. A more detailed study of this gene is necessary to understand its role in leaf development in flax.

Further, flax EST database was also used to obtain expression evidence for the identified miRNA genes. Expression evidence for 19% of the miRNA genes was obtained and most of these genes were expressed in developing seed tissue. Most of the miRNAs were expressed in stem (ST) and torpedo stage embryo (TE), followed by globular embryo (GE) and heart shaped embryo (HE). Likewise, only one miRNA each was expressed in stem peel (PE) and etiolated seedlings (ES). Among all the expressed miRNAs, lus-MIR159b and lus-MIR159c were expressed in most diverse tissue types including stem, stem peel as well as developing seed tissues. The over-representation of developing seed tissues in miRNA expression could possibly be because of the large numbers of ESTs were derived from the developing seed tissues. Moreover, the ESTs were obtained from size-selected cDNA libraries, wherein the low molecular weight RNA was excluded. This also explains the lower expression evidence for the miRNA genes, which might increase when diverse tissue types are specifically assayed by small RNA sequencing.

3.4.3 Identification and expression of target genes

To understand the function of the identified miRNAs, the potential miRNA targets of miRNA genes were predicted using the psRNATarget webserver. A total of 142 putative genes were targeted by 105 miRNA genes belonging to 22 families. Almost all of the flax miRNAs were predicted to target more than four target genes and this is in accordance with the report that miRNAs might have approximately 100 target sites within the protein-coding genes (Brennecke et al., 2005). Therefore, flax miRNAs might target multiple genes involved in a wide variety of biological and metabolic pathways.

A majority of the identified miRNAs target transcription factors, suggesting their roles in post-transcriptional regulation and transcription networks. For example, miR156 targets SBP transcription factors in monocots and dicots (Zhang et al., 2008); while miR167 and miR160 target auxin response factor (ARF) transcription factors in Arabidopsis and maize, respectively (Wu et al., 2006) and were also captured by our analyses. Additionally, flax miR172 targeting APETALA2-like (AP2) proteins were also predicted, which ultimately promote floral organ maturation (Aukerman and Sakai, 2003). Besides transcription factors, other predicted target genes were involved

in diverse physiological and metabolic processes. Furthermore, the GO enrichment analysis revealed that the miRNA targeted genes were involved in regulation of cellular processes (78%), biological processes (42%) and metabolic processes (63%). Therefore, the miRNAs targeting these genes need to be experimentally validated to understand their functions in flax development.

Further evidence about the possible roles of nine miRNAs in flax were obtained, by comparing the expression profiles of the corresponding pre-miRNAs and their potential targets in various flax tissues. Almost all the genes, except miR167a and miR167h, were negatively correlated with their corresponding miRNAs; indicating that these miRNAs likely regulate their corresponding predicted target genes. The miRNA-target pair experiments further confirmed that they play crucial roles in plant development (Wu et al., 2006). For example, the expression of miR167-resistant ARF6 leads to arrested ovule development and indehiscent anthers (Wu et al., 2006). In the present study, the miR167h targets the LIM gene, which may act as a transcriptional activator of lignin biosynthesis, pollen tube growth and/or as actin binding and bundling proteins (Arnaud et al., 2012) and the pair expression (miRNA and target) was positively correlated. Such positive correlation indicates that the miRNAs might be co-transcribed with their host genes and targets and self-regulate their host genes. Similar positive correlation was also observed for some miRNA-target pair in *Glycine max* (Wang and Li, 2009).

Both genetic as well as biochemical studies revealed that miR166 regulate the expression of HD-ZIP III genes, a class of important transcription factors that control leaf development in Arabidopsis (Zhong and Ye, 2007) and organ polarity in plants (Yang et al., 2007). The flax miR166b and miR166d target the leucine zipper protein HOX32 and homeodomain transcription factor genes, respectively and their expression was repressed in anther, indicating their roles in the anther development. The AP2 domain-containing transcription factor was predicted to be the target of lus-miR172e and miR172i. The miR172 genes repress the AP2 genes thereby playing a crucial role in the identity of floral organs and flowering time. The present study indicates that the target gene expression was repressed in the anther tissue, possibly playing a role in anther development in flax. In addition to miR172, miR156 also targets SBP genes and regulate flowering time or floral organ development; although their roles in floral transition needs confirmation (Jones-Rhoades et al., 2006). The flax miR156a was predicted to target the SBP and repressed its expression in the leaf

and etiolated seedlings. Further study is necessary to understand its role in leaf development in flax.

3.4.4 Role of miRNA genes and their targets in oil biosynthesis and accumulation

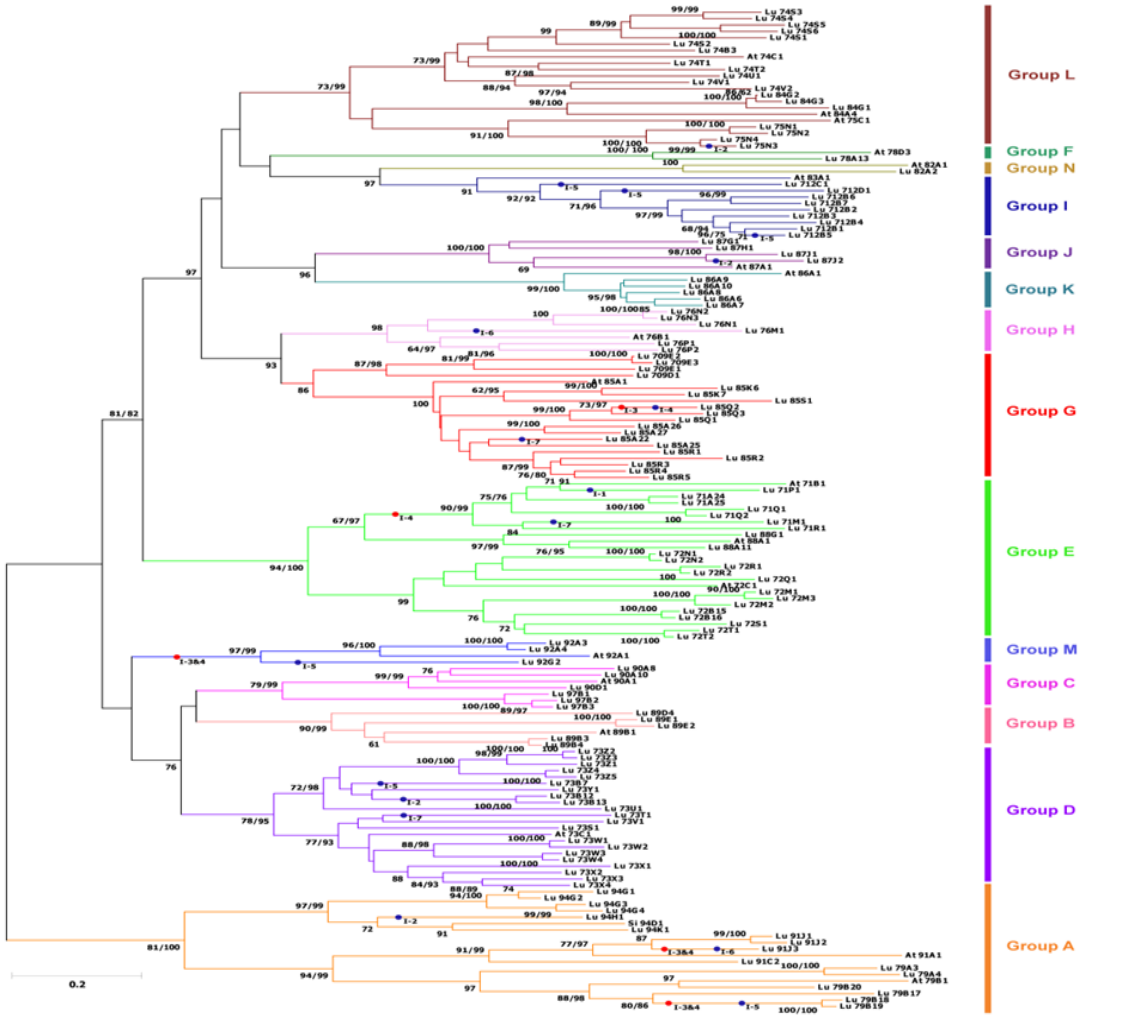
MicroRNA profiling in two *Brassica napus* cultivars substantiated the role of miRNAs in oil biosynthesis and accumulation (Zhao et al., 2012). Four miRNA families, namely miR169, miR390, miR394, and miR6028, had consistently higher expression in the low oil cultivar of *Brassica napus*. Of the four families; miR169, miR390 and miR394 were also identified in the present analysis and might have roles in oil accumulation. Most of the identified target transcripts coded for transcription factors controlling various metabolic pathways including oil biosynthesis. The transcription factors, which are targets for miRNAs, like squamosa promoter binding protein-like (SPL), auxin response factor (ARF) and VIP1 had proven to be involved in oil metabolism (Zhao et al., 2012). The same transcription factor has also been identified as a target of miRNAs in the present study and can be considered as a candidate gene for manipulating oil yield in flax. Overexpression of endogenous miRNA in oil producing green algae *Chlamydomonas reinhardtii* gave enhanced oil yield and further supporting the role of miRNA in oil metabolism (Maor et al., 2011). the role of glabra2 transcription factor in enhancing oil content of Arabidopsis seeds has also been very well established (Shi et al., 2012). Glabra2 silenced genotypes could produce more oil by redirecting carbon flux from mucilage biosynthesis to FA biosynthesis. Glabra2 transcription factor belongs to the family of homeodomain leucine zipper transcription factor (HD- ZIP TF). In the present analysis, miRNA family 166 targets HD-ZIP transcription factors and overexpression of these miRNA genes possibly result in increased flax oil content. Either overexpression or silencing of particular miRNA gene might result into increased or altered oil composition; however, identifying and targeting specific miRNA will be a challenging task.

In summary, the present study identified 116 conserved miRNAs belonging to 23 families from the flax genome using a computational approach. The precursor miRNAs varied in length; while most of the mature miRNAs were 21 nucleotide long, intergenic and showed conserved signatures of RNA polymerase II transcripts in their upstream regions. Promoter region analysis of these miRNA genes indicated prevalence of MYB transcription factor binding sites. Four miRNA gene clusters

containing members of three phylogenetic groups were identified. Further, 142 target genes were predicted for these miRNAs and most of these represent transcriptional regulators. The miRNA encoding genes were expressed in diverse tissues as determined by digital expression analysis as well as real-time PCR. The expression of 14 miRNAs and nine target genes was independently validated using the reverse transcription quantitative PCR. This study suggests that a large number of conserved plant miRNAs are also found in flax and these may play important roles in growth and development of flax. The present study also suggested that these miRNA genes and/ or target genes can be genetically modified to get desired changes such as enhanced or altered oil composition.

CHAPTER 4

Phylogenomic analysis of UDP glycosyltransferase 1 multigene family in flax regulating secondary metabolite accumulation



A part of this work has been published in *BMC Genomics* (2012) 13: 175

4.1 Introduction

The glycosylation process is catalyzed by glycosyltransferase enzymes (GTs), which are highly divergent, polyphyletic, belong to a multigene family and found in all living organisms (Mackenzie et al., 1997). GTs from diverse species have been classified into 92 families based on the amino acid sequence similarities, catalytic mechanisms and the presence of conserved sequence motifs (http://www.cazy.org/GlycosylTransferases.html). Among these, the glycosyltransferase family 1 is the largest family, the enzymes of which generally catalyze transfer of the glycosyl group from nucleoside diphosphate-activated sugars (e.g., UDP-sugars) to a diverse array of substrates, including hormones, secondary metabolites and xenobiotics such as pesticides and herbicides (Jones and Vogt, 2001; Ross et al., 2001). The glycosylation is terminal and regulatory process in secondary metabolite biosynthesis. Glycosylation modulates solubility and stability of compounds resulting in accumulation of particular secondary metabolite. Phenylpropanoid pathway has been efficiently controlled by UDP glycosyltransferases (UGT) types of enzymes (Aksamit-Stachurska et al., 2008). The plant UGT enzymes are characterized by a unique, well-conserved sequence of 44 amino acid residues designated as the plant secondary product glycosyltransferases (PSPG) box (Paquette et al., 2003) and a catalytic mechanism that inverts the anomeric configuration of a transferred sugar (Wang and Hou, 2009). This important class of regulatory enzymes was not studied from flax, although flax seed has around 800 times higher concentration of lignans than any other related plants.

The GT family 1 has been extensively studied in various plants species, as well as in humans. In mammals, UGTs coordinate the activity of signal molecules such as steroid hormones and detoxify xenobiotic compounds taken up from the environment (Tukey and Strassburg, 2000). Polymorphisms among these UGTs have been shown to be associated with increased susceptibility to certain diseases in humans (Strassburg et al., 2002). Studies in model plants have shown that the plant genomes contain a great diversity of gene sequences predicted to be involved in glycosylation (Li et al., 2001; Geisler-Lee et al., 2006). The occurrence of a wide range of glycosylated products in flax (Eliasson et al., 2003) suggests the presence of a large number of UGTs. The availability of the flax genome sequence (http://www.linum.ca), tissue specific ESTs (http://www.ncbi.nlm.nih.gov/nuccore?

term=Linum%20usitatissimum) and microarray expression dataset (Fenart et al., 2010) (http://www.ncbi.nlm.nih.gov/projects/geo/) of flax provide an opportunity to analyze the diversity of expressed glycosyltransferase family genes in this economically important oilseed crop.

4.2 Materials and Methods

4.2.1 Probing the flax genome for UGT genes

The presently available draft genome sequence of flax represents 85% genome coverage, which is derived from the low-copy fraction of the genome (Wang et al., 2012). This coverage is consistent with the length of the entire low-copy fraction previously estimated by reassociation kinetics (Cullis, 1981). The predicted protein database available at http://www.linum.ca was used to identify flax UGT genes. The 44 amino acid conserved sequence of the PSPG box that characterizes plant UGTs was used as a query against the 47,912 predicted flax gene models. The resulting scaffolds were analyzed to identify the genes, ORFs, intron positions and sizes using the GBrowse tool available on the same website.

4.2.2 PCR amplification, cloning and sequencing

Genomic DNA from a flax variety, NL260, was extracted using CTAB method. Total RNA from developing seeds was extracted using Spectrum Plant Total RNA kit (Sigma-Aldrich, USA) and treated with DNaseI (Promega, USA), followed by first strand cDNA synthesis using AMV Reverse Transcriptase (Promega, USA). To confirm the reading frames, primers were designed to amplify full length genes including the start and stop codons **(Table S 4.1A**). For intron-less genes, 50 ng genomic DNA, and for intron containing genes, 1.5 µl pooled cDNA from developing seeds was used as template for PCR amplification using AccuPrime™ *Pfx* DNA Polymerase (Invitrogen, USA). PCR was performed using the annealing temperatures mentioned in **Table S 4.1A**. The PCR amplicons were analyzed on 1.0% agarose gels and eluted using GenElute gel extraction kit (Sigma-Aldrich, USA) followed by cloning into pGEM-T Easy vector (Promega, USA). Plasmid DNA was isolated using GenElute plasmid extraction kit (Sigma-Aldrich, USA) and sequenced using MegaBACE 500 DNA analysis system (GE Healthcare, UK).

4.2.3 Sequence alignment and phylogenetic analysis

The predicted amino acid sequences of the UGT genes were initially aligned using ClustalW with default gap penalties (Thompson et al., 1994). These alignments were visually inspected for indels and to minimize insertion/deletion events in unalignable regions. Trees were constructed from 409 alignable amino acid positions (60.41%) for all the sequences. Distance as well as Parsimony analyses were performed using MEGA5 (Tamura et al., 2011). Only the regions of unambiguous alignments were used in the phylogenetic analyses with Dayhoff substitution matrix (PAM250) and trees were constructed by neighbour-joining algorithm (Saitou and Nei, 1987) with bootstrapping (1000 replicates). Eighteen Arabidopsis UGT sequences, one from each UGT family and one sesame sequence (UGT94D1) were also included in the analyses (**Table S 4.2**).

4.2.4 Intron mapping and organization

A flax UGT intron map was constructed by determining the intron splice sites, phases and positions. The introns were serially numbered relative to their positions in the amino acid sequence produced by aligning all the flax UGTs. Intron phases were determined as follows: introns positioned between two codons as phase 0, introns positioned after the first base in the codon as phase 1, and introns positioned after the second base in the codon as phase 2.

4.2.5 Detection of orthologs of flax UGTs in four sequenced dicots

Blast2Go (Conesa et al., 2005) was used to search the orthologs for flax UGTs in four sequenced dicots, *Ricinus communis* (*Euphorbiaceae*), *Populus trichocarpa* (*Salicaceae*), *Vitis vinifera* (*Vitaceae*) and *Arabidopsis thaliana* (*Brassicaceae*), using default parameters except for E value cut off of $<e^{-100}$. These four dicots were selected based on the genome homologies with flax as reported by Ragupathy et al. (2011).

4.2.6 Digital expression analysis

The putative UGT coding sequences were BLAST searched against the *Linum usitatissimum* NCBI-EST dataset (dated: June, 2011; 2,86,895 sequences; http://www.ncbi.nlm.nih.gov/nucest?term=Linum%20usitasimum) to identify transcriptional evidence for individual UGT genes and to estimate the number of

ESTs expressed per tissue type and gene model. These tissue types include flower (FL), globular embryo (GE), heart embryo (HE), torpedo embryo (TE), bent embryo (BE), mature embryo (ME), seed coat at globular stage (GC), seed coat at torpedo stage (TC), pooled endosperm (EN), etiolated seedling (ES), stem (ST), leaf (LE), peeled stem (PS) (Venglat et al., 2011), 12days DAF bolls and outer fibrous stem tissue. Additionally, microarray expression data for 48,021 flax unigenes (http://www.ncbi.nlm.nih.gov/geo/ query/acc.cgi?acc=GSE21868) were also used. RMA - normalized, averaged gene-level signal intensity (log2) values for the unigenes exhibiting specified sequence similarity were used from all the biological as well as technical replicates and averaged further. A heat map for digital expression analysis was constructed with these values using TIGR MultiExperiment Viewer (MeV, http://www.tm4.org/mev.html).

4.2.7 Reverse transcription quantitative real time PCR

Total RNA from mature leaves (ML), stem (ST), root (RT), etiolated seedling (ES), flower (FL) and seed developmental stages (4, 8, 12, 16, 22, 30, 48 days after flowering) of flax variety NL260 was isolated using spectrum™ plant total RNA kit (Sigma, USA) as manufacturer's instructions. DNaseI treated total RNA was reverse transcribed using oligo(dT) primer and MultiScribe™ reverse transcriptase (Applied Biosystems, USA). Gene specific primers for 10 glycosyltransferase genes (**Table S 4.1B**) were designed using Primer3 (Rozen and Skaletsky, 2000). PCR conditions were optimized for annealing temperature and primer concentration. Primers used for real-time PCR are listed in **Table S 4.1B**. Real-time PCR was carried out in 7900HT Fast real-time PCR system (Applied Biosystems, USA) using FastStart universal SYBR green master mix (Roche, USA). Each 10 µL real-time PCR cocktail contained 0.125-0.4 µM (**Table S 4.1B**) concentrations of both forward and reverse gene-specific primers, 4 µL of 1:16 diluted first strand cDNA, ⅓ SYBR green master mix and sterile milliQ water to make up the reaction volume. Real-time PCR amplification reactions were performed with following conditions: 95 °C denaturation for 10 min, followed by 40 cycles of 95 °C for 3s, with primer annealing and extension at 60 °C for 30s. Following amplification, a melting dissociation curve was generated using a 62–95 °C ramp with 0.4 °C increment per cycle in order to monitor the specificity of each primer pair. Eukaryotic translation initiation factor 5A (*ETIF5A*) gene from flax was used as a housekeeping or reference gene for all the real-time PCRs (Huis et al.,

2010). The housekeeping gene was selected after confirming the stability of this gene across all the tissue types used in the study. For each biological replicate, two independent technical replications were performed and averaged for further calculations. PCR conditions were optimized such that PCR efficiencies of the housekeeping gene and the gene of interest will be closer to two. PCR efficiencies were calculated using LinRegPCR v12.x (Ramakers et al., 2003). Relative transcript abundance calculations were performed using comparative C_T (ΔC_T) method as described by Schmittgen and Livak (2008). For each of the three biological replicates, two independent technical replications were performed.

4.3 Results

4.3.1 Identification of flax UGT genes

BlastP search against the 47,912 flax gene models (http://www.linum.ca) using the conserved PSPG box sequence resulted in the identification of 179 scaffolds. Family 1 UGTs usually utilize low molecular weight compounds as acceptor substrates and UDP-sugars as donors (Li et al., 2001) and commonly possess a carboxy terminal consensus sequence (PSPG box) believed to be involved in binding to the UDP moiety of the sugar nucleotide donor (Vogt and Jones, 2000; Wang and Hou, 2009). Taking these characteristics into account, 137 sequences (**GenBank accession numbers JN088282-JN088418; Table S 4.3**) having lengths of 375–530 amino acids and 0–2 introns were selected and subjected to phylogenetic and digital expression analysis. In order to confirm the open reading frame (ORF) sequence of these genes, 11 genes expressed in seed tissue were randomly selected, isolated using PCR, cloned and sequenced, which revealed that they were 100% identical to the putative UGT gene sequences identified.

4.3.2 Phylogenetic analysis

All the identified putative UGT genes were classified as per the recommendations of the UGT Nomenclature Committee (Mackenzie et al., 1997) **(Table S 4.4**). As expected, the PSPG signature motif was present in all the UGT sequences and the overall sequence similarity among them varied substantially from 36 to 98% (**Table S 4.5**). A total of 409 amino acid positions (60.41% of the sequences) were aligned for all the genes analyzed and used to construct a phylogenetic tree. Fourteen major

groups (A-N) were defined by both the neighbour-joining (NJ) and parsimony methods with high bootstrap supports (>85) (**Figure 4.1**). The tree topology and grouping of the UGTs were similar as described for the Arabidopsis UGT genes (Bowles, 2002), e.g. group L consists of the UGTs belonging to the families 74, 75 and 84. However, in four groups, A, C, G and I, sequences from additional UGT families were observed *viz.* LuUGT94, LuUGT97, LuUGT709 and LuUGT712, respectively. The number of genes (1–22) as well as the sequence diversity varied considerably within each group (**Table S 4.5**).

4.3.3 Detection of orthologs and duplicated genes

The orthologs of flax UGTs identified in the four selected dicots are listed in the **Table S 4.6**. Of the 137 sequences, orthologs were identified for 130 UGTs from at least one of the four dicots. However, for 72 sequences, orthologs were identified from all the four species. The maximum number of orthologs (125) was identified in case of *Vitis vinifera*, while the lowest (80) were detected in case of *Arabidopsis thaliana*. Seven flax diverged UGTs were identified (*LuUGT94G1*, *LuUGT94G2*, *LuUGT94G3*, *LuUGT94G4*, *LuUGT94H1*, *LuUGT75N3* and *LuUGT79A4*) and 22 gene duplication events with sequence similarity of ~90% were observed (**Table S 4.7**).

4.3.4 Analysis of intron gain/loss events

Among the 137 sequences, 55 were intron less, while 72 and 10 had one and two introns each, respectively (**Table S 4.3**). Total 92 introns were detected in the 137 UGTs, with an average of 0.67 intron per gene. Seven independent intron insertion events were observed when the intron positions were compared with the sequence relationship predicted by the phylogenetic analysis (**Figure 4.2**). An intron was considered conserved if its position in a particular sequence was within 40–45 amino acids of its mean recorded position across the sequences (for complete sequence alignment see **Figure S 4.1**). Two conserved introns (intron 3 and intron 4) were identified, of which intron 3 was observed in 44 UGTs belonging to the A, C and F-J phylogenetic groups, while intron 4 was observed in 27 UGTs belonging to the D, E, K and L phylogenetic groups. *LuUGT79A4* from group A and *LuUGT709E3* from group G both had the conserved introns.

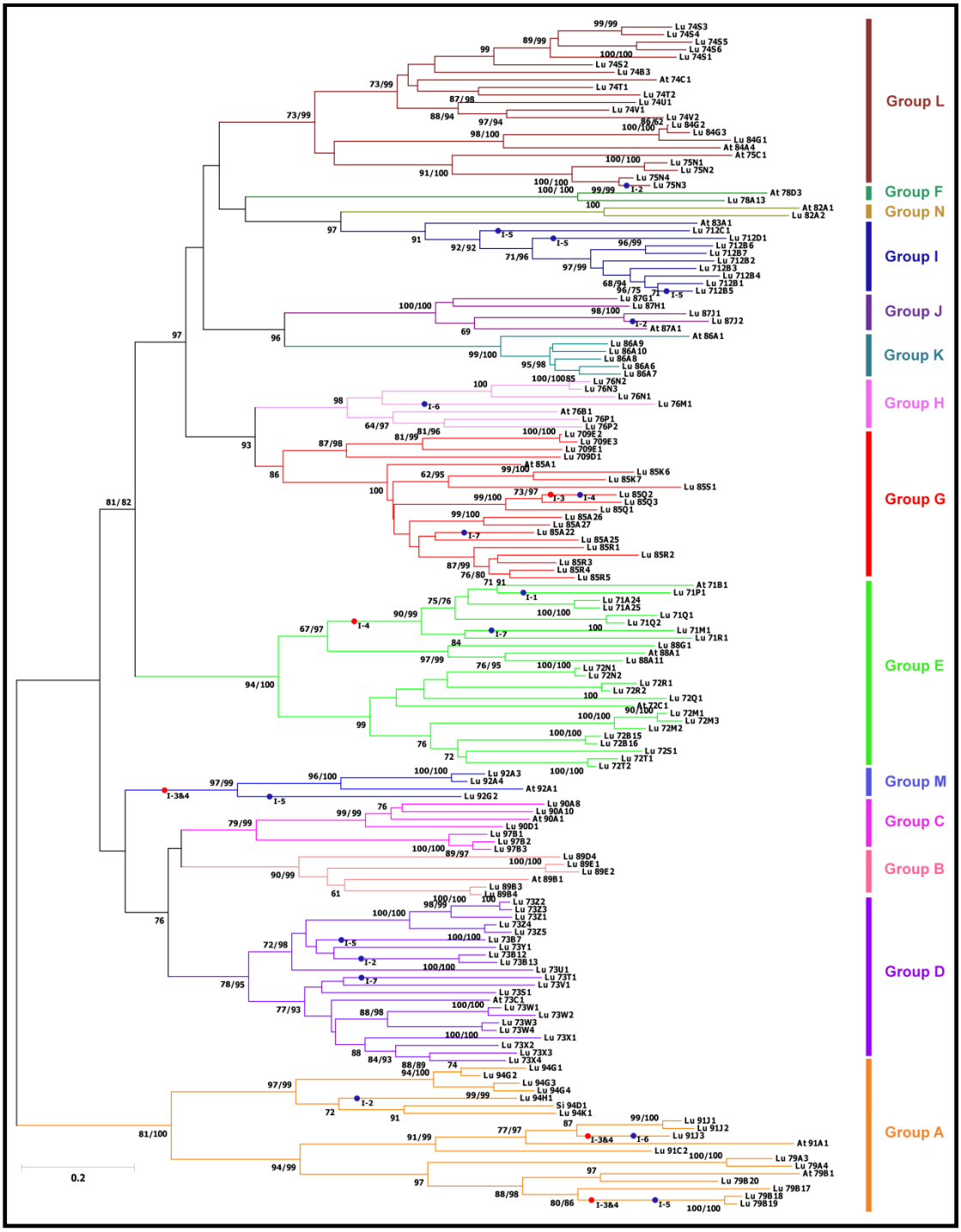


Figure 4.1 Phylogenetic analysis of the flax UGT family genes. The tree was derived by neighbour-joining distance analysis of alignable regions comprising ~60% of the UGT sequences using MEGA5. Bootstrap values over 60% are indicated at the nodes, with the number on the *left* for neighbour-joining and *right* for parsimony methods. Hypothetical positions of intron gain and loss are indicated by *dots* followed by intron number and it is assumed that introns 3 and 4 were gained prior to diversification of flax UGTs (see **Figure 4.2)**. Postulated intron gains are indicated by

blue dots and intron losses by red dots. Eighteen Arabidopsis and one Sesame UGT sequences from each UGT family were included in the analysis **(Accession numbers given in Table S 4.2)**

Alternatively, group M showed absence of both the conserved introns, while *LuUGT92G2* from group M showed gain of intron 5. Within the members of groups F-J and N, intron 3 was predominant, except in *LuUGT85Q2* and *LuUGT87J2*. In comparison, the members of groups K and L had intron 4, while only one member of L group (*LuUGT74S1*) showed the presence of intron 3. All other introns were either found only within a single restricted group of closely related sequences or in only a single gene. Group B members were intron less.

Many sequences showed loss of the conserved introns and gain of other introns. For example, within group A, three members from family 79 and one member from family 91 (*LuUGT91J3*) showed loss of conserved introns 3 and 4, and gain of introns 5 and 6, respectively. Similarly, within group D, four members of family 73 lost conserved intron 4 and few members gained introns 2, 5 and 7. Likewise, in group E, all the members of the family 71 showed loss of conserved intron 4 while gain of introns 1, 7 and 8 in few members.

Most of the conserved introns were either in phase 1 (49 genes) or phase 0 (15 genes) **(Table S 4.3**). The intron sizes of flax UGTs ranged from 65 bp to 2258 bp with an average of 406bp for both the introns. About 28% of the flax UGT introns were in the size range of 65–99bp **(Figure S 4.2)**.

In Arabidopsis, 37 out of 88 UGT genes contained introns while, three genes had two introns. By comparing the intron positions with sequence relationships predicted by phylogenetic analysis, a minimum of nine independent intron insertion events appear to have happened in the course of UGT evolution in Arabidopsis. Intron 2 was found to be widespread and oldest intron and was present in all of the 23 UGT sequences in groups F–K in Arabidopsis (Li et al., 2001). Similarly in flax, the introns 3 and 4 have been found in most members of the groups F-J and K, respectively and could be considered as the oldest introns.

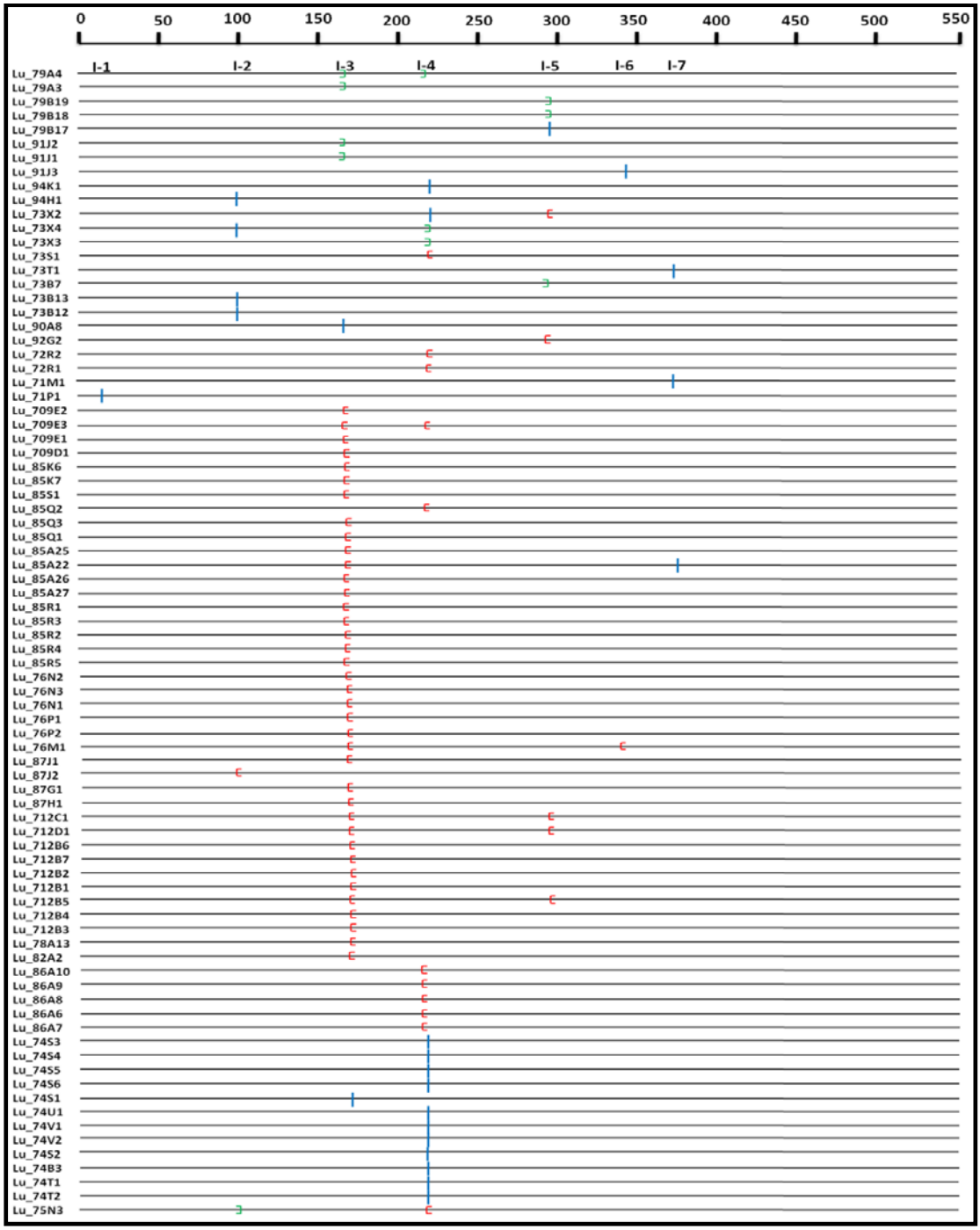


Figure 4.2 Distribution of introns among 82 UGT genes of Flax. The introns are mapped and numbered to the alignment of their amino acid sequences. It is hypothesized that the introns 3 and 4 were gained prior to diversification of flax UGTs and the gain and loss of other introns within a phylogenetic genetic group are indicated by the colored mark found or predicted in the corresponding genes. The numbers on the top of the map show the intron insertion number occurred on each gene. Intron phases are indicated by blue bar, red open bracket and green close bracket for zero, one and two, respectively.

4.3.5 Expression analysis of flax UGT genes using EST data

Expression of the identified UGT genes was analyzed using the available EST and microarray data of flax. Of the 137 genes, 100 genes showed expression evidence based on either or both the datasets. Among these, 85 genes (62.04%) were expressed based on the EST data; while the microarray data indicated expression evidence for 60 genes (43.79%) (**Table S 4.8**). Similarly for 45 genes, the expression evidence was present in both the datasets. Further, the ESTs from various flax tissues were mapped onto the 137 flax UGT gene models to estimate their gene expression levels. This analysis identified that a total of 325 ESTs mapped to 85 flax UGT sequences with an average of 3.82 ESTs per gene. The frequency of ESTs varied greatly from 1 to 54 per UGT gene model. Among the various tissue types, flower (FL, 18.46%) and seed coat at torpedo stage (TC, 15.69%) had the largest number of highly expressed genes, while globular embryo (GE) stage had the lowest (2, 0.61%) number of expressed genes.

The highest number of (91) ESTs were mapped to 13 sequences of group G, followed by 69 ESTs mapping to 15 members of group E. On the contrary, only one EST was mapped to a single group N member. On an average, the highest of 7.00 ESTs were mapped per UGT sequence of family G, followed by 4.60 ESTs per gene of family E. The percentage of the genes expressed per phylogenetic group or family varied from 28 to 100% (**Table S 4.8**). Among all the genes expressed, *LuUGT85Q2* and *LuUGT74S1* showed the highest expression in flower (FL) and seed coat at torpedo stage (TC), respectively (**Table S 4.8**).

4.3.6 Expression analysis of flax UGT genes using microarray data

In addition to the sequence based expression analysis method, publicly available microarray data were also used (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE21868) under the platform GSE21868, which profiles expression patterns for various flax tissues and seed developmental stages, *viz*., roots (R), leaves (L), stem outer tissues: vegetative stage (SOV), stem outer tissues: green capsule stage (SOGC), stem inner tissues: vegetative stage (SIV), stem inner tissues: green capsule stage (SIGC), seeds: 10–15 days after flowering (DAF) (S1), seeds: 20–30 DAF (S2) and seeds: 40–50 DAF (S3) (Fenart et al., 2010). The Robust Multichip Average (RMA)-normalized, averaged gene-level log2 values were used for analyzing expression

evidence of UGTs and to construct a heat map (**Figure 4.3**). Hierarchical clustering with Pearson correlation matrix highlighted co-expression of specific gene family members in specific tissue types. Only 60 of the 137 (43.79%) flax UGTs represented on the array showed expression evidence (**Table S 4.8**). Three genes were highly expressed in seed stages S2 and S3 (averaged gene-level log2 value: *LuUGT85R2* (11.11 and 11.30), *LuUGT709E2* (10.57 and 10.76), and *LuUGT709E3* (10.57 and 10.76), respectively; while one gene (*LuUGT85Q3*, averaged gene-level log2 value: 11.53) showed the highest expression in leaf tissue (**Figure 4.3**). The number of genes having higher expression in different tissues (averaged gene-level log2 values >6.96) varied from 14 (S1) to 24 (SOGC) (**Table S 4.8**). Among the different tissues, SOGC had the largest number of highly expressed genes, while S3 had the lowest (23%) (**Table S 4.8**). Surprisingly, the two contrasting varieties, Drakkar and Belinka did not show any difference in the expression of these 60 UGTs (**Figure 4.3**).

4.3.7 Expression profiling using RT-qPCR

The RT-qPCR is currently the most accurate method for detecting differential gene expression. The 12 tissue types selected for UGT expression profiling cover all plant parts and seed developmental stages from fertilization to seed maturation. *Eukaryotic translation initiation factor 5A* (*ETIF5A*, GenBank ID GR508912) was selected as a reference gene after confirming the stability of this gene across all the tissue types used in the study (Huis et al., 2010). Single dissociation curves were observed for all the flax UGT genes and *ETIF5A*, confirming amplification specificity of the primers. The ΔC_T method (Schmittgen and Livak, 2008) was used to express the results relative to the reference gene. A validation experiment was conducted to ensure similar amplification efficiencies of all the genes analyzed.

Relative transcript abundance of 10 flax UGT genes was profiled and is graphically represented in **Figure 4.4**. All the selected genes had EST expression evidence and covered six phylogenetic groups. The *LuUGT71M1* transcript was detected in mature leaves, stem, etiolated seedling and 48 DAF; however, the relative expression level compared to other UGT genes was very low. *LuUGT94G1* is expressed constitutively in almost all tissues types; specifically it showed maximum expression in stem. Its expression was also supported by ESTs from stem peel library.

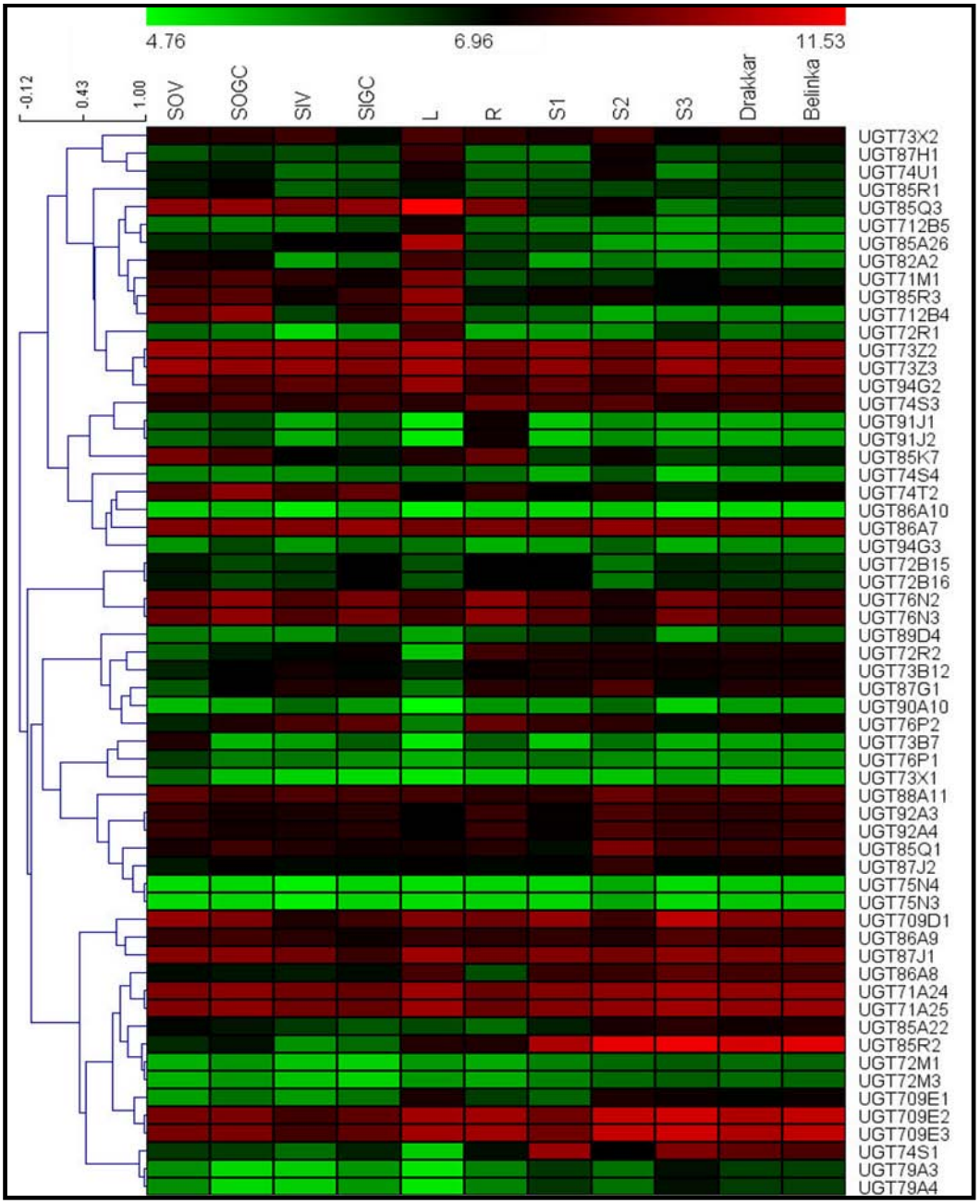


Figure 4.3 Expression levels for flax UGT genes in various tissues by microarray analysis. The RMA-normalized, average log2 signal values of flax UGTs in various tissues and seed developmental stages (listed at the top of heat map) were used for construction of the heat map. The left side of the heat map shows hierarchical clustering based on Pearson correlation matrix. The colour scale (representing log2 signal values) is shown at the top. Microarray data from stem outer tissues; vegetative stage (SOV), stem outer tissues, green capsule stage (SOGC), stem inner tissues; vegetative stage (SIV), stem inner tissues; green capsule stage (SIGC), leaves (L), roots (R), seeds, 10–15 DAF (S1), seeds, 20–30 DAF (S2) and seeds, 40–50 DAF (S3) were used for constructing the expression heat map.

LuUGT72N1 expressed in flower, 4 and 8 DAF with peak at 4 DAF. *LuUGT85Q2* had 54 ESTs mapped from flower EST library and RT-qPCR analysis confirmed its high expression in flower. Expression of *LuUGT89B3* gene was observed in later stages of seed development *viz.* 30 and 48 DAF and supported by two EST clones identified in torpedo seed coat stage. Gene *LuUGT72M2* expressed in mature leaves, flowers and early seed developmental stages whereas *LuUGT72R1* and *LuUGT712B1* were highly expressed in various seed developmental stages. Gene *LuUGT85Q1* belonged to family 85 which is known to be involved in glycosylation of cyanogenic compounds (Thorsoe et al., 2005). The abundance of cyanogenic compounds and higher expression of *LuUGT85Q1* in stem, root and mature seed (i.e. 48 DAF) suggest the putative function as cyanogenic glycosyltransferases (Shahidi and Wanasundara, 1997). The *LuUGT74S1* expressed highly in developmental seed stages and peaked at 12 DAF i.e. torpedo stage of embryo. Flax has a major lignan secoisolariciresinol diglucoside which is phenylpropanoid and accumulates in seed coat (Hano et al., 2006). UGTs belonging to the gene family 74 glycosylate phenylpropanoid group of compounds. About 25 ESTs clones from torpedo stage seed coat library were mapped on *LuUGT74S1* gene indicating its putative *in planta* function as secoisolariciresinol glycosyltransferase. Expression profiles of the 10 selected genes analyzed using RT-qPCR, matched well with the digital expression results.

4.4 Discussion

Glycosylation mediated by glycosyltransferase enzymes (GTs) is a critical step in metabolic pathways with diverse roles in cellular processes and homeostasis (Jaeken and Matthijs, 2001). Recent studies involving functional characterization of plant GTs suggest their important roles in growth, development and interaction with the environment (Wang and Hou, 2009). The activities of many GTs from a variety of plants and biological roles of their products have been known for a long time (Schneider and Schliemann, 1994). However, the methods for identification of UGTs based on biochemical and classical genetic approaches are slow and difficult (Vogt and Jones, 2000). Recent developments in plant genomics stimulated the use of strategies such as differential display methods and/or homology-based screening of cDNA libraries for identification and isolation of novel UGT genes (Yamazaki et al.,

1999; Martin et al., 2001; Ono et al., 2006), although the roles of many UGTs still remain uncertain. Availability of whole genome sequence of many plants enabled a thorough and detailed analysis of multigene families. For example, in Arabidopsis, genome-wide search using PSPG motif identified 120 putative UGT genes. Similarly, a whole genome survey of six plant species resulted in identification of 56 (*Carica papaya*) to 242 (*Glycine max*) UGTs (Yonekura-Sakakibara and Hanada, 2011).

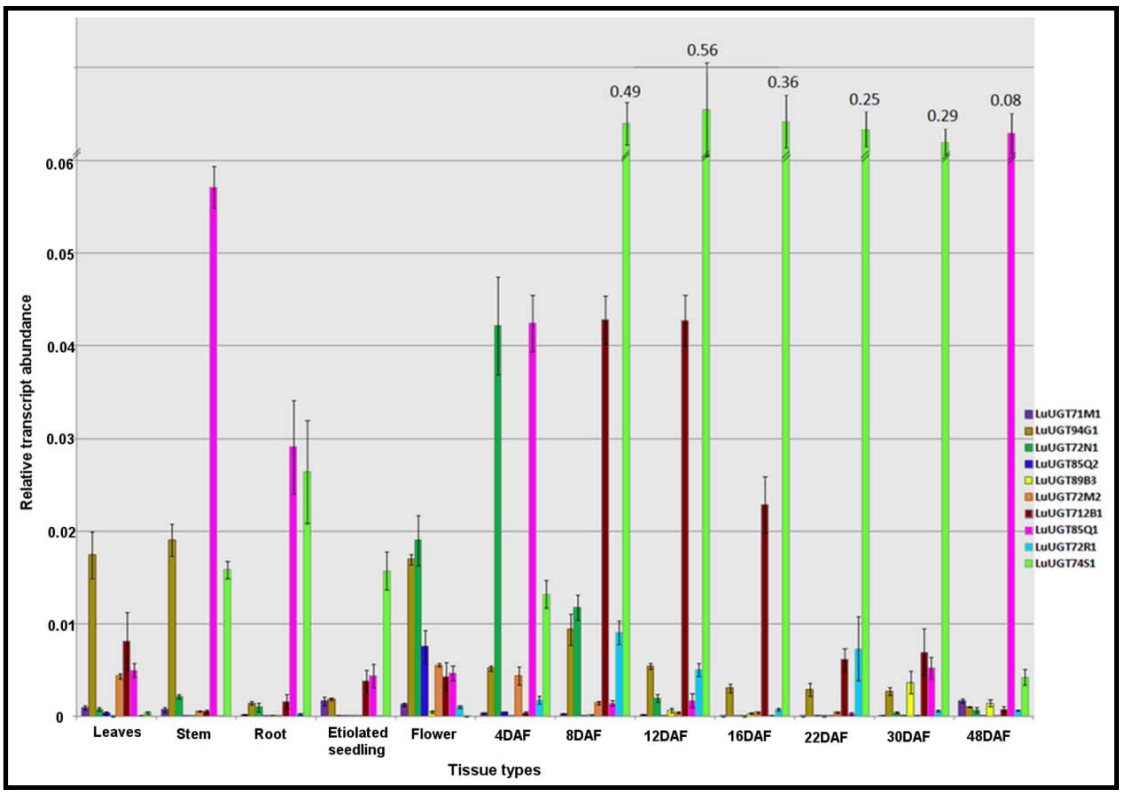


Figure 4.4 RT-qPCR expression profile of 10 selected flax UGT genes in 12 different tissue types. Tissue types analysed for LuUGT expression include; mature leaves (ML), stem (ST), root (RT), etiolated seedling (ES), flower (FL) and seed developmental stages (4, 8, 12, 16, 22, 30, 48 DAF). These graphs show the relative transcript abundance of each gene in comparison with the reference gene, *Linum usitatissimum* ETIF5A (GR508912). Expression values are reported as the average of three biological and two technical replicates. Values correspond to the mean and standard error of biological triplicates.

The recently published draft genome sequence and the extensive tissue specific EST library collections of flax provided an opportunity to investigate the diversity in flax UGT multigene family in a greater detail. One hundred and thirty

seven flax UGTs were identified, which were more than that identified in Arabidopsis but less than that discovered in rice, grapevine and Medicago (Yonekura-Sakakibara and Hanada, 2011). All the identified UGTs contain two major domains, a conserved C-terminal domain and a variable N-terminal domain, although the overall sequence diversity was high among the genes.

4.4.1 Flax UGT family resembles the phylogenetic group structure of Arabidopsis UGTs

A phylogenetic tree provides a framework to compare the properties of gene family members and to identify similarities and differences among them (Jung et al., 2008). In the present study, the flax genome revealed 22 UGT families including four new families (94, 97, 709 and 712), not reported in Arabidopsis. However, phylogenetic analysis of flax UGTs clustered them in 14 groups (A-N) as reported in Arabidopsis (Li et al., 2001; Ross et al., 2001) and interestingly, the four new flax UGT families did not form any additional groups. Moreover, all the six sequences of the UGT94 family clustered with the *Sesamum indicum* UGT94D1 sequence [BAF99027 (Noguchi et al., 2008)]. UGT94D1 and UGT94B1 [AB190262 (Sawada et al., 2005)] are the only UGT94 family sequence reported till now. A phylogenetic tree constructed by Bowles et al. (2005) using 22 UGT sequences reported from other plant species along with the Arabidopsis UGT sequences, mostly resulted in 14 groups, while an additional group of cytokinin GTs was identified containing the *Phaseolus vulgaris* and *Zea mays* UGT sequences (Hou et al., 2004; Bowles et al., 2005). Based on the phylogenetic analysis of Arabidopsis UGTs, it has been shown that it might be possible to correlate, to a large extent, the regiospecificity of glycosylation to the phylogenetic groups (Cartwright et al., 2008). The exception to this might be due to regioswitching events taking place during evolution. In some cases, phylogenetically closely related UGTs show distinct regiospecific differences towards a common acceptor. For example, *Arabidopsis thaliana* UGTs, *AtUGT74F1* and *AtUGT74F2*, share ~82% amino acid sequence identity, and while *AtUGT74F1* glucosylates the phenolic hydroxyl group of 2-hydroxy benzoic acid, *AtUGT74F2* glucosylates both the carboxyl and hydroxyl groups of 2-hydroxy benzoic acid (Lim et al., 2002). On the contrary, in some cases (e.g. UGT85B1), the genes have been shown to exhibit a broad specificity toward acceptors *in vitro*; however, a member of this group (UGT85Q1) in *Sorghum bicolor* specifically catalyzes the conversion of *p*-

hydroxymandelonitrile into dhurrin *in vivo* (Hansen et al., 2003). This analysis, along with amino acid sequence similarity of UGT families within a group, might be useful for predicting substrates (Lim et al., 2003; Bowles et al., 2005). For example, Osmani *et al.* (2009) reported that the group G members glycosylate terpenoids; while the members of groups D, E and L glycosylate flavaonoids, tepenoids and benzoates.

However, a study of several *Medicago truncatula* UGTs highlighted the difficulties in assigning substrate specificity based on phylogeny. Biochemical and phylogenetic studies of *MtUGT78G1* and *MtUGT85H2* showed that substrate specificity could not be predicted by their clustering with biochemically characterized UGTs belonging to the same family (Modolo et al., 2007). Although, few genomes such as rice, poplar, grapevine and Medicago have been screened and annotated for GT genes, they have not been assigned to GT groups and families so far. Apart from the model plant Arabidopsis (Li et al., 2001), this is the first attempt to classify GT genes into groups and families from a crop plant flax, as per the standardized system recommended by the UGT Nomenclature Committee (Mackenzie et al., 1997). Thus, the present analysis of flax UGT genes might help to narrow down the substrate choice of a specific gene.

4.4.2 Detection of orthologs and functional divergence of unique flax UGTs

Detection of orthologs is critically important for accurate functional annotation and has been widely used to facilitate the studies on comparative and evolutionary genomics (Chen et al., 2007). Several methods such as the BlastP (Altschul et al., 1990), inparanoid (Remm et al., 2001) and reciprocal smallest distance (Wall et al., 2003) have been reported to detect orthologs. In the present study, BlastP program was used to identify the orthologs for flax UGTs from four sequenced dicots (*Ricinus communis*, *Populus trichocarpa*, *Vitis vinifera* and *Arabidopsis thaliana*). Of the 137 flax UGTs, 130 UGTs had orthologs from the four dicots and seven flax-diverged UGTs were detected. Based on the microarray and EST data, 95 of these 130 orthologs (73%) showed expression evidence; while, five of the seven flax diverged UGTs revealed expression evidence, suggesting their functional divergence. Thus, the flax diverged UGTs, with significantly different primary sequences than those of other surveyed dicots, might have evolved independently since the last common ancestor between flax and these dicots. As the number of flax diverged UGTs identified in our analysis is small, other methods such as inparanoid search need to be

conducted to identify more flax diverged UGTs that the present analysis might have missed. However, this analysis was not performed, as the flax chromosome sequences were not available for conducting the inparanoid search.

4.4.3 Intron mapping to understand the evolution of UGT family

To understand the evolution of a gene family within phylogenetic groups, introns, more specifically their position, phase, loss and gain, can serve as an important tool (Stoltzfus et al., 1997). Therefore, intron mapping was conducted in the 137 flax UGTs among which 40.14% sequences were intron less. This percentage is less than that observed in Arabidopsis, wherein >50% genes were intron less (Li et al., 2001). In flax UGTs, a total of seven intron positions were identified with the number of introns per family in the range of one to four. Most families showed the presence of conserved introns 3 (53.65%) and 4 (32.92%), which could probably be considered as the oldest among the seven introns identified. Intron 3 was present in almost all members of the groups F-J and N; while intron 4 was dominant in groups L and K. Interestingly, in these groups wherever intron 3 was present, intron 4 was absent and *vice versa* except in case of *LuUGT709E3*, where both the introns were present; while in case of *LuUGT87J2*, both were absent. In other groups, the introns 3 and 4 were absent in some members of groups A, D, M and E. This suggests that either of these introns was gained prior to diversification of flax UGTs. This is also supported by the observation that most of the conserved introns were in the same phase.

It is a commonly held view that the majority of conserved introns are ancient elements and their phases usually remain unchanged (Roy and Gilbert, 2005). In fact, it has been further suggested that the intron sliding or shifts of intron-exon boundary over a few nucleotides causing change of intron phase are rare events and introns retain their phase for a long evolutionary time (Rogozin et al., 2000). Furthermore, the introns other than the conserved introns were found only within a single restricted group of closely related sequences or in only a single gene, suggesting a general pattern of intron gain during evolution of the flax UGT gene family. A clear case of loss of a conserved intron and gain of intron 5 was seen in the subfamily of closely related genes *LuUGTB17-LuUGTB19* from group A. Similarly, in case of *LuUGT73B12* and *LuUGT73B13*, loss of conserved introns and gain of intron 2 was also observed. Thus, analysis of the evolution of the flax UGT multigene family

provides evidence for both intron gain and loss and thereby strongly supports the "intron-late" theory of intron evolution (Palmer and Logsdon, 1991).

4.4.4 Expressed flax UGTs: Identified by digital expression analysis and supported by RT-qPCR

Functional divergence among duplicated genes is one of the most important sources of evolutionary innovation in complex organisms. Interestingly, among the 22 duplicated genes, five pairs of genes *LuUGT94G3* and *LuUGT94G4*, *LuUGT73B12* and *LuUGT73B13*, *LuUGT712B1* and *LuUGT712B5*, *LuUGT86A8* and *LuUGT86A9* and *LuUGT74S5* and *LuUGT74S6*, showed evidence of differential expression. For example, *LuUGT74S5* showed seed coat specific expression, while its duplicated counterpart, *LuUGT74S6*, remained unexpressed. Evidence for differential expression was also provided by the duplicated gene pair *LuUGT86A8* and *LuUGT86A9*. This suggests that after duplication, the genes acquired either differential or tissue specific expression patterns. In an earlier study, Haberer et al. (2004) estimated that about two thirds of duplicate gene pairs had divergent expression in Arabidopsis.

To predict and understand the role of these UGT genes in various tissue types, gene expression pattern analysis is very helpful as one can infer which gene family members are expected to perform distinct or similar roles. With this aim, expression analysis of flax UGTs was performed using EST libraries, microarray data and RT-qPCR. About 62% flax UGTs showed expression evidence based on the EST data and one or more ESTs were detected per tissue type, providing strong evidence that most of the flax UGT genes were expressed in varied tissue types. The expression patterns analyzed using RT-qPCR very well correlated with the digital expression analysis.

The frequency of ESTs per UGT gene ranged from 1–54 among the UGTs, suggesting varied expression levels. Among the different tissue types, seed and stem tissues showed the highest number of expressed UGTs. It is known that flax seeds and stem contain a large number of secondary metabolites and hence could explain the abundance of UGTs in these tissues (Kozlowska et al., 1983; Kraushofer and Sontag, 2002). However, this could also be due to a large number of EST libraries available for these tissue types (seed: 9 EST libraries, 2,20,724 ESTs and stem: 3 EST libraries, 32,184 ESTs; June 2011). This study also identified two genes, *LuUGT85Q2* and *LuUGT74S1*, belonging to groups G and L, respectively, which showed high expression in flower and seed coat from the torpedo stage. The members of these

groups are predicted to glycosylate terpenoids, flavanoids and benzoates classes (Osmani et al., 2009); and hence, they can be considered as potential targets for screening against these predicted classes to identify their substrates.

Compared to the sequence based expression analysis method, microarray provides a high-throughput tool for simultaneous analysis of expression at the whole transcriptome level. As per the microarray data, 44% flax UGTs showed expression evidence in various tissue types (**Figure 4.3**). Three genes from seed stage and one gene from leaf showed high expression, suggesting possible involvement of these genes in seed and leaf secondary metabolite glycosylation. Microarray data from two contrasting flax varieties, Drakkar and Belinka were also analyzed. Drakkar produces better quality fibres than Belinka, and is more resistant to the fungal pathogen *Fusarium* (Fenart et al., 2010). However, UGT having variety specific expression pattern was not detected. Although, plant UGTs have been reported to be involved in defence mechanism (Langlois-Meurinne et al., 2005), the available microarray data were not generated by exposing the varieties to any pathogen. The difference in expression of the UGTs between the EST and microarray datasets might have resulted from the differences in the number of tissue types, size of each dataset and varieties used for data generation. The EST dataset was larger compared to the microarray dataset, therefore expression evidence might have been obtained for more genes using the EST dataset. Moreover, the long sequence reads of ESTs provide fairly unambiguous evidence of gene expression, compared with the hybridization based microarray data and hence EST profiling could be considered as a more reliable method for transcriptomic analysis as also suggested by Geisler-Lee et al. (2006) and Moreau et al. (2005).

Regarding the 37 unexpressed flax UGTs, it is possible that some or most of these genes might express at very low levels in particular tissue type or express only under specific conditions such as biotic or abiotic stresses. Hence, they might have not been represented in the EST and microarray data as the data were generated from unchallenged libraries. Even in the large Arabidopsis EST collection gathered over several years, only 64.5% of the genes had corresponding ESTs (Rudd, 2003). Absence of an EST for a corresponding gene implies that it is either inactive or expressed at undetectable level in the tissues sampled or that it is a non-functional gene *per se*.

In summary, the present study identified 137 UDP glycosyltransferase (UGT) genes from flax using a conserved signature motif. Phylogenetic analysis of these protein sequences clustered them into 14 major groups (A-N). Expression patterns of these genes were investigated using publicly available expressed sequence tag (EST), microarray data and reverse transcription quantitative real-time PCR (RT-qPCR). Seventy-three per cent of these genes (100 out of 137) showed expression evidence in 15 tissues examined and indicated varied expression profiles. The RT-qPCR results of 10 selected genes were also coherent with the digital expression analysis. Interestingly, five duplicated UGT genes were identified, which showed differential expression in various tissues. Of the seven intron loss/gain positions detected, two intron positions were conserved among most of the UGTs, although a clear relationship about the evolution of these genes could not be established. Comparison of the flax UGTs with orthologs from four other sequenced dicot genomes indicated that seven UGTs were flax diverged.

CHAPTER 5

Summary and future directions

The present work includes systematic profiling of developing seed proteome and characterization of functional metabolic pathways operating during seed development in flax. The objectives of the study were to identify and characterize the regulatory events controlling biosynthesis and accumulation of oil and lignans in developing flax seeds. Enzymatic regulation and regulatory post translational modifications were investigated by proteome analysis. Upstream and terminal regulators of oil and lignin biosynthetic pathways such as microRNAs and UDP-glycosyltransferases were comprehensively studied and their regulatory role in flax seed metabolite accumulation has been thoroughly investigated.

5.1 Developing seed proteome analysis

This study provides a global proteomics perspective of the complex metabolic processes occurring during flax seed development. Among various protein prefractionation methodologies employed to simplify the seed proteome complexity, SDS-PAGE coupled with Nano-LC-ESI-MS^E was found to be the most efficient to cover developing seed proteome. A total of 1716 seed specific non-redundant proteins were identified; of which, several proteins were involved in synthesis of the health promoting compounds. Their temporal expression patterns revealed the developmental stages important for their accumulation. The present study demonstrated that three major carbon assimilatory routes for *de novo* FA synthesis were functional in flax. Further studies of the enzymes involved in these routes might help in understanding the major routes of carbon flow for *de novo* FA synthesis. Lipid profile of developing seed was analysed by using GC-FID and it was observed that, fatty acid accumulation was associated with oleosin and lipoxygenase expression as also confirmed by RT-qPCR. Developing seed also showed the existence of lipid degradation pathways like β-oxidation and glyoxylate pathway, which might be responsible for preferential accumulation of ALA in flax seed. In addition, the study also revealed an unusual diversity of flax seed storage proteins. As in all other dicotyledonous plants, the globulin type of seed storage proteins was highly abundant protein, which was also confirmed by RT-qPCR. As reported in other oilseeds, methionine metabolism was enhanced during the transition from embryogenesis to seed filling, indicating a switch from the active growth phase to the quiescence phase. The enzymes of metabolic pathway for phenylpropanoid and flavonoids class of

compounds were also detected in this study and their importance with respect to accumulation has also been discussed. Cyanogenic glycosides are the main undesired compounds accumulated in the flax seed. The enzymes involved in biosynthesis and degradation were expressed throughout seed development; however, abundance of transcript expression for anabolic enzymes was very high, resulting in accumulation at the end of development process. Moreover, about 81% of the identified proteins had transcriptional evidence in the form of ESTs from the developing seeds of flax. Further, RT-qPCR of selected 19 genes was carried out to understand their roles during seed development. This confirms gene expression at the transcript level and would be helpful for candidate gene isolation and characterization. More in-depth studies of the identified proteins will be useful for better understanding of the complexities of flax seed development.

5.2 Identification and characterization of miRNAs and their target genes

In this study, we computationally identified 116 conserved flax miRNAs belonging to 23 families. They showed characteristics of the miRNA genes like small length (~21 nt), high MFEI index and signatures of RNA polymerase II and were mostly intergenic. The miRNA family distribution was similar to other flax related plant species analyzed. Four gene clusters were identified, which included miRNA169 and miRNA399 as reported in earlier studies in other plant species. Target prediction revealed that majority of the identified miRNA targets were transcription factors. Using RT-qPCR, the expression of 14 of these predicted miRNAs and nine target genes in eight flax tissues was confirmed. Majority of the identified miRNAs were involved in the flower development and repressed the expression of the target genes suggesting that they cause the degradation of the target genes and have organ specific expression. Further characterization of these genes will help to define their roles in flax flower development. Role of the identified miRNAs and their target genes in manipulating various metabolic pathways was also investigated. The candidate genes for overexpresion or silencing have been suggested to increase oil content or change oil composition. However, more in-depth studies need to be conducted to understand the involvement of these genes in various metabolic pathways.

5.3 Phylogenomic analysis of UGT multigene family

The present study identified a large number of UGT genes from the flax genome including few flax diverged ones. These genes were clustered into 14 distinct evolutionary groups based on the phylogenetic analysis. Two new UGT family members not observed in Arabidopsis were identified. Expression patterns of these genes were investigated using publicly available ESTs, microarray data and RT-qPCR. Seventy-three per cent of these genes (100 out of 137) showed expression evidence in 15 tissues examined and indicated varied expression profiles. The RT-qPCR results of 10 selected genes were also coherent with the digital expression analysis. Further, expression analysis suggested that gene *LuUGT74S1* highly expressed in developing seed tissue might glycosylate secoisolariciresinol. Two conserved introns were observed, indicating evolution of flax UGTs from two lineages. Intron architecture analysis supported the "Intron late" theory of intron evolution, indicating that introns were already present in the UGT genes and during the course of evolution, some of them might have been lost or gained. The phylogenetic tree can be useful for understanding the structure-function relatedness of the UGT family members and might further facilitate their functional analysis. Moreover, this study identified tissue and condition specific repertoire of UGT genes. Based on the phylogenetic and expression analysis, UGT involved in secoisolariciresinol glycosylation has been proposed. This study would facilitate precise selection of candidate genes and their further characterization of substrate specificities and *in planta* functions.

5.4 Future directions

The present study opened up new avenues and suggested various approaches to improve desirable characteristics of flax crop. Large number of proteins were identified and annotated with their putative roles in seed development and biosynthetic pathways and some of these could potentially be used to improve oil content and other agronomical traits using genetic manipulation techniques. Proteome analysis of embryo developing stages such as globular, heart shaped, torpedo and cotyledonary embryo would provide deeper prospective of regulation of various metabolic pathways. Moreover, metabolome analysis of same seed developmental

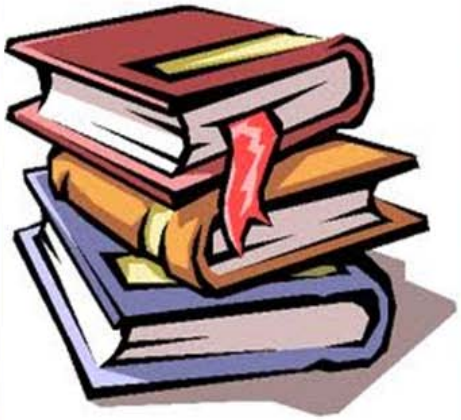
stages would give more comprehensive understanding of identified enzymes and abundance of their metabolites.

In the present study, miRNA mediated down-regulation of target transcripts was supported with RT-qPCR analysis of pre-miRNA and their corresponding target genes. Further confirmation of target transcripts could be performed using a high throughput technique such as degradome sequencing. Another method, RNA ligase mediated rapid amplification of cDNA ends (RLM-RACE) could be employed for confirmation of target transcripts one by one. The present study identified only conserved miRNAs from flax. For identification of novel flax specific miRNAs, small RNA isolation and sequencing needs to be carried out. Small RNA profiling using next generation sequencing of high and low oil/ALA containing varieties would provide better understanding of role of specific miRNA in oil accumulation and would also identify flax specific novel miRNAs. The miRNAs and their targets like transcription factors may be silenced or overexpressed to understand their function in oil biosynthesis and other important traits.

UDP-glycosyltransferase is the terminal regulatory enzyme for secondary metabolite biosynthesis in plants. Silencing of these genes might accumulate their respective substrates in plants. The present study suggested secoisolarisiresinol glycosyltransferase to be involved in glycosylation/biosynthesis of SDG. The function of this gene was independently confirmed by another study. Hence, silencing of this gene might accumulate secoisolarisiresinol, which has more anti-oxidative, anti-carcinogenic and bioavailability properties. In future, the roles of each UGT can be investigated based on their tissue specific expression and individual substrate specificity, *in vitro* cross reactivity, and *in planta* role in defense and xenobiotic detoxification.

The present study suggests various approaches for genetic improvement of a long-time neglected crop such as flax. The candidate genes identified in this study could be used to improve and manipulate desired traits in flax thereby further enhancing its nutritional and commercial importance. This would result in higher demand for the crop and make it more profitable for the flax farmers, improving their livelihoods.

BIBLIOGRAPHY



Adlercreutz H, Mousavi Y, Clark J, Hockerstedt K, Hamalainen E, Wahala K, Makela T, Hase T (1992) Dietary phytoestrogens and cancer: *in vitro* and *in vivo* studies. Journal of Steroid Biochemistry and Molecular Biology **41(3-8):** 331-337

Agrawal GK, Hajduch M, Graham K, Thelen JJ (2008) In-depth investigation of the soybean seed-filling proteome and comparison with a parallel study of rapeseed. Plant Physiology **148:** 504-518

Aksamit-Stachurska A, Korobczak-Sosna A, Kulma A, Szopa J (2008) Glycosyltransferase efficiently controls phenylpropanoid pathway. BMC Biotechnology **8:** 25

Alexandrov NN, Troukhan ME, Brover VV, Tatarinova T, Flavell RB, Feldmann KA (2006) Features of Arabidopsis genes and genome discovered using full-length cDNAs. Plant Molecular Biology **60:** 69-85

Allaby RG, Peterson GW, Merriwether DA, Fu YB (2005) Evidence of the domestication history of flax (*Linum usitatissimum* L.) from genetic diversity of the sad2 locus. Theoretical and Applied Genetics **112:** 58-65

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic Local Alignment Search Tool. Journal of Molecular Biology **215:** 403-410

Altuvia Y, Landgraf P, Lithwick G, Elefant N, Pfeffer S, Aravin A, Brownstein MJ, Tuschl T, Margalit H (2005) Clustering and conservation patterns of human microRNAs. Nucleic Acids Research **33:** 2697-2706

Amarowicz R, Wanasundara PKJPD, Shahidi F (1994) Chromatographic separation of flaxseed phenolics. Nahrung-Food **38:** 520-526

Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, Chen XM, Dreyfuss G, Eddy SR, Griffiths-Jones S, Marshall M, Matzke M, Ruvkun G, Tuschl T (2003) A uniform system for microRNA annotation. RNA **9:** 277-279

Andersen MD, Busk PK, Svendsen I, Moller BL (2000) Cytochromes P-450 from cassava (*Manihot esculenta* Crantz) catalyzing the first steps in the biosynthesis of the cyanogenic glucosides linamarin and lotaustralin - Cloning, functional expression in Pichia pastoris, and substrate specificity of the isolated recombinant enzymes. Journal of Biological Chemistry **275:** 1966-1975

Arnaud D, Dejardin A, Leple JC, Lesage-Descauses MC, Boizot N, Villar M, Benedetti H, Pilate G (2012) Expression analysis of LIM gene family in poplar, toward an updated phylogenetic classification. BMC Res Notes **5:** 102

Aukerman MJ, Sakai H (2003) Regulation of flowering time and floral organ identity by a microRNA and its APETALA2-like target genes. Plant Cell **15:** 2730-2741

Axtell MJ, Bartel DP (2005) Antiquity of microRNAs and their targets in land plants. Plant Cell **17:** 1658-1673

Ayad AA (2010) Characterization and properties of flaxseed protein fractions. McGill University, Montreal

Babu PR, Rao KV, Reddy VD (2013) Structural organization and classification of cytochrome P450 genes in flax (*Linum usitatissimum* L.). Gene **513:** 156-162

Bak S, Kahn RA, Nielsen HL, Moller BL, Halkier BA (1998) Cloning of three A-type cytochromes p450, CYP71E1, CYP98, and CYP99 from *Sorghum bicolor* (L.) Moench by a PCR approach and identification by expression in *Escherichia coli* of CYP71E1 as a multifunctional cytochrome p450 in the biosynthesis of the cyanogenic glucoside dhurrin. Plant Molecular Biology **36:** 393-405

Banik M, Duguid S, Cloutier S (2011) Transcript profiling and gene characterization of three fatty acid desaturase genes in high, moderate, and low linolenic acid genotypes of flax (*Linum usitatissimum* L.) and their role in linolenic acid accumulation. Genome **54:** 471-483

Barakat A, Wall PK, Diloreto S, Depamphilis CW, Carlson JE (2007) Conservation and divergence of microRNAs in Populus. BMC Genomics **8:** 481

Bartel DP (2004) MicroRNAs: Genomics, biogenesis, mechanism, and function. Cell **116:** 281-297

Baud S, Boutin JP, Miquel M, Lepiniec L, Rochat C (2002) An integrated overview of seed development in *Arabidopsis thaliana* ecotype WS. Plant Physiology and Biochemistry **40:** 151-160

BeMiller JN (1973) Quince seed, psyllium seed, flax seed and okra gums. *In* RL Whistler, JN BeMiller, eds, In Industrial Gums. Academic Press, New York, pp 331-337

Bentwich I (2005) Prediction and validation of microRNAs and their targets. FEBS Letters **579:** 5904-5910

Bhatty RS, Cherdkiatgumchai P (1990) Compositional analysis of laboratory prepared and commercial samples of Linseed meal and of hull isolated from flax. Journal of the American Oil Chemists Society **67:** 79-84

Bickel C, Gadani S, Lukacs M, Cullis C (2011) SSR markers developed for genetic mapping in flax (*Linum usitatissimum* L.). Research and Reports in Biology **2011:2:** 23-29

Bonnet E, Wuyts J, Rouze P, Van de Peer Y (2004) Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. Bioinformatics **20:** 2911-2917

Borriello SP, Setchell KDR, Axelson M, Lawson AM (1985) Production and metabolism of lignans by the human fecal flora. Journal of Applied Bacteriology **58:** 37-43

Bowles D (2002) A multigene family of glycosyltransferases in a model plant, *Arabidopsis thaliana*. Biochemical Society Transactions **30:** 301-306

Bowles D, Isayenkova J, Lim EK, Poppenberger B (2005) Glycosyltransferases: managers of small molecules. Current Opinion in Plant Biology **8:** 254-263

Bradford MM (1976) Rapid and sensitive method for quantitation of microgram quantities of protein utilizing principle of protein-dye binding. Analytical Biochemistry **72:** 248-254

Brennecke J, Stark A, Russell RB, Cohen SM (2005) Principles of MicroRNA-target recognition. Plos Biology **3:** 404-418

Calvino M, Bruggmann R, Messing J (2011) Characterization of the small RNA component of the transcriptome from grain and sweet sorghum stems. BMC Genomics **12:** 356

Cartwright AM, Lim EK, Kleanthous C, Bowles DJ (2008) A kinetic analysis of regiospecific glucosylation by two glycosyltransferases of *Arabidopsis thaliana* - Domain swapping to introduce new activities. Journal of Biological Chemistry **283:** 15724-15731

Casey R, Domoney C, Forster C, Hedley C, Hitchin E, Wang T (1998) The effect of modifying carbohydrate metabolism on seed protein gene expression in peas. Journal of Plant Physiology **152:** 636-640

Chen F, Mackey AJ, Vermunt JK, Roos DS (2007) Assessing performance of orthology detection strategies applied to eukaryotic genomes. Plos One **2:** 4

Chen MJ, Mooney BP, Hajduch M, Joshi T, Zhou MY, Xu D, Thelen JJ (2009) System analysis of an Arabidopsis mutant altered in de novo fatty acid synthesis reveals diverse changes in seed composition and metabolism. Plant Physiology **150:** 27-41

Chia TYP, Pike MJ, Rawsthorne S (2005) Storage oil breakdown during embryo development of *Brassica napus* (L.). Journal of Experimental Botany **56:** 1285-1296

Cloutier S, Niu Z, Datla R, Duguid S (2009) Development and analysis of EST-SSRs for flax (*Linum usitatissimum* L.). Theoretical and Applied Genetics **119:** 53-63

Cloutier S, Ragupathy R, Miranda E, Radovanovic N, Reimer E, Walichnowski A, Ward K, Rowland G, Duguid S, Banik M (2012) Integrated consensus genetic and physical maps of flax (*Linum usitatissimum* L.). Theoretical and Applied Genetics **125:** 1783-1795

Cloutier S, Ragupathy R, Niu ZX, Duguid S (2011) SSR-based linkage map of flax (*Linum usitatissimum* L.) and mapping of QTLs underlying fatty acid composition traits. Molecular Breeding **28:** 437-451

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics **21:** 3674-3676

Cui W, Mazza G (1996) Physicochemical characteristics of flaxseed gum. Food Research International **29:** 397-402

Cui X, Xu SM, Mu DS, Yang ZM (2009) Genomic analysis of rice microRNA promoters and clusters. Gene **431:** 61-66

Cullis CA (1981) DNA-sequence organization in the flax genome. Biochimica Et Biophysica Acta **652:** 1-15

Czemplik M, Kulma A, Bazela K, Szopa J (2012) The biomedical potential of genetically modified flax seeds overexpressing the glucosyltransferase gene. BMC Complementary and Alternative Medicine **12:** 251

Dabrowski KJ, Sosulski FW (1984) Composition of free and hydrolyzable phenolic acids in defatted flours of ten oilseeds. Journal of Agricultural and Food Chemistry **32:** 128–130

Dai XB, Zhao PX (2011) psRNATarget: a plant small RNA target analysis server. Nucleic Acids Research **39:** 155-159

Daskalchuk T, Ahiahonu P, Heath D, Yamazaki Y (2006) Biotechnology in Agriculture and Forestry. *In* K Saito, RA Dixon, L Willmitzer, eds, Plant Metabolomics, Vol 57. Springer-Verlag, Berlin Heidelberg

Daun JK, DeClercq DR (1994) Sixty years of Canadian flaxseed quality surveys at the Grain Research Laboratory. *In* Proceeding of the Flax Institute of the United States, Fargo, ND US., pp 192-200

Davin LB, Lewis NG (2000) Dirigent proteins and dirigent sites explain the mystery of specificity of radical precursor coupling in lignan and lignin biosynthesis. Plant Physiology **123:** 453-462

Day A, Fenart S, Neutelings G, Hawkins S, Rolando C, Tokarski C (2013) Identification of cell wall proteins in the flax (*Linum usitatissimum*) stem. Proteomics **13:** 812-825

DeClercq DR, Daun JK (2002) Quality of Western Canadian Flaxseed. Canadian Grain Commission. *In* http://www.grainscanada.gc.ca/flax-lin/harvest-recolte/2011/hqf11-qrl11-eng.htm,

Delange F, Ekpechi LO, Rosling H (1994) Cassava cyanogenesis and iodine defciency disorders. Acta Horticulturae **375:** 289-293

Deng W, Nickle DC, Learn GH, Maust B, Mullins JI (2007) ViroBLAST: a stand-alone BLAST web server for flexible queries of multiple databases and user's datasets. Bioinformatics **23:** 2334-2336

Deng X, Long SH, He DF, Li XA, Wang YF, Hao DM, Qiu CS, X.B. C (2011) Isolation and characterization of polymorphic microsatellite markers from flax (*Linum usitatissimum* L.). African Journal of Biotechnology **10:** 734-739

Dennis DT, Miernyk JA (1982) Compartmentation of non-photosynthetic carbohydrate metabolism. Annual Review of Plant Physiology and Plant Molecular Biology **33:** 27-50

Denyer K, Smith AM (1988) The capacity of plastids from developing Pea cotyledons to synthesize acetyl CoA. Planta **173:** 172-182

Dev DK, Quensel E (1989) Functional properties of linseed protein products containing different levels of mucilage in selected food systems. Journal of Food Science **54:** 183-186

Diederichsen A, Hammer K (1995) Variation of cultivated Flax (*Linum usitatissimum* L subsp usitatissimum) and Its wild progenitor pale flax (subsp Angustifolium (Huds) Thell). Genetic Resources and Crop Evolution **42:** 263-272

Dietz KJ (1987) Control function of hexosemonophosphate isomerase and phosphoglucomutase in starch synthesis in leaves. *In* J Biggins, ed, Proceedings of the 7th International Congress on Photosynthesis, Vol 3. Martinus Nijhoff Dr. W.

Junk Publications, The Hague, The Netherlands, pp 329–332

Dribnenki JCP, Green AG (1995) LinolaTM '947' low linolenic acid flax. Canadian Journal of Plant Science **75:** 201-202

Du Z, Zhou X, Ling Y, Zhang ZH, Su Z (2010) AgriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Research **38:** W64-W70

Eastmond PJ, Rawsthorne S (2000) Coordinate changes in carbon partitioning and plastidial metabolism during the development of oilseed rape embryo. Plant Physiology **122:** 767-774

Eaton SB, Konner M (1985) Paleolithic nutrition. A consideration of its nature and current implications. New England Journal of Medicine **312:** 283-289

Eliasson C, Kamal-Eldin A, Andersson R, Aman P (2003) High-performance liquid chromatographic analysis of secoisolariciresinol diglucoside and hydroxycinnamic acid glucosides in flaxseed by alkaline extraction. Journal of Chromatography A **1012:** 151-159

Fang JJ, Ramsay A, Paetz C, Tatsis EC, Renouard S, Hano C, Grand E, Fliniaux O, Roscher A, Mesnard F, Schneider B (2013) Concentration kinetics of secoisolariciresinol diglucoside and its biosynthetic precursor coniferin in developing flaxseed. Phytochemical Analysis **24:** 41-46

Fang Y, Robinson DP, Foster LJ (2010) Quantitative analysis of proteome coverage and recovery rates for upstream fractionation methods in proteomics. Journal of Proteome Research **9:** 1902-1912

Fenart S, Ndong YPA, Duarte J, Riviere N, Wilmer J, van Wuytswinkel O, Lucau A, Cariou E, Neutelings G, Gutierrez L, Chabbert B, Guillot X, Tavernier R, Hawkins S, Thomasset B (2010) Development and validation of a flax (*Linum usitatissimum* L.) gene expression oligo microarray. BMC Genomics **11:** 592

Ford JD, Huang KS, Wang HB, Davin LB, Lewis NG (2001) Biosynthetic pathway to the cancer chemopreventive secoisolariciresinol diglucoside-hydroxymethyl glutaryl ester-linked lignan oligomers in flax (*Linum usitatissimum*) seed. Journal of Natural Products **64:** 1388-1397

Frazier TP, Xie FL, Freistaedter A, Burklew CE, Zhang BH (2010) Identification and characterization of microRNAs and their target genes in tobacco (*Nicotiana tabacum*). Planta **232:** 1289-1308

Frehner M, Scalet M, Conn EE (1990) Pattern of the cyanide-potential in developing fruits - Implications for plants accumulating cyanogenic monoglucosides (*Phaseolus lunatus*) or cyanogenic diglucosides in their seeds (*Linum usitatissimum, Prunus amygdalus*). Plant Physiology **94:** 28-34

Fu YB (2005) Geographic patterns of RAPD variation in cultivated flax. Crop Science **45:** 1084-1091

Fu YB (2011) Genetic evidence for early flax domestication with capsular dehiscence. Genetic Resources and Crop Evolution **58:** 1119-1128

Fu YB, Allaby RG (2010) Phylogenetic network of Linum species as revealed by non-coding chloroplast DNA sequences. Genetic Resources and Crop Evolution **57:** 667-677

Fu YB, Diederichsen A, Richards KW, Peterson G (2002) Genetic diversity within a range of cultivars and landraces of flax (*Linum usitatissimum* L.) as revealed by RAPDs. Genetic Resources and Crop Evolution **49:** 167-174

Fu YB, Peterson GW (2010) Characterization of expressed sequence tag-derived simple sequence repeat markers for 17 Linum species. Botany-Botanique **88:** 537-543

Gallardo K, Firnhaber C, Zuber H, Hericher D, Belghazi M, Henry C, Kuster H, Thompson R (2007) A combined proteome and transcriptome analysis of developing *Medicago truncatula* seeds. Molecular & Cellular Proteomics **6:** 2165-2179

Gallardo K, Job C, Groot SPC, Puype M, Demol H, Vandekerckhove J, Job D (2001) Proteomic analysis of Arabidopsis seed germination and priming. Plant Physiology **126:** 835-848

Gallardo K, Job C, Groot SPC, Puype M, Demol H, Vandekerckhove J, Job D (2002) Importance of methionine biosynthesis for Arabidopsis seed germination and seedling growth. Physiologia Plantarum **116:** 238-247

Gallardo K, Le Signor C, Vandekerckhove J, Thompson RD, Burstin J (2003) Proteomics of Medicago truncatula seed development establishes the time frame of diverse metabolic processes related to reserve accumulation. Plant Physiology **133:** 664-682

Ganjewala D, Kumar S, Devi SA, Ambika K (2010) Advances in cyanogenic glycosides biosynthesis and analyses in plants: A review. Acta Biologica Szegediensis **54:** 1-14

Geisler-Lee J, Geisler M, Coutinho PM, Segerman B, Nishikubo N, Takahashi J, Aspeborg H, Djerbi S, Master E, Andersson-Gunneras S, Sundberg B, Karpinski S, Teeri TT, Kleczkowski LA, Henrissat B, Mellerowicz EJ (2006) Poplar carbohydrate-active enzymes. Gene identification and expression analyses. Plant Physiology **140:** 946-962

Germain V, Rylott EL, Larson TR, Sherson SM, Bechtold N, Carde JP, Bryce JH, Graham IA, Smith SM (2001) Requirement for 3-ketoacyl-CoA thiolase-2 in peroxisome development, fatty acid beta-oxidation and breakdown of triacylglycerol in lipid bodies of Arabidopsis seedlings. Plant Journal **28:** 1-12

Gill KS, Yermanos DM (1967) Cytogenetic studies on genus *Linum* L. hybrids among taxa with 15 as haploid chromosome number. Crop Science **7:** 623-627

Gonzalez LG, Deyholos MK (2012) Identification, characterization and distribution of transposable elements in the flax (*Linum usitatissimum* L.) genome. BMC Genomics **13:** 644

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Research **40:** D1178-1186

Gou JY, Felippes FF, Liu CJ, Weigel D, Wang JW (2011) Negative Regulation of Anthocyanin Biosynthesis in Arabidopsis by a miR156-Targeted SPL Transcription Factor. Plant Cell **23:** 1512-1522

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2008) miRBase: tools for microRNA genomics. Nucleic Acids Research **36:** D154-D158

Gutierrez L, Conejero G, Castelain M, Guenin S, Verdeil JL, Thomasset B, Van Wuytswinkel O (2006) Identification of new gene expression regulators specifically expressed during plant seed maturation. Journal of Experimental Botany **57:** 1919-1932

Haberer G, Hindemitt T, Meyers BC, Mayer KFX (2004) Transcriptional similarities, dissimilarities, and conservation of cis-elements in duplicated genes of arabidopsis. Plant Physiology **136:** 3009-3022

Hajduch M, Casteel JE, Hurrelmeyer KE, Song Z, Agrawal GK, Thelen JJ (2006) Proteomic analysis of seed filling in *Brassica napus*. Developmental characterization of metabolic isozymes using high-resolution two-dimensional gel electrophoresis. Plant Physiology **141:** 1159-1159

Hajduch M, Ganapathy A, Stein JW, Thelen JJ (2005) A systematic proteomic study of seed filling in soybean. Establishment of high-resolution two-dimensional reference maps, expression profiles, and an interactive proteome database. Plant Physiology **137:** 1397-1419

Hajduch M, Hearne LB, Miernyk JA, Casteel JE, Joshi T, Agrawal GK, Song Z, Zhou MY, Xu D, Thelen JJ (2010) Systems analysis of seed filling in arabidopsis: Using general linear modeling to assess concordance of transcript and protein expression. Plant Physiology **152:** 2078-2087

Hajduch M, Matusova R, Houston NL, Thelen JJ (2011) Comparative proteomics of seed maturation in oilseeds reveals differences in intermediary metabolism. Proteomics **11:** 1619-1629

Hall C, Tulbek MC, Xu TY (2006) Flaxseed. *In* Advances in Food and Nutrition Research, Vol 51. Academic Press, pp 1-97

Hammer K (1986) Linaceae. *In* J Schultze-Motel, ed, Rudolf Mansfeld directory in agricultural and horticultural crops, Vol 2. Akademie-Verlag, Berlin, pp 710-713

Hano C, Laine E, Martin I, Fliniaux O, Legrand B, Gutierrez L, Arroo RRJ, Mesnard F, Lamblin F (2006) Pinoresinol-lariciresinol reductase gene expression and secoisolariciresinol diglucoside accumulation in developing flax (Linum usitatissimum) seeds. Planta **224:** 1291-1301

Hansen KS, Kristensen C, Tattersall DB, Jones PR, Olsen CE, Bak S, Moller BL (2003) The in vitro substrate regiospecificity of recombinant UGT85B1, the cyanohydrin glucosyltransferase from *Sorghum bicolor*. Phytochemistry **64:** 143-151

Hanson AD, Roje S (2001) One-carbon metabolism in higher plants. Annual Review of Plant Physiology and Plant Molecular Biology **52:** 119-137

Harlan JR (1971) Agricultural origins: centers and noncenters. Science **174:** 468-474

Harwood JL (1988) Fatty acid metabolism Annual Review of Plant Physiology and Plant Molecular Biology **39:** 101-138

Hattenbach A, Heineke D (1999) On the role of chloroplastic phosphoglucomutase in the regulation of starch turn over. Planta **207:** 527–532

Haynes PA, Fripp N, Aebersold R (1998) Identification of gel-separated proteins by liquid chromatography electrospray tandem mass spectrometry: Comparison of methods and their limitations. Electrophoresis **19:** 939-945

Hays DB, Reid DM, Yeung EC, Pharis RP (2000) Role of ethylene in cotyledon development of microspore-derived embryos of *Brassica napus*. Journal of Experimental Botany **51:** 1851-1859

Heinz E (1993) Biosynthesis of polyunsaturated fatty acids. *In* TS Moore, ed, Lipid Metabolism in Plants. CRC Press, Boca Raton, FL, pp 33-90

Helbæk H (1959) Notes on the evolution and history of Linum. *In* KUML. Stiftsbogtrykkeri, pp 103-129

Higo K, Ugawa Y, Iwamoto M, Korenaga T (1999) Plant cis-acting regulatory DNA elements (PLACE) database. Nucleic Acids Research **27:** 297-300

Hofacker IL (2003) Vienna RNA secondary structure server. Nucleic Acids Research **31:** 3429-3431

Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K (2007) WoLF PSORT: protein localization predictor. Nucleic Acids Research **35:** W585-W587

Hotte NSC, Deyholos MK (2008) A flax fibre proteome: identification of proteins enriched in bast fibres. BMC Plant Biology **8:** 52

Hou BK, Lim EK, Higgins GS, Bowles DJ (2004) N-glucosylation of cytokinins by glycosyltransferases of *Arabidopsis thaliana*. Journal of Biological Chemistry **279:** 47822-47832

Houston NL, Hajduch M, Thelen JJ (2009) Quantitative proteomics of seed filling in castor: comparison with soybean and rapeseed reveals differences between photosynthetic and nonphotosynthetic seed metabolism. Plant Physiology **151:** 857-868

Hradilova J, Rehulka P, Rehulkova H, Vrbova M, Griga M, Brzobohaty B (2010) Comparative analysis of proteomic changes in contrasting flax cultivars upon cadmium exposure. Electrophoresis **31:** 421-431

Huang AHC (1992) Oil bodies and oleosins in seeds. Annual Review of Plant Physiology and Plant Molecular Biology **43:** 177-200

Huang Y, Niu BF, Gao Y, Fu LM, Li WZ (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics **26:** 680-682

Hudson BJF, Lewis JI (1983) Polyhydroxy flavonoid antioxidants for edible oils. Structural criteria for activity. Food Chemistry **10:** 47-55

Huis R, Hawkins S, Neutelings G (2010) Selection of reference genes for quantitative gene expression normalization in flax (*Linum usitatissimum* L.). BMC

Plant Biology **10:** 71

Hutchins AM, Slavin JL (2003) EVects of flaxseed on sex hormone metabolism. *In* LU Thompson, SC Cunnane, eds, Flaxseed in human nutrition. AOCS Press, Champaign, IL, pp 126–149

Hutton D, Stumpf PK (1969) Fat metabolism in higher plants XXXVII. Characterization of the beta -oxidation systems from maturing and germinating castor bean seeds. Plant physiology **44:** 508–516

Hyun TK, Kumar D, Cho YY, Hyun HN, Kim JS (2013) Computational identification and phylogenetic analysis of the oil-body structural proteins, oleosin and caleosin, in castor bean and flax. Gene **515:** 454-460

Jaeken J, Matthijs G (2001) Congenital disorders of glycosylation. Annual Review of Genomics and Human Genetics **2:** 129-151

Jaworski JG, Clough RC, Barnum SR (1989) A cerulenin insensitive short chain 3-ketoacyl-acyl carrier protein synthase in *Spinacia oleracea* leaves. Plant Physiology **90:** 41-44

Johnsson P, Peerlkamp N, Kamal-Eldin A, Andersson RE, Andersson R, Lundgren LN, Aman P (2002) Polymeric fractions containing phenol glucosides in flaxseed. Food Chemistry **76:** 207-212

Jones-Rhoades MW, Bartel DP, Bartel B (2006) MicroRNAs and their regulatory roles in plants. Annual Review of Plant Biology **57:** 19-53

Jones P, Vogt T (2001) Glycosyltransferases in secondary plant metabolism: tranquilizers and stimulant controllers. Planta **213:** 164-174

Joshi K, Lad S, Kale M, Patwardhan B, Mahadik SP, Patni B, Chaudhary A, Bhave S, Pandit A (2006) Supplementation with flax oil and vitamin C improves the outcome of Attention Deficit Hyperactivity Disorder (ADHD). Prostaglandins Leukotrienes and Essential Fatty Acids **74:** 17-21

Journet EP, Douce R (1985) Enzymic capacities of purified cauliflower bud plastids for lipid synthesis and carbohydrate metabolism. Plant Physiology **79:** 458-467

Jung JH, Seo PJ, Kang SK, Park CM (2011) miR172 signals are incorporated into the miR156 signaling pathway at the SPL3/4/5 genes in Arabidopsis developmental transitions. Plant Molecular Biology **76:** 35-45

Jung KH, An GH, Ronald PC (2008) Towards a better bowl of rice: assigning function to tens of thousands of rice genes. Nature Reviews Genetics **9:** 91-101

Kale SM, Pardeshi VC, Barvkar VT, Gupta VS, Kadoo NY (2013) Genome-wide identification and characterization of nucleotide binding site leucine-rich repeat genes in linseed reveal distinct patterns of gene structure. Genome **56:** 1-9

Kale SM, Pardeshi VC, Kadoo NY, Ghorpade PB, Jana MM, Gupta VS (2012) Development of genomic simple sequence repeat markers for linseed using next-generation sequencing technology. Molecular Breeding **30:** 597-606

Kangas N, Niehaus M, Hall CA (2006) Utlization of flaxseed gum in dairy foods. *In* JF Carter, ed, Proceedings of the 61st Flax Institute of the US, North Dakota State University, Fargo, ND, pp 205–212

Kitts DD, Yuan YV, Wijewickreme AN, Thompson LU (1999) Antioxidant activity of the flaxseed lignan secoisolariciresinol diglycoside and its mammalian lignan metabolites enterodiol and enterolactone. Molecular and Cellular Biochemistry **202:** 91-100

Klubicova K, Bercak M, Danchenko M, Skultety L, Rashydov NM, Berezhna VV, Miernyk JA, Hajduch M (2011a) Agricultural recovery of a formerly radioactive area: I. Establishment of high-resolution quantitative protein map of mature flax seeds harvested from the remediated Chernobyl area. Phytochemistry **72:** 1308-1315

Klubicova K, Danchenko M, Skultety L, Berezhna VV, Hricova A, Rashydav NM, Hajduch M (2011b) Agricultural recovery of a formerly radioactive area: II. Systematic proteomic characterization of flax seed development in the remediated Chernobyl area. Journal of Proteomics **74:** 1378-1384

Klubicova K, Danchenko M, Skultety L, Miernyk JA, Rashydov NM, Berezhna VV, Pret'Ova A, Hajduch M (2010) Proteomics analysis of Flax grown in chernobyl area suggests limited effect of contaminated environment on seed proteome. Environmental Science & Technology **44:** 6940-6946

Kosinska A, Penkacik K, Wiczkowski W, Amarowicz R (2011) Presence of caffeic acid in flaxseed lignan macromolecule. Plant Foods for Human Nutrition **66:** 270-274

Kozlowska H, Zadernowski R, Sosulski FW (1983) Phenolic acids in oilseed flours. Nahrung-Food **27:** 449-453

Kratzer FH, Vohra P (1996) The use of flaxseed as a poultry feedstuff. *In*. University of California, Avian Sciences Department,, Davis, CA 95616

Kraushofer T, Sontag G (2002) Determination of matairesinol in flax seed by HPLC with coulometric electrode array detection. Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences **777:** 61-66

Kumar S, You FM, Cloutier S (2012) Genome wide SNP discovery in flax through next generation sequencing of reduced representation libraries. BMC Genomics **13:** 684

Langlois-Meurinne M, Gachon CMM, Saindrenan P (2005) Pathogen-responsive expression of glycosyltransferase genes UGT73B3 and UGT73B5 is necessary for resistance to *Pseudomonas syringae pv* tomato in Arabidopsis. *Plant Physiology* **139:** 1890-1901

Lee J, Garrett WM, Cooper B (2007) Shotgun proteomic analysis of *Arabidopsis thaliana* leaves. Journal of Separation Science **30:** 2225-2230

Lee RC, Feinbaum RL, Ambros V (1993) The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell **75:** 843-854

Lee RE, Manthey FA, Hall CA (2004) Content and stability of hexane extractable lipid at various steps of producing macaroni containing ground flaxseed. Journal of Food Processing and Preservation **28:** 133-144

Li GZ, Vissers JPC, Silva JC, Golick D, Gorenstein MV, Geromanos SJ (2009) Database searching and accounting of multiplexed precursor and product ion spectra from the data independent analysis of simple and complex peptide mixtures. Proteomics **9:** 1696-1719

Li Y, Baldauf S, Lim EK, Bowles DJ (2001) Phylogenetic analysis of the UDP-glycosyltransferase multigene family of *Arabidopsis thaliana*. Journal of Biological Chemistry **276:** 4338-4343

Lim EK, Baldauf S, Li Y, Elias L, Worrall D, Spencer SP, Jackson RG, Taguchi G, Ross J, Bowles DJ (2003) Evolution of substrate recognition across a multigene family of glycosyltransferases in Arabidopsis. Glycobiology **13:** 139-145

Lim EK, Doucet CJ, Li Y, Elias L, Worrall D, Spencer SP, Ross J, Bowles DJ (2002) The activity of Arabidopsis glycosyltransferases toward salicylic acid, 4-hydroxybenzoic acid, and other benzoates. Journal of Biological Chemistry **277:** 586-592

Lu C, Meyers BC, Green PJ (2007) Construction of small RNA cDNA libraries for deep sequencing. Methods **43:** 110-117

Mackenzie PI, Owens IS, Burchell B, Bock KW, Bairoch A, Belanger A, FournelGigleux S, Green M, Hum DW, Iyanagi T, Lancet D, Louisot P, Magdalou J, Chowdhury JR, Ritter JK, Schachter H, Tephly TR, Tipton KF, Nebert DW (1997) The UDP glycosyltransferase gene superfamily: Recommended nomenclature update based on evolutionary divergence. Pharmacogenetics **7:** 255-269

Maor R, Sitbon E, Avniel A (2011) Compositions and methods for enhancing oil content in plants. *In* USPTO, ed. Rosetta Genomics Ltd., Israel

Martin RC, Mok MC, Habben JE, Mok DWS (2001) A maize cytokinin gene encoding an O-glucosyltransferase specific to cis-zeatin. Proceedings of the National Academy of Sciences of the United States of America **98:** 5922-5926

Masoudi-Nejad A, Tonomura K, Kawashima S, Moriya Y, Suzuki M, Itoh M, Kanehisa M, Endo T, Goto S (2006) EGassembler: online bioinformatics service for large-scale processing, clustering and assembling ESTs and genomic DNA fragments. Nucleic Acids Research **34:** W459-W462

Mathews DH, Sabina J, Zuker M, Turner DH (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. Journal of Molecular Biology **288:** 911-940

Mazza G, Biliaderis CG (1989) Functional properties of flaxseed mucilage. Journal of Food Science **54:** 1302-1305

McCarthy FM, Wang N, Magee GB, Nanduri B, Lawrence ML, Camon EB, Barrell DG, Hill DP, Dolan ME, Williams WP, Luthe DS, Bridges SM, Burgess SC (2006) AgBase: a functional genomics resource for agriculture. BMC Genomics **7:** 229

Megraw M, Baev V, Rusinov V, Jensen ST, Kalantidis K, Hatzigeorgiou AG (2006) MicroRNA promoter element discovery in Arabidopsis. RNA **12:** 1612-1619

Meyers BC, Axtell MJ, Bartel B, Bartel DP, Baulcombe D, Bowman JL, Cao X, Carrington JC, Chen XM, Green PJ, Griffiths-Jones S, Jacobsen SE, Mallory AC, Martienssen RA, Poethig RS, Qi YJ, Vaucheret H, Voinnet O, Watanabe Y, Weigel D, Zhui JK (2008) Criteria for annotation of plant microRNAs. Plant Cell **20:** 3186-3190

Miernyk JA, Trelease RN (1981) Control of enzyme-activities in cotton cotyledons during maturation and germination: IV. Beta-oxidation. Plant Physiology **67:** 341-346

Minas IS, Tanou G, Belghazi M, Job D, Manganaris GA, Molassiotis A, Vasilakakis M (2012) Physiological and proteomic approaches to address the active role of ozone in kiwifruit post-harvest ripening. Journal of Experimental Botany **63:** 2449-2464

Modolo LV, Blount JW, Achnine L, Naoumkina MA, Wang XQ, Dixon RA (2007) A functional genomics approach to (iso)flavonoid glycosylation in the model legume *Medicago truncatula*. Plant Molecular Biology **64:** 499-518

Moire L, Rezzonico E, Goepfert S, Poirier Y (2004) Impact of unusual fatty acid synthesis on futile cycling through beta-oxidation and on gene expression in transgenic plants. Plant Physiology **134:** 432-442

Moreau C, Aksenov N, Lorenzo MG, Segerman B, Funk C, Nilsson P, Jansson S, Tuominen H (2005) A genomic approach to investigate developmental cell death in woody tissues of Populus trees. Genome Biology **6:** 4

Mori TA (2006) Omega-3 fatty acids and hypertension in humans. Clinical and Experimental Pharmacology and Physiology **33:** 842-846

Morris DH (2007) Flax - A health and nutrition primer. *In*, Ed Fourth Edition. Flax Council of Canada

Moss TY, Cullis CA (2012) Computational prediction of candidate microRNAs and their targets from the completed *Linum ussitatissimum* genome and EST Database. Journal of Nucleic Acids Investigation **3:e2**

Moxon S, Schwach F, Dalmay T, MacLean D, Studholme DJ, Moulton V (2008) A toolkit for analysing large-scale plant small RNA datasets. Bioinformatics **24:** 2252-2253

Mozaffarian D (2005) Does alpha-linolenic acid intake reduce the risk of coronary heart disease? A review of the evidence. Alternative Therapies, Health and Medicine **11:** 24-30; quiz 31, 79

Narisawa T, Fukaura Y, Yazawa K, Ishikawa C, Isoda Y, Nishizawa Y (1994) Colon-cancer prevention with a small amount of dietary perilla oil high in alpha-linolenic acid in an nimal-model. Cancer **73:** 2069-2075

Neutelings G, Fenart S, Lucau-Danila A, Hawkins S (2012) Identification and characterization of miRNAs and their potential targets in flax. Journal Plant Physiology **169:** 1754-1766

Neutelings G, Fenart S, Lucau-Danila A, Hawkins S (2012) Identification and characterization of miRNAs and their potential targets in flax. Journal of Plant

Physiology **169:** 1754-1766

Niedzwiedz-Siegien I (1998) Cyanogenic glucosides in *Linum usitatissimum*. Phytochemistry **49:** 59-63

Noguchi A, Fukui Y, Iuchi-Okada A, Kakutani S, Satake H, Iwashita T, Nakao M, Umezawa T, Ono E (2008) Sequential glucosylation of a furofuran lignan, (+)-sesarninol, by Sesamum indicum UGT71A9 and UGT94D1 glucosyltransferases. Plant Journal **54:** 415-427

Ohlrogge J, Browse J (1995) Lipid biosynthesis. Plant Cell **7:** 957-970

Ono E, Fukuchi-Mizutani M, Nakamura N, Fukui Y, Yonekura-Sakakibara K, Yamaguchi M, Nakayama T, Tanaka T, Kusumi T, Tanaka Y (2006) Yellow flowers generated by expression of the aurone biosynthetic pathway. Proceedings of the National Academy of Sciences of the United States of America **103:** 11075-11080

Oomah BD (2001) Flaxseed as a functional food source. Journal of the Science of Food and Agriculture **81:** 889-894

Oomah BD, Mazza G (1993) Flaxseed proteins - a review. Food Chemistry **48:** 109-114

Osborne TB (1924) The vegetable proteins, Ed 2. Longsman-Green, London

Osmani SA, Bak S, Moller BL (2009) Substrate specificity of plant UDP-dependent glycosyltransferases predicted from crystal structures and homology modeling. Phytochemistry **70:** 325-347

Palmer JD, Logsdon JMJ (1991) The recent origins of introns. *Current Opinions in Genetics and Development* **1:** 470-477

Pan A, Yu DX, Demark-Wahnefried W, Franco OH, Lin X (2009) Meta-analysis of the effects of flaxseed interventions on blood lipids. American Journal of Clinical Nutrition **90:** 288-297

Paquette S, Moller BL, Bak S (2003) On the origin of family 1 plant glycosyltransferases. Phytochemistry **62:** 399-413

Pellizzon MA, Billheimer JT, Bloedon LT, Szapary PO, Rader DJ (2007) Flaxseed reduces plasma cholesterol levels in hypercholesterolemic mouse models. Journal of the American College of Nutrition **26:** 66-75

Periappuram C, Steinhauer L, Barton DL, Taylor DC, Chatson B, Zou J (2000) The plastidic phosphoglucomutase from Arabidopsis. A reversible enzyme reaction with an important role in metabolic control. Plant Physiology **122:** 1193-1200

Plaxton WC, Podesta FE (2006) The functional organization and control of plant respiration. Critical Reviews in Plant Sciences **25:** 159-198

Pohjanheimo TA, Hakala MA, Tahvonen RL, Salminen SJ, Kallio HP (2006) Flaxseed in breadmaking: Effects on sensory quality, aging, and composition of bakery products. Journal of Food Science **71:** S343-S348

Porta H, Rocha-Sosa M (2002) Plant lipoxygenases. Physiological and molecular features. Plant Physiology **130:** 15-21

Post-Beittenmiller D, Jaworski JG, Ohlrogge JB (1991) In vivo pools of free and acylated acyl carrier proteins in spinach. Evidence for sites of regulation of fatty acid biosynthesis. Journal of Biological Chemistry **266:** 1858-1865

Post-Beittenmiller D, Roughan G, Ohlrogge JB (1992) Regulation of plant fatty acid biosynthesis: analysis of acyl-coenzyme a and acyl-acyl carrier protein substrate pools in spinach and pea chloroplasts. Plant Physiology **100:** 923-930

Prakash V, Rao MS (1986) Physicochemical properties of oilseed proteins. CRC Crit Rev Biochem **20:** 265-363

Prasad K (2000) Antioxidant activity of secoisolariciresinol diglucoside-derived metabolites, secoisolariciresinol, enterodiol, and enterolactone. International Journal of Angiology **9:** 220-225

Prufer K, Stenzel U, Dannemann M, Green RE, Lachmann M, Kelso J (2008) PatMaN: rapid alignment of short sequences to large databases. Bioinformatics **24:** 1530-1531

Rachinskaia OA, Lemesh VA, Muravenko OV, Iurkevich O, Guzenko EV, Bol'sheva NL, Bogdanova MV, Samatadze TE, Popov KV, Malyshev SV, Shostak NG, Heller K, Khotyleva LV, Zelenin AV (2011) Genetic polymorphism of flax Linum usitatissimum based on use of molecular cytogenetic markers. Genetika **47:** 65-75

Ragupathy R, Rathinavelu R, Cloutier S (2011) Physical mapping and BAC-end sequence analysis provide initial insights into the flax (*Linum usitatissimum* L.) genome. BMC Genomics **12:** 217

Rajagopalan R, Vaucheret H, Trejo J, Bartel DP (2006) A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis thaliana*. Genes & Development **20:** 3407-3425

Rajwade AV (2012) Molecular and biochemical characterization of high and low α-linolenic acid containing Indian flax (*Linum usitatissimum* L.) varieties. University

of Mumbai, Pune, INDIA

Rajwade AV, Arora RS, Kadoo NY, Harsulkar AM, Ghorpade PB, Gupta VS (2010) Relatedness of Indian flax genotypes (*Linum usitatissimum* L.): An inter-simple sequence repeat (ISSR) primer assay. Molecular Biotechnology **45:** 161-170

Ramakers C, Ruijter JM, Deprez RHL, Moorman AFM (2003) Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. Neuroscience Letters **339:** 62-66

Ranocha P, McNeil SD, Ziemak MJ, Li CJ, Tarczynski MC, Hanson AD (2001) The S-methylmethionine cycle in angiosperms: ubiquity, antiquity and activity. Plant Journal **25:** 575-584

Rao S, Abdel-Reheem M, Bhella R, McCracken C, Hildebrand D (2008) Characteristics of high alpha-linolenic acid accumulation in seed oils. Lipids **43:** 749-755

Ravanel S, Gakiere B, Job D, Douce R (1998) The specific features of methionine biosynthesis and metabolism in plants. Proceedings of the National Academy of Sciences of the United States of America **95:** 7805-7812

Remm M, Storm CEV, Sonnhammer ELL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. Journal of Molecular Biology **314:** 1041-1052

Roach MJ, Deyholos MK (2007) Microarray analysis of flax (*Linum usitatissimum* L.) stems identifies transcripts enriched in fibre-bearing phloem tissues. Molecular Genetics and Genomics **278:** 149-165

Roach MJ, Deyholos MK (2008) Microarray analysis of developing flax hypocotyls identifies novel transcripts correlated with specific stages of phloem fibre differentiation. Annals of Botany **102:** 317-330

Rogozin IB, Lyons-Weiler J, Koonin EV (2000) Intron sliding in conserved gene families. Trends in Genetics **16:** 430-432

Romans JR, Johnson RC, Wulf DM, Libal GW, Costello WJ (1995) Effects of ground flaxseed in swine diets on pig performance and on physical and sensory characteristics and omega-3 fatty acid content of pork: I. Dietary level of flaxseed. Journal of Animal Science **73:** 1982-1986

Roose-Amsaleg C, Cariou-Pham E, Vautrin D, Tavernier R, Solignac M (2006) Polymorphic microsatellite loci in *Linum usitatissimum*. Molecular Ecology Notes **6:** 796-799

Ross J, Li Y, Lim EK, Bowles DJ (2001) Higher plant glycosyltransferases. Genome Biology **2:** 2

Roy SW, Gilbert W (2005) Rates of intron loss and gain: Implications for early eukaryotic evolution. Proceedings of the National Academy of Sciences of the United States of America **102:** 5773-5778

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol **132:** 365-386

Ruan MB, Zhao YT, Meng ZH, Wang XJ, Yang WC (2009) Conserved miRNA analysis in *Gossypium hirsutum* through small RNA sequencing. Genomics **94:** 263-268

Rudd S (2003) Expressed sequence tags: alternative or complement to whole genome sequences? Trends in Plant Science **8:** 321-329

Ruuska SA, Girke T, Benning C, Ohlrogge JB (2002) Contrapuntal networks of gene expression during Arabidopsis seed filling. Plant Cell **14:** 1191-1206

Saitou N, Nei M (1987) The Neighbor-Joining Method - a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution **4:** 406-425

Sammour RH (1999) Proteins of linseed (*Linum usitatissimum* L.), extraction and characterization by electrophoresis. Botanical Bulletin of Academia Sinica **40:** 121-126

Sammour RH, Elshourbagy MN, Aboshady AM, Abasary AM (1994) The Seed Proteins of Linseed *(Linum usitatissimum* L.). Botanical Bulletin of Academia Sinica **35:** 171-177

Sawada S, Suzuki H, Ichimaida F, Yamaguchi M, Iwashita T, Fukui Y, Hemmi H, Nishino T, Nakayama T (2005) UDP-glucuronic acid : anthocyanin glucuronosyltransferase from red daisy (Bellis perennis) flowers - Enzymology and phylogenetics of a novel glucuronosyltransferase involved in flower pigment biosynthesis. Journal of Biological Chemistry **280:** 899-906

Schmittgen TD, Livak KJ (2008) Analyzing real-time PCR data by the comparative C_t method. Nature Protocols **3:** 1101-1108

Schneider G, Schliemann W (1994) Gibberellin conjugates - an overview. Plant Growth Regulation **15:** 247-260

Schwender J, Ohlrogge J, Shachar-Hill Y (2004) Understanding flux in plant metabolic networks. Current Opinion in Plant Biology **7:** 309-317

Schwender J, Shachar-Hill Y, Ohlrogge JB (2006) Mitochondrial metabolism in developing embryos of *Brassica napus*. Journal of Biological Chemistry **281:** 34040-34047

Shahidi F, Wanasundara PKJPD (1997) Cyanogenic glycosides of flaxseeds. Antinutrients and Phytochemicals in Food **662:** 171-185

Shewry PR, Napier JA, Tatham AS (1995) Seed storage proteins: structures and biosynthesis. Plant Cell **7:** 945-956

Shi JS, Zhen Y, Zheng RH (2010) Proteome profiling of early seed development in *Cunninghamia lanceolata* (Lamb.) Hook. Journal of Experimental Botany **61:** 2367-2381

Shi L, Katavic V, Yu YY, Kunst L, Haughn G (2012) Arabidopsis glabra2 mutant seeds deficient in mucilage biosynthesis produce more oil. Plant Journal **69:** 37-46

Shimakata T, Stumpf PK (1983) Purification and characterization of beta-ketoacyl-ACP synthetase I from *Spinacia oleracea* leaves. Archives of Biochemistry and Biophysics **220:** 39-45

Shimakata T, Stumpf PK (1983) The purification and function of acetyl coenzyme-a - acyl carrier protein transacylase. Journal of Biological Chemistry **258:** 3592-3598

Sim JS, Cherian G (1994) Alpha-linlenic acid for designing poultry products for human consumption: Canadian designer eggs and nutritional significance. Proceedings of the 55th Flax Institute, Fargo, North Dakota**:** 106-122

Simcox PD, Reid EE, Canvin DT, Dennis DT (1977) Enzymes of the glycolytic and pentose phosphate pathways in proplastids from the developing endosperm of *Ricinus communis* L. Plant Physiology **59:** 1128-1132

Simopoulos A (1995) Evolutionary aspects of diet: fatty acids, insulin resistance and obesity *In* T VanItallie, A Simopoulos, eds, Obesity: new directions in assessment and management. Charles Press, Philadelphia, pp 241-261

Simopoulos AP (2002) Omega-3 fatty acids in inflammation and autoimmune diseases. Journal of the American College of Nutrition **21:** 495-505

Skorkowska-Telichowska K, Kulma A, Szopa J (2012) The response of diabetic foot to a new type of dressing. International Archives of Medicine **5:** 33

Smale ST (2001) Core promoters: active contributors to combinatorial gene regulation. Genes & Development **15:** 2503-2508

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SBH, Hood LE (1986) Fluorescence detection in automated DNA-sequence

analysis. Nature **321:** 674-679

Solovyev V, Salamov A (1997) The Gene-Finder computer tools for analysis of human and model organisms genome sequences. *In* C Rawling, D Clark, R Altman, L Hunter, T Lengauer, S Wodak, eds, Fifth International Conference on Intelligent Systems for Molecular Biology, Proceedings. AAAI Press, Halkidiki, Greece., pp 294-302

Somerville C, Browse J (1991) Plant lipids - metabolism, mutants, and membranes. Science **252:** 80-87

Song C, Jia Q, Fang J, Li F, Wang C, Zhang Z (2010) Computational identification of citrus microRNAs and target analysis in citrus expressed sequence tags. Plant Biology **12:** 927-934

Soto-Cerda BJ, Saavedra HU, Navarro CN, Ortega PM (2011) Characterization of novel genic SSR markers in *Linum usitatissimum* (L.) and their transferability across eleven Linum species. Electronic Journal of Biotechnology **14:** 6

Stoltzfus A, Logsdon JM, Palmer JD, Doolittle WF (1997) Intron "sliding" and the diversity of intron positions. Proceedings of the National Academy of Sciences of the United States of America **94:** 10739-10744

Strack D (1997) Phenolic metabolism. *In* PM Dey, JB Harborne, eds, Plant Biochemistry. Academic Press, London, UK, pp 387–416

Strassburg CP, Vogel A, Kneip S, Tukey RH, Manns MP (2002) Polymorphisms of the human UDP-glucuronosyltransferase (UGT) 1A7 gene in colorectal cancer. Gut **50:** 851-856

Sun G (2012) MicroRNAs and their diverse functions in plants. Plant Molecular Biology **80:** 17-36

Tammes T (1928) The genetics of the genus *Linum*. Bibliographica Genetica **4:** 1-36

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution **28:** 2731-2739

Tang GL (2010) Plant microRNAs: An insight into their gene structures and evolution. Seminars in Cell & Developmental Biology **21:** 782-789

Thakur V, Wanchana S, Xu M, Bruskiewich R, Quick WP, Mosig A, Zhu XG (2011) Characterization of statistical features for plant microRNA prediction. BMC Genomics **12:** 108

Thompson JD, Higgins DG, Gibson TJ (1994) Clustal-W - improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research **22:** 4673-4680

Thompson LU (1995) Flaxseed, lignans, and cancer. *In* SC Cunnane, LU Thompson, eds, Flaxseed in human nutrition. AOCS Press, Chicago, IL, pp 219–236

Thompson LU, Robb P, Serraino M, Cheung F (1991) Mammalian Lignan Production from Various Foods. Nutrition and Cancer-an International Journal **16:** 43-52

Thorsoe KS, Bak S, Olsen CE, Imberty A, Breton C, Moller BL (2005) Determination of catalytic key amino acids and UDP sugar donor specificity of the cyanohydrin glycosyltransferase UGT85B1 from Sorghum bicolor. Molecular modeling substantiated by site-specific mutagenesis and biochemical analyses. Plant Physiology **139:** 664-673

Toure A, Xu XM (2010) Flaxseed lignans: source, biosynthesis, metabolism, antioxidant activity, bio-active components, and health benefits. Comprehensive Reviews in Food Science and Food Safety **9:** 261-269

Troufflard S, Roscher A, Thomasset B, Barbotin JN, Rawsthorne S, Portais JC (2007) *In vivo* ^{13}C NMR determines metabolic fluxes and steady state in linseed embryos. Phytochemistry **68:** 2341-2350

Tukey RH, Strassburg CP (2000) Human UDP-glucuronosyltransferases: Metabolism, expression, and disease. Annual Review of Pharmacology and Toxicology **40:** 581-616

Turner M, Yu O, Subramanian S (2012) Genome organization and characteristics of soybean microRNAs. BMC Genomics **13:** 169

Turner MA, Yang XD, Yin D, Kuczera K, Borchardt RT, Howell PL (2000) Structure and function of S-adenosylhomocysteine hydrolase. Cell Biochemistry and Biophysics **33:** 101-125

Unver T, Budak H (2009) Conserved microRNAs and their targets in model grass species *Brachypodium distachyon*. Planta **230:** 659-669

Venglat P, Xiang D, Qiu S, Stone SL, Tibiche C, Cram D, Alting-Mees M, Nowak J, Cloutier S, Deyholos M, Bekkaoui F, Sharpe A, Wang E, Rowland G, Selvaraj G, Datla R (2011) Gene expression analysis of flax seed development. BMC Plant Biology **11:** 74

Vigeolas H, Waldeck P, Zank T, Geigenberger P (2007) Increasing seed oil content in oil-seed rape (*Brassica napus* L.) by over-expression of a yeast glycerol-3-phosphate dehydrogenase under the control of a seed-specific promoter. Plant Biotechnology Journal **5:** 431-441

Vittal K, Kerkhi S, Chary GR, Sankar GM, Ramakrishna Y, Srijaya T, Samra J (2005) District based promising technologies for rainfed linseed based production system in India. *In*, Vol 71. All India Co-ordinated Research Project for Dryland Agriculture, Central Research Institute for Dryland Agriculture, Indian Council of Agricultural Research, Hyderabad 500 059

Vogt T, Jones P (2000) Glycosyltransferases in plant natural product synthesis: characterization of a supergene family. Trends in Plant Science **5:** 380-386

Wall DP, Fraser HB, Hirsh AE (2003) Detecting putative orthologs. Bioinformatics **19:** 1710-1711

Wang J, Hou B (2009) Glycosyltransferases: key players involved in the modification of plant secondary metabolites. Frontiers of Biology in China **4:** 36-46

Wang JW, Czech B, Weigel D (2009) miR156-Regulated SPL transcription factors define an endogenous flowering pathway in *Arabidopsis thaliana*. Cell **138:** 738-749

Wang W, Vignani R, Scali M, Cresti M (2006) A universal and rapid protocol for protein extraction from recalcitrant plant tissues for proteomic analysis. Electrophoresis **27:** 2782-2786

Wang YP, Li KB (2009) Correlation of expression profiles between microRNAs and mRNA targets using NCI-60 data. BMC Genomics **10:** 218

Wang Z, Hobson N, Galindo L, Zhu S, Shi D, McDill J, Yang L, Hawkins S, Neutelings G, Datla R, Lambert G, Galbraith DW, Grassa CJ, Geraldes A, Cronk QC, Cullis C, Dash PK, Kumar PA, Cloutier S, Sharpe A, Wong GK, Wang J, Deyholos MK (2012) The genome of flax (*Linum usitatissimum*) assembled de novo from short shotgun sequence reads. Plant J **72:** 461-473

Weber MJ (2005) New human and mouse microRNA genes found by homology search. FEBS Journal **272:** 59-73

Wightman B, Ha I, Ruvkun G (1993) Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in *C. elegans*. Cell **75:** 855-862

Williams D, Verghese M, Walker LT, Boateng J, Shackelford L, Chawan CB (2007) Flax seed oil and flax seed meal reduce the formation of aberrant crypt foci (ACF) in azoxymethane-induced colon cancer in Fisher 344 male rats. Food and Chemical Toxicology **45:** 153-159

Wise MJ, Tunnacliffe A (2004) POPP the question: what do LEA proteins do? Trends in Plant Science **9:** 13-17

Wu MF, Tian Q, Reed JW (2006) Arabidopsis microRNA167 controls patterns of ARF6 and ARF8 expression, and regulates both female and male reproduction. Development **133:** 4211-4218

Xie FL, Xiao P, Zhang BH (2011) Target-align: a Tool for Plant MicroRNA Target Identification. In Vitro Cellular & Developmental Biology-Animal **47:** S71-S72

Xie ZX, Allen E, Fahlgren N, Calamar A, Givan SA, Carrington JC (2005) Expression of Arabidopsis MIRNA genes. Plant Physiology **138:** 2145-2154

Yamazaki M, Gong Z, Fukuchi-Mizutani M, Fukui Y, Tanaka Y, Kusumi T, Saito K (1999) Molecular cloning and biochemical characterization of a novel anthocyanin 5-O-glucosyltransferase by mRNA differential display for plant forms regarding anthocyanin. Journal of Biological Chemistry **274:** 7405-7411

Yang TW, Xue LG, An LZ (2007) Functional diversity of miRNA in plants. Plant Science **172:** 423-432

Yonekura-Sakakibara K, Hanada K (2011) An evolutionary view of functional diversity in family 1 glycosyltransferases. Plant Journal **66:** 182-193

Zagrobelny M, Bak S, Rasmussen AV, Jorgensen B, Naumann CM, Moller BL (2004) Cyanogenic glucosides and plant-insect interactions. Phytochemistry **65:** 293-306

Zeist van W, Bakker-Heeres JAH (1975) Evidence for linseed cultivation before 6000 bc. Journal of Archaeological Science **2:** 215-219

Zeng CY, Wang WQ, Zheng Y, Chen X, Bo WP, Song S, Zhang WX, Peng M (2010) Conservation and divergence of microRNAs and their functions in Euphorbiaceous plants. Nucleic Acids Research **38:** 981-995

Zhang BH, Pan XP, Anderson TA (2006) Identification of 188 conserved maize microRNAs and their targets. FEBS Letters **580:** 3753-3762

Zhang BH, Pan XP, Cannon CH, Cobb GP, Anderson TA (2006) Conservation and divergence of plant microRNA genes. Plant Journal **46:** 243-259

Zhang BH, Pan XP, Cox SB, Cobb GP, Anderson TA (2006) Evidence that miRNAs are different from other RNAs. Cellular and Molecular Life Sciences **63:** 246-254

Zhang BH, Pan XP, Stellwag EJ (2008) Identification of soybean microRNAs and their targets. Planta **229:** 161-182

Zhang BH, Wang QL, Wang KB, Pan XP, Liu F, Guo TL, Cobb GP, Anderson TA (2007) Identification of cotton microRNAs and their targets. Gene **397:** 26-37

Zhang LF, Chia JM, Kumari S, Stein JC, Liu ZJ, Narechania A, Maher CA, Guill K, McMullen MD, Ware D (2009) A genome-wide characterization of microRNA genes in Maize. PLoS Genetics **5:** 11

Zhao BT, Ge LF, Liang RQ, Li W, Ruan KC, Lin HX, Jin YX (2009) Members of miR-169 family are induced by high salinity and transiently inhibit the NF-YA transcription factor. BMC Molecular Biology **10:** 29

Zhao YT, Wang M, Fu SX, Yang WC, Qi CK, Wang XJ (2012) Small RNA profiling in two *Brassica napus* cultivars identifies microRNAs with oil production and development correlated expression and new small RNA classes. Plant Physiology **158:** 813-823

Zhong R, Ye ZH (2007) Regulation of HD-ZIP III Genes by MicroRNA 165. Plant Signaling & Behavior **2:** 351-353

Zhou M, Sun J, Wang QH, Song LQ, Zhao GA, Wang HZ, Yang HX, Li X (2011) Genome-wide analysis of clustering patterns and flanking characteristics for plant microRNA genes. FEBS Journal **278:** 929-940

Zhu RS, Li X, Chen QS (2011) Discovering numerical laws of plant microRNA by evolution. Biochemical and Biophysical Research Communications **415:** 313-318

Zybailov B, Friso G, Kim J, Rudella A, Rodriguez VR, Asakura Y, Sun Q, van Wijk KJ (2009) Large scale comparative proteomics of a chloroplast Clp protease mutant reveals folding stress, altered protein homeostasis, and feedback regulation of metabolism. Molecular & Cellular Proteomics **8:** 1789-1810

CURICULUM VITAE

Mr. Vitthal T. Barvkar

Division of Biochemical Sciences,
CSIR-NCL, Pune- 411008, Maharashtra,
India.

Email: vt.barvkar@ncl.res.in
vbarvkar@gmail.com
Mobile No. +917588797433
Phone No. +91-020-

Academic qualifications

Master of Science *in Biotechnology* (2009), from VSBT, Baramati, University of Pune, with **69.84%**.

Bachelor of Science *in Biotechnology* (2007), from VSBT, Baramati, University of Pune, with **71.68%**.

Higher School Certificate from D.B.F. Dayanand College Solapur (2004) with *Biology*, *Mathematics*, *Physics and Chemistry* major subjects, from Pune Divisional Board, with **86.50%**.

Publications

1. **Barvkar VT**, Pardeshi VC, Kale SM, Kadoo NY, Gupta VS (2012) Phylogenomic analysis of UDP glycosyltransferase 1 multigene family in *Linum usitatissimum* identified genes with varied expression patterns. **BMC Genomics 13:** 175 (**IF: 4.07**)
2. **Barvkar VT**, Pardeshi VC, Kale SM, Giri AP, Kadoo NY, Gupta VS (2012) Proteome profiling of flax (*Linum usitatissimum* L.) seed: Characterization of functional metabolic pathways operating during seed development. **Journal of Proteome Research 11** (12) 6264–6276 (**IF: 5.113**)
3. **Barvkar VT**, Pardeshi VC, Kale SM, Qiu S, Rollins M, Datla R, Gupta VS, Kadoo NY (2013) Genome-wide identification and characterization of microRNA genes and their targets in flax (*Linum usitatissimum*). **Planta** (**In press** DOI: 10.1007/s00425-012-1833-5 **IF: 3.0**)
4. Kale SM, Pardeshi VC, **Barvkar VT**, Gupta VS, Kadoo NY (2012) Identification and Characterization of Nucleotide-Binding Site-Leucine-Rich Repeat Genes in linseed. **Genome 56**: 1–9 (**IF: 1.653)**
5. Kale SM, Srivastava RL, Singh PK, Dubey SD, Pardeshi VC, **Barvkar VT**, Gupta VS, Kadoo NY (2012) Development of core collection of linseed (*Linum*

usitatissimum L.) and exploring its genetic diversity and population structure using SSR markers, communicated to **Genetica. IF:2.148**

6. Kale SM, Pardeshi VC, **Barvkar VT**, Kadoo NY, Gupta VS (2012) *In-silico* analysis of EST-SSRs in linseed (*Linum usitatissimum*) for indentifying potential markers for different traits. Manuscript ready for submission

Patent

Gupta VS, Kadoo NY, **Barvkar VT**, Pardeshi VC, Kale SM, MicroRNA sequences from Flax

IN Patent Filed (Reference number: 3239DEL2012) Dated 18th Oct.2012

Techniques known

- ✓ Polymerase Chain Reaction (PCR)
- ✓ DNA and Protein Sequencing
- ✓ Gene cloning and heterologous expression in bacterial and yeast system
- ✓ Recombinant enzyme purification and enzyme kinetics studies
- ✓ Real time PCR for gene expression analysis
- ✓ Chromatographic separation of fatty acids using GC-FID
- ✓ LC-MS/MS and MALDI for proteomic studies
- ✓ LC-MS for metabolite analysis

Bioinformatics tools used

- ✓ Knowledge of various databases for biological research, homology searches with protein and nucleic acid sequences and sequence comparisons.
- ✓ Whole genome mining and analysis: MIRA, CAP3, MEME, MAST, Blast2Go, MultiExperiment Viewer, etc.
- ✓ Phylogenetic analysis: Phylip, MEGA, etc
- ✓ Proteomics data analysis: ProteinLynx Global SERVER (PLGS), Mascot

Awards and honours

1. Awarded NCL RF-KEERTHI SANGORAM MEMORIAL ENDOWMENT AWARD for **"Best Research Scholar"** from biological Sciences for year 2012.
2. Awarded NCL RF-GUPTA-PARDESHI - SAINANI AWARD**"** for **"Best Published Research Paper in Biological Sciences"** with highest impact factor for the year 2012

Scholarships awarded

1. **Senior Research Fellowship July-2011** by Council of Scientific and Industrial Research (CSIR), New Delhi for Life Sciences.
2. **Junior Research Fellowship July-2009** by Council of Scientific and Industrial Research (CSIR), New Delhi for Life Sciences.
3. Qualified highly competitive **National Eligibly Test (CSIR-NET) Dec-2008** essential for the doctoral studies as well as postgraduate teaching in **Life Sciences**.

Invitation as a resource person for workshop

Invited as a resource person for workshop on **"Use of Molecular Markers in Plants"** sponsored by DBT-Star college project at Department of Biotechnology, PVP College, Pravaranagar, July 2012.

Conference participation and Poster Prsentations

1. Participated:

A. In 93rd **Indian Science Congress** Theme: Integrated Rural Development: Science and Technology **2006** held at **ANGRU Hyderabad**, India.

B. Attended six months certificate course on "**Patents for Researchers**" organized by **Intellectual Property Facilitation Center** (IPFace) Venture Center, NCL, Pune, India.

C. Volunteered and attended seminar on **"Food Safety Issues"** with specific emphasis on **GM Food Crops** on November 14 to 16, 2010 at NCL, Pune, India.

2. Poster presentation in International Conferences

A. **Barvkar VT**, Pardeshi VC, Kale SM, Kadoo NY, Giri AP, Gupta VS. Label-free Quantitative Proteome Analysis of Flax (*Linum usitatissimum*) Seed Developmental Stages and Expression Profiles of Enzymes Involved in Oil Biosynthesis. At **World Congress on Biotechnology** during 21-23 March 2011, Hyderabad, India. (**DOI:10.4172/2153-0602.10000S1**)

B. **Barvkar VT**, Pardeshi VC, Kale SM, Gupta VS,Kadoo NY. *In silico* Identification and Characterization of Flax (*Linum usitatissimum*) MicroRNA Genes**.** At C-DAC Symposium on **Accelerating Biology 2012: Computing to Decipher** during 15-17 February 2012, Pune, India.

C. **Barvkar VT**, Pardeshi VC, Kale SM, Kadoo NY, Giri AP, Gupta VS. Proteome profiling of flax (*Linum usitatissimum*) seed: Characterization of functional metabolic pathways operating during seed development. At **International Symposium on Proteomics Beyond ids... and 4th Annual Meeting of PSI** during 22-24 November 2012, Pune, India.

Personal Details

Date of Birth	:	27th December, 1986
Marital status	:	Married
Language to Speak	:	English, Hindi, & Marathi.

References

Dr. Vidya S. Gupta, Chief Scientist and Chair, Division of Biochemical Sciences, National Chemical Laboratory, Pune - 411008, India. vs.gupta@ncl.res.in
Ph. +91-020-25902237

Dr. Ashok P. Giri, Senior Scientist, Division of Biochemical Sciences, National Chemical Laboratory, Pune - 411008, India. ap.giri@ncl.res.in Ph. +91-020-25902710

Dr. Narendra Y. Kadoo, Senior Scientist, Division of Biochemical Sciences, National Chemical Laboratory, Pune - 411008, India. ny.kadoo@ncl.res.in
Ph. +91-020-25902724