Development and application of DNA markers for genetic improvement of linseed

A thesis submitted to the
University of Pune

For the degree of
DOCTOR OF PHILOSOPHY

In
BIOTECHNOLOGY

By
Sandip M. Kale

Research Guide
Dr. VIDYA S. GUPTA

Plant Molecular Biology Group
Division of Biochemical Sciences
CSIR-National Chemical Laboratory
Pune 411008 (INDIA)

March 2014

This work was performed at

CSIR-NATIONAL CHEMICAL LABORATORY, PUNE, INDIA

Scientists involved in this activity

CSIR-NCL: Dr. Vidya S. Gupta (supervisor)

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

PDKV, Nagpur: Dr. P. B. Ghorpade

CSAUAT: Drs. R. L. Srivastava & P. K. Singh

CERTIFICATE

Certified that the work in the Ph.D. thesis entitled "**Development and application of DNA markers for genetic improvement of linseed**" submitted by **Mr. Sandip M. Kale** was carried out by the candidate under our supervision. The material obtained from other sources has been duly acknowledged in the thesis.

Date:
Place: Pune

Dr. Vidya S. Gupta	**Dr. Narendra Y. Kadoo**
(Research Guide)	(Research Co-guide)

DECLARATION

I hereby declare that the thesis entitled **“Development and application of DNA markers for genetic improvement of linseed”** 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:

Sandip M. Kale
Division of Biochemical Sciences,
CSIR-National Chemical Laboratory,
Pune 411008

Dedicated

to

My beloved

family

&

friends

Contents

Chapter 3: Development and characterization of EST-SSR markers for linseed and their comparative analysis with related species

Chapter 4: Development of core collection of linseed and analysis of genetic diversity and population structure using SSR markers

Chapter 7: Summary and future directions

Acknowledgements

The PhD program that I have been through is truly a once in a lifetime experience. There are many people who deserve to be thanked for what I have been achieving and for what I have become now.

First of all, I would like to express my sincere gratitude to the person directly responsible for my PhD research- my supervisor Dr. Vidya Gupta. It has been immense pleasure working under your guidance and your simplicity and passion towards Molecular markers has been an inspiration to me. The constant moral support, constructive advice and criticism, untiring interest throughout the pursuit of the research were really helpful. I learnt a lot about the subject and also about myself during our conversations.

My co-supervisor Dr. Narendra Kadoo also deserves my gratitude for sharing the highs and lows that are inevitably linked to research presented in this thesis. His great efforts in computational analysis, improvement in manuscripts need special thanks. I am also very grateful to Dr. Ashok Giri for his moral and emotional support throughout my PhD and financial help during my visit to USA.

I would like to acknowledge Dr. P. B. Ghorpade and his team at AICRP on Linseed, College of Agriculture, Nagpur for providing the seed material and field facility for growing linseed varieties used in my study. I also thank Dr. R. L. Srivastava and Dr. P. K. Singh from PC Linseed unit, CSAUAT, Kanpur for providing phenotypic data of linseed accessions and seed samples. I am especially thankful to Dr. Arron Lorenz and Dr. Diego Jarquin from University of Nebraska-Lincoln, USA for helping me analyzing SNP data and allowing me to attend "Population Genetics" class. Without their help, this work wouldn't have been possible. Special thanks to Dr. Jeffrey P. Mower for providing space on Linux server.

I am especially indebted to Dr. Dipak Santra for providing me an opportunity to work as a "Visiting Scholar" at the University of Nebraska-Lincoln (UNL), USA. That international experience will be very helpful to build my career. Also, help of Dr. Devin Rose and Dr. Vicki Schlegel in data analysis during my visit to UNL is greatly appreciated.

I can't move forward without expressing my gratitude to Dr. Varsha Pardeshi for her contribution in designing my PhD objectives and for her suggestions and constant encouragement throughout the work. I want to express my deepest gratitude towards Dr. Vitthal Barvkar, my friend and my colleague for his help and for constantly supporting me in all way since my graduation. I would also like to thank Dr. Ashwini Rajwade for helping me in initial experiment and teaching basic molecular biology techniques. Dr. Gayatri Gurjar's help in learning fungal DNA isolation, while Dr. Ramya's help in QTL work is also greatly acknowledged.

There are no words to describe my heartfelt thanks to my friends, Dr. Vishal Dawkar, Dr. Ashok Chougale, and Yashwant. I think of them as elder brothers and have invaluable role in my Ph.D. Also, my sincere thanks to Dr. Purushottam, Rakesh, Kedar, Sachin and Tejas for your emotional support. I will always remember the wonderful days spent in your company.

I sincerely thank Dr. Bhushan Dholakia for healthy scientific discussions, sound advice and good company and also critically reviewing my thesis. I would like to seize this opportunity to thank all my PMB friends Ajit, Amey, Ashish, Atul, Hemangi, Gouri, Hemlata, Medha, Priyanka, Radhika, Ram, Rasika, Reema, Rahul, Neha K., Neha M., Sheon, Yojana, Ramesha, Smrati, Amol, Sucheta, Yamini and all other lab mates 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 the present Directors of CSIR-NCL, Pune 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 my PhD at CSIR-NCL, Pune.

Last but not the least, sincere thanks to all my family members for their support and encouragement throughout. Aai and Kaka, I am speechless to express my gratitude towards your naive support, love and faith in me. I am profoundly thankful of my wife, Priti, for her trust and unconditional support. A special mention goes to my parent-in-law for her support. Special thanks to my brother, Shankar and his wife, Rekha, my sister, Sujata (Mai) and her husband (Kishor) for providing encouragement, moral and financial support.

I fully understand that my memory serves me well only to a certain extent rendering a limitation to my recollection of the good deeds that many other people have done to and for me. For those who I have failed to mention, I sincerely apologize, and I thank you from the bottom of my heart. Above all, I thank God almighty for giving me the strength and courage at every step of life.

Sandip

List of Abbreviations

AA	Amino acid
AFLP	Amplified fragment length polymorphism
AHC	Agglomerative hierarchical clustering
ALA	α-Linolenic acid
AM	Association mapping
AMOVA	Analysis of molecular variance
ANOVA	Analysis of variance
ARI	Agharkar Research Institute
BAC	Bacterial artificial chromosome
BL	Breeding lines
bp, kb, Mb, Gb	Base pair, kilo-base pair, megabase pair, gigabase pair
CC	Core collection
cDNA	Complementary deoxyribonucleic acid
CNL	Coiled-coil-NBS-LRR
CPP	Capsules per plant
CR	Coincidence rate of range
CSAUAT	Chandra Shekhar Azad University of Agriculture and Technology
Ct	Thresh hold cycles
CTAB	Hexadecyl-trimethyl-ammonium bromide
DAF	Days to 50% flowering
DAI	Days after inoculation
DH	Doubled haploid
DHA	Docosahexaenoic acid
DNA	Deoxy-ribose nucleic acid
DNR	Di-nucleotide repeats
dNTPs	Deoxyribonucleotide triphosphate
DTM	Days to maturity
DUS	Distinctness, uniformity and stability
EC	Exotic collections
ED	Enterodiol
EDTA	Ethylene diamine tetra acetate

EFA	Essential fatty acids
EL	Enterolactone
EL	Elite lines
EPA	Eicosapentanoic acid
ESTs	Expressed sequence tags
FA	Fatty acids
FAMEs	Fatty acid methyl esters
FDR	False discovery rate
FIASCO	Fast Isolation by AFLP of Sequences COntaining repeats
FID	Flame ionization detector
GBS	Genotyping by sequencing
GC	Gas chromatography
GLM	General linear model
GMU	Germplasm maintenance and use
GWAS	Genome-wide association scan/study
Ha	Hectare
HMM	Hidden Markov model
Ho	Heterozygosity
IRAP	Inter-retrotransposon amplified polymorphism
ISSR	Inter simple sequence repeats
Kbp	Kilo base pairs
t, Kg, g, mg, µg, ng	Tones, kilogram, gram, milligram, microgram, nanogram
l, ml, µl	Liter, milliliter, microliter
LA	Linoleic acid
LD	Linkage disequilibrium
LOESS	Locally weighted scatter plot smoothing
LRR	Leucine-rich repeat
LZ	Leucine zipper
M, mM, µM, N	Molar, millimolar, micromolar, normal
MAF	Minor allele frequency
MAS	Marker-assisted selection
MD	Mean difference percentage
MLM	Mixed linear model

MTA	Marker trait association
NBS	Nucleotide binding site
NCBI	National Center for Biotechnology Information
NJ	Neighbor joining
OA	Oleic acid
ORF	Open reading frame
PA	Palmitic acid
PAGE	Polyacrylamide gel electrophoresis
PC	Primitive cultivars
PCA	Principal component analysis
PCoA	Principal co-ordinate analysis
PCR	Polymerase chain reaction
PDA	Potato dextrose agar
PGRC	Plant Gene Resources of Canada
PH	Plant height
PIC	Polymorphism information content
PIMA	PCR isolation of microsatellite arrays
PUFA	Poly-unsaturated fatty acids
qRT-PCR	Quantitative real-time polymerase chain reaction
QTLs	Quantitative trait loci
RAD	Restriction-site associated DNA
RAPD	Randomly amplified polymorphic DNA
RE	Restriction enzyme
RFLP	Restriction fragment length polymorphism
RNA	Ribonucleic acid
SA	Stearic acid
SD	Standard deviation
SDG	Secoisolariciresinol diglucoside
SE	Standard error
SNPs	Single nucleotide polymorphism
SPC	Seeds per capsule
SSRs	Simple sequence repeats
TDA	Tridecanoic acid

TIR	Toll like Interleukin Receptor
TNL	TIR-NBS-LRR
TNR	Tri-nucleotide repeats
TPH	Technical plant height
TW	Thousand seed weight
USA	United States of America
VLC-PUFA	Very long chain polyunsaturated fatty acids
WGS	Whole genome sequence
YPP	Seed yield per plant

List of Figures

List of Tables

Note: Those 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 Tables

Thesis abstract

Linseed or flax (*Linum usitatissimum* L., 2n=30) is a self-pollinating annual crop species, grown commercially for fiber and seed. It is the third largest natural fiber crop and one of the major oilseed crop worldwide; with production of 2.05 million tons from 3.07 million ha land (FAOSTAT, 2012). It is not only one of the richest sources of α-linolenic acid (ALA) and lignans, but also is an essential source of high quality protein and soluble fiber. Therefore, linseed has a potential to emerge as a crop for food, feed and fiber. Inspite of its high economic and medicinal value, linseed is largely neglected crop as compared to other oilseed crop. Genetic and genomic resources play an important role in genetic improvement of a crop and such resources are limited in linseed. Keeping the nutraceutical importance and need to develop genetic and genomic recourses in linseed, I initiated my research project towards development of DNA markers and subsequently they were used to analyze genetic diversity, population structure and association mapping studies in linseed.

Development of genomic simple sequence repeat markers for linseed

Three microsatellite enriched genomic libraries *viz.* PCR Isolation of Microsatellite Arrays (PIMA), 5’-anchored PCR method and Fast Isolation by AFLP of Sequences COntaining repeats (FIASCO) were prepared. The amplified products from the three methods were pooled and sequenced using 454 GS-FLX platform. A total of 36,332 reads were obtained, which assembled into 2,183 contigs and 2,509 singletons. The contigs and singletons contained 1,842 microsatellite motifs with dinucleotide (54%) motifs as the most abundant repeat type followed by trinucleotide (44%) motifs. Out of the 290 markers designed, 52 were evaluated using a panel of 25diverse linseed genotypes. Among the three enrichment methods, the 5’-anchored PCR method was the most efficient method for isolation of microsatellites; while FIASCO was more efficient to develop SSR markers. In this study, we show the utility of the next generation sequencing technology to efficiently discover a large number of new microsatellite markers in non-model plant.

Development and characterization of EST-SSR markers for linseed and their comparative analysis with related species

The EST resources were used to develop genic markers which are still limiting in number for linseed and also compared with nine closely related species. The expressed sequence tags (ESTs) (2,86,882) and unigenes (59,626) were assembled resulting in 31,928 contigs and 19,400 singletons, which were mined for the presence of putative simple sequence repeats (SSRs). Trinucleotide repeats were abundant with the GAA repeat motif being predominant and could be explained by GC1 codon preference in linseed. Total, 2.47% (1272) class-I SSR-containing ESTs were identified out of which 927 (class-I repeat) primer pairs were designed. Further, virtual PCR using these linseed EST-SSRs against EST dataset of nine selected plant species showed varied amplification rate (0-7.11%), indicated low transferability rate. Thirty randomly selected EST SSRs were evaluated for polymorphism on a diverse set of linseed cultivars and showed allelic diversity. Comparison of EST-SSR distribution among 10 plant genomes chosen for this study revealed that trinucleotides are predominant in most with the frequency of class B (AAG) being the highest. In all the species under study, tri-, hepta- and octa-nucleotide repeats showed similar distribution pattern while others showed varied distribution.

Development of core collection of linseed and analysis of genetic diversity and population structure using SSR markers

This study describes development of core collection (CC) of linseed comprising 222 accessions from the linseed germplasm collection of 2,239 accessions based on morphological and agronomic traits. Agglomerative hierarchical clustering with Ward's minimum variance and proportional random sampling strategy were used to construct the core collection and statistical analysis proved its representativeness to the entire collection and homogeneity. The core collection germplasm was analyzed for the contents of five fatty acids (palmitic, stearic, oleic, LA and ALA) and genetic diversity and population structure using morphological and 29 SSR markers. Further, to demonstrate the practical utility of the core collection, we identified genetically diverse potential parents for seven economic traits for genetic improvement of linseed. This core collection provides an effective mechanism for proper exploitation

of Indian linseed germplasm resources for genetic improvement of this nutraceutically important crop.

Genotyping by sequencing in linseed: development of genome-wide SNP markers and preliminary association mapping for agronomic traits

In this study, we developed 13,280 single nucleotide polymorphism (SNP) markers using advanced genotyping by sequencing (GBS) technology and analyzed the population structure and linkage disequilibrium (LD) in 95 diverse linseed accessions. The neighbor-joining and principal co-ordinate analysis (PCoA) showed weak population structure, which was further supported by low population differentiation (F_{st} = 0.006), indicating that the population is ideal for association mapping (AM). The average LD among significant SNPs was 0.16, while LD decay of over 75 Kbp was observed with reference to the draft genome sequence of linseed. Genome-wide association scan (GWAS) for eight agronomic traits was conducted for the 95 accessions using the SNP markers, which identified 24 SNPs associated with three traits *viz*., number of capsules per plant (CPP), technical plant height (TPH) and plant height (PH) at 5% false discovery rate (FDR).The study successfully demonstrates the application of GBS technology to develop large numbers of SNP markers. Also, preliminary GWAS identify markers associated with agronomic traits in an underutilized crop like linseed. The GWAS can be extended to locate stable quantitative trait loci (QTLs) for different traits of interest.

Genome-wide identification and characterization of nucleotide binding site leucine rich repeat genes in linseed

Plants employ different disease-resistance genes to detect pathogens and induce defense responses. The largest class of these genes encodes proteins with nucleotide binding site (NBS) and leucine-rich repeat (LRR) domains. To identify the putative NBS-LRR encoding genes from linseed, we analyzed the recently published linseed genome sequence and identified 147 NBS-LRR genes. The NBS domain was used for phylogeny construction and these genes were classified into two well-known families, non-TIR (CNL) and TIR related (TNL) and formed eight clades in the neighbor joining bootstrap tree. Eight different gene structures were observed among these

genes. An unusual domain arrangement was observed in the TNL family members, predominantly in the TNL-5 clade members belonging to class D. About 12% linseed specific genes were observed. The study indicated that the linseed NBS-LRR genes probably have an ancient origin with few progenitor genes. Quantitative expression analysis of five genes showed inducible expression. The *in silico* expression evidence was obtained for few of these genes and the expression was not correlated with the presence of any particular regulatory element or with unusual domain arrangement in those genes. This study would help in understanding the evolution of these genes and their disease resistance mechanism which might lead to the development of disease resistant varieties of linseed in future.

CHAPTER 1

Introduction and Review of literature

Flaxseed oil contains
- Phytoestrogen
(balances hormonal levels)
- Omega-3 fatty acids
(improve circulation, strengthen body, and reduce cholesterol)
- Anti-inflammatory properties
(reduce cysts in breasts and inflammation in bowels, digestive tracts, and joints)
shutterstock
"Look at it this way; books don't crash, they don't get viruses, and they don't need screensavers."
"MAPLE LEAF" 38
Old Process
LINSEED OIL CAKE MEAL
MINIMUM PROTEIN 38%
GOOD FOR ALL LIVE STOCK
THE CANADA LINSEED OIL MILLS
LIMITED
MANUFACTURERS
TORONTO 3
CANADA
PURE LINSEED
CAKE MEAL

1.1 Linseed

Linum usitatissimum L., the botanical name for linseed or flax, suitably explains its versatile use and importance to economical and social human development. The word *Linum* is originated from Celtic word *lin* or "thread", while *usitatissimum* is a Latin meaning "most useful" (Koledzeijczyk and Fedec 1995). The word "line" is derived from Latin or Greek ancestor *linum* while other words like linen, lining, lineage and linear are derived from word "line" (Judd 1995). Depending on the region, the terms *flaxseed* and *linseed* have particular meanings. In Europe, flax refers to the seed grown for fiber (linen), while linseed refers to oilseed flax grown for industrial and nutritional uses; although both flax and linseed are the same species. Fiber flax is grown mainly in Northern Europe, Russia and China, but linseed is primarily grown in Canada, USA, Argentina and India as well as Russia and China (Gill 1987; Marchenkov et al. 2003). The common name "linseed" is maintained throughout the thesis to describe both, fiber flax and linseed.

1.2 Taxonomic status

Linseed is annual, self-pollinating crop of the genus *Linum* and family Lineaceae. The Lineaceae family comprises 22 genera. Worldwide, over 300 different species of this family are present having widespread geographic distribution. The systemic classification of family Lineaceae is as follow: Kingdom: *Plantae*, Subkingdom: *Tracheobionta*, Super-division: *Spermatophyta*, Division: *Magnoliophyta*, Class: *Magnoliopsida*, Subclass: *Rosidae* and Order: *Linales*. Within the family, the genus *Linum* belongs to the tribe Linoideae H.Winkl and is the largest genus within the family. The genus Linum is divided into six subsections *viz.*; *Linum*, *Linastrum, Cathartolinum*, *Dasylinum, Syllinum* and *Cliococca* of which the subsection *Linum* contains the cultivated species *L. usitatissimum* L. and the ornamentals *L. grandiflorum* and *L. perenne*. However, the latter two species are of little economic importance. The number of chromosomes of the *Linum* species shows a wide range varying from 2n = 16 to 2n = 72 (Fedorov 1974); however, *L. usitatissimum* and its wild relatives contain 2n = 30 chromosomes (Muravenko et al. 2003). All species are predominantly self-pollinated (Zohary and Hopf 2000); however, cross-pollination may occur via honey bees or by artificial means.

1.3 Origin and domestication

The center of origin of linseed is uncertain, though it is believed to be the Middle East. Secondary diversity centers were identified in the Mediterranean basin, Ethiopia, Central Asia, and India (Vavilov 1951; Zohary and Hopf 2000). A single domestication origin for all existing cultivated linseed, in spite of its wide geographical range, has also been reported based on molecular analysis of the *steroyl-ACP-desaturase II* (sad2) locus, involved in unsaturated fatty acid synthesis (Allaby et al. 2005). As like other oilseed crops, linseed appears to have evolved from wild or weedy form that are either perennial or annual. Some researchers consider *L. bienne* as the progenitor of small seeded flax, which has same chromosome number (2n = 30), periwinkle blue flowers and dehiscent bolls or capsules of seed, originating from Kurdistan and Iran, whereas others consider *L. angustifolium* containing high oil content and seed weight, as progenitor, originating from the Mediterranean region (Zeven and deWet 1975). Other researchers suggest that *L. bienne* and *L. angustifolium* are the same species, and are widely distributed over Western Europe, the Mediterranean basin, North Africa, the Near East, Iran and Caucasus (Tutin et al. 1968; Zohary and Hopf 2000; Fu et al. 2002). Phytogeographic, cytogenetic, phenotypic and recent molecular marker studies also support this hypothesis, and suggest pale flax (*L. angustifolium* and *L. bienne*) to be the wild progenitor of cultivated linseed (Hjelmquist 1950; Gill and Yermanos 1967; Diederichsen and Hammer 1995; Fu et al. 2002).

Linseed is considered a founding crop because it was among the first domesticated plants. Its cultivation might have begun in the fertile valley of the Tigris and Euphrates rivers in Mesopotamia (Zohary and Hopf 1993; Smith 1995). Archeological evidences suggest that linseed probably has been used by humans for about 10,000 years and cultivated for some 8,000 years. Recently, 30,000-year-old processed and colored flax fiber has been found, indicating its even earlier utilization by humans (Kvavadze et al. 2009). Over the time, its cultivation spread from the Near East to Europe, the Nile Valley and West Asia through trades and communication (Zohary and Hopf 1993). Since the domestication of linseed, there has been a preference for growing it either for its fiber or oil. In India, it is mainly grown for its oil.

1.4 Uses of linseed

Historically, linseed has been consumed like a cereal and valued for its medicinal qualities, while the oil from seed has served as a frying medium for food, a lamp oil and a preservative in products such as paints and floorings. The fabric made from linseed stalk is the best documented example of the earliest use of linseed. Almost every part of the linseed plant is utilized commercially, either directly or after processing (**Figure 1.1**). In 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. For a long time, linseed has been cultivated as a dual-purpose crop, but nowadays fiber flax and linseed represent different gene pools (Vromans 2006). This section describes the ancient and contemporary uses of linseed.

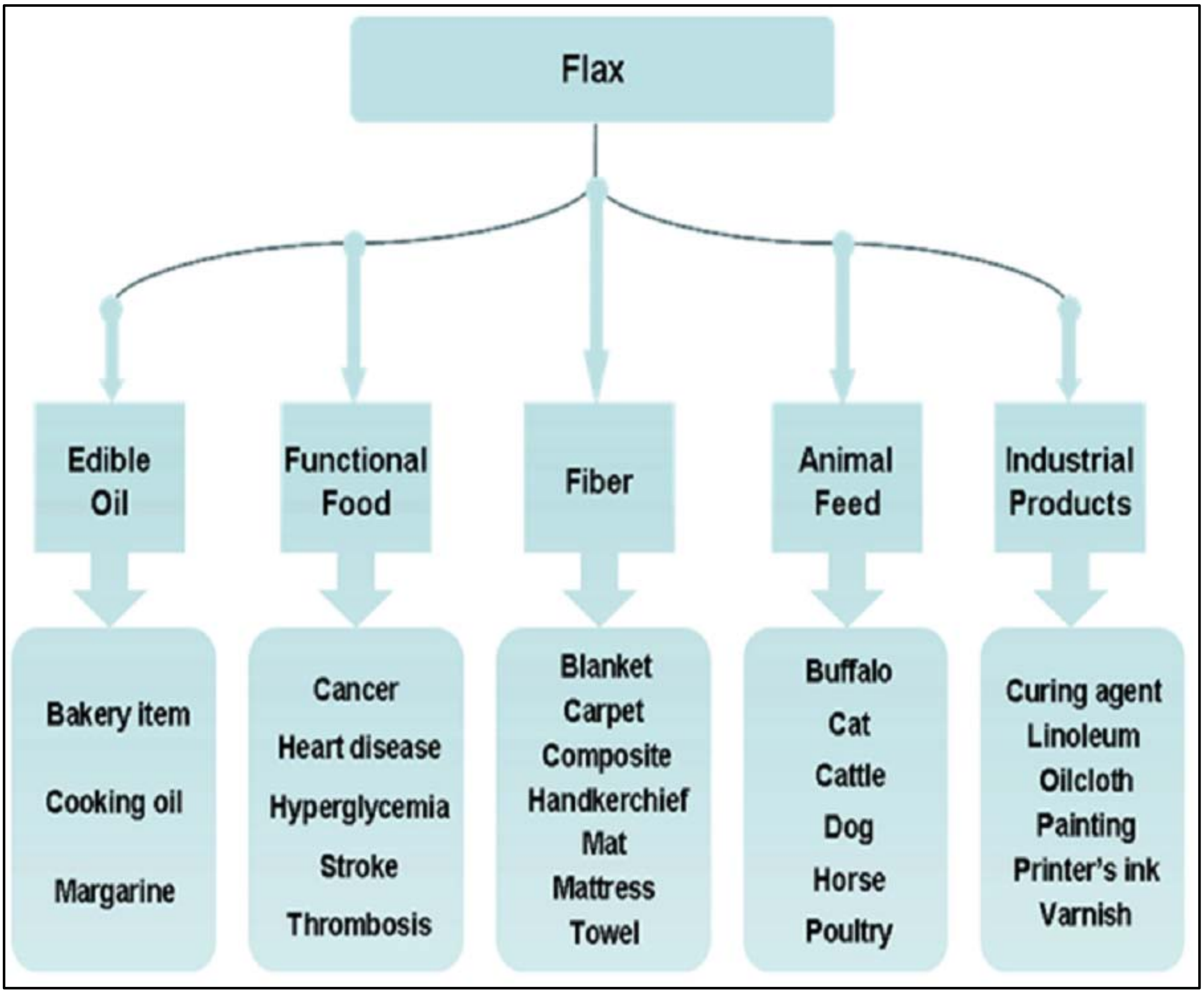


Figure 1.1: Uses of linseed

1.4.1 Nutritional composition of linseed

Linseed has recently been found to be one of the key sources of phytochemicals and has gained importance in the functional food. It is not only one of the richest sources of α-linolenic acid oil and lignans, but also is an essential source of high quality protein and soluble fiber. It has a considerable potential as a source of phenolic compounds (Oomah 2001). The composition of linseed seed as reported by various sources is listed in **Table 1.1**. The following section describes the composition of linseed seed and use of linseed as food and feed crop.

Table 1.1: Proximate composition of linseed based on common measures (Source: Flax Council of Canada)

Form of linseed	Weight (g)	Common measure	Energy Kcal	Total fat (g)	ALA (g)	Protein (g)	Total cho^{cd} (g)	Dietary fiber (g)
Proximate analysis	100	–	450	41	23	20	29	28
Whole Seed	180	1 cup	810	74	41	36	52	50
	11	1 tbsp	50	4.5	2.5	2.2	3	3
	4	1 tsp	18	1.6	0.9	0.8	1.2	1.1
Milled Seed	130	1 cup	585	53	30	26	38	36
	8	1 tbsp	36	3.3	1.8	1.6	2.3	2.2
	2.7	1 tsp	12	1.1	0.6	0.5	0.8	0.8
Linseed oil	100	–	884	100	57	–	–	–
	14	1 tbsp	124	14	8	–	–	–
	5	1 tsp	44	5	2.8	–	–	–

c = CHO = Carbohydrate.
d = Total Carbohydrate includes carbohydrates like sugars and starches (1 g) and total dietary fiber (28 g) per 100 g linseed seeds.

1.4.1.1 Linseed oil

Fatty acids (FAs) are carboxylic acids with a long unbranched aliphatic chain and classified as saturated or unsaturated based on absence or presence of double bonds. The unsaturated fatty acids are further classified as mono-unsaturated and poly-unsaturated fatty acids (PUFA) based on number of double bonds presents. Mono-unsaturated fatty acid contains single alkenyl functional group or double bond while

PUFA contains more than one alkenyl functional group or double bonds. Linseed contains 5 to 6% palmitic acid (16:0), 3 to 6% stearic acid (18:0), 19 to 29% oleic acid (18:1n-9), 14 to 18% linoleic acid (LA)(18:2n-6), and 45 to 52% α- linolenic acid (ALA)(18:3n-3) (Bhatty 1995). The fatty acid profile of linseed and its oil as reported by various researchers are tabulated in **Table 1.2**. The important feature of linseed oil is that, it is low in saturated fat (9% of total fatty acids), moderate in mono-saturated fatty acid (18%), and rich in polyunsaturated fatty acid (PUFA) (73%) (Cunnane et al. 1993). The PUFA content comprises about 16% ω-6 fatty acids, primarily as linoleic acid (LA), and 57% ω-3 fatty acid, ALA and is thus, the richest agricultural source of ALA (Bhatty and Cherdkiagumchai 1990). 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 by a series of alternating desaturations and elongations steps, into very long chain polyunsaturated fatty acids (VLC-PUFA), eicosapentanoic acid (EPA), and Docosahexaenoic Acid (DHA) 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. Another important feature of linseed oil is its lower ratio of ω-6 to ω-3 fatty acids. 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. Linseed has a lower ratio of ω-6/ω-3 and hence helps to improve this FA balance in human body.

Table 1.2: Fatty acid profile of linseed (Source: Mazza, 2008)

Fatty acids	**Whole linseed (%)**			**Linseed oil (%)**	
Palmitic acid (C16:0)	4.6–6.3	—	4.21–8.71	6.0	5.0
Stearic acid (C18:0)	3.3–6.1	—	3.52–8.17	2.5	3.6
Oleic acid (C18:1)	19.3–29.	4 3.6 g	22.17–41.72	19.0	19.5
Linoleic acid (C18:2)	14.0	3.2 g	4.82–19.13	24.1	15.6
Linolenic acid (C18:3)		11.4g	33.22–54.79	47.4	55.8

1.4.1.2 Phenolics

Phenolic compounds in general possess an aromatic ring bearing one or more hydroxyl substituent's and may be found in free state, conjugated with sugars or esters or polymerized (Shahidi and Naczk 2004). In plants, phenols play an important role in protection against photo-oxidation, attracting insect for pollination and disease resistance (Antolovich et al. 2000). Many phenolics appear to have anticancer and antioxidant effects in humans. Linseed contains at least three types of phenolics: lignans, flavonoids, and phenolic acids **(Figure 1.2)**.

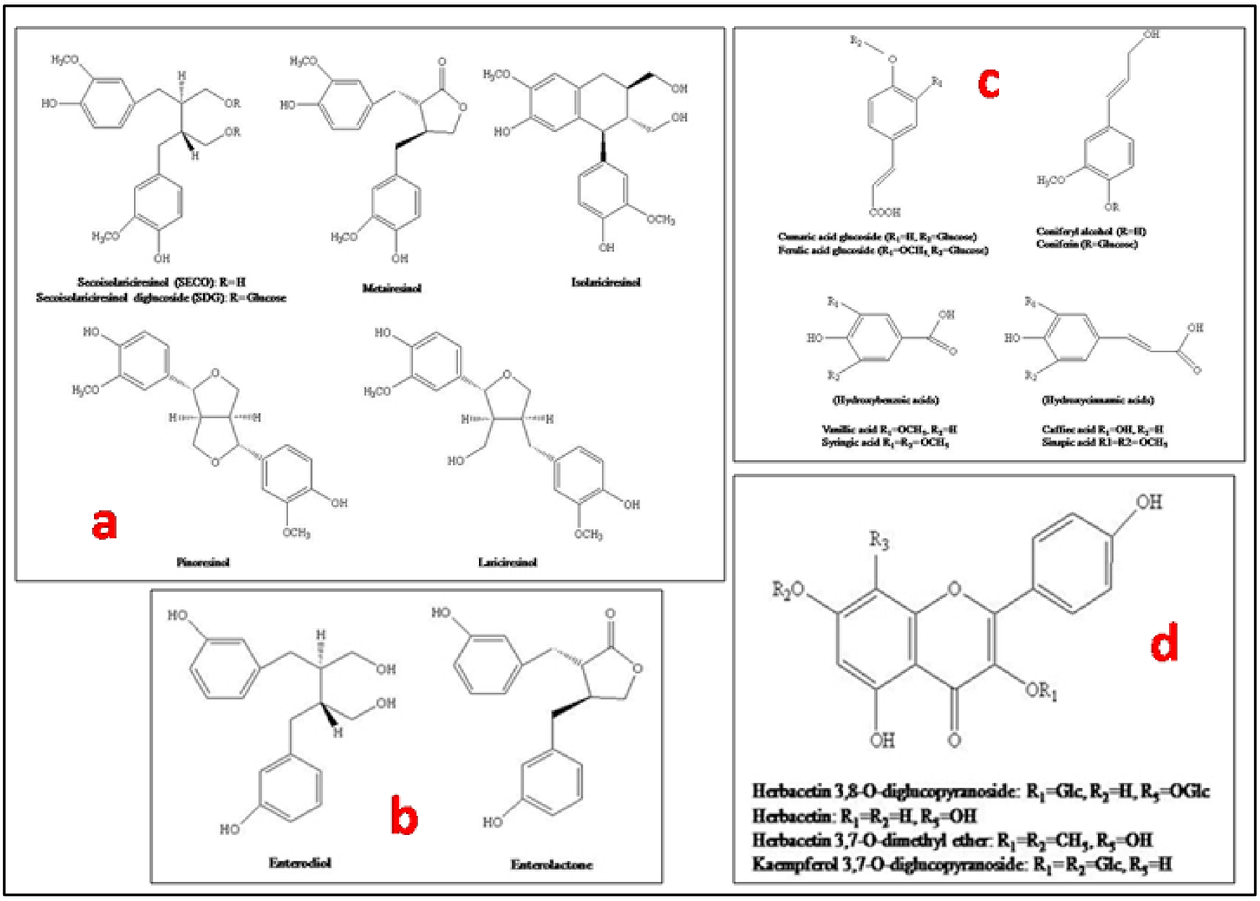

a
Secoisolariciresinol (SECO): R=H
Secoisolariciresinol diglucoside (SDG): R=Glucose
Matairesinol
Isolariciresinol
Pinoresinol
Lariciresinol
b
Enterodiol
Enterolactone
c
Coumaric acid glucoside (R_1=H, R_2=Glucose)
Ferulic acid glucoside (R_1=OCH_3, R_2=Glucose)
Coniferyl alcohol (R=H)
Coniferin (R=Glucose)
(Hydroxybenzoic acids)
(Hydroxycinnamic acids)
Vanillic acid R_1=OCH_3, R_2=H
Syringic acid R_1=R_2=OCH_3
Caffiec acid R_1=OH, R_2=H
Sinapic acid R1=R2=OCH_3
d
Herbacetin 3,8-O-diglucopyranoside: R_1=Glc, R_2=H, R_3=OGlc
Herbacetin: R_1=R_2=H, R_3=OH
Herbacetin 3,7-O-dimethyl ether: R_1=R_2=CH_3, R_3=OH
Kaempferol 3,7-O-diglucopyranoside: R_1=R_2=Glc, R_3=H

Figure 1.2: Chemical structure of linseed Phenolics (Meagher et al., 1999)

1.4.1.2.1 Lignans

Plant lignans are the biologically important class of phenolic compounds. They belong to a group of phenols which are characterized by coupling of two phenylpropanoid units (Willfor et al. 2006). Lignans are found in a variety of plant materials including linseed seed, pumpkin seed, sesame seed, soybean, broccoli and some berries. The major lignan in the linseed is secoisolariciresinol diglucoside (SDG) (Bambagiotti-alberti et al. 1994; Cardoso carraro et al. 2012). In addition to SDG, smaller quantities of other type lignans such as matairesinol, isolariciresinol, lariciresinol and pinoresinol have also been identified in the linseed (Meagher et al. 1999; Sicilia et al. 2003) **(Figure 1.2a)**. Among foods, linseed is the richest source of SDG, which contains 75–800 times more SDG than any other foods (Mazur et al. 1996; Westcott and Muir 1996). SDG was first time isolated from linseed by Bakke and Klosterman (1956). SDG is then converted into mammalian lignans, enterodiol (ED) and enterolactone (EL) **(Figure 1.2b)** by colon bacteria (Borriello et al. 1985; Wang et al. 2000).

The health benefits of linseed 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 (Mazur et al. 1996). The behavior of the lignans depends on the biological levels of estradiol. At the normal estradiol levels, lignans act as estrogen antagonists, but in postmenopausal women (at the low estradiol levels) they can act as weak estrogens therefore, inhibit hormone dependent cancers (Hutchins and Slavin 2003).

1.4.1.2.2 Phenolic acids

Linseed was reported to contain 8-10 g/kg total phenolic acids, about 5 g/kg of esterified phenolic acids and 3-5 g/kg of etherified phenolic acids (Oomah 2001). They are either in free and/or bound forms. Free phenolic acids are mainly composed of trans and cis-sinapic, o-coumaric, p-droxybenzoic, trans-p-coumaric and vanillic acids (Kozlowska et al. 1983; Babrowski and Sosulski 1984) **(Figure 1.2c).** However, most of the linseed phenolic acids such as p-hydroxybenzoic, trans-ferulic and trans-p-coumaric acids are ester bound. Among these phenolic acids, ferulic and p-coumaric acid glucosides were accumulated at high concentrations in the linseed (Beejmohun et al. 2007). In addition, phenolic acid like caffiec acid and their glucosides were also reported in linseed (Babrowski and Sosulski 1984). Variations in phenolic acid content in linseed were largely attributed to seasonal effects (Oomah 2001).

1.4.1.2.3 Flavonoids

Flavonoids are the polyphenols, with C_6-C_3-C_6 skeleton that consists of two aromatic rings joined by a three-carbon link. Flavonoids generally include anthocyanins, flavanols, flavones, flavanones and flavonols. Depending upon growing and cultivar conditions, linseed possesses about 0.3-0.71 g of total flavonoids per kg of linseed (Oomah et al. 1995). In linseed, flavonoids are in the form of their glucoside such as herbacetin 3, 8-O-diglucopynanoside, herbacetin 3, 7-O-dimethyl ether, and kaempferol 3, 7-O-diglucopyranoside (Qiu et al. 1999) **(Figure 1.2d)**.

1.4.1.3 Protein

Protein content of linseed varies from 20 to 30%, constituting mainly globulins (linin and conlinin), glutelin, but no albumin (Care 1954). The amino acid pattern of linseed protein is similar to that of soybean protein, which is viewed as one of the most nutritious of the plant proteins. The non-protein nitrogen in the seed forms 21.7% of the total nitrogen. Linseed proteome contains a relatively higher proportion of arginine and glutamic acid. The total nitrogen content in linseed is 3.25 g/ 100 g of seed (Gopalan et al. 2007). Lysine is the most limiting amino acid in linseed. The amino acid composition of linseed is shown in **Table 1.3**. Anonymous (1962) reported that linseed proteins possess high digestibility coefficients (89.6%) at 8% level of protein intake and biological value (77.4%).

1.4.1.4 Dietary fibers

Linseed is different from other oilseeds in having a relatively high content of a mucilaginous material composed of acidic and neutral polysaccharides **(Table 1.4)** (BeMiller 1973; Mazza and Biliaderis 1989b).The proportion of soluble to insoluble fiber in linseed varies between 20:80 and 40:60 (Oomah et al. 1995; Morris 2003). The major insoluble fiber fraction consists of cellulose and lignin, while the soluble fiber fractions are the mucilage gums (Mazza and Biliaderis 1989a; Vaisey-Genser and Morris 2003). 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. The mucilage accounts for about 8% of the linseed weight. The acid hydrolysis products of these polysaccharides are L-galactose, D-xylose, L-arabinose, L-rhamnose, D-galacturonic acid, and perhaps traces of D-glucose (Mazza and Biliaderis 1989b).

Mucilage gums are polysaccharides that become viscous when mixed with water or other fluids and have an important role in laxatives.

Table 1.3: Amino acid composition of linseed (Bhatty et al., 1995)

Amino acid	Linseed cultivar		Soy flour
	Brown linseed (norlin)	Yellow linseed (Omega) G/100 g protein	
Alanine	4.4	4.5	4.1
Arginine	9.2	9.4	7.3
Aspartic acid	9.3	9.7	11.7
Cystine	1.1	1.1	1.1
Glutamic acid	19.6	19.7	18.6
Glycine	5.8	5.8	4
Histidine*	2.2	2.3	2.5
Isoleucine*	4	4	4.7
Leucine*	5.8	5.9	7.7
Lysine*	4	3.9	5.8
Methionine*	1.5	1.4	1.2
Phenylalanine*	4.6	4.7	5.1
Proline	3.5	3.5	5.2
Serine	4.5	4.6	4.9
Threonine*	3.6	3.7	3.6
Tryptophan*	1.8	NR	NR
Tyrosine	2.3	2.3	3.4
Valine*	4.6	4.7	5.2

NR = Not reported.* Essential amino acids for humans.

Table 1.4: Composition of relative neutral sugars in linseed and commercial gums (Mazza and Biliaderis 1989b)

	Linseed seed gums			Commercial gums		
	NorMan	Omega	Foster (%)	Arabic	Guar	Xanthan
Rhamonse	21.2	27.2	25.6	34	0	0
Fucose	5	7.1	5.8	0	0	0
Arabinose	13.5	9.2	11	24	24	0
Xylose	37.4	28.2	21.1	0	0	0
Galactose	20	24.4	28.4	45	33	0
Glucose	2.1	3.6	8.2	0	0	50.7
Mannose	0	0	0	0	67	49.3

1.5 Linseed as a crop

1.5.1 Linseed as a food crop

Linseed is believed to be the earliest and best documented oil and fiber crop cultivated some 8000 years ago (Zohary and Hopf 1993; Smith 1995). Linseed has three major components making it beneficial in human nutrition: (1) a very high content of alpha linolenic acid (omega-3 fatty acid) essential for humans; (2) a high percentage of dietary fiber, both soluble and insoluble; and (3) the highest content of plant “lignans” of all plant or seed products used for human food. Because of this, the National Cancer Institute evaluated linseed, along with a number of other potential food ingredients, as a component of “designer foods” (Stitt 1990). Designer foods may be defined as those foods composed of one or more ingredients that contribute essential nutrients for health but also protect against certain diseases such as cancer and coronary heart disease.

The Greeks preferred to mix roasted linseed, barley and coriander with salt for making breads. In Ethiopia, it was a key ingredient in stews, porridges and drinks. Roasted seeds of linseed were mixed with pulses to make a stew called "w' et". A popular porridge was also made from roasted, crushed and cooked linseeds, to which salt and red-pepper were added. In India, dry roasted linseed seeds are used to make chutney, which can be served as mixed with raw oil to have with parathas or chapati. Nowadays, whole linseed seeds, oils, capsules and food products made with added linseeds are available in supermarkets and health food stores. Whole or milled linseed seed is popular addition to baked products like bagels, nutrition bars and multigrain breads.

1.5.2 Linseed as a feed crop

Linseed cake approximately contains 30.5% protein, 6.6% fat, 43.2% nitrogen–free extract, 9.5% crude fiber, and 7% mineral matter (Brown 1953). It is the most valuable feeding cake; perhaps the most favorite cattle feed. The linseed cake is valued for its appetite stimulating and slightly laxative effects and good for animals, both as a feed component and as a nutritional additive. Other sources of fat such as micronized soybeans and Megalac can be completely substituted by whole untreated linseed as the fat source in the diet of early lactating cows without any adverse effect on production and linseed increased milk protein percentage and its ϖ-6 to ϖ -3 fatty acids ratio (Petit 2002). Holstein cows when fed with a control diet with no linseed, raw linseed, a micronized linseed diet, and an extruded diet showed that linseed supplementation improved total nutrient utilization with no adverse effects on ruminal fermentation. It is known that feeding ϖ-3 enriched diets to poultry increases the ϖ-3 content of eggs and meat and thus enriched poultry products offer consumers an alternative to enhance their ϖ–3 daily intake (Leskanich and Noble 1997). Feeding linseed to laying hens increases the ϖ-3 fatty acid in the egg by 6 to 8 times, making one egg equal to 113 (4 oz) of cold water fish as some of the omega-3 fatty acids. Moreover, feeding hens with the enriched diets did not result in any significant changes in the quality parameters of eggs, laying efficiency, and the consumption of feed per egg.

1.5.3 Linseed as a fiber crop

Linseed stem fiber is now being processed and used for a number of products. In addition to cigarette paper, linseed fibers are being used for pulp and paper, erosion control mats, reinforcing materials in plastics and particle composite products (Domier and Kerr 2000). In Europe, there is considerable interest in the use of natural fibers (such as linseed) in interior panels, visors, and other parts of automobiles. Natural fibers like linseed are blended with polypropylene or other synthetic fibers then needle-punched into a mat, a cover material can be added and then the composite can be hot pressed in one operation. Panels and molded products made from linseed fiber/polypropylene mats may be very suitable for applications such as dairy plants, abattoirs, food processing facilities, etc. Mats made from linseed with or without the addition of other materials such as polypropylene, polyethylene, cotton, wool, may be suitable for use as insulation, filters, upholstery padding, carpet backing, geotextiles for erosion control, and horticultural applications.

1.5.4 Linseed: production

1.5.4.1 Worldwide scenario

Traditionally, linseed has been grown for its oil, which is used in industry for paints, varnish etc. Recently, a mutant variety, Linola, is produced which contains less than 5% ALA and thus suitable for culinary uses. Linseed is currently grown on about 3 million hectare land worldwide, with the majority of the linseed production concentrated in three countries, Canada (23.83%),Russian Federation (18%) and

China (17%) (FAOSTAT 2012) **(Table 1.5)**. Canada is the world's largest producer of oilseed flax, with Saskatchewan producing about 80% of the flax grown in western Canada with a current production of 0.49 million tons (http://www.flaxcouncil.ca). Linseed is the third most important oilseed crop in Canada. Canada exports most of its production to Europe, Japan, Korea and USA (Flax council of Canada, 2012).All linseed varieties registered in Canada are brown-seeded and have high levels of alpha-linolenic fatty acid (ALA). Initially, wilt and rust were the major diseases of linseed; however, recently, linseed yield is affected more due to various diseases like pasmo, powdery mildew and others (Flax council of Canada).

1.5.4.2 Indian scenario

In India, linseed is grown in Rabi season i.e. cultivated during winter (September-October) while harvested in spring (February-March) although optimum time of sowing varies from region to region. In the peninsular region, it is generally sown early, whereas in the north, it is generally sown little late. Sowing delay beyond optimum time adversely affects plant growth which in turn affects yield. Though it is mainly cultivated at low altitude, it can be successfully grown up to 770 meters. It requires cool temperature during vegetative growth while moderate temperatures (21-26°C) during harvesting. High temperature above 32°C along with drought during the flowering stage reduces the yield, the oil content of the seed and also the quality of the oil. It requires less water and grow best in area where annual rainfall ranges from 45-75 cm. Linseed is very flexible as per soil is concerned, however, the yield is high on heavier soil with more water retention capacity. Also, it tolerates wide range of soil pH conditions. In Madhya Pradesh and Maharashtra linseed is grown largely on the black cotton soils having high clay and lime content, however, can be grown on the light alluvial soils of Uttar Pradesh, Bihar and West Bengal. The linseed plants mature within 120 -150 days after sowing. Stem of plant becomes yellow while capsules and leaves begin to dry at maturity. After harvesting, plant takes three to four days to dry completely.

In India, linseed is mainly grown as a rotating crop with the main winter crops. It is cultivated over 0.5 million hectare area mainly in Uttar-Pradesh, Madhya-Pradesh, Maharashtra, Chhatisgarh and Jharkhand. The annual production of linseed in India is decreasing. In 2012, the total production was 0.13 million tons, while it was 0.17 million tons in 2009. Presently, it contributes only about 6% to total world production (FAOSTAT, 2012). Madhya Pradesh leads in linseed production, followed by Uttar Pradesh, Bihar and Maharashtra **(Table 1.6)**. Low economic returns and low productivity of the native varieties are the main reasons for linseed under cultivation in India. Also diseases like wilt, rust, powdery mildew and Alternaria blight are causing lot of losses to this crop. The reported losses in yield due to diseases and pests in severe conditions are 40-100% by rust, 60% and above by powdery mildew, 27-60% by alternaria blight, upto 80% by wilt and 97% by linseed bud fly (Vittal et al. 2005).

1.6 Goals of linseed breeding

The average productivity of linseed at national (0.4 t/ha) as well as at global level (0.85 t/ha) is low in comparison to other oil crops like soybean, rapeseed mustard and groundnut. Hence, the breeding strategies for yield enhancement need immediate attention. Along with that, the linseed breeding in India is mainly concentrated on following objectives:

1. Development of short duration varieties (105 days or less).
2. Development of linseed varieties resistant to budfly, alternaria blight and powdery mildew.
3. Value addition and product diversification for pharmaceutical and nutraceutical purpose.

Most of the above traits are complex and polygenic in nature, controlled by several quantitative trait loci (QTLs). Conventional breeding can be used to address these, however, it has inherent problem of long selection cycle and enormous resources required for successful outcome. In contrast, marker-assisted selection (MAS) has the potential to reduce the time required for the development of new and improved varieties. MAS has been successfully utilized in almost all type of crops including cereals, oilseed and legumes (Xu and Crouch 2008). The basic requirement of MAS is the availability of a large number of molecular markers and diverse germplasm collection for trait of interest.

Table 1.5: Top ten linseed producing countries (FAOSTAT 2012)

Country	**Production (tones)**				**Country**	**Area Harvested (Hectare)**			
	2009	**2010**	**2011**	**2012**		**2009**	**2010**	**2011**	**2012**
Canada	930100	423000	368300	489000	**Russian Federation**	80700	226500	472700	558300
Russian Federation	102620	178213	471220	369043	**India**	407900	342000	338810	500000
China	318135	352812	358641	350000	**Kazakhstan**	58400	225200	310000	370000
Kazakhstan	47650	94610	273000	158000	**Canada**	623300	353300	273200	360000
United States of America	188550	230030	70890	146360	**China**	336930	324400	322100	350000
India	169200	153700	147000	130000	**China, mainland**	336930	324400	322100	350000
Ethiopia	156079	150629	65420	112761	**United States of America**	127070	169160	70010	135980
Ukraine	37300	46800	51100	65000	**Ethiopia**	180873	140801	73688	116541
France	43113	40795	30403	49357	**France**	66178	73285	77220	79372
United Kingdom	54000	72000	71000	42000	**Ukraine**	46800	56300	58700	55000

Table 1.6: State-wise production of linseed in India

States	Production ('000 tons)							
	2000-01	2001-02	2002-03	2003-04	2004-05	2005-06	2006-07	2007-onwords
Andhra Pradesh	1.1	1.0	1.0	1.0	1.0	1.0	1.0	0.0
Assam	5.3	5.0	5.0	5.0	4.8	4.1	4.0	4.0
Bihar	28.5	25.6	22.2	27.2	23.2	26.4	24.3	23.5
Chhatisgarh	15.8	22.2	19.7	23.1	16.5	17.9	16.9	18.2
Himachal Pradesh	0.7	0.7	0.4	1.0	2.0	0.1	0.4	0.4
Jammu & Kashmir	0.3	0.3	0.2	0.1	0.3	0.2	2.3	0.1
Jharkhand	4.0	4.0	4.0	5.0	4.0	4.8	12.1	12.3
Karnataka	6.4	6.6	6.1	1.0	3.0	4.0	4.0	5.0
Madhya Pradesh	54.1	62.7	50.3	61.7	52.4	55.9	49.2	32.8
Maharashtra	16.3	21.0	12.0	15.0	11.0	18.0	16.0	19.0
Nagaland	3.0	5.0	9.0	5.0	5.6	4.8	5.7	6.8
Orissa	8.0	8.5	5.4	7.2	8.5	9.7	10.5	11.9
Punjab	0.6	0.3	0.2	0.2	0.2	0.1	0.1	0.1
Rajasthan	5.2	3.1	0.6	2.3	2.3	1.5	1.0	0.6
Uttar Pradesh	50.2	38.4	36.9	39.8	32.9	22.1	19.2	27.1
West Bengal	4.0	4.7	3.7	1.9	2.0	1.9	1.2	1.6

The recent advancement in molecular biology has increased the number and decreased the cost of molecular marker development in various crop species. This advancement along with expansion in biotechnology, genomic research and conventional plant breeding has created the foundation for molecular plant breeding, an interdisciplinary science that has revolutionized crop improvement in the 21st century. Though molecular markers have been widely used to map agronomically important genes/traits and to assist breeding in many plant species, use of them in linseed is still at preliminary stage. For example, the first linkage map in linseed was constructed in 1998 (Spielmeyer et al. 1998), while it took next 13 years to publish SSR based linkage map (Cloutier et al. 2011). This might be due to less contribution of this crop to the agricultural economy and unawareness about its medicinally important properties. However, with increasing health awareness, many research groups are now actively involved in linseed research. As a result of this, together with new, highly informative and high-throughput sequencing technologies, the ample of genetic resources *viz*., complete genome sequence information and gene expression data are now available from linseed which has enormous potential for application in the genetic analysis and breeding. Following section summarizes the available molecular marker resources in linseed and discusses the possible future development.

1.7 Molecular markers in linseed breeding: frontier and prospects

1.7.1 Molecular markers in linseed

Morphological and biochemical markers are traditionally used to answer specific questions in commercial plant breeding later followed by the molecular markers. Use of molecular markers in linseed started in the late 1990s. Gorman and Parojcic (1992) used 16 RAPD, and 22 RFLP markers along with 4 isozymes and 4 morphological polymorphisms to develop initial version of linseed genetic map. Later Cullis et al. (1995) used more RAPD (69) markers to develop a preliminary genetic map which was later modified by Oh et al. (2000). The discovery of PCR based marker system especially provides new opportunities to identify a large number of DNA polymorphisms within genomes of homozygous species such as linseed. One of such techniques called AFLP in combination with RFLP was used to develop the first comprehensive linkage map of linseed (Spielmeyer et al. 1998). Further, simple sequence repeats (SSRs) also called as microsatellite markers became popular because of their highly polymorphic and robust nature and relatively simple and inexpensive analysis.

The actual development of SSR markers in linseed was started in 2006 (Roose-Amsaleg et al. 2006), though Wiesner et al. (2001) developed microsatellite specific - PCR (MP-PCR) primers for linseed. Different methods such as SSR-enriched library screening and, more recently, through the more economical mining of EST or genomic sequence data have been used to develop these markers. For example, 1) Roose-Amsaleg et al. (2006) used a method in which genomic DNA was first digested using restriction enzyme and the digested fragments were cloned and transformed. The repeat containing sequences were captured using biotinylated probes and streptavidin coated magnetic beads and sequenced. 2) Deng et al. (2010) used PIMA (PCR Isolation of microsatellite arrays) method developed by Lunt et al. (1999), which uses RAPD primers for initial PCR amplification. He identified 88 genomic SSR markers. 3) Soto-Cerda et al. (2011) mined publically available genomic sequences to develop 88 SSR markers. More recently, 4) Cloutier et al., used expressed sequence tags and bacterial artificial chromosome (BAC) end sequences to develop SSR markers (Cloutier et al. 2009; Cloutier et al. 2012a). Although the number of publically available SSRs is increasing in linseed, they are still very less in comparison with other important crop species such as rice, wheat, maize etc. Till date, 2,629 genomic and EST-SSRs have been reported as follows: 10 (Wiesner et al. 2001), 28 (Roose-Amsaleg et al. 2006), 662 (Cloutier et al. 2009), 88 (Deng et al. 2010), 92 (Bickel et al. 2011), 92 (Deng et al. 2011), 88 (Soto-Cerda et al. 2011a), 43 (Soto-Cerda et al. 2011b), 20 (Rachinskaya et al. 2011), and 1506 (Cloutier et al. 2012a). Out of these, only 1,317 were found to be polymorphic on various linseed cultivars **(Table 1.7)**.

Nowadays, single nucleotide polymorphism markers (SNPs), which are highly abundant and known to be present in high frequency in the genome, are gaining very much attention**.** Kumar et al., (2012), have developed 55, 465 SNPs using reduced representation libraries of eight linseed genotypes, and 4,706 SNPs have been validated. However, more such markers need to be developed.

Table 1.7: List of genomic and EST-SSR markers developed for linseed

Author	Year	No. of SSRs	Polymorphic
Wiesner et al.	2001	10	10
Roose-Amsaleg et al.	2006	28	23
Cloutier et al.	2009	662	248
Deng et al.	2010	88	35
Deng et al	2011	92	38
Soto-Cerda et al.	2011a	88	60
Soto-Cerda et al.	2011b	43	23
Bickel et al.	2011	92	42
Rachinskaya et al.	2011	20	20
Cloutier et al.	2012a	1506	818
Total		2629	1317

1.7.2 Application of molecular markers

1.7.2.1 Genetic diversity analysis

Availability of diverse germplasm serves various purposes in plant breeding. For example, diverse germplasm can also act as a panel for association mapping (AM) studies and diverse genotypes can be selected as potential parents for biparental mapping etc. Advantages of molecular markers along with morphological and biochemical markers to analyze germplasm diversity has already been reported in many crop plants (Li et al. 2004). Various markers such as RAPD, RFLP, ISSR and SSRs have been used to study genetic diversity and very low genetic diversity within linseed germplasm has been reported (Oh et al. 2000; Fu 2005; Cloutier et al. 2009; Rajwade et al. 2010; Cloutier et al. 2012a). Fu et al., (2002) used RAPD markers to analyze genetic diversity and relationships in 61 (22 Canadian cultivars, 29 world cultivars and 10 landraces) accessions of flax and reported low genetic diversity. However, they observed clear distinction between fiber and oil type accessions, thus suggesting their distinct genetic makeup. Fu et al., (2003) also analyzed diversity within 54 North American linseed cultivars using RAPD markers. Additionally, Diederichsen and Fu., 2006 analyzed diversity within 3,101 cultivated flax accessions from flax collection held by Plant Gene Resources of Canada (PGRC). The accessions were grouped into four infra-specific groups, *viz*. dehiscent flax, fiber flax, large-seeded flax, and intermediate flax; and within and between group diversity was analyzed using RAPD primers. Moreover, Wiesnerova and Wiesner (2004); Rajwade et al. (2010); and Uysal et al. (2010) analyzed genetic diversity independently from different linseed germplasm using ISSR markers and found low diversity within linseed accessions. Recently, Smykal et al. (2011) used inter-retrotransposon amplified polymorphism (IRAP) markers to analyze genetic diversity within 708 accessions and also reported the same observation.

Besides the advantages of molecular markers over morphological and biochemical markers, use of these markers on a very large collection is not very efficient (Li et al. 2004). In such cases, development of core collection (CC) which is a true representative of entire collection would be more beneficial. Large collections of linseed germplasm are available in different countries such as Germany **(**http://fox-serv.ipk-gatersleben.de/), Russia (Zhuchenko and Rozhmina 2000), United States of America etc., including the world collection of linseed maintained by Plant Gene Resources of Canada (PGRC). Diederichsen (2007) reported the presence of more than 45,000 *Linum* accessions worldwide. However, there have been limited efforts to establish a CC for linseed. During 1998-2001, the Centre for Genetic Resources, The Netherlands (CGN) has developed a core collection of 83 accessions from 947 accessions of both, fiber flax and linseed. India also has a large collection of linseed germplasm (2,239 accessions) maintained at the Project Coordinating Unit (Linseed), C.S. Azad University of Agriculture and Technology Campus, Kanpur. Inspite of this impressive collection of germplasm, there has been limited use of these accessions in the genetic improvement of Indian linseed; thereby the extent of genetic diversity present in Indian germplasm collection remains unknown. Therefore, development of core collection of Indian linseed accession will be more useful in future for efficient utilization of genetic diversity present within accessions.

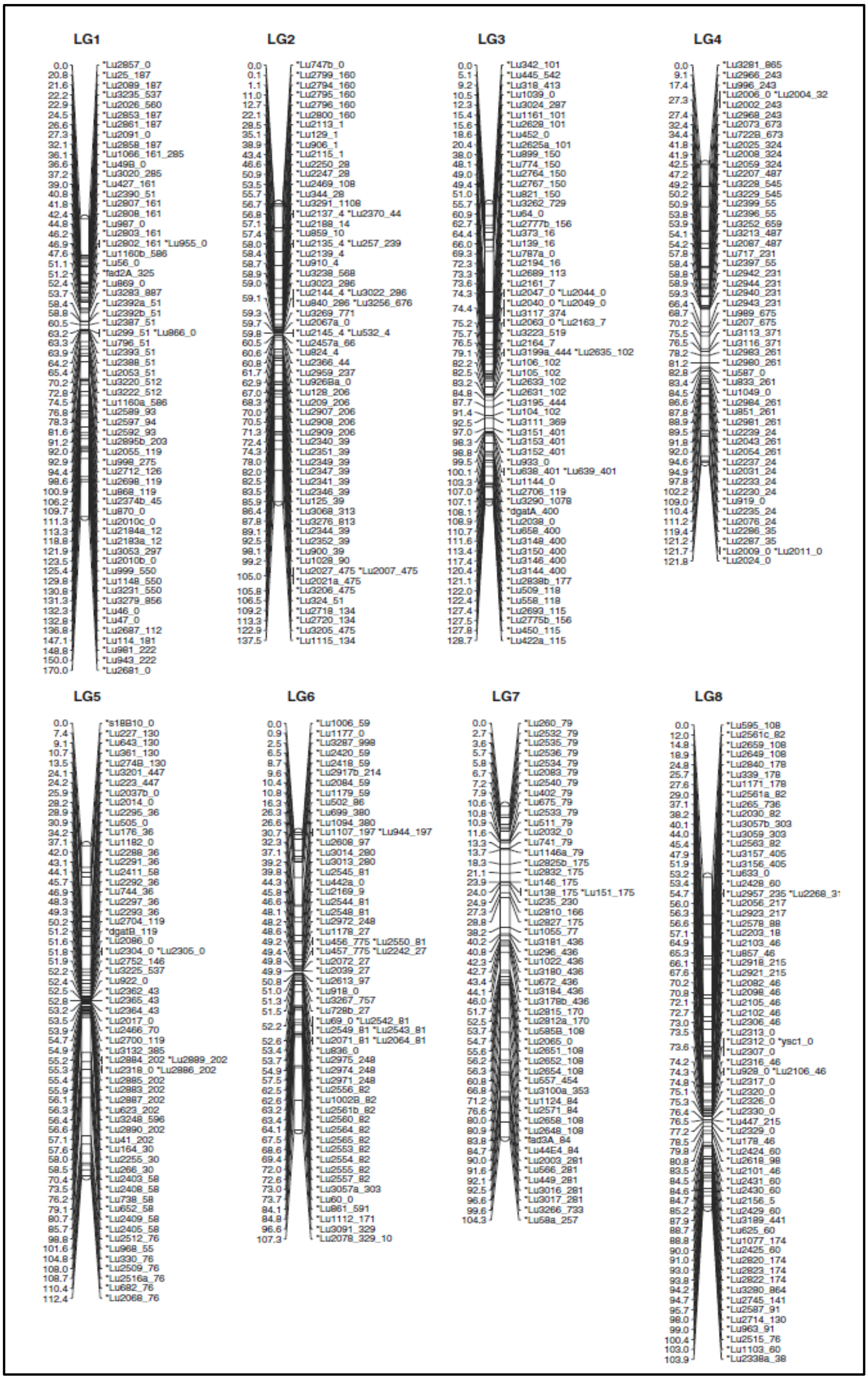
1.7.2.2 Linkage map development

The application of molecular markers for linkage mapping in linseed is at early stage. Various markers such as AFLP, RFLP, RAPD and SSRs have been used to develop linkage maps in linseed. Spielmeyer et al. 1998 constructed a linkage map using 213 AFLP markers and 143 double haploid lines covering 1400 cM of the linseed genome and comprising 18 linkage groups (LGs). The study identified two Quantitative trait loci (QTLs) on linkage group 6 and linkage group 10 explaining 38% and 26%, respectively of the total phenotypic variation. Further, a linkage map comprising 15

LGs and covering 1,000 cM of linseed genome was constructed using RFLP/RAPD markers and F_2 population (Oh et al. 2000). In addition to this, Cloutier et al., 2011 constructed a linkage map using 114 EST-SSRs and five SNP markers and a double haploid population of 78 individuals. The map covered 883.8 cM of linseed genome and comprised 24 linkage groups detecting two major QTLs for linoleic and linolenic acids and one for palmitic acid. However, about 27% markers showed distortion from expected 1:1 ratio in DH population. Recently, Cloutier et al., 2012b constructed a consensus map of linseed using 770 SSR markers and comprising 15 LGs and spanning 1,151 cM with a mean marker density of 2.0 cM **(Figure 1.3).** The map was constructed from three linkage maps using three different populations containing about 385-469 mapped markers each. A total of 670 markers were anchored to 204 of the 416 fingerprinted contigs of the physical map (Ragupathy et al. 2011) corresponding to ~274 Mb or 74% of the estimated linseed genome size of 370 Mb. However, an internationally accepted consensus nomenclature has to date not been agreed. Such high density map can be used to study genome evolution and organization, anchoring of whole genome sequences and for comparative genomic studies.

1.7.2.3 Comparative genomics

Recently, Wang et al. 2012 performed whole genome shotgun sequencing of CDC Bethune cultivar. Seven paired end libraries were sequenced using an Illumina genome analyzer and assembled *de novo* (with very deep coverage approximately 69 % of filtered reads) to generate a 302 Mb of non-redundant sequence representing an estimated 81% genome coverage of total 375 Mb linseed genome. Availability of whole genome sequence (WGS) has opened the doors of comparative genomics in linseed. For example, WGS of linseed has been used for genome wide identification of glucosyl transferase (Barvkar et al. 2012) and miRNA genes (Barvkar et al. 2013). Further, linseed genome shares 80% homology with that of poplar indicates great future for comparative mapping in linseed. This can be utilized to understand genome evolution and chromosome organization in linseed. Also, comparative genomics can be used to find allele variants of a functional gene, which can be used to develop functional markers.


LG1
LG2
LG3
LG4
LG5
LG6
LG7
LG8

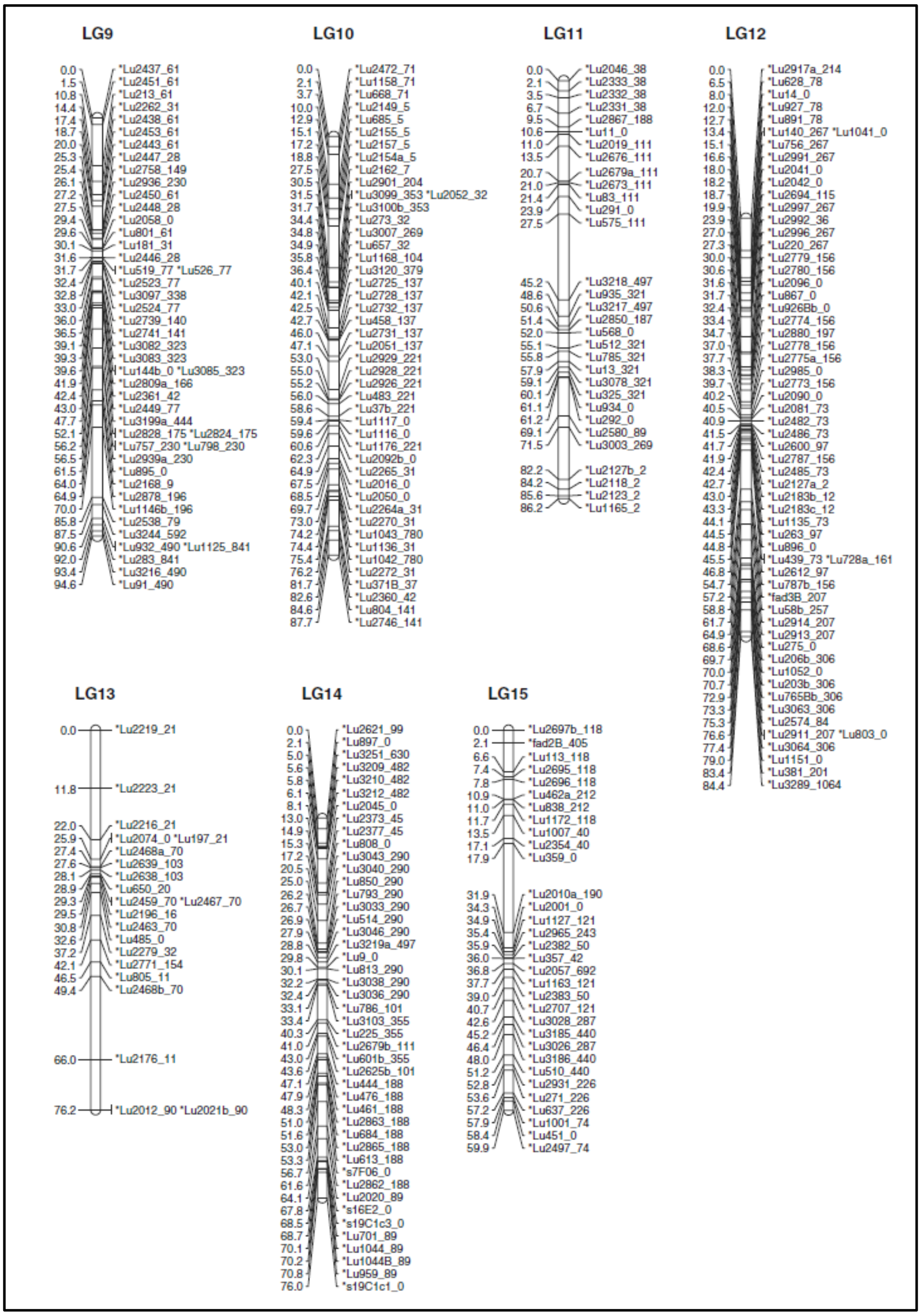


Figure 1.3: Consensus genetic map of linseed integrated from three mapping populations. Numbers to the left of each linkage group represent Kosambi map units (cM). Locus names followed by their FPC contig anchor separated by an underscore are on the right. Linkage groups are in decreasing size order (Cloutier et al., 2012b).

Various diseases like Alternaria, bud fly, pasmo etc. adversely affect linseed yield and developing a resistant variety would certainly help to increase crop productivity. During evolution, plants have developed a repertoire of resistance (*R*) genes containing various conserved domains. Many such genes conferring resistance to a wide range of pathogens and pests have been identified and cloned from numerous plant species (Ellis and Jones 2003). A *L6* gene for linseed rust resistance in linseed was identified and cloned by transposon tagging method (Lawrence et al. 1995). Among various R gene classes, Nucleotide binding site - Leucine rich repeat (NBS-LRR) class is predominant. The NBS-LRR gene contains conserved NBS domain, which have been extensively utilized to identify resistance gene analogs (RGAs) in model plants and various crop species (Yaish et al. 2004; Palomino et al. 2006) and to understand the genomic architecture of this gene family. Genome sequencing efforts of plant species have facilitated genome-level investigation of the NBS-encoding gene family in various plants; for example, *Arabidopsis* (Meyers et al. 2003; Tan et al. 2007), rice (Monosi et al. 2004), *Medicago* (Ameline-Torregrosa et al. 2008), poplar (Kohler et al. 2008), grape (Yang et al. 2008), sorghum (Paterson et al. 2009), papaya (Porter et al. 2009) etc. However, very little is known about the NBS-LRR genes in linseed, as such studies have not yet been reported in this crop. Genomic evaluation of NBS-LRR gene homologs can help in identifying the putative resistance genes and understanding the mechanism of disease resistance in linseed, which in future will help to develop disease resistant variety of linseed.

1.7.2.4 Association mapping: Current status and future prospects

Till now, linkage mapping or QTL mapping, which works on an experimental population derived from a cross of bi-parents divergent for a trait of interest, is the most common approach in plants to detect quantitative trait loci (QTLs) corresponding to complex traits. However, as a biparental mapping population is often derived from a restricted number of meiotic events, the genetic resolution of QTL maps often remains confined to a range of 10-30 cM (Flint-Garcia et al. 2005; Zhu et al. 2008). Moreover, linkage analysis is limited to only the alleles for which the two parents differ, which is very small as compared to the distribution of alleles in natural population. An alternative approach, association mapping (AM) or linkage disequilibrium (LD) mapping can exploit the entire pool of genetic diversity existing in natural populations and overcomes the limitation inherent to linkage mapping. With the intrinsic nature of exploiting historical recombination events, association mapping offers increased mapping resolution to polymorphisms at sequence level and therefore, enhances the efficiency of gene discovery and facilitate marker assisted selection (MAS) in plant breeding (Gupta et al. 2005; Moose and Mumm 2008). Genome-wide association mapping has been successfully applied to elucidate the genetic basis of agronomic traits in rice (Agrama et al. 2007) and maize (Huang et al. 2010), flowering time genes in barley (Stracke et al. 2009), the *PsyI-AI* locus in wheat (Singh et al. 2009), the *rhg-1* gene in soybean (Li et al. 2009), and a series of candidate genes in Arabidopsis (Zhao et al. 2007; Ehrenreich et al. 2009)(reviewed by Zhu et al. (2008). One such study using 390 Canadian flax accessions and 460 SSR markers was carried out in linseed (Soto-Cerda et al. 2013). Based on this, 12 significant SNPs for traits like thousand seed weight, plant branch, lodging, etc. were identified. However, more such studies using high density SNP markers and for various agronomically important traits need to be carried out.

1.8 Genesis of thesis and organization

Linseed is a next generation crop for food, feed and fiber because of its nutraceutical, medical and industrial properties. However, it is a largely neglected crop compared to other oilseed crops. The gradual decrease in annual production supports this fact. This might be because of lack of genetic and genomic resources and also, its susceptibility to various diseases like Alternaria blight, pasmo, budfly, etc. Although recent advancement in technology and active involvement of various researchers have generated many of these, they are still not sufficient for genetic improvement of linseed. The present study was, therefore, undertaken to generate more such resources efficiently using conventional and highly advance technology and computational approaches.

The thesis is organized in seven chapters and the content of each chapter is as follows:

Chapter 1: Introduction and review of literature (Current chapter)

Chapter 2: Development of genomic simple sequence repeat markers for linseed

This chapter describes the strategies used for development of genomic SSR markers in linseed and the distribution of identified motifs.

Chapter 3: Development and characterization of EST-SSR markers for linseed and their comparative analysis with related species

This chapter describes the method used for development of EST-SSR markers in linseed and the comparison of repeat motifs with nine closely related species. Also, transferability of the linseed SSRs in other nine species has been estimated.

Chapter 4: Development of core collection of linseed and analysis of genetic diversity and population structure using SSR markers.

This chapter describes the strategy used for core collection development based on morphological data, evaluation of core collection using various statistical methods and analysis of genetic diversity using SSR markers.

Chapter 5: Genotyping by sequencing in linseed: development of genome-wide SNP markers and genetic diversity and population structure study.

This chapter describes the development of SNPs and genotyping of linseed accessions using genotyping by sequencing technology. Further, population structure and genetic diversity analysis within linseed accessions and finally association mapping to identify significant SNPs for various agronomic traits has also been discussed.

Chapter 6: Genome-wide identification and characterization of nucleotide binding site leucine rich repeat genes in linseed

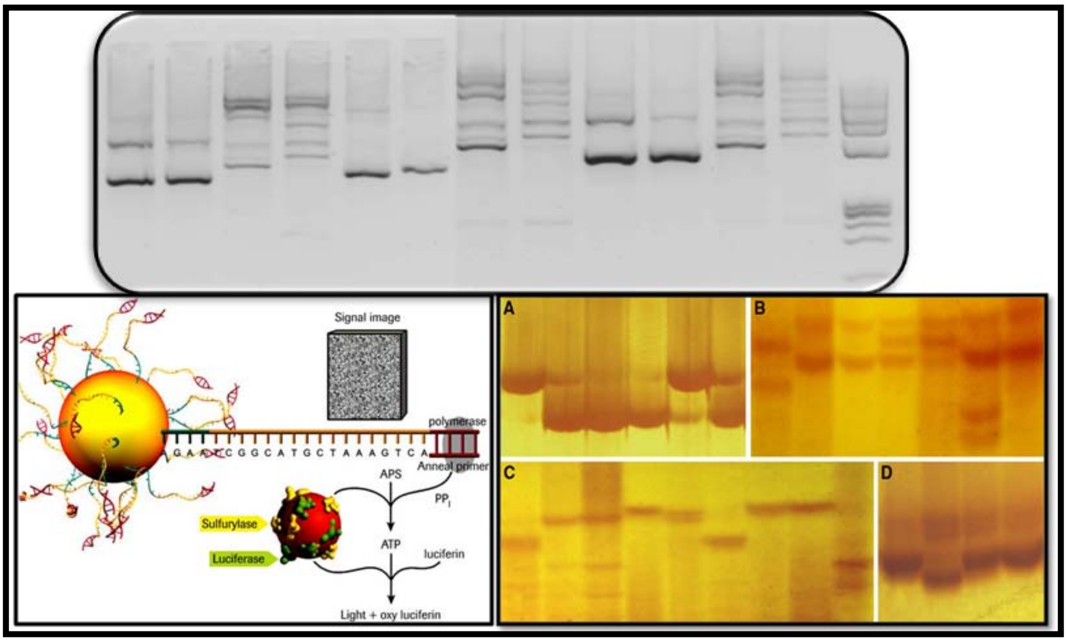
This chapter describes genome-wide identification of NBS-LRR genes in linseed, their phylogenetic and structure analysis. Moreover, the *in vitro* and *in vivo* expression analysis of NBS-LRR genes has also been discussed.

Chapter 7: Summary and future directions

BIBLIOGRAPHY

CHAPTER 2

Development of genomic simple sequence repeat markers for linseed



A part of this work has been published in Molecular Breeding (2012) 30:597–606

2.1 Introduction

Most of the important agronomic traits are complex and polygenic in nature, controlled by several quantitative trait loci (QTLs) (Varshney et al. 2005). Although traditional breeding continues to play an important role in enhancing the yield and quality of crop plants, it is hindered by the long selection cycle and the enormous resources required for the successful outcome. Marker-assisted selection (MAS) has the potential to reduce the time required for the development of new and improved varieties. However, the basic requirement of MAS is the availability of a large number of molecular markers with reasonable level of polymorphism to allow construction of saturated genetic map. A wide range of molecular markers has been developed and used over the past two decades. Among these, microsatellites or simple sequence repeats (SSR) are widely used markers because of their co-dominant, multi-allelic, highly polymorphic nature and easy genotyping (Weber and May 1989). However, generation of SSR markers is technically demanding due to the primary need of their *de novo* development by construction and sequencing of genomic libraries. As there are very few (189) genomic SSRs reported till the initiation of this work (Cloutier et al. 2009; Deng et al. 2010; Roose-Amsaleg et al. 2006; Soto-Cerda et al. 2011a; Wiesner et al. 2001), the main purpose of this study was to develop a large number of useful SSR markers to facilitate future marker based studies in this crop.

The traditional method of isolation of microsatellites by constructing and sequencing a genomic library, is both time and cost consuming, with relatively low efficiency of microsatellite detection. Therefore, various microsatellite enrichment methods were developed for isolation of microsatellites (Squirrell et al. 2003; Zane et al. 2002). Three widely used methods *viz.*, PCR isolation of microsatellite sequences (PIMA, Lunt et al. 1999), 5’-anchored method (Fisher et al. 1996) and Fast Isolation by AFLP of Sequences COntaining repeats (FIASCO, Zane et al. 2002), were evaluated for their efficiency in identifying microsatellites and developing useful SSR markers in linseed. A small modification was done wherein construction of genomic libraries was bypassed and the pooled amplicons generated by these methods were directly sequenced using the 454 GS-FLX next generation sequencing platform (Roche, USA). We chose 454 GS-FLX, principally because of longer read lengths, which are necessary for the *de novo* assembly of microsatellites-rich regions.

Bypassing library preparation and employing the next-generation sequencing technology enabled us to develop a large number of additional SSR markers for linseed in a highly efficient and cost-effective manner. A subset of these markers was experimentally evaluated to demonstrate the practical utility of the method.

2.2 Materials and methods

2.2.1 Plant material and DNA extraction

NL-97, a commercial variety of linseed, was grown at College of Agriculture, Nagpur, India and young leaf tissue was collected. DNA was extracted using the modified CTAB method (Ghosh et al. 2009) and the quality and quantity of DNA were checked visually on 0.8% agarose gel stained with GelRed (Biotium, USA) as well as spectrophotometrically using Nanodrop 1000 spectrophotometer (Thermo Scientific, USA).

2.2.2 Microsatellite enrichment

Three methods, *viz.*, PCR isolation of microsatellite sequences (PIMA, Lunt et al. 1999), 5'-anchored method (Fisher et al. 1996) and Fast Isolation by AFLP of Sequences COntaining repeats (FIASCO, Zane et al. 2002), were used to isolate microsatellite rich genomic regions from linseed. The experimental details are given below.

2.2.2.1 PCR isolation of microsatellite sequences (PIMA)

Amplification of 20 ng DNA was performed in 25 μl reaction volume containing 0.2 mM of each dNTP, 1.5 mM $MgCl_2$, 1X PCR buffer, 0.9U *Taq* DNA polymerase (Banglore Genei, India) and 20 pmoles of RAPD primer. PCR was performed on Verity thermal cycler (Applied Biosystems, USA) using following cycling conditions: initial denaturation at 94°C for 3 min; followed by 40 cycles of denaturation at 94°C for 1 min, annealing at 37°C for 1 min and extension at 72°C for 1 min, with a final extension at 72°C for 7 min. Forty RAPD primers from the series, OPAA, OPAB, OPAK and OPAL (Operon Technologies, USA) were screened. Of these, four primers (*viz.*OPAA7, OPAA9, OPAA14and OPAB20) showed good amplification and were used for isolation of microsatellites. A modification in the original protocol was made to facilitate sequencing of the amplicons using the 454 sequencing platform; instead of cloning and screening the clones with ISSR primers prior to sequencing, the amplicons were enriched using repeat containing 5'-biotinylated probes [$(AG)_{10}$, $(CT)_{10,}$ $(GA)_{10,}$ $(AC)_{10,}$ and $(AGA)_{10}$] and isolated using streptavidin coated magnetic beads (Roche, USA).

2.2.2.2 The 5'-anchored PCR method

Nine degenerate 5'-anchored primers [Flaxdeg1 (5'-KKVRVRV$(AG)_{10}$-3'), Flaxdeg2 (5'-KKVRVRV$(CT)_{10}$-3'), Flaxdeg3 (5'-KKVRVRV$(GA)_{10}$-3'), Flaxdeg4 (5'-KKVRVRV$(AAG)_{10}$-3'), Flaxdeg5 (5'-KKVRVRV$(AGA)_{10}$-3'), Flaxdeg6 (5'-KKVRVRV$(GAA)_{10}$-3'), Flaxdeg7 (5'-KKHBHBH$(AC)_{10}$-3'),Flaxdeg8 (5'-KKHBHBH$(AG)_{10}$-3') and Flaxdeg9 (5'-KKHBHBH$(GA)_{10}$-3'); where K = G/T, V = G/C/A, R = G/A, H = A/C/T and B = G/C/T], with different repeat motifs were used for PCR amplification to capture repeat motifs of the linseed genome.

2.2.2.3 Fast Isolation by AFLP of Sequences COntaining repeats (FIASCO)

About 700 ng genomic DNA of NL-97 was completely digested with *Mse*I and then ligated to an *Mse*I AFLP adaptor (5'-TACTCAGGACTCAT-3' / 5'-GACGATGAGTCCTGAG-3'). The digestion–ligation mixture was amplified with adaptor-specific primer (5'-GATGAGTCCTGAGTAAN-3') using the Verity thermal cycler (Applied Biosystems, USA). The amplified DNA fragments, with a size range of 200–800 bp, were enriched for repeats with 5'-biotinylated probes [$(AG)_{10}$, $(CT)_{10}$, $(GA)_{10}$, $(AC)_{10}$, $(AGA)_{10}$, $(GAA)_{10}$ and $(AAG)_{10}$] using streptavidin coated magnetic beads (Roche, USA). The enriched fragments were amplified again with adaptor-specific primers before sequencing.

2.2.3 Sequencing and comparison of efficiency of microsatellite isolation

The repeat enriched amplicons generated by the above three methods using various probes and primers were mixed **(Table 2.1)** and sequenced using 1/16th area of the PicoTiter Plate device on the 454 GS-FLX platform at Eurofins MWG Operon (Germany). The sequence reads were sorted by the enrichment method using the sequences of adaptor-specific primers (FIASCO) and RAPD primers (PIMA). The Cd-hit-est (Li and Godzik 2006) software was used to obtain unique sequences within each of these methods at a threshold of 0.95 and the SSR containing sequences were identified using MISA (http://pgrc.ipk-gatersleben.de/misa/misa.html). Comparison among the three enrichment methods was performed on the basis of the number of unique sequences obtained, number of microsatellite containing sequences and the number of sequences suitable for development of SSR primers.

Table 2.1: Sample used for amplicon sequencing

Sr No.	Modified PIMA		FIASCO		5'Anchored PCR	
	Repeat Motif	**Amplicon (ng)**	**Repeat Motif**	**Amplicon (ng)**	**Repeat Motif**	**Amplicon (ng)**
1	$(AC)_{10}$	700	$(AC)_{10}$	700	$(AC)_{10}$	700
2	$(AG)_{10}$	700	$(AG)_{10}$	700	$(AG)_{10}$	700
3	$(GA)_{10}$	1000	$(GA)_{10}$	1000	$(GA)_{10}$	1000
4	$(CT)_{10}$	700	$(CT)_{10}$	700	$(CT)_{10}$	700
5	$(GAA)_7$	700	$(GAA)_7$	700	$(GAA)_7$	700
6			$(AAG)_7$	700	$(AAG)_7$	700
7			$(AGA)_7$	700	$(AGA)_7$	700
	Total	**3800**	**Total**	**5200**	**Total**	**5200**

2.2.4 Identification and classification of microsatellites

The sequence reads were assembled using CAP3 software (Huang and Madan 1999) with default parameters, except for the following: base quality cut-off: 12, overlap length cut-off: 40 and overlap percent identity cut-off: 80. The resulting contigs and singlets were then used for microsatellite detection using SSR Locator software (Carlos daMaia et al. 2008). The SSRs were selected using the criteria of≥8 repeats for dinucleotide motifs and≥5 repeats for other repeat classes (tri-, tetra-, penta-, and hexanucleotides). Mononucleotide repeats were not considered due to the difficulty of distinguishing *bona fide* microsatellites from sequencing or assembly errors that are associated with the 454 sequencing method. The SSR containing sequences were analyzed for repeat class and length and categorized as described by Katti et al. (2001). In brief, (AC)n was considered equivalent to (CA)n, (TG)n, and (GT)n. Thus, four classes were possible for dinucleotide repeats (DNR), 10 for trinucleotide repeats (TNR), and 33 for tetranucleotide repeats. Individual repeat frequencies were determined for all of these classes in the sequences selected for primer designing.

2.2.5 Designing SSR primers and evaluation of polymorphism

SSR primers were designed from the assembled sequences of repeat containing contigs and singlets (GenBank accession No. HQ883990-HQ888681) **(Table S2.1)** using SSRLocator software according to the program’s default parameters, with the following exceptions: preferred product size range: 150 to 400 bp, optimum primer size: 18, melting temperature (Tm): 50-60 °C and GC content: 40-60%. Virtual PCR was carried out with the same software using the sequences as template with three PCR simulations and the best primers were selected. About 10-15% primers were randomly selected from DNR and TNR classes. These motifs/classes were selected based on their abundance in the primers designed. In case of TNRs, the primers were designed from the sequences belonging to five classes, thus representing minimum 50% of total TNR classes. The primers were synthesized (Eurofins MWG Operon, Germany) for 52 contigs and singlets with different microsatellite motifs [DNR:- class-2: CT (1), AG (3), class-3: AC (9); TNR:- class-1: AAT (2), class-2: AAG (10), AGA (19), class-3: AAC (1), ACA (2), class-6: GAG (1), class-9: ACC (1); tetra nucleotide repeats:- AAAG (1), AAGA (1) and pentanucleotide repeats:- AAATC (1)]. These primers were empirically evaluated for reproducible amplification using standard PCR conditions with appropriate annealing temperatures using the DNA of NL-97 linseed variety and later validated using a set of 27 diverse linseed genotypes **(Table 2.2)**. The selection of the genotypes was based on the analysis of 95 linseed accessions using phenotypic and ISSR data (data not shown). The amplicons were separated on 6% polyacrylamide gel and visualized using silver staining (Sanguinetti et al. 1994). The sequences used for SSR primer development were also searched against the NCBI database to identify novel SSRs.

Table 2.2: List of genotypes used for polymorphism screening of designed SSRs

Sr. No.	Name of genotypes	Sr. No.	Name of genotypes	Sr. No.	Name of genotypes
1	EC22596	**10**	EC41599	**19**	EC5873
2	EC13220	**11**	EC41526	**20**	EC41525
3	EC22715	**12**	EC41576	**21**	EC41585
4	EC12082	**13**	EC41615	**22**	NL-260
5	EC41636	**14**	EC41623	**23**	AYOGI
6	EC41577	**15**	EC41495	**24**	NL-97
7	EC23572	**16**	EC41528	**25**	JRF-5
8	EC41559	**17**	EC22684	**26**	NEELUM
9	EC41572	**18**	EC29006	**27**	PADMINI

2.3 Results

2.3.1 Comparison of microsatellites enrichment methods

Three methods were evaluated for isolation of microsatellite sequences from the linseed genome. The probes and primers used for microsatellites isolation contained the same repeat units **(Table 2.1)** in the three methods, except the PIMA method wherein two trinucleotide probes (AAG and AGA) were not used. The amplicons produced by the PIMA and the 5'-anchored methods generated multiple bands in the range of 200-1,000 bp, while smear in the range of 200-1,200 bp was observed in the FIASCO method, which was then size selected to obtain fragments in the range of 200-800 bp. Pooled samples from all the three methods **(Table 2.1)** were used for nucleotide sequencing.

When the sequence reads were sorted according to the enrichment method, 73% of them were from 5'-anchored method, 24% from FIASCO method while only 3%were contributed by the PIMA method **(Table 2.3).** The Cd-hit-est software was used to remove redundant sequences obtained using each of the enrichment methods separately and 41%, 12% and 2% unique sequences resulted from the three enrichment methods, respectively. Among these sequences, the percentages of repeat containing sequences for the three methods were 56%, 14% and 14%, respectively **(Table 2.3),** while the actual numbers of SSRs obtained were 2,198, 569 and 4, respectively. When these repeat containing sequences were analyzed for suitability of primer development, 30% of the repeats from FIASCO method and 18% from 5'-anchored method contained sufficient flanking regions for primer development **(Table 2.3).**

Table 2.3: Comparison of efficiency of microsatellite isolation methods

Sr No.	**Enrichment method**	**Sequences obtained**	**Redundant sequences**	**Unique sequences**	**SSR containing sequences**	**Number of SSRs identified**	**Sequences suitable for designing primers**
1	**PIMA**	816 (2.5%)	795 (97.42%)	21 (2.57%)	3 (14.28%)	4	1
2	**5'-anchored PCR**	26,631 (73.29%)	23,462 (88.10%)	3,169 (11.89%)	1,803 (56.89%)	2,198	328 (18.19%)
3	**FIASCO**	8,885 (24.45%)	5,275 (59.36%)	3,610 (40.63%)	529 (14.63%)	569	169 (30%)
	Total	**36,332**	**29,532**	**6,800**	**2,335**	**2,771**	**498**

2.3.2 Microsatellite mining

A total of 36,332 reads with an average size of 232 bp and totalling 8.8Mb **(Table 2.4)** were obtained after sequencing, which were assembled into 2,183 contigs and 2,509 singlets. This assembly totaled to 1.17 Mb, representing approximately 0.17% of the estimated size of the linseed genome (Bennett 2005). The length of the contigs and singlets ranged from 100 bp to 1,387 bp with an average of 249 bp. For mining the SSR containing sequences, two software, MISA and SSR Locator, were used which produced similar results. However, as SSRLocator can also be used for primer designing and virtual PCR, we exploited it for further analysis. About 30% (1,442) of the sequences harboured SSR motifs in which, 97.5% (1,407) contained perfect repeats, while only 2.5% (35) contained compound repeats. Overall, the sequences contained 1,842 SSR motifs withdinucleotides (54%, 1,004) as the most abundant repeat type followed by trinucleotides (44%, 827) and other repeats (0.59%, 11). The highest SSR length frequencies observed were: 20 (27%) and 21 (20%) bp **(Figure 2.1).** Forty nine percent of the DNRs had 8-10 repeat units, while 50% had 11 or more repeat units within the respective loci. In case of TNR, most of the loci had 7-24 repeat units (71%), while 28% had 5-6 repeat units.

Table 2.4: Sequence data generated and microsatellites characterized

Source DNA	**NL-97 linseed variety**
Sequence data generated	8.8 Mb
Sequence reads	36,332
Assembled contigs	2,183
Singlets	2,509
Total size of contigs and singlets	1.17 Mb
Sequences containing microsatellites	1,442
Sequences containing perfect microsatellites	1,407
Sequences containing compound microsatellites	35
Sequences with amplifiable microsatellite regions	350
Total Number of SSR primers designed	290

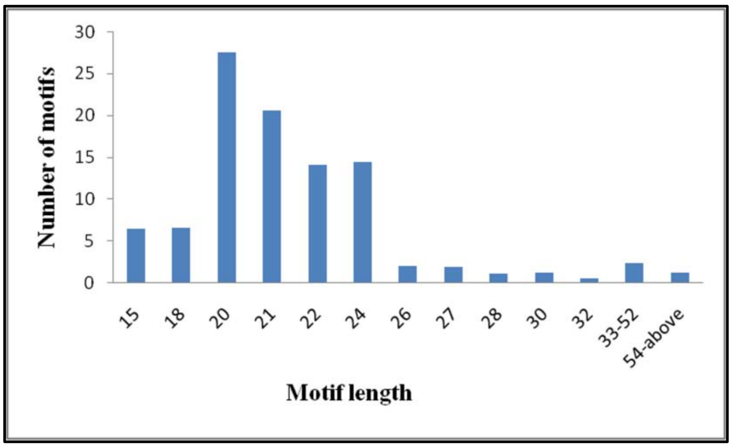


Figure 2.1: Frequency distribution of the linseed genomic SSRs based on the motif length (motif length and number of repeats)

2.3.3 Development of SSR markers

Of the 1,842 microsatellite loci identified, flanking primers could not be designed for 76% of the microsatellite sequences because of the lack of suitable flanking sequence or Tm constraints. In all, 350 sequences (24%) contained the microsatellite motifs that were flanked by the sequences suitable for designing primers. However, 60 of these were eliminated due to Tm constraints or having too high or too low GC. Thus, 290 primer pairs were designed **(Table S2.1),** among which, 298 SSR motifs were present (TNR-60%, DNR-39% and others-2%). Among the DNR motifs, class-2: AG was the most abundant (53%), whereas class-1: AT and class-4: GC motifs were absent **(Table 2.5).** In the class-2, the frequency of $(AG)_n$ motif was maximum (26%); whereas in class-3, maximum frequency of 20% was observed for $(AC)_n$ **(Figure 2.2a).** Similarly, among the TNR motifs, class-2: AAG was predominant (63%) **(Table 2.5 and Figure 2.2b)** and class-10: GGC had the least frequency **(Table 2.5).** We also observed considerable non-targeted TNR repeats in the following classes; class-1: AAT, class-3: AAC, class-4: ATG, class-5: AGT and class-6: AGG **(Table 2.5).** Ninety-five per cent of the SSR containing sequences did not find any similarity with the NCBI database and thus were considered novel.

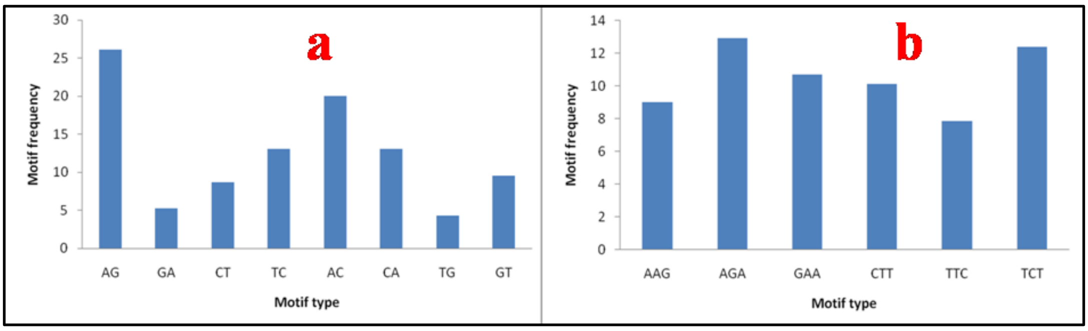


Figure 2.2: Frequencies of the motifs of **a)** dinucleotide class-2 and -3 **b)** trinucleotide class-3 in sequences used for primer development

Table 2.5: Frequencies of the dinucleotide and trinucleotide repeat motifs in sequences used for primer designing

Class	Motif type	Percentage
	Dinucleotide (DNR)	
1	AT/TA	0.0
2	AG/GA/CT/TC	53.0
3	AC/CA/TG/GT	47.0
4	GC/CG	0.0
	Trinucleotide (TNR)	
1	AAT/ATA/TAA/ATT/TTA/TAT	3.9
2	AAG/AGA/GAA/CTT/TTC/TCT	62.9
3	AAC/ACA/CAA/GTT/TTG/TGT	5.6
4	ATG/TGA/GAT/CAT/ATC/TCA	6.7
5	AGT/GTA/TAG/ACT/CTA/TAC	1.7
6	AGG/GGA/GAG/CCT/CTC/TCC	10.1
7	AGC/GCA/CAG/GCT/CTG/TGC	2.2
8	ACG/CGA/GAC/CGT/GTC/TCG	3.4
9	ACC/CCA/CAC/GGT/GTG/TGG	2.8
10	GGC/GCG/CGG/GCC/CCG/CGC	0.6

In virtual PCR analysis, most of the 290 primer pairs showed single amplicons. However, 11 primer pairs showed non-specific amplifications, indicating less than 100% complimentarity of the primer binding sites to the target sequences. In order to corroborate the virtual PCR results, wet-lab PCR experiments were performed. For this, 52 primer pairs **(Table S2.1)** were randomly selected, synthesized and used for amplification of genomic DNAs of 27 diverse linseed genotypes **(Table 2.2).** Allelic amplification was obtained for 43 markers across the analyzed genotypes. Of the amplified primers, 31 (72%) contained TNR motifs and 10 (23%) contained DNR motifs, whereas nine primers (21%) were polymorphic **(Table S2.1 and Figure 2.3)**.

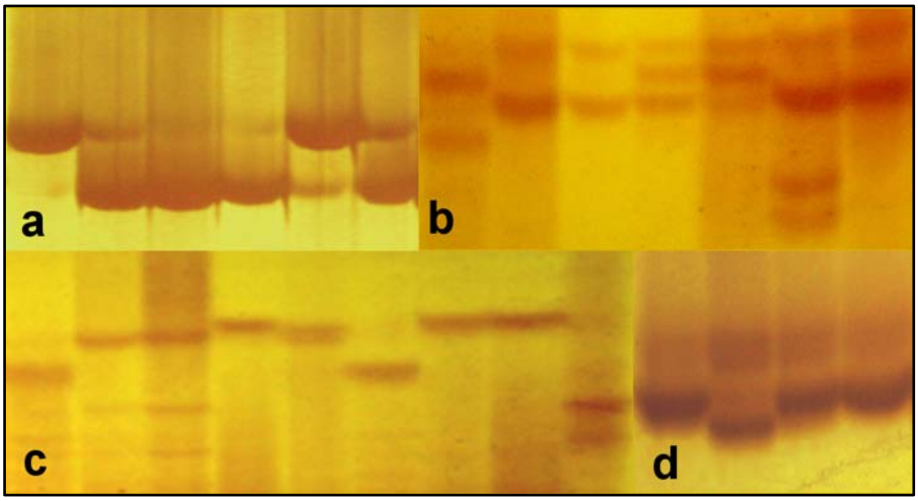


Figure2.3: Allele patterns obtained from four genomic SSR markers a) NCL_29,b) NCL_32, c) NCL_34, and d) NCL_39

2.4 Discussion

Molecular markers are powerful tools for gaining insight into the inheritance of complex quantitative characters, and are being used for both, the dissection of complex agronomic traits as well as development of marker-assisted breeding strategies. Till date, several types of molecular markers have been used in linseed (Bickel et al. 2011; Cloutier et al. 2009; Cloutier et al. 2011; Deng et al. 2010; Roose-Amsaleg et al. 2006; Soto-Cerda et al. 2011a; Wiesner et al. 2001); however, their numbers are quite limited. Among the various markers, SSRs are preferred in linkage mapping, gene tagging and QTL studies, because of their abundance, genome-wide distribution, co-dominance and highly polymorphic nature. In addition, they may show high inter-species transferability and thus are widely utilized in plant genomics studies (Collard et al. 2005; He et al. 2003; Varshney et al. 2005). However, till the initiation of this work, only 189 genomic SSR and 248 EST-SSR markers were reported in linseed, which are very limited (Cloutier et al. 2009; Deng et al. 2010; Roose-Amsaleg et al. 2006; Soto-Cerda et al. 2011a; Wiesner et al. 2001). The EST-SSR markers reveal lower polymorphism compared to genomic SSRs and hence are not as efficient as the latter for distinguishing closely related genotypes (Joshi et al. 1999). Although Rajwade et al. (2010) showed that ISSR markers could be effectively used to distinguish the Indian linseed varieties, they are dominant in nature and more polymorphic co-dominant markers would be beneficial. Therefore, the present study was aimed at establishing new linseed specific SSR markers using the next generation sequencing technology.

Construction and sequencing of SSR-enriched genomic libraries has been the principal and efficient means of discovering genomic SSRs (Zane et al. 2002). From previous study of diversity analysis in linseed (Rajwade et al. 2010) in our laboratory, the ISSR primers with $(GA)_n$ repeats were found to be highly polymorphic. In addition, Cloutier et al. (2009) also observed that $(AG)_n$ and $(GA)_n$ dinucleotide repeats and $(GAA)_n$, $(AAG)_n$ and $(AGA)_n$ trinucleotide repeats were predominant in EST-SSRs from linseed. We, therefore, used these repeat motifs to design degenerate primers and probes to capture a large number of repeats from the linseed genome by employing three microsatellites enrichment methods.

The next generation sequencing technology has directly or indirectly enabled a rapid progress in the development of SSR markers in several plant species. A large number of EST-SSRs have been identified in pigeonpea (Singh et al. 2011), chickpea (Hiremath et al. 2011), lentil (Kaur et al. 2011) etc. by analysis of the transcriptome sequence data generated by the next generation sequencing technology. Santana et al. (2009) used this technology to develop genomic microsatellite markers for three unrelated organisms: *Fusarium circinatum, Sirexnoctilio* and *Deladenussiricidicol* using two microsatellite enriched libraries. Their study indicated that the cost of generating amplifiable microsatellite and unique sequences through pyrosequencing was 276% and 62% lower than that of Sanger sequencing for *F. circinatum*. To the best of our knowledge, this was the first report where the next generation sequencing technology was used for identification of microsatellites from a plant, linseed (*Linum usitatissimum* L.) using three methods of enrichment and bypassing the development of genomic libraries.

2.4.1 Comparison of efficiency of the microsatellites enrichment methods

Surprisingly, a large number of sequences were obtained from 5’-anchored method (73%); although the FIASCO method produced higher number of unique sequences; suggesting that the 5’-anchored (88%) and PIMA (97%) methods showed higher redundancy. These results are in conformity with those obtained by Tang et al. (2009), wherein 75% redundancy was observed using 5’-anchored method. However, direct comparison of these results is not possible due to the differences in the number of sequences analyzed. Moreover, PIMA and 5’-anchored methods are known to suffer from redundancies as has been observed in other studies (KÖlliker et al. 2001; Rallo et al. 2000; Squirrell et al. 2003). Thus, we found that the 5’-anchored method was the most efficient for isolation of repeat containing sequences; whereas, FIASCO was the most effective for successful primer development. Similar results have also been observed in other cases (Santana et al. 2009). This is largely because of the use of degenerate repeat primers in the 5’-anchored method, which leads to efficient capturing of more number of microsatellite containing sequences; however, the lack of sufficient region on one side of the repeat prevents designing specific flanking primers (Fisher et al. 1996; Tang et al. 2009).

2.4.2 Development of SSR markers

Assembly of the 36,332 reads resulted in 2,183 contigs and 2,509 singlets and 30% of them harboured SSRs. As expected, most of the microsatellite motifs were located at the terminal region of the sequences and hence were rendered unsuitable for marker development. The sequences having suitable flanking regions for primer development were analyzed further using the SSR Locator software. Total 290 SSR markers were developed, which are much larger in number than all the previously reported genomic SSR markers for linseed (Deng et al. 2010; Roose-Amsaleg et al. 2006; Soto-Cerda et al. 2011a; Wiesner et al. 2001). The SSR containing sequences were searched against the NCBI nucleotide database restricted to linseed, wherein about 95% of the sequences did not yield matches and thus were considered as novel.

2.4.3 The occurrence and features of SSRs

A total of 1,842 SSR motifs were detected in 1,442 repeat containing sequences and most of the repeats were perfect repeats (97%). The dinucleotide repeats were higher than trinucleotide repeats, which is consistent with the results from *A. thaliana*(Katti et al. 2001), cucumber (Cavagnaro et al. 2010) and eight underutilised crop species (Jae-Woong et al. 2009). The relative abundance of DNRs and TNRs detected is also a function of the SSR searching criteria and the software used for SSR database mining, which could have partly contributed to the observed differences (Aggarwal et al. 2007; Varshney et al. 2005). Besides DNR and TNR motifs, other motifs were also observed; however, their frequency was insignificant. This might be because the tetra-nucleotide probes were not used in the enrichment procedure and were therefore, non-targeted motifs. Cloutier et al. (2009) also obtained a small proportion (0.5%) of tetra-nucleotide repeats in EST-SSR studies in linseed. It is also possible that these repeats might be rare in the linseed genome; however, a further study using these probes is necessary to confirm this observation.

Among the 298 motifs detected in the sequences used for marker development, a high representation of motifs belonging to the classes AG, AC and AAG was observed, since we used these motifs for microsatellite isolation. However, in addition to these classes, many other classes were also observed. Gimenes et al. (2007) also observed 37% SSRs with different repeats that were not totally complementary to the oligonucelotide probes used. However, the classes GC and GGC were not represented in our study, which could be due to the limitation in isolation of these repeats because of their self-complementary nature. Within the TNRs, AAG class was predominant as also reported by Toth et al. (2000) and Katti et al. (2001). In dinucleotide motifs, AG motif (26%) was present in high percentage and this is consistent with the results from most of the plant species (Cardle et al. 2000; McCouch et al. 1997; Newcomb 2006; Powell et al. 1996; Rabello et al. 2005). However, it needs to be seen if the results obtained in this study using a small fraction of the linseed genome could generally be applicable to similar analysis using the whole genome sequence of linseed. It would also reveal if the strategy used in this study is efficient enough to develop SSR markers for other non-model species, whose genomes are yet to be sequenced. Also, the over-representation of sequences from the 5’-anchored method and under-representation of sequences from PIMA method is intriguing. Independent sequencing of the amplicons obtained by the three methods will reveal if this is associated with any of the microsatellite isolation method or there are some other reasons. A true comparison of the efficiency of the three methods could be made by sequencing and analyzing the amplicons separately.

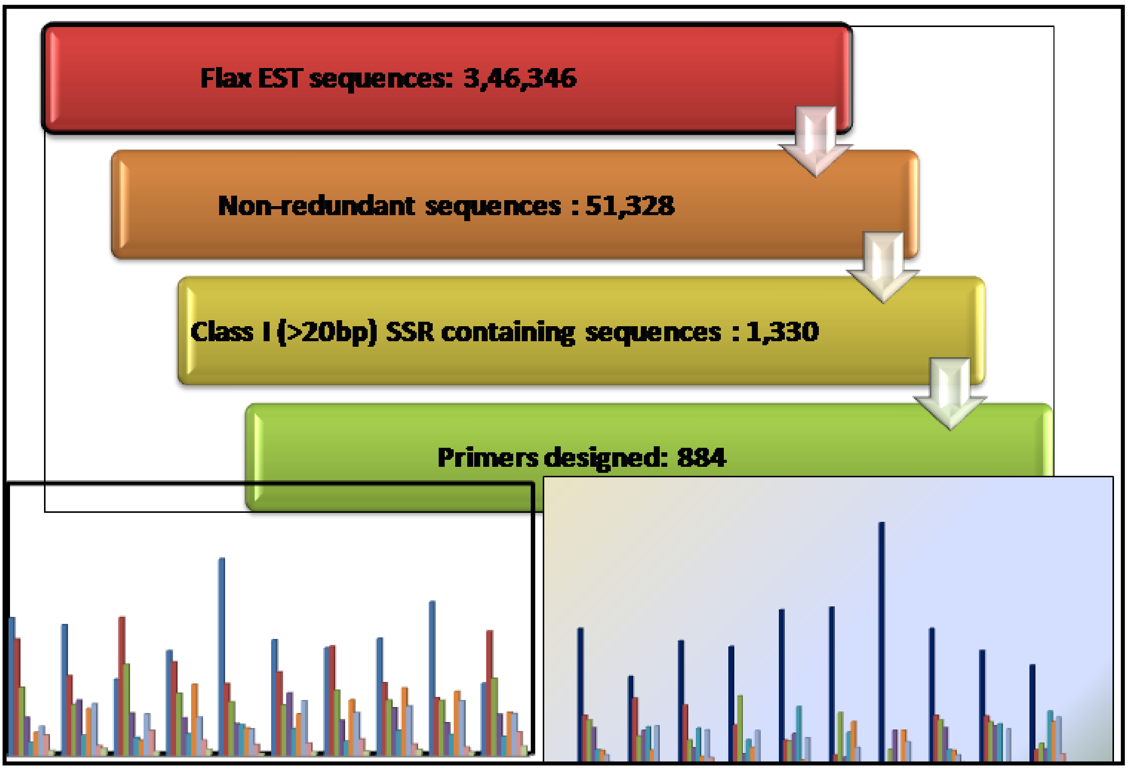
2.4.4 Practical utility of the developed SSR markers in linseed

To evaluate the practical utility of the sequenced microsatellite regions for development of the polymorphic markers, 52 primer pairs were synthesized and used for amplifying the DNA of 27 diverse linseed genotypes. About 82% primers produced amplifications, wherein the trinucleotide motifs showed the highest success rate (59%).The polymorphism observed in the present study (21%) was lower than that reported earlier in linseed for genomic SSRs (36%, Bickel et al., 2011**;** and 39.8%, Deng et al. 2010) and other oil seed plant such as peanut (32%, He et al. 2003). Several possible reasons could contribute to this:(i) low genetic diversity within the genotypes analyzed, (ii) the level of polymorphism of SSRs often increases with increasing SSR length and the number of repeat units, as observed in studies on maize, pepper and rice (McCouch et al. 1997; Sharopova et al. 2002) and (iii) in general, DNR markers are more polymorphic than TNR markers (Gadaleta et al. 2006). In the present study, 28% of the amplified primers were of $(TNR)_5$ type and only 25% primers were targeted towards dinucleotide repeats.

In conclusion, our study demonstrated that the 454 sequencing technology was much faster and cost effective for microsatellite discovery than cloning and sequencing individual libraries, which is substantially expensive, time consuming and delivers few microsatellite markers. Using this technology, we developed a large number of genomic microsatellite markers for linseed, which is greater than the total number of SSR markers reported till date. Thus, a large number of new microsatellite markers are now available for linseed genomic and genetic research.

CHAPTER 3

Development and characterization of EST-SSR markers for linseed and their comparative analysis with related species



Flax EST sequences: 3,46,346

Non-redundant sequences : 51,328

Class I (>20bp) SSR containing sequences : 1,330

Primers designed: 884

This work has been communicated to

Bioinformation

3.1 Introduction

Rapid adoption of next generation sequencing technologies and increasing emphasis on functional genomics have led to the development of large datasets of ESTs from various plant species. With evolving bioinformatics tools, it is now possible to identify and develop EST-SSR markers at a large scale in a time and cost-effective manner. EST-SSRs also have a higher probability of being in linkage disequilibrium with genes/QTLs controlling economic traits, making them more useful in studies involving marker-trait association, QTL mapping and genetic diversity analysis (Gupta et al. 2003). Increasing numbers of EST-SSR markers are now being identified and used for a variety of applications in a number of plant species like grapes (Scott et al. 2000), sugarcane (Cordeiro et al. 2001) and cereals such as wheat, barley, rye and rice (Varshney et al. 2005). Unlike genomic SSRs, they may be used across a number of related species. EST resources can also be used for genome to genome comparisons and synteny analysis, which can serve several functions, including identification of gene-clusters, potentially syntenic chromosomal regions and single nucleotide polymorphism within the compared sequences. For example, a comparison of soybean ESTs with those from corn, rice, sorghum, barley, potato, tomato and *Medicago* (Schlueter et al. 2004) focused on evolutionary distance and synteny between ESTs. In addition, soybeans ESTs have been compared with lupin and Arabidopsis (Francki and Mullan 2004) detailing gene structure and expression. Moreover, it is possible to transfer these markers from model to orphan species for their precise improvement, e.g. wheat EST-SSRs showed high transferability to a wide range of plant species including orphan wild species such as *Agropyrum* (Zhang et al. 2005).

In the present study, the linseed EST dataset was mined for identification, development and characterization of SSR markers in linseed. The primers were designed for class I EST-SSRs and their transferability was evaluated using virtual PCR and few were experimentally validated using diverse linseed genotypes. Further, linseed ESTs along with ESTs from nine closely related plant species were surveyed for occurrence pattern of SSRs and their comparison was performed.

3.2 Materials and methods

3.2.1 Plant material and DNA extraction

Twelve genetically diverse linseed accessions **(Table 3.1)** were selected for validation of the synthesized primers. DNA was extracted from young leaves using the modified CTAB method (Ghosh et al. 2009) and the quality and quantity of DNA was checked visually on 0.8% agarose gel stained with GelRed (Biotium, USA) as well as spectrophotometrically using a Nanodrop 1000 spectrophotometer (Thermo Scientific, USA).

Table 3.1: Origin information of linseed accessions used in this study

Sr. No.	Accession Name	Origin
1	A-375	Australia, LLR
2	BR-1	Sabour, Bihar, India 6 x NP 121, Primitive cultivar
3	CI-1427	Ethiopia, LLR
4	LCK-152	Kanpur, UP, India Sel3 x EC 1552
5	LMH-379	Mauranipur, UP, India LC 216 x BR1, BL
6	NO-356	New Pusa, New Delhi, India Local selection
7	NP-59	New Pusa, New Delhi, India Local selection
8	NPHY-29	New Pusa, New Delhi, India No. 3 x NP 48, PC
9	NPHY-38	New Pusa, New Delhi, India NP 59 x No. 55, PC
10	NPHY-39	New Pusa, New Delhi, India NP 8x NP 103, PC
11	RKY-14	Not available
12	SHIKHA	Kanpur, UP, IndiaHira x Crista, RC

3.2.2 EST sequence sources and processing

The EST sequences of *Linum usitatissimum* (linseed) were retrieved from dbEST database (2, 86,882) of NCBI (http://www.ncbi.nlm.nih.gov/). Additional unigenes (59,626) were downloaded from http://urgi. versailles. inra. fr/index. php/urgi/Species site. The EST sequences were initially processed to remove any vector sequences, adaptors and low quality sequences using SeqTrim (Falgueras et al. 2010) and assembled using CAP3 software (Huang and Madan 1999), with default parameters. The assembled contigs and singletons were combined with the unigenes (59,626) and again assembled as mentioned above. The resultant non-redundant dataset of EST sequences was used for further analysis. The non-redundancy of assembled sequences was confirmed using cd-hit-est software (Weizhong and Adam 2006), at 90%

sequence identity. The BLASTX analysis was performed with all linseed non redundant EST dataset to identify closely related species and top 9 species *viz. Ricinus communis* (castor bean), *Populus trichocarpa* (black cottonwood, poplar), *Arabidopsis thaliana* (thale cress, Arabidopsis), *Vitis vinifera* (grape), *Glycine max* (soybean), *Gossypium hirsutum* (cotton), *Medicago truncatula* (medick or burclover, Medicago), *Jatropha curcas* (Jatropha) and *Nicotiana tabacum* (tobacco) were selected for comparative analysis. The EST sequences of these plant species were retrieved from dbEST database of NCBI and assembled as described above separately for each species.

3.2.3 Detection and characterization of SSR motifs

The SSRLocator software (Carlos daMaia et al. 2008) was used to identify SSR loci, to analyze motif statistics, amino acid distribution and to design primers from linseed. The mononucleotide repeats were not considered in order to reduce the sequence ambiguity due to sequencing errors. The parameters for SSR identifications were adjusted to have the motif length ≥ 15 bp . The repeats obtained were classified as described by Katti et al. (2001). For example, dinucleotide repeats were divided into four classes while trinucleotide repeats into 10 classes.

3.2.4 Transferability and validation of EST-SSRs

To check transferability of the linseed EST-SSRs, virtual PCR was carried out using EST sequences from the selected 9 species mentioned above as targets using the e-PCR software (Rotmistrovsky et al. 2004). The results were analyzed for the variation in product size and percentage polymorphism.

Thirty primers were randomly selected and synthesized (50% TNR and 50% DNR) (Eurofins MWG Operon, Germany). These primers were empirically evaluated for reproducible amplification using standard PCR conditions with appropriate annealing temperatures and later validated using a set of 12 diverse linseed genotypes selected based on the analysis of 222 linseed core germplasm of India using phenotypic and SSR data (data not shown). The amplicons were separated on 6% polyacrylamide gel and visualized using silver staining (Sanguinetti et al. 1994).

3.3 Results

3.3.1 Search for linseed EST containing SSR motifs

In this study, 2, 86,882 ESTs and 59,626 unigenes sequences of linseed were used for the analysis of microsatellites. All the GenBank ESTs (2, 86,882) were first assembled into contigs and singletons which were again assembled along with the 59,626 unigenes **(Figure 3.1)**. This combined assembly resulted in 31,928 contigs and 19, 400 singletons which totaled ≈ 30.4 Mb representing approximately 4.34% of the estimated 700 Mb flax genome (Benneth and Leitch2005) and revealed 6.8 fold redundancy in the EST dataset.

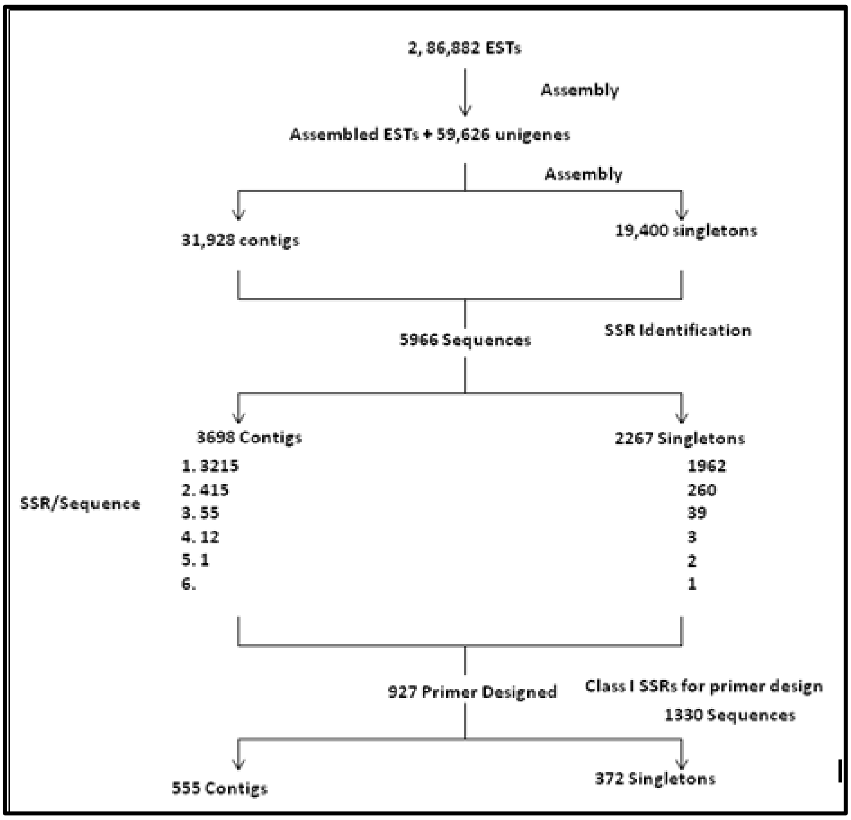


Figure 3.1: Diagram illustrating the EST assembly, the SSR identification and the primer design of the flax EST-SSRs

SSRLocator detected 5,966 SSRs containing sequences from 51,328 unigenes, of which majority (5,177/5,966, 86.8%) of the sequences had a single putative SSR, while 788 (13.2%) of them had 2, 3, 4, 5 or 6 putative SSRs per sequence **(Figure**

3.1) for a total of 6, 890 putative SSRs. This approach allowed the identification of one putative SSR in every eight examined sequences (11.62% of 51,328 unigenes) and the frequency of occurrence for EST-SSRs averaged one SSR per 16.5 kb of EST sequence. The average frequency of SSR motif in the linseed sequence collection was 1.99% **(Table 3.2)**. The number of repeat motifs ranged from 2 to 45. Two was the most frequent repeat number followed by six and then five. An inverse relation between the frequencies of putative EST-SSRs was found with increasing number of repeats. The distribution pattern of EST-SSRs based on the length of SSRs (motif size x number of repeats) was also studied. The highest frequency was found for SSRs that were 15 (29.8%), 16 (22.6%) and 18 (16.4%) nucleotides in length. Frequency decreased dramatically for SSR lengths of motif above 24 bases **(Figure 3.2a and 3.2b)**. The occurrence of individual types of repeat motifs in linseed ESTs also varied with the trinucleotide motifs (TNR) being the most abundant (43.66%) followed by octanucleotides (16%) with few representations of deca and heptanucleotide motifs **(Figure 3.3)**. Among the TNR motifs, class B (AAG) was predominant (33%) and class H (ACG) had the lowest frequency (2%). Within, class B, the frequency of the (GAA) n motif was the largest (20.68%).

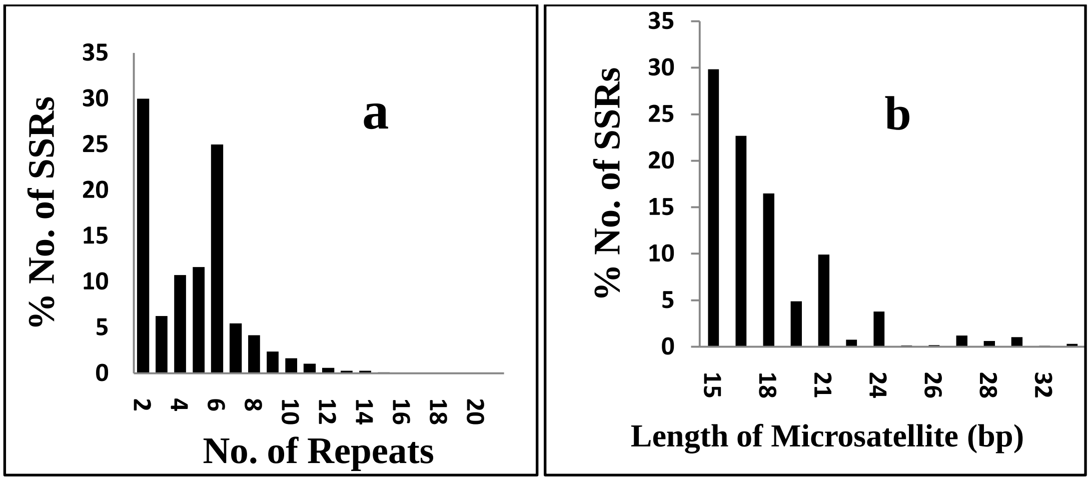


Figure3.2: Frequency distribution of the putative EST-SSRs from flax ESTs based a) number of repeats and b) length of microsatellite (motif length X numbers of repeats)

Table 3.2: EST database size, non-redundant sequences generated, avg. bp count, % SSR/EST data in 10 species analyzed

Species	EST database count (%)	Contigs	Singletons	Total non-redundant seq.	Number of SSR loci	Avg. bp count per ESTs	SSR/EST database (%)
Arabidopsis thaliana	15,29,700 (32.059)	22,063	47,226	69,289	2,314	120	0.151
Glycine max	14,61,624 (30.632)	43,227	83,907	1,27,134	15,153	746	1.036
Gossypiumhirsutum	2,73,779 (5.7378)	20,574	25,644	46,218	8,830	751	3.225
Jatropha curcas	42,747 (0.89588)	5,940	9,950	15,890	2,742	839	6.414
Linum usitatissimum	3,46,508 (7.262)	31,928	19,400	51,328	6,890	596	1.99
Medicago truncatula	2,69,238 (5.642)	11,339	13,911	25,250	4,908	564	1.822
Nicotiana tabacum	3,32,667 (6.971)	37,363	1,00,968	1,38,331	14,984	746	4.504
Populus trichocarpa	8,9,943 (1.88)	11,865	13,488	25,353	4,900	555	5.447
Ricinus communis	62,592 (1.3117)	5,449	7,869	13,318	3,753	540	5.995
Vitis vinifera	3,62,674 (7.6)	23,323	54,594	77,917	15,654	521	4.31
Total	47,71,472 (100)	1,81,143	3,57,557	5,38,700	80,128		

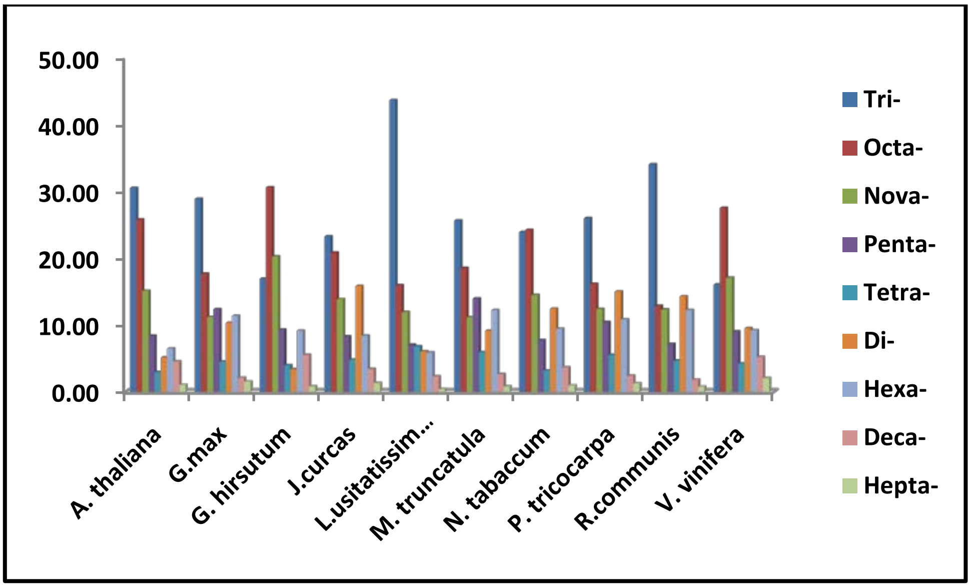


Figure 3.3: Distribution of repeat motifs in species studied

3.3.2 Primer development, transferability and validation of EST-SSRs

As class-I SSRs (≥20bp; 1,340) are more polymorphic/ informative than class-II SSRs (<20bp - ≥15bp), they were considered for primer designing resulting in development of 927 primer pairs (**Table S3.1**). Virtual PCR was carried out using these primers on non-redundant EST sequences of selected nine species to check their transferability. Percentage of primers showing amplification ranged from 0.00 to 7.11, the highest being from Tobacco while the lowest was from Jatropha **(Table 3.3)**. When the expected product size calculated from linseed ESTs was compared with observed product size using ESTs from target species, Arabidopsis, grape, tobacco, poplar and Medicago showed 100% deviation from expected product size. In castor, maximum 15.38% sequences had same product size as expected, followed by soybean (4.55%) and cotton (3.85%) **(Figure 3.4).**

Table 3.3: e-PCR to show transferability of linseed EST-SSRs across 9 species studied

Species	Positive primers	Transferability (%)
Jatropha curcas	**0**	**0.00**
Arabidopsis thaliana	**3**	**0.32**
Ricinus communis	**13**	**1.40**
Medicago truncatula	**22**	**2.37**
Populus trichocarpa	**22**	**2.37**
Glycine max	**22**	**2.37**
Gossypiumhirsutum	**52**	**5.60**
Vitis vinifera	**60**	**6.47**
Nicotiana tabacum	**66**	**7.11**

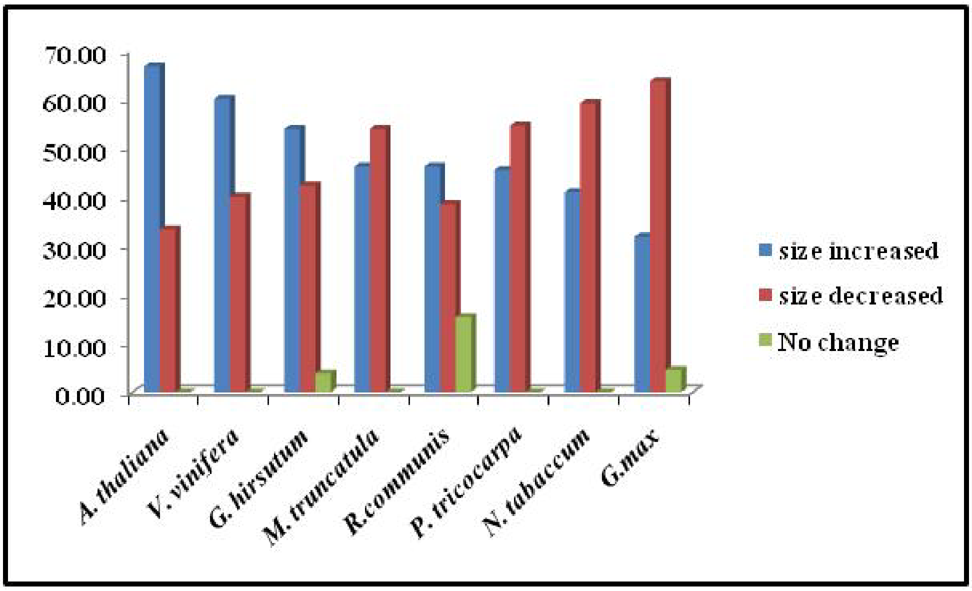


Figure 3.4: Deviation in virtual PCR product size observed in different species from that observed in linseed

In order to corroborate the virtual PCR results, the wet-lab PCR experiments were performed. For this, 30 randomly selected primer pairs **(Table S3.2)** out of the 927markers designed **(Table S3.1)** in this study were synthesized, and used for amplification of genomic DNAs of 12 diverse linseed accessions**.** Twenty-seven EST-SSRs produced repeatable and reliable amplifications, eight of which were located in ORF regions. Out of seven polymorphic markers, only one was located in ORF. Interestingly, seven polymorphic EST-SSR marker loci except one had an associated putative function related to stress responses or other hypothetical functions. Two amplicons seemed to include intron sequences (as deduced from a product amplification size larger than the one expected from the published EST sequence), which were larger than 500 bp. Of the amplified primers, 15 (55%) contained TNR motifs and 12 (45%) contained DNR motifs. However, DNR motif containing primers were more polymorphic compared with that of TNR motif **(Table S3. 1).**

3.3.3 Comparative EST-SSR analysis among selected 10 species

A total of 47, 71,472 ESTs from 10 plant species were analyzed in this study, wherein the highest sequences were from Arabidopsis (32.05%) while the lowest were from Jatropha (0.9%) **(Table 3.2)**. The average bp count per EST sequence ranged from 120-839 bp, with Jatropha having the highest value while Arabidopsis having the lowest bp count per EST. The frequency of SSRs per EST ranged from 6.4% to 0.15%, in which the Jatropha had the highest (6.4%) and Arabidopsis had the lowest frequency (0.15%) **(Table 3.2)**. Total 12,122 sequences contained more than one SSR. Among the plant species, castor and Arabidopsis showed extreme distributions with 31.31% and 7.00% sequences containing one or more repeat motifs, respectively. The average motif length was 18.20 bp, wherein castor (18.82 bp) showed the highest while it was the lowest in Arabidopsis (17.29 bp) **(Table 3.4)**.

Total 80,128 SSR loci were identified from 10 species, in which 16,644 were of class I type; while 63,546 were of class II type. Out of 80,128 loci, 76,747 loci were perfect while 1,397 loci were compound repeats. Maximum percentage of compound repeats was observed in grape (3.10%); while the minimum was in linseed (0.76%) **(Table3. 4)**.

Table 3.4: SSR containing ESTs, overall occurrence of SSRs, percentage and avg. length motif

Species	**EST sequences with SSRs (%)**	**Seq. containing more than one repeat motif (%)**	**Sequences containing perfect repeat motif (%)**	**Sequences containing compound repeat motif (%)**	**Avg. motif length**
Arabidopsis thaliana	2158 (3.11)	7.00	97.88	1.04	17.29
Glycine max	13,201 (10.38)	14.56	93.54	1.43	17.78
Gossypium hirsutum	7,498 (16.22)	17.56	97.90	1.04	17.66
Jatropha curcas	2,373 (14.93)	17.66	95.62	2.19	18.78
Linum usitatissimum	5,966 (11.62)	13.86	97.71	0.76	17.79
Medicago truncatula	4,090 (16.2)	19.61	97.23	1.34	18.24
Nicotiana tabacum	13,307 (9.62)	12.37	97.18	1.37	18.52
Populus trichocarpa	4,144 (16.35)	17.88	95.82	2.04	18.63
Ricinus communis	2,855 (21.44)	31.31	94.78	2.56	18.82
Vitis vinifera	12,226 (15.69)	27.83	93.67	3.10	18.55

Presence of di- to deca-nucleotide repeat motifs was identified in all the species. Among all the motifs analyzed, tri and octa-nucleotide repeat motifs were predominantly present in all the species except grape and cotton; whereas heptanucleotide repeat motifs were unanimously low in all the species analyzed. The abundance pattern of rest of repeat motifs varied among the 10 species. Among all, in linseed, the trinucleotide repeat motifs were most abundant while the frequency of penta, hexa and hepta-nucleotide repeats was the lowest. Interestingly, Medicago, castor and grape showed comparatively higher abundance of the penta, hexa and hepta-nucleotide repeat motifs, respectively **(Figure 3.3)**.

Further, the DNR and TNR motifs were classified into four and ten classes, respectively as described by Katti et al. (2001) **(Figures 3.5 and 3.6)**. The distribution of DNR motif classes followed the same pattern in all the 10 species. The BB (AG/GA) class repeat was the most predominant, followed by AA (AT/TA) and CC (AC/CA) classes. Surprisingly, the DNR motif class DD (GC/CG) was completely absent in all the species. The frequency of BB class was the highest in Arabidopsis and the lowest in cotton. Conversely, the frequency of AA class was the highest in cotton and the lowest in Arabidopsis whereas the frequency of CC class was the highest in cotton and the lowest in grape. Overall, the repeat TC was predominant in dinucleotide. Among the TNR classes, the frequency of class B was the highest in all the species and similarity in the distribution pattern for other classes was also observed. The frequency of B class was the highest in Arabidopsis and the lowest in poplar whereas significant difference in the frequencies for class A was observed in all the ten species. In majority of the cases, the classes B, G, D, F and A each contributed >10% of total repeat motifs. The classes J, E and H were among the lowest contributors, contributing from 1% to 7% additively. The Arabidopsis, grape and Jatropha showed the lowest frequencies while soybean, Medicago and poplar showed the highest frequencies for classes J, E and H, respectively.

The predicted amino acid content for SSR loci observed in the 10 species is shown in (**Figure 3.7)**. The distribution pattern of 20 amino acids was similar in all the species studied; except Jatropha and grape. In all the species except these two, the frequency of leucine was the highest; while in Jatropha and grape, the frequency of serine was the highest. Among the 20 amino acids, the frequency of metheonine was the lowest in the 10 species **(Figure 3.7)**.

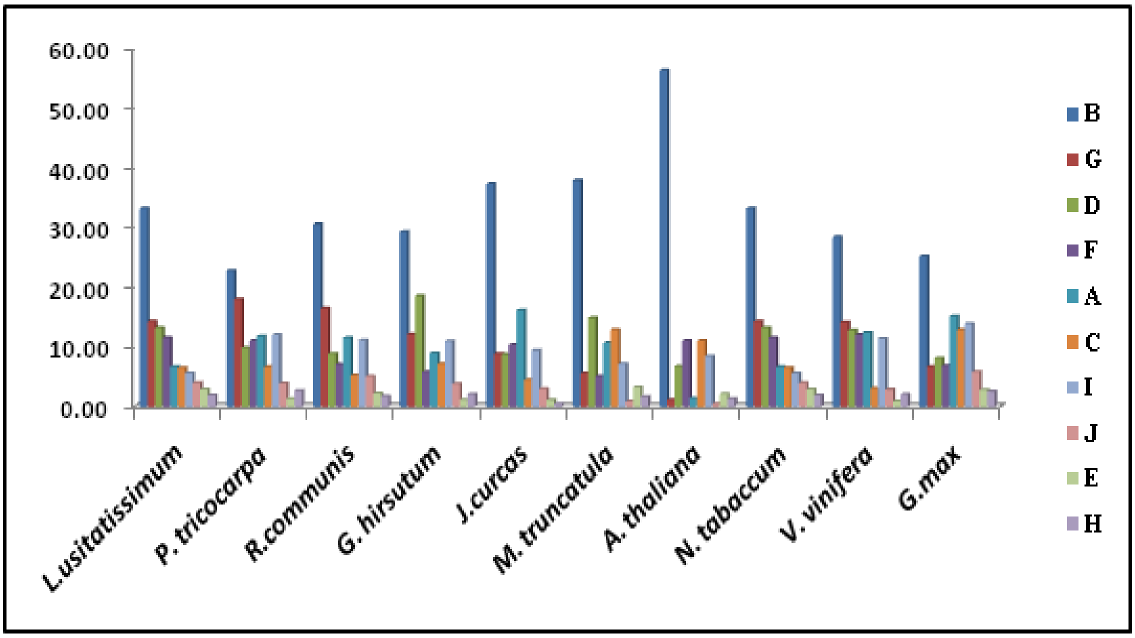


Figure 3.5: Distribution of trinucleotide repeats motifs into 10 different trinucleotide classes

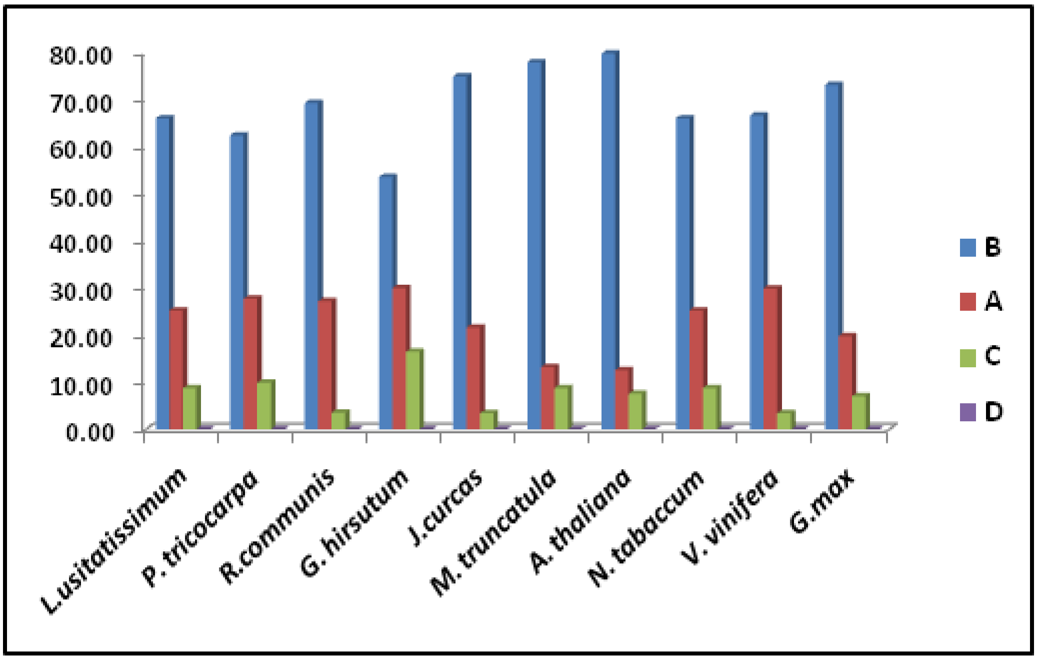


Figure 3.6: Distribution of dinucleotide repeat motifs into 4 different trinucleotide classes

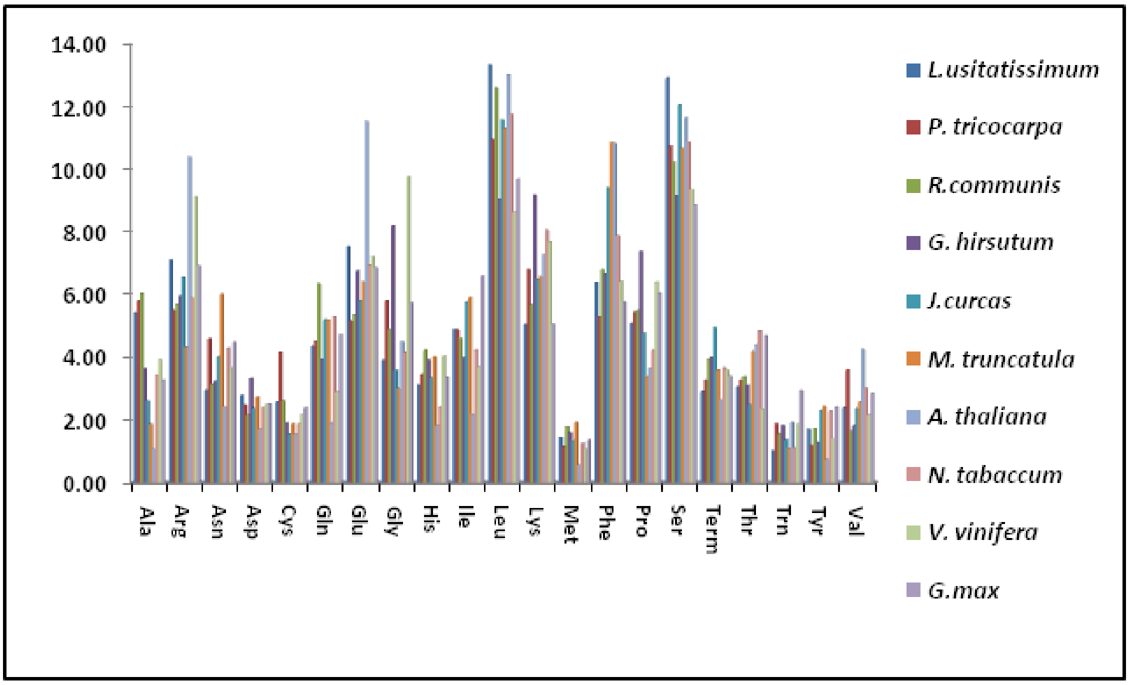


Figure 3.7: Predicted amino acid occurrences in SSR loci within plant species studied

3.4 Discussion

Linseed is an important oilseed crop because of presence of medicinally important ALA and lignan. However, the high yielding varieties are susceptible to various biotic and abiotic stresses. Therefore, there is a need for genetic improvement of this crop. Classical breeding is being used for this purpose since ancient times and it is well-known that the molecular marker technology in combination with classical breeding reduces the time and effort required for genetic improvement of any plant species. Availability of large number of molecular markers is the prerequisite for the development of saturated linkage maps which in turn helps in the genetic improvement of the species. In linseed, various molecular markers as EST-SSRs (Cloutier et al. 2009; Soto-Cerda et al. 2011a), genomic SSRs (Deng et al. 2010; Roose-Amsaleg et al. 2006; Soto-Cerda et al. 2011b), AFLP (Everaert et al. 2001), RFLP (Spielmeyer et al. 1998) and ISSR (Rajwade et al. 2010; Wiesner et al. 2001) have been developed, however, the number is not sufficient to develop a saturated map required to identify QTLs. Three genetic maps were developed using AFLP, RFLP, RAPD and SSR markers (Cloutier et al. 2011; Oh et al. 2000; Spielmeyer et al. 1998), but very few markers were used in map construction and hence coverage was not enough to identify promising QTLs for various biotic and abiotic stresses. It is reported that environmental factors affect plant height; weight and DNA content of linseed varieties (Cullis 1973) indicating a need for genetic analysis of abiotic stress response. Among all the markers, abundant nature of microsatellites makes them the most popular markers. Although genomic SSRs are extensively utilized, they occur in a non-coding regions of the genome which are less conserved thereby exhibit less transferability across species and hence cannot be utilized to study species within genus and other related species. It is also difficult to understand the evolutionary process using these markers as they show homoplasy, where identical band sizes may not be identical by descent (Thiel et al. 2003).

One of the major impediments in the molecular analyses of the linseed is the generation of informative as well as transferable molecular markers. The ability of EST-SSRs to associate with many phenotypes makes them potential markers to identify phenotypic and genetic changes. They are cost-effective, co-dominant and useful for studying functional diversity, comparative mapping, candidate gene mapping and interspecific linkage map construction (Varshney et al. 2005). Moreover, EST sequences are conserved among species as they are derived from coding regions and exhibit a high rate of transferability to other species (Ellis and Burke 2007; Eujayl et al. 2002; Iniguez-Luy et al. 2008; Varshney et al. 2005). The characterization of EST-SSRs within and between different plant families could facilitate developing genetic markers from model plants to orphan species.

In linseed the occurrence and characteristics of EST-SSRs were first reported by Cloutier et al. 2009 which were later utilized for the microsatellite based linkage map development (Cloutier et al. 2011). Some of them were also used to characterize their transferability to *Linum* species (Fu and Peterson 2010; Soto-Cerda et al. 2011a). Such study is important as little is known about the evolutionary relationship among *Linum* species (approximately 200) and its taxonomic position is still unclear. Thus, EST-SSRs collection provides new molecular tools for *Linum* studies. Nonetheless, they are still limited in number. Therefore, to understand genetics of linseed, more informative genic SSRs should be developed. In the present study, the available linseed EST resources were used to develop more genic SSR for this species. At the same time, comparative analysis of linseed genic SSR was performed with 9 closely related species. Such comparative analysis would be useful to understand evolution as well as for future studies of transferability of molecular markers.

3.4.1 Distribution of EST-SSR in linseed

Total 51, 328 unigenes were deduced from a total of 3,46,508 sequences analyzed, indicating 6.8 fold redundancy in the EST dataset which is similar to what was found by Cloutier et al. , 2009. However, it was less than that reported by Venglat et al., 2011 (8.9), possibly due to inclusion of unigenes during assembly. Out of the 51,328 unigene set examined, 11.62% sequences contained 6,890 genic SSRs which account for an average density of one SSR per 4.52 Kb. Similar studies conducted by Cloutier et al. (2009) detected 3.5% SSR containing ESTs. The discrepancy could be due to different dataset, software and parameters used. However, it is similar to other dicot species (2.65% - 16.82%) (Kumpatla and Mukhopadhyay 2005). The overall percentage of EST-SSR sequences was 1. 99% which matches the expected range as it is estimated that 2-5% of all plant-derived ESTs harbor SSRs (Pashley et al. 2006; Varshney et al. 2005). Among all the motifs analyzed in linseed, the trinucleotides (43%) were the most abundant followed by octa- (15.92%) and Nova- (11.87%). The high abundance of trinucleotides might result in their positive selection since they do not cause frame shift mutation in coding region. The TNR class B was the most abundant with the predominance of GAA motif (20.7%). The results are in accordance with those obtained by Cloutier et al. (2009) and similar distribution was noticed in most of the plant species studied so far (Gong et al. 2010; Kumpatla and Mukhopadhyay 2005; Roorkiwal and Sharma 2011). However, predominance of other trimeric repeats was also reported in other dicot plant species (Mahalakshmi et al. 2002; Rabello et al. 2005; Thiel et al. 2003; Varshney et al. 2002). The variation in length of an SSR motif was also considered for its usefulness as a marker (Varshney et al. 2002). Hence, only class-I repeats were selected for the development of SSRs and 927 markers were developed in the present study.

3.4.2 *In silico* transferability of linseed EST-SSR markers across selected plant species

As EST SSRs originate from expressed regions and hence are more conserved across a number of related species, across-species transferability of EST-SSRs is greater than genomic SSRs (Varshney et al. 2005). However, linseed EST-SSR primers showed lower success rate of transferability across the plant species tested. This might be because of the presence of large number of unique genes. Analysis of linseed transcriptome revealed that nearly one fifth of the linseed transcriptome is unique (Venglat et al. 2011). The low transferability rate might also result from the stringency of the virtual PCR conditions. Surprisingly, the transferability to castor and poplar, plant species closely related to linseed, was low while maximum for Tobacco. Although, Tobacco showed 100% deviation in expected product sizes while castor had maximum expected product sizes. The reason for this is unclear at present. It is quite possible that the currently available EST dataset of castor and poplar is not sufficient to cover the complete genome and hence the low transfer rate. However, similar low transferability across the related species was also observed by Victoria et al. (2011).

To evaluate the practical utility of the developed SSRs, 30 primer pairs were synthesized and amplified in a panel of 12 diverse linseed accessions. About 90% of primers produced amplifications, in which the trinucleotide motifs showed the highest success rate (55%). The polymorphism observed in the present study (23%) was lower than that reported earlier in linseed for EST as well as genomic SSRs (37.5 % Cloutier et al. , 2009; 53.5%, Soto-Cerda et al. , 2011a. 36%, Bickel et al. 2011; and 39.8%, Deng et al. 2010). However, the results are comparable with our previous studies conducted using genomic SSRs. The primary and most important reason for low polymorphism could be due to low genetic diversity within the Indian genotypes. In general, DNR markers are more polymorphic than the TNR markers (Thiel et al. 2003). In the present study, 55% of the amplified primers were of the $(TNR)_5$ type, while 45% primers were targeted towards dinucleotide repeats. Low variability of trimeric EST-SSR loci was also reported in *Oryza sativa* (Cho et al. 2000) and *Pinustaeda* (Liewlaksaneeyanawin et al. 2004). On the contrary, dimeric EST-SSRs with high numbers of repeats seem to have high polymorphism, as do genomic SSRs. The relationship between polymorphism and the number of repeats has been reported for EST-SSRs in barley (Thiel et al. 2003) and G-SSRs in cultivated linseed (Soto-Cerda et al. 2011a).

3.4.3 Patterns of EST –SSRs distribution in transcript data of plant species

In the present study, the survey was carried out for occurrence, distribution pattern and comparative analysis of EST-SSR sequences using the non-redundant EST data from ten plant species. The average bp count per EST ranged from 120-839 bp, with Arabidopsis having the lowest bp count per EST. The shorter bp size of Arabidopsis

EST could be because of incomplete representation of genes, errors in sequencing, sequencing technology used and parameters used for sequence assembly (Goldberg et al. 2006). However, it should be taken into consideration that average gene sizes could vary among species, e. g. rice fl-cDNAs (1,747 bp) are 14% longer than Arabidopsis fl-cDNAs (1,532 bp) (Victoria et al. 2011). The overall bp count was comparable to other studies in plants (Von Stackelberg et al. 2006). The average motif length for Arabidopsis was 17.29, which was lower than the reported (26.5 bp). This could be because of difference in number of sequences analyzed, type of repeats analyzed and parameters used for repeat identification. The number of sequences containing more than one SSRs were in range of 7.0% (*Arabidopsis thaliana*) - 31.31% (*Ricinus communis*). The Arabidopsis value was higher than the reported (3.46%) possibly due to differences in the parameters used for SSR identification (da Maia et al. 2009). Overall, the range is consistent with the report by Victoria et al. (2011) (3.46% to 37.34%). Around 0.76% to 3.26% loci were compound repeats and maximum were from linseed. About ~4% of compound repeats were observed in higher plants (Victoria et al. 2011) supporting the observation in the present study.

3.4.4 Common pattern of motif distribution among selected plant species

The identified microsatellite loci in 10 species were compared to understand the distribution pattern of repeat motifs among the species. Trinucleotides are reported to be present at higher frequency in higher plants (da Maia et al. 2009; Lawson and Zhang 2006; Morgante et al. 2002). The same trend was observed in the present study, with the frequency of trinucleotide repeat motif being the highest in all except grape and cotton. The frequency of dinucleotide repeat motif ranges from 3.39% to 15.79%, which was lower than that of octa and nova-nucleotide repeat motif frequency. This might be because of parameters used for repeat identification and also, could be due the observation that EST sequences contain higher frequencies of Tri, hexa and nova-nucleotide repeat motifs (Jiang et al. 2006; Lawson and Zhang 2006; Tóth et al. 2000). The di-nucleotide repeat motifs and trinucleotide repeat motifs were classified into 4 and 10 classes as described by Katti et al. (2001). Among the dinucleotide repeat motifs, repeat motif TC was found to be predominant while repeat motifs GC/CG were completely absent. The same distribution was also reported by Moccia et al. (2009) in *White Campion* and by Riju et al. (2009) in cacao plants. Among trinucleotide repeat motif classes, the class B (AAG/AGA/GAA) was predominant in all the species as also observed by Cloutier et al. (2009) and Victoria et al. (2011).

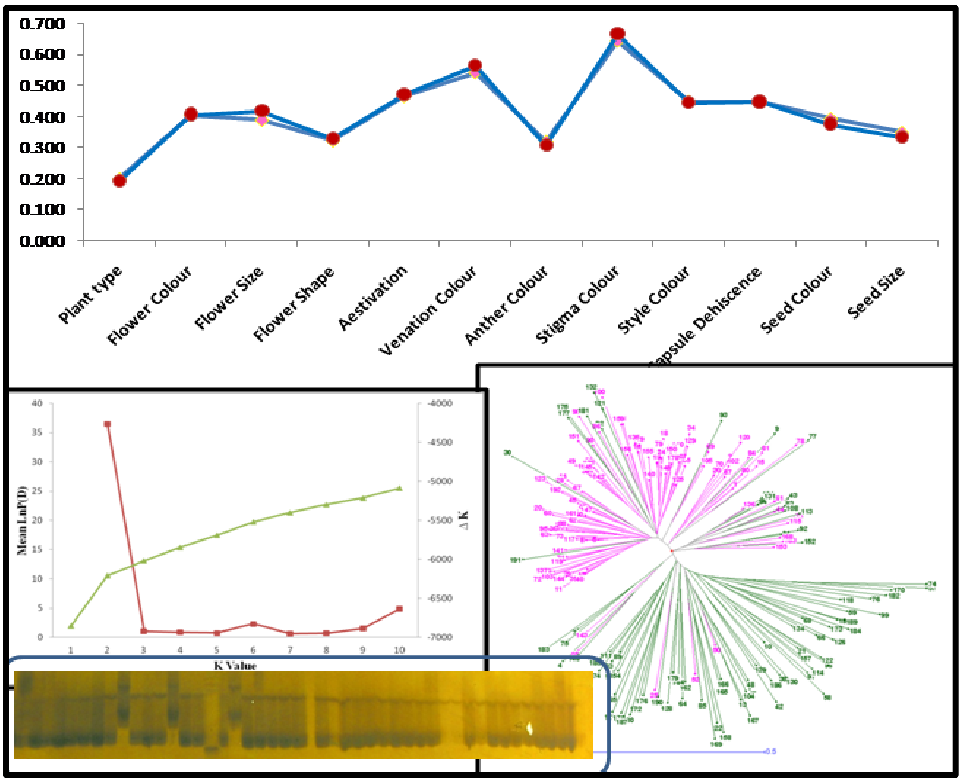
Overall, the frequency of tetra-, penta- and hexa-nucleotide repeats was very low in all the plant genomes investigated in the present study. In almost all the species, the frequency of penta nucleotide repeat was higher than that of the other tetra- and hexa- nucleotides. This is contradictory to the observation made by Victoria et al., (2011) in eleven species where both tetra and penta were rare while hexa nucleotide repeat motifs were in abundance in higher plants. Variation in the distribution could be related to strategy used by researcher (software, repeat number and motif type) (da Maia et al. 2009). There was no similarity in the pattern of distribution of tetra, penta and hexa-nucleotide repeat motif frequency. This could be due to variation in the sizes of the database and genome for each species. However, studies on grass genome indicated no correlation of genome size and tandem repeat loci content (da Maia et al. 2009).

The distribution of amino acids coding for SSRs were assessed using SSRLocator software. The amino acids leucine, serine, glutamic acid, phenyl-alanine, arginine and glycine were predominant agreeing with the previous reports from flowering plants (da Maia et al. 2009; Jung et al. 2005; La Rota et al. 2007; Parida et al. 2006). This is because the codons for these amino acids are GC rich and GC rich repeats were found to be predominant in the EST sequences (Victoria et al. 2011). In previous study, arginine was reported as the predominant amino acid in Arabidopsis (10.46%) whereas in the present study, leucine was observed as predominant amino acid, while the frequency of arginine was similar as reported earlier. This might be because of different parameters used for SSR identification.

In summary, the *in silico* data mining was found to be useful, efficient and inexpensive method to develop EST-SSRs in linseed and provided information about the frequency, type and distribution of EST-SSRs in linseed genome. The developed markers add to the existing genomic tools of linseed and would be useful for various applications such as genetic mapping, diversity analysis and Marker Assisted Selection (MAS) and related plant species. A lot of genomic research programs on linseed are being carried out to develop genomic resources for it as it is commercially grown worldwide for oil and fiber and contains many nutraceutical compounds. Therefore, the developed SSR primers will be useful for these programs for in depth study of linseed genome.

CHAPTER 4

Development of core collection of linseed and analysis of genetic diversity and population structure using SSR markers



This work has been communicated to *Indian Journal of Genetics and Plant Breeding*

4.1 Introduction

Availability of a diverse germplasm collection is a primary requirement of conventional breeding for enhanced production and quality to select diverse genotypes as potential parents for development of new varieties and mapping populations. Further, a diverse germplasm collection is also essential to develop a panel for association mapping (AM) studies. Large collections of linseed germplasm have been reported from different countries such as Germany (http://fox-serv.ipk-gatersleben.de/), Russia (Zhuchenko and Rozhmina 2000), United States of America etc., including the world collection of linseed maintained by Plant Gene Resources of Canada (PGRC). Diederichsen (2007) reported the presence of more than 45,000 *Linum* accessions worldwide. Establishment of a core collection (CC) has proven to be a favored approach to facilitate efficient exploration of novel variation from the available genetic resources (Ellis et al. 1998; Holbrook et al. 2000; Malvar et al. 2004). However, there have been limited efforts to establish a CC for linseed. During 1998-2001, the Centre for Genetic Resources, Netherlands (CGN) has developed a CC of 83 accessions from 947 accessions of both, fiber flax and linseed.

India also has a large collection of linseed germplasm (2,239 accessions) maintained at the Project Coordinating Unit (Linseed), C.S. Azad University of Agriculture & Technology Campus, Kanpur. Inspite of this impressive collection of linseed germplasm, there has been limited use of these accessions in genetic improvement of Indian linseed; thereby the extent of genetic diversity present in Indian germplasm collection remains unknown. Estimation of the genetic diversity is a critical step for any crop improvement program for which morphological and molecular markers as well as agronomic and eco-geographical traits have been widely used. However, the phenotypic characters controlled by polygenes and influenced by environment, may not reflect the true genetic diversity within populations. Molecular markers overcome these limitations and hence they have rapidly become popular to aid construction of CC based on the underlying genetic diversity (Li et al. 2004a). In case of linseed, genetic diversity has been analyzed using a variety of marker systems, including RAPD (Diederichsen and Fu 2006; Fu et al. 2003a), ISSR (Rajwade et al. 2010; Wiesnerova and Wiesner 2004), SSRs (Cloutier et al. 2009; Fu 2011; Soto-Cerda et al. 2012) and retrotransposon based markers (Smykal et al. 2011). However, use of molecular markers on such a huge collection is impractical. In such case, a CC which is true representative of entire collection would be beneficial.

The present study details construction of a linseed CC using 2,239 germplasm accessions based on eight quantitative morphological traits. The homogeneity, distribution pattern, trait-associations and diversity of the CC were evaluated using chi-square test, *z*-test, Wilcoxon rank-sum test and Shannon’s diversity indices using data of twelve qualitative characters. The CC was analyzed for contents of five fatty acids and evaluated using simple sequence repeat (SSR) markers to characterize the population structure and genetic diversity.

4.2 Materials and methods

4.2.1 Plant material

A total of 2,239 linseed accessions were analyzed in this study; among which, 1,890 accessions were of Indian origin including indigenous collection, local landraces and breeding lines; while 349 accessions were of exotic origin. The accessions were grown in single rows of 3m length with a row spacing of 40 cm and plant to plant spacing of 6-8 cm at Germplasm Maintenance and Use (GMU) unit, C.S. Azad University of Agriculture & Technology, Kanpur, India. Recommended agronomic practices were followed to raise a healthy crop.

4.2.2 Analysis of morphological and agronomic traits

The data on eight quantitative traits, *viz*. days to 50% flowering (DTF), days to maturity (DTM), plant height (cm) (PH), technical plant height (cm) (TPH), capsules per plants (CPP), seeds per capsule (SPC), 1000 seeds weight (g) (TW) and seed yield per plant (g) (YPP) and twelve qualitative traits, *viz*. plant type, flower color, flower size, flower shape, aestivation type, venation color, anther color, stigma color, style color, capsule dehiscence, seed color and seed size were recorded. The data on qualitative traits were recorded as per the approved DUS (distinctness, uniformity and stability) test guidelines for linseed (http://agricoop.nic.in/SeedTestguide/linseed1.htm).

4.2.3 Construction of core collection

The quantitative data of eight morphological traits were used for construction of the CC from a total of 2,239 germplasm accessions. The CC was constructed as described by Upadhyaya et al. (2003). The data were first standardized to have the mean 0 and standard deviation 1. The standardized data were used for calculating a distance matrix using Euclidean distance method and clustering was performed using agglomerative hierarchical model based on Ward's minimum variance method (Ward 1963). From each cluster, approximately 10% accessions were randomly selected for inclusion into the CC.

4.2.4 Evaluation of the core collection

The means of the entire germplasm collection and CC for the eight quantitative morphological traits used for construction of the CC were compared using t-tests. The CC was also evaluated for mean difference percentage (MD) and coincidence rate of range (CR) parameters using the following formulae (Hu et al. 2000):

$MD = \left({}^{S_t}\!/_{n}\right) \times 100$ where S_t is the number of traits which have a significant difference (α=0.05) between their means in the initial collection and in the CC and *n* is total number of traits, and

$CR = \frac{1}{n}\sum_{i=1}^{n}\frac{R_{C(i)}}{R_{I(i)}}\times 100$ where $R_{C\ (i)}$ is the range of the i^{th} trait in the CC; $R_{I\ (i)}$ is the range of the corresponding trait in the entire germplasm collection and *n* is total number of traits. The CC is considered to be representative of the initial collection under the following conditions: (1) no more than 20% of the traits have different means in the CC from the initial collection; and (2) the CR retained by the CC is no less than 80% (Hu et al. 2000). Phenotypic correlations among the traits in the CC were estimated to discern whether these associations, which might be under genetic control; were appropriately represented in the CC. The qualitative data of twelve morphological traits were used to evaluate distribution homogeneity using chi-square test, representativeness using Wilcoxon rank-sum non-parametric test (Wilcoxon 1945) and diversity in the entire collection and the CC using the diversity index (H') of Shannon and Weaver (1949). For this, the data in the entire collection and the CC were ranked. The smallest value was given a rank of 1 and the next small value was ranked as 2, and so on. The ranks for the entire collection and the CC were summed up and used to test the null hypothesis of no differences in distributions of entire collection and CC. Morphological traits of core population were also analyzed using NJ clustering analysis. Standardized morphological data were used for calculation of

Euclidian distance using NTSYS-PC v2.2 (Rohlf 2006) and were subjected to cluster analysis based on the neighbor-joining (NJ) clustering method and visualized using DARwin v5.0.158 (Perrier and Jacquemoud-Collet 2006).

4.2.5 Evaluation of fatty acid contents of core collection

The contents of five fatty acids (palmitic acid, PA; stearic acid, SA; oleic acid, OA; linoleic acid, LA and α-linolenic acid, LA) were determined in the CC germplasm. Extraction and analysis of fatty acid methyl esters (FAMEs) from the seeds of the CC germplasm was performed according to the method described by Rajwade et al. (2010) with some modifications. Gas chromatography was performed using the AutoSystem XL GC (PerkinElmer, USA) with SP-2330 Supelco capillary column, 30 m long and 0.32 mm diameter. Each fatty acid was identified by comparison with a known standard and absolute quantification was done by comparison with the known internal standard, Tridecanoic acid (TDA). Pearson's correlation coefficients among the fatty acid data of the CC genotypes were calculated using XLSTAT v2012.6.08.

4.2.6 Identification of sources of economic traits

The phenotypic data on seven quantitative traits, *viz*. days to 50% flowering (DTF), days to maturity (DTM), plant height (cm) (PH),capsules per plant (CPP), seeds per capsule (SPC), 1000 seeds weight (g) (TW) and seed yield per plant (g) (YPP) were filtered separately for individual traits using the "Top 10" filter of Microsoft Excel v2010 and the genetic distances of the resultant genotypes were determined using XLSTAT v2012.6.08. For each of the economically important traits, four genotypes were identified as potential parents and two crosses were suggested considering their trait values, geographic origins, breeding material types and genetic distance between them.

4.2.7 DNA isolation and SSR genotyping

Total genomic DNA of all the accessions in the CC was isolated from leaves of 14-day old seedlings grown in green house using the DNeasy Plant Mini Kit (Qiagen, USA). The quality and quantity of DNA was assessed visually on 0.8% agarose gel stained with GelRed (Biotium, USA) as well as spectrophotometrically using Nanodrop 1000 spectrophotometer (Thermo Scientific, USA). Twenty-nine SSR primers showing high polymorphism among selected cultivars (data not shown), were used for evaluating polymorphism among the CC germplasm. PCR products were separated by 6% 7.5 M urea polyacrylamide gel (PAGE) electrophoresis and were visualized by silver staining (Bassam et al. 1991).

4.2.8 Statistical data analysis

The amplification products were scored as present (1) or absent (0) for each of the accessions to convert SSR data in binary data. Diversity parameters were computed for all the accessions and also within subgroups of the accessions using Power Marker v3.23 (Liu and Muse 2005). The diversity measures included the major allele frequency and the average number of alleles per SSR locus, observed heterozygosity (H_O), gene diversity (i.e., expected heterozygosity, He), and polymorphism information content (PIC).

4.2.9 Evaluation of population structure

The model-based program, STRUCTURE v2.3.1 (http://pritch.bsd.uchicago.edu/software/structure_v.2.3.1.html), was used to determine *K*, the number of structured groups (Falush et al. 2003, 2007; Pritchard et al. 2000), within the core collection. Using the ancestry model with admixture and the correlated allele frequency option, multiple runs of STRUCTURE were performed by setting *K* from 1 to 10. The length of burn-in was set at 10,000 followed by 100,000 iterations, and each run was replicated 10 times. To identify optimal *K*; two procedures, an *ad hoc* procedure described by Pritchard et al. (2000) and another procedure developed by Evanno et al. (2005), were employed. Runs with the highest *Ln*P(D) probability (estimated log probability of data) and ΔK value (the rate of change in the log probability of data between successive *K* values) were considered for each *K* and graphical outputs were visualized to determine the most appropriate value of *K*. To confirm these results, principal component analysis (PCA) and neighbor-joining (NJ) clustering were performed. The PCA was carried out using GenAlEx v6.1 (Peakall and Smouse 2006) and the first two principal components were plotted. The genetic distance matrix was computed using NTSYS-PC v2.2 (Rohlf 2006). It was subjected to cluster analysis based on the NJ clustering method and visualized using DARwin v5.0.158 (Perrier and Jacquemoud-Collet 2006).

To measure the goodness of fit for cluster analysis, a cophenetic correlation value between the original similarity matrix and the cophenetic matrix was compared using the MXCOMP procedure of NTSYS-PC v2.2. The significance of correlation between the matrices was tested using the normalized Mantel *Z*-statistics (Mantel 1967).Mantel test was also implemented for correlation analysis of morphological and genotypic data. Further, Analysis of Molecular Variance (AMOVA) was performed to partition the molecular genetic variance within and among the accessions and between the populations as determined by STRUCTURE analysis using Arlequin v3.5.1.2 (Excoffier and Lischer 2010). In addition, global F_{ST} and pair-wise F_{ST} values were estimated to infer the pattern of population structure according to Weir and Cockerham (1984) and significance of estimates were calculated with FSTAT v2.9.3 (Goudet 2001).

4.3 Results

4.3.1 Phenotypic variation in entire germplasm collection and development of core collection

Eight quantitative characters were analyzed in 2,239 accessions of linseed and found a large variation among them **(Table 4.1)**. Based on the agglomerative hierarchical clustering of the quantitative data, the entire collection was grouped into 20 major clusters. The numbers of accessions in the largest and the smallest clusters were 186 and 19, respectively. Approximately 10% of the accessions were randomly selected from each cluster to incorporate into the CC. The CC thus formed, contained 222 accessions from the total of 2,239 accessions.

Geographic origin and qualitative data were available for 192 accessions of the CC and hence they were used for determining fatty acid contents and molecular marker analysis. Interestingly, nearly half of the 192 CC accessions (88, 46%) were of exotic origin, while the remaining (104, 54%) accessions were of Indian origin. Thus, the proportion of exotic genotypes in the CC (46%) was much higher compared to their proportion in the entire germplasm collection (16%).The CC mostly comprised landraces (130, 67.71%) followed by breeding lines (24, 12.50%) and primitive cultivars (22, 11.46%) **(Table S4.1)**.

Table 4.1: Trait values for eight quantitative characters in the entire collection (EC) and core collection (CC) of linseed

Trait^	Collection	Sample size	Minimum	Maximum	Mean ± Std. error	Std. deviation	CV	Pvalue*
DTF	EC	2239	54.000	110.000	82.946±0.182	8.629	10.403	0.550
	CC	222	55.000	105.000	82.604±0.580	8.643	10.463	
DTM	EC	2239	106.890	175.890	139.300±0.168	7.929	5.692	0.942
	CC	222	116.000	158.000	139.338±0.514	7.652	5.492	
PH	EC	2239	23.060	151.000	62.142±0.249	11.792	18.977	0.318
	CC	222	27.000	111.000	62.923±0.742	11.055	17.569	
TPH	EC	2239	8.000	104.000	33.666±0.193	9.134	27.132	0.950
	CC	222	15.000	88.000	33.680±0.589	8.778	26.062	
CPP	EC	2239	22.000	364.000	106.724±0.959	45.381	42.522	0.829
	CC	222	22.000	364.000	107.437±3.156	47.025	43.770	
SPC	EC	2239	3.000	10.560	7.749±0.027	1.279	16.502	0.919
	CC	222	4.000	11.000	7.802±0.097	1.448	18.558	
TW	EC	2239	1.600	21.940	6.391±0.031	1.451	22.708	0.187
	CC	222	3.000	10.600	6.470±0.090	1.348	20.829	
YPP	EC	2239	0.030	16.750	5.505±0.051	2.404	43.657	0.875
	CC	222	0.460	15.000	5.479±0.154	2.293	41.853	
PA	CC	192	0.668	4.217	1.418±0.030	0.457	32.225	-
SA	CC	192	0.349	2.707	1.007±0.030	0.345	34.253	-
OA	CC	192	0.946	11.286	3.898±0.110	1.466	37.623	-
LA	CC	192	0.590	7.518	2.284±0.060	0.866	37.915	-
ALA	CC	192	2.866	24.296	10.325±0.260	3.576	34.634	-

^: See text for trait abbreviations; *: The *P* value is the probability of difference in the trait means of the entire collection and CC

4.3.2 Evaluation of the core collection

The *z*-test was used to test the significance of difference between the means of the entire collection and the CC derived from it, for the eight quantitative morphological traits. The differences between the means of the entire collection and the CC were non-significant for all the characters indicating true representativeness of the CC to the entire collection. The *P*value was >0.8 for five (DTM, TPH, CPP, SPC and YPP) of the eight quantitative characters **(Table 4.1)**. The values of MD and CR were 0.00% and 82.35%, which further support the representativeness of the CC to the entire collection. Frequency distribution analysis of the 12 qualitative morphological traits evaluated in the 1,965 initial accessions and 192 CC accessions indicated homogeneity of distribution for all the traits. The highest *P*value (0.99) was observed for capsule dehiscence; while seed color had the lowest *P*value (0.150) **(Table 4.2)**. The Wilcoxon rank-sum test indicated that most of the traits except flower color (*P*=0.003), anther color (*P*=0.036), stigma color (*P*=0.006) and style color (*P*=0.009) had similar distribution in the core and entire collections. The Shannon and Weaver diversity index (H') in the CC ranged from 0.192 – 0.666 with an average of 0.324; while that of the entire collection ranged from 0.198 – 0.640 with an average of 0.330, which were very similar indicating that the diversity in the entire collection was appropriately represented in the CC **(Figure 4.1)**.

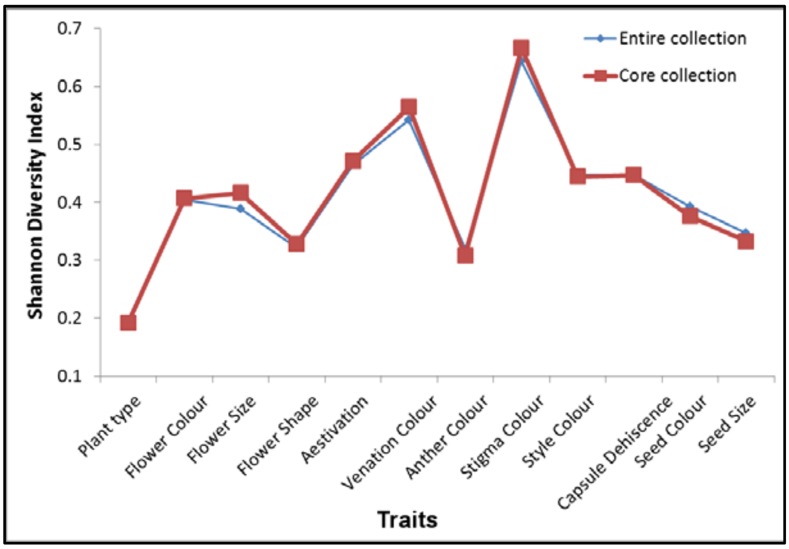


Figure 4.1: Distribution of the Shannon-Weaver diversity index for 12 morphological traits in the entire and core collections

Table 4.2: Results of chi-square test and Wilcoxon rank-sum nonparametric test to evaluate distribution of morphological traits in the entire and core collections of linseed

Sr. No.	Trait	No. of classes	Distribution for morphological traits		Wilcoxon test *P*value
			χ^2	*P*value	
1	Plant type	2	3.841	0.777	0.371
2	Flower color	12	19.675	0.800	0.003
3	Flower size	4	7.815	0.354	0.100
4	Flower shape	3	5.991	0.213	0.181
5	Aestivation	3	5.991	0.559	0.181
6	Venation color	8	14.067	0.624	0.014
7	Anther color	6	11.070	0.918	0.036
8	Stigma color	10	16.919	0.689	0.006
9	Style color	9	15.507	0.941	0.009
10	Capsule dehiscence	3	5.991	0.990	0.181
11	Seed color	5	9.488	0.150	0.059
12	Seed size	3	5.991	0.734	0.181

4.3.3 Diversity for phenotypic and biochemical traits in the core collection

There were vast differences in values of the eight quantitative characters and fatty acid contents among the CC accessions **(Table 4.1)**. The traits, capsules per plants (CPP) and yield per plant (YPP), along with contents of all the five fatty acids showed high variation, indicating diverse nature of the accessions. Further, to estimate the phenotypic diversity, clustering was performed using Euclidian distance and the accessions were divided into five clusters **(Figure 4.2)**. Correlations among the morphological traits were examined and many traits were significantly correlated. For example, PH was significantly correlated with DTF (*r*=0.409, *P*<0.0001), DTM (*r*=0.448, *P*<0.0001) and CPP (*r*=0.337, *P*<0.0001), while YPP was significantly correlated with DTM (*r*=0.182, *P*<0.01) and CPP (*r*=0.197, *P*<0.01) **(Table 4.3)**. Correlations among contents of five fatty acids were also determined, which revealed highly significant (*P*<0.0001) and positive correlations among all the fatty acids **(Table 4.4)**. The minimum correlation value of 0.762 was observed between OA and LA contents, while the highest correlation of 0.898 was observed between contents of PA and SA. Overall, the content of PA, which is the first fatty acid in the omega-3 fatty acid biosynthetic pathway, highly influenced contents of the rest of the four fatty acids.

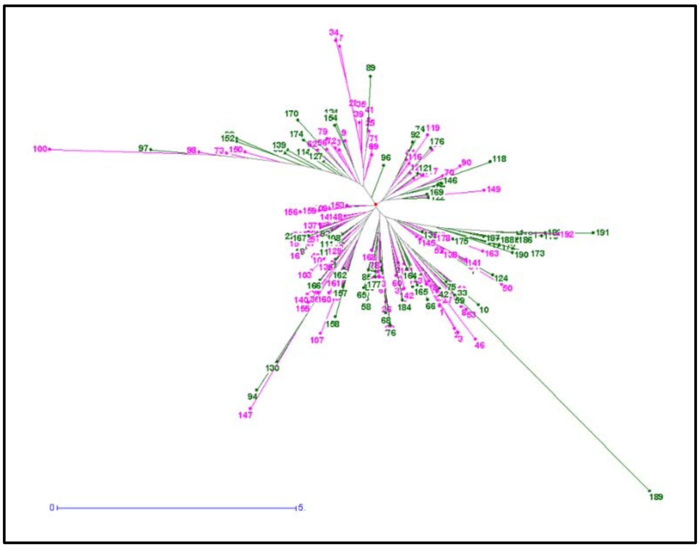


Figure 4.2: Neighbor-joining tree constructed using Euclidian distance matrix based on phenotypic data. The genotypes in *Pink* belong to Pop1 and those in *Green* belong to Pop2.

Table 4.3: Correlation coefficients among morphological traits in the linseed CC (A correlation > P=0.235 will be significant at P=0.01)

Trait	DTF	DTM	PH	TPH	CPP	SPC	TW
DTM	(0.260)***						
PH	(0.409 ***	(0.448)***					
TPH	(0.363)***	(0.285)***	(0.812)***				
CPP	-0.106	-0.135	(0.337)***	(0.211)***			
SPC	(-0.070)	(-0.080)	(-0.040)	-0.088	(-0.08)		
TW	(-0.040)	-0.102	-0.008	(-0.020)	(-0.05)	(-.060)	
YPP	(-0.040)	(0.182)**	-0.113	-0.058	(0.197)**	-0.121	-.023

Table 4.4: Correlation coefficients among contents of five fatty acids in seeds of the CC of linseed

Fatty acid	PA	SA	OA	LA
SA	0.898***			
OA	0.828***	0.794***		
LA	0.884***	0.781***	0.762***	
ALA	0.887***	0.781***	0.793***	0.894***

Note: A correlation > P=0.139 will be significant at P=0.01

4.3.4 Genetic diversity of the core collection

A total of 29 SSRs were used to assess the genetic diversity in the core collection. A representative PAGE pattern has been shown in **(Figure 4.3)**. These SSRs detected total 78 alleles and the number of alleles per SSR locus ranged from 2 to 6 with an average of 2.7 alleles per primer pair **(Table S4.2)**. The average observed heterozygosity was 0.0070, which ranged from 0 to 0.0395. It was slightly higher in POP1 (0.06) than POP2 (0.03) (POP1 and POP2 are described in the following paragraph).The PIC ranged from 0.1214 (Lu787) to 0.7365 (LU9) with a mean of 0.3328. Gene diversity ranged from 0.1298 to 0.7714 with a mean of 0.3864. Based on the estimates of gene diversity, the most informative markers were Lu868, Lu125a and Lu273, while the markers Lu787, Lu452 andNCL_Flx_1 were least informative. Among all, 17 loci showed no heterozygosity, while three loci detected <1%heterozygosity among the 192 accessions **(Table S4.2)**.

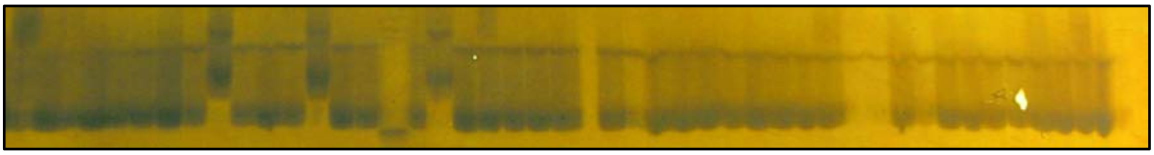


Figure 4.3: Allele patterns obtained from SSR markers NCL_flax_2

4.3.5 Population structure

Population structure of the CC (192 accessions) was inferred based on the analysis of genotypic data of 78 SSR alleles. The Bayesian based clustering method, as implemented in the STRUCTURE software, revealed K = 2 as the most appropriate number of inferred clusters based on the values of both, *Ln* P (D) and Δ*K*, indicating the presence of two subpopulations in the CC **(Figure 4.4)**. For assignments of individuals to their respective subgroups inferred by the STRUCTURE software, the criterion of 75% membership probability was used. This resulted in assignment of 53 accessions to group K1 (designated as POP1) and 71 accessions to group K2 (designated as POP2), while 68 accessions showed admixture ancestry. Thus, 124 accessions (64.58%) were strongly assigned to one or the other population; while 68 (35.41%) had mixed ancestry. To assign the accessions with mixed ancestry to either of these populations, a cut off value was set as follows; POP1: 0-50% proportion of genome from cluster K1 and POP2: 0-50% proportion of genome from cluster K2. This finally resulted in 86 and 106 accessions in POP1 and POP2, respectively.

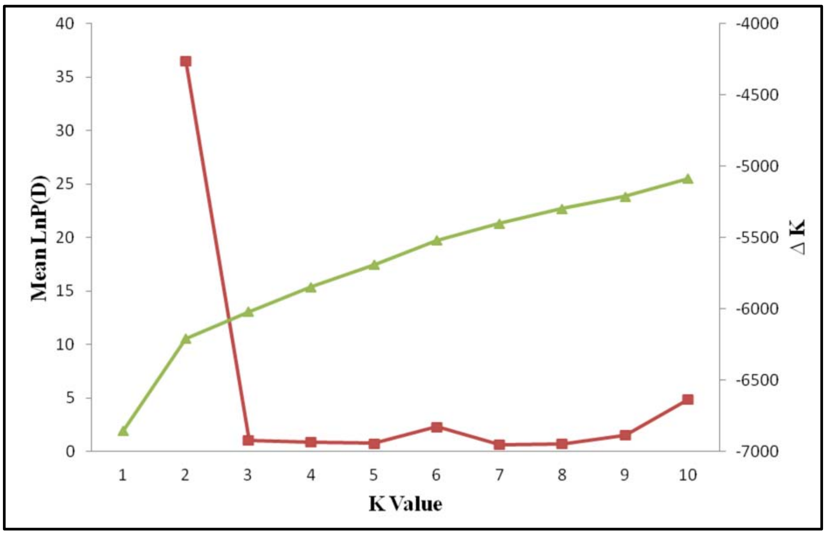


Figure 4.4: Average log likelihood values (mean LnP (D)) and ad-hoc statistic Δ*K* for *K* values ranging from 1 to 10

To evaluate the results of the STRUCTURE software, we also performed PCA and neighbor-joining clustering of the CC germplasm. The maximum SSR variation was explained by the first three principal components as 28.24%, 20.09% and 14.61%, respectively among the 192 genotypes. Plotting the first two principal components, with genotypes color coded according to their subgroups, showed separation of the accessions with little overlap **(Figure 4.5)**. Similar pattern was also obtained when a neighbor-joining tree was constructed to visualize the relationships among the CC accessions **(Figure 4.6)**. The t-test indicated significant differences between POP1 and POP2 for all the parameters except heterozygosity and allele number **(Table 4.5)**. However, Mantel test did not show significant correlation between the morphology and genetic diversity based matrices (data not shown).

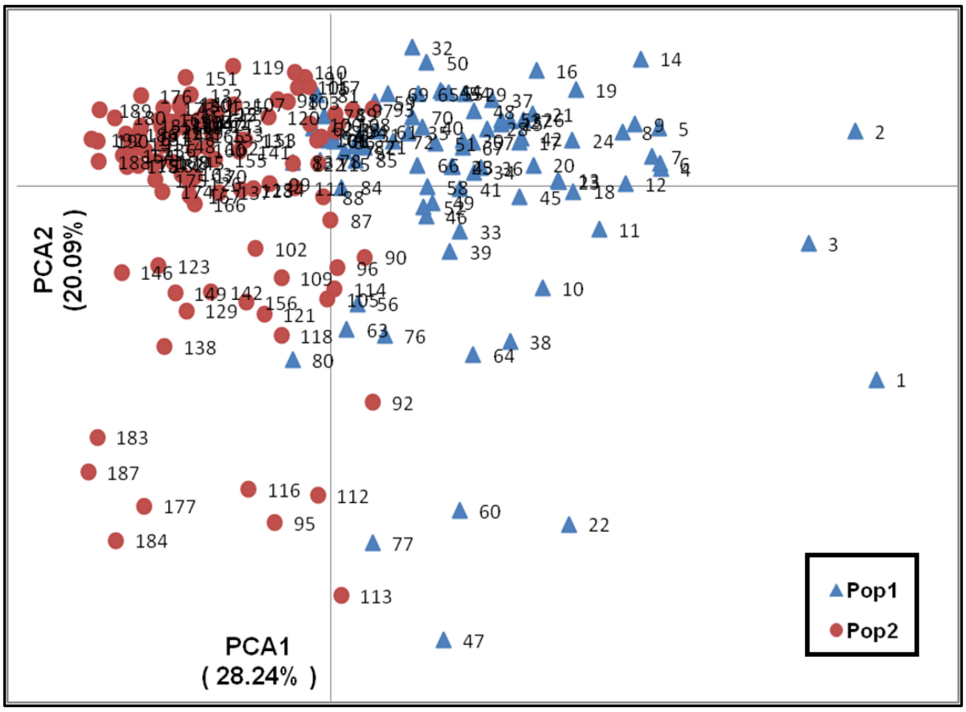


Figure 4.5: PCA of 192 linseed accessions based on molecular marker data showing segregation of the accessions into two major groups

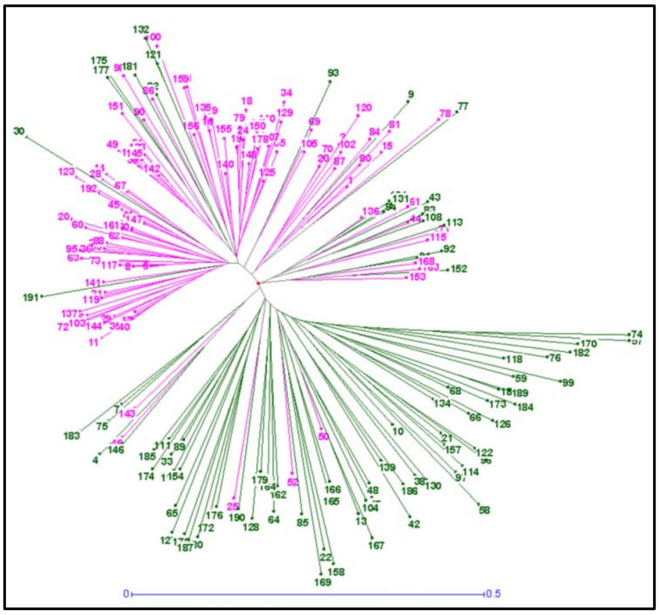


Figure 4.6: Neighbor-joining tree constructed using genetic distance matrix based on molecular marker data. The genotypes in *Pink* belong to Pop1 and those in *Green* belong to Pop2.

Table 4.5: Summary statistics for the two sub-populations identified by STRUCTURE analysis

	Total	**POP1**	**POP2**	***P* value**
Frequency of major allele	0.72	0.64	0.82	0.00004*
Number of alleles	2.69	2.66	2.62	0.89898 n.s.
Gene diversity (**He**)	0.39	0.46	0.27	0.00005*
Heterozygosity (**Ho**)	0.01	0.01	0.01	0.22953 n.s.
PIC	0.33	0.39	0.23	0.00031*

The AMOVA **(Table 4.6)** indicated that 14.75% of the total genetic variance was partitioned between POP1 and POP2, while 83.36%was among the accessions within the populations and the remaining (1.89%) variation resided within accessions. Although most of the genetic variance was among accessions within populations, the differentiation of the groups was highly significant (*P*=0.000). Such genetic differentiation of populations was also supported by the estimates based on the Wright's F_{ST} (0.145), being significantly different from zero.

Table 4.6: AMOVA for the two sub-populations identified by STRUCTURE analysis

Source of variation	**DF**	**Sum of squares**	**Variance components**	**Percentage of variation**
Among populations	1	155.320	0.77141	14.75
Within populations	190	1675.091	4.35866	83.36
Within individuals	192	19.000	0.09896	1.89
Total	**383**	**1849.411**	**5.22903**	

4.3.6 Identification of sources of economic traits

For each of the seven quantitative traits *viz.* days to 50% flowering (DTF), days to maturity (DTM), plant height (cm) (PH),capsules per plant (CPP), seeds per capsule (SPC), 1000 seeds weight (g) (TW) and seed yield per plant (g) (YPP), 10-11 genotypes with high trait values were identified **(Table S4.3)**. For each trait, two crosses involving four genetically diverse potential parents having superior trait values were identified. In all, 14 crosses were suggested for the seven economic traits, which involved 20 different parents. Interestingly, the potential parent genotype A-199 was involved in four different crosses for the traits DTF and YPP. Likewise, the genotype Ayogi was involved in three different crosses for the traits TPH and CPP. Moreover, these genotypes were of exotic origin, Australia and The Netherlands, respectively.

4.4 Discussion

Genetic diversity in populations can be analyzed using various data including morphological, agronomic and eco-geographical traits or molecular and biochemical markers. Each of these criteria has its advantages and disadvantages for estimating genetic diversity. In the present study, the data on morphological and agronomic traits and molecular markers were exploited to analyze the Indian linseed germplasm and establish a core collection.

4.4.1 Development of CC and phenotypic diversity

Selecting an appropriate stratification method plays an important role in constructing a core collection. Various methods such as stratified sampling, M (maximization) method (Charmet and Balfourier 1995; Peeters and Martinelli 1989; Schoen and Brown 1995; Spagnoletti Zeuli and Qualset 1993) and Ward's minimum variance methods (Ward 1963) have been reported for constructing a core collection. The Ward's agglomerative clustering method was used as it optimizes the objective function by minimizing the sum of squares within groups and maximizing the sum of squares among groups. Likewise, sample size also plays an important role in constructing a CC, which in turn depends upon the genetic diversity and degree of genetic redundancy present in the entire collection. Generally, small sampling percentage is suitable for large initial population and *vice versa.* Brown (1989) suggested that a CC with 10% sampling percentage could represent 70% genetic diversity of the initial population when the number of the initial accessions was over 3,000. In the present study, 10% samples were randomly selected from each cluster resulting in the CC of 222 accessions from an initial collection of 2,239 accessions. Interestingly, the proportion of exotic genotypes in the CC (46%) was much higher compared to their proportion in the entire germplasm collection (16%). In addition, the CC mostly comprised landraces, followed by breeding lines and primitive cultivars. Both these facts indicate that most of the genetic diversity of the linseed germplasm lies in the exotic germplasm and landraces and they should be used more often in breeding programs to utilize the diversity for developing improved varieties.

The CC thus constructed was evaluated for homogeneity, representativeness and diversity by various statistical methods such as *z*-test, Chi-square test, Wilcoxon rank-sum test, Shannon diversity index etc. The results of these analyses confirmed that the CC contained almost all the diversity present in the entire collection for all the characters and also it was homogeneous. Further, the difference between the means of the entire collection and the CC was non-significant for all the characters, as revealed by the *z*-test. The values of MD and CR were 0.00% and 82.35%, which further supported the representativeness of the CC to the entire collection.

Frequency distribution analysis of the 12 qualitative morphological traits evaluated in the 1,965 initial accessions and 192 accessions from the CC indicated homogeneity of distribution for all the traits. The Wilcoxon rank-sum test indicated that most of the traits, except flower color, anther color, stigma color and style color, had similar distribution in the core and entire collections. Further, the Shannon and Weaver diversity index (H') in the CC and the entire collection were very similar, indicating that the diversity in the entire collection was appropriately represented in the CC.

Further, the data of eight quantitative traits and five fatty acids were analyzed to understand the extent of diversity for these traits in the core collection. Wide variation was observed in the quantitative morphological traits, specifically for number of capsules per plant and LA content, which could be due to a variety of factors such as, disruptive selection for fiber type or oil type varieties, environmental influence on these characters during phenotypic evaluation, due to the ancient cultivation history of linseed resulting in accumulation of variation or it might be due to the diverse geographical origins of the accessions. Similar high morphological diversity was also observed by Diederichsen and Fu (2006).

4.4.2 Genetic diversity and population structure analysis

Several studies have been attempted to reveal the genetic relationships, extent of variation and genetic erosion across various linseed collections using a variety of markers such as RAPD (Diederichsen and Fu 2006; Fu et al. 2003b), ISSR (Rajwade et al. 2010; Wiesnerova and Wiesner 2004), SSR (Cloutier et al. 2009; Fu 2011; Soto-Cerda et al. 2012) and retrotransposon based markers (Smykal et al. 2011). However, very few Indian accessions were analyzed and there was no report of analyzing diversity within the linseed germplasm of India, which is in fact considered as one of the origins of linseed (Ahlawat 2008).

The CC was analyzed using 78 SSR alleles with an average of 2.7 alleles per marker. The PIC value (average 0.332) as well as the average gene diversity (0.386) indicated low genetic diversity. The frequency distribution analysis further confirmed the above results. These observations support earlier molecular marker studies, which also reported low genetic diversity in linseed (Rajwade et al. 2010; Smykal et al. 2011; Soto-Cerda et al. 2012).This indicates narrow genetic base of linseed germplasm possibly due to its self-pollinated nature, limited gene flow and breeding methods (Albertini et al. 2011).

Further, to understand the population structure of the core collection, the CC was subjected to STRUCTURE analysis using the molecular marker data, which identified existence of two sub-populations (POP1 and POP2) consistent with the PCA and NJ results. Additionally, the presence of statistically significant population structure was confirmed by AMOVA and population-specific F_{ST} analyses. The results of AMOVA indicated that 15% of the total genetic variance was partitioned between the two populations, while 83% variance was among accessions within populations. These results are in conformity with those reported earlier (7 to 24% genetic variance being partitioned among populations and the remaining 76 to 93% variance being partitioned among accessions within populations) (Mansby et al. 2000; Rajwade et al. 2010; Smykal et al. 2011; Soto-Cerda et al. 2012). However, the underlying reason of this differentiation remains unclear. One possible explanation for this could be different ancestry of these accessions. However, the pedigree data were not available for these accessions and hence it is difficult to comment on it at present. The CC contained about 45% (88) accessions from other countries, which were dispersed in both POP1 and POP2 sub-populations indicating sharing of genes among the accessions of different geographical origins. In conclusion, the results from the three independent methods confirm sub-division of the CC into two sub-populations.

The significant and positive correlation of technical plant height with days to maturity underscores the long duration nature of fiber flax varieties and shorter duration of oilseed varieties. In case of fatty acids, PA, which is the first fatty acid in the omega-3 fatty acid biosynthetic pathway, highly influenced contents of the rest of the four fatty acids. Significant correlation of the molecular data with days to maturity, technical plant height, and contents of OA and PA was observed. As expected, the clusters obtained using the molecular data were of better resolution than the clusters based on morphological data, which indicated an under-estimate of genetic relationships with morphological traits. Hence, the correlation between the molecular data and few of the morphological and biochemical traits, although significant, needs to be treated with caution. As the SSR markers are distributed throughout the genome, they represent the overall genomic diversity and not phenotypic or functional diversity. Hence, they might show wide range of genetic distances between orthologous genomic regions in the germplasm. On the contrary, the morphological and biochemical traits might be governed by multiple genes and are mostly subjected to strong direct or indirect selection during the breeding process. Using a large number of SSRs or other types of makers to analyze the CC might explain the basis of sub-division of the CC into two sub-populations and also give a better correlation with the phenotypic data.

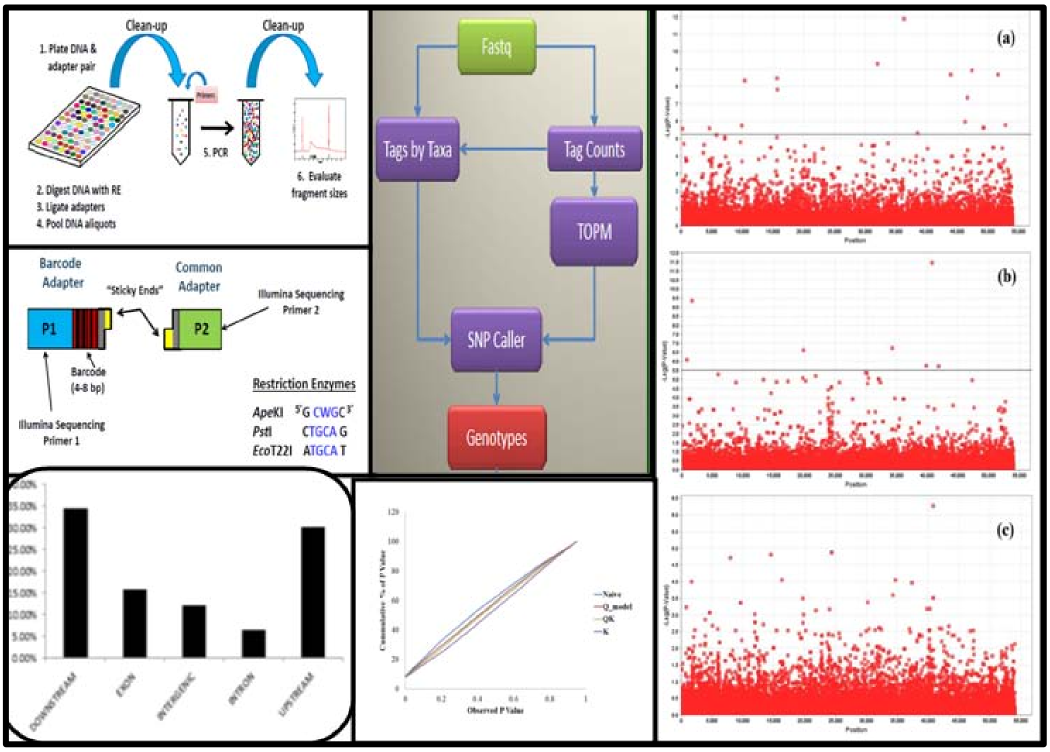
4.4.3 Identification of sources of economic traits

To demonstrate the practical utility of the CC developed in this study, we identified genetically diverse genotypes as potential parents for linseed breeding. For seven economically important traits, we suggested two crosses involving four genetically diverse parents having superior trait values. We also considered the geographic origin and breeding material type of the genotypes while determining the potential parental combinations. In majority of the crosses, the parents were from different countries and/or different breeding material type, suggesting usefulness of the strategy used to identify the parents. Only in case of the trait technical plant height (TPH), both the potential parents of Cross I were local landraces from The Netherlands. However, they showed high genetic distance as well as were phenotypically diverse. Two genetically diverse exotic genotypes (A-199 and Ayogi) were identified as potential parents for four economically important traits. Therefore, we suggest utilizing these genotypes for genetic improvement of Indian flax for the identified traits. Crossing such genetically diverse parents with high trait values is expected to improve the trait in the resulting progeny. Likewise, using a similar strategy, it is also possible to identify the genotypes with contrasting phenotypes for different traits. These genotypes could be crossed to generate mapping populations to map the genes and quantitative trait loci (QTLs) for different traits. Furthermore, these 20 genetically diverse genotypes having high trait values could be crossed in diallel fashion to develop a segregating population, from which superior genotypes combining different traits could be selected. This strategy is expected to widen the genetic base of the linseed germplasm and allow its exploitation to develop improved varieties with desired traits.

In summary, a CC of 222 accessions was developed from a collection of 2239 gemplasm accessions using phenotypic data. Further, the representativeness and homogeneity of developed CC was evaluated using various statistical methods *viz.;* z-test, chi-square test, %MD, %CR etc. and results proved its homogeneity and representativeness. Moreover, genetic diversity and population structure within CC was analyzed using SSR markers. Total 78 alleles were obtained with 29 polymorphic SSRs. Structure analysis divides the CC into two sub-populations, which was confirmed by neighbor-joining and PCoA study. However, very low *F*st value showed weak substructure. This CC would be helpful for efficient utilization of Indian germplasm for genetic improvement of linseed.

CHAPTER 5

Genotyping by sequencing in linseed: development of genome-wide SNP markers and genetic diversity and population structure study



This work has been communicated to *Molecular Breeding*

5.1 Introduction

Understanding the genetic factors controlling variability in agronomic traits constitutes the basis for efficient management of genetic resources, which is helpful not only in the short term breeding prospective, but also to conserve genetic diversity available in that species. However, the complex nature of agronomic traits results in greater difficulty for discerning the underlying genetic differences. Linkage mapping or quantitative trait loci (QTL) mapping, which works on an experimental population derived from a cross of bi-parents divergent for a trait of interest, is the most common approach in plants to detect QTL corresponding to complex traits. However, as a biparental mapping population is often derived from a restricted number of meiotic events, the genetic resolution of QTL maps often remains confined to a range of 10-30 cM (Flint-Garcia et al. 2005; Zhu et al. 2008). Moreover, linkage analysis is limited to only the alleles for which the two parents differ, which is very small as compared to the distribution of alleles in natural population.

An alternative approach, association mapping (AM) or linkage disequilibrium (LD) mapping can exploit the entire pool of genetic diversity existing in natural populations and overcomes the limitation inherent to linkage mapping. With the intrinsic nature of exploiting historical recombination events, AM offers increased mapping resolution to polymorphisms at sequence level and should therefore, enhance the efficiency of gene discovery and facilitate marker assisted selection (MAS) in plant breeding (Gupta et al. 2005; Moose and Mumm 2008).

Availability of large number of closely placed markers is the fundamental in success of LD mapping (Flint-Garcia et al. 2005; Mackay and Powell 2007). Till now, microsatellites or simple sequence repeats (SSR) markers are widely used (Weber and May 1989), however, nowadays, single nucleotide polymorphism markers (SNPs), which are highly abundant and known to be present in high frequency in the genome, are gaining very much attention (Deschamps et al. 2012**.** Further, with the advancement in next generation sequencing (NGS) technology, it is now possible to develop millions of genome-wide SNPs within sufficiently short duration and in cost-effective manner (Cortés et al. 2011).To list a few, large number of SNPs have been developed using NGS in humans (Altshuler et al. 2000), insect (*Drosophilla malanogaster*) (Berger et al. 2001) and from various plant species viz., wheat (Allen et al. 2011), eggplant (Barchi et al. 2011), rice (Feltus et al. 2004), *Arabidopsis* *thaliana* (Zhang and Borevitz 2009),rapeseed (Bus et al. 2012), and maize (Jones et al. 2009). In linseed, Kumar et al. 2012 developed more than 50,000 SNPs using reduced genome representation (RGR) libraries and Illumina sequencing platform of eight linseed genotypes.

Once developed, genotyping of SNPs can be performed simultaneously using various array based systems as KASPar (Nijman et al. 2008), GoldenGate (Fan et al. 2003) and Infinium assays (Gunderson 2009). Recently, genotyping-by-sequencing (GBS), has come forward, where SNP development in a large population is combined with SNP scoring, enabling a rapid and direct study of its diversity, population structure and LD mapping (Deschamps et al. 2012; Elshire et al. 2011)**.**

The present study was undertaken with the following objectives: (i) to develop a large number of genome wide SNPs in linseed using the GBS technology, (ii) to use the SNPs to determine population structure, LD (iii) to demonstrate use of GBS for genome-wide association study (GWAS) in linseed. More than 54,000 SNPs in 95 diverse linseed accessions were developed using a highly advanced GBS approach. Further, genetic diversity and population structure analysis was carried out to study the population admixture. Finally, a preliminary study of GWAS was carried out.

5.2 Materials and methods

5.2.1 Germplasm and genotyping

A collection of 95 linseed accessions from different countries, majority from India (49), was used for this study. These include local land races (LRR, 65%), primitive cultivars (PC, 11%), breeding lines (BL, 11%), exotic collections (EC, 6%), elite lines (EL, 5%) and released cultivars (RC, 2%) **(Table S5.1)**. These accessions are being maintained at the Project Coordinating Unit (Linseed), Chandrashekhar Azad University of Agriculture & Technology, Kanpur, India.

The genotyping by sequencing (GBS) technique was used for SNP development and genotyping. Genomic DNA was extracted from young leaves of 16 days old plants grown in green house (4-5 plants per accession) using DNeasy Plant Mini Kit (Qiagen). The DNAs were dried at 54°C and processed at the Institute for Genomic Diversity, Cornell University, Ithaca, NY, USA. Three restriction enzymes (REs), *ApeKI, PstI* and *EcoT22I,* were used for initial standardization. Of these, *ApeKI* showed good results and used for library preparation. DNAs were digested individually with *ApeKI* (recognition site: G/CWCG), and 96-plex GBS libraries were prepared (Elshire et al. 2011). DNA sequencing was performed on the Illumina Genome Analyzer-IIx. Sequence tags, 64-bp sequences that included a leading 4-bp C[T/A]GC signature from the cut site, were identified and tags with at least 10Xtotal coverage were retained. The draft genome sequence of linseed variety, CDC Bethune (Wang et al. 2012c), was used as reference to map the tags using BWA (Li et al. 2009), and SNPs were called with the TASSEL ver. 3.0 GBS pipeline (Bradbury et al. 2007) (http://www.maizegenetics.net/tassel/). Missing data were imputed with TASSEL ver. 3.0 (Bradbury et al. 2007). The SNPs were filtered with minor allele frequency cut-off of 0.1 before the analysis.

5.2.2 Functional annotation of SNPs

Linseed/flax genome annotation information was retrieved from phytozome (http://www.phytozome.com/) in GFF3 format. The genome annotation provided predicted gene structure and verified exon/intron boundaries using the ESTs and cDNA data. Therefore, we used the SnpEff ver. 3.0 program (http://snpeff.sourceforge.net/) (Cingolani et al. 2012) to co-locate all the SNPs with gene models and predicted their structural and functional relevance in the genome. SnpEff was also used for variant annotation and effect prediction of SNPs. The SNPs were described on the basis of their structural occurrence in the intergenic region, exons, introns and exon-intron splicing sites.

5.2.3 Genetic diversity, population structure and LD analysis

The model-based program STRUCTURE ver. 2.3.1 (http://pritch.bsd.uchicago.edu/software/structure_v.2.3.1.html) was used to determine ‘*K*’, the number of structured groups or subpopulations (Falush et al. 2003, 2007; Pritchard et al. 2000) within the present samples. Using the ancestry model with admixture and the correlated allele frequency option, multiple runs of STRUCTURE were performed by setting ‘*K*' from 1 to 10. The length of burn-in was set at 50,000 followed by 50,000 iterations, and each run was replicated thrice. The optimal number of subpopulations was identified using two different procedures *viz*: an *ad hoc* procedure described by Pritchard et al. (2000) and another procedure developed by Evanno et al. (2005). To confirm these results, principal coordinate analysis (PCoA) and neighbor-joining (NJ) clustering were performed. The PCoA was carried out using GenAlEx ver. 6.1 (Peakall and Smouse 2006) and the first two principal components were plotted, whereas DARwin ver. 5.0.158 was used to construct a NJ tree based on Nei's genetic distance matrix (Perrier and Jacquemoud-Collet 2006). Pair-wise F_{ST} comparisons were made using R script (http://www.evachan.org/rscripts.html) to determine the genetic differentiation among the inferred genetic groups.

The LD between all pairs of SNP loci was measured using the r^2 statistic implemented in TASSEL ver. 4.0 (http://www.maizegenetics.net/tassel/). The r^2cut-off of 0.1 was selected to consider loci in LD (Newell et al. 2011). We used sequential approach as described by Esteras et al. (2013) for calculating preliminary LD decay. Three contigs having the highest number of mapped SNPs were selected. The pair-wise r^2 values between SNPs were calculated separately and by combining all the SNPs from three contigs, using TASSEL ver. 4.0 (Bradbury et al. 2007). The LD decay over physical distance was studied by plotting r^2 against distances in Kbp, fitting the data using a second-degree locally weighted scatter plot smoothing (LOESS); (Breseghello and Sorrells 2006)) using ggplot2 ver. 0.9.3.1 package implemented in R (Ginestet 2011). The minimum distance for LD detection was defined at the distance where the LOESS fitted curve reached a plateau.

5.2.4 Phenotypic evaluation

The linseed accessions were grown in single rows of 3m length with a row spacing of 40 cm and plant to plant spacing of 6-8 cm at Germplasm Maintenance and Use (GMU) unit, C.S. Azad University of Agriculture & Technology, Kanpur, India. Recommended agronomic practices were followed to raise a healthy crop. The data on eight quantitative traits, *viz*. days to 50% flowering (DTF), days to maturity (DTM), plant height (cm) (PH), technical plant height (the height of the stem till the first node, from where branching starts; determines the length of the bast fibers that can be extracted from the plant) (cm) (TPH), capsules per plants (CPP), seeds per capsule (SPC), 1000 seeds weight (g) (TW) and seed yield per plant (g) (YPP) were recorded as per the approved DUS (distinctness, uniformity and stability) test guidelines for linseed (http://agricoop.nic.in/SeedTestguide/linseed1.htm).

5.2.5 Genome-wide association scan (GWAS)

Four models referring to the population structure (Q) and kinship (K), a pair-wise relationship matrix, were selected to test marker trait association (MTA). Results were compared to determine the best model for our analysis. This was done by plotting the end ranked P-values from GWAS in a cumulative way for each model by using capsule per plant (CPP) as phenotypic trait and a model in which P values follow uniform distribution with less deviation from expected P values, which is considered as ideal (Kang et al. 2008). The general linear model (GLM) included the Q model, and a naive model that did not control for Q. The mixed linear model (MLM) comprised the K model and the Q + K model. All the analysis was performed using TASSEL ver. 4.0 (http://www.maizegenetics.net/tassel). Results from all the models were compared to determine the best model for the present study. The 5% FDR (false discovery rate) value was calculated using q-value package (Storey 2002) and used to identify significant associations.

5.3 Results

5.3.1 Development and annotation of SNP markers

Three different REs (*ApeKI*, *PstI* and *EcoT22I*) were evaluated for their ability to digest linseed genomic DNA. Of these, the enzyme *ApeKI* produced a large fraction of DNA fragments in the 100-400 bp range and was selected for GBS library preparation (**Figure 5.1)**. For each accession, *ApeKI*-reduced representation libraries were constructed and sequenced using the Illumina Genome AnalyzerIIx, which produced about 25 Gbp of GBS data. A total of 1.84 million unique 64-bp tags were identified across all the linseed accessions, of which only 38.73% tags aligned to the draft linseed genome and 53.32%tags remain unaligned. Eight percent of the tags aligned to multiple positions on the linseed genome.

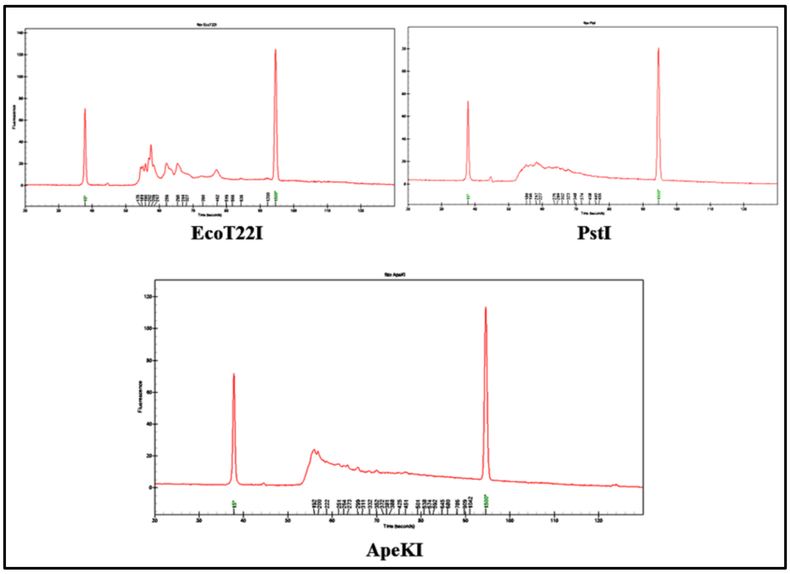


Figure 5.1: RE digestion profile of linseed genomic DNA a) *EcoT22I*, b) *PstI* c) *ApeKI*

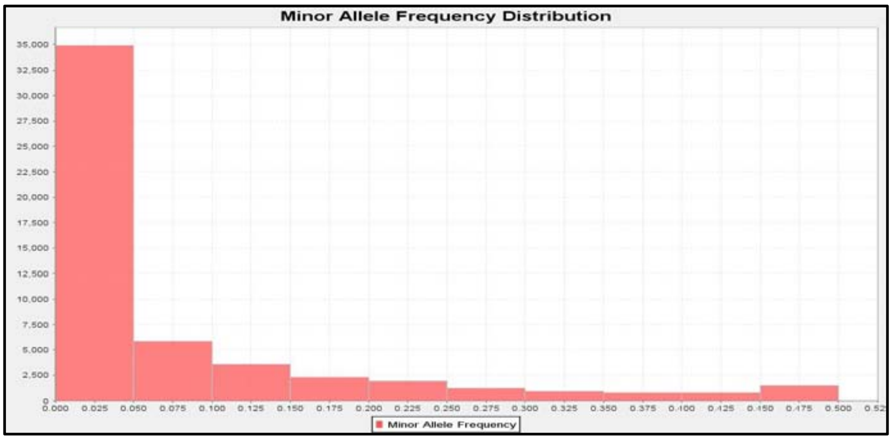


Figure 5.2: Minor allele frequency distribution of SNPs among the accessions

From the aligned tags 72,758 SNPs were identified. After filtering for tag coverage (>10% of taxa) about 54,000 SNPs were obtained. The minor allele frequency (MAF) among them varied from 0.025 to 0.5 **(Figure 5.2)**. The SNPs were further filtered with MAF cutoff ≥ 0.1 and 13,280 SNPs were retained, with an average density of one SNP per 27.86 kbp **(Table 5.1)**. Moreover, the structural and functional relevance of SNPs were also investigated by comparing the location of the 13,280 SNPs with the coordinates of predicted genes **(Figure 5.3)**. It revealed that 76.75% of these SNPs resided in intergenic regions, of which 30.08% were within 5 kb immediately upstream and 34.48% within 5 kb immediately downstream of an open reading frame (ORF). The remaining SNPs (23.25%) were located in exons, introns or untranslated regions (UTRs) within coding sequences. A higher percentage of SNPs was observed in exonic regions (15.72%) than in introns (6.46%). Hence, these SNPs can be used for candidate gene studies for the respective genes. In addition, 23 SNPs were observed at intron splicing sites, which could potentially alter the function of these genes.

Table5.1: Summary of GBS of 95 linseed accessions

Description	Number	Percent
Tags aligned to unique position	714992	38.73
Tags aligned to multiple position	146718	7.95
Unaligned tags	984359	53.32
Total number of tags	1846069	100.00
Total Number of SNPs identified	72758	
Filtered SNPs	13280	

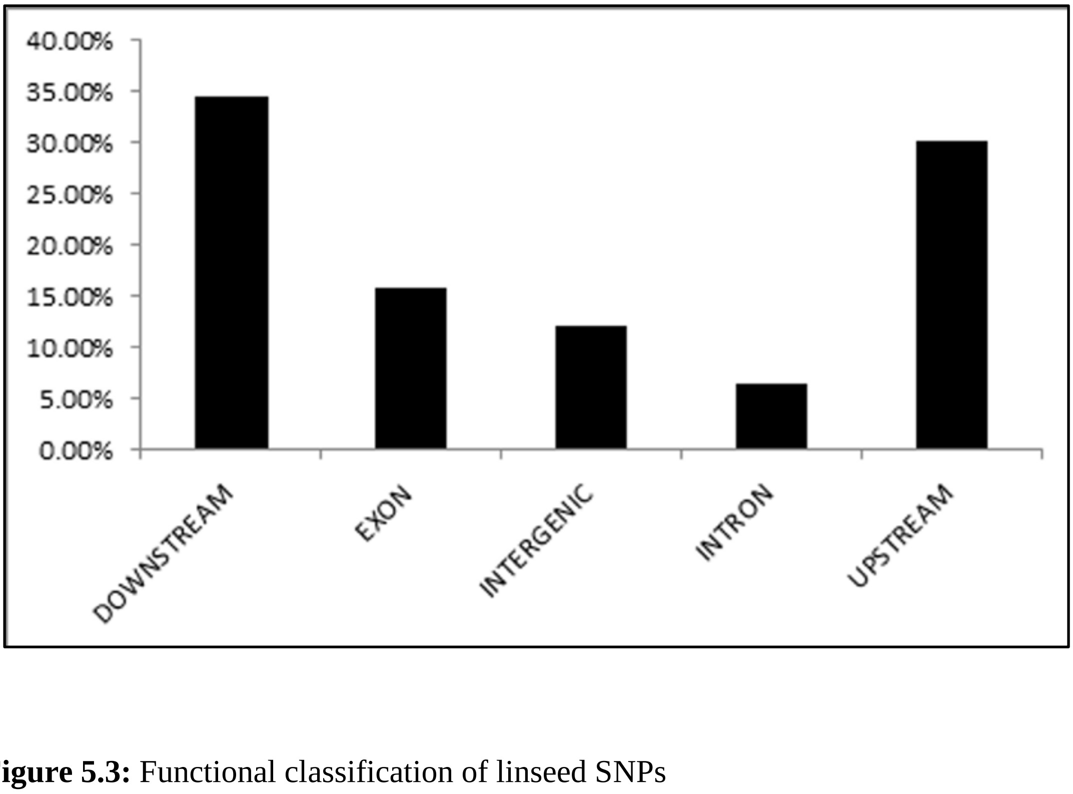


Figure 5.3: Functional classification of linseed SNPs

5.3.2 Population structure and linkage disequilibrium study

Population structure of the accessions was inferred based on the analysis of genotypic data of 13,280 SNPs. The Bayesian based clustering method, as implemented in the STRUCTURE software, revealed $K = 2$ as the most appropriate number of inferred clusters based on the values of both, *Ln* P (D) and ΔK, indicating the presence of two subpopulations in the accessions **(Figure 5.4)**. For assignments of individuals to their respective subgroups inferred by the STRUCTURE software, the criterion of 70% membership probability was used. This resulted in assignment of 13 accessions to group K1 (designated as POP1) and 54 accessions to group K2 (designated as POP2), while 28 accessions showed mixed ancestry. Thus, 67 accessions (70.52%) were strongly assigned to either of the subpopulations; while 28 (29.48%) had mixed ancestry. To assign the accessions with mixed ancestry to either of these subpopulations, a cut off value was set as follows; POP1: 0-50% proportion of genome from cluster K1 and POP2: 0-50% proportion of genome from cluster K2. This finally resulted in 24 and 71 accessions in POP1 and POP2, respectively. The NJ-tree and PCoA analysis also showed two major groups and the accessions from each subpopulation fairly grouped together **(Figures 5.5 and 5.6)**, thus supporting the predicted population structure. The very low coefficient of population differentiation (*F*st = 0.014) between major groups further support the above results.

Comparison of 88,172,560SNP pairs was performed from 13,280 SNPs to investigate LD in the set of accessions studied. About 4,715,000 (5.3%) loci were in significant LD. The mean r^2 for significant loci was 0.16. There were about 79,000 pairs with r^2 value greater than 0.4. These SNPs would be important in association mapping. The sequential approach was used for LD decay analysis. Three contigs with the highest number of SNPs mapped, *viz.*, contig_25 (2818.9 Kbp, 197 SNPs), contig_67 (2844.03 Kbp, 182 SNPs) and contig_123 (2472.80Kbp, 281 SNPs) were selected for the analysis. LD decayed to around 75 Kbp, 112 Kbp and 100 Kbp for the contigs, respectively **(Figure 5.7)**; while the overall LD decay was around 75 Kbp for the entire accessions analyzed (data not shown).

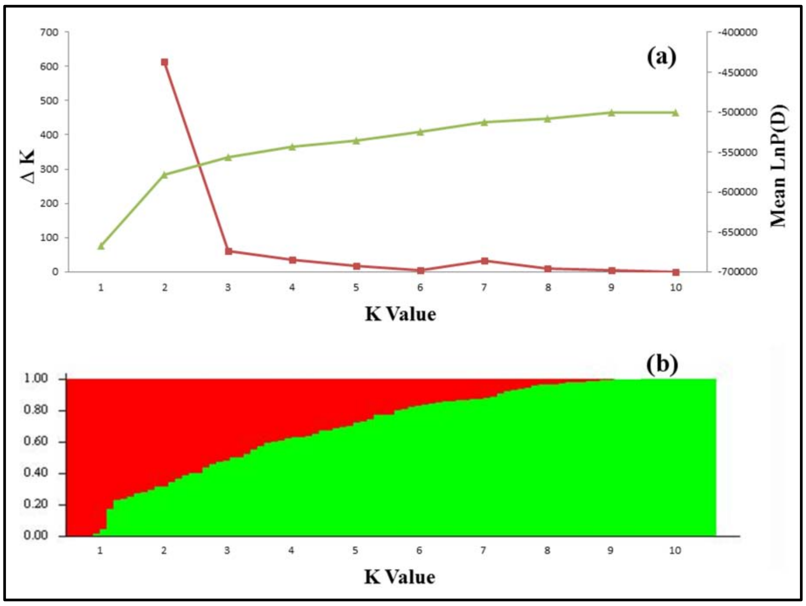


Figure 5.4: a)Average log likelihood values (mean LnP (D)) and ad-hoc statistic ΔK (Evanno et al. (2005)) for K values ranging from 1 to 10 b) Bayesian model-based analysis of SNP data analyzed using STRUCTURE (K = 2) software.

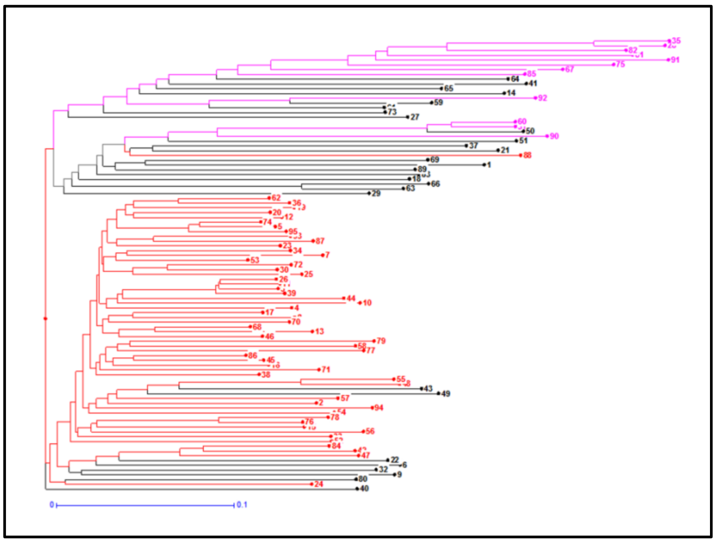


Figure 5.5: Neighbor- joining tree constructed using genetic distance matrix calculated based on molecular marker data

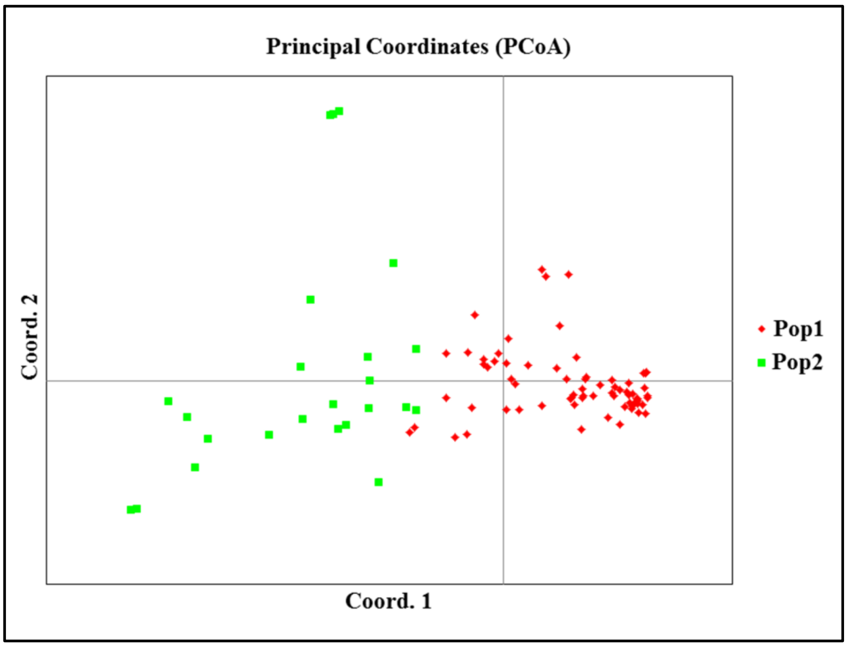


Figure 5.6: PCoA analysis of 95 linseed accessions based on SNP data showing partitioning of accessions into two distinct groups

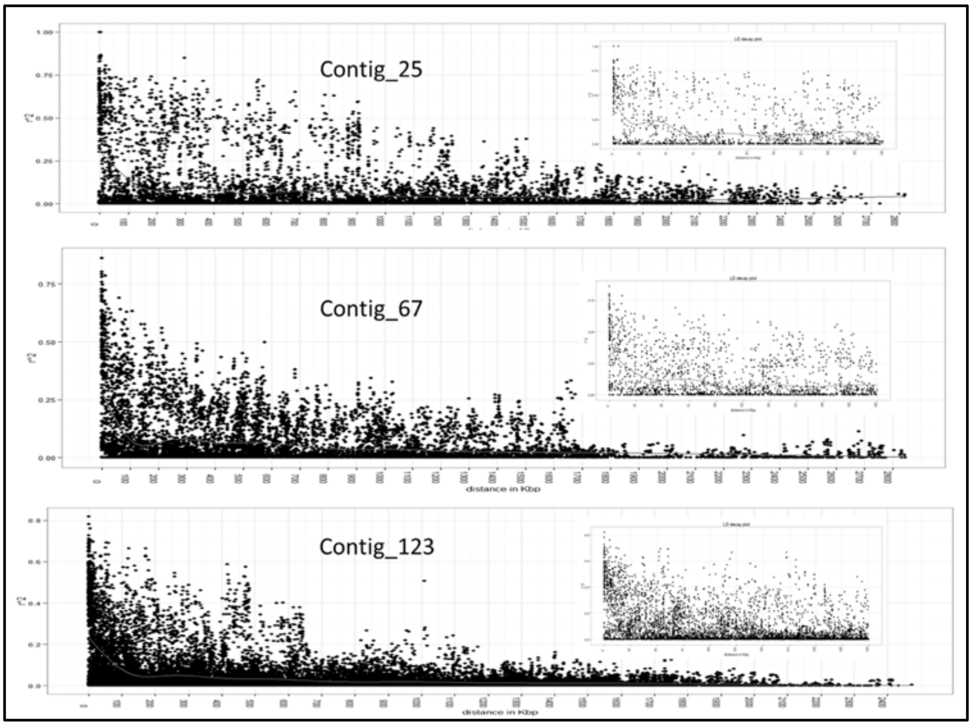


Figure 5.7: Linkage disequilibrium (r^2) versus physical distance (Kbp) between linseed accessions on different contigs. Curves were fitted by second degree LOESS.

5.3.3 Phenotypic diversity

There were vast differences in values of the eight quantitative characters among the accessions. Several agronomic characters, like number of capsules per plant (CPP), seed yield per plant (YPP), technical plant height (TPH) and plant height (PH), showed large standard deviations, indicating diverse nature of the accessions **(Table 5.2)**. Correlations among the morphological traits were examined and many traits were significantly correlated. For example, PH was significantly correlated with DTF (*r*=0.48, *P*=0.0001), DTM (*r*=0.40, *P*=0.0001) and CPP (*r*=0.23, *P*=0.0001), while YPP was significantly correlated with DTM (*r*=0.13, *P*=0.01) **(Table 5.3)**.

Table 5.2: Minimum and maximum values for eight quantitative characters in the linseed accessions

Trait	Min	Max	Average	S. D.
PH	40.00	111.00	65.21	11.84
TPH	20.00	88.00	35.24	9.98
CPP	22.00	364.00	112.55	49.57
SPC	4.00	10.56	7.72	1.46
TW	3.00	21.94	6.46	2.07
YPP	0.97	10.60	5.17	1.82
DAF	54.67	105.00	82.07	9.08
DTM	120.22	157.00	139.42	7.12

Table 5.3: Correlation coefficients between morphological descriptors in the linseed accessions (A correlation > *P*=0.235 will be significant at *P*=0.01).

Trait*	PH	TPH	CPP	SPC	TW	YPP	DAF
TPH	0.83***						
CPP	0.23*	0.12					
SPC	-0.04	0.05	-0.03				
TW	-0.08	-0.02	-0.09	0.06			
YPP	0.00	-0.10	0.04	0.07	0.13		
DAF	0.48***	0.46***	0.10	-0.11	-0.02	-0.21	
DTM	0.40***	0.33**	0.05	-0.09	-0.04	0.13	0.30**

*****: Please refer text for trait abbreviations

5.3.4 Genome-wide Association Scan (GWAS)

To demonstrate utility of GBS for GWAS in linseed, a preliminary analysis with identified SNPs and phenotypic data of one year and single location was carried out.

The reanalysis with replicated and multi-location phenotypic data will help to identify stable and promising SNPs.

5.3.4.1 Selection of the Best Model

Different statistical models were used to calculate *P*-values for associating each marker with the trait of interest, along with accounting for population structure to avoid spurious associations by TASSEL v.4.0. We followed the formula y = Xβ +M + Zu + e, where y is a response vector for phenotypic values, β is a vector of fixed effects regarding population structure, α is the vector of fixed effect for marker effects, u is the vector of random effects for co-ancestry and e is the vector of residuals. X can be either the Q-matrix or the PCs from Principal Component Analysis (PCA), M denotes the genotypes at the marker and Z is an identity matrix. Four models, i) Naive model: GLM without any correction for population structure; ii) Q-model: GLM with Q-matrix as correction for population structure; iii) QK model: MLM with Q-matrix and K-matrix as correction for population structure and iv) K-model: MLM with K-matrix as correction for population structure, comprising both general linear models (GLM) and mixed linear models (MLM) were selected to test the marker-trait-associations (MTA).

The results were compared to determine the best model for our analysis. As weak population structure was evident, we observed comparatively less difference in *P* value distribution among all the models **(Figure 5.8)**. Among the four models, QK and K showed good fit of *P* values; while in the naïve and Q models, excess amount of small *P* values relative to QK and K model, which indicate spurious association, was observed. As expected, the naïve model showed the highest number of small *P* values, while K model performed similar to QK model in displaying highly uniform distribution of *P* values and at the same time requiring less computational time. Irrespective of the model, major marker trait associations were constantly detected. However, the more stringent the model was, the less spurious background associations were detected; hence, K model was used for AM.

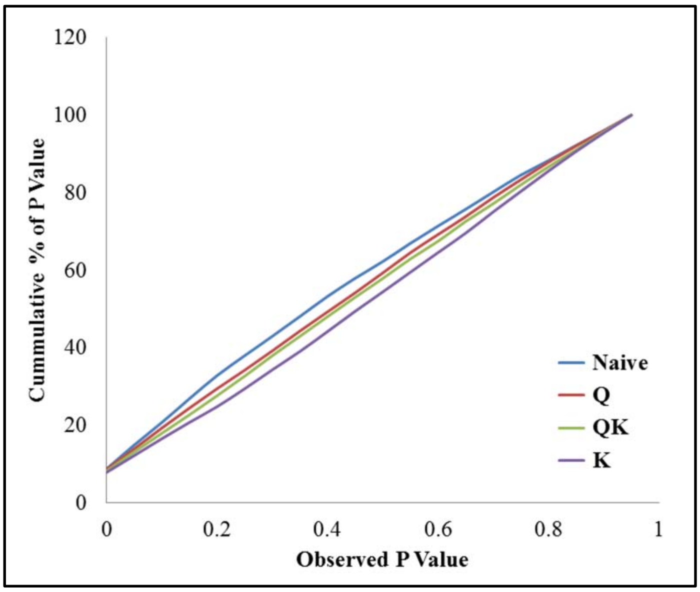


Figure 5.8: Comparison of different GWAS models. The model (K) showing uniform distribution with less number of low P values was selected for analysis

5.3.4.2 Association Mapping

We used very stringent criteria of 5% FDR to test the significance of SNPs and hence, were able to identify 24 significant SNPs for TPH, PH and CPP **(Table 5.4)**. CPP is directly related with yield and thus is an important agronomic trait. This trait was associated with a maximum of 16 SNP loci **(Table 5.4, Figure 5.9)**. Majority of the loci were dispersed over different contigs, which indicates that the genes too, are spread over different contigs. There were two loci each on contig_165 and contig_1376.The technical plant height is an important trait with respect to fiber length. Seven loci were significantly associated with this trait. We found a single significant locus for plant height and it was common with that of TPH **(Table 5.4, Figure 5.9)**. This might be because these are highly correlated traits (r^2 = 0.83).

Table 5.4: Particulars of the SNPs associated with agronomic traits (CPP, PH and TPH)

Sr No.	Trait*	Marker	Contig	P-value	Marker R^2
1	CPP	S1_216681484	604	$1.3428E^{-12}$	0.45
2	CPP	S1_190627014	543	$5.1793E^{-10}$	0.37
3	CPP	S1_272084742	1486	$1.2037E^{-09}$	0.36
4	CPP	S1_292322126	3345	$2.0638E^{-09}$	0.35
5	CPP	S1_255428550	863	$2.0918E^{-09}$	0.35
6	CPP	S1_99077443	165	$3.3418E^{-09}$	0.35
7	CPP	S1_65947019	176	$4.6843E^{-09}$	0.34
8	CPP	S1_99416150	165	$1.5193E^{-08}$	0.32
9	CPP	S1_268035379	1253	$4.4506E^{-08}$	0.31
10	CPP	S1_266207356	1123	$1.0881E^{-06}$	0.26
11	CPP	S1_317592846	8159373	$1.6936E^{-06}$	0.25
12	CPP	S1_58699858	86	0.000001795	0.25
13	CPP	S1_280473547	1376	0.000002381	0.25
14	CPP	S1_280473569	1376	0.000002381	0.25
15	CPP	S1_28215937	98	$2.5904E^{-06}$	0.24
16	CPP	S1_1332555	6	$2.7156E^{-06}$	0.24
17	PH	S1_239193874	883	$5.4111E^{-07}$	0.27
18	TPH	S1_239193874	883	$3.5507E^{-12}$	0.44
19	TPH	S1_9536658	34	$4.4288E^{-10}$	0.37
20	TPH	S1_205470003	464	$1.8669E^{-07}$	0.29
21	TPH	S1_123040863	196	$2.4116E^{-07}$	0.28
22	TPH	S1_4380573	8	$8.0373E^{-07}$	0.26
23	TPH	S1_234085689	888	$1.8082E^{-06}$	0.25
24	TPH	S1_246082832	924	$1.9045E^{-06}$	0.25

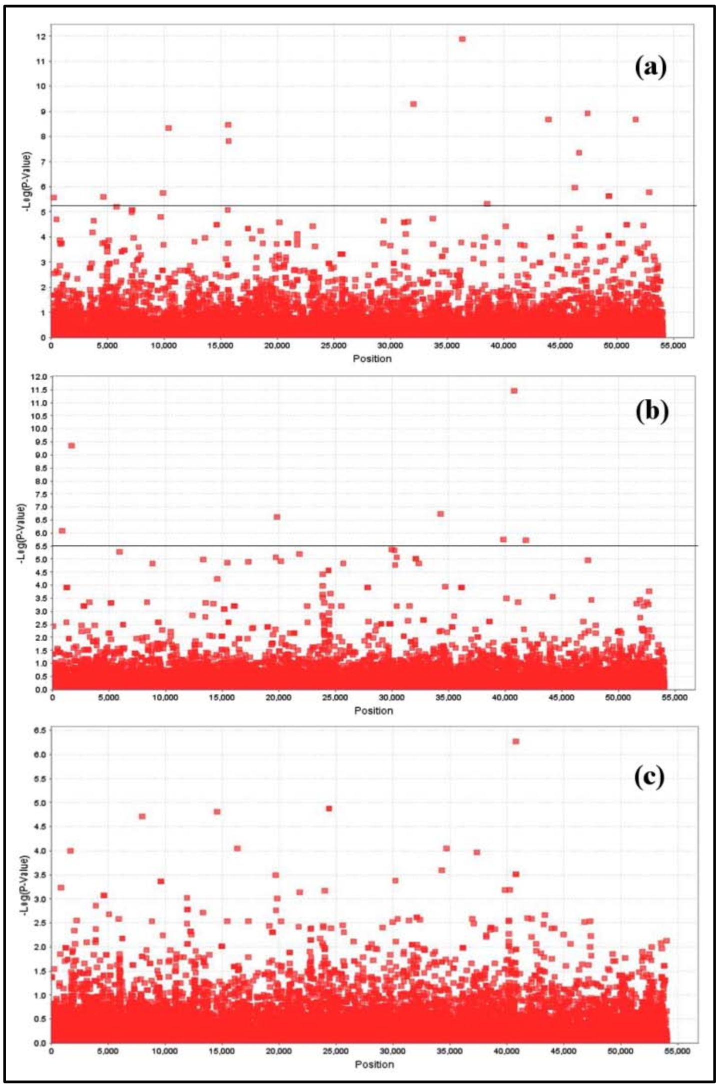


Figure 5.9: GWAS for different traits [a) CPP, b) TPH, and c) PH] using 13,280 SNPs with K model

5.4 Discussion

The advancement in sequencing technology dramatically reduces SNP identification cost. Further emergence of massively parallel array system has enabled immediate scoring of upto thousands of markers specifically SNPs in plants (Gupta et al. 2008). However, array based genotyping techniques suffer with many limitations viz., requiring prior sequence information, identification of polymorphism, validation and array production and thus establishment of such system in orphan plant is still comparatively a costly affair. An alternative technique, GBS, where sequencing is combined with allele scoring, therefore, bypassing the entire marker assay development is thus gaining interest. Since its invention, GBS has been used in many commercial and under-utilized crop plants for diversity analysis, population structure analysis and for GWAS (Deschamps et al. 2012). In present study also, GBS was used for SNP development, genetic diversity and population structure analysis. Along with that, utility of GBS for GWAS in linseed is demonstrated.

5.4.1 SNP analysis

Selection of the appropriate restriction enzyme (RE) is a critical step in developing a GBS library for a given species. As there was no size selection step during library preparation, it was important to maximize the proportion of restriction fragments that fall within the desired size range (100–400 bp) for sequencing. Out of three REs tested *viz., ApeKI*, *PstI* and *EcoT22I*, the enzyme *ApeKI* produced a large fraction of DNA fragments in the 100-400 bp range and was selected for GBS library preparation. In maize, *ApeKI* was found to preferentially cut in the low copy fraction of the genome and also widely used in other GBS studies (Elshire et al. 2011). The paired end reads obtained were mapped against reference genome of CDC Bethune (Wang et al. 2012c) variety of linseed and the SNPs were called using TASSEL ver. 3.0 GBS pipeline and after filtration, 13, 280 SNPs were identified, with an average density of one SNP per 27.86 Kbp. These SNPs could be sufficient for GWAS in linseed considering its smaller genome size of 375 Mbp (Ragupathy et al. 2011; Wang et al. 2012c). Additionally, because of simultaneous SNP discovery and genotyping, this sequencing-based SNP map is expected to have little ascertainment bias and greater power for mapping studies (Brachi et al. 2011). Further, the structural and functional relevance of SNPs were also investigated by comparing the location of the 13,280 SNPs with the coordinates of predicted genes. These SNPs can be used for candidate gene studies for the respective genes.

5.4.2 Population structure and linkage disequilibrium study

Understanding the genetic relationships and structure of the accessions is critical to control false positives in association mapping (Myles et al. 2009). Thepopulation structure within 95 accession used in present study was analysed using STRUCTURE software and optimum number of subpopulation were determined using the method descibered by (Evanno et al. 2005; Pritchard et al. 2000). Two subpopulations were identifed by both the method. Further, NJ and PCoA analyis also divided the accessions into two major subgroups, conforming the STRUCTURE results. However, very low *F*st value (0.006) showed weak population structure. We did not observe any geographic or phenotypic correlation among the accessions within each subpopulation. This could be because a majority of the accessions (65%) were local land races (LLR), whose specific origin could not be assigned. Additionally, the passport data may be occasionally weak where the donor country is considered as the country of origin. Soto-Cerda et al. (2013a) also observed two major groups in globally distributed flax accessions, supporting our results. The two major groups supported by our combined approach showed weak population subdivision making it ideal for association mapping.

The SNP data were also used for LD analysis. Three methods *viz*. D, D' and r2 are widely used for LD estimation. We used the correlation squared (r^2) method because: (a) it is not much influenced by small sample sizes and low allele frequencies (Flint-Garcia et al. 2003), and (b) it is relevant for QTL mapping because it relates the amount of variance explained by the marker to the amount of variance generated by the associated QTL (Zhu et al. 2008). About 5.3% loci were in significant LD with mean r^2 value for significant loci was 0.16. Further, as chromosome wise genome sequence is not available for linseed, a sequential approach was used for LD decay analysis. Three contigs with the highest number of SNPs mapped, *viz*., contig_25 (2818.9 Kbp, 197 SNPs), contig_67 (2844.03 Kbp, 182 SNPs) and contig_123 (2472.80Kbp, 281 SNPs) were selected for analysis. Overall LD decay was around 75 Kbp which was respectively around 75 Kbp, 112 Kbp and 100 Kbp for the contigs, 25, 67 and 123.

Though we cannot directly compare our results with that of Soto-Cerda et al. (2013a) due to different marker systems used, they also observed low LD among linseed accessions. Hence, the low LD could be because of narrow genetic diversity within the linseed germplasm, as LD is influenced by the level of genetic variation captured by the target population. For example, in wild barley (*Hordeum vulgare* ssp. *spontaneum*), in spite of its high rate of self-fertilization (~98%), LD decayed within 2 kb, a value similar to that observed in maize, an out-crossing species (Morrell et al. 2005). The low LD means large numbers of markers are required for marker trait association and SNPs are the most suitable markers in such situation.

5.4.3 Genome-wide association mapping

A preliminary GWAS study with SNP developed and phenotypic data of single location was carried out to demonstrate utility of GBS. In future, a detailed study with multi-location and replicated phenotypic data needs to be done in order to identify stable and promising SNPs. Population structure within the mapping population severely affect the LD and thus in turn results in spurious associations (Cappa et al. 2013). Therefore, selection of the best model plays an important role in GWAS to remove spurious associations. Comparatively less difference in P value distribution among all the models was observed. This might be because of weak population structure present within accessions selected. In contrast, Pasam et al. (2012) observed excess of small P values using the naïve (simple) model for AM in a highly structured barley germplasm. Out of four models tested, model QK and K showed good fit of P values. However, K model works comparatively faster and thus selected for GWAS. Other studies in plants also found K as more suitable model for AM (Pasam et al. 2012; Wang et al. 2012a). Nevertheless, it should be kept in mind that population structure corrections not only help in reducing the frequency of false positives, but also may entail false negatives in situations where a character state is strongly correlated with population structure (Cockram et al. 2010).

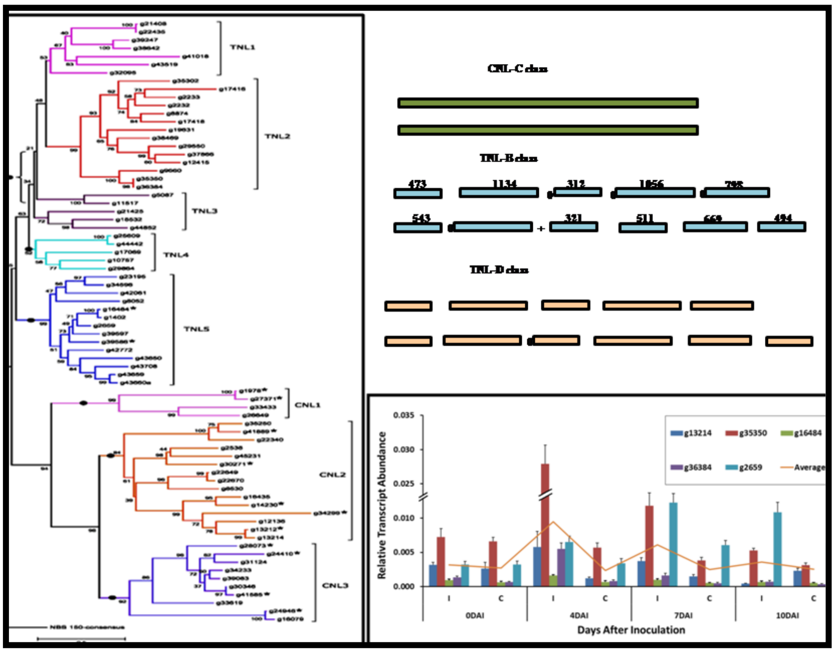
Although, it is very well known that accurate and replicated phenotyping at multiple environment need to be performed for understanding G and E contribution to the phenotypic variation, we performed preliminary analysis to demonstrate the utility of GBS for GWAS in this crop. Total 24 SNPs were identified at 5% FDR level, with maximum 16 for CPP, seven for TPH and one for PH. The marker R^2 value (percentage of genetic trait variation explained) ranged from 0.24 to 0.45.Although we cannot quantify the environment effect or GXE effects due to single year data, such lowR^2 values were also reported in other GWAS studies. For e.g., Roy et al. (2010), reported R^2 values to range from 0.2% to 3.95% in GWAS in plants. Many GWAS in humans have reported low R^2 values, while the remaining unexplained variation was termed as unexplained missing heritability (Manolio et al. 2009; Wang et al. 2012b). Several explanations have been proposed for this “missing heritability” including: i) insufficient marker coverage, in cases where the causal polymorphism is not in perfect LD with the genotyped SNP, which reduces the power to detect associations and in turn, the variation explained by such a SNP marker; ii) rare alleles

(MAF < 5%) with a major effect might go undetected in cases where they are associated; iii) the expression of a character or trait depends on a large number of genes/QTLs with small individual effects, which escape statistical detection; iv) inadequacy of the available statistical approaches to detect epistatic interactions in GWAS; and v) underestimated effect size of associated SNPs due to incomplete linkage with causal variants (Frazer et al. 2009; Gibson 2010; Maher 2008; Manolio et al. 2009; Wang et al. 2012b). Although these reasons were mainly discussed with respect to GWAS in humans, they also pertain to GWAS in plants and other organisms. In addition, the statistical model employed for the analysis will affect the variation explained by the SNPs (Pasam et al., 2012). As the stringency and threshold of the models increases, the power of detecting small effect SNPs will be reduced as the larger portion of the trait variation is explained by the model itself and the less variation is left to be explained by genetic effects. Hence, GWAS in inbreeding crops will depend on the careful optimization of the model regarding sensitivity vs. selectivity.

In summary, we developed a large number of SNPs for linseed and used for genetic diversity and population structure analysis in linseed accessions. Further, the use of GBS for GWAS in linseed was demonstrated. A detailed analysis with multiple year and multi-location phenotypic data will help to identify stable and promising SNPs for respective traits.

CHAPTER 6

Genome-wide identification and characterization of nucleotide binding site leucine rich repeat genes in linseed



A part of this work has been published in

Genome, 2013, 56(2): 91-99

6.1 Introduction

Plants have developed a repertoire of resistance (*R*) genes containing various conserved domains. Many such genes conferring resistance to a wide range of pathogens and pests have been identified and cloned from numerous plant species. Based on the presence of conserved domains, the *R* gene products have been grouped into five major classes, *viz*. detoxifying enzymes, kinases, nucleotide binding site-leucine rich repeat (NBS-LRR) proteins, extracellular receptors and receptor kinases (McDowell and Woffenden 2003). Among these, the NBS-LRR class is the most abundant and has been identified from a wide range of plant species, from non-vascular plants to angiosperms (Kohler et al. 2008; Mun et al. 2009; Xue et al. 2012). The NBS domain of the *R* genes is responsible for the binding and hydrolysis of ATP and GTP (Tameling et al. 2002), whereas LRR is typically involved in protein-protein interactions and is partly responsible for recognition specificity (Leister and Katagiri 2000). In addition, the ARC domain lying between the NBS and LRR domains was identified in potato and reported to play a role in recruitment of the LRR domain to the N-terminal region of NBS to maintain the molecule either in active or inactive states (Rairdan and Moffett 2006).

The plant NBS-LRR genes have further been categorized into two subgroups; TIR and non-TIR, based on the presence (TIR) or absence (non-TIR) of an N-terminal domain with homology to the receptor domain of the innate immunity factors Toll and Interleukin-1 found in animals (Parker et al. 1997). The non-TIR NBS-LRR proteins alternatively contain N-terminal coiled-coil (CC) or leucine zipper (LZ) motifs (Dangl and Jones 2001) and display conspicuous differences in amino acid motifs within the NBS domain. Functions of these domains have not been clearly elucidated, but they are predicted to be involved in signal transduction pathways (Dangl and Jones 2001; McDowell and Woffenden 2003). The NBS domain of diverse *R* genes is fairly conserved and contains various distinct motifs such as P-loop/Kin 1a, RNBS-A, Kinase-2, RNBS-B, RNBS-C, GLPL, RNBS-D-TIR and RNBS-D-non-TIR (Meyers et al. 1999); of which P-loop, Kinase 2, GLPL and RNBS-Dare the most conserved (Xue et al. 2012). These motifs have been extensively utilized to identify resistance gene analogs (RGAs) in model plants and various crop species (Palomino et al. 2006; Yaish et al. 2004) and to understand the genomic architecture of this gene family. Genome sequencing efforts of plant species have facilitated genome-level investigation of the NBS-encoding gene family in various plants; for example, *Arabidopsis* (Meyers et al. 2003; Tan et al. 2007), rice (Monosi et al. 2004), *Medicago* (Ameline-Torregrosa et al. 2008), poplar (Kohler et al. 2008), grape (Yang et al. 2008b), sorghum (Paterson et al. 2009), papaya (Porter et al. 2009) etc. However, very little is known about the NBS-LRR genes in linseed, as such studies have not yet been reported in this crop.

The yield of linseed is severely affected by various fungal diseases caused by *Fusarium* spp., *Alternaria* spp., *Melamspora lini* etc., and pests such as bud fly (*Dasyneura lini*). Classical genetics studies reported the presence of three resistance genes (*L6*, *M* and *P*) conferring resistance to *Melamspora lini* in linseed (Dodds et al. 2001). However, this strategy of resistance gene identification is very difficult and slow; and as a result, introducing durable disease resistance into agronomically superior varieties has been a major challenge in linseed breeding. Genomic evaluation of disease-resistance gene homologs can help in identifying the putative resistance genes and understanding the mechanism of disease resistance in linseed. Therefore, the main objective of this study was identification of putative NBS-LRR encoding resistance genes from the linseed genome using bioinformatics approaches, followed by confirmation of expression using quantitative real-time polymerase chain reaction (qRT-PCR).Recently sequenced genome of linseed variety CDC Bethune (http://www.linum.ca/) was analyzed to identify 147 NBS-LRR encoding *R* genes. The predicted linseed NBS-LRR genes were examined for the presence of conserved domains and motifs, and their gene structure, expression and phylogeny were studied to examine their relationships and evolution.

6.2 Materials and methods

6.2.1Inoculation procedure and sample collection

Seeds of the linseed variety Ayogi (tolerant to *Alternaria* spp.) were surface sterilized using sodium hypochlorite (0.1%) for 15 min, washed thoroughly with sterile distilled water and sown. The plants were grown in plastic pots containing mixture of soilless compost in controlled environment chambers at 18 ± 2°C. The *Alternaria lini* culture was obtained from C.S. Azad University of Agriculture and Technology (CSAUAT), Kanpur, India and was sent for confirmation to the Fungal Identification Services Division, Agharkar Research Institute (ARI), Pune, India. After confirmation, the isolates were maintained in slants of potato dextrose agar (PDA) at 4°C. The inoculation procedure and the disease conducive conditions in the growth chamber were as described by Vloutoglou et al. (1999). Linseed plants at growth stage (GS) 5–6 (16–18 true leaves) were artificially inoculated (sprayed until run-off) with the conidial suspension of *A. lini* isolates (10^6 spores/ml). Approximately 20 ml of the conidial suspension was sprayed onto the plants in each pot (10 plants per pot). Mock inoculation was similarly performed using sterile distilled water. The inoculation procedure lasted for ~30 min and at the end of this period, no conidial germination was observed. The plants were inoculated at the beginning of a dark period (approximately 16:00 h) and the pots were covered with plastic bags to maintain high humidity and then wrapped with aluminum foil to create darkness. Samples from inoculated and mock-inoculated plants were collected at 0, 4, 7 and 10 days after inoculation (DAI), immediately frozen into liquid nitrogen and stored at -80°C until use.

6.2.2 Identification of NBS-LRR genes in linseed

All the 47,912predicted gene models of the linseed genome (http://www.linum.ca/) were used to identify the NBS-LRR genes. The gene models were searched against the Pfam database using Pfam ver. 24 software with an e-value of 0.1 (Bateman et al. 2002), and the results were analyzed to identify TIR, NBS and LRR containing sequences. The NBS containing sequences were further analyzed using the COILS software (Lupas et al. 1991) at a threshold of 0.9 to detect coiled-coil (CC) domains. Alternatively, the 114 curetted *R* gene sequences (http://www.prgdb.org/) were also searched against the 47,912predicted gene models of linseed with blastp, using a threshold e-value of 10^{-15}. The obtained hits were searched against the Pfam database and COILS software as described above to identify TIR, NBS, LRR and CC domains. Non-redundant NBS-LRR genes were determined by eliminating the duplicates.

6.2.3 Identification of homologs of linseed NBS-LRR genes

The whole genome protein sequences of *Arabidopsis* and poplar were downloaded from ftp://ftp.tigr.org/pub/data/a_thaliana/ath1/SEQUENCES and http://genome.jgi-psf.org/Poptr1/Poptr1.download.ftp.html, respectively and protein databases were created on an in-house server. Viroblast (ver. 2.2+) (Deng et al. 2007) was configured to use the stand-alone suite of BLAST programs (ver. 2.2.25+), and was used to search the *Arabidopsis* and poplar protein databases with the predicted protein sequences of linseed *R* genes as well as the predicted NBS part of these genes using blastp. The sequences of *Arabidopsis* and poplar orthologs with the highest match scores were imported in Microsoft Excel (ver. 2007) and further analyses were performed.

6.2.4 Phylogenetic analysis of linseed NBS-LRR genes

Only the NBS domain was used for phylogenetic analysis as it was conserved in both CNL and TNL genes. The NBS domain sequences of all the genes were extracted and the sequences having NBS domain length ≥150 aa were used for alignment. The HMMER3 software (http://hmmer.janelia.org) was initially used for building the profile of NBS domain and then for alignment. The alignment was edited using MEGA ver. 5.0 software (Tamura et al. 2011), and the sequences containing fewer than 75% of the hidden Markov model (HMM) match-state residues were retained for subsequent analysis. Indels and poorly aligned regions were removed by trimming the regions outside the HMM match-states prior to phylogenetic tree construction. Phylogenies were calculated using parsimony and bootstrapped neighbor joining algorithms. The trees were constructed using protpars in the Phylip suite ver. 3.6 with bootstrap value of 100 using the Mobyle portal (http://mobyle.pasteur.fr/). The input sequence order was jumbled five times and topologies were calculated based on each sequence order. One most-parsimonious tree was chosen at random to serve as the basis for branch length calculations. Maximum likelihood branch lengths were calculated on the parsimony topologies with TreePuzzle ver. 5.2 (Schmidt et al. 2002) using the substitution model of Muller and Vingron (2000). Amino acid frequencies were calculated from the input trees and rate heterogeneity was allowed with four γ rate categories. A neighbor joining tree of the cleaned alignment from hmmalign was calculated using MEGA ver. 5.0 (Tamura et al. 2011) with 1,000 bootstrap replicates. In addition, two more trees were constructed using: (i) *Arabidopsis*, poplar and linseed NBS-LRR sequences, and (ii) curetted NBS containing *R* gene sequences (http://www.prgdb.org) and the linseed NBS-LRR sequences.

6.2.5 Analysis of conserved motif structures

The domain and motif analyses were carried out to explore the structural diversity among the predicted NBS-LRR genes. Accordingly, the sequences were divided into three classes as those containing the N-terminal motif (sequence upstream the P-loop of NBS domain); the NBS domain (P-loop to WMA) and the LRR-C-terminal domain (sequence downstream the WMA of NBS domain). The three classes of sequences were separately analyzed using the MEME/MAST system (http://meme.sdsc.edu/meme/website/intro.html) (Bailey et al. 2006) to investigate the protein motifs in more detail. The MEME (Multiple Expectation Maximization for Motif Elicitation) motif analyses were performed on each of the subgroup separately (e.g., TNL [containing the domains **T**IR, **N**BS and **L**RR], CNL [containing the domains **C**C, **N**BS and **L**RR]) and with settings designed to identify 15, 20, 30 or 50 motifs. MAST (Motif Alignment and Search Tool) (Bailey and Gribskov 1998) was used to assess the correlations between MEME motifs and the associated distance matrices.

6.2.6 Analysis of promoter regions of NBS-LRR genes

For each NBS-LRR putative gene, 2 kb upstream region was selected according to the position of the genes provided by Gbrowse annotation (http://www.linum.ca/) on the scaffold sequences of linseed. The extracted sequences were screened against the PLACE database (http://www.dna.affrc.go.jp/PLACE/) to identify the *cis*-acting regulatory elements. The regulatory elements overrepresented in the dataset and those known to be involved in regulation during the resistance response and under stressed conditions were selected for further analysis (Jang et al. 2006). Among them, WBOX [TGAC (C/T)], CBF (GTCGAC), DRE [(G/A) CCGAC] and GCC boxes were retained for further analysis as they were reported to be present in the promoters of resistance genes (Jang et al. 2006).

6.2.7 *In silico* and qRT-PCR expression analysis

The putative NBS-LRR genes were BLAST searched against the NCBI linseed EST dataset (dated: April 11, 2011; 2,86,894 sequences, http://www.ncbi.nlm.nih.gov/nucest?term=Linum%20usitatissimum) to obtain transcriptional evidence for individual NBS-LRR genes and to estimate the number of ESTs expressed per tissue type and gene model. A >85% sequence identity criterion was used for any EST to map onto a gene model. These tissue types include flower (FL), seed coat at globular stage (GC), pooled endosperm (EN), seed coat at torpedo stage (TC), heart embryo (HE), globular embryo (GE), torpedo embryo (TE), bent embryo (BE), mature embryo (ME), etiolated seedling (ES), stem (ST), leaf (LE), peeled stem (PS) (Venglat et al. 2011), 12 DAF bolls and outer fibrous stem tissue.

Total RNA from both inoculated and control samples was extracted using Spectrum Plant Total RNA kit (Sigma-Aldrich, USA) and treated with DNaseI (Promega, USA). The RNA was reverse transcribed using MultiScribe™ reverse transcriptase (Applied Biosystems, USA) and oligo (dT) primer. Nineteen gene models were randomly selected for quantitative expression analysis. Primers spanning various regions of RGAs were designed using Primer3 (Rozen and Skaletsky 2000).qRT-PCR was carried out in 7900HT Fast real-time PCR system (Applied Biosystems, USA). Each 10 µl qRT-PCR cocktail contained 0.25 µM each of forward and reverse gene-specific primers, 4 µl of 1:10 diluted first strand cDNA, 1× FastStart universal SYBR green master mix (Roche, USA) and sterile milliQ water.

Following cycling conditions were used: 95°C denaturation for 10 min, followed by 40 cycles of 95°C for 3s, primer annealing at 55°C for 15s 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 to monitor the specificity of each primer pair. The eukaryotic translation initiation factor 5A (*ETIF5A*) gene from linseed was used as a reference gene for the qRT-PCR (Huis et al. 2010). The housekeeping gene was selected after confirming its stability of expression across the tissues used in the study. Two technical replicates for each of the three biological replicates were performed. PCR efficiencies of the housekeeping gene and the gene of interest were calculated using LinRegPCR (Ramakers et al. 2003) and the qRT-PCR conditions were optimized such that the efficiencies were similar and close to 2.0. Relative transcript abundance calculations were performed using comparative CT ($^{\Delta}$CT) method described by Schmittgen and Livak (2008).

6.3 Results

6.3.1 Mining the linseed genome for NBS-LRR genes

The scaffold assembly of the linseed genome (http://www.linum.ca/) contained 302 Mb of non-redundant sequences representing estimated 85% genome coverage. This coverage is consistent with the length of the entire low-copy fraction previously estimated by reassociation kinetics and captures nearly the entire gene-rich portion of the linseed genome (http://www.linum.ca/). The 47,912 linseed gene models were searched against the Pfam database to identify TIR, NBS and LRR domains. Additionally, the gene models were also searched using the 114 curetted *R* gene sequences from PRGdb (http://www.prgdb.org/) as blast queries using e-value of 10^{-15}, which identified 2,220 *R* gene like sequences. These hits were searched against the Pfam database and COILS software to identify TIR, NBS, LRR and CC domains. The results of both the analyses were the same and 147 non-redundant NBS-LRR encoding genes were identified from the linseed genome (**Table S6.1**). To identify partial and truncated NBS-LRR genes, a minimum size of 24 amino acids was considered for the NBS domain.

6.3.2 Gene structure and intron-exon configurations of linseed NBS-LRR genes

We analyzed the gene structure, intron positions and phases of the linseed CNL and TNL genes to assess the diversity within and between each family **(Table S6.2)**. The intron/exon boundaries of the linseed genes (available at http://www.linum.ca/) were analyzed for the 147 NBS-LRR genes. To determine their intron/exon configurations, blastp search was carried out using the full protein sequence as well as only NBS region of the predicted linseed *R* proteins against all the *Arabidopsis* proteins. The structures of the top matched *Arabidopsis* genes were compared with those of the corresponding linseed genes, which revealed that the number of introns was less in CNLs (0-13) than in TNLs (2-16) (**Table S6.2 and S6.3; Figure6.1**).

Distinct patterns of gene structure were identified in CNLs [CNL-A (8) and CNL-C (40)] and TNLs [TNL-B (21), C (13) and D (56)] **(Table 6.1 and Figure 6.1)**. Most CNLs (25) showed a common gene structure as observed in the *Arabidopsis* CNL-C protein At3g14470. The TNL family genes showed complex structures usually consisting of some short exons in the LRR region and most of them were interrupted by a phase-0 intron **(Figure 6.1)**. In these, various domains (TIR, NB-ARC, LRR) were often separated by introns. Interestingly, the linseed proteins, which matched with high confidence to a type ‘A’ protein (At4g36140) from *Arabidopsis* having unusual structure (TNTNL), had only TN or TNL domains. Similarly, the linseed *R* genes (g43650, g43708 and g8052) having high similarity to the At1g27170 *Arabidopsis* protein (a TNL-C protein with canonical structure), either lacked the LRR domain or had unusual structures.

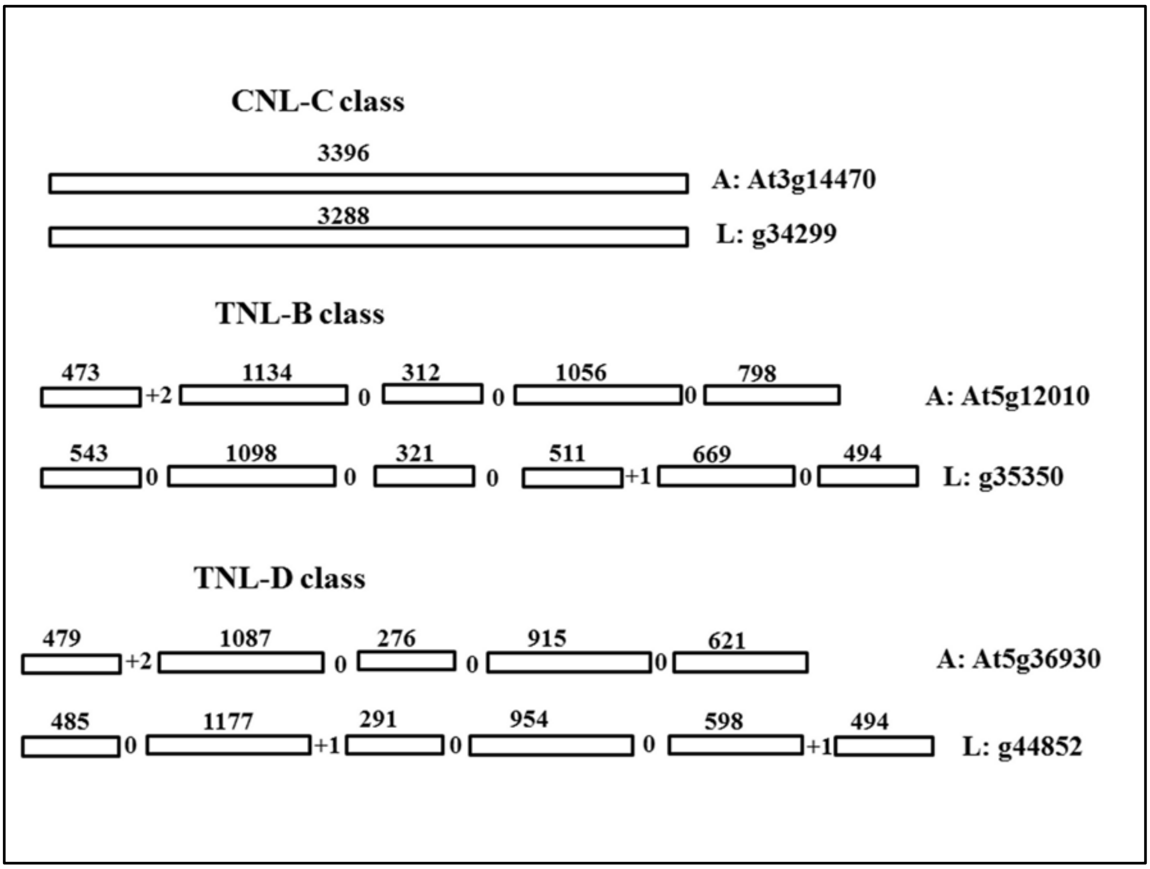


Figure 6.1: Comparison of the intron/exon configuration of selected *R* genes from linseed (‘L’) with related *Arabidopsis* (‘A’) genes. Numbers above exons indicate the size of the exons in base-pairs. Numbers between exons denote intron phases.

6.3.3 Phylogenetic analysis of NBS-LRR genes

To assess the sequence diversity and evolution of the linseed *R* genes, a phylogenetic tree was constructed using only the NBS domain of linseed *R* proteins as it is conserved in both CNL and TNL proteins and contains numerous conserved motifs **(Figure 6.2)**. The NBS consensus sequence developed from the extended HMM model was used as an out-group to root the tree. To understand the evolution of linseed genes, *Arabidopsis* (44) and poplar (104) proteins having high sequence similarity to the linseed TNL and CNL proteins as well as the curetted NBS containing *R* proteins from PRGdb (http://www.prgdb.org) were included in the phylogenetic analysis. Both parsimony- and distance-matrix-methods yielded very similar results (data not shown). From the values obtained in the bootstrap analysis, it was apparent that most of the deep nodes of the tree have high statistical significance.

The tree showed long branch lengths and closely clustered nodes, reflecting a high level of sequence divergence **(Figure 6.2)**. As reported in earlier studies, the phylogenetic tree was divided into two main clusters, TNL and CNL (Meyers et al. 1999) that were further divided into five TNL (TNL-1, TNL-2, TNL-3, TNL-4 and TNL-5) and three CNL (CNL-1, CNL-2 and CNL-3) clades, respectively **(Figure 6.2)**. Interestingly, the truncated *R* genes (TN, CN, NL and N) were dispersed among various clades throughout the tree, indicating a diverse rather than a monophyletic origin due to domain loss. In the CNL family, CNL-2 was the largest clade (15) and was composed of the type 'C' linseed *R* proteins **(Figure 6.2)**. It mainly contained poplar *R* proteins and five known *R* proteins of diverse origin **(Figure 6.3)**. The CNL-3 clade was also composed of type 'C' linseed *R* proteins (10), and showed high similarity with *Arabidopsis* RPM1 *R* protein. Interestingly, all the linseed *R* proteins of CNL-1 clade (4) were of type 'A' and showed high similarity to *Arabidopsis R* proteins (At5g66920 and At4g433300). However, they did not cluster with any characterized *R* proteins **(Figure 6.4)**.

Among the TNL family, clades TNL-2 (14) and TNL-5 (14) contained majority of *R* proteins. The TNL-2 clade composed of mainly 'B' type *R* proteins and clustered with the well characterized P2 protein of linseed and poplar *R* protein; whereas, TNL-5 clade composed of both 'C' and 'D' (3 with unusual domain architecture) type *R* proteins, in equal proportion and showed high similarity to the previously characterized linseed *R* genes (L6: AAA91022 and M: AAB47618). The clades TNL-1 and TNL-3 clustered with WRKY transcription factor 52 of *Arabidopsis* and resistant proteins from Solanum family, respectively. Both the clades composed of TNL-D type linseed *R* proteins and clustered mainly with poplar *R* proteins, except one linseed *R* protein with unusual domain architecture (g21425; TNLL), which clustered with *Arabidopsis R* proteins. The TNL-4 clade contained five *R* proteins and appeared to be paraphyletic in origin **(Figure 6.2)** and did not cluster with any characterized *R* proteins. Interestingly, the 'D' type *R* proteins were observed in all the TNL clades.

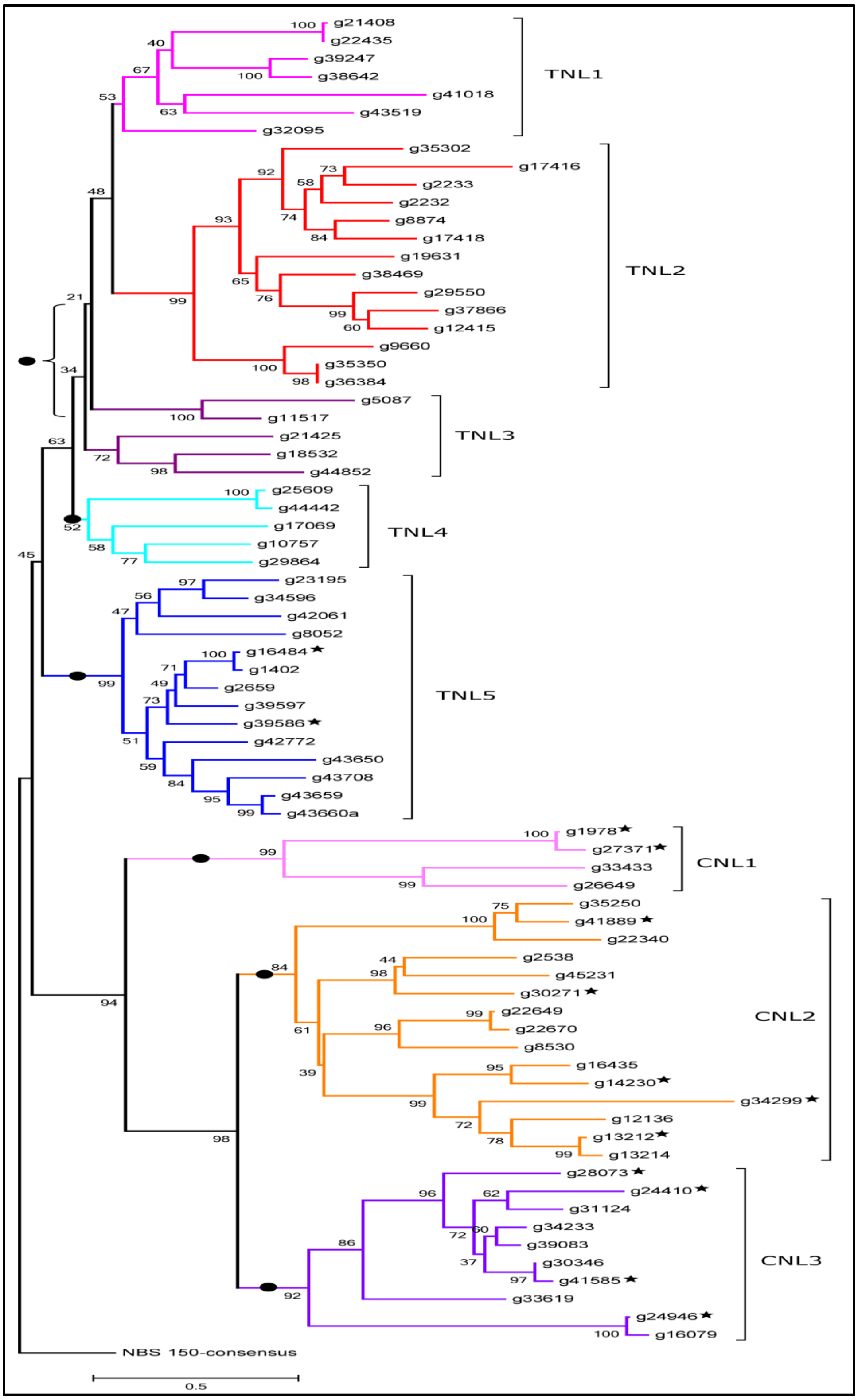


Figure 6.2: Neighbor joining bootstrap tree of NBS domains of NBS-LRR predicted proteins from linseed. The dots indicate approximate coalescence points for NBS sequences from *Arabidopsis* and/or poplar. The '*' indicate truncated genes.

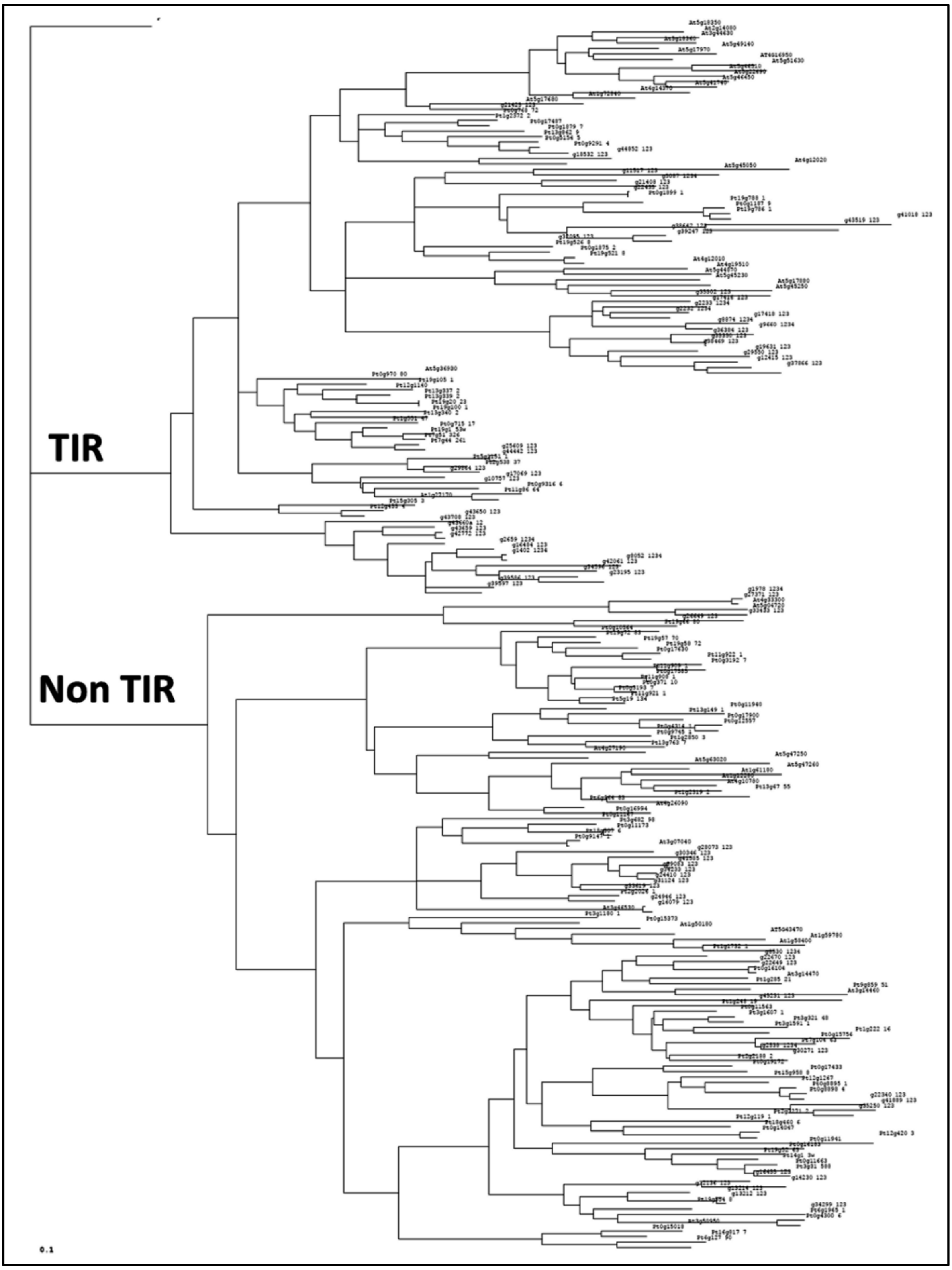


Figure 6.3: Maximum parsimony tree with maximum likelihood branch length generated using NBS domains of NBS-LRR predicted proteins from linseed, *Arabidopsis* and poplar.

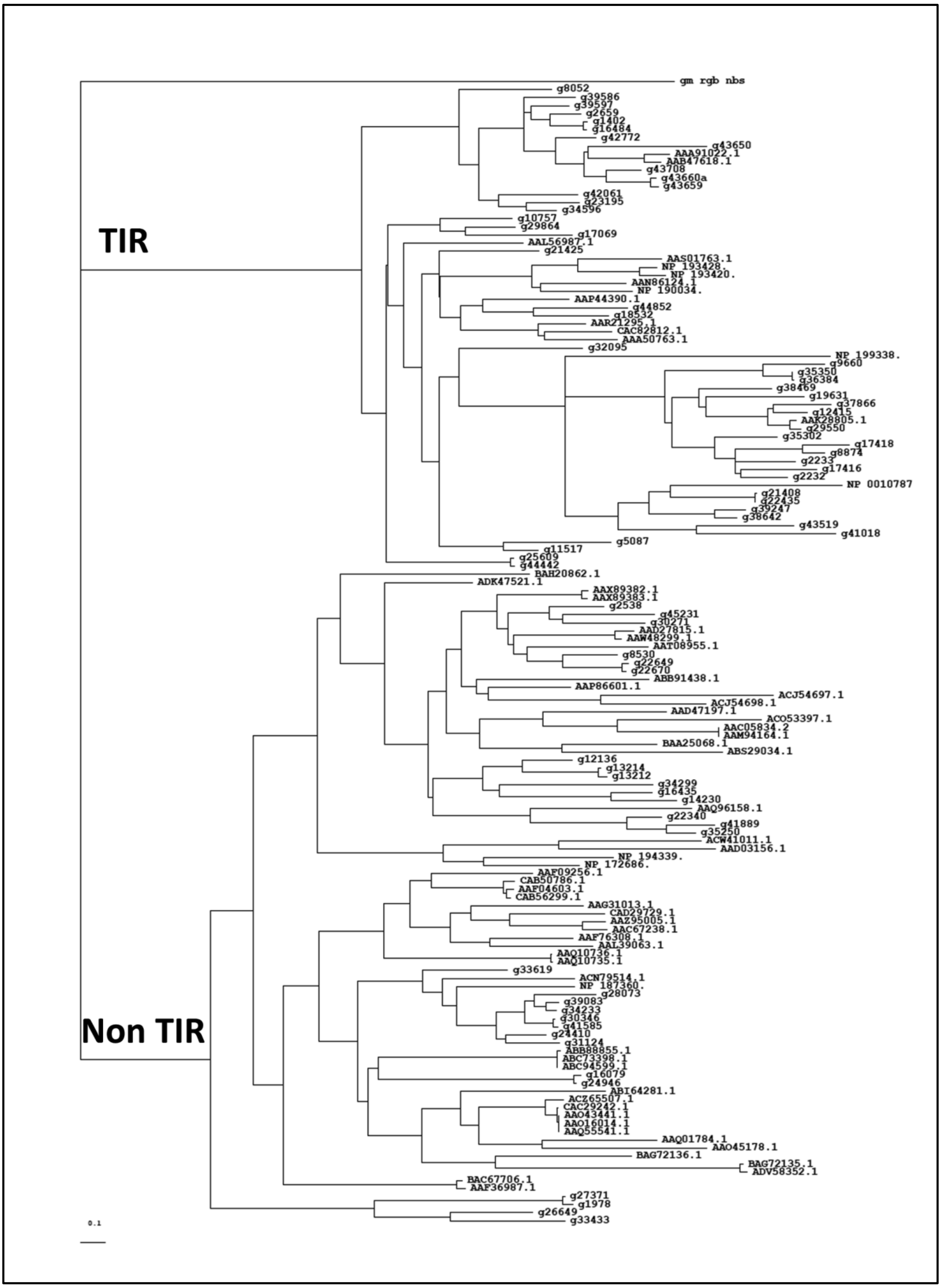


Figure 6.4: Maximum parsimony tree with maximum likelihood branch length generated using NBS domains of NBS-LRR predicted proteins from linseed, and that of curetted *R* genes from PRGdb

6.3.4 Conserved domains and motifs in linseed NBS-LRR genes

Various domains used for classification of NBS-LRR genes such as TIR, NBS, LRR, CC etc. were identified in the putative 147 NBS-LRR genes as described before. Based on the Pfam analysis of the predicted proteins as well as homology with previously described motifs within the NBS domain (Meyers et al. 1999), the 147 genes were divided into two major families, CNL (49) and TNL (98) **(Table 6.1)**. From the 49 CNLs, 13 sequences lacked LRR (CN and N), while five sequences were devoid of CC domains (NL and N). The fusion of RpW8 and PRK domain with the C-terminal region of genes belonging to CNL family was observed **(Table S6.1)**. On the contrary, diverse domain arrangement was observed in the TNL family. The most common type was typical TNL 57/98 (58.17%) followed by TN (22/98, 22%) **(Table 6.1)**. In addition, other domains, *viz*. F box, gal binding lectin and pg-box, were also identified in the gene models of TNL class.

Further, the N terminal, NBS and C-terminal regions of the putative linseed *R* gene proteins were analyzed separately using the program MEME (Bailey et al. 2006)**.** The N-terminal region ranged from 148-248 and 144-1168 amino acids in CNL and TNL families, respectively; and both TIR and CC domains were identified in this region. Fifteen distinct motifs were observed in the N-terminal and NBS regions of CNL and TNL families **(Tables S6.4 and S6.5)**. Several conserved motifs (TIR 1-4) were observed and their order was also conserved in the N-terminal region of the TNL linseed *R* proteins **(Table S6.4)**. The length of the NBS domain ranged from 243-294 amino acids and the crucial motifs of the NBS domain (P-loop/kinase-1a, kinase-2, and kinase-3a and GLPL) were highly conserved among the predicted linseed *R* proteins. On the contrary, the LRR C-terminal region was highly variable in sequence and size (1-898 amino acids in CNL and 26-1854 amino acids in TNL family, respectively) and 50 motifs were identified in both CNL and TNL sequences.

Table 6.1: NBS-LRR genes classified into different classes and their domain compositions

Class		**Domains**																
		CN	**CNL**	**CNNL**	**CNNNL**	**N**	**NL**	**TCNL**	**TLTNL**	**TN**	**TNL**	**TNLTNL**	**TNN**	**TNNL**	**TNNN**	**TNTNL**	**TTNL**	**Total**
C	**CNLA**	2	5	-	-	1	-	-	-	-	-	-	-	-	-	-	-	**8**
N	**CNLB**	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	**1**
L	**CNLC**	8	26	2	1	1	2	-	-	-	-	-	-	-	-	-	-	**40**
T	**TNLA**	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	**1**
N	**TNLB**	-	-	-	-	1	-	-	-	6	12	-	1	-	-	1	-	**21**
L	**TNLC**	-	-	-	-	1	2	-	-	2	8	-	-	-	-	-	-	**13**
	TNLD	-	-	-	-	1	1	2	1	8	36	1	2	1	1	1	1	**56**
	TNLH	-	-	-	-	-	-	-	-	6	1	-	-	-	-	-	-	**7**
Total		**10**	**31**	**2**	**1**	**7**	**5**	**2**	**1**	**22**	**57**	**1**	**3**	**1**	**1**	**2**	**1**	**147**

6.3.5 Promoter region analysis of NBS-LRR genes

To understand regulation of the NBS-LRR genes, 2 kb region upstream the predicted NBS-LRR genes was used to identify promoter sequences **(Table S6.6)**. Four regulatory elements (WBOX cassettes, Type II MYB and G boxes associated with the WRKY transcription factors (Dong et al. 2003), CBF and DRE boxes (Sakuma et al. 2006) and the GCC motif associated with the ERF-type transcription factors (Ohme-Takagi et al. 2000) were overrepresented. These regulatory sequences are associated with responses to biotic or abiotic stresses. Among all, WBOX elements were the most abundant, averaging 4.59 for the CNL and 4.49 for the TNL families, respectively. Of the predicted NBS-LRR genes, 66% contained between 4 and 10 copies of WBOX elements, with four and six copies being the most common. On the contrary, the average numbers of other element types were 0.62 (CBF), 0.62 (GCC), and 0.28 (DRE) and 1.24 (associated with WRKY transcription factors) and observed only once per upstream region. In few cases, multiple boxes of these regulatory elements were observed. Interestingly, the WBOX elements were fairly uniformly distributed across the tree **(Table S6.6)**. No groups showed systematic excesses or deficits of the less numerous CBF, GCC or DRE box motifs. No significant correlation between the arrangement of these promoter sequences (WBOX, CBF, GCC and DRE) and *in silico* expression was observed **(Table S6.6)**.

6.3.6 *In silico* and qRT-PCR expression analysis

In silico expression analysis of the identified NBS-LRR genes was performed by comparing these genes against the 2,86,894 EST sequences from 13 libraries downloaded from GenBank. A criterion of >85% nucleotide identity between ESTs and NBS-LRR genes was applied to gain EST support for the genes. In cases where one EST matched with multiple candidate genes, only the gene with the highest alignment score was considered. At this threshold, 35/147 genes (representing 23% of the predicted NBS-LRR genes) exhibited EST support **(Table S6.7)**. In all, 56 EST matches were identified and the frequency of ESTs varied from 1 to 10 per NBS-LRR gene model. These genes were expressed in a wide range of cDNA libraries, including those constructed from various development stages and tissue types. Among the various tissue types, peeled stem (PS, 32%) had the largest number of expressed genes. The percentage of the genes expressed per clade varied from 2.7% to 47% and even between highly similar genes within the same group **(Figure 6.2; Table S6.7)**. A majority of the TNL genes had EST support and the highest of 18 ESTs were mapped to seven genes of the TNL-5 clade, whereas only three ESTs mapped to three genes from the CNL-2 clade. Among all the expressed genes, g2659 showed the highest expression in stem (ST), leaf (LE) and mature embryo (ME) **(Table S6.7)**.

The qRT-PCR was performed to investigate expression of the putative NBS-LRR genes. Five of the nineteen gene models **(Table 6.2)** showing stable and consistent amplification were selected for quantitative expression analysis. The expression profiles of the selected gene models are presented in **Figure 6.5**. The highest level of expression was observed at 4 and 7 DAI in inoculated tissues, while in the control tissue, g2659 over-expressed at 7 and 10 DAI, although its expression was higher in other control tissues as well.

Table 6.2: Primers used for qRT-PCR analysis

Gene model	Forward (5'-3')	Reverse (5'-3')
g13214	TGTCTGCCGGAGTGGCTTCA	TCCTTGAGTGTGTGCATCCAGATTG
g35350	ACCTCAGCTTCGTCGGTTGC	CCCACCGTACGGAGATTGGA
g16484	TGTCCGTTGTGAAGGGTGGA	CCTCCGAGCGATTGTCATCA
g36384	CGGATTCGTCACAGCCCTCT	GGGACAACGGTCAGGTTTGC
g2659	GAAGCATCCTCGTCGTGTGC	CGCAAACAGCGTCAGTCCAT

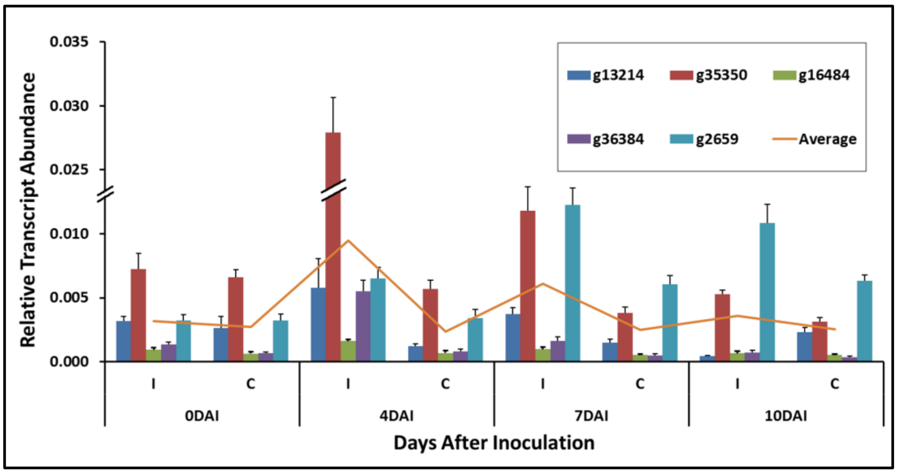


Figure 6.5: Expression profiles of five *R* genes in inoculated and control linseed tissues collected at 0, 4, 7 and 10 days after inoculation (DAI)

6.4 Discussion

Linseed is being grown since ancient times for fiber (flax) and oil-rich seed (linseed). Such a long cultivation history of linseed in a wide range of environmental conditions and exposure to various biotic and abiotic stresses suggests that numerous *R* genes for diverse biotic stresses might be present in the linseed genome. Therefore, there is a need for better understanding of the genomic organization of NBS-LRR encoding genes in the linseed genome and characterization of their genetic and molecular mechanisms. Recently, abundant genomic resources in the form of ESTs and genome sequence have been developed for this crop (Ragupathy et al. 2011; Venglat et al. 2011; Wang et al. 2012c). These resources were used for genome-wide discovery of NBS-encoding *R* genes from the linseed genome. Analysis of predicted domain structure, promoter regions, phylogeny and *in silico* gene expression revealed a number of prominent features as reported for NBS-LRR genes from the other plant species.

6.4.1 Evolution of linseed NBS-LRR genes

A total of 147 NBS-LRR coding genes representing about 0.3% of all the predicted linseed proteins were identified. This percentage is less than that reported in other plant genomes (0.6–1.8%) (Ameline-Torregrosa et al. 2008; Kohler et al. 2008; Meyers et al. 2003; Monosi et al. 2004; Yang et al. 2008b; Yu et al. 2005). This could be because of several reasons: i) due to assembly of short sequence reads- We used the genome sequences assembled from Illumina short sequence reads (44-75 bp) with 95X coverage (http://www.linum.ca). However, the NBS-LRR genes are reported to be multi-copy and highly diversified genes (Kuang et al. 2004; Yang et al. 2008a) and cannot be assembled yet by short sequence reads. Hence, the linseed genome might contain a large number of NBS-LRR genes than the 147 such genes identified in this study. ii) due to misannotation-nearly 50% of the initially identified genes (73 out of 147) were smaller in size and possibly resulted due to missed start and/or stop codons. About 36% errors in automated annotations were also observed in *Arabidopsis* (Meyers et al. 2003). iii) deletion of redundant NBS-encoding gene components after ancient whole genome duplication, as reported in other plants [Rice (Yu et al. 2005), grape (Yang et al. 2008b)**,** *Arabidopsis* (Nobuta et al. 2005) and poplar (Kohler et al. 2008)]. Likewise, it has been suggested that linseed might have undergone an ancient polyploidization event (Ragupathy et al. 2011). iv) distribution of NBS-LRR genes: It is well known that most NBS-LRR coding genes are unevenly distributed and clustered in specific chromosomal regions (Ameline-Torregrosa et al. 2008; Meyers et al. 2003; Monosi et al. 2004). v) Linseed may harbor a smaller number of NBS-LRR genes compared to other plant species. As the linseed genome is yet to be fully sequenced, there is a possibility that the genomic regions containing major NBS-encoding gene clusters are yet to be sequenced. This percentage might change when the linseed genome is fully sequenced and annotated. However, the estimates are similar to those obtained in *B. napa* (Mun et al. 2009) as well as papaya (Porter et al. 2009).

As observed in other dicots, phylogenetic analysis of linseed NBS-LRR genes showed two distinct families, the TIR-NBS-LRR related genes and the non-TIR NBS-LRR related genes, reflecting the well-known ancient differentiation of NBS-LRR genes into two major families (Meyers et al. 2003). Each family was further divided into multiple clades based on their clustering in phylogenetic tree and classes based on distinct domain organizations as described by Meyers et al. (2003) in *Arabidopsis* (TNL A-D and CNL A and C). The phylogenetic tree indicated that the linseed NBS-LRR genes probably originated from a small number of progenitor sequences in contrast to that in *Arabidopsis* and poplar. Alternatively, it is also possible that many progenitor genes might have subsequently been lost during linseed evolution. The above clades are linseed specific asthey contain very few of the *Arabidopsis* or poplar *R* genes. However, this needs to be validated by including more *R* genes from other closely related families of the order Malphygials. We could not attempt this ourselves due to the lack of genomic sequence data of these family members.

Many of the linseed *R* genes showed high sequence similarity in some clades (e.g. TNL-2 and CNL-3) indicated by the presence of many sequences with small branch lengths. The CNL family of proteins appeared to be more ancient than the TNL family proteins due to their long branch lengths. Similar results have also been reported for other plant genomes (Kohler et al. 2008; Meyers et al. 2003). Within seven CNL family genes, C terminal fusion of Rpw8 domain was observed. Similarly, C terminal fusion of Rpw8 domain has also been observed in the CU013515_1.4/Mt5g1164 gene of *Medicago* (Ameline-Torregrosa et al. 2008) and At5g66910 gene of *Arabidopsis*. Thus, it appears that these genes have remained as single copy genes in their respective genomes since the last ancestor of these plant lineages. The Rpw8 gene in *Arabidopsis* and other dicots provides broad-spectrum mildew resistance (Xiao et al. 2001). Thus, it can be presumed that such genes in linseed might also provide a broad spectrum disease resistance. In general, the CNL proteins are predicted to have basic defense roles as regional adaptation to the specific biotic stresses (Yang et al. 2008b).

The TNL genes in linseed share the largest proportion of the NBS-LRR genes and show large domain diversity compared to the CNL genes, as also reported in *Arabidopsis* (Meyers et al. 2003)**,** *Medicago* (Ameline-Torregrosa et al. 2008)and poplar (Kohler et al. 2008). Such greater diversity can be explained partly by the exon-intron and domain structure of TNL family genes. The linseed TNL proteins contained more introns than CNL proteins as observed in *Arabidopsis* (Meyers et al. 2003)**,** poplar (Kohler et al. 2008) and cereals (Bai et al. 2002). Most of the additional introns and domains in these TNL genes occur at the 3' end of the genes. It has been proposed that each of the TIR, CC, NBS and LRR domains evolved independently with distinct functions and later fused to form new proteins. For example, the potato CNL protein Rx can act in *trans* with the CC or the LRR domains expressed from separate proteins to produce hypersensitive response (Moffett et al. 2002).

Interestingly, most of the unusual domains identified in the present study were observed in the TNL-5 clade and all were type 'D' proteins. Thus, this clade or the type 'D' proteins seem to tolerate more domain arrangements than other groups. The extra domains present in the TNL-D might have fused during the gene/domain duplication or recombination events resulting in the diversity of the proteins. The evidence of exon additions or fusions was observed in WRKY-related domains and some metallopeptidases in *Arabidopsis* (Meyers et al. 2003) as well as BED/DUF1544 domain in poplar (Tuskan et al. 2006) genomes. The presence of TNTNL in *Arabidopsis* is predicted to arise from the fusion of TN and TNL genes. Interestingly, linseed TNL-A proteins showing high similarity to the *Arabidopsis* TNTNL protein (At4g36140) exhibited either TN or TNL domains, further supporting the evidence of exon fusion hypothesis. The occurrence of diverse genes within a single clade in TNL family (e.g. TNL-5 contains both type C and D proteins) was also observed. This might have resulted from clustered arrangement of NBS-LRR genes and their recombination. Similar clustering of diverse genes has also been observed in *Arabidopsis* (Meyers et al. 2003). Overall, the high diversity of TNL family genes indicates its role in recognizing "species-specific" pathogens (Yang et al. 2008b).

Several classes of truncated *R* genes [CC-NBS (CN), TIR-NBS (TN) and NBS-LRR/ NBS (NL/N)] were also observed in linseed. These truncated proteins were dispersed among various clades and classes, and scattered throughout the phylogenetic tree **(Figure 5.2)**. It is possible that some of these genes were misassembled or some additional unique domains are yet to be identified. Such truncated *R* genes are suggested to act as adapter molecules in the resistance response through recruitment or interaction with NBS-LRR genes (Belkhadir et al. 2004). Similarly, NBS proteins lacking an LRR were also identified in *Arabidopsis* and *Medicago* (Ameline-Torregrosa et al. 2008; Meyers et al. 2003). Genes containing only the NBS domains were also observed, which are reported to be either absent or reduced in other dicot genomes (Bai et al. 2000). However, three such truncated genes (g1978, g16484 and g27371); interestingly showed expression evidence based on the EST data of linseed **(Table S6.7)**. Furthermore, the unusual TIR-NBS-containing genes (TNCL and TCN) similar to those previously reported in *Arabidopsis* (Meyers et al. 2003) and *Medicago* (Ameline-Torregrosa et al. 2008) were also observed in linseed. In addition, few other domains were also identified, indicating that the gain of these domains has occurred in this family of proteins in the linseed lineage **(Table S6.1)**.

6.4.2 Origin of the linseed NBS-LRR genes

The coalescent points on the phylogenetic tree provide relative time references (Ameline-Torregrosa et al. 2008)**,** indicating the clades that have probably expanded in linseed. The linseed *R* proteins were clustered with similar proteins from poplar, a closely related species; suggesting that they evolved from a common ancestor and later expanded separately within each species **(Figure 6.3)**. The TNL family proteins show recent expansion compared to the CNL family proteins as evident from the tree. As indicated by the coalescent points, almost all the clades from TNL family expanded at similar times, whereas variation in the time scale was observed in the expansion of CNL family genes. When phylogenetic analysis of linseed *R* proteins was carried out with the curetted *R* proteins having NBS domains (http://www.prgdb.org), various clade proteins clustered with many curetted *R* genes. Interestingly, proteins from CNL-1 and TNL-4 clades did not cluster with any characterized *R* proteins and represented 13.5% of the predicted linseed *R* proteins. This indicates the unique evolutionary history and function of these proteins. The

TNL-4 genes showed paraphyletic origin **(Figure 6.2)** and their characterization might reveal some interesting and new NBS-LRR genes from linseed. All the members of CNL-2 clade clustered mostly with *R* proteins from various plant families, whereas CNL-3 clade clustered with mainly resistant protein RPM1 (*Arabidopsis thaliana*), indicating that these proteins are ancient in origin. Within the TNL family, the TNL-2 and 5 clades clustered with the known linseed proteins P2 and M and L6, respectively. Thus, they are the closest homologues of these characterized genes and probably represent linseed specific NBS-LRR genes. Interestingly, TNL-3 clustered with *R* proteins characterized mainly from the *Solanaceae* family.

6.4.3 Analysis of gene expression and promoter regions

The quantitative expression analysis showed inducible expression of the RGAs. The gene model g2659 showed the highest expression among the five gene models evaluated at 7 and 10 DAI and both control and infected tissue. This might be due to constitutive plus inducible expression of this gene. Constitutive expression of RGAs was also reported in soybean (Graham et al. 2002). The same gene also showed higher expression in control tissue, which further supports the above hypothesis. Analysis of promoter regions of the linseed NBS-LRR genes revealed uniformity in the numbers of four overrepresented cis elements (WBOX, CBF, DRE and GCC boxes) across all the clades studied in both TNL and CNL families. In *Arabidopsis*, tandemly duplicated genes show higher levels of conservation of *cis* elements compared to segmentally duplicated genes (Haberer et al. 2004). The same observation was also reported in *Medicago* (Ameline-Torregrosa et al. 2008). All the linseed *R* genes contained the WBOX domain and there was little difference in the frequency of occurrence of this motif between the two families (TNL and CNL). WBOX motifs are present in the NPR1 gene (Yu and Goh 2001) and most of the pathogen response genes in *Arabidopsis* (Li et al. 2004). It has also been shown that they activate NBS-LRR genes in *Arabidopsis* and grape (Marchive et al. 2007; Zheng et al. 2007). The conservation of WBOX motif indicates its importance in regulation of NBS-LRR gene family. However, it does not mean identical regulatory mechanism and other less conserved factors might be involved in fine regulation of these genes.

Based on the EST data **(Table S6.7)**, 23% of the linseed NBS-LRR genes provided expression evidence, with the peeled stem library exhibiting the highest number of expressed genes. Such low expression of NBS-LRR genes has also been reported in poplar (Kohler et al. 2008) and it has been suggested that these genes are either expressed at a very low levels or they are induced only during specific conditions in specific tissues. Hence, they might go undetected or escape during library preparation. Further, the expression patterns of these linseed genes could not be correlated with the presence of any of the regulatory elements or domain arrangements. However, the gene g2659 with high EST expression contained an extra F box like domain that is known to be involved in signal transduction and protein-protein interaction. However, no correlation between gene expression and promoter sequence was observed. An in-depth analysis of such genes could reveal finer details of the disease resistance mechanism.

In summary, total 147 NBS-LRR genes were identified from the linseed genome having limited domain novelty, with almost all the novelty existing in the TNL subfamily, and most of that within the type D proteins. Analysis of domain structure, promoter regions, phylogeny and *in silico* expression revealed typical features of the NBS-LRR genes as reported from other plant species. Further characterization of these genes might help in understanding disease resistance mechanism in linseed.

CHAPTER 7

Summary and future directions

Linseed has a potential to emerge as a crop for food, feed and fiber. However, the genetic and genomic resources are very limited in this crop as compared to other oilseed crops. The present research was an attempt to develop few such resources for linseed which could linseed breeding efforts for the development of varieties with biotic and abiotic stress resistance, high and better quality oil and better nutraceutical.

The major objectives for linseed geneticists and breeders in near future would be: (i) to develop more advanced markers, (ii) to align existing genetic and physical maps in order to generate a consensus map with a standardized nomenclature and (iii) to compile and integrate relevant morphological and agronomic data with allelic information for linseed germplasm, in order to enable the development of association mapping techniques for the exploitation of available genetic resources outside the narrow linseed gene pool.

7.1 Development of genomic and EST-SSRs

Availability of large number of polymorphic SSRs is the basic requirement of molecular plant breeding. Lack of such markers in linseed lead to initiation of this thesis work on the development of SSR markers for linseed and total 290 genomic and 927 EST-SSRs were identified. To develop genomic SSRs, three widely used methods were exploited to obtain SSR enriched amplicons. A small modification in existing protocols was done wherein construction of genomic libraries was bypassed and the pooled amplicons generated by these methods were directly sequenced using the 454 GS-FLX next generation sequencing platform that resulted in development of large number of SSRs within sufficiently short time. While, the EST-SSRs were developed from publically available linseed ESTs using bioinformatics approach. Further, comparative analysis of the EST-SSR motifs distribution among nine other related species was also studied, which showed that, TNRs are predominantly abundant in linseed ESTs. To check the practical utility of developed SSRs, 51 genomic SSRs and 30 EST-SSRs were screened for polymorphism and respectively, 11 and seven polymorphic SSRs were identified.

7.2 Development of core collection of linseed and genetic diversity and population structure analysis using SSRs

In the present study, a core collection (CC) of 222 accessions was developed from a collection of 2,239 germplasm accessions using phenotypic data of eight quantitative characters. Further, the representativeness and homogeneity of developed CC core collection was evaluated using various statistical methods *viz.;* z-test, chi-square test, %MD, %CR etc. and results proved its homogeneity and representativeness. Moreover, genetic diversity and population structure within CC was analyzed using SSR markers. Total 78 alleles were obtained with 29 polymorphic SSRs. The number of allele ranged from 2 to 6 with average alleles per marker as 2.5 showing less diversity within core collection. Structure analysis divided the core collection into two sub-populations, which was confirmed by neighbor-joining and PCoA study. However, very low *F*st value showed weak substructure.

7.3 SNP development and genome-wide association study (GWAS)

In this study, 95 diverse accessions were selected from CC based on phenotypic and genotypic diversity analysis and a highly advanced Genotyping By Sequencing (GBS) technique was used to develop more than 50, 000 SNPs. Out of that, 13, 280 SNPs having MAF value ≥ 0.1 were used for genetic diversity, population structure and preliminary genome-wide association study. The distribution of these SNPs was also studied and 15% SNPs were found to be present in exonic region. Population structure analysis using SNP also divided the accessions into 2 sub-populations, which was confirmed by neighbor-joining and PCoA analysis. The *F*st value was very low (0.006) again suggesting weak substructure making the panel ideal for GWAS. Various models were tested to avoid spurious association and K model (Mix linear model with Kinship matrix) was used for association study with phenotypic data of 8 characters. Preliminary, GWAS identified 24 significant SNPs, 16 for trait capsule per plants (CPP), 7 for trait technical plant height (TPH) and one SNP for trait plant height (PH).

7.4 Identification and characterization of NBS-LRR genes in linseed

This study was undertaken to identify the NBS-LRR *R* genes using the draft genome sequence of linseed. We identified 147 NBS-LRR genes from the linseed genome having limited domain novelty, with almost all the novelty existing in the TNL subfamily, and most of that within the type D proteins. Interestingly, only a few genes were linseed specific and further characterization of these might identify novel resistance genes in linseed. Analysis of domain structure, promoter regions, phylogeny and *in silico* expression revealed typical features of the NBS-LRR genes as reported from other plant species. However, some features were distinctive to linseed, like: (i) most of the discovered linseed NBS genes are probably of ancient origin, (ii) only a limited variation in the domain arrangements was observed; although as reported earlier, TNL family showed comparatively higher variation, (iii) uniformity of promoter regions across the gene family was detected, and (iv) many linseed specific genes from the CNL and TNL families were identified. Further analysis of the complete genome sequence with accurate annotation and manual verification would provide more insight into the evolution of NBS-encoding genes in linseed.

7.5 Future directions

7.5.1 Development and use of DNA markers

The present study developed various genetic resources which can be utilized for genetic improvement of linseed. A large number of genomic and EST-SSRs were developed. Polymorphism screening of these SSRs will develop more polymorphic markers for linseed, which can be used for various applications like marker assisted breeding, germplasm conservation etc. Also, the transferability study will help to develop SSRs for orphan species.

7.5.2 Use of core collection for association mapping

The core collection developed in this study can have wide applications in linseed breeding. We show that most of the diversity in the linseed germplasm was manifested in the exotic germplasm and landraces, and hence they need to be utilized to develop improved linseed varieties. The analysis of genotypic and phenotypic diversity based on genetic structure would facilitate parent selection for broadening the genetic base of linseed cultivars via breeding. In fact, we already identified genetically diverse genotypes with high trait values as potential parents for seven economically important traits. It would be interesting to see if these genotypes, when crossed, indeed produce improved varieties with better yield and quality. The information of genetic and phenotypic differentiation could also be helpful for association mapping of genes of interest. Few accessions from each cluster can be used to form a panel for genome wide association studies using next generation sequencing technology, which in combination with phenotypic data; will help to identify the promising QTLs for various traits of agronomic importance. The strategies discussed in this study could also be applicable to other crops.

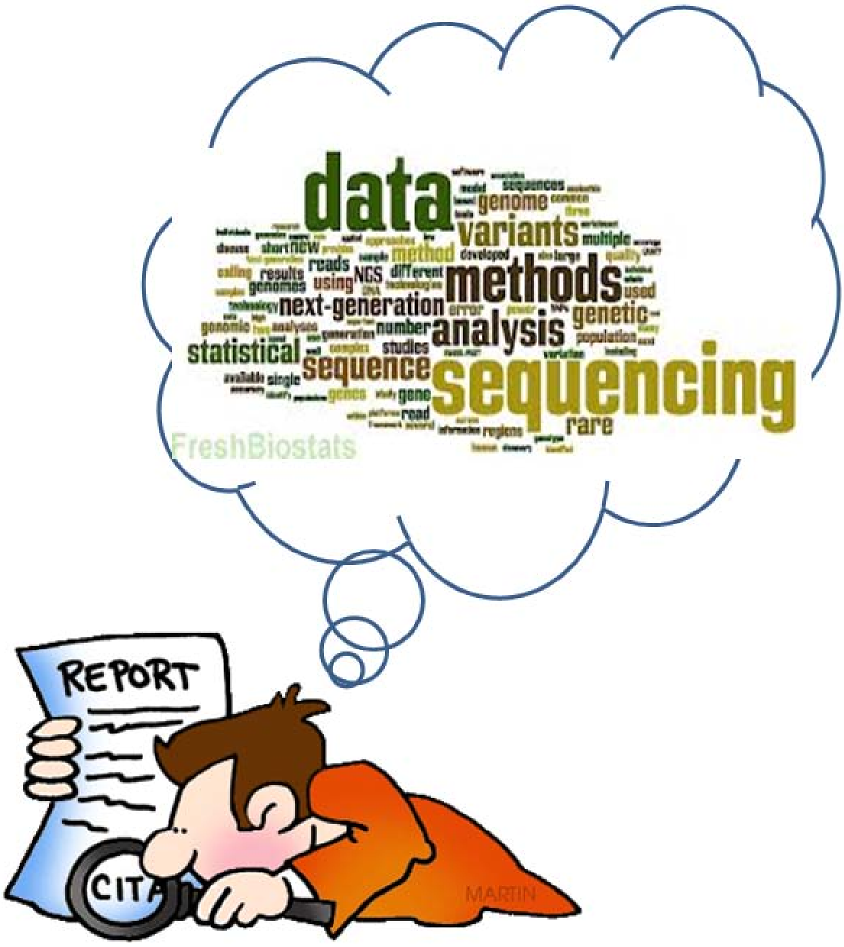
The GBS of remaining accessions from core collections will help to identify more SNPs for linseed. Those SNP present in the exonic region can have direct application in marker assisted selection and need further characterization. Further, the multi-location phenotypic data will help to identify stable SNPs for the trait studied. Moreover, the phenotypic data for more agro-economically important traits and from multiple locations combined with SNP data would be helpful to identify SNPs for those traits.

7.5.3 Identification and characterization of disease resistance genes

The NBS-LRR is the predominant class of R genes present in plants. The present analysis identify many linseed specific NBS-LRR genes, further characterization of these genes will help to understand the disease resistance mechanisms in linseed and to generate novel disease resistance specificities in this nutritionally important crop. Additionally, these genes can be employed to develop disease resistance varieties of linseed by molecular breeding.

In the view of input requirements for linseed breeding amalgamation of such high throughput technologies along with traditional breeding practices are of utmost importance in linseed. Many more genome-wide studies such as identification and characterization of other gene families, synteny mapping with other closely related plant species, evolution of gene families in the background of linseed genome instability need to be performed. Thus, the present work is a step towards such genome-wide studies in linseed.

BIBLIOGRAPHY



Bibliography

Aggarwal R, K, Prasad S, H, Varshney R, K, Prasanna R, B, Krishnakumar V, Lalji S (2007) Identification, characterization and utilization of EST-derived genic microsatellite markers for genome analyses of coffee and related species. Theoretical and Applied Genetics **114**:359–372

Agrama HA, Eizenga GC, Yan W (2007) Association mapping of yield and its components in rice cultivars. Molecular Breeding **19**:341-356

Ahlawat IPS (2008) Linseed. In: Ahlawat IPS (ed) Agronomy – Rabi Crops

Albertini E, Torricelli R, Bitocchi E, Raggi L, Marconi G, Pollastri L, . . . Veronesi F (2011) Structure of genetic diversity in *Olea europaea* L. cultivars from central Italy. Molecular Breeding **27**:533-547

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

Allen AM, Barker GL, Berry ST, Coghill JA, Gwilliam R, Kirby S, . . . Edwards KJ (2011) Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (*Triticum aestivum* L.). Plant Biotechnology Journal **9**:1086–1099

Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, Lander ES (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature **407**:513–516

Ameline-Torregrosa C, Wang BB, O'Bleness MS, Deshpande S, Zhu HY, Roe B, . . . Cannon SB (2008) Identification and characterization of nucleotide-binding site-leucine-rich repeat genes in the model plant *Medicago truncatula*. Plant Physiology **146**:5-21

Antolovich M, Prenzler P, Robards K, Ryan D (2000) Sample preparation in the determination of phenolic compounds in fruits. Analyst **125**:989–1009

Babrowski KJ, Sosulski FW (1984) Composition of free and hydrolyzable phenolic acids in defatted flours of ten oilseeds. Journal of Agricultural and Food Chemistry 128-130

Bai JF, Choi SH, Ponciano G, Leung H, Leach JE (2000) *Xanthomonas oryzae* pv. *oryzae* avirulence genes contribute differently and specifically to pathogen aggressiveness. Molecular Plant-Microbe Interactions **13**:1322-1329

Bai JF, Pennill LA, Ning JC, Lee SW, Ramalingam J, Webb CA, . . . Hulbert SH (2002) Diversity in nucleotide binding site-leucine-rich repeat genes in cereals. Genome Research **12**:1871-1884

Bailey TL, Gribskov M (1998) Combining evidence using p-values: application to sequence homology searches. Bioinformatics **14**:48-54

Bailey TL, Williams N, Misleh C, Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Research **34**:369-373

Bakke JE, Klosterman HJ (1956) A new diglucoside from flaxseed. Proceedings of North Dakota Academy of Science pp 18–22

Bambagiotti-alberti M, Coran SA, Ghiara C, Moneti G, Raffaelli A (1994) Investigation of mammalian lignan precursors in flax seed: first evidence of secoisolariciresinol diglucoside in two isomeric forms by liquid chromatography/mass spectrometry. Rapid Communications in Mass Spectrometry **8**: 929-932

Barchi L, Lanteri S, Portis E, Acquadro A, Vale G, Toppino L, Rotino G (2011) Identification of SNP and SSR markers in eggplant using RAD tag sequencing. BMC Genomics **12**:304

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**

Barvkar VT, Pardeshi VC, Kale SM, Qiu S, Rollins M, Datla R, . . . Kadoo NY (2013) Genome-wide identification and characterization of microRNA genes and their targets in flax (Linum usitatissimum): Characterization of flax miRNA genes. Planta **237**:1149-1161

Bassam BJ, Caetanoanolles G, Gresshoff PM (1991) Fast and sensitive silver staining of DNA in polyacrylamide gels. Analytical Biochemistry **196**:80-83

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, . . . Sonnhammer ELL (2002) The Pfam protein families database. Nucleic Acids Research **30**:276-280

Beejmohun V, Fliniaux O, Grand E, Lamblin F, Bensaddek L, Christen P, . . . Mesnard F (2007) Microwave-assisted extraction of the main phenolic compounds in flaxseed. Phytochemical Analysis **18**:275–282

Belkhadir Y, Subramaniam R, Dangl JL (2004) Plant disease resistance protein signaling: NBS-LRR proteins and their partners. Current Opinions in Plant Biology **7**:391-399

BeMiller JN (1973) Quince seed, psyllium seed, flax seed and okra gums. In: Whistler RLB, J. N. (ed) Industrial Gums. Academic Press, New York, pp 331-337

Bennett MD (2005) Nuclear DNA Amounts in Angiosperms: Progress, Problems and Prospects. Annals of Botany **95**:45-90

Berger J, Suzuki T, Senti KA, Stubbs J, Schaffner G, Dickson BJ (2001) Genetic mapping with SNP markers in Drosophila Nature Genetics **29**:475–481

Bhatty RS (1995) Nutritional composition of whole flaxseed and flaxseed meal. In: Cunnane SC, Thompson LH (eds) Flaxseed in HumanNutrition AOCS Press, Champaign, IL, pp 22–45

Bhatty RS, Cherdkiagumchai 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 CL, Gadani S, Lukacs M, Cullis CA (2011) SSR markers developed for genetic mapping in flax (*Linum usitatissimum L.*). Research and Reports in Biology **2**:23–29

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

Brachi B, Morris GP, Borevitz JO (2011) Genome-wide association studies in plants: the missing heritability is in the field. Genome Biology **12**:232

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics **23**:2633-2635

Breseghello F, Sorrells ME (2006) Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. Genetics **172**:1165-1177

Brown AHD (1989) Core collections - a practical approach to genetic resources management. Genome **31**:818-824

Brown F (1953) The tocopherol content of farm–feeding stuffs. Journal Science Food Agricultural **4**:161

Bus A, Hecht J, Huettel B, Reinhardt R, Stich B (2012) High-throughput polymorphism detection and genotyping in *Brassica napus* using nextgeneration RAD sequencing. BMC Genomics **13**:281

Cappa EP, El-Kassaby YA, Garcia MN, Acuña C, Borralho NMG, Grattapaglia D, Marcucci Poltri SN (2013) Impacts of Population Structure and Analytical Models in Genome-Wide Association Studies of Complex Traits in Forest Trees: A Case Study in Eucalyptus globulus. PloS One

Cardle L, Ramsay L, Milbourne D, Macaulay M, Marshall D, Waugh R (2000) Computational and experimental characterization of physically clustered simple sequence repeats in plants. Genetics **156**:847–854

Cardoso carraro JC, Inês de souza dantas M, Rocha espeschit AC, Duarte martino HS, Rocha ribeiro SM (2012) Flaxseed and Human Health: Reviewing Benefits and Adverse Effects. Food Reviews International **28**:203-230

Care AD (1954) Goitrogenic properties of iinseed. Nature **173**:172-173

Carlos daMaia L, Dario Abel Palmieri DA, Velci Queiroz de Souza VQ, Marini Kopp M, F´ elix de Carvalho FI, de Oliveira AC (2008) SSR Locator: Tool for Simple Sequence Repeat Discovery Integrated with Primer Design and PCR Simulation. International Journal of Plant Genomics

Cavagnaro FP, Senalik DA, Yang L, Simon PW, Harkins TT, Kodira CD, . . . Weng Y (2010) Genome-wide characterization of simple sequence repeats in cucumber (Cucumis sativus L.). BMC Genomics **11**

Charmet G, Balfourier F (1995) The use of geostatistics for sampling a core collection of perennial ryegrass populations. Genetic Resources and Crop Evolution **42**:303-309

Cho YG, Ishii T, Temnykh S, Chen X, Lipovich L, McCouch SR, . . . Cartinhour S (2000) Diversity of microsatellites derived from genomic libraries and GenBank sequences in rice (*Oryza sativa L.*). Theoretical and Applied Genetics **100**:713-722

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, . . . Ruden DM (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w(1118); iso-2; iso-3. Fly **6**:80-92

Cloutier S, Niu ZX, 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, Niu Z, 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

Cloutier S, Miranda E, Ward K, Radovanovic N, Reimer E, Walichnowski A, . . . Ragupathy R (2012a) Simple sequence repeat marker development from bacterial artificial chromosome end sequences and expressed sequence tags of flax (*Linum usitatissimum L.*). Theoretical and Applied Genetics **125**:685–694

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

Cockram J, White J, Zuluaga DL, Smith D, Comadran J, Macaulay M, . . . Consortium A (2010) Genome-wide association mapping to candidate polymorphism resolution in the unsequenced barley genome. Proceedings of the National Academy of Sciences of the United States of America **107**:21611-21616

Collard B, C,Y, Jahufer M, Z, Brouwer J, B, Pang E, C, K (2005) An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: the basic concepts. Euphytica **142**:169–196

Cordeiro GM, Casu R, McIntyre CL, Manners JM, Henry RJ (2001) Microsatellite markers from sugarcane (*Saccharum spp.*) ESTs cross transferable to erianthus and sorghum. Plant Science **160**:1115-1123

Cortés A, Chavarro M, Blair M (2011) SNP marker diversity in common bean (*Phaseolus vulgaris* L.) Theoriotical and Applied Genetics **123**:827–845

Cullis CA (1973) DNA Differences between Flax Genotrophs. Nature **243**:515-516

Cullis CA, Oh TJ, Gorman MB (1995) Genetic mapping in flax (Linum usitatissimum). Breeding for Fiber and Oil Quality in Flax, Valery en Caux, France: Centre Technique pour l'Etude et l'Amelioration du Lin pp 161–169

Cunnane SC, Ganguli S, Menard C, Liede AC, Hamadeh MJ, Chen ZY, . . . Jenkins DJA (1993) High α –linolenic acid flaxseed (*Linum usitatissimum*): some nutritional properties in humans. British Journal of Nutrition **49**:443–453

da Maia LC, de Souza VQ, Kopp MM, de Carvalho FIF, de Oliveira AC (2009) Tandem repeat distribution of gene transcripts in three plant families. Genetics and Molecular Biology **32**:822-833

Dangl JL, Jones JDG (2001) Plant pathogens and integrated defence responses to infection. Nature **411**:826-833

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 S, He D, Li X, Wang Y, Hao D, . . . Chen X (2011) Isolation and characterization of polymorphic microsatellite markers from flax (*Linum usitatissimum L.*). African journal of Biotechnology **10**:734–739

Deng X, Long SH, He DF, Li X, Wang YF, Liu J, Chen XB (2010) Development and characterization of polymorphic microsatellite markers in *Linum usitatissimum*. Journal of Plant Research **123**:119-123

Deschamps S, Llaca V, May GD (2012) Genotyping-by-Sequencing in Plants. Biology **1**:460 - 483

Diederichsen A (2007) Ex situ collections of cultivated flax (*Linum usitatissimum* L.) and other species of the genus Linum L. Genetic Resources and Crop Evolution **54**:661-678

Diederichsen A, Fu YB (2006) Phenotypic and molecular (RAPD) differentiation of four infraspecific groups of cultivated flax (*Linum usitatissimum L.* subsp usitatissimum). Genetic Resources and Crop Evolution **53**:77-90

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

Dodds PN, Lawrence GJ, Ellis JG (2001) Six amino acid changes confined to the leucine-rich repeat beta-strand/beta-turn motif determine the difference between the P and P2 rust resistance specificities in flax. Plant Cell **13**:163-178

Domier K, Kerr N (2000) The potential for agricultural fibers. Proceedings of the 58th Flax Institutes, pp 138–140

Dong JX, Chen CH, Chen ZX (2003) Expression profiles of the Arabidopsis WRKY gene superfamily during plant defense response. Plant Molecular Biology **51**:21-37

Ehrenreich IM, Hanzawa Y, Chou L, Roe JL, Kover PX, Purugganan MD (2009) Candidate gene association mapping of Arabidopsis flowering time. Genetics **183**:325-335

Ellis JG, Jones DA (2003) Plant Disease Resistance Genes. In: Ezekowitz RAB, Hoffmann JA (eds) *Infectious Disease: Innate Immunity*. Humana Press Inc.,, Totowa, NJ

Ellis JR, Burke JM (2007) EST-SSRs as a resource for population genetic analyses. Heredity **99**:125–132

Ellis PR, Pink DAC, Phelps K, Jukes PL, Breeds SE, Pinnegar AE (1998) Evaluation of a core collection of Brassica oleracea accessions for resistance to Brevicoryne brassicae, the cabbage aphid. Euphytica **103**:149-160

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS One **6**:e19379

Esteras C, Formisano G, Roig C, Diaz A, Blanca J, Garcia-Mas J, . . . Pico B (2013) SNP genotyping in melons: genetic variation, population structure, and linkage disequilibrium. Theoretical and Applied Genetics **126**:1285-1303

Eujayl I, Sorrells ME, Baum M, Wolters P, Powell W (2002) Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat. Theoretical and Applied Genetics **104**:399-407

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology **14**:2611-2620

Everaert I, De Riek J, De Loose M, Van Waes J, Van Bockstaele E (2001) Most similar variety grouping for distinctness evaluation of flax and linseed (*Linum usitatissimum L.*) varieties by means of AFLP and morphological data. Plant Varieties and Seeds **14**:69-87

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources **10**:564-567

Falgueras J, Lara AJ, Fernandez-Pozo N, Canton FR, Perez-Trabado G, Claros MG (2010) SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read. BMC Bioinformatics **11**

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics **164**:1567-1587

Falush D, Stephens M, Pritchard JK (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. Molecular Ecology Notes **7**:574-578

Fan JB, Oliphant A, Shen R, Kermani BG, Garcia F, Gunderson KL, . . . Chee MS (2003) Highly parallel SNP genotyping. Cold Spring Harbor Symposia on Quantitative Biology **68**:69–78

Fedorov AA (1974) Chromosome numbers of flowering plants: 412-414

Feltus FA, Wan J, Schulze SR, Estill JC, Jiang N, Paterson AH (2004) An SNP resource for rice genetics and breeding based on subspecies indica and japonica genome alignments. Genome Research **14**:1812–1819.

Fisher P, J , Gardner R, C, Richardson T, E (1996) Single Locus Microsatellites Isolated Using 5′ Anchored PCR. Nucleic Acids Research **24**:4369–4371

Flint-Garcia SA, Thuillet AC, Yu JM, Pressoir G, Romero SM, Mitchell SE, . . . Buckler ES (2005) Maize association population: a high-resolution platform for quantitative trait locus dissection. Plant Journal **44**:1054-1064

Francki MG, Mullan DJ (2004) Application of comparative genomics to narrow-leafed lupin (*Lupinus angustifolius* L.) using sequence information from soybean and Arabidopsis. Genome **47**:623-632

Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. Nature Review Genetics **10**:241-251

Fu Y, Peterson G, Diederichsen A, Richards KW (2002) RAPD analysis of genetic relationships of seven flax species in the genus *Linum* L. Genetic Resources and Crop Evolution **49**:253–259

Fu Y, Rowland GG, Duguid SD, Richards KW (2003a) RAPD analysis of 54 North American flax cultivars. Crop Science **43**:1510–1515

Fu YB (2002) Redundancy and distinctness in flax germplasm as revealed by RAPD. Plant Genetic Resources **4**:117–124

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, Guerin S, Peterson GW, Diederichsen A, Rowland GG, Richards KW (2003b) RAPD analysis of genetic variability of regenerated seeds in the Canadian flax cultivar CDC Normandy. Seed Science and Technology **31**:207-211

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

Gadaleta A, Mangini G, Mulè G, Blanco A (2006) Characterization of dinucleotide and trinucleotide EST-derived microsatellites in the wheat genome. Euphytica **153**:73-85

Ghosh R, Paula S, Kumar S, a G, Roy A (2009) An improved method of DNA isolation suitable for PCR-based detection of begomoviruses from jute and other mucilaginous plants. Journal of Virological Methods **159,**:34-39

Gibson G (2010) Hints of hidden heritability in GWAS. Nature Genetics **42**:558-560

Gill KS (1987) Linseed. In: division Pai (ed). Incdian Council of Agriculture Research, New Delhi, p 386

Gill KS, Yermanos DM (1967) Cytogenetic studies on the genus Linum. Crop Science **7**: 623–631

Gimenes MA, Hoshino AA, Barbosa AVG, Palmieri DA, Lopes CR (2007) Characterization and transferability of microsatellite markers of the cultivated peanut (*Arachis hypogaea*). BMC Plant Biology **7**:9

Ginestet C (2011) ggplot2: Elegant graphics for data analysis. Journal of Royal Statatistic Society of America **174**:245-245

Goldberg SMD, Johnson J, Busam D, Feldblyum T, Ferriera S, Friedman R, . . . Venter JC (2006) A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. Proceedings of the National Academy of Sciences of the United States of America **103**:16057-16057

Gong YM, Xu SC, Mao WH, Hu QZ, Zhang GW, Ding J, Li YD (2010) Developing new SSR markers from ESTs of pea (*Pisum sativum L.*). Journal of Zhejiang University-Science B **11**:702-707

Gopalan C, Ramasastri BV, Subramanian SC (2007) Nutritive Value of Indian Foods. National Institute of Nutrition, Hyderabad, India

Gorman M, Parojcic M (1992) Genomic mapping in flax (*Linum usitatissimum*) Plant Genome I Conference, Town & Country Conference Center, SanDiego, CA.

Goudet J (2001) FSTAT- a program to estimate and test gene diversities and fixation indices (version 2.9.3). Available from http://www.unil.ch/izea/softwares/fstat.html Updated from Goudet (1995)

Graham MA, Marek LF, Shoemaker RC (2002) Organization, Expression and Evolution of a Disease Resistance Gene Cluster in Soybean. Genetics **162**:1961-1977

Gunderson KL (2009) Whole-genome genotyping on bead arrays. Methods in Molecular Biology **529**:197–213

Gupta PK, Rustgi S, Kulwal PL (2005) Linkage disequilibrium and association studies in higher plants: Present status and future prospects. Plant Molecular Biology **57**:461-485

Gupta PK, Rustgi S, Mir RR (2008) Array-based high-throughput DNA markers for crop improvement. Heredity **101**:5-18

Gupta PK, Rustgi S, Sharma S, Singh R, Kumar N, Balyan HS (2003) Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. Molecular Genetics and Genomics **270**:315-323

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

He GH, Meng RH, Newman M, Gao GQ, Pittman RN, Prakash CS (2003) Microsatellites as DNA markers in cultivated peanut (*Arachis hypogea*). BMC Plant Biology **3**:8

Hiremath PJ, Farmer A, Cannon SB, Woodward J, Kudapa H, Tuteja R, . . . Varshney RK (2011) Large-scale transcriptome analysis in chickpea (*Cicer arietinum L.*), an orphan legume crop of the semi-arid tropics of Asia and Africa. Plant Biotechnology Journal **9**:922-931

Hjelmquist H (1950) The flax weeds and the origin of cultivated flax. Botaniska Notiser 257–298

Holbrook CC, Timper P, Xue HQ (2000) Evaluation of the core collection approach for identifying resistance to Meloidogyne arenaria in peanut. Crop Science **40**:1172-1175

Hu J, Zhu J, Xu HM (2000) Methods of constructing core collections by stepwise clustering with three sampling strategies based on the genotypic values of crops. Theoretical and Applied Genetics **101**:264-268

Huang XH, Wei XH, Sang T, Zhao QA, Feng Q, Zhao Y, . . . Han B (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nature Genetics **42**:961-976

Huang XQ, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Research **9**:868-877

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**

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

Iniguez-Luy FL, Voort AV, Osborn TC (2008) Development of a set of public SSR markers derived from genomic sequence of a rapid cycling Brassica oleracea L. genotype. Theoretical and Applied Genetics **117**:977-985

Jang CS, Kamps TL, Skinner DN, Schulze SR, Vencill WK, Paterson AH (2006) Functional classification, genomic organization, putatively cis-acting regulatory elements, and relationship to quantitative trait loci, of sorghum genes with rhizome-enriched expression. Plant Physiology **142**:1148-1159

Jiang D, Zhong GY, Hong QB (2006) Analysis of microsatellites in Citrus Unigenes. Acta Genetica Sinica **33**:345– 353

Jones E, Chu W-C, Ayele M, Ho J, Bruggeman E, Yourstone K, . . . Smith SM (2009) Development of single nucleotide polymorphism (SNP) markers for use in commercial maize (Zea mays L.) germplasm. Molecular Breeding **24**:165–176

Joshi SP, Ranjekar PK, Gupta VS (1999) Molecular markers in plant genome analysis. Current Science **77**:230-240

Judd A (1995) Flax-Some historical consideration. In: Cunnance SC, Thompson LU (eds) Flaxseed in human nutrition. AOCS press, Champaign, IL, pp 1-10

Jung S, Abbott A, Jesudurai C, Tomkins J, Main D (2005) Frequency, type, distribution and annotation of simple sequence repeats in Rosaceae ESTs. Functional & integrative genomics **5**:136-143.

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E (2008) Efficient control of population structure in model organism association mapping. Genetics **178**:1709-1723

Katti MV, Ranjekar PK, Gupta VS (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. Molecular Biology and Evolution **18**:1161-1167

Kaur S, Cogan NO, Pembleton LW, Shinozuka M, Savin KW, Materne M, Forster JW (2011) Transcriptome sequencing of lentil based on second-generation technology permits large-scale unigene assembly and SSR marker discovery. BMC Genomics **12**

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

Kohler A, Rinaldi C, Duplessis S, Baucher M, Geelen D, Duchaussoy F, . . . Martin F (2008) Genome-wide identification of NBS resistance genes in *Populus trichocarpa*. Plant Molecular Biology **66**:619-636

Koledzeijczyk PP, Fedec P (1995) Processing flaxseed for human consumption In: Cunnance SC, Thompson LU (eds) Flaxseed in human nutrition AOCS press, Champaign, IL, pp 261-280

KÖlliker R, Jones ES, Drayton MC, Dupal MP, Forster JW (2001) Development and characterization of simple sequence repeat (SSR) markers for white clover (*Trifolium repens* L.). Theoretical and Applied Genetics **102**:8

Kozlowska H, Zadernowski R, Sosulski FW (1983) Phenolic acids in oilseed flour. Nahrung 449-453

Kuang H, Woo SS, Meyers BC, Nevo E, Michelmore RW (2004) Multiple genetic processes result in heterogeneous rates of evolution within the major cluster disease resistance genes in lettuce. Plant Cell **16**:2870-2894

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

Kumpatla SP, Mukhopadhyay S (2005) Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. Genome **48**:985-998

Kvavadze E, Bar-Yosef O, Belfer-Cohen A, Boaretto E, Jakeli N, Matskevich Z, Meshveliani T (2009) 30,000-year-old wild flax fibers. Science 325:1359

La Rota M, Kantety RV, Yu JK, Sorrells ME (2007) Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. BMC Genomics **6**

Lawrence GJ, Finnegan EJ, Ayliffe MA, Ellis JG (1995) The L6 gene for flax rust resistance is related to the Arabidopsis bacterial-resistance gene Rps2 and the tobacco viral resistance gene-N. Plant Cell **7**:1195-1206

Lawson MJ, Zhang LQ (2006) Distinct patterns of SSR distribution in the Arabidopsis thaliana and rice genomes. Genome Biology **7**

Leister RT, Katagiri F (2000) A resistance gene product of the nucleotide binding site - leucine rich repeats class can form a complex with bacterial avirulence proteins *in vivo*. Plant Journal **22**:345-354

Leskanich CO, Noble RC (1997) Manipulation of the n-3 polysaturated fatty acid composition of avian eggs and meat. World’s Poulty Science **15**:183 –189

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, . . . Proc GPD (2009a) The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**:2078-2079

Li Y, Shi YS, Cao YS, Wang TY (2004a) Establishment of a core collection for maize germplasm preserved in Chinese National Genebank using geographic distribution and characterization data. Genetic Resources and Crop Evolution **51**:845-852

Li Y, Shi YS, Cao YS, Wang TY (2004b) Establishment of a core collection for maize germplasm preserved in Chinese National Genebank using geographic distribution and characterization data. Genetic Resources and Crop Evolution **51**:845-852

Li YH, Zhang C, Gao ZS, Smulders MJM, Ma ZL, Liu ZX, . . . Qiu LJ (2009b) Development of SNP markers and haplotype analysis of the candidate gene for *rhg1*, which confers resistance to soybean cyst nematode in soybean. Molecular Breeding **24**:63-76

Liewlaksaneeyanawin C, Ritland CE, El-Kassaby YA, Ritland K (2004) Single-copy, species-transferable microsatellite markers developed from loblolly pine ESTs. Theoretical and Applied Genetics **109**:361-369

Liu KJ, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics **21**:2128-2129

Lunt DH, Hutchinson WF, Carvalho GR (1999) An efficient method for PCR-based isolation of microsatellite arrays (PIMA). Molecular Ecology **8**:2

Lupas A, Vandyke M, Stock J (1991) Predicting coiled coils from protein sequences. Science **252**:1162-1164

Willfor SM, Smeds AI, Holmbom BR (2006) Chromatographic analysis of lignans.
Journal of Chromatography A **1112**:64–77

Mackay I, Powell W (2007) Methods for linkage disequilibrium mapping in crops. Trends in Plant Science **12**:57-63

Mahalakshmi V, Aparana P, Ramadevi S, Ortiz R (2002) Genomic sequence derived simple sequence repeat markers - Case study with *Medicago spp*. Electronic Journal of Biotechnology **5**:233–242

Maher B (2008) Personal genomes: The case of the missing heritability. Nature **456**:18-21

Malvar RA, Butron A, Alvarez A, Ordas B, Soengas P, Revilla P, Ordas A (2004) Evaluation of the European Union maize landrace core collection for resistance to Sesamia nonagrioides (*Lepidoptera : Noctuidae*) and Ostrinia nubilalis (*Lepidoptera : Crambidae*). Journal of Economic Entomology **97**:628-634

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, . . . Visscher PM (2009) Finding the missing heritability of complex diseases. Nature **461**:747-753

Mansby E, Diaz O, von Bothmer R (2000) Preliminary study of genetic diversity in Swedish flax (*Linum usitatissimum*). Genetic Resources and Crop Evolution **47**:417-424

Mantel NA (1967) The detection of disease clustering and a generalized regression approach. Cancer Research **27**:209 - 220

Marchenkov A, Rozhmina T, Uschapovsky I, Muir AD (2003) Cultivation of flax. In: Muir AD, Westcott ND (eds) Flax: the genus Linum. CRC, New York, pp 74-91

Marchive C, Mzid R, Deluc L, Barrieu F, Pirrello J, Gauthier A, . . . Lauvergeat V (2007) Isolation and characterization of a *Vitis vinifera* transcription factor, VvWRKY1, and its effect on responses to fungal pathogens in transgenic tobacco plants. Journal of Experimental Botany **58**:1999-2010

Mazur W, Fotsis T, Wähälä K, Ojala S, Salakka A, Adlercreutz H (1996) Isotope dilution gas chromatographic-mass spectrometric method for the determination of isoflavonoids, coumestrol, and lignans in food samples. Analytical Biochemistry **233**:169-180

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

McCouch SR, Chen X, Panaud O, Temnykh S, Y X, Y.G C, . . . Blair M (1997) Microsatellite marker development, mapping and applications in rice genetics and breeding. Plant Molecular Biology **35**:10

McDowell JM, Woffenden BJ (2003) Plant disease resistance genes: recent insights and potential applications. Trends in Biotechnology **21**:178-183

Meagher LP, Beecher GR, Flanagan VP, Li BW (1999) Isolation and characterization of the lignans, isolariciresinol and pinoresinol in flaxseed meal. Journal of Agricultural and Food Chemistry 3173-3180

Meyers BC, Dickerman AW, Michelmore RW, Sivaramakrishnan S, Sobral BW, Young ND (1999) Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily. Plant Journal **20**:317-332

Meyers BC, Kozik A, Griego A, Kuang HH, Michelmore RW (2003) Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis. Plant Cell **15**:1683-1683

Moccia MD, Oger-Desfeux C, Marais GAB, Widmer A (2009) A White Campion (Silene latifolia) floral expressed sequence tag (EST) library: annotation, EST-SSR characterization, transferability, and utility for comparative mapping. BMC Genomics **10**

Moffett P, Farnham G, Peart J, Baulcombe DC (2002) Interaction between domains of a plant NBS-LRR protein in disease resistance-related cell death. EMBO Journal **21**:4511-4519

Monosi B, Wisser RJ, Pennill L, Hulbert SH (2004) Full-genome analysis of resistance gene homologues in rice. Theoretical and Applied Genetics **109**:1434-1447

Moose SP, Mumm RH (2008) Molecular plant breeding as the foundation for 21st century crop improvement. Plant Physiology **147**:969-977

Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nature Genetics **30**:194-200

Morris DH (2003) Flax -A Health andNutrition Primer. Winnipeg, Manitoba, Flax Council of Canada

Muller T, Vingron M (2000) Modeling amino acid replacement. Journal of Computational Biology **7**:761-776

Mun JH, Yu HJ, Park S, Park BS (2009) Genome-wide identification of NBS-encoding resistance genes in *Brassica rapa*. Molecular Genetics and Genomics **282**:617-631

Muravenko OV, Lemesh VA, Samatadze TE, Amosova AV, Grushetskaya ZE, Popov KV, . . . Zelenin AV (2003) Genome comparisons with chromosomal and molecular markers for three closely related flax species and their hybrids. Russian Journal of Genetics **39:** 414-421

Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang ZW, Costich DE, Buckler ES (2009) Association mapping: critical considerations shift from genotyping to experimental design. Plant Cell **21**:2194-2202

Newcomb RD (2006) Analysis of expressed sequence tags from apple. Plant Physiology **141**:147-166

Newell MA, Cook D, Tinker NA, Jannink JL (2011) Population structure and linkage disequilibrium in oat (*Avena sativa* L.): implications for genome-wide association studies. Theoretical and Applied Genetics **122**:623-632

Nijman IJ, Kuipers S, Verheul M, Guryev V, Cuppen E (2008) A genome-wide SNP panel for mapping and association studies in the rat. BMC Genomics **9**

Nobuta K, Ashfield T, Kim S, Innes RW (2005) Diversification of non-TIR class NB-LRR genes in relation to whole-genome duplication events in Arabidopsis. Molecular Plant-Microbe Interactions **18**:103-109

Oh TJ, Gorman M, Cullis CA (2000) RFLP and RAPD mapping in flax (*Linum usitatissimum*). Theoretical and Applied Genetics **101**:590-593

Ohme-Takagi M, Suzuki K, Shinshi H (2000) Regulation of ethylene-induced transcription of defense genes. Plant and Cell Physiology **41**:1187-1192

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

Oomah BD, Kenaschuk EO, Mazza G (1995) Phenolic acids in flaxseed. Journal of Agriculture and Food Chemistry 2016-2019

Palomino C, Satovic Z, Cubero JI, Torres AM (2006) Identification and characterization of NBS-LRR class resistance gene analogs in faba bean (*Vicia faba* L.) and chickpea (*Cicer arietinum* L.). Genome **49**:1227-1237

Parida SK, Kumar KAR, Dalal V, Singh NK, Mohapatra T (2006) Unigene derived microsatellite markers for the cereal genomes. Theoretical and Applied Genetics **112**:808-817

Parker JE, Coleman MJ, Szabo V, Frost LN, Schmidt R, vanderBiezen EA, . . . Jones JDG (1997) The Arabidopsis downy mildew resistance gene RPP5 shares similarity to the toll and interleukin-1 receptors with N and L6. Plant Cell **9**:879-894

Pasam RK, Sharma R, Malosetti M, van Eeuwijk FA, Haseneyer G, Kilian B, Graner A (2012) Genome-wide association studies for agronomical traits in a world wide spring barley collection. BMC Plant Biology **12**:16

Pashley CH, Ellis JR, McCauley DE, Burke JM (2006) EST databases as a source for molecular markers: lessons from Helianthus. The Journal of heredity **97**:381-388

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, . . . Rokhsar DS (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature **457**:551-556

Peakall R, Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecology Notes **6**:288-295

Peeters JP, Martinelli JA (1989) Hierarchical Cluster-Analysis as a Tool to Manage Variation in Germplasm Collections. Theoretical and Applied Genetics **78**:42-48

Perrier X, Jacquemoud-Collet JP (2006) DARwin software. http://darwin.cirad.fr/

Petit HV (2002) Digestion, Milk production, Milk composition and blood composition of dairy cows fed whole flaxseed. Journal of Dairy Science **85**:1482–1490

Porter BW, Paidi M, Ming R, Alam M, Nishijima WT, Zhu YJ (2009) Genome-wide analysis of *Carica papaya* reveals a small NBS resistance gene family. Molecular Genetics and Genomics **281**:609-626

Powell W, Machray GC, Provan J (1996) Polymorphism revealed by simple sequence repeats. Trends in Plant Science

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. American Journal of Human Genetics **67**:170-181

Qiu S, Lu Z, Luyengi L, Lee SK, Pezzuto JM, Farnsworth NR, . . . Fong HS (1999) Isolation and characterization of flaxseed (*Linum usitatissimum*) constituents. Pharmaceutical Biology **37**:1–7

Rabello E, Nunes de Souza A, Saito D, Tsai S (2005) In silico characterization of microsatellites in *Eucalyptus* spp.: abundance, length variation and transposon associations. Genetics and Molecular Biology **28**:582–588

Rachinskaya OA, Lemesh VA, Muravenko OV, Yurkevich OY, Guzenko EV, Bol'sheva NL, . . . Zelenin AV (2011) Genetic polymorphism of flax (*Linum usitatissimum*) based on the use of molecular cytogenetic markers. Russian Journal of Genetics **47**:56–65

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**

Rairdan GJ, Moffett P (2006) Distinct domains in the ARC region of the potato resistance protein Rx mediate LRR binding and inhibition of activation. Plant Cell **18**:2082-2093

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

Rallo R, Dorado G, Martin A (2000) Development of simple sequence repeats (SSRs) in olive tree (*Olea europacea L.*). Theoretical and Applied Genetics **101**:984–989

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

Riju A, Rajesh MK, Sherin PTPF, Chandrasekar A, Apshara SE, Arunachalam V (2009) Mining of expressed sequence tag libraries of cacao for microsatellite markers using five computational tools. Journal of Genetics **88**:217-225

Rohlf JF (2006) NTSYS-pc 2.2. Numerical taxonomy and multivariate analysis system.Version 2.2. State university of New York, Stony Brook

Roorkiwal M, Sharma PC (2011) Mining functional microsatellites in legume unigenes. bioinformation **7**

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

Rotmistrovsky K, Jang W, Schuler GD (2004) A web server for performing electronic PCR. Nucleic Acids Research **32**:W108-W112

Roy JK, Smith KP, Muehlbauer GJ, Chao SM, Close TJ, Steffenson BJ (2010) Association mapping of spot blotch resistance in wild barley. Molecular Breeding **26**:243-256

Sakuma Y, Maruyama K, Osakabe Y, Qin F, Seki M, Shinozaki K, Yamaguchi-Shinozaki K (2006) Functional analysis of an Arabidopsis transcription factor, DREB2A, involved in drought-responsive gene expression. Plant Cell **18**:1292-1309

Sanguinetti CJ, Dias Neto E, Simpson AJG (1994) Rapid silver staining and recovery of PCR products separated on polyacrylamide gels. Biotechniques **17**:915-919

Santana Q, Coetzee M, Steenkamp E, Mlonyeni O, Hammond G, Wingfield M, Wingfield B (2009) Microsatellite discovery by deep sequencing of enriched genomic libraries. Biotechniques **46**:217-223

Schlueter JA, Dixon P, Granger C, Grant D, Clark L, Doyle JJ, Shoemaker RC (2004) Mining EST databases to resolve evolutionary events in major crop species. Genome **45**:868-876

Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics **18**:502-504

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

Schoen DJ, Brown AHD (1995) Maximising genetic diversity in core collections of wild relatives of crop species. In: Hodgkin T, Brown AHD, van Hintum TJL, Morales EAV (eds) Core collections genetic resources. John Wiley & Sons, Chichester, UK, pp 55-77

Scott KD, Eggler P, Seaton G, Rossetto M, Ablett EM, Lee LS, Henry RJ (2000) Analysis of SSRs derived from grape ESTs. Theoretical and Applied Genetics **100**:723-726

Shahidi F, Naczk M (2004) Biosynthesis, classification, and nomenclature of phenolics in food and nutraceuticals. Phenolics in Food and Nutraceuticals. CRC Press Boca Raton, FL, pp 1–16

Shannon CE, Weaver W (1949) A mathematical model of communication. University of Illinois Press, Urbana, IL

Sharopova N, McMullen MD, Schultz L, Schroeder S, Sanchez-Villeda H, Gardiner J, . . . Coe EH (2002) Development and mapping of SSR markers for maize. Plant Molecular Biology **48**:463-481

Sicilia T, Niemeyer HB, Honig DM, Metzler M (2003) Identification and stereochemical characterization of lignans in flaxseed and pumpkin seeds. Journal of Agricultural and Food Chemistry **51**:1181-1188

Singh A, Reimer S, Pozniak CJ, Clarke FR, Clarke JM, Knox RE, Singh AK (2009) Allelic variation at *Psy1-A1* and association with yellow pigment in durum wheat grain. Theoretical and Applied Genetics **118**:1539-1548

Singh NK, Dutta S, Kumawat G, Singh BP, Gupta DK, Singh S, . . . Cook DR (2011) Development of genic-SSR markers by deep transcriptome sequencing in pigeonpea [*Cajanus cajan (L.)* Millspaugh]. BMC Plant Biology **11**

Smith BD (1995) The emergence of Agriculture. Scientific American Library New York, pp 152-153

Smykal P, Bacova-Kerteszova N, Kalendar R, Corander J, Schulman AH, Pavelek M (2011) Genetic diversity of cultivated flax (*Linum usitatissimum*

L.) germplasm assessed by retrotransposon-based markers. Theoriotical and Applied Genetics **122**:1385–1397

Soto-Cerda BJ, Carrasco RA, Aravena GA, Urbina HA, Navarro CS (2011a) Identifying novel polymorphic microsatellites from cultivated flax (*Linum usitatissimum L.*) following data mining. Plant Molecular Biology Reporter

Soto-Cerda BJ, Diederichsen A, Ragupathy R, Cloutier S (2013a) Genetic characterization of a core collection of flax (*Linum usitatissimum* L.) suitable for association mapping studies and evidence of divergent selection between fiber and linseed types. BMC Plant Biology **13**:78

Soto-Cerda BJ, Duguid S, Booker H, Rowland G, Diederichsen A, Cloutier S (2013b) Genomic regions underlying agronomic traits in linseed (*Linum usitatissimum* L.) as revealed by association mapping. Journal of Integrative Plant Biology **(In press)**

Soto-Cerda BJ, Maureira-Butler I, Mun˜oz G, Rupayan A, Cloutier S (2012) SSR-based population structure, molecular diversity and linkage disequilibrium analysis of a collection of flax (*Linum usitatissimum* L.) varying for mucilage seed-coat content. Molecular Breeding **30**:875-888

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

Spagnoletti Zeuli PL, Qualset CO (1993) Evaluation of five strategies for obtaining a core subset from a large genetic resource collection of durum wheat. Theoretical and Applied Genetics **87**:295-304

Spielmeyer W, Green AG, Bittisnich D, Mendham N, Lagudah ES (1998) Identification of quantitative trait loci contributing to Fusarium wilt resistance on an AFLP linkage map of flax (*Linum usitatissimum*). Theoretical and Applied Genetics **97**:633-641

Squirrell J, Hollingsworth PM, Woodhead M, Russell J, Lowe AJ, Gibby M, Powell W (2003) How much effort is required to isolate nuclear microsatellites from plants? Molecular Ecology **12**:1339-1348

Stitt P (1990) Flax the ideal "designer food" ingredient 53rd Flax Institute pp 20–24

Storey JD (2002) A direct approach to false discovery rates. Journal of Royal Statistical Society B **64**:479-498

Stracke S, Haseneyer G, Veyrieras JB, Geiger HH, Sauer S, Graner A, Piepho HP (2009) Association mapping reveals gene action and interactions in the determination of flowering time in barley. Theoretical and Applied Genetics **118**:259-273

Tameling WIL, Elzinga SDJ, Darmin PS, Vossen JH, Takken FLW, Haring MA, Cornelissen BJC (2002) The tomato R gene products I-2 and Mi-1 are functional ATP binding proteins with ATPase activity. Plant Cell **14**:2929-2939

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

Tan X, Calderon-Villalobos LIA, Sharon M, Zheng CX, Robinson CV, Estelle M, Zheng N (2007) Mechanism of auxin perception by the TIR1 ubiquitin ligase. Nature **446**:640-645

Tang S-J, Liu Z-Z, Tang W-Q, Yang J-Q (2009) A simple method for isolation of microsatellites from the mudskipper (Boleophthalmus pectinirostris), without constructing a genomic library. Conservation Genetics **10**:1957-1959

Thiel T, Michalek W, Varshney RK, Graner A (2003) Exploiting EST databases for the development of cDNA derived microsatellite markers in barley (*Hordeum vulgare L.*). Theoretical and Applied Genetics **1**:411-422

Tóth G, Gáspári Z, Jurka J, (2000) Microsatellites in different eukaryotic genomes: survey and analysis. Genome Research **10**:967–981

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

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, . . . Rokhsar D (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science **313**:1596-1604

Tutin TG, Heywood VH, Burges NA, Moore DM, Valentine DH, Walters SM, Webb DA (1968) Flora Europaea Rosaceae to Umbelliferae

Upadhyaya HD, Ortiz R, Bramel PJ, Singh S (2003) Development of a groundnut core collection using taxonomical, geographical and morphological descriptors. Genetic Resources and Crop Evolution **50**:139-148

Uysal H, Fu YB, Kurt O, Peterson GW, Diederichsen A, Kusters P (2010) Genetic diversity of cultivated flax (*Linum usitatissimum L.*) and its wild progenitor pale flax (*Linum bienne* Mill.) as revealed by ISSR markers. Genetic Resources and Crop Evolution **57**:1109–1119

Vaisey-Genser M, Morris DH (2003) Introduction: History of the cultivation and uses of flaxseed. In: Muir AD, Westcott ND (eds) Flax, the Genus Linum, Taylor and Francis, London, pp 1–21.

Varshney RK, Graner A, Sorrells ME (2005) Genic microsatellite markers in plants: features and applications. Trends in Biotechnology **23**:48-55

Varshney RK, Thiel T, Stein N, Langridge P, Graner A (2002) In silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species. Cellular & Molecular Biology Letters **7**:537-546

Vavilov NI (1951) The origin, variation, immunity and breeding of cultivated plants. Chronica Botanica, pp 1-366

Venglat P, Xiang DQ, Qiu SQ, Stone SL, Tibiche C, Cram D, . . . Datla R (2011) Gene expression analysis of flax seed development. BMC Plant Biology **11**

Victoria FC, da Maia LC, de Oliveira AC (2011) *In silico* comparative analysis of SSR markers in plants. BMC Plant Biology **11**

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: Agriculture AIC-oRPfD (ed). Central Research Institute for Dryland Agriculture, Indian Council of Agricultural Research, Hyderabad, Hyderabad

Vloutoglou I, Fitt BDL, Lucas JA (1999) Infection of linseed by *Alternaria linicola*; effects of inoculum density, temperature, leaf wetness and light regime. European Journal of Plant Pathology **105**:585-595

Von Stackelberg MV, Rensing SA, Reski R (2006) Identification of genic moss SSR markers and a comparative analysis of twnty-four algal and plant gene indices reveal species-specific rather than group-specific characteristics of microsatellites. BMC Plant Biology **6**

Vromans J (2006) Molecular genetic studies in flax (*Linum usitatissimum L.*). Production Ecology and Resource Conservation. Wageningen University, The Netherlands, The Netherlands

Wang LQ, Meselhy MR, Li Y, Qin GW, Hattori M (2000) Human intestinal bacteria capable of transforming secoisolariciresinol diglucoside to mammalian lignans, enterodiol and enterolactone. Chemical and Pharmaceutical Bulletin **48**:606–610

Wang MH, Jiang N, Jia TY, Leach L, Cockram J, Waugh R, . . . Luo ZW (2012a) Genome-wide association mapping of agronomic and morphologic traits in highly structured populations of barley cultivars. Theoretical and Applied Genetics **124**:233-246

Wang YP, Gjuvsland AB, Vik JO, Smith NP, Hunter PJ, Omholt SW (2012b) Parameters in dynamic models of complex traits are containers of missing heritability. Plos Computational Biology **8**

Wang ZW, Hobson N, Galindo L, Zhu SL, Shi DH, McDill J, . . . Deyholos MK (2012c) The genome of flax (*Linum usitatissimum*) assembled *de novo* from short shotgun sequence reads. Plant Journal **72**:461-473

Ward JH (1963) Hierarchical Grouping to Optimize an Objective Function. Journal of American Statistical Association **58**:236-244

Weir BS, Cockerham CC (1984) Estimating F-Statistics for the Analysis of Population-Structure. Evolution **38**:1358-1370

Weizhong L, Adam G (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics **22**:1658–1659

Westcott ND, Muir AD (1996) Variation in the concentration of the flaxseed lignan concentration with variety, location and year. Procedings of the Flax Institute of the United States **56**:77-80

Wiesner I, Wiesnerova D, Tejklova E (2001) Effect of anchor and core sequence in microsatellite primers on flax fingerprinting patterns. Journal of Agricultural Science **137**:37-44

Wiesnerova D, Wiesner I (2004) ISSR-based clustering of cultivated flax germplasm is statistically correlated to thousand seed mass. Molecular Biotechnology **26**:207-214

Wilcoxon F (1945) Individual comparisons by ranking methods. Biometrics Bull **1**:80-83

Xiao SY, Ellwood S, Calis O, Patrick E, Li TX, Coleman M, Turner JG (2001) Broad-spectrum mildew resistance in *Arabidopsis thaliana* mediated by RPW8. Science **291**:118-120

Xu Y, Crouch JH (2008) Marker-Assisted Selection in Plant Breeding: From Publications to Practice. Crop Science **48**:391-407

Xue Y, Wang Y, Wu P, Wang Q, Yang L, Pan X, . . . Chen J (2012) A primary survey on Bryophyte species reveals two novel classes of Nucleotide-Binding Site (NBS) genes. PloS One **7**

Yaish MWF, de Miera LES, de la Vega MP (2004) Isolation of a family off resistance gene analogue sequences of the nucleotide binding site (NBS) type from *Lens* species. Genome **47**:650-659

Yang SH, Gu TT, Pan CY, Feng ZM, Ding J, Hang YY, . . . Tian DC (2008a) Genetic variation of NBS-LRR class resistance genes in rice lines. Theoretical and Applied Genetics **116**:165-177

Yang SH, Zhang XH, Yue JX, Tian DC, Chen JQ (2008b) Recent duplications dominate NBS-encoding gene expansion in two woody species. Molecular Genetics and Genomics **280**:187-198

Yu H, Goh CJ (2001) Molecular genetics of reproductive biology in orchids. Plant Physiology **127**:1390-1393

Yu J, Dixit A, Ma K-H, Chung J-W, Park Y-J (2009) A study on relative abundance, composition and length variation of microsatellites in 18 underutilized crop species. Genetic Resources and Crop Evolution **56**:237–246

Yu J, Wang J, Lin W, Li SG, Li H, Zhou J, . . . Yang HM (2005) The genomes of *Oryza sativa*: A history of duplications. PLoS Biology **3**:266-281

Zane L, Bargelloni L, Patarnello T (2002) Strategies for microsatellite isolation : a review Molecular Ecology **11**:1-16

Zeven AC, de Wet JMJ (1975) Dictionary of cultivated plants and their regions of diversity. In: Documentation CfAPa (ed). Pudoc, Wageningen, The Netherlands, p 263

Zhang LY, Bernard M, Leroy P, Feuillet C, Sourdille P (2005) High transferability of bread wheat EST-derived SSRs to other cereals. Theoretical and Applied Genetics **111**:677-687

Zhang X, Borevitz JO (2009) Global analysis of allele-specific expression in *Arabidopsis thaliana*. Genetics **182**:943–954

Zhao KY, Aranzana MJ, Kim S, Lister C, Shindo C, Tang CL, . . . Nordborg M (2007) An Arabidopsis example of association mapping in structured samples. Plos Genetics **3**

Zheng ZY, Mosher SL, Fan BF, Klessig DF, Chen ZX (2007) Functional analysis of Arabidopsis WRKY25 transcription factor in plant defense against *Pseudomonas syringae*. BMC Plant Biology **7**

Zhu CS, Gore M, Buckler ES, Yu JM (2008) Status and prospects of association mapping in plants. Plant Genome **1**:5-20

Zhuchenko AA, Rozhmina TA (2000) Mobilizacija genetičeskich resursov l'na. [Mobilization of Flax Genetic Resources] RASCHNIL, VILAR and VNIIL, Starica, Russia

Zohary D, Hopf M (1993) Oil and fiber crops. Charendon press Oxford

Zohary D, Hopf M (2000) Domestication of plants in the Old World: the origin and spread of cultivated plants in West Asia, Europe and the Nile Valley. Oxford University Press, Oxford:316

Sandip Kale

PMB Group, Division of Biochemical Sciences, CSIR-National Chemical Laboratory (NCL),
Pune-411008, Maharashtra, India; Ph. +919028321544; Email: sandipmkale@gmail.com
http://ncl-india.academia.edu/SandipKale

Education and Research

CSIR-National Chemical Laboratory (University of Pune) at Pune, Maharashtra, India

Ph.D.in Biotechnology (Plant Breeding and Genetics) with Dr. Vidya Gupta and Dr. Narendra Kadoo

Thesis: “***Development and application of DNA markers for genetic improvement of linseed (Linum usitatissimum L.)***”

- Developed genomic SSR Markers using novel 454- Amplicon sequencing technique.
- Constructed a core collection of 222 accessions from 3000 accessions using phenotypic data.
- Developed SNPs using GBS techniques and genome-wide association study was carried out to dissect agro economically important traits.
- Genome-wide identification of disease resistance genes in linseed and their expression in resistant variety was studied
- Genome-wide identification and characterization of glucosyltransferase genes
- Proteome profiling of seed development in linseed

University of Nebraska at Lincoln, Nebraska, USA

Visiting Scholar with Dr. Devin Rose and Dr. Dipak Santra, Jan.2013- Dec.2013

Project: ***Molecular and biochemical characterisation of fenugreek accessions (Trigonella foenum-graecum)***

- More than 200 Fenugreek accessions were evaluated for medicinally important compounds *viz.* Galactomannan, 4-Hydroxy-isoleucine and Diosgennin

Vidya Pratishthan’s School of Biotechnology (University of Pune) Maharashtra, India

Graduate research with Dr. Sushma Chaphalkar

Master of Science in Biotechnology, June 2007 – July 2009

Project: “*Virus indexing of Citrus tristeza virus from Citrus plants*”

Bachelor of Science in Biotechnology, June 2004 – May 2007
Project: “***Study of antibacterial activity of limonoids from Rutaceae and Maleaceae families***”

Others:

- Supervised five M.Sc. students
- Worked with colleague on proteome profiling of seed development and characterisation of microRNA genes in linseed
- Attended population genetics class at University of Nebraska, Lincoln
- **Computer skills**: a) Completed basic course on Linux b) Completed Computing data analysis and Statistic one course emphasizing use of R available on Coursera c) Extensive knowledge of NGS data assembly and analysis.

Publications:

1. **S. M. Kale**, V.C. Pardeshi, N. Y. Kadoo, P. B. Ghorpade, M. M. Jana and V. S. Gupta **(2012)**. Development of simple sequence repeat markers in linseed using next generation sequencing technology. ***Molecular Breeding***, 30 , 596-606

2. V. T. Barvkar, V. C. Pardeshi, **S. M. Kale,** N. Y. Kadoo and V. S. Gupta. (**2012)**. Phylogenomic analysis of UDP glycosyltransferase 1 multigene family in Linum usitatissimum identified genes with varied expression patterns. ***BMC Genomics***, 13:175

3. **S.M. Kale**, S.G. Kale, V.C. Pardeshi, G.S. Gurjar, V.S. Gupta, R.T. Gohokar, P.B. Ghorpade and N.Y. Kadoo. **(2012).** Inter-simple sequence repeat markers reveal high genetic diversity among *Alternaria alternata* isolates of Indian origin**. *Journal of mycology and plant pathology*,** 42(2): 194-200

4. V. T. Barvkar, V. C. Pardeshi, **S. M. Kale,** N. Y. Kadoo and V. S. Gupta (**2012**). Proteome profiling of flax (*Linum usitatissimum*) seed: Characterization of functional metabolic pathways operating during seed development. ***Journal of Proteome Research***11 (12), 6264-6276

5. **S. M. Kale,** V. C. Pardeshi, V. T. Barvkar, V. S. Gupta and N. Y Kadoo. (**2013**) Genome-wide identification and characterization of nucleotide binding site leucine-rich repeat genes in linseed reveal distinct patterns of gene structure. ***Genome***, 56:91-99, 10.1139/gen-2012-0135

6. V. T. Barvkar, V.C. Pardeshi, **S.M. Kale**, S. Qiu, M. Rollin S, R. Datla, V. S. Gupta and N. Y. Kadoo. (**2013**). Genome-wide identification and

characterization of microRNA genes and their targets in flax (Linum usitatissimum). ***Planta***. 237:1149–1161

7. **S.M. Kale**, R.L. Srivastava, P.K. Singh, V.C. Pardeshi, V.T. Barvkar, N.Y. Kadoo, V.S. Gupta. Development of core collection of linseed and Genetic diversity, population structure analysis using SSR markers. **Submitted to *Indian Journal of Genetics and Plant Breeding***

8. **S.M. Kale**, D. Jarquin R.L. Srivastava, P.K. Singh, V.C. Pardeshi, V.T. Barvkar, A. Lowrenz, N.Y. Kadoo, V.S. Gupta. Genome wide association mapping in linseed identifies significant loci for different agronomic traits. **Submitted to Molecular Breeding.**

9. **S.M. Kale**, V.C. Pardeshi, V.T. **Barvkar**, N.Y. Kadoo, V.S. Gupta (2012) *In-silico* analysis of EST-SSRs in linseed (*Linum usitatissimum*) for indentifying potential markers for different traits. **Submitted to Bioinformation.**

Techniques learned:

Polymerase Chain Reaction (PCR), PAGE gel and multiplexing Sequencing, Real time PCR, GC, HPLC, LC-MS, NGS data analysis

Scholarships awarded:

Senior Research Fellowship 2011 by Council of scientific and Industrial research (CSIR), New Delhi for Life Sciences
Junior Research Fellowship 2009 by Council of scientific and Industrial research (CSIR), New Delhi for Life Sciences

Conference participation and Awards:

Poster presentation: S.M. Kale, R. Zbasnik, V. Schegel, D. Rose and D. Santra (2013) Biochemical evaluation of fenugreek (*Trigonella foenum-graecum* L) germplasm at **AACCI Annual meeting** held at Albuquerque New Mexico, USA.

Workshop: "Genotyping by Sequencing" (2013) at Institute for Genomic Diversity, Cornell University, Ithaca, NY, USA.

Oral presentation: S.M. Kale, V.C. Pardeshi, N.Y. Kadoo, P.B. Ghorpade and V.S. Gupta (2011) Assessing genetic relationship among linseed genotypes using morphological and molecular markers at **International conference on biodiversity conservation** held at modern college pune-411005, India

Poster presentation: S. M. Kale, V.C. Pardeshi, N. Y. Kadoo, P. B. Ghorpade and V. S. Gupta (2011) Next generation sequencing technology for development of simple sequence repeat markers in linseed at **World congress on biotechnology** held at Hyderabad, India

Poster presentation: S.M. Kale, S.G. Kale, V.C. Pardeshi, G.S. Gurjar, N.Y. Kadoo, R. T. Gohokar, P.B. Ghorpade and V.S. Gupta (2010) Assessing the genetic diversity of the *Alternaria alternata* isolates of Indian origin using Inter simple sequence repeat markers at **National conference on molecular approaches for management of fungal diseases** held at IIHR, Bangalore, India

Participated: In symposium on **'Accelerating biology'** held at CDAC, pune (India) in 2010

References:

Dr. Vidya S. Gupta, Division of Biochemical Sciences, National Chemical Laboratory, Pune-411008, India. [Email: vs.gupta@ncl.res.in] (Ph. +91-020-25902237)
Dr. Narendra Y. Kadoo, Division of Biochemical Sciences, National Chemical Laboratory, Pune-411008, India. [Email: ny.kadoo@ncl.res.in] (Ph. +91-020-25902724)
Dr. Dipak Santra, Department of Agronomy and Horticulture, University of Nebraska, Lincoln, USA [Email: dsantra2@unl.edu] (Ph. +1308-632-1244)

Personal:

Birth date: 13th June 1986

Marital status: Married with no kids

Language speaks: English, Hindi, and Marathi