Research Article **Open Access**

Genome-wide Identification and Characterization of the Gene Families Controlling Fatty Acid Biosynthesis in Flax (*Linum usitatissimum* L)

Frank M You[1]*, Pingchuan Li[1], Santosh Kumar[2], Raja Ragupathy[2], Zhengnan Li[1†] and Sylvie Cloutier[2,4]*

[1]*Cereal Research Centre, Agriculture and Agri-Food Canada, Morden, Manitoba R6M 1Y5, Canada*
[2]*Plant Science Department, University of Manitoba, Winnipeg, Manitoba R3T 2N6, Canada*
[3]*Plant Gene Resources of Canada, Agriculture and Agri-Food Canada, Saskatoon, Saskatchewan S7N 0X2, Canada*
[4]*Eastern Cereal and Oilseed Research Centre, Agriculture and Agri-Food Canada, Ottawa, Ontario K1A 0C6, Canada*
[†]*Current address: Northwest Agriculture & Forestry University, China*

Abstract

Flax (*Linum usitatissimum* L.) is an important crop with many characteristic features such as its abundant essential ω-3 fatty acids for human nutrition. Fatty acid (FA) biosynthesis in plants, including flax, involves several consecutive steps governed by different gene families. Using *in silico* gene mining and comparative analysis, genome-wide gene identification and characterization were performed for six gene families related to FA biosynthesis, including *KAS, SAD, FAD, KCS* and *FAT.* We identified 91 FA-related genes from flax cv. CDC Bethune genome, from which seven previously cloned genes were validated. The newly identified 84 FA-related genes include 14 novel genes from the *KAS* family, two from the *SAD* family, 13 from the *FAD2* family, three from the *FAD3* family, 38 from the *KCS* family and 14 from the *FAT* family. Out of the 91 genes identified, 88 were duplicated as a consequence of recent whole genome duplication events, in which 13 *FAD2* genes were hypothesized to have evolved from tandem gene duplication events followed by a whole genome duplication event and, more recently, by a single gene deletion. The six gene families described here are highly conserved in plants and have diverged anciently. These newly identified flax genes will be a useful resource for further research on FA gene cloning and expression, QTL identification, marker development and marker-assisted selection.

Keywords: Flax; *KAS*; *SAD*; *FAD2*; *FAD3*; *KCS*; *FAT*; Omega-3 fatty acid

Abbreviations: ACP: Acyl Carrier Protein; ALA: α-Linolenic Acid; BAC: Bacterial Artificial Chromosome; BES: BAC-end Sequence; CDS: Coding Sequence; ECR: *Trans*-2,3-enoyl-CoA Reductase; EFA: Essential Fatty acids; EMS: Ethyl Methane Sulfonate; ER: Endoplasmic Reticulum; EST: Expressed Sequence Tag; FA: Fatty Acid; FAD: Fatty Acid Desaturase; FAE: Fatty Acid Elongase; FAT: Fatty acyl-ACP Thioesterase; FPC: Fingerprint Contig; HCD: 3-hydroxacyl-CoA Dehydratase; IME: Intron-mediated Enhancement; KAS: β-ketoacyl-ACP Synthase; KCR: 3-ketoacyl-CoA Reductase; KCS: 3-ketoacyl-CoA synthase; LIO: Linoleic acid; LIN: Linolenic Acid; MY: Million years; MYA: Million Years Ago; OLE: Oleic Acid; PAL: Palmitic Acid; PUFA: Polyunsaturated Fatty Acid; SAD: Stearoyl-ACP Desaturase; STE: Stearic Acid; TAG: Triacylglycerol; TE: Acyl-ACP thioesterase; VLCFA: Very Long Chain Fatty Acid; WGS: Whole Genome Shotgun

Introduction

Flax *(Linum usitatissimum* L.) is a self-pollinated diploid (2n=2x=30) crop from the *Linaceae* family. Its use by humans dates back to the Paleolithic era, nearly 30,000 years ago, but it was domesticated for its stem fibers and seed oil only ~7,000 years ago [1]. The oilseed morphotype is referred to as linseed or flaxseed and the currently grown varieties have oil contents up to 50% [2] with unique fatty acid (FA) compositions. The eighteen carbon FAs constitute the major FAs of linseed with stearic (STE; C18:0, where the C*x:y* denotes a FA with *x* carbons and *y* double bonds), oleic (OLE; C18:1cis$^{\Delta 9}$), linoleic (LIO; C18:2cis$^{\Delta 9,12}$) and linolenic (LIN; C18:3 cis$^{\Delta 9,12,15}$) acid contents of approximately 4.4%, 24.2%, 15.3% and 50.1%, respectively [3]. LIN is also referred to as α-linolenic acid (ALA), an omega-3 FA. LIO and LIN are essential fatty acids (EFAs), precursors of the omega (ω)-6 and ω-3 families, respectively [4]. The majority of the oilseed crops contain LIO but LIN is present only in oils from certain fish, microalgae and crops such as canola (rapeseed or oilseed rape) and

linseed [4]. The ALA content in linseed can be as high as 60 to 73% in high-linolenic acid varieties making this crop the richest source of plant based omega-3 FAs [5]. LIO and LIN, collectively called polyunsaturated FAs (PUFAs), regulate plant metabolism, hormone signaling and contribute to membrane integrity in addition to their role as an energy reservoir in the form of triacylglycerols (TAGs) [6]. The nutraceutical industry promotes linseed as a rich source of ALA and lignan, improving cardiovascular and brain health [7-9]. In animal husbandry and in the food industry, animal feed is fortified with linseed or linseed meal to enrich the ALA content in products such as meat, milk and eggs [10].

The FA biosynthesis pathway in plants involves the sequential elongation and desaturation of FA precursors (Figure 1). The monofunctional FA synthases use acetyl-CoA as the starting substrate and malonyl-acyl-carrier-protein (ACP) as the elongator. This initial reaction is catalyzed by 3-ketoacyl-ACP synthase type III (KAS III). The malonyl-thioester undergoes recurring condensation with acetyl-CoA up to C16:0-ACP which is catalyzed by KAS I/KAS B isoforms. The
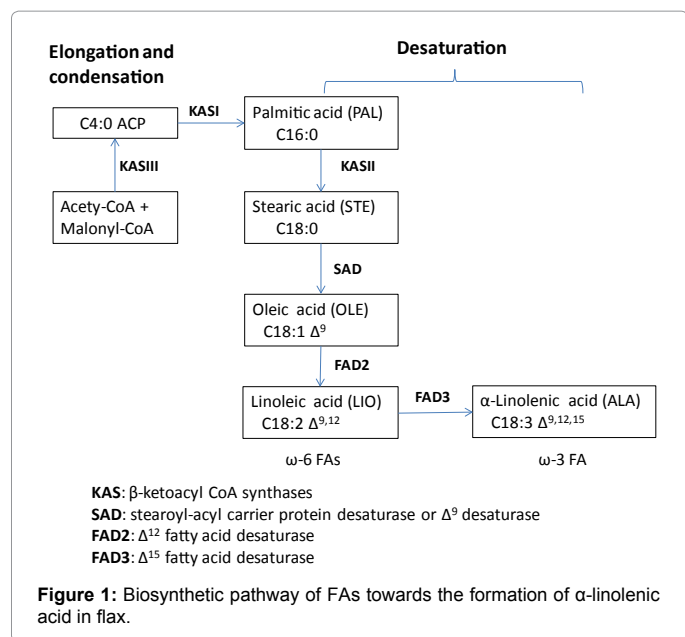
Figure 1: Biosynthetic pathway of FAs towards the formation of α-linolenic acid in flax.

KAS II/KAS A isoforms finally elongate 16:0-ACP to C18:0-ACP [11]. The step-by-step desaturation of C18:0 FA determines the saturated to unsaturated FA ratio which, in turn, influences the end use of the oil in food source and industrial applications [12,13].

The desaturation of FAs is carried out by desaturases which insert double bonds in the linear hydrocarbon chain of FAs [14,15]. Many of the genes encoding for these and other enzymes involved in FA biosynthesis in flax have been identified and characterized [16-22].

The enzyme stearoyl-ACP desaturase (SAD) introduces a double bond at the Δ9 position of stearoyl-ACP to convert it to oleoyl-ACP and thereby increases the unsaturated FA content of plants [23]. Two paralogous *SAD* loci, *SAD1* and *SAD2*, have been previously identified in flax [24]. The cDNA sequences encoding the SAD proteins were isolated and characterized from flax cultivars Glenelg [25] and AC McDuff [18]. *SAD1* and *SAD2*, cloned and characterized from various plant species such as soybean and *Arabidopsis*, show highly conserved exon structure; however, they are structurally unrelated to their animal and fungal homologues [26].

FA desaturase (FAD) enzymes introduce additional double bonds into the mono-unsaturated OLE [15]. FAD2 desaturates OLE into LIO by adding a double bond at the Δ12 position. Two closely related *FAD2* genes, *FAD2-2* and *FAD2*, were cloned and characterized from flax genotypes NL97 and Nike [19,20]. The FAD3 enzymes add an additional double bond at the Δ15 position in LIO to produce LIN. Three *FAD3* genes previously identified in the flax genome include *FAD3a* and *FAD3b* from cultivar Normandy [22] and *FAD3c* from flax cultivars AC McDuff and breeding lines UGG5-5 and SP2047 [16]. FAD3a and FAD3b are the major enzymes controlling the LIN portion of the storage lipids in flax seeds [22]. Both FAD2 and FAD3 are membrane-bound proteins containing three highly conserved histidine (HIS)-box motifs essential for enzyme activity [15,27]. The exact function of *FAD3c* is yet to be established but a gene expression study indicated that it likely did not play a major role in LIN accumulation in seeds [16].

A recent study on 120 flax accessions representing a broad range of germplasm including some ethyl methane sulfonate (EMS) mutant lines identified a total of six alleles for *SAD1* and *SAD2*, 21 for *FAD2a*, 5 for *FAD2b*, 15 for *FAD3a* and 18 for *FAD3b* corresponding to 4, 2, 3, 4, 6 and 7 isoforms, respectively [28]. The study also found significant correlation between *SAD* and *FAD* isoforms and both FA composition and oil content [28]. Genes encoding desaturases involved in FA biosynthesis have also been cloned and characterized from other plant species such as *Arabidopsis* [29], oilseed rape [30], soybean [31], peanut [32] as well as cyanobacteria [33] and algae [34]. These desaturases exhibit conservation across species in both sequence and domain architecture [28].

Very-long-chain fatty acids (VLCFAs) are FAs longer than 18 carbons (C18) in length. VLCFAs are required in all plant cells for the production of sphingolipids and phospholipids, and in specific cell types for the synthesis of other VLCFA derivatives such as cuticular waxes, pollen coats and suberin [35]. FA elongation to VLCFAs initiates from C18 FA and requires four successive reactions catalyzed by four different enzymes coordinated in an endoplasmic reticulum (ER)-associated complex. The first FA elongation reaction is the condensation of a long chain acyl-CoA with a malonyl-CoA by 3-ketoacyl-CoA synthase (KCS). The resulting 3-ketoacyl-CoA is then reduced by 3-ketoacyl-CoA reductase (KCR) to 3-hydroxyacyl-CoA which is then dehydrated by 3-hydroxacyl-CoA dehydratase (HCD) to form *trans*-2,3-enoyl-CoA. The last elongation reaction is the reduction of *trans*-2,3-enoyl-CoA by *trans*-2,3-enoyl-CoA reductase (ECR) to form a two-carbon elongated acyl-CoA [35]. The KCS enzyme has been hypothesized to play a role in determining the substrate and tissue specificities of FA elongation whereas the three other enzymes are thought to have broad substrate specificities [36,37]. Some *KCS* genes coding KCS enzymes have been identified in *Arabidopsis*. An *FAE1*, isolated from the *Arabidopsis thaliana* mutant *fae1* [38,39], was the first seed-specific *KCS* gene characterized for the extension of C18 to C20 or C22 in storage lipids [40]. Subsequently twenty-one further *FAE1*-like *KCS* genes have been identified and characterized in *Arabidopsis* [41,42]. The *FAE1*-like *KCS* genes have also been recently identified in many other plant species such as *Gossypium raimondii* [43] and soybean [31].

The acyl-ACP thioesterases (TEs) differ from KASs and KCSs by their role in the termination of fatty acyl group extension by hydrolyzing the acyl moiety from the anabolically active acyl-ACP at an appropriate chain length and, eventually, releasing the free FAs [44-46]. TEs are nuclear encoded enzymes that mature in the plastid by N-terminal transit peptide hydrolysis [47]. A comparison of more than 30 plant TE sequences revealed that they can be grouped into two distinct classes of fatty acyl-ACP thioesterases: FatA and FatB [47]. The FatA class is specific for unsaturated 18:1-ACP substrates with minor activities on 18:0- and 16:0-ACPs, whereas FatB shows marked activity on the saturated acyl-ACPs with chain length varying between 8 and 16 carbons [47-49]. The conserved and ubiquitous *FatA* or *FatB* genes have been identified and characterized in many plants [49] including *Arabidopsis* [50-52], brassica [53], *Cuphea* [47] and maize [54].

With the rapid advance of next generation sequencing technologies, whole genome sequences are becoming publicly available for an increasing number of plant genomes (http://www.phytozome.net/). *In silico* gene mining approaches using genome-wide gene annotations, expressed sequence tags (ESTs) and RNA-Seq data have been successfully applied to identify and characterize plant gene families [55,56]. The availability of the whole genome shotgun (WGS) sequence of flax [57] provides an opportunity to systematically analyze the gene families controlling FA biosynthesis in this species. Here,

we report on genome-wide *in silico* and comparative analyses of the first iteration of the flax genome sequence to identify and characterize several important gene families controlling FA biosynthesis, including *KAS*, *SAD*, *FAD2*, *FAD3*, *KCS* and *FAT*. Complete information on FA biosynthesis related gene families is essential for a better understanding of the FA biosynthesis pathway and for effective genetic improvement of flax seed oil profiles.

Materials and Methods

Flax genome sequences

The draft flax scaffold sequences and their gene annotations were downloaded from the Phytozome (v 9.0) database (ftp://ftp.jgi-psf.org/pub/compgen/phytozome/v9.0/Lusitatissimum/). A total of 88,420 unsorted nucleotide scaffold sequences and their predicted 43,484 annotated genes were analyzed.

In silico identification of FA gene families

A comparative gene analysis approach was used to identify gene families controlling FA biosynthesis in linseed cv. CDC Bethune. Six major classes of FA biosynthesis related genes, *KAS*, *SAD*, *FAD2*, *FAD3*, *KCS* and *FAT*, were investigated. Some previously identified flax FA genes and their conserved orthologs in other plant species were downloaded from the NCBI nucleotide and protein databases. These genes are *KAS* from flax (*KAS I*, CD760578.1; *KAS II*, CD760581.1), *Arabidopsis* (*AtKAS*, AT2G04540.1; *AtKAS I*, AT5G46290.3; *AtKAS II*, AT1G74960.1; *AtKAS III*, AT1G62640.1), soybean (*GmKAS I*, Glyma08g08910, Glyma05g36690, Glyma08g02850, Glyma18g10220, Glyma05g25970; *GmKAS II*, Glyma17g05200; *GmKAS II*, Glyma13g17290) and castor bean (*RcKAS III*, A6N6J4); *SAD* from flax (*SAD*, X70962; *SAD1*, AJ006957; *SAD2*, JN653452, AJ006958); *FAD2* from flax (EU660502), Ethiopian mustard (*Brassica carinata*, AF124360.2), *Arabidopsis* (NP187819.1) and field pepperwort (*Lepidium campestre*) (FJ907546.1) and *FAD3* from flax (ABA02172, ADV92268 and ADV92272). For *KCS* and *FAT*, twenty-one *KCS* genes from *Arabidopsis*, ten *KCS* genes from soybean, and twenty-five *FAT* genes from twenty-one different plant species (*Arabidopsis thaliana*, *Bradyrhizobium japonicum*, *Brassica napus*, *Cinnamonum camphorum*, *Capsicum chinense*, *Cuphea hookeriana*, *Cuphea lanceolata*, *Cuphea palustris*, *Coriandrum sativum*, *Carthamus tinctorius*, *Cuphea wrightii*, *Elaeis guineensis*, *Gossypium hirsutum*, *Garcinia mangostana*, *Helianthus annuus*, *Iris germanica*, *Iris tectorum*, *Myristica fragrans*, *Triticum aestivum*, *Ulmus Americana* and *Umbellularia californica*) were downloaded from GenBank and used for comparative analyses.

cDNA and protein sequences of known flax FA genes and their orthologs were aligned against the flax genome using BLASTN and BLASTP with an E-value of 1e-30 and 1e-10, respectively. Flax genes with hits in both BLASTN and BLASTP were aligned using CLC Sequence Viewer v6.8.1 (CLC Bio, Aarhus, Denmark) for both cDNA and protein sequences, and, phylogenetic analyses were carried out using the Neighbor-Joining (NJ) algorithm implemented in MEGA 6.0 [58]. Gene structures were compared to existing FA genes and orthologs. Finally, genes with significant FA gene features (such as HIS-box, dilysine motif, etc) and that clustered with the known FA genes and orthologs in phylogenetic trees, were considered.

Digital differential transcription of identified FA genes

Flax ESTs have been generated from cv. CDC Bethune [59], the same cultivar used for WGS sequencing [57]. A total of 261,272 ESTs from 13 libraries including many stages of developing embryos, seed coat, endosperm, flowers, etiolated seedlings, leaves, and stem tissue (LIBEST_026995 - LIBEST_027011) were downloaded from GenBank (www.ncbi.nlm.nih.gov). Additionally, 11,640 ESTs from bolls 12 days after flowering in flax cv. AC McDuff and from outer fiber-bearing tissues at mid-flowering stage in flax cv. Hermes were also downloaded from GenBank. BLASTN searches of the coding sequences of the identified FA genes were performed against the ESTs at an E-value threshold of 1e-30. The best EST hits were counted for each FA gene. For ESTs that hit multiple FA genes, only the one with the largest bit score or the smallest E-value was assigned to the FA gene. The total numbers of EST hits per FA gene were used to characterize digital expression levels of FA transcripts.

Gene duplication

Gene duplication analysis of identified FA genes was performed. The cDNA sequences of the FA genes were searched against themselves (self-BLASTN) using a threshold E-value of 1e-30. Pairs of FA genes returning reciprocal top hits of each other and having identical or fairly similar gene structures were considered to be duplicate copies. Duplication time was calculated using the molecular clock proposed for evolution of duplicate genes [60]. The cDNA sequences of two duplicate genes were aligned and synonymous substitution (Ks) was calculated with the MEGA software (v6.0) [58]. The evolutionary distance between two duplicate genes was calculated based on the Ks corrected with the Nei-Gojobori model of nucleotide evolution which accounts for multiple substitutions per site [58]. The divergence (k) of a pair of duplicated genes can be converted into duplication or divergence time (t) in million years (MY) using $t=k/(2r)/10^6$, where r is the substitution rate of 6.5×10^{-9} substitutions per synonymous site per year [61].

Identity of pairs of duplicate genes was calculated by alignment with ClustalW (http://www.genome.jp/tools/clustalw/) followed by identity calculation with the SIAS server (http://imed.med.ucm.es/Tools/sias.html). Identity was defined as the number of identical positions divided by the length of the alignment. Gaps in alignments were taken into account.

Chromosome location of FA gene families

The draft WGS sequence of flax was reported [57] but it was not sorted based on chromosomes or linkage groups. In order to align these scaffolds, the flax physical map, the bacterial artificial chromosome (BAC) end sequences (BESs) [62] and the consensus SSR map with 15 linkage groups [63] were used. The physical map consists of 416 contigs spanning ~368 Mb, representing approximately 98.7% of the haploid genome (373 Mb) [57]. A total of 43,776 BACs from the CDC Bethune BAC library were end sequenced to generate 87,552 BESs which covered all physical map contigs and almost all BACs [62]. This end sequencing enabled the anchoring of scaffold sequences to all contigs through the mapping of the BESs to the scaffold sequences and the subsequent mapping of the physical map contigs to the linkage groups using the consensus SSR map which included SSR markers developed from scaffold sequences. In order to locate the identified genes in linkage groups, we adopted the following procedure: (1) anchoring the flax draft WGS sequence, i.e., scaffolds onto the flax fingerprint contig (FPC) map [62] by mapping BESs using BLASTN; (2) anchoring FPCs onto the fifteen flax linkage groups using the consensus SSR map [63]; (3) locating the identified FA genes on scaffolds and corresponding linkage groups. A detailed methodology to order the flax genome sequences will be published separately.

Results and Discussion

KAS gene family

The *KAS* gene family includes *KAS I, KAS II* and *KAS III*. The first two are involved in the synthesis of palmitic acid (PAL) and STE, respectively, whereas *KAS III* enzymes control the initial reaction to form C4:0 ACP (Figure 1). Three pairs of *KAS I*, one pair of *KAS II* and two pairs of *KAS III* genes were identified in flax. Each member of the *KAS* gene family was observed in two copies as a consequence of a recent genome duplication event that occurred 4.4-16.6 MYA (Table 1). Each pair of *KAS* genes had identical or similar numbers of introns and exons (Figure 2A). Also, a pair of mitochondrial *KAS* genes (*mtKAS-1* and *mtKAS-2*) located on LG3 and LG15 was identified. They had 13 and 14 exons, respectively, and they displayed high similarity at the protein level to soybean mitochondrial orthologs (Glyma13g19010, Glyma10g04680) and to an *Arabidopsis mtKAS* (AT2G04540) (Figure 3). Plant FAs are biosynthesized in plastids and further modified in the ER. However, other cellular compartments, like mitochondria, also have the ability to *de novo* biosynthesize FAs. This pathway is assumed to be conserved in species such as *Neurospora*, the species from which the *de novo* FA synthesis in mitochondria was originally elucidated [64]. Other species including pea and *Arabidopsis* are also reported to harbor FA synthesis components like mtKAS and ACP [64-69].

All *KAS I, KAS II* and *mtKAS* genes shared high sequence similarity in the conserved region of the proteins (Figure S1). Also all *KAS* genes differed at the N-terminus (Figure S1). Phylogenetic analysis showed that *KAS I, KAS II, KAS III* and *mtKAS* genes clustered with their orthologs in *Arabidopsis, Glycine max, Jatropha curcas, Populus trichocarpa* or *Ricinus communis* as an individual branch, suggesting function conservation of each class of *KAS* genes (Figure 3). *KAS III* diverged from *KAS I, KASII* and *mtKAS* approximately 91.9 -106.9 million years ago (MYA), substantially earlier than the divergence among *KAS I, KAS II* and *mtKAS* which was estimated at 36.7-58.1 MYA (Table S1).

KAS I, KAS II and *mtKAS* in flax and other plant species share two protein domain families, PF02801 (β-ketoacyl-ACP synthase, C-terminal domain) and PF00108 (thiolase, N-terminal domain) (Figure 3). They had a strictly conserved active site triad, Cys-His-His [70,71] such as Cys220-His360-His396 of KAS Ia-Lu10025390 (Figure

S1). Although the structure comparison of *KAS I, KAS II* and *mtKAS* could not reveal the basis of chain length specificity [71], the protein sequences of the three gene families had their own conserved regions (Figure S1). *KAS III* genes showed conservation in two protein domain families, PF08541 and FP08545 which are both 3-Oxoacyl-[ACP] synthase III (Figure 3), and in a conserved active site triad Cys-His-Asn (e.g., Cys182-His334-Asn364 of KAS IIIa-Lu10024608 or Cys220-His360-Asn364 of KAS Ia-Lu10025390) (Table 1 and Figure S1). Arg339 of KAS IIIa-Lu10024608 (or Arg399 of KASIa-Lu10025390) is another conserved residue critical in the interaction between KAS III and ACP [72], In addition, KAS III proteins possess the highly conserved motif GNTSAAS in flax and other plant species (Figure S1) [70]. The motif GNTSAAS was proposed to be responsible for the binding of acyl-ACPs. Absence of the tetrapeptide GNTS changes the secondary structure and results in complete loss of condensing activity of KAS III [73].

SAD gene family

Stearoyl-ACP Δ9-desaturase (SAD) is the only known soluble desaturase present in the chloroplast stroma [25] involved in the synthesis of ALA. It can convert the stearate into ACP-bound oleic acid (18:1) by introducing the first double bond at the 9th position from the carboxylic end (α-end). *SAD* genes have been identified in several plants [15,26,74] including seven *DES-like SAD* genes in *Arabidopsis* (At1g43800, At2g43710, At3g02610, At3g02620, At3g02630, At5g16230, At5g16240).

In flax, three *SAD* genes have been reported. The first flax *SAD* (X70962) [25] was isolated from cv. Glenelg by hybridization with the cDNA-derived castor *SAD* probe [25]. Two isoforms each from *SAD1* (AJ006957) and *SAD2* (AJ006958) were deduced from flax cv. McGregor by promoter cloning [24]. *SAD2* (JN653452) was also cloned from Turkish flax germplasm Uw15 [75]. In addition, one EST of *linSAD1* (CD760586) [18], as well as a truncated *linSAD2* (CD760587) [18] corresponding to *SAD1* and *SAD2* were reported (Figure S3).

In the multiple protein sequence alignments of *SAD* genes, four (X70962, AJ006957, AJ006958 and JN653452) shared high similarity with Lus10027486 and Lus10039241, indicating that all six genes belonged to the same family (Figure 4A, Figure S2). The DNA sequence alignment showed a deletion of the dinucleotide GA starting at position

No	Gene	Gene ID	Chromosome location	Start position in scaffold	Genomic sequence length (bp)	CDS sequence length (bp)	Amino acid sequence length (aa)	No of exons	CDS identity (%)	Amino acid identity (%)	Duplication time (MYA)	Ks	Catalytic site				
													C220	H360	H396	N413	R399
1	*KAS Ia-1*	Lus10025390	scaffold46 (LG12)	408940	2,695	1,410	469	7	91.7	94.9	9.5	0.123	+	+	+	G	K
2	*KAS Ia-2*	Lus10015267	scaffold924 (LG12)	234076	2,638†	1,425	470	6					+	+	+	G	K
3	*KAS Ib-1*	Lus10040883	scaffold156 (LG3)	2043083	3,012	1,425	474	7	68.4	67.3	16.8	0.219	+	+	+	G	K
4	*KAS Ib-2*	Lus10004935	scaffold858 (LG8)	182334	1,923	1,092	363	7					+	+	+	G	K
5	*KAS Ic-1*	Lus10001814	scaffold3494 (LG7)	17887	2,893	1,719	582	8	75.7	77.7	15.0	0.195	+	+	+	G	K
6	*KAS Ic-2*	Lus10003195	scaffold1056 (LG9)	24875	2,847	1,455	484	6					+	+	+	G	K
7	*KAS II-1*	Lus10034886	scaffold66 (LG5)	1423486	4,672	1,692	563	13	77.5	76.6	4.4	0.057	+	+	+	G	H
8	*KAS II-2*	Lus10033422	scaffold488 (LG3)	1269585	4,837	1,401	466	12					+	+	+	G	H
9	*KAS IIIa-1*	Lus10024608	scaffold349 (LG8)	281086	2,572	1,227	408	8	96.6	97.1	6.2	0.081	+	+	N	+	+
10	*KAS IIIa-2*	Lus10032246	scaffold291 (LG5)	266433	2,572	1,227	408	8					+	+	N	+	+
11	*KAS IIIb-1*	Lus10004342	scaffold1134 (LG9)	133755	2,362	1,212	403	8	94.7	93.1	7.5	0.098	+	+	N	+	+
12	*KAS IIIb-2*	Lus10028925	scaffold540 (LG7)	890572	2,453	1,209	402	8					+	+	N	+	+
13	*mtKAS-1*	Lus10014622	scaffold398 (LG3)	39249	2,888	1,428	475	13	86.8	86.4	5.5	0.071	+	+	+	G	T
14	*mtKAS-2*	Lus10033810	scaffold222 (LG15)	385235	3,987	1,584	527	14					+	+	+	G	T

†Corrected length. Catalytic site: +indicates the presence of the residues of the putative catalytic sites. The residue position is based on the sequence of *KAS Ia-1*.

Table 1: *KAS* gene family involved in fatty acid elongation and condensation identified in flax cv. CDC Bethune.
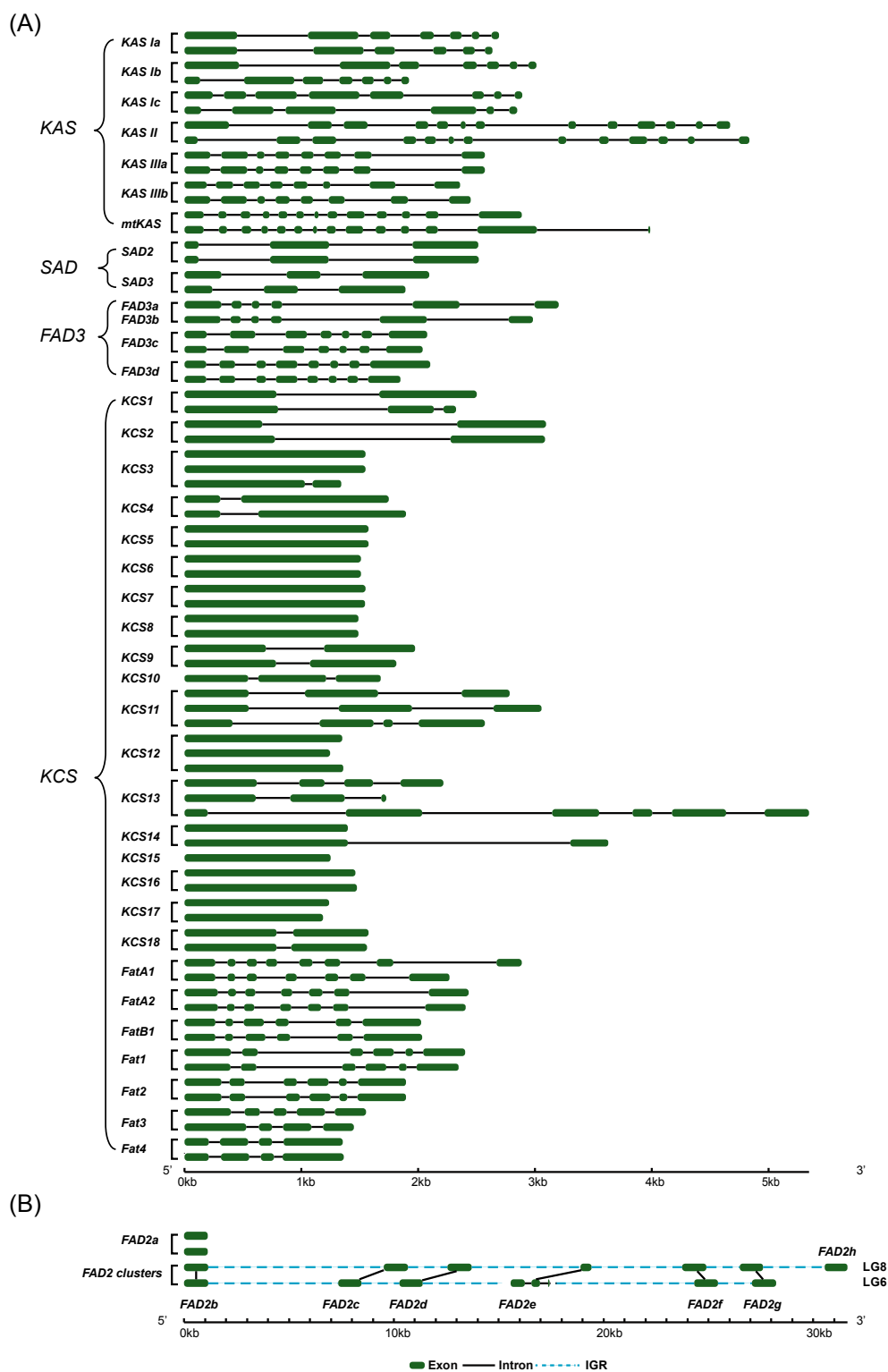
Figure 2: Gene structure of the *KAS*, *SAD*, *FAD3*, *KCS*, *FAT* (A) and *FAD2* (B) gene families identified from flax cv. CDC Bethune. Green boxes represent exons. Solid black lines between green boxes represent introns and dashed blue lines represent intergenic regions in the case of the intronless *FAD2* gene family except for *FAD2e-2* which had three exons and two introns. Pairs of duplicated genes are presented where the top one represents *copy 1* and the bottom one represents *copy 2* in Tables 1, 2, 5 and 6; for example, *mtKAS-1* and *mtKAS-2*. Some genes have a third copy such as *KCS3*. Gene structure was drawn using Gene Structure Display Server v2.0 (http://gsds.cbi.pku.edu.cn/).
IGR: intergenic region.

Figure 3: Phylogenetic tree of the *KAS* gene family from flax cv. CDC Bethune. Multiple sequence alignments were generated using the predicted amino acid sequences. The unaligned 160 amino acids of the N-terminus with variable deduced protein lengths were truncated to facilitate the NJ tree construction. Orthologous *KAS* genes previously identified from other plants species are labeled in red and the bootstrap support expressed as the percentage of the 1,000 bootstrap replicates is shown. At: *Arabidopsis*; Gm: *Glycine max* (soybean); Jc: *Jatropha curcas*; Pt: *Populus trichocarpa* (black cottonwood); Rc: *Ricinus communis* (castor bean).

132 bp of X70962 (Figure S3). This two base pair deletion resulted in a frame shift mutation between the 45th and 82nd amino acid residues of X70962 (Figure S2).

Phylogenetic analysis using cDNA and protein sequences (Figures 4A and 5), as well as multiple sequence alignments (Figures S2 and S3), showed that both AJ006957 and X70962, derived from different varieties, corresponded to Lus10027486 (LG10) and, JN653452 and AJ006958 corresponded to Lus10039241 (LG 11), whereas Lus10027486 and Lus10039241 formed a pair of duplicated genes. Based on our analysis, we concluded that the *SAD1* and *SAD2* identified from cv. McGregor and Uw15 were duplicate copies of the same gene, which are designated *SAD2* in order to remain consistent with previous gene nomenclature.

We identified two new *SAD* genes Lus10018926 and Lus10028627. These *SAD* genes share more than 95% DNA and protein sequence similarity suggesting their duplicate nature (Table 2 and Figure S3). Based on the phylogenetic analysis (Figure 4A), Lus10018926 and

Lus10028627 belonged to a separate cluster. The *SAD2* genes and their new paralogs shared similar intron/exon structure and length (Figure 2A). In addition, two genes grouped with the *SAD6* gene (At1g43800) from *Arabidopsis* (Figure 4A). We designated them *SAD3-1* and *SAD3-2*, respectively. The multiple protein sequence alignment between *SAD2* and *SAD3* suggests that the two *SAD3* genes varied at the N-terminus compared to *SAD2-1* and *SAD2-2* (Figure S2). *SAD3-2* (SAD3-Lus10028627) lacked 34 and 26 amino acids from the N-terminal region compared with *SAD2* and *SAD3-1* (Figure S2). We further searched the flax EST database and found that the coding sequence (CDS) of SAD3-Lus10028627 (*SAD3-2*) was 100% similar to EST JG184015 which had the complete N-terminal fragment absent in SAD3-Lus10028627. Re-annotation of the genomic sequence of this gene confirmed that improper determination of the starting codon caused exclusion of the N-terminal fragment in *SAD3-2*. The corrected SAD3-Lus10028627 had 95% CDS similarity with *SAD3-1* (SAD3-Lus10018926) (Table 2). To determine whether the *SAD* genes were targeted to the plastid where FAs are synthesized, we used ChloroP1.1, an online chloroplast

No	Gene	Gene ID	Chromosome location	Start position in scaffold	Genomics sequence length (bp)	CDS sequence length (bp)	Amino acid sequence length (aa)	No of exons	CDS identity (%)	Amino acid identity (%)	Duplication time (MYA)	*Ks*
SAD family												
1*	*SAD2-1*	Lus10027486	scaffold96 (LG10)	605098	2,515	1,191	396	3	97.1	99.5	7.4	0.096
2*	*SAD2-2*	Lus10039241	scaffold33 (LG11)	150467	2,519	1,191	396	3				
3	*SAD3-1*	Lus10018926	scaffold103 (LG1)	558807	2,096	1,185	394	3	95.3	96.4	12.5	0.163
4	*SAD3-2*	Lus10028627	scaffold346 (LG14)	461462	1,894†	1,107	368	3				
FAD2 family												
1*	*FAD2a-1*	Lus10012007+ Lus10012008	scaffold931 (LG1)	161638	1,137†	1,137†	378	1	99.8	100	0.6	0.008
2*	*FAD2a-2*	Lus10029283+ Lus10029284	scaffold360 (LG1)	463408	1,137†	1,137†	378	1				
3	*FAD2b-1*	Lus10021051	scaffold155 (LG8)	296486	1,149	1,149	382	1	95.7	99.5	3.8	0.050
4	*FAD2b-2*	Lus10004175	scaffold2404 (LG6)	134640	1,149	1,149	382	1				
5	*FAD2c-1*	Lus10021050	scaffold155 (LG8)	286964	1,143	1,143	380	1	89.7	91.1	17.8	0.232
6	*FAD2c-2*	Lus10004176	scaffold2404 (LG6)	141984	1,119	1,119	372	1				
7	*FAD2d-1*	Lus10021049	scaffold155 (LG8)	283912	1,155	1,155	384	1	91.1	92.7	12.2	0.159
8	*FAD2d-2*	Lus10004177	scaffold2404 (LG6)	144905	1,125	1,125	374	1				
9	*FAD2e-1*	Lus10021048	scaffold155 (LG8)	278212	516	516	171	1	40.3	42.1	9.4	0.122
10	*FAD2e-2*	Lus10004178	scaffold2404 (LG6)	150204	1,864	1,864	392	3				
11	*FAD2f-1*	Lus10021047	scaffold155 (LG8)	272730	1,146	1,146	381	1	92.7	95.8	7.7	0.100
12	*FAD2f-2*	Lus10004180	scaffold2404 (LG6)	158962	1,146	1,146	381	1				
13	*FAD2g-1*	Lus10021046	scaffold155 (LG8)	270018	1,125	1,125	374	1	90.5	91.9	3.7	0.048
14	*FAD2g-2*	Lus10004181	scaffold2404 (LG6)	161731	1,146	1,146	381	1				
15	*FAD2h*	Lus10021045	scaffold155 (LG8)	265986	1,110	1,110	369	1				
FAD3 family												
1*	*FAD3a*	Lus10038321	scaffold28 (LG7)	701556	3,208	1,179	392	6	94.8	93.9	8.6	0.112
2*	*FAD3b*	Lus10036184	scaffold27 (LG12)	60680	2,989	1,176	391	6				
3*	*FAD3c-1*	Lus10040660	scaffold 156 (LG3)	906049	2,080	1,179	392	7	95.4	97.2	7.7	0.100
4	*FAD3c-2*	Lus10018245	scaffold183 (LG12)	202288	2,039	1,161	386	7				
5	*FAD3d-1*	Lus10027809	scaffold1143 (LG1)	472802	2,106	1,359	452	8	80.4	81.4	6.2	0.081
6	*FAD3d-2*	Lus10005039	scaffold1982	44775	1,851	1,122	373	8				

*Previously identified genes. †Corrected length.

Table 2: Genes related to the fatty acid desaturation including S*AD*, *FAD2* and *FAD3* gene families identified in flax cv. CDC Bethune.

targeting protein prediction tool [76]. Chloroplast targeting sequences were identified in both flax *SAD2* genes and *Arabidopsis SAD6* but not in the *SAD3* genes. We noticed that *FAD3* genes lacked a fragment in the N-terminus existing in the *SAD2* and *SAD6* genes where the chloroplast targeting sequence is located (Figure S2). It is hypothesized that the function of the *SAD3* genes might have evolved or have been lost during evolution.

FAD2 gene family

FA desaturase (FAD) is a membrane-bound protein generally targeted to the ER. FAD2 enzymes can utilize OLE as a substrate to synthesize LIO by adding the 2nd double bond at position Δ^{12} towards the synthesis of ALA/LIN [13]. Two closely related *FAD2* genes, *FAD2a* and *FAD2b*, were cloned and characterized from flax genotypes Nike and NL97 [19,20]. In addition, *FAD2* genes have been identified in many other plant species such as *Arabidopsis* [77], rice [78], *Brassica napus* [79], grape [80] and safflower [81]. Thus, an attempt was made to identify the full complement of *FAD2* genes in flax. Multiple homologs of *FAD2* from other plant species were employed to query for a homology search against the flax WGS reference sequence. A total of 15 *FAD2* genes were identified including seven pairs (*FAD2a-g*) of duplicate genes (distinguished by the suffix after the gene name, see Table 2 for more details) and one single gene, *FAD2h*. Except for the duplicated *FAD2a*, located on LG1, the remaining thirteen *FAD2* family genes were present in two clusters of tandem repeats on LG6 (scaffold

155) and LG8 (scaffold 2404), respectively (Figure 2B and Table 2). Of these 13 *FAD2* genes, 12 are actually six pairs of syntenic duplicate genes as suggested by the highly conserved DNA/protein sequence similarity (Figures S4 and S5) and the phylogenetic analysis (Figure 4B) in which two copies of each designated *FAD2* gene pair grouped together. The previously identified *FAD2* genes from various plant species clustered with the two flax *FAD2a* duplicate genes, implying high conservation of *FAD2a* in plant species.

We calculated the duplication time for each pair of *FAD2* genes and between *FAD2* gene pairs (Table 3). *FAD2a* diverged from *FAD2b-h* approximately 93 MYA, indicating that *FAD2a* and the ancestor of *FAD2b-h* were the two most ancient copies of *FAD2* genes. The seven genes of *FAD2b-h* resulted from a tandem repeat duplication event in scaffold 155 (LG8) because duplication times between all pairwise *FAD2b-h* genes were similar (the only exception was between *FAD2f* and *FAD2g*), averaging 33 MYA. *FAD2b* might be the ancestral copy of the tandemly duplicated genes (*FAD2b-h*) based on the closer evolutionary relationship between *FAD2a* and *FAD2b* as compared to others (Figure 4B). Another tandem duplication copy of *FAD2b-g* was generated in scaffold 2404 (LG6) between 3.7 and 17.8 MYA with an average of 9.1 MYA to form six duplicated gene pairs but no corresponding duplicate member for *FAD2h* was found. This structure suggests that a tandem duplication of seven *FAD2* genes (*FAD2b-h*) in scaffold 155 (LG8) was followed by a genome duplication event to create another tandem duplication copy of seven genes further followed

	FAD2a	FAD2b	FAD2c	FAD2d	FAD2e	FAD2f	FAD2g
FAD2a	0.6						
FAD2b	86.4	3.8					
FAD2c	108.0	34.4	17.8				
FAD2d	73.8	30.2	35.9	12.2			
FAD2e	120.5	29.0	31.4	32.3	9.4		
FAD2f	76.6	36.1	37.5	32.0	44.9	7.7	
FAD2g	100.3	38.5	38.1	32.6	42.9	8.8	3.7
FAD2h	86.2	25.2	31.5	25.3	30.4	37.0	39.4

The diagonal elements represent duplication time between pairs of FA duplicate genes whereas the others show the average duplication time between any two pairs of genes.

Table 3: Duplication time (MYA) of *FAD2* genes.

Gene	Scaffold	LG	TSS	5'UTR intron	Start codon position	Strand
FAD2a-1	Scaffold931	1	163503	162667-163412	162634	Minus
FAD2a-2	Scaffold360	1	465421	464614-465329	464581	Minus
FAD2b-2	Scaffold2404	6	137814	135809-137697	135788	Minus

LG: linkage group;TSS: transcription start site

Table 4: 5'UTR introns of *FAD2* genes determined by alignment with flax ESTs.

by a deletion of *FAD2h* in scaffold 2404 (LG6). Thus, it can be inferred that the *FAD2* genes in scaffold155 (LG8) should be the ancient copy paralogous to the derived copy in scaffold 2404 (LG6), supporting the two genome duplication events hypothesized to have occurred during the flax genome evolution [57,82].

The *FAD2a* genes on LG1 were incorrectly annotated into two separate genes (Lus10012007 and Lus10012008 for *FAD2a-1* and Lus10029283 and Lus10029284 for *FAD2a-2*) in the original WGS genome annotation [57]. The intact *FAD2a-1* should be represented by merging Lus10012007 and Lus10012008 and *FAD2a-2* by merging Lus10029283 and Lus10029284. To verify this correction, we compared the gene sequence from BAC clone LuBAC346C18 containing the *FAD2a-2* locus with Lus10029283 and Lus10029284 in scaffold 360. *FAD2a-2* had a fragment length of 1,137 bp in the BAC. The multiple sequence alignment of this sequence fragment, *FAD2-2* derived from flax NL97 (EU660502) and the sequences of *FAD2a* in both scaffolds, indicated that one insertion, possibly caused by mis-assembly, was observed in scaffold 360 between base pair 441 and 477 (Figure S6), whereas a sequence fragment between 440 and 616 bp in scaffold 931 was missing (Figure S6). We speculated that the missing region in scaffold 931 was incorrectly removed during the assembly process. Thus the insertion and deletion in both scaffolds most likely caused mis-annotation of the *FAD2a* genes in the original report [57].

The plant *FAD2* genes have no introns as reported in flax [19,20], *Arabidopsis* [77], *Brassica rapa* [83] and *Brassica napus* [84]; however, a unique intron existing in the 5'UTR was reported in *Arabidopsis* *FAD2*-At3G12120 [77], sesame *SeFAD2* [85], cotton *FAD2-4* [86], camelina *CsFAD2* [87], grape *VlFAD2* [80] and safflower *FAD2* [81]. A recent comprehensive study indicated that these highly conserved introns in multicellular plants, preferentially located within the 5'UTR, can function via the intron-mediated enhancement (IME) mechanism to enhance gene expression level in *Arabidopsis* [88] and other plants [89-92]. This unique feature of the gene has been widely utilized in wet lab gene reconstruction for transformation, and an IME prediction tool specifically designed for rice and *Arabidopsis* named IMEter [92] was developed. We verified the presence of 5'UTR in *FAD2* genes by using the public flax EST data set from GenBank [59] for alignment with flax *FAD2* genes. The alignments confirm that *FAD2a-1, FAD2a-2* and
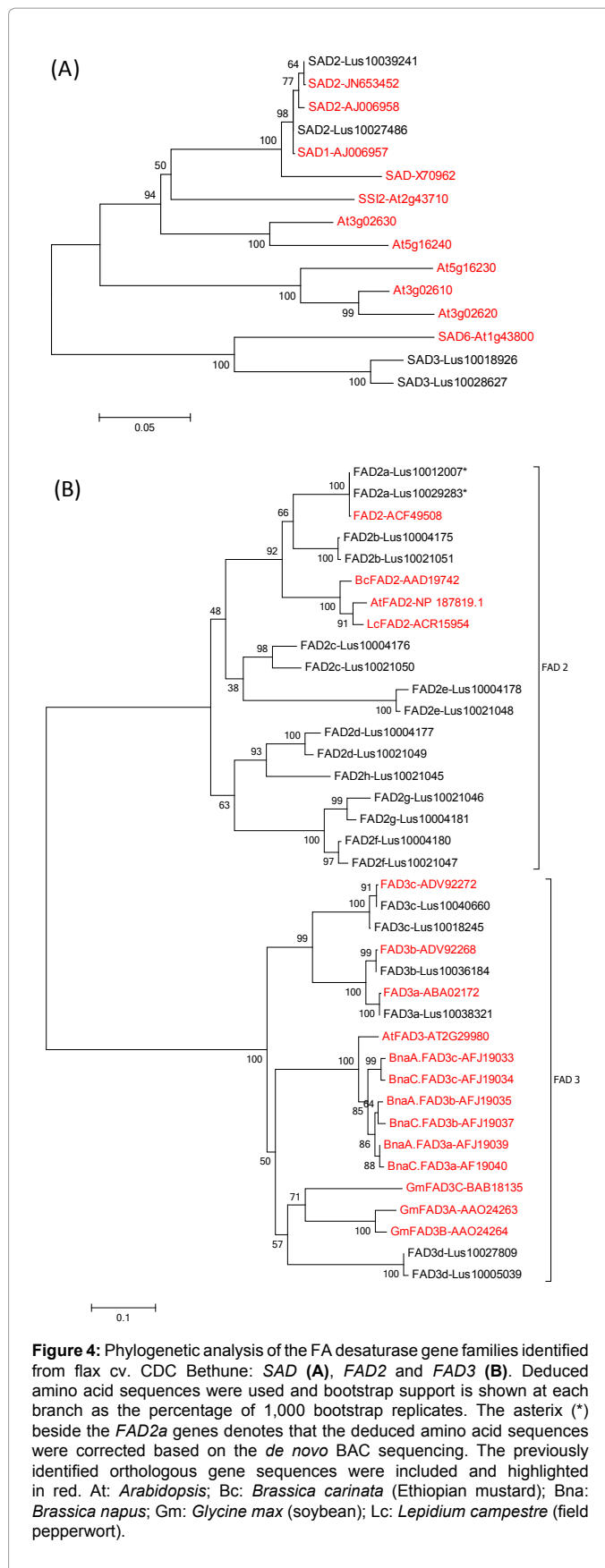

Figure 4: Phylogenetic analysis of the FA desaturase gene families identified from flax cv. CDC Bethune: *SAD* **(A)**, *FAD2* and *FAD3* **(B)**. Deduced amino acid sequences were used and bootstrap support is shown at each branch as the percentage of 1,000 bootstrap replicates. The asterix (*) beside the *FAD2a* genes denotes that the deduced amino acid sequences were corrected based on the *de novo* BAC sequencing. The previously identified orthologous gene sequences were included and highlighted in red. At: *Arabidopsis*; Bc: *Brassica carinata* (Ethiopian mustard); Bna: *Brassica napus*; Gm: *Glycine max* (soybean); Lc: *Lepidium campestre* (field pepperwort).

FAD2b-2 carry a 5'UTR intron and a transcription start site (Table 4). The remaining *FAD2* genes had no EST hits matching upstream of their 5' CDS, possibly as a result of the developmental stage specific *FAD2* gene expression, the relatively low EST sequencing coverage or simply because they are not expressed. Thus the flax *FAD2* genes expressed in developing seeds have the specific intron at their 5'UTR, which may regulate their tissue specificity and temporal expression patterns.

As the most important component of the catalytic center [27], three conserved HIS-boxes were observed within the 15 *FAD2* sequences (Figure S5). The amino acid alignment of the *FAD2* genes revealed that two genes, *FAD2d* (Lus10021049) and *FAD2g* (Lus10004181), have a single mutation causing a histidine to glutamine or asparagine mutation, respectively, and where both were in the first HIS-box. Mutations of histidine residues within the conserved motifs alter the desaturase functionality [93], hence, FAD2d and FAD2g may not be functional desaturases.

FAD3 gene family

FAD3 is the third desaturase gene, which inserts the double bond at the Δ^{15} position resulting in the synthesis of ALA/LIN [94]. To date, three microsomal linseed *FAD3* genes namely *FAD3a*, *FAD3b* and *FAD3c* have been reported [16,22]. *FAD3a* and *FAD3b* were identified from flax cv. Normandy [22] and *FAD3c* was recently identified from flax cultivars AC McDuff, UGG5-5 and SP2047 [16]. Each of these *FAD3* genes carried three conserved HIS-boxes that are required for the activity of *FAD3* [95], as well as the conserved dilysine sequence (-KKXX- or -KXKXX-) at the C-terminal end that is involved in the subcellular localization [96]. Specific mutations in HIS-boxes were shown to cause loss of catalytic activity of *FAD3b* [97].

Our informatics analysis indicated that *FAD3a*, *FAD3b* and *FAD3c* are identical to three annotated genes Lus10038321, Lus10036184 and Lus10040660 in the flax WGS reference genome (Table 2, Figures 4B, S7 and S8). *FAD3a* and *FAD3b* turned out to be a pair of duplicated genes. Thus it is not surprising that *FAD3a* and *FAD3b* were simultaneously observed as the dominant contributors of accumulation of ALA during seed development among the three *FAD3* genes [16]. Also, *FAD3a* and *FAD3b* had the highest numbers of mutations (SNPs and indels) compared to *SAD* and *FAD2* genes as previously surveyed in 120 flax accessions [28].

In addition to the three previously identified *FAD3* genes, we found three more, including one duplicated copy of *FAD3c*. To avoid confusion with previous nomenclature, we re-designated the previously identified *FAD3c* as *FAD3c-1*, and the newly identified paralog as *FAD3c-2*. The remaining two *FAD3* genes were also duplicated genes that we denoted as *FAD3d-1* and *FAD3d-2* (Table 2, Figure 4B). Although six of the
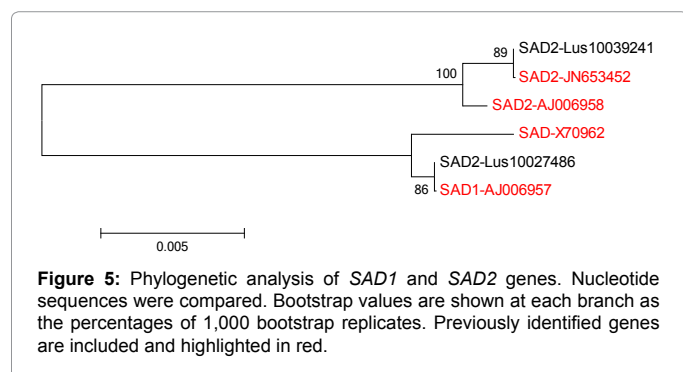
FAD3 genes harbored the conserved HIS-boxes (Figure S8), protein similarity indicated that both *FAD3d* genes diverge substantially from *FAD3a*, *b* and *c*, especially at the N-terminus (Figure S8). As a result, they clustered into a distinct clade from the *FAD3a-c* sub-clades (Figure 4B). On the other hand, the conserved dilysine motif at the C-terminus of *FAD3d* seems to have been lost during evolution. No report has been found regarding *FAD3d*, but extra amino acid insertions were detected in the N-terminus compared to *FAD3a-c* (Figure S8). Our analysis also indicated that there was a PEST-like motif in the N-terminus of the *FAD3* gene family that was identified by employing the Emboss tool ePESTfind (http://emboss.bioinformatics.nl/cgi-bin/emboss/epestfind). The PEST enriched motif, functioning as a *cis*-acting signal of protein degradation, was largely responsible for protein turn-over [98]. A potential PEST-like motif in the N-terminus of FAD3c as well as a poorer candidate motif in the N-terminus of FAD3a and FAD3b were identified (Figure S8). O'Quin et al. [99] found that the *cis*-acting degradation signal associated ubiquitin-proteasomal pathway could result in the increased half-life of *Brassica napus* FAD3 protein at cooler temperatures, indicating that this N-terminal motif may contribute to improved adaptive capability to abiotic stresses.

Although *SAD*, *FAD2* and *FAD3* belong to the same class of the FA desaturases, they differed in substrate specificity, gene structure (Figure 2 and Table 2) and functional motifs. Their divergence is ancient, approximately 124-128 MYA (Table S2).

KCS gene family

To date, no intact *KCS* gene has been successfully cloned from flax but two ESTs (CD760578 and CD760581) were reported to be partial transcripts of a KCS [18]. Based on the sequence similarity to the previously identified *FAE1*-like *KCS* genes from *Arabidopsis* and soybean [38-40,100], 38 flax KCS genes were identified (Table 5). Thirteen genes appeared as paired paralogs evolved from recent genome or segmental duplication events (Table 2). Four genes *KCS3*, *KCS11*, *KCS12*, and *KCS13* had three paralogous members according to their divergence time and duplication events. For example, the three members of *KCS13* (*KCS13-1*, *KCS13-2* and *KCS13-3*) were hypothesized to be the result of a tandem duplication event whereas *KCS3* likely encountered two separate duplication events: a tandem duplication and a genome or fragmental duplication. Only two genes, *KCS10* and *KCS15*, had only one member. All gene pairs or triplets were attributed to recent duplication events (Table 5). These *KCS* genes were scattered on 11 different linkage groups or chromosomes (LG1, 3, 4, 5, 6, 7, 8, 10, 11, 12 and 15) and several unanchored scaffolds (Table 5).

Phylogenetic analysis showed that the 38 *KCS* genes clustered into seven groups (I-VII) (Figure 6). Flax *KCS* genes of six of the groups clustered with at least one orthologous *KCS* gene from *Arabidopsis* or soybean. Group I included six pairs of genes and one triplet with one to three exons. Group II contained three genes of a triplet and one single gene with three or four exons. This group corresponds to the ε subclass of the *KCS* genes in *Arabidopsis* that also have three exons [42]. Group III, IV and VI, corresponding to the α, γ, and θ subclass of the *Arabidopsis KCS* genes, included four (two pairs), two (one pair) and three (one triplet) genes, respectively, none of which had introns. Group III clustered with *AtFAE1/AtKCS18* which was the first *KCS* gene identified in *Arabidopsis* [38,39]. Group V had three genes (a triplet) that did not cluster with any other orthologous genes. Group VII consisted of eight genes (two pairs, one triplet and one single) with 1-4 exons that clustered with the θ subclass of the *Arabidopsis KCS* genes which was divided into two subgroups in flax (VI and VII). To

Figure 5: Phylogenetic analysis of *SAD1* and *SAD2* genes. Nucleotide sequences were compared. Bootstrap values are shown at each branch as the percentages of 1,000 bootstrap replicates. Previously identified genes are included and highlighted in red.

No	Clade	Gene	Gene ID	Chromosome location	Start position in scaffold	Genomic sequence length (bp)	CDS sequence length (bp)	Amino acid sequence length aa	No of exons	CDS identity (%)	Amino acid identity (%)	Duplication time (MYA)	Ks	Catalytic site C248	H416	N449	S307
1		*KCS1-1*	Lus10006637	Scaffold244 (LG8)	250465	2,499	1,623	540	2	76.95	72.4	13.7	0.178	+	+	+	+
2		*KCS1-2*	Lus10039401	Scaffold33 (LG10)	1236391	2,321	1,308	435	3					+	+	+	+
3		*KCS2-1*	Lus10021895	Scaffold164 (LG3)	494574	3,092	1,425	474	2	85.57	85.79	7.1	0.092	+	+	+	+
4		*KCS2-2*	Lus10041177	Scaffold280 (LG15)	1287373	3,083	1,584	527	2					+	+	+	+
5		*KCS3-1*	Lus10019446	Scaffold906 (LG6)	285487	1,551	1,551	516	1	97.73	98.44	1.9	0.024	+	+	+	+
6		*KCS3-2*	Lus10043300	Scaffold25 (LG12)	2296451	1,551	1,551	516	1					+	+	+	+
7	I	*KCS3-3*	Lus10019448	Scaffold906 (LG6)	290816	1,343	1,281	426	2	89.87	83.13	1.7	0.022		+	+	+
8		*KCS4-1*	Lus10002691	Scaffold1347 (LG15)	106658	1,748	1,572	523	2	92.93	95.21	14.1	0.183	+	+	+	+
9		*KCS4-2*	Lus10030209	Scaffold217 (LG4)	231412	1,893	1,572	523	2					+	+	+	+
10		*KCS5-1*	Lus10010108	Scaffold722 (LG1)	258593	1,575	1,575	524	1	96.06	98.28	4.9	0.063	+	+	+	+
11		*KCS5-2*	Lus10012611	Scaffold101 (LG15)	126220	1,575	1,575	524	1					+	+	+	+
12		*KCS9-1*	Lus10006636	Scaffold244 (LG8)	241675	1,974	1,482	493	2	81.37	77.51	23.3	0.302	+	+	+	T
13		*KCS9-2*	Lus10039399	Scaffold33 (LG10)	1219216	1,811	1,524	507	2					+	+	+	T
14		*KCS10*	Lus10028105	scaffold132 (LG5)	1068330	1,678	1,518	505	3	88.50	90			+	L	+	+
15	II (ε)	*KCS11-1*	Lus10016528	scaffold279 (LG12)	371666	2,781	1,587	529	3	96.84	97.73	6.0	0.078	+	+	+	+
16		*KCS11-2*	Lus10040796	scaffold156 (LG3)	1494523	3,052	1,587	528	3					+	+	+	+
17		*KCS11-3*	Lus10009799	scaffold1	186142	2,565	1,527	508	4	88.50	90	22.6	0.293	+	+	+	+
18		*KCS6-1*	Lus10001657	scaffold227 (LG8)	14836	1,512	1,512	503	1	96.75	97.21	4.2	0.055	+	+	+	+
19	III (α)	*KCS6-2*	Lus10002533	scaffold81 (LG6)	60705	1,512	1,512	503	1					+	+	+	+
20		*KCS7-1*	Lus10026345	scaffold898 (LG11)	718114	1,551	1,551	516	1	95.35	98.64	5.2	0.067	+	+	+	+
21		*KCS7-2*	Lus10042318	scaffold123 (LG1)	1671369	1,548	1,548	515	1					+	+	+	+
22	IV (γ)	*KCS8-1*	Lus10034319	scaffold310 (LG1)	596075	1,491	1,491	496	1	95.7	98.79	6.3	0.082	+	+	+	+
23		*KCS8-2*	Lus10041452	scaffold272 (LG4)	142994	1,491	1,491	496	1					+	+	+	+
24		*KCS12-1*	Lus10020662	scaffold303 (LG4)	74137	1,350	1,350	449	1	65.72	62.3	36.0	0.469	+	+	+	+
25	V	*KCS12-2*	Lus10031486	scaffold863	523726	1,248	1,248	415	1					+	+	+	+
26		*KCS12-3*	Lus10029880	scaffold416	487616	1,362	1,362	453	1	65.72	52.35	23.4	0.304	R	+	+	+
27		*KCS16-1*	Lus10005434	scaffold847 (LG7)	119164	1,464	1,464	487	1	91.49	92.9	8.9	0.116	+	+	+	T
28	VI (θ)	*KCS16-2*	Lus10015222	scaffold924 (LG12)	18075	1,479	1,479	492	1					+	+	+	T
29		*KCS17-1*	Lus10004918	scaffold858 (LG8)	122399	1,239	1,239	412	1	63.45	62.83	6.3	0.082	+	+	+	T
30		*KCS17-2*	Lus10040873	scaffold156	1981104	1,188	1,188	395	1					+	+	+	T
31		*KCS13-1*	Lus10033626	scaffold701 (LG12)	874027	2,211	1,455	484	4	68.71	65.36	14.9	0.194	+	+	+	T
32		*KCS13-2*	Lus10033627	scaffold701 (LG12)	878209	1,722	1,122	373	3					+	+	+	T
33		*KCS13-3*	Lus10033625	scaffold701 (LG12)	867459	5,336	2,271	756	4	49.28	44.28	12.0	0.156	+	+	+	T
30	VII (θ)	*KCS14-1*	Lus10002191	scaffold1010	67381	1,401	1,401	466	1	76.57	76.48	11.0	0.144	+	+	+	T
31		*KCS14-2*	Lus10039906	scaffold15 (LG4)	2153982	3,627	1,725	574	2					+	+	+	T
32		*KCS15*	Lus10002192	scaffold1010	69983	1,254	1,254	417	1					+	+	Y	T
37		*KCS18-1*	Lus10023458	scaffold1216 (LG7)	288113	1,575	1,437	478	2	93.73	96.02	14.0	0.183	+	+	+	T
38		*KCS18-2*	Lus10040333	scaffold86	1404742	1,560	1,437	478	2					+	+	+	T

For genes with three duplicate members, the estimates of sequence identity, duplication time and *Ks* of the third member were calculated as the average of the first two members vs. the third one. Clade: the clades are determined according to the phylogenetic analysis of *KCS* genes (Figure 6). The clades in parentheses correspond to classification of *KCS* genes in *Arabidopsis* [42]. Catalytic site: +represents the presence of the residues of putative catalytic sites. The positions of residues, Cys248, His416, Asn449 and Ser307, are based on the sequence of *KCS1-1*.
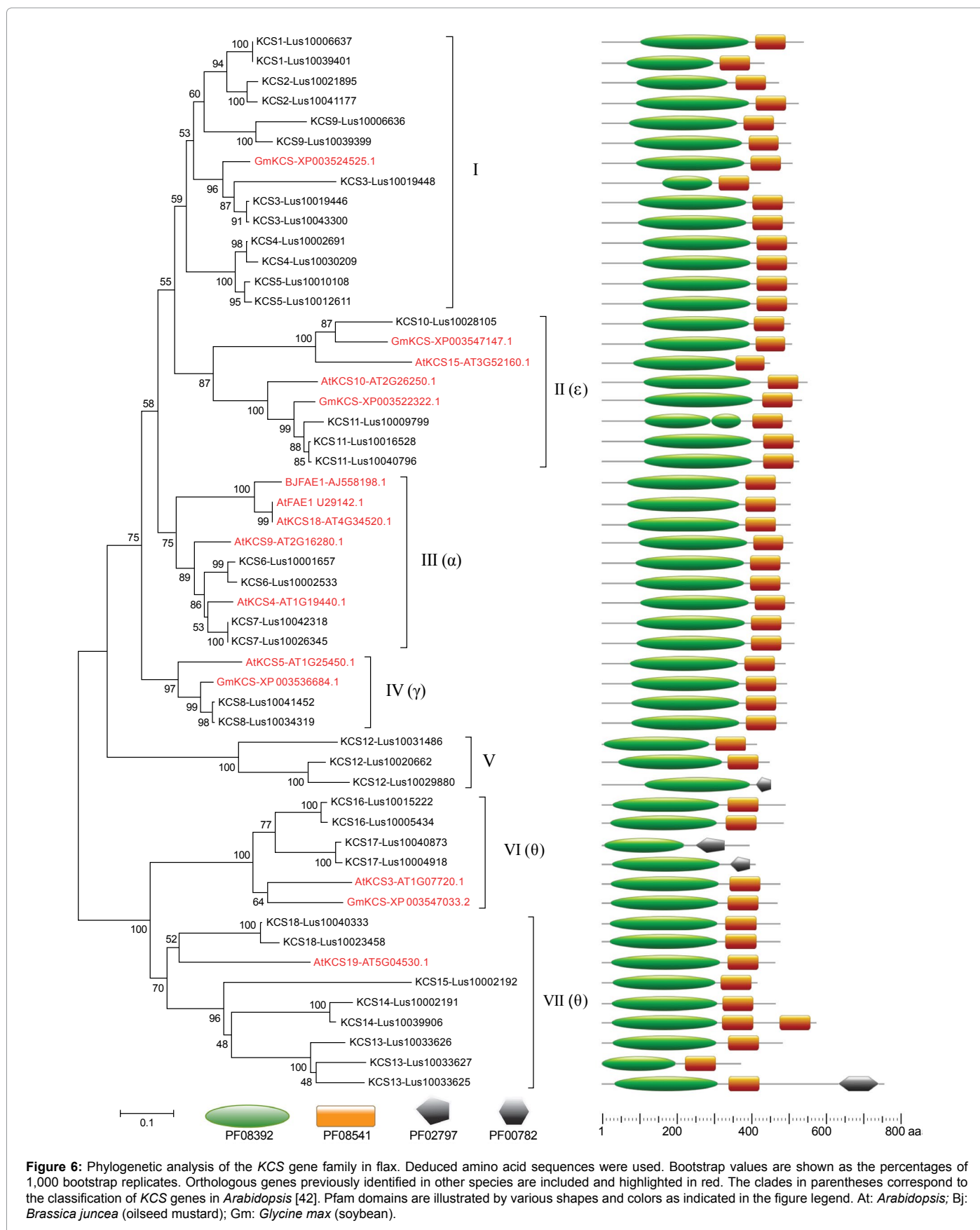
Table 5: *KCS* gene family related to the fatty acid elongation to VLCFA identified in flax cv. CDC Bethune.

compare the relationship of the seven groups of genes, we calculated the divergence time of genes among and within the groups (Table S3). We found that the *KCS* gene pairs were duplicated most recently (1.7-23.4 MYA) and hypothesized this to be to the result of a recent whole genome duplication event (Table 5). The between group gene duplications appeared to be more ancient, i.e., more than 100 MYA, which precedes the divergence time between *SAD*, *FAD2* and *FAD3*, between *FAD2a* and *FAD2b*, and between *KAS I*, *KAS II* and *KAS III* (Tables S1, S2 and S3). The amino sequence alignment of the predicted KCS enzymes demonstrated that most of regions of the KCS proteins were highly conserved (Figure S9).

All *KCS* genes displayed two common domains, PF08392 (FAE1/Type III polyketide synthase-like protein) and PF08541 (3-Oxoacyl-[acyl-carrier-protein (ACP)] synthase III C terminal) with the exception of KCS12-Lus10029880, KCS17-Lus10040873, and KCS17-

10004818) which had a PF08541 domain and an alternative domain PF02797 (chalcone and stilbene synthases, C-terminal domain) (Figure 6). KCS13-Lus10033625 had an additional domain PF00782 (protein tyrosine phosphatase). In addition, all *KCS* genes maintained a putative catalytic triad (Cys248-His416-Asn449 based on the sequence of *KCS1-1*) that plays a critical role in KCS catalysis [101-104] (Table 5, Figure S9). An additional putative active site (Ser307 of KCS1-1) has been demonstrated to be essential for the activity of the FAE1 enzyme in *Brassica napus* [105]. This residue was also conserved for all the *KCS* genes except two *KCS9* genes, all 12 genes in Group VI and VII, and all 3 *KCS* genes in the θ subclass of *Arabidopsis* which had a threonine instead of a serine at the same position (Table 5).

KCSs show homology with the known KAS III that can catalyze the reaction between malonyl-ACP and acyl-ACP to synthesize the C4 3-ketoacyl-ACP in *E. coli*, spinach and *Brassica napus* [106,107]. We

Figure 6: Phylogenetic analysis of the *KCS* gene family in flax. Deduced amino acid sequences were used. Bootstrap values are shown as the percentages of 1,000 bootstrap replicates. Orthologous genes previously identified in other species are included and highlighted in red. The clades in parentheses correspond to the classification of *KCS* genes in *Arabidopsis* [42]. Pfam domains are illustrated by various shapes and colors as indicated in the figure legend. At: *Arabidopsis;* Bj: *Brassica juncea* (oilseed mustard); Gm: *Glycine max* (soybean).

also observed amino acid sequence similarity between KCS and KAS III families in flax and other plant species (Figure S9). KCS and KAS III families are thought to have similar catalytic mechanisms [35], but their substrate specificity differs substantially. KCSs use malonyl-CoA as a substrate for decarboxylation in lieu of malonyl-ACP for KAS III. *KCS* and *KAS* genes clustered into two distinct groups (Figure S10) and *KAS III* was evolutionarily much closer to *KAS I* and *KAS II* than *KCS* in term of their p-distances (the number of amino acid differences per site) (Table S4).

FAT gene family

Two classes of *FAT* genes coding TEs, *FatA* and *FatB*, control the termination of FA chain extension in plants. Facciotti and Yuan [108] summarized features of FatA and FatB. Both classes share a conserved core region of ~210 amino acids featuring the protein family of TEs (Pfam: PF01643) and an ~60 amino acid transit peptide in the N-terminus. Two unique and short regions surrounding the catalytic sites differentiate FatA and FatB. Also, *FatB* genes have an additional conserved hydrophobic region [108].

We identified 14 flax *FAT* genes (Table 6). The similarity analysis based on the protein domain family database (Pfam 27.0) revealed that all 14 *FAT* genes shared a conserved region (Pfam: PF01643) characteristic of the TE family (Figure 7). The phylogenetic analysis grouped the 14 flax genes into four distinct clusters. Two pairs of duplicated genes (Lus10038190 vs. Lus10025912, and Lus10022772 vs. Lus10011839) formed the *FatA* group based on their evolutionary distance associated with the previously identified *FatA* genes [109] whereas another pair of flax duplicated genes (Lus10017751 vs. Lus10033072) constituted the *FatB* group. The Pfam analysis showed that the FatB proteins shared a similar region (Pfam: PF12590, Acyl-ATP thioesterase) in the N-terminus which is supposed to overlap with the conserved FatB hydrophobic region. However, four pairs of flax *FAT*-like genes (Lus10013480 vs. Lus10007942, Lus10034617 vs Lus10000365, Lus10035901 vs. Lus10025762, and Lus10035900 vs. Lus10025763) (Table 6), did not properly group into either FatA or FatB because of insufficient biochemical enzyme activity assay evidence (Figure 7). The amino acid sequence alignment (Figure S11) indicated that the two pairs of flax duplicated genes (Lus10013480 vs. Lus10007942 and Lus10034617 vs. Lus10000365) were closer to the

FatA class in the phylogenetic tree (Figure 7); however, two typically conserved regions (labeled A and B in Figure S11) featured in FatA were totally different from either FatA or FatB. An additional extended C-terminal tail featured in FatB was present in the four genes. They also lacked the PF12590 domain family of a typical FatA gene (Figure 7). Thus, we temporarily named these two flax gene pairs *Fat1* and *Fat2* and grouped them into FAT I (Table 6 and Figure 7).

The other two gene pairs (Lus10035901 vs. Lus10025762, and Lus10035900 vs. Lus10025763) had a less conserved hydrophobic region than other plant FatB proteins. In addition, they had two gaps in the N-terminal transit peptide. One pair of genes (Lus10035900 and Lus10025763) lacked the extended C-terminal tail. These two pairs of genes on LG 1 and LG 11 appeared to have arisen from a tandem duplication event and a whole genome/fragmental duplication event at around the same time (5.2-6.6 MYA), evolutionary speaking (Table 6 and Table S5). Thus we named these two flax gene pairs *Fat3* and *Fat4*, respectively, grouping them into Fat II (Table 6 and Figure 7).

All seven pairs of *FAT* genes were the result of more recent gene duplications (Table 6); however, four groups of genes appeared to be highly conserved in flax and other plant species. Flax *FatA* and *FatB* genes clustered with corresponding *FatA* or *FatB* genes from other species (Figure 7), indicative of their somewhat more ancient divergence approximately 40 MYA (Table S5). The genes in the Fat I and Fat II groups might be specific in flax because no orthologous *FAT* genes clustered into these two groups. Furthermore, they all maintained the two conserved catalytic sites of Cys275 and His310 (based on the sequence of *FatA1-1*) (Table 6 and Figure S11), indicating that they all may have the biological function to terminate the synthesis of FAs.

Differential transcription of the identified FA genes

The overall impact of a gene family is a combination of the size of the gene family, individual structural differences among family members and differential transcription within individual members [56]. Digital expression analysis (Figure 8) shows the number of ESTs assigned to all identified FA genes; however, because of the high similarity of the paired genes it was not always possible to unequivocally assign a specific EST to one member of a gene pair. Thus, except for *FAD3a* and *FAD3b*, ESTs were assigned to pairs of duplicated genes. We observed

No	Clade	Gene	Locus	Chromosome location	Start position in scaffold	Genomic sequence length (bp)	CDS sequence length (bp)	Amino sequence length (aa)	No. of exons	CDS identity (%)	Amino acid identity (%)	Duplication time (MYA)	Ks	Catalytic site	
														H275	C310
1	FatA	*FatA1-1*	Lus10038190	scaffold28 (LG11)	210541	2,873	1,125	370	8	84.71	79.94	6.2	0.08	+	+
2		*FatA1-2*	Lus10025912	scaffold605 (LG1)	914799	2,259	1,113	374	7					+	+
3		*FatA2-1*	Lus10022772	scaffold8 (LG10)	41937	2,422	1,131	376	7	96.64	95.74	2.7	0.035	+	+
4		*FatA2-2*	Lus10011839	scaffold754 (LG9)	29342	2,395	1,131	376	7					+	+
5	FatB	*FatB1-1*	Lus10017751	scaffold212 (LG6)	296331	2,015	1,257	418	6	97.93	98.80	0.5	0.006	+	+
6		*FatB1-2*	Lus10033072	scaffold306 (LG12)	739174	2,024	1,257	418	6					+	+
7	Fat I	*Fat1-1*	Lus10013480	scaffold230 (LG3)	386737	2,394	1,257	418	6	92.91	93.54	3.4	0.044	+	+
8		*Fat1-2*	Lus10007942	scaffold281 (LG3)	75576	2,338	1,251	416	6					+	+
9		*Fat2-1*	Lus10034617	scaffold9 (LG13)	987681	1,889	1,230	409	6	94.89	93.41	2.2	0.029	+	+
10		*Fat2-2*	Lus10000365	scaffold2160	5905	1,888	1,233	410	6					+	+
11	Fat II	*Fat3-1*	Lus10035901	scaffold76 (LG11)	200209	1,547	1,161	383	5	89.49	87.82	5.2	0.067	+	+
12		*Fat3-2*	Lus10025762	scaffold605 (LG1)	301503	1,444	1,152	386	4					+	+
13		*Fat4-1*	Lus10025763	scaffold605 (LG1)	304576	1,350	1,077	363	4	90.29	85.95	5.5	0.071	+	+
14		*Fat4-2*	Lus10035900	scaffold76 (LG11)	197570	1,357	1,092	358	4					+	+

Clade: the clades are determined according to the phylogenetic analysis of *FAT* genes (Figure 7). Catalytic site: + represents the presence of the residues of putative catalytic sites. The positions of residues, His275 and Cys310, are based on the sequence of *FatA1-1*.

Table 6: *FAT* gene family involved in the fatty acid chain termination identified in flax cv. CDC Bethune.
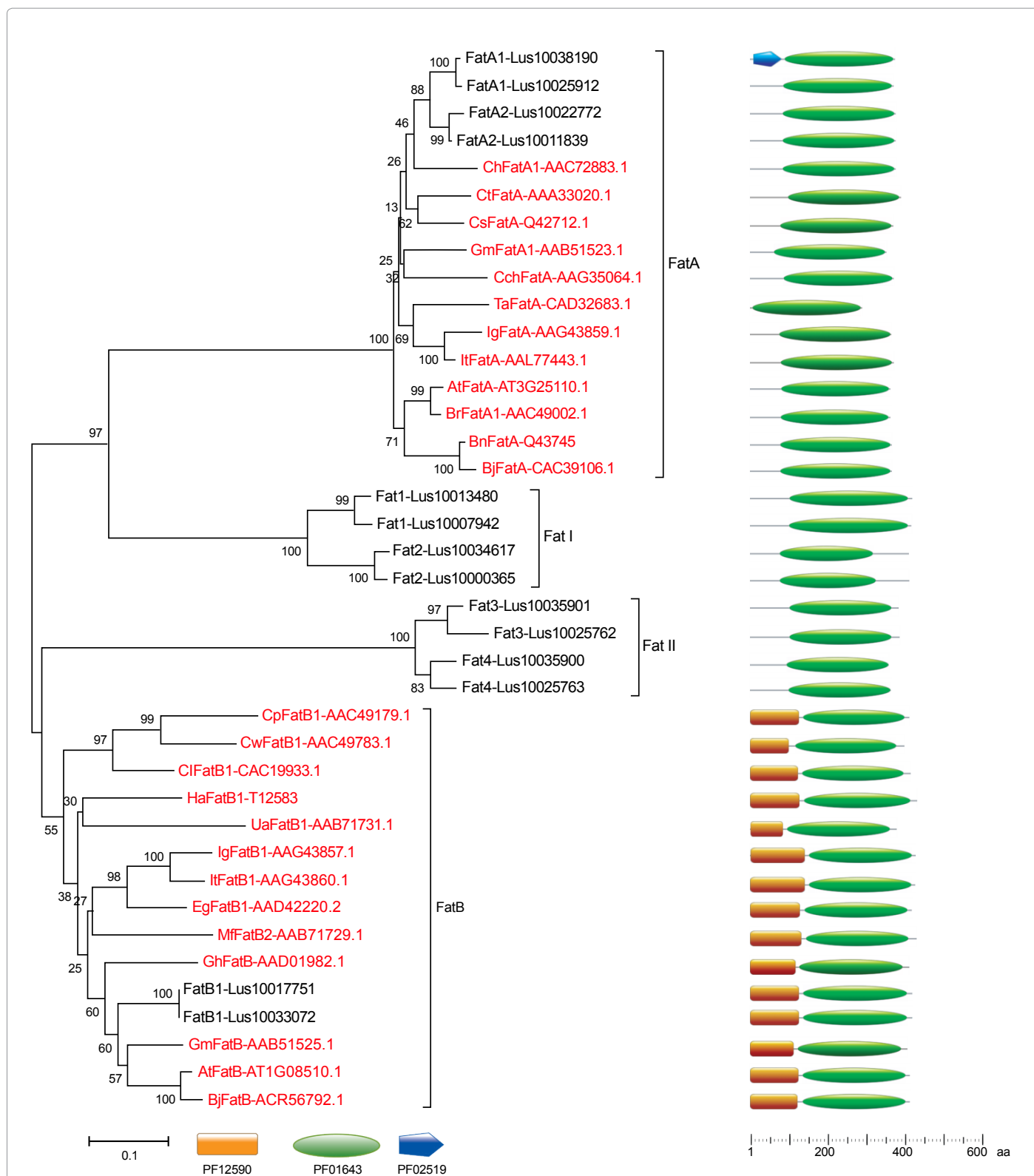
Figure 7: Phylogenetic analysis of the *FAT* gene family in flax. Deduced amino acid sequences were used. Bootstrap values are shown as the percentages of 1,000 bootstrap replicates. Orthologous genes previously identified in other species are included and highlighted in red. At: *Arabidopsis thaliana*; Bj: *Bradyrhizobium japonicum*; Bn: *Brassica napus*; Cc: *Cinnamonum camphorum*; Cch: *Capsicum chinese*; Ch: *Cuphea hookeriana*; Cl: *Cuphea lanceolata*; Cp: *Cuphea palustris*; Cs: *Coriandrum sativum*; Ct: *Carthamus tinctorius*; Cw: *Cuphea wrightii*; Eg: *Elaeis guineensis*; Gh: *Gossypium hirsutum*; Gm: *Garcinia mangostana*; Ha: *Helianthus annuus*; Ig: *Iris germanica*; It: *Iris tectorum*; Mf: *Myristica fragrans*; Ta: *Triticum aestivum*; Ua: *Ulmus americana*; Uc: *Umbellularia californica*.

large differences in the number of ESTs in different gene pairs in the *KAS* families (condensing genes) as well as in the *SAD/FAD2/FAD3* families (desaturation genes) (Figure 8A). Independent Chi-square tests to assess the statistical significance of distribution of EST counts corresponding to condensing genes and desaturation genes, and, their comparison between the two separate categories of genes, were performed. The results showed that differences between EST hits within desaturation genes and within condensing genes as well as between these two categories were not derived by chance ($p<0.01$). Desaturation genes had ten-fold more ESTs than condensing genes, suggesting that the higher expression level of desaturation genes play a major role with the respect to the biosynthesis of unsaturated FAs in flax.

In terms of condensing genes, the *KAS Ia* and *KAS II* gene pairs had more EST hits than other genes which came mostly from torpedo embryo, endosperm and torpedo stage seed coat tissues (Figure 8A). We failed to identify any known EST associated with the *mtKAS* and *KAS IIIa* gene pairs. Within desaturation genes, the *FAD2* gene pairs had the most abundant EST hits followed by *FAD3* and *SAD* (Figure 8). The ESTs associated with these genes were mostly derived from the mature embryo where OLE, LIO and LIN are synthesized (Figure 8A). Within the *FAD2* gene family, *FAD2b* showed more EST hits than *FAD2a* while other gene pairs (*FAD2c-h*) had few ESTs associated with them. Within the *FAD3* gene family, the *FAD3a/b* pair had more EST hits than other members (Figure 8A), consistent with previous reports [16,97]. Using real time (RT) PCR, Banik et al. [16] quantified the expression levels of the *FAD3a, b* and *c* genes and reported that

the expression of *FAD3a* and *FAD3b* transcripts was modulated during seed development, however *FAD3c* expression remained low throughout these stages. Radovanovic et al. [97] measured the conversion rate of OLE into LIO and LIO into LIN by heterologous expression of FAD2 and FAD3 isoforms in yeast and showed that the conversion rate of FAD2 exceeded that of FAD3 and, that FAD2b had a 10% higher conversion rate than FAD2a.
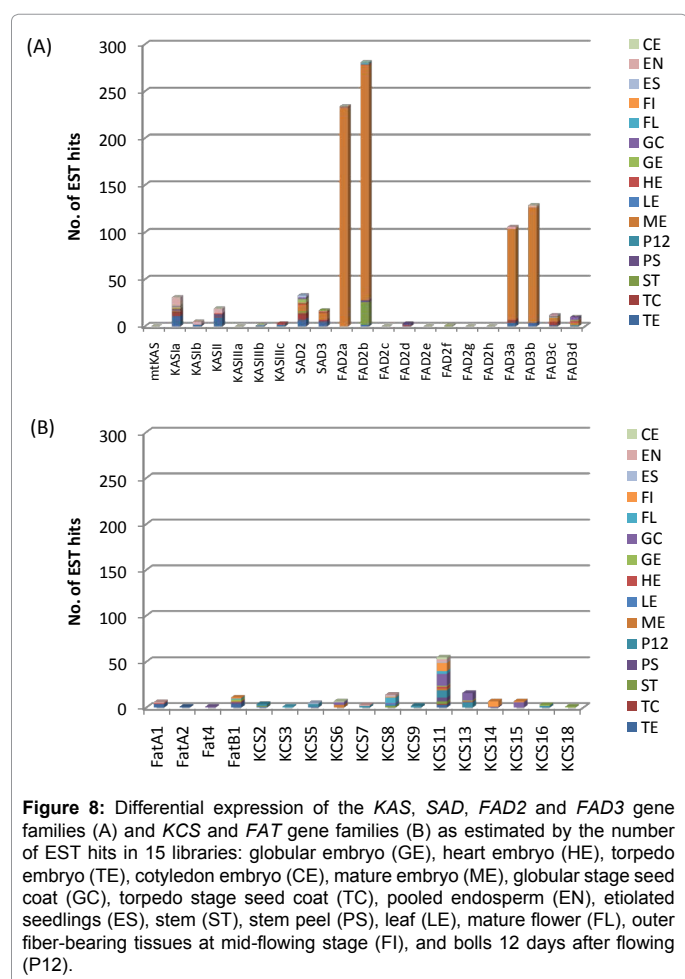
FAD2c-h and *FAD3d* genes were only slightly expressed as quantified by their EST hits (Figure 6). It is presumed that, during evolution, these genes have lost their functions or have acquired neofunctionalization through their gene structure change after duplication. For instance, among *FAD2c-h* gene members, no 5' UTR intron was present whereas the more recently formed genes had the feature, which may help plants cope with climate change associated stress [88]. *FAD3d* also lost its conserved dilysine motif at the C-terminus (Figure S8). Similarly, several condensing genes (*KAS Ib*, and *KAS IIIa*) with very few EST hits are relatively older duplicated copies (10.8-16.3 MYA) than *KAS Ia* and *KAS II* which were more highly expressed and were also more recently duplicated (1.5-9.2 MYA) (Table 1 and Figure 6).

Our results and previous studies [16,97] confirmed that the three pairs of duplicated genes *FAD2a-1/FAD2a-2*, *FAD2b-1/FAD2b-2* and *FAD3a/FAD3b* are highly expressed and play key roles in the FA profile of flax, indicating that these most recently duplicated copies may co-express and have additive effects to improve phenotypic performance as reported in soybean [110].

KCS is the largest gene family identified in this research with 38 genes. We observed that 13 pairs of *KCS* displayed EST hits; however, only *KCS11* showed significant higher digital expression level than other *KCS* genes and *FAT* genes (Figure 8B). The hit ESTs were mostly derived from seed coat (GC), bolls (P12), and fiber-bearing tissues (FI). *KCS11* clustered with *AtKCS10* in *Arabidopsis* which is required for normal development of the epidermis [111] and, is expressed in all tissues except for root, with the highest expression level in stems and siliques [42]. For *FAT* genes, no significant expression from EST hits was observed (Figure 8B).

Concluding Remarks

Through *in silico* gene mining of the WGS reference sequence of cv. CDC Bethune, we identified 84 new and validated seven previously cloned flax genes hypothesized to be involved in FA elongation, desaturation and the termination of FA chain elongation in flax and belonging to the following gene families: *KAS*, *SAD*, *FAD2*, *FAD3*, *KCS* and *FAT*. Fourteen β-ketoacyl-ACP synthases reported here include one pair of mitochondria targeting *mtKAS*, two pairs of *KAS I*, one pair of *KAS II* and three pairs of *KAS III*. These synthase enzymes are involved in the stepwise elongation of FAs to form 18-carbon polyunsaturated FAs. *SAD*, *FAD2* and *FAD3* are three gene families encoding desaturases responsible for insertion of double bonds at the Δ^9, Δ^{12} and Δ^{15}, respectively, to ultimately enrich flax seeds in LIO and LIN. Apart from seven genes cloned previously, eighteen desaturase genes were newly identified in the form of gene pairs with the exception of *FAD2h* which had no duplicated copy. As the largest gene family in this study, the 38 *KCS* genes identified here represented one of the elongase enzymes involved in the extension of FA chains from C18 to VLCFA. Furthermore 14 *FAT* genes, another important class of FA genes responsible for the termination of the FA chain elongation were also described. The majority of the identified FA synthesis genes are duplicated gene pairs caused by recent whole genome or fragmental duplication events [82], but the six gene families are highly conserved

Figure 8: Differential expression of the *KAS*, *SAD*, *FAD2* and *FAD3* gene families (A) and *KCS* and *FAT* gene families (B) as estimated by the number of EST hits in 15 libraries: globular embryo (GE), heart embryo (HE), torpedo embryo (TE), cotyledon embryo (CE), mature embryo (ME), globular stage seed coat (GC), torpedo stage seed coat (TC), pooled endosperm (EN), etiolated seedlings (ES), stem (ST), stem peel (PS), leaf (LE), mature flower (FL), outer fiber-bearing tissues at mid-flowing stage (FI), and bolls 12 days after flowing (P12).

in flax and other plants. They were hypothesised to have diverged anciently.

The new flax FA genes were identified from a single flax cultivar (CDC Bethune), and more FA related genes may be discovered upon investigation of a diverse flax germplasm [75]. Such investigation is practically feasible given the recent advances in next generation sequencing. Although digital prediction of expression patterns for the identified genes were made based on the flax ESTs developed from thirteen libraries [59], gene expression analysis by the RT-PCR method is expected to enhance the validation of our findings regarding the contribution of gene families to FA biosynthesis [16]. Together, these efforts will generate essential knowledge and provide useful genomic resources for further gene cloning, characterization, marker development and marker assisted selection in flax breeding. The CDS and amino acid sequences of FA related genes identified in this research are available as supplementary files (Supplementary file 2 and 3).

Acknowledgement

Supplementary files

1. Supplementary file 1: Figures S1-S11.

2. Supplementary file 2: Tables S1-S5.

3. Supplementary file 3: CDS sequences of FA biosynthesis related genes identified in flax cv. CDC Bethune.

4. Supplementary file 4: Amino acid sequences of FA biosynthesis related genes identified in flax cv. CDC Bethune.

References

1. Zohary D (1999) Monophyletic vs. polyphyletic origin of the crops on which agriculture was founded in the Near East. Genet Resour Crop Evol 46: 133-142.

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

3. Muir A, Westcott N (2003) Flax: the genus Linum. Taylor & Francis, London, UK.

4. Simopoulos AP, Leaf A, Salem N Jr. (1999) Essentiality of and recommended dietary intakes for omega-6 and omega-3 fatty acids. Ann Nutr Metab 43: 127-130.

5. Fofana B, Ragupathy R, Cloutier S (2010) Flax lipids: classes, biosynthesis, genetics and the promise of applied genomics for understanding and altering of fatty acids. In: Gilmore PL (Ed) (2011) Lipids: Categories, biological functions and metabolism, nutrition, and health. Nova Science Publishers Inc, New York.

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

7. Leyva DR, Zahradka P, Ramjiawan B, Guzman R, Aliani M, et al. (2011) The effect of dietary flaxseed on improving symptoms of cardiovascular disease in patients with peripheral artery disease: Rationale and design of the FLAX-PAD randomized controlled trial. Contemp Clin Trials 32: 724-730.

8. Kim Y, Ilich JZ (2011) Implications of dietary alpha-linolenic acid in bone health. Nutrition 27: 1101-1107.

9. Touré A, Xueming X (2010) Flaxseed lignans: source, biosynthesis, metabolism, antioxidant activity, bio-active components, and health benefits. Comp Rev Food Sci Food Saftey 9: 261-269.

10. Woods VB, Fearon AM (2009) Dietary sources of unsaturated fatty acids for animals and their transfer into meat, milk and eggs: A review. Livest Sci 126: 1-20.

11. Baud S, Lepiniec L (2010) Physiological and developmental regulation of seed oil production. Prog Lipid Res 49: 235-249.

12. Knutzon DS, Thompson GA, Radke SE, Johnson WB, Knauf VC, et al. (1992) Modification of Brassica seed oil by antisense expression of a stearoyl-acyl carrier protein desaturase gene. Proc Natl Acad Sci USA 89: 2624-2628.

13. Mikkilineni V, Rocheford TR (2003) Sequence variation and genomic organization of fatty acid desaturase-2 (fad2) and fatty acid desaturase-6 (fad6) cDNAs in maize. Theor Appl Genet 106: 1326-1332.

14. Warude D, Joshi K, Harsulkar A (2006) Polyunsaturated fatty acids: biotechnology. Crit Rev Biotechnol 26: 83-93.

15. Shanklin J, Cahoon EB (1998) Desaturation and related modifications of fatty acids1. Annu Rev Plant Physiol Plant Mol Biol 49: 611-641.

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

17. Fofana B, Cloutier S, Duguid S, Ching J, Rampitsch C (2006) Gene expression of stearoyl-ACP desaturase and delta12 fatty acid desaturase 2 is modulated during seed development of flax (*Linum usitatissimum*). Lipids 41: 705-712.

18. Fofana B, Duguid S, Cloutier S (2004) Cloning of fatty acid biosynthetic genes β-ketoacyl CoA synthase, fatty acid elongase, stearoyl-ACP desaturase, and fatty acid desaturase and analysis of expression in the early developmental stages of flax (*Linum usitatissimum* L.) seeds. Plant Sci 166: 1487-1496.

19. Khadake RM, Ranjekar PK, Harsulkar AM (2009) Cloning of a novel omega-6 desaturase from flax (*Linum usitatissimum* L.) and its functional analysis in *Saccharomyces cerevisiae*. Mol Biotechnol 42: 168-174.

20. Krasowska A, Dziadkowiec D, Polinceusz A, Plonka A, Lukaszewicz M (2007) Cloning of flax oleic fatty acid desaturase and its expression in yeast. J Am Oil Chem Soc 84: 809-816.

21. Sørensen BM, Furukawa-Stoffer TL, Marshall KS, Page EK, Mir Z, et al. (2005) Storage lipid accumulation and acyltransferase action in developing flaxseed. Lipids 40: 1043-1049.

22. Vrinten P, Hu Z, Munchinsky MA, Rowland G, Qiu X (2005) Two FAD3 desaturase genes control the level of linolenic acid in flax seed. Plant Physiol 139: 79-87.

23. Ohlrogge JB, Jaworski JG (1997) Regulation of fatty acid synthesis. Annu Rev Plant Physiol Plant Mol Biol 48: 109-136.

24. Jain RK, Thomson RG, Taylor DC, MacKenzie SL, McHughen A, et al. (1999) Isolation and characterization of two promoters from linseed for genetic engineering. Crop Sci 39: 1696-1701.

25. Singh S, McKinney S, Green A (1994) Sequence of a cDNA from *Linum usitatissimum* encoding the stearoyl-acyl carrier protein desaturase. Plant Physiol 104: 1075.

26. Shanklin J, Somerville C (1991) Stearoyl-acyl-carrier-protein desaturase from higher plants is structurally unrelated to the animal and fungal homologs. Proc Natl Acad Sci USA 88: 2510-2514.

27. Los DA, Murata N (1998) Structure and expression of fatty acid desaturases. Biochim Biophys Acta 1394: 3-15.

28. Thambugala D, Duguid S, Loewen E, Rowland G, Booker H, et al. (2013) Genetic variation of six desaturase genes in flax and their impact on fatty acid composition. Theor Appl Genet 126: 2627-2641.

29. Beisson F, Koo AJ, Ruuska S, Schwender J, Pollard M, et al. (2003) *Arabidopsis* genes involved in acyl lipid metabolism. A 2003 census of the candidates, a study of the distribution of expressed sequence tags in organs, and a web-based database. Plant Physiol 132: 681-697.

30. Scheffler JA, Sharpe AG, Schmidt H, Sperling P, Parkin IAP, et al. (1997) Desaturase multigene families of *Brassica napus* arose through genome duplication. Theor Appl Genet 94: 583-591.

31. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, et al. (2010) Genome sequence of the palaeopolyploid soybean. Nature 463: 178-183.

32. Chen Z, Wang ML, Barkley NA, Pittman RN (2010) A simple allele-specific PCR assay for detecting FAD2 alleles in both A and B genomes of the cultivated peanut for high-oleate trait selection. Plant Mol Biol Rep 28: 542-548.

33. Chi X, Yang Q, Zhao F, Qin S, Yang Y, et al. (2008) Comparative analysis of fatty acid desaturases in cyanobacterial genomes. Comp Funct Genomics.

34. Lu Y, Chi X, Li Z, Yang Q, Li F, et al. (2010) Isolation and characterization of a stress-dependent plastidial delta12 fatty acid desaturase from the Antarctic microalga *Chlorella vulgaris* NJ-7. Lipids 45: 179-187.

35. Haslam TM, Kunst L (2013) Extending the story of very-long-chain fatty acid elongation. Plant Sci 210: 93-107.

36. Fehling E, Mukherjee KD (1991) Acyl-CoA elongase from a higher plant (*Lunaria annua*): metabolic intermediates of very-long-chain acyl-CoA products and substrate specificity. Biochim Biophys Acta 1082: 239-246.

37. Millar AA, Kunst L (1997) Very-long-chain fatty acid biosynthesis is controlled through the expression and specificity of the condensing enzyme. Plant J 12: 121-131.

38. Lemieux B, Miquel M, Somerville C, Browse J (1990) Mutants of *Arabidopsis* with alterations in seed lipid fatty acid composition. Theor Appl Genet 80: 234-240.

39. James DW Jr, Dooner HK (1990) Isolation of EMS-induced mutants in *Arabidopsis* altered in seed fatty acid composition. Theor Appl Genet 80: 241-245.

40. Kunst L, Taylor DC, Underhill EW (1992) Fatty acid elongation in developing seeds of *Arabidopsis thaliana*. Plant Physiol Biochem 30: 425-434.

41. Costaglioli P, Joubès J, Garcia C, Stef M, Arveiler B, et al. (2005) Profiling candidate genes involved in wax biosynthesis in *Arabidopsis thaliana* by microarray analysis. Biochim Biophys Acta 1734: 247-258.

42. Joubès J, Raffaele S, Bourdenx B, Garcia C, Laroche-Traineau J, et al. (2008) The VLCFA elongase gene family in *Arabidopsis thaliana*: phylogenetic analysis, 3D modelling and expression profiling. Plant Mol Biol 67: 547-566.

43. Wang K, Wang Z, Li F, Ye W, Wang J, et al. (2012) The draft genome of a diploid cotton *Gossypium raimondii*. Nat Genet 44: 1098-1103.

44. De Renobales M, Rogers L, Kolattukudy PE (1980) Involvement of a thioesterase in the production of short-chain fatty acids in the uropygial glands of mallard ducks (*Anas platyrhynchos*). Arch Biochem Biophys 205: 464-477.

45. Pollard MR, Anderson L, Fan C, Hawkins DJ, Davies HM (1991) A specific acyl-ACP thioesterase implicated in medium-chain fatty acid production in immature cotyledons of *Umbellularia californica*. Arch Biochem Biophys 284: 306-312.

46. Smith S (1980) Mechanism of chain length determination in biosynthesis of milk fatty acids. J Dairy Sci 63: 337-352.

47. Jones A, Davies HM, Voelker TA (1995) Palmitoyl-acyl carrier protein (ACP) thioesterase and the evolutionary origin of plant acyl-ACP thioesterases. Plant Cell 7: 359-371.

48. Ginalski K, Rychlewski L (2003) Detection of reliable and unexpected protein fold predictions using 3D-Jury. Nucleic Acids Res 31: 3291-3292.

49. Voelker T (1996) Plant acyl-ACP thioesterases: chain-length determining enzymes in plant fatty acid biosynthesis. Genet Eng (N Y) 18: 111-133.

50. Moreno-Pérez AJ, Venegas-Calerón M, Vaistij FE, Salas JJ, Larson TR, et al. (2012) Reduced expression of FatA thioesterases in *Arabidopsis* affects the oil content and fatty acid composition of the seeds. Planta 235: 629-639.

51. Salas JJ, Ohlrogge JB (2002) Characterization of substrate specificity of plant FatA and FatB acyl-ACP thioesterases. Arch Biochem Biophys 403: 25-34.

52. Hobbs DH, Flintham JE, Hills MJ (2004) Genetic control of storage oil synthesis in seeds of *Arabidopsis*. Plant Physiol 136: 3341-3349.

53. Hellyer A, Leadlay PF, Slabas AR (1992) Induction, purification and characterisation of acyl-ACP thioesterase from developing seeds of oil seed rape (*Brassica napus*). Plant Mol Biol 20: 763-780.

54. Zheng P, Babar MD, Parthasarathy S, Gibson R, Parliament K, et al. (2014) A truncated FatB resulting from a single nucleotide insertion is responsible for reducing saturated fatty acids in maize seed oil. Theor Appl Genet 127: 1537-1547.

55. Anderson OD (2013) The B-hordein prolamin family of barley. Genome 56: 179-185.

56. Anderson OD, Huo N, Gu YQ (2013) The gene space in wheat: the complete γ-gliadin gene family from the wheat cultivar Chinese Spring. Funct Integr Genomics 13: 261-273.

57. Wang Z, Hobson N, Galindo L, Zhu S, Shi D, et al. (2012) The genome of flax (*Linum usitatissimum*) assembled de novo from short shotgun sequence reads. Plant J 72: 461-473.

58. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol 30: 2725-2729.

59. Venglat P, Xiang D, Qiu S, Stone SL, Tibiche C, et al. (2011) Gene expression analysis of flax seed development. BMC Plant Biol 11: 74.

60. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. Nat Genet 20: 43-45.

61. Gaut BS, Morton BR, McCaig BC, Clegg MT (1996) Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. Proc Natl Acad Sci USA 93: 10274-10279.

62. 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.

63. Cloutier S, Ragupathy R, Miranda E, Radovanovic N, Reimer E, et al. (2012) Integrated consensus genetic and physical maps of flax (*Linum usitatissimum* L.). Theor Appl Genet 125: 1783-1795.

64. Mikolajczyk S, Brody S (1990) De novo fatty acid synthesis mediated by acyl-carrier protein in Neurospora crassa mitochondria. Eur J Biochem 187: 431-437.

65. Wada H, Shintani D, Ohlrogge J (1997) Why do mitochondria synthesize fatty acids? Evidence for involvement in lipoic acid production. Proc Natl Acad Sci USA 94: 1591-1596.

66. Yasuno R, von Wettstein-Knowles P, Wada H (2004) Identification and molecular characterization of the β-ketoacyl-[acyl carrier protein] synthase component of the *Arabidopsis* mitochondrial fatty acid synthase. J Biol Chem 279: 8242-8251.

67. Zhang L, Joshi AK, Hofmann J, Schweizer E, Smith S (2005) Cloning, expression, and characterization of the human mitochondrial beta-ketoacyl synthase. Complementation of the yeast CEM1 knock-out strain. J Biol Chem 280: 12422-12429.

68. Christensen CE, Kragelund BB, von Wettstein-Knowles P, Henriksen A (2007) Structure of the human beta-ketoacyl [ACP] synthase from the mitochondrial type II fatty acid synthase. Protein Sci 16: 261-272.

69. Cronan JE, Fearnley IM, Walker JE (2005) Mammalian mitochondria contain a soluble acyl carrier protein. FEBS Lett 579: 4892-4896.

70. Li MJ, Li AQ, Xia H, Zhao CZ, Li CS, et al. (2009) Cloning and sequence analysis of putative type II fatty acid synthase genes from *Arachis hypogaea* L. J Biosci 34: 227-238.

71. von Wettstein-Knowles P, Olsen J, Arnvig Mcguire K, Larsen S (2000) Molecular aspects of beta-ketoacyl synthase (KAS) catalysis. Biochem Soc Trans 28: 601-607.

72. Zhang YM, Wu B, Zheng J, Rock CO (2003) Key residues responsible for acyl carrier protein and beta-ketoacyl-acyl carrier protein reductase (FabG) interaction. J Biol Chem 278: 52935-52943.

73. Abbadi A, Brummel M, Spener F (2000) Knockout of the regulatory site of 3-ketoacyl-ACP synthase III enhances short- and medium-chain acyl-ACP synthesis. Plant J 24: 1-9.

74. Hildebrand DF, Yu K, McCracken C, Rao SS (2005) Biology, utilisation and manipulation. In: Murphy DJ (Ed) Plant lipids: Biology, Utilisation and Manipulation. Blackwell Publishing.

75. Fu YB (2012) Population-based resequencing revealed an ancestral winter group of cultivated flax: implication for flax domestication processes. Ecol Evol 2: 622-635.

76. Emanuelsson O, Nielsen H, von Heijne G (1999) ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. Protein Sci 8: 978-984.

77. Okuley J, Lightner J, Feldmann K, Yadav N, Lark E, et al. (1994) Arabidopsis FAD2 gene encodes the enzyme that is essential for polyunsaturated lipid synthesis. Plant Cell 6: 147-158.

78. Cao YP, Shi JL, Li Z, Ming F (2010) Isolation of OsFAD2, OsFAD6 and FAD family members response to abiotic stresses in *Oryza sativa* L. Yi Chuan 32: 839-847.

79. Yang Q, Fan C, Guo Z, Qin J, Wu J, et al. (2012) Identification of FAD2 and FAD3 genes in *Brassica napus* genome and development of allele-specific markers for high oleic and low linolenic acid contents. Theor Appl Genet 125: 715-729.

80. Lee KR, Kim SH, Go YS, Jung SM, Roh KH, et al. (2012) Molecular cloning and functional analysis of two FAD2 genes from American grape (*Vitis labrusca* L.). Gene 509: 189-194.

81. Cao S, Zhou XR, Wood CC, Green AG, Singh SP, et al. (2013) A large and functionally diverse family of Fad2 genes in safflower (*Carthamus tinctorius* L.). BMC Plant Biol 13: 5.

82. Sveinsson S, McDill J, Wong GK, Li J, Li X, et al. (2014) Phylogenetic pinpointing of a paleopolyploidy event within the flax genus (Linum) using transcriptomics. Ann Bot 113: 753-761.

83. Tanhuanpää P, Vilkki J, Vihinen M (1998) Mapping and cloning of FAD2 gene to develop allele-specific PCR for oleic acid in spring turnip rape (*Brassica rapa* ssp. oleifera). Mol Breeding 4: 543-550.

84. Hu X, Sullivan-Gilbert M, Gupta M, Thompson SA (2006) Mapping of the loci controlling oleic and linolenic acid contents and development of fad2 and fad3 allele-specific markers in canola (*Brassica napus* L.). Theor Appl Genet 113: 497-507.

85. Kim MJ, Kim H, Shin JS, Chung CH, Ohlrogge JB, et al. (2006) Seed-specific expression of sesame microsomal oleic acid desaturase is controlled by combinatorial properties between negative cis-regulatory elements in the SeFAD2 promoter and enhancers in the 5'-UTR intron. Mol Genet Genomics 276: 351-368.

86. Zhang D, Pirtle IL, Park SJ, Nampaisansuk M, Neogi P, et al. (2009) Identification and expression of a new delta-12 fatty acid desaturase (FAD2-4) gene in upland cotton and its functional expression in yeast and *Arabidopsis* thaliana plants. Plant Physiol Biochem 47: 462-471.

87. Kang J, Snapp AR, Lu C (2011) Identification of three genes encoding microsomal oleate desaturases (FAD2) from the oilseed crop *Camelina sativa*. Plant Physiol Biochem 49: 223-229.

88. Parra G, Bradnam K, Rose AB, Korf I (2011) Comparative and functional analysis of intron-mediated enhancement signals reveals conserved features among plants. Nucleic Acids Res 39: 5328-5337.

89. Chung BY, Simons C, Firth AE, Brown CM, Hellens RP (2006) Effect of 5'UTR introns on gene expression in *Arabidopsis thaliana*. BMC Genomics 7: 120.

90. Schröpfer S, Kobbe D, Hartung F, Knoll A, Puchta H (2014) Defining the roles of the N-terminal region and the helicase activity of RECQ4A in DNA repair and homologous recombination in *Arabidopsis*. Nucleic Acids Res 42: 1684-1697.

91. Akua T, Shaul O (2013) The *Arabidopsis thaliana* MHX gene includes an intronic element that boosts translation when localized in a 5' UTR intron. J Exp Bot 64: 4255-4270.

92. Lopes-Caitar VS, de Carvalho MCCG, Darben LM, Kuwahara MK, Nepomuceno AL, et al. (2013) Genome-wide analysis of the Hsp20 gene family in soybean: comprehensive sequence, genomic organization and expression profile analysis under abiotic and biotic stresses. BMC Genomics 14: 577.

93. Shanklin J, Whittle E, Fox BG (1994) Eight histidine residues are catalytically essential in a membrane-associated iron enzyme, stearoyl-CoA desaturase, and are conserved in alkane hydroxylase and xylene monooxygenase. Biochemistry 33: 12787-12794.

94. Browse J, McConn M, James D Jr, Miquel M (1993) Mutants of *Arabidopsis* deficient in the synthesis of a-linolenate. Biochemical and genetic characterization of the endoplasmic reticulum linoleoyl desaturase. J Biol Chem 268: 16345-16351.

95. Avelange-Macherel MH, Macherel D, Wada H, Murata N (1995) Site-directed mutagenesis of histidine residues in the delta 12 acyl-lipid desaturase of Synechocystis. FEBS Lett 361: 111-114.

96. McCartney AW, Dyer JM, Dhanoa PK, Kim PK, Andrews DW, et al. (2004) Membrane-bound fatty acid desaturases are inserted co-translationally into the ER and contain different ER retrieval motifs at their carboxy termini. Plant J 37: 156-173.

97. Radovanovic N, Thambugala D, Duguid S, Loewen E, Cloutier S (2014) Functional characterization of flax fatty acid desaturase FAD2 and FAD3 isoforms expressed in yeast reveals a broad diversity in activity. Mol Biotechnol 56: 609-620.

98. Rechsteiner M, Rogers SW (1996) PEST sequences and regulation by proteolysis. Trends Biochem Sci 21: 267-271.

99. O'Quin JB, Bourassa L, Zhang D, Shockey JM, Gidda SK, et al. (2010) Temperature-sensitive post-translational regulation of plant omega-3 fatty-acid desaturases is mediated by the endoplasmic reticulum-associated degradation pathway. J Biol Chem 285: 21781-21796.

100. Schlüter PM, Xu S, Gagliardini V, Whittle E, Shanklin J, et al. (2011) Stearoyl-acyl carrier protein desaturases are associated with floral isolation in sexually deceptive orchids. Proc Natl Acad Sci USA 108: 5696-5701.

101. Ghanevati M, Jaworski JG (2001) Active-site residues of a plant membrane-bound fatty acid elongase beta-ketoacyl-CoA synthase, FAE1 KCS. Biochim Biophys Acta 1530: 77-85.

102. Ghanevati M, Jaworski JG (2002) Engineering and mechanistic studies of the *Arabidopsis* FAE1 beta-ketoacyl-CoA synthase, FAE1 KCS. Eur J Biochem 269: 3531-3539.

103. Lassner MW, Lardizabal K, Metz JG (1996) A jojoba beta-Ketoacyl-CoA synthase cDNA complements the canola fatty acid elongation mutation in transgenic plants. Plant Cell 8: 281-292.

104. Heath RJ, Rock CO (2002) The Claisen condensation in biology. Nat Prod Rep 19: 581-596.

105. Katavic V, Mietkiewska E, Barton DL, Giblin EM, Reed DW, et al. (2002) Restoring enzyme activity in nonfunctional low erucic acid *Brassica napus* fatty acid elongase 1 by a single amino acid substitution. Eur J Biochem 269: 5625-5631.

106. Tai H, Jaworski JG (1993) 3-Ketoacyl-acyl carrier protein synthase III from spinach (*Spinacia oleracea*) is not similar to other condensing enzymes of fatty acid synthase. Plant Physiol 103: 1361-1367.

107. Tsay JT, Oh W, Larson TJ, Jackowski S, Rock CO (1992) Isolation and characterization of the beta-ketoacyl-acyl carrier protein synthase III gene (fabH) from *Escherichia coli* K-12. J Biol Chem 267: 6807-6814.

108. Facciotti MT, Yuan L (1998) Molecular dissection of the plant acyl-acyl carrier protein thioesterases. Eur J Lipid Sci Tech 100: 167-172.

109. Mayer KM, Shanklin J (2007) Identification of amino acid residues involved in substrate specificity of plant acyl-ACP thioesterases using a bioinformatics-guided approach. BMC Plant Biol 7: 1.

110. Curtin SJ, Kantar MB, Yoon HW, Whaley AM, Schlueter JA, et al. (2012) Co-expression of soybean Dicer-like genes in response to stress and development. Funct Integr Genomics 12: 671-682.

111. Pruitt RE, Vielle-Calzada JP, Ploense SE, Grossniklaus U, Lolle SJ (2000) FIDDLEHEAD, a gene required to suppress epidermal cell interactions in *Arabidopsis*, encodes a putative lipid biosynthetic enzyme. Proc Natl Acad Sci USA 97: 1311-1316.