The developmental transcriptome atlas of the biofuel crop Camelina sativa

Kagale, Sateesh; Nixon, John; Khedikar, Yogendra; Pasha, Asher; Provart, Nicholas J.; Clarke, Wayne E.; Bollina, Venkatesh; Robinson, Stephen J.; Coutu, Cathy; Hegedus, Dwayne D.; Sharpe, Andrew G.; Parkin, Isobel A. P.

Publisher's version / Version de l'éditeur:
https://doi.org/10.1111/tpj.13302
The Plant Journal, 88, 5, pp. 879-894, 2016-12

the plant journal

SEB Society for Experimental Biology

The Plant Journal (2016) **88,** 879–894 doi: 10.1111/tpj.13302

RESOURCE

The developmental transcriptome atlas of the biofuel crop *Camelina sativa*

Sateesh Kagale[1,2], John Nixon[1], Yogendra Khedikar[1], Asher Pasha[3], Nicholas J. Provart[3], Wayne E. Clarke[1], Venkatesh Bollina[1], Stephen J. Robinson[1], Cathy Coutu[1], Dwayne D. Hegedus[1], Andrew G. Sharpe[2] and Isobel A. P. Parkin[1,*]

[1]*Agriculture and Agri-Food Canada, 107 Science Place, Saskatoon, SK, Canada,*
[2]*National Research Council Canada, 110 Gymnasium Place, Saskatoon, SK, Canada, and*
[3]*Department of Cell and Systems Biology, University of Toronto, Toronto, ON, Canada*

Received 21 September 2015; revised 1 August 2016; accepted 4 August 2016.
*For correspondence (e-mail isobel.parkin@agr.gc.ca).

SUMMARY

Camelina sativa **is currently being embraced as a viable industrial bio-platform crop due to a number of desirable agronomic attributes and the unique fatty acid profile of the seed oil that has applications for food, feed and biofuel. The recent completion of the reference genome sequence of *C. sativa* identified a young hexaploid genome. To complement this work, we have generated a genome-wide developmental transcriptome map by RNA sequencing of 12 different tissues covering major developmental stages during the life cycle of *C. sativa*. We have generated a digital atlas of this comprehensive transcriptome resource that enables interactive visualization of expression data through a searchable database of electronic fluorescent pictographs (eFP browser). An analysis of this dataset supported expression of 88% of the annotated genes in *C. sativa* and provided a global overview of the complex architecture of temporal and spatial gene expression patterns active during development. Conventional differential gene expression analysis combined with weighted gene expression network analysis uncovered similarities as well as differences in gene expression patterns between different tissues and identified tissue-specific genes and network modules. A high-quality census of transcription factors, analysis of alternative splicing and tissue-specific genome dominance provided insight into the transcriptional dynamics and sub-genome interplay among the well-preserved triplicated repertoire of homeologous loci. The comprehensive transcriptome atlas in combination with the reference genome sequence provides a powerful resource for genomics research which can be leveraged to identify functional associations between genes and understand the regulatory networks underlying developmental processes.**

Keywords: polyploidy, transcription factors, tissue specificity, gene expression, gene diversification, eFP browser, *Camelina sativa*.

INTRODUCTION

Camelina sativa, also known as false flax or gold-of-pleasure, is an oilseed crop from the family Brassicaceae that has been cultivated since prehistoric times. With the introduction of high-yielding rapeseed varieties during the middle of the 20th century, production of *C. sativa* had almost completely been eliminated. However, there is now renewed interest in developing *C. sativa* as a bio-platform crop, primarily due to its numerous valuable agronomic attributes which support sustainable agriculture (reviewed in Eynck and Falk, 2013). *Camelina sativa* grows well under a wide range of climatic and soil conditions. Its moderate-to-low requirement for nutrients and low seeding rate keep production costs low compared with other oilseed crops. A short growing season (85–100 days), low input costs and ease of transformation, with a simple floral dip procedure similar to that used for the model plant *Arabidopsis thaliana* (Lu and Kang, 2008), make *C. sativa* an attractive host for genetic and metabolic engineering. *Camelina sativa* has a unique seed oil profile with potential applications in food, feed and biofuel industries. Its seeds contain 36–47%

oil that is enriched with exceptionally high levels of essential fatty acids (Moser, 2012) that have been shown to be beneficial for human health. Over 90% of the fatty acids in *C. sativa* are unsaturated, and as such are generally believed to be susceptible to oxidation. However, the presence of high levels of antioxidants such as phenolics and tocopherols (vitamin E) in *C. sativa* oil make it more stable to oxidative degeneration than other unsaturated oils (Budin *et al.*, 1995; Abramovic and Abram, 2005). The recent success in developing transgenic *C. sativa* lines with high levels of omega-3 and -7 fatty acids, poly-3-hydroxybutyrate, unusual acetyl triacylglycerols, such as 3-acetyl-1,2-diacyl-*sn*-glycerols, and wax esters (Petrie *et al.*, 2014; Ruiz-Lopez *et al.*, 2014; Iven *et al.*, 2015; Liu *et al.*, 2015; Malik *et al.*, 2015; Nguyen *et al.*, 2015) has demonstrated that the content and quality of *C. sativa* lipids can be easily manipulated according to the needs of the bioproducts industry. *Camelina sativa* oil has been successfully converted into biodiesel and renewable jet fuel. A blend of *C. sativa*-derived biojet fuel and conventional petroleum-derived jet fuel has successfully powered several military, public and private aviation test flights. Consequently, there has been much interest in developing this emerging oilseed crop not only as a biofuel platform but also for the production of bio-lubricants and bio-plastics.

Development of *C. sativa* as a sustainable bioenergy feedstock will require a significant increase in crop productivity and improvement of oil composition for industrial applications. Genetic and genomic tools that aid in deciphering the complete gene complement, their structure and organization and the patterns of gene expression are key to the successful advancement of this novel oilseed crop through breeding of high-yielding cultivars with re-engineered oil composition. The recent sequencing of the *C. sativa* genome established a foundational resource for *C. sativa* genomics research and provided invaluable insights into the novel highly undifferentiated hexaploid structure of the genome (Kagale *et al.*, 2014a). The genome of *C. sativa* encodes 89 418 genes and has the highest density of genes observed thus far in sequenced plant genomes. The high gene density is a result of the evolution of *C. sativa* from the hybridization of three progenitors with lower chromosome number, with limited evidence for gene fractionation following polyploidization (Kagale *et al.*, 2014a). The notably conserved allopolyploid structure of the *C. sativa* genome suggests that modification of crop phenotypes using both traditional breeding and gene manipulation approaches will be challenging.

Transcriptional programs play a key role in the development and evolution of complex polyploid plant genomes. A thorough understanding of the transcriptional landscape of *C. sativa* is essential for precise and context-dependent interpretation of gene functions, to facilitate targeted traditional breeding, mutagenic and/or transgenic strategies for improvement of crop yield and oil quality. Previous efforts to characterize the *C. sativa* transcriptome have focused on characterization of either single or only a few life-cycle stages (Liang *et al.*, 2013; Nguyen *et al.*, 2013; Mudalkar *et al.*, 2014), and thus much of the *C. sativa* transcriptome remains to be explored. Additionally, the lack of a reference genome at the time impeded such global gene expression analyses.

Here, we report a comprehensive analysis of the developmental transcriptome of *C. sativa* enabled by RNA sequencing (RNAseq) of 12 different tissues covering four distinct developmental stages during its life cycle. The purpose of this RNAseq analysis was two-fold. Firstly, to establish a comprehensive catalogue of expressed transcripts in *C. sativa*, which served as experimental evidence for better delineation of exons, introns, untranslated regions and splice sites, and significantly improved the accuracy of gene prediction during the annotation of the recently published reference genome sequence (Kagale *et al.*, 2014a). The second objective, which is the focus of this paper, was to generate an integrated overview of gene expression dynamics during development of *C. sativa* as well as to establish a searchable database, or gene expression atlas, as a complementary resource to the genome sequence for future genomics research. Analysis of global gene expression provided a preliminary view of spatio-temporal gene expression patterns active during the development of *C. sativa* and identified clusters of tissue-specific co-expressed genes. Additionally, the results present evidence for alternative splicing and tissue-specific genome dominance among *C. sativa* genes. A high-quality census of transcription factors (TFs) and their expression analysis identified 4527 genomic loci that encode putative TFs in *C. sativa*, revealed their tissue-specific expression profiles and provided evidence for both functional conservation and diversification in the polyploid genome. The publically available genome-wide expression profiling data presented here together with the pre-existing reference genome sequence will greatly facilitate our understanding of gene regulatory mechanisms underlying developmental processes in *C. sativa*.

RESULTS AND DISCUSSION

The global landscape of *C. sativa* developmental transcriptome

To comprehensively characterize expression dynamics throughout *C. sativa* development, RNAseq libraries were constructed and sequenced from 12 different tissue samples collected at four major developmental stages (Table 1), including early growth (germinating seed and cotyledon), vegetative growth (young leaf, root, stem and senescing leaf), flower development (buds and flower) and seed development (early, early mid, late mid and late

Table 1 Summary statistics for transcriptome sequencing of diverse tissue types in genome-wide developmental expression analysis of *Camelina sativa*

Tissue	Total reads (million)	Mapped reads (%)	Mapped bases					Unmapped reads (%)
			Total (GB)	Coding (%)	UTR (%)	Intronic (%)	Intergenic (%)	
Early growth								
Germinating seed (GS)	137.83	90.09	12.39	44.08	8.53	2.15	45.24	9.91
Cotyledon (C)	99.38	91.28	9.07	77.45	11.45	3.38	7.71	8.72
Vegetative growth								
Young leaf (YL)	55.98	91.22	5.12	76.53	11.52	3.54	8.41	8.78
Root (R)	81.21	90.67	7.38	75.97	12.32	3.92	7.79	9.33
Stem (S)	63.54	90.38	5.76	75.72	12.67	3.68	7.94	9.62
Senescing leaf (SL)	41.63	90.17	3.76	75.74	12.17	3.96	8.13	9.83
Flower development								
Bud	32.71	90.01	2.95	72.52	14.53	4.25	8.70	9.99
Flower (F)	43.83	88.07	3.86	73.57	14.45	3.56	8.42	11.93
Seed development								
Early seed development (ESD)	38.81	89.01	3.47	74.96	13.70	3.13	8.21	10.99
Early mid seed development (EMSD)	46.60	87.61	4.10	76.91	10.98	3.02	9.09	12.39
Late mid seed development (LMSD)	42.63	88.44	3.79	76.23	12.05	3.24	8.47	11.56
Late seed development (LSD)	43.24	90.04	3.91	73.07	10.26	3.52	13.15	9.96
Total	727.39		65.55					
Average		89.75		72.73	12.05	3.45	11.77	10.25

UTR, untranslated region.

seed development). Three independent biological replicates were assessed per tissue type. Using the Illumina HiSeq2000 platform, about 727 million 100-bp paired-end reads (73 Gb) were generated, corresponding to an average of 61 million reads per tissue sample (Table 1).

Low-quality, adapter contaminated and short reads were removed and the remaining high-quality reads from each sample were aligned to the *C. sativa* reference genome (Kagale *et al.*, 2014a) using TOPHAT2 (Kim *et al.*, 2013). Approximately 90% of the filtered reads from each tissue sample mapped to the genome sequence (Table 1). A total of 65.55 Gb of sequence data mapped to the reference genome, representing a 567-fold coverage of the annotated *C. sativa* exome. On average, 73% of the aligned bases mapped to the coding sequence (CDS), while the remaining bases mapped to untranslated regions UTRs (12%), introns (3%) and intergenic regions (12%; Table 1). In germinating seed, the proportion of bases that aligned to the intergenic regions (45%) was substantially higher than in any other tissue (7–13%; Table 1). The distribution of intergenically mapped reads from this tissue revealed broad intergenic expression across the length of each chromosome, except for centromeric and pseudo-centromeric regions. There was no evidence for enrichment of such reads within annotated repeat regions as only 1.5% of the intergenically mapped reads aligned to annotated repeats in the *C. sativa* genome. Further, approximately 11% of the intergenic reads from germinating seeds mapped to first/last exon junctions and unannotated UTR regions, while an additional 10% mapped to plastid-like sequences embedded within the nuclear genome. Considering all samples, the proportion of reads that mapped to intergenic regions is likely lower than the average 12% observed across tissue types, as intergenic regions might contain as-yet unidentified genes, long non-coding RNAs, unannotated UTRs and poorly annotated genes/splice variants. During the past decade, it has become increasingly evident that the vast majority of the genome is pervasively transcribed (Consortium 2007; Berretta and Morillon, 2009; Clark *et al.*, 2011). The transcripts arising from intergenic regions could also be due to the prevalence of pervasive transcription in *C. sativa*, particularly in germinating seeds.

To explore the dynamics of gene expression during the development of *C. sativa*, normalized transcription levels (fragments per kilobase of exon per million fragments mapped, FPKM) for each gene/gene model were calculated using CUFFLINKS (Trapnell *et al.*, 2012). A matrix representing individual *C. sativa* genes and their expression values (FPKM) across individual tissue types is presented in Data S1 in the Supporting Information. FPKM normalization is dependent on the effective library length, which is invariably different between samples, and hence could limit comparisons of FPKM values across libraries. New expression profiling tools are being developed to use transcripts per kilobase million (TPM) to normalize counts to a consistent number per library (1 million), arguably improving comparison of relative expression values across samples (Wagner *et al.*, 2012). In addition to the Tuxedo

package (TOPHAT2 and CUFFLINKS), KALLISTO (Bray *et al.*, 2016) was used to estimate relative transcript abundance (TPM) in individual tissues/replicates. While KALLISTO provides an alternative estimate of transcript abundance, it relies upon alignment of sequence reads to a fully annotated transcriptome, and thus will be limited by errors or omissions in the current annotation, and precludes downstream analyses of alternative splice variants. A matrix representing individual *C. sativa* genes and their relative transcript abundance (TPM) across individual tissue types is presented in Data S2. The correlation of expression (FPKM versus TPM) for each gene across all tissue types was calculated to determine if there was a significant difference between the FPKM and TPM datasets. These analyses revealed a concentration of correlation values closer to one (Figure S1), thus indicating high correlation between TPM and FPKM values. All subsequent analyses were performed using TPM values.

Genes with TPM values greater than 0.1 in at least two of the three replicates were considered expressed. Expression measurements from the three biological replicates for each tissue were highly correlated, with an average Pearson's correlation coefficient (R) value of 0.96 $\pm$ 0.003 (Data S3), indicating the high reproducibility of the biological replicates. Pearson correlation values between randomly assorted samples/replicates were generally very low (ranging from 0.04 to 0.94; Data S3).

The *C. sativa* genome was predicted to contain 89 418 genes and 94 495 gene models (Kagale *et al.*, 2014a). We found that 82 810 (88% of the total) annotated gene models were expressed in at least one of the 12 tissue types (Figure 1a). Detection of expression for such a high

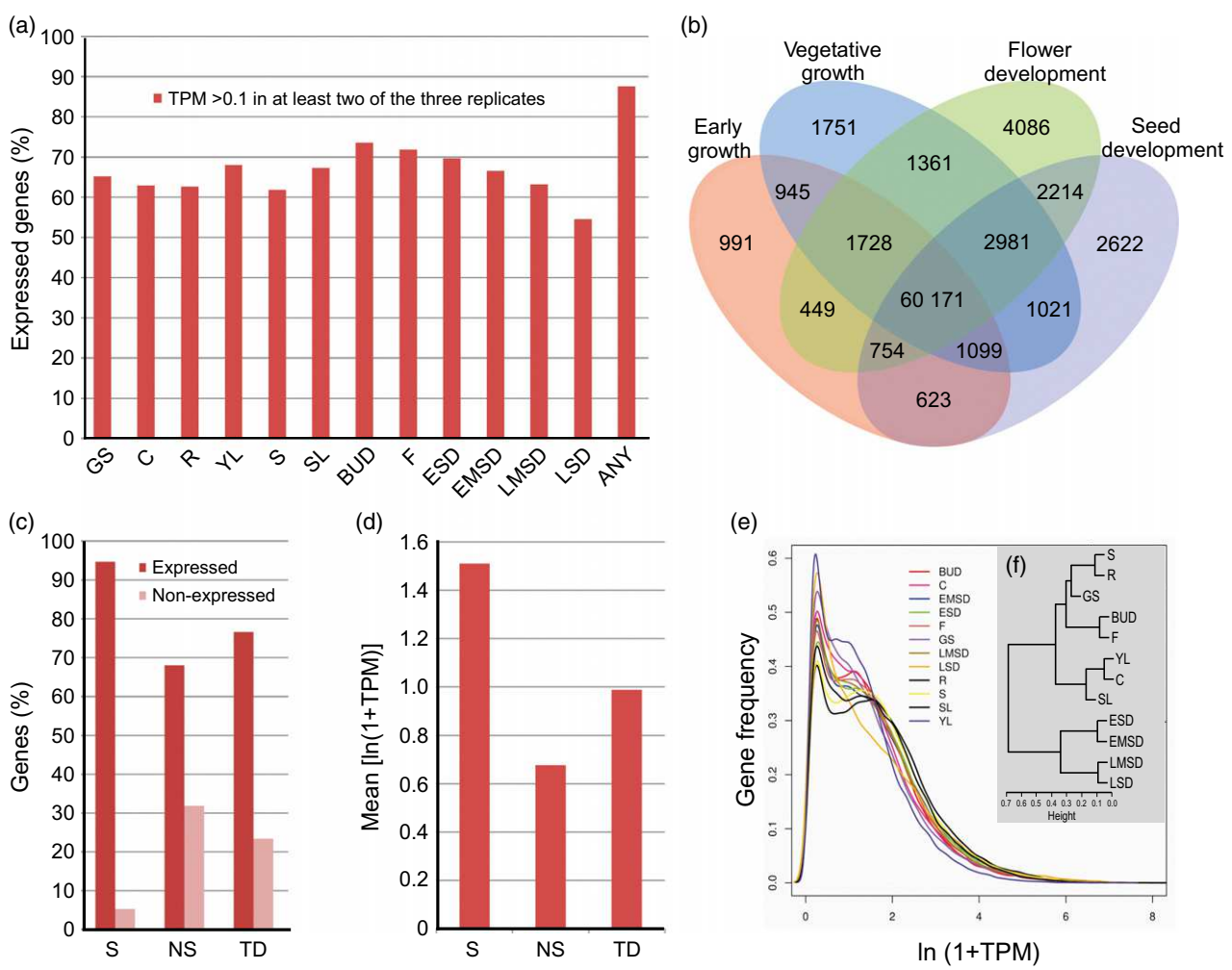


Figure 1. Global gene expression patterns in different tissues of *Camelina sativa*.
(a) Percentage of genes expressed in each of the 12 tissue samples. For abbreviated tissue names see Table 1. (b) Number of genes detected in the four distinct developmental stages. (c) Proportion of expressed and non-expressed syntenic (S), non-syntenic (NS) and tandem duplicate (TD) genes. (d) Average expression levels of S, NS and TD genes. (e) Distribution of expression levels for each tissue type. To improve the legibility of the figure, *x*-axis values less than 0.1 were excluded. (f) Cluster dendrogram showing the global relationship among different tissue samples.

proportion of genes may be attributed to the diversity of tissues included in the study as well as high sequencing depth and coverage. The expression of 11 685 gene models (12% of the total) was either not detected in any of the tissues or detected below background levels (TPM < 0.1). These gene models could represent pseudogenes or they may be expressed under specific biotic or abiotic stress conditions. The proportion of gene models for which expression was detected in individual tissues ranged from 55% to 73% (Figure 1a). A total of 60 171 gene models (73% of the expressed genes) produced mRNAs in all four developmental stages (Figure 1b), indicating a remarkable correspondence of gene expression among the spatio-temporally distinct tissue types. The remaining 27% of the gene models exhibited either specific or shared expression patterns among different vegetative or reproductive growth stages (Figure 1b).

Based on synteny with *A. thaliana*, *C. sativa* genes were classified as syntenic (62 277 genes), non-syntenic (16 349) or tandem duplicate (10 792) genes (Kagale *et al.*, 2014a). Comparative assessment of gene expression revealed that about 32 and 23% of the non-syntenic and tandem duplicate genes, respectively, were not expressed in any of the 12 different samples (Figure 1c). In contrast, only about 5% of the syntenic genes were not expressed. Furthermore, a notable difference in the level of expression of syntenic [mean ln(1 + TPM) = 1.51] and non-syntenic/tandem-duplicates (0.67/0.98) was observed (Figure 1d). Compared with their syntenic counterparts, non-syntenic genes were also found to be shorter, with fewer introns, and were evolving with higher K_a/K_s ratios (Table S1). These features are expected evolutionary signatures of relaxed selection constraints leading to pseudogenes (Zou *et al.*, 2009; Yang *et al.*, 2011) indicating that the non-syntenic genes in *C. sativa* are potentially targeted for pseudogenization. A similar observation was made for such genes in *Brassica oleracea* (Parkin *et al.*, 2014).

The distribution of expression values of all expressed genes in *C. sativa* revealed subtle differences between many of the different tissues. Most of the tissue types, except for the germinating seed and late seed development, exhibited bimodal expression distributions with two prominent peaks, one near a ln(1 + TPM) value of 0.1 and the other at around 1.8 (Figure 1e). The expression distributions of germinating seed and late seed development lacked the secondary peak at ln(1 + TPM) of 1.8 (Figure 1e). Classification of tissue samples by clustering based on mean expression values for all genes revealed global relationships among the different tissues (Figure 1f). There were two main clusters: the first cluster contained the four seed developmental stages whereas the second cluster was represented by all vegetative tissues and also floral buds and flowers. Within the respective clusters there was a clear distinction between the early and late seed developmental stages. As expected, transcriptomes of floral buds and flowers were more similar to each other than to other tissues. Similarly, the transcriptomes of root and shoot were more similar to each other than to any other tissue. The cotyledon and leaf transcriptomes formed a separate sub-cluster. This distinction of closely related tissue types indicated that the transcriptome data captured unique characteristics of each developmental phase in *C. sativa* and could yield valuable insights into the transcriptional programs underlying developmental processes.

Tissue-specific expression dynamics

Despite the qualitative similarities in gene expression amongst various tissues (Figure 1), dynamic changes in gene expression levels and patterns were observed as about 75% (71 096 genes) of all *C. sativa* genes showed significant differential expression [ANOVA, *P*-value adjusted for false discovery rate (FDR) < 0.01; Data S4] between tissues. A search among the differentially expressed dataset identified 6657 genes showing specific expression in individual tissues (Data S5). The greatest number of tissue-specific genes were expressed in floral buds (2217 genes), whereas the young leaf (103 genes) and cotyledons (108 genes) had the smallest number of uniquely expressed genes (Figure 2a). The other tissues with higher numbers of specifically expressed genes included early seed development (1093 genes), germinating seed (859 genes), root (700 genes) and flower (585 genes; Figure 2a). By contrast, there were very few specifically expressed genes in cotyledon, senescing leaf, stem and mid–late seed developmental stages (Figure 2a).

In light of the importance of *C. sativa* as an oilseed crop, we examined the expression of seed-specific genes in more detail. A total of 2622 genes were identified to be preferentially expressed during one or more of the four seed developmental stages examined in this study (Figure 2a, b, Data S5). Of these, 263 genes were expressed in all four seed developmental stages (Figure 2b). Interestingly, about 42% (1093 genes) of the seed-specific genes were uniquely expressed during the early seedling development (ESD; Figure 2b) stage. The number of preferentially expressed genes decreased with the progression of seed development as the late seed developmental stages had relatively the lowest number of specifically expressed genes (Figure 2b). These data imply that distinct transcriptional repertoires and highly coordinated expression dynamics play a crucial role in regulating spatio-temporal progression of seed development.

Comparison of the seed-specific genes (2622 genes) with putative genes encoding proteins and regulatory factors involved in acyl-lipid metabolism (Li-Beisson *et al.*, 2013) in *C. sativa* revealed a small subset of the lipid metabolism genes (51 genes) that were expressed specifically during

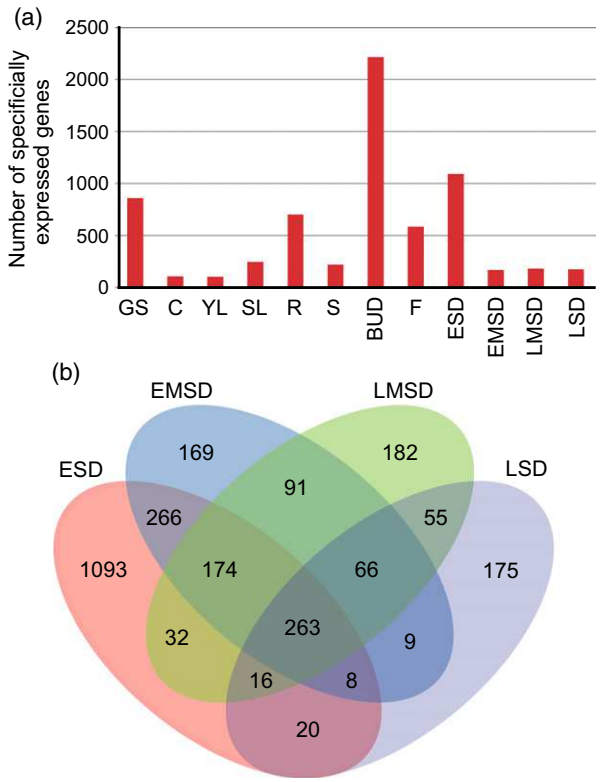


Figure 2. Tissue-specific genes in *C. sativa*.
(a) Number of genes specifically expressed in different tissues. For abbreviated tissue names see Table 1.
(b) Seed-specific genes were classified based on shared and specific expression profiles in multiple seed developmental stages, including early seed development (ESD), early mid seed development (EMSD), late mid seed development (LMSD) and late seed development (LSD).

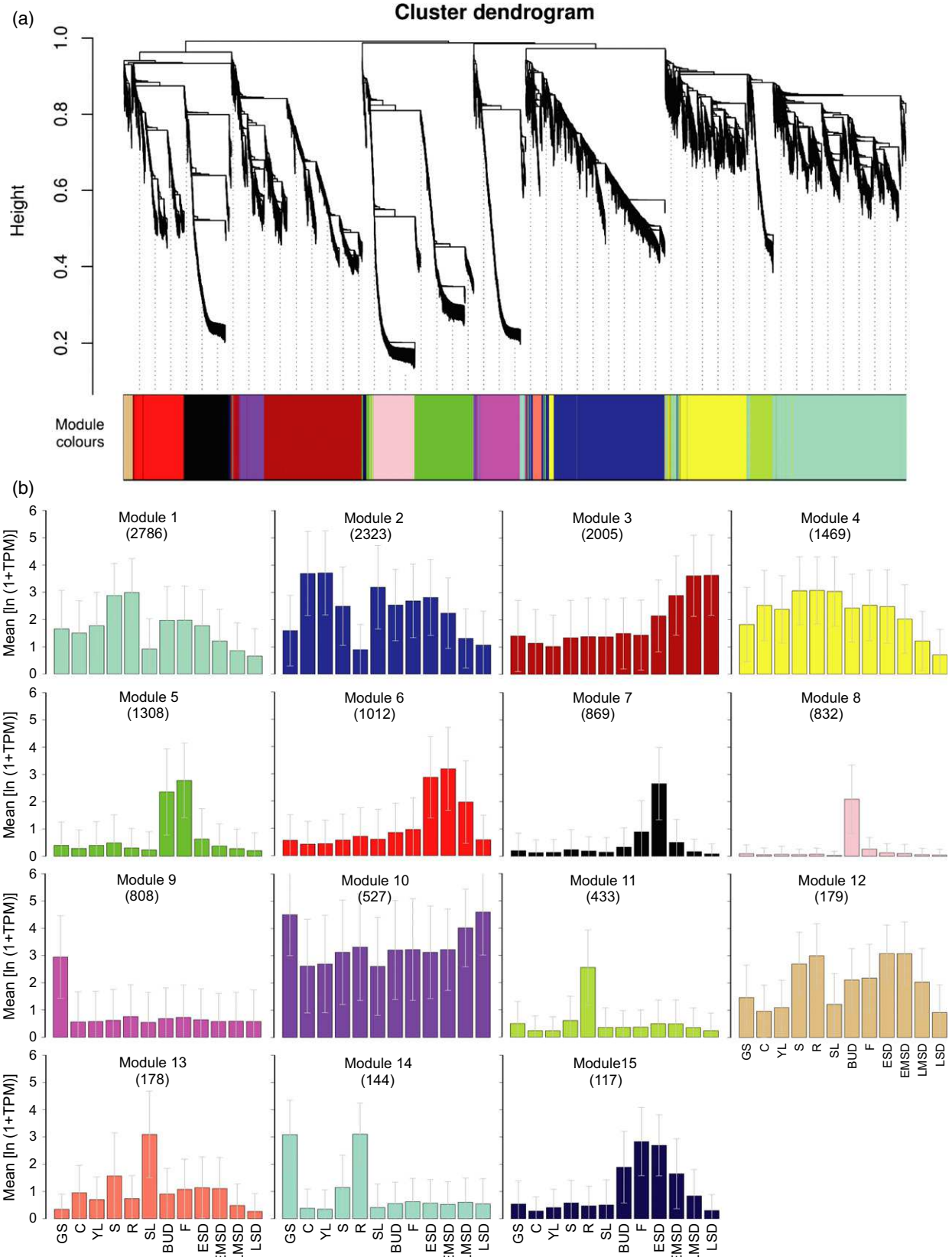
seed developmental stages (Table S2). Functional classification of these genes revealed that most encode proteins/regulatory factors involved in fatty acid elongation, wax biosynthesis and triacylglycerol biosynthesis (Table S2). It may be possible to use these seed-specific lipid metabolism genes as direct targets for modification of the oil profile and/or as sources of seed-specific promoters for transgenic research.

Co-expression network analysis

Gene regulatory networks can provide systematic understanding of molecular mechanisms underlying specific biological pathways/processes (Ihmels *et al.*, 2002; Stuart *et al.*, 2003). Co-expression analysis was established in response to the theory that a high degree of similarity in gene expression profiles can with certain caveats correlate with relatedness of gene functions (Rhee and Mutwil, 2014). Recent advances in high-resolution sequencing and statistical genomics have motivated a shift from gene-centric data analysis approaches to network (module)-based systems approaches (Ponomarev *et al.*, 2012). Weighted gene co-expression network analysis (WGCNA) is one of the most established and widely used systems approaches for identifying highly interconnected modules of co-expressed genes.

Module detection by WGCNA analysis is a computationally intensive process and allowed analysis of relatively few genes (up to 15 000 genes at a time). Based on the statistically determined ranking of the differentially expressed genes (ANOVA, minimum FDR adjusted *P*-value $< 2.88 \times 10^{-13}$), WGCNA was used to build gene co-expression networks for the 15 000 highest ranked genes. Fifteen distinct transcription modules (Figure 3a) associated with various developmental phases in *C. sativa* were identified. The module size ranged from 117 to 2786 genes. To explore the functional relevance of these gene expression modules, tissue-wise means and variances of transformed TPM values [ln(1 + TPM)] for genes in each module were plotted (Figure 3b). The majority of the modules contained genes predominantly expressed in either a single tissue or a few developmentally related tissues. For example, modules 3, 6, 7 and 15 were composed of genes predominantly co-expressed during various seed developmental stages, whereas genes in modules 5 and 8 were co-expressed in flower and/or floral buds (Figure 3b). Similarly, modules 9, 11 and 13 identified germinating seed-, root- and senescing leaf-specific co-expression clusters (Figure 3b), respectively. This is consistent with earlier analysis using hierarchical clustering, where tissue-specific gene expression clusters were identified among homeologous genes showing significant sub-genome × tissue-type interaction (Kagale *et al.*, 2014a). The identification of tissue-specific transcriptional modules suggests that delineation of expression between tissues could play a role in demarcation of developmental phases in *C. sativa*.

Figure 3. Co-expression network analysis of developmental genes in *Camelina sativa*.
(a) Hierarchical cluster tree showing co-expression modules identified by weighted gene co-expression network analysis of the 15 000 most differentially expressed genes. The major tree branches constitute 15 distinct co-expression modules labelled by different colours as indicated by a colour band underneath the tree.
(b) Tissue-specificity of genes belonging to individual co-expression modules. Histogram bars are coloured according to the corresponding module colour. The histograms show the mean and one standard deviation error bars for the transformed expression values (mean over replicates of ln(1 + TPM)l; TPM, transcripts per kilobase million) for each tissue. (Note: the order of tissues in module histograms 1–11 follows the same order as indicated for modules 12–15.) The number of genes in each module is indicated in parenthesis below the module number.



Analysis of divergence of tissue-dependent homeologous gene expression

Polyploid genomes can display dynamic changes in gene expression as a genomic shock response to intergenomic interactions (Levy and Feldman, 2002; Chen and Ni, 2006). Regulatory mechanisms that orchestrate changes in gene expression in polyploids often favour overexpression of individual genes specifically from the least-fractionated or most conserved sub-genome (Schnable *et al.*, 2011; Cheng *et al.*, 2012; Tang *et al.*, 2012). For allohexaploid *C. sativa*, the comparison of genome-wide transcript abundance of homeologous genes indicated partial dominance of the sub-genome Cs-G3 over the other two sub-genomes (Kagale *et al.*, 2014a). However, the extent of divergence of genome dominance between various tissues is largely unknown.

In this study, we analysed tissue-specific genome dominance in *C. sativa* based on transcript abundance of orthologous genes between the sub-genomes of *C. sativa,* in 12 different tissue types. Only the fully retained 18 565 homeologous triplets (homeologous gene loci that had a representative member in all three sub-genomes of *C. sativa*) were included in this analysis to avoid discrepancies due to biased fractionation (Kagale *et al.*, 2014a). The homeologous triplets for which a statistically significant difference ($P < 0.05$; ANOVA) between sub-genomes was observed were identified, and the frequency of the highest-expressing genes in each sub-genome was enumerated. Comparison of the frequencies of the highest expressing genes revealed significant expression divergence ($P < 0.05$; χ^2 test) between the three sub-genomes in various tissues (Figure 4, Table S3). Notably, the genome-wide expression differences between sub-genomes Cs-G1 and Cs-G3 as well as Cs-G2 and Cs-G3 were highly significant across all 12 tissue types, with the sub-genome Cs-G3 having a clear expression advantage over the other two (Figure 4). The comparison of expression differences between Cs-G1 and Cs-G2 revealed dominance of Cs-G1 over Cs-G2, specifically during reproductive growth stages, including flowers, and the late seed developmental stage (Figure 4). During our earlier analysis of genome dominance using the combined whole-plant expression dataset, expression divergence between Cs-G1 and Cs-G2 was not evident (Kagale *et al.*, 2014a). Subtle tissue-specific changes in the levels of expression for a subset of genes may have been missed or overshadowed by larger changes in expression of these genes at a whole plant level. The reproductive growth-specific dominance of Cs-G1 over Cs-G2 confirms the existence of tissue-specific expression variation among homeologues and suggests that a relatively complex architecture of transcriptional mechanisms may govern the development of reproductive organs in *C. sativa*.

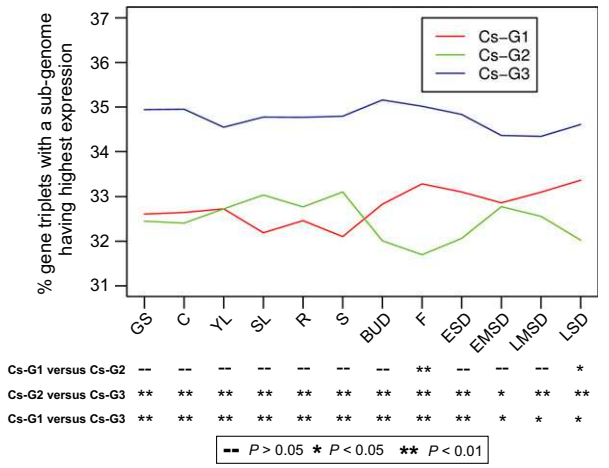
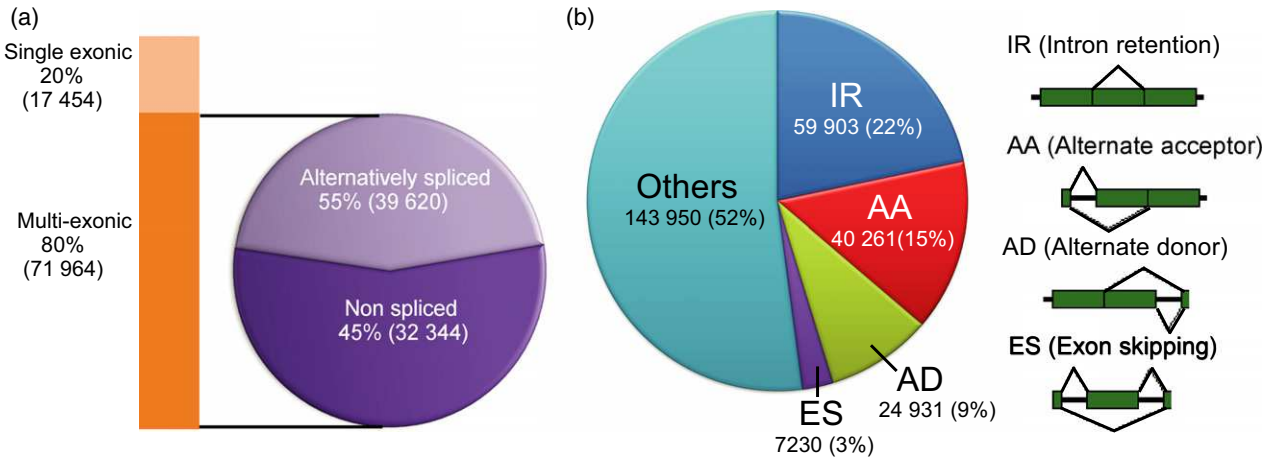


Figure 4. Tissue-specific genome expression dominance.
This figure shows frequencies of genes belonging to the three sub-genomes (Cs-G1/-G2/-G3) within *C. sativa* having the highest expression across the homeologous gene triplet in individual tissues types. Significant expression divergence among homeologous genes assessed by Pearson χ^2 test is indicated below the graph by asterisk(s). The actual χ^2 and *P*-values associated with each comparison are presented in Table S3. For abbreviated tissue names see Table 1.

Alternative splicing

Alternative splicing (AS) is an important post-transcriptional gene regulatory mechanism that regulates gene expression and increases the proteome complexity of higher eukaryotes (Black, 2003; Chen and Manley, 2009; Reddy *et al.*, 2013). High-throughput RNAseq analyses in various plants, including Arabidopsis (Filichkin *et al.*, 2010; Marquez *et al.*, 2012; Syed *et al.*, 2012), rice (Zhang *et al.*, 2010), soybean (Shen *et al.*, 2014), maize (Thatcher *et al.*, 2014), grape (Zenoni *et al.*, 2010), cucumber (Guo *et al.*, 2010) and *Brachypodium distachyon* (Walters *et al.*, 2013), revealed that up to 60% of intron-containing genes can be alternatively spliced. In the *C. sativa* genome annotation, 4753 (6.6% of the multi-exonic genes in *C. sativa*) genes were predicted to have alternatively spliced variants (ASV; Kagale *et al.*, 2014a), a relatively low proportion compared with the incidence of AS in other plant species. The transcripts derived from combined RNAseq data from all 12 tissues were analyzed using ASTALAVISTA (alternative splicing transcriptional landscape visualization tool; Foissac and Sammeth, 2007) to identify the complete repertoire of ASVs in *C. sativa.* This identified a total of 276 275 AS events, distributed across 39 620 genes, accounting for 55% of the multi-exonic genes in *C. sativa* (Figure 5a).

AS can be mainly classified as intron retention (IR), alternative acceptor (AA) sites, alternative donor (AD) sites and exon skipping (ES; Black, 2003; Sammeth *et al.*, 2008). Of the 276 275 AS events identified, IR (22% of the total) was



(c)

Tissue	IR	ES	AA	AD	Others	Total
Germinating seed	13 944	3613	11 389	7926	27 998	64 870
Cotyledon	15 737	4027	12 201	8137	32 388	72 490
Young leaf	12 637	3312	9664	6683	24 996	57 292
Senescing leaf	13 722	3593	10,701	7415	28 259	63 690
Root	19 930	4380	14,256	9261	39 903	87 730
Stem	16 705	4103	12,297	8262	34 354	75 721
Buds	14 421	3713	10,384	7281	26 959	62 758
Flower	13 741	3888	11,468	7754	27 949	64 800
Early seed development	10 122	2966	8072	6005	19 574	46 739
Early-mid seed development	10 527	3211	8721	6447	22 631	51 537
Late-mid seed development	11 524	3361	9756	7116	24 703	56 460
Late seed development	11 765	2645	7322	5141	20 149	47 022

Figure 5. Alternative splicing events in the *Camelina sativa* genome.
(a) Proportions of alternatively spliced and non-spliced multi-exonic genes in *C. sativa*.
(b) Frequency of different alternative splicing events during *C. sativa* development.
(c) Frequency of different alternative splicing events in individual tissues.

found to be the most abundant type followed by AA (15%), AD (9%) and ES (3%; Figure 5b). These results are consistent with the observations made in Arabidopsis (Marquez *et al.*, 2012), soybean (Shen *et al.*, 2014) and grape (Vitulo *et al.*, 2014), thus supporting the idea that IR is the most common AS event in plants.

To determine if AS was prevalent in any one tissue, the RNAseq transcripts from individual tissues were analysed. Although the frequency of AS types in each tissue (Figure 5c) was similar to the global pattern of AS (Figure 5b), the number of AS events varied in different tissues. There were relatively more AS events in early growth, vegetative (except young leaf) and flower developmental stages (buds and flowers) than in seed developmental stages (Figure 5c). The lower number of AS events during seed developmental stages might be due to the relatively lower number of genes expressed during late seed development (Figure 1a); however, such a correlation was not evident in vegetative tissues (e.g. young and senescing leaf tissues, and cotyledons). The unequal distribution of AS events during *C. sativa* developmental stages might be regulated by tissue-specific cues.

To determine the cause of variance in tissue-wise AS, we performed differential expression analysis of alternative splicing factor genes. *Arabidopsis thaliana* encodes 395 splicing-related proteins (Wang and Brendel, 2004). A search for syntelogs (syntenic orthologues) identified 1000 orthologous genes encoding splicing-related proteins in *C. sativa*, of which 970 genes were expressed in at least one tissue (Data S6). Further analysis revealed that about 93% of these genes were differentially expressed (ANOVA, FDR-adjusted P-value < 0.01) in various *C. sativa* tissues (Data S6). Thus, the high level of differential expression of splicing-related genes in *C. sativa* could potentially be one

of the causes for the variance in AS events observed in different tissues.

It has been suggested that a large proportion of AS events in plants are induced only in response to abiotic and biotic stresses (Staiger and Brown, 2013). As the RNA-seq samples were collected mainly from different developmental stages and did not include samples from stress conditions, AS in *C. sativa* might be more prevalent than was detected in this study.

Transcription factor genes in the *C. sativa* genome

The coordinated regulation of gene expression in plants and other higher eukaryotes is essential for determining the majority of cellular and biological processes, at the heart of which are transcription factors (TFs) that create focal points in complex gene regulatory networks (Massie and Mills, 2008). TF families are strongly conserved across eukaryotic organisms, especially in plants (Libault *et al.*, 2009; Jin *et al.*, 2014). The availability of complete genome sequences as well as expression data has enabled identification of TF repertoires from 83 plant species covering major lineages of green plants (Plant Transcription Factor Database, PlantTFDB; Jin *et al.*, 2014). Given the central role of TFs in regulating plant gene expression, a census of TFs in *C. sativa* and analysis of their expression is essential for understanding various processes involved in its development, differentiation and responses to the environment.

The *A. thaliana* genome encodes 1716 transcription factors classified into 58 families based on the structure of their DNA-binding domains (PlantTFDB; Jin *et al.*, 2014). A complete repertoire of TFs in the *C. sativa* genome was identified by surveying the syntelog matrix representing individual *A. thaliana* genes and the corresponding *C. sativa* syntelogs (Kagale *et al.*, 2014a) (Data S7). This resulted in a high-confidence dataset of 4527 genes (*c.* 5% of the total number of protein-coding genes) that encode TFs in *C. sativa*. The family-wise classification and count of TFs in *C. sativa* is presented in Figure 6(a). Enumeration of syntelogs revealed that a vast majority of TFs (*c.* 82%) were retained in three copies, whereas the remaining TFs lost one (3%), two (7%) or all three homeologues (8%; Figure 6b). TFs from a subset of families, including AP2, CAMPTA, CO-like, DBB, EIL, GRF, HB, HRT-like, LFY, LSD, NF-YA, Nin-like, SPL, VOZ, Whirly and YABBY, have evaded fractionation and retained all three copies (Figure 6a). In contrast, there were a few TF families, such as B3, GeBP, M-type and NAC, which experienced exceptionally high levels of fractionation (with loss of all homeologues in some cases; Figure 6a).

Examination of tissue-dependent expression revealed that approximately 94% (4259 out of 4527) of TFs were expressed in at least one tissue (Figure 6c). The proportion of expressed TFs varied greatly between tissue types, ranging from approximately 53% in late seed development to 78% in floral buds (Figure 6c). The proportion of expressed TFs relative to all expressed genes was remarkably stable at about 5% across all tissue samples (Figure 6c). Maintaining lower expression levels of TFs is thought to ensure correct recognition of target sites and avoid undesired binding (Liu and Clarke, 2002; Vaquerizas *et al.*, 2009). Moreover, activating or repressing a regulatory event by altering the concentration of TFs would be easier if their expression levels are kept low (Vaquerizas *et al.*, 2009). As might be expected based on function, the annotated TFs were found to be expressed at slightly lower levels compared with all expressed genes in *C. sativa* (Figure 6d).

Hierarchical clustering of TFs in *C. sativa* based on patterns of gene expression across all 12 tissue types revealed 13 major clusters (Figure S2). The number of TFs belonging to each cluster ranged from 28 to 1265. Remarkably, TFs belonging to each cluster were predominantly expressed or suppressed in one or a few tissue types, which could suggest tissue-specific functions (Figure S2). This is consistent with the tissue-specific gene expression patterns revealed by the genome-wide WGCNA analysis (Figure 3) and earlier hierarchical clustering analysis (Kagale *et al.*, 2014a). However, the sub-genome-wide classification of TFs belonging to each cluster revealed that genes belonging to about 36% of homeologous triplets were separated across two or three independent clusters. Homeologues with altered expression patterns could provide evidence for functional diversification of a subset of triplicated TFs within *C. sativa* through evolutionary mechanisms such as neofunctionalization, the acquisition of new functions, or subfunctionalization, where the homeologue acquires an additional functional role.

Conservation of expression patterns of TF orthologues in *C. sativa* and *A. thaliana*

It is widely accepted that orthologues conserved across different species often perform the same or similar biological functions; in such cases they also are expected to maintain tissue-specific expression patterns despite millions of years of evolution. The availability of large expression datasets for comparable tissue samples in different plant species has enabled analysis of conservation or diversification of gene function in plant lineages. Comparison of expression data from *A. thaliana*, rice and *Medicago truncatula* revealed conservation of expression context in 77 and 62% of orthologous genes, respectively (Benedito *et al.*, 2008; Movahedi *et al.*, 2011). To assess potential functional conservation among orthologous genes from *A. thaliana* and *C. sativa,* pair-wise comparison of expression of TFs in matching tissue samples (cotyledon, leaf, senescing leaf, root, flower and seed developmental stages) (Schmid *et al.*, 2005) from the two species was performed. Spearman correlation analysis revealed a concentration of *P*-values close to zero in pair-wise comparison of

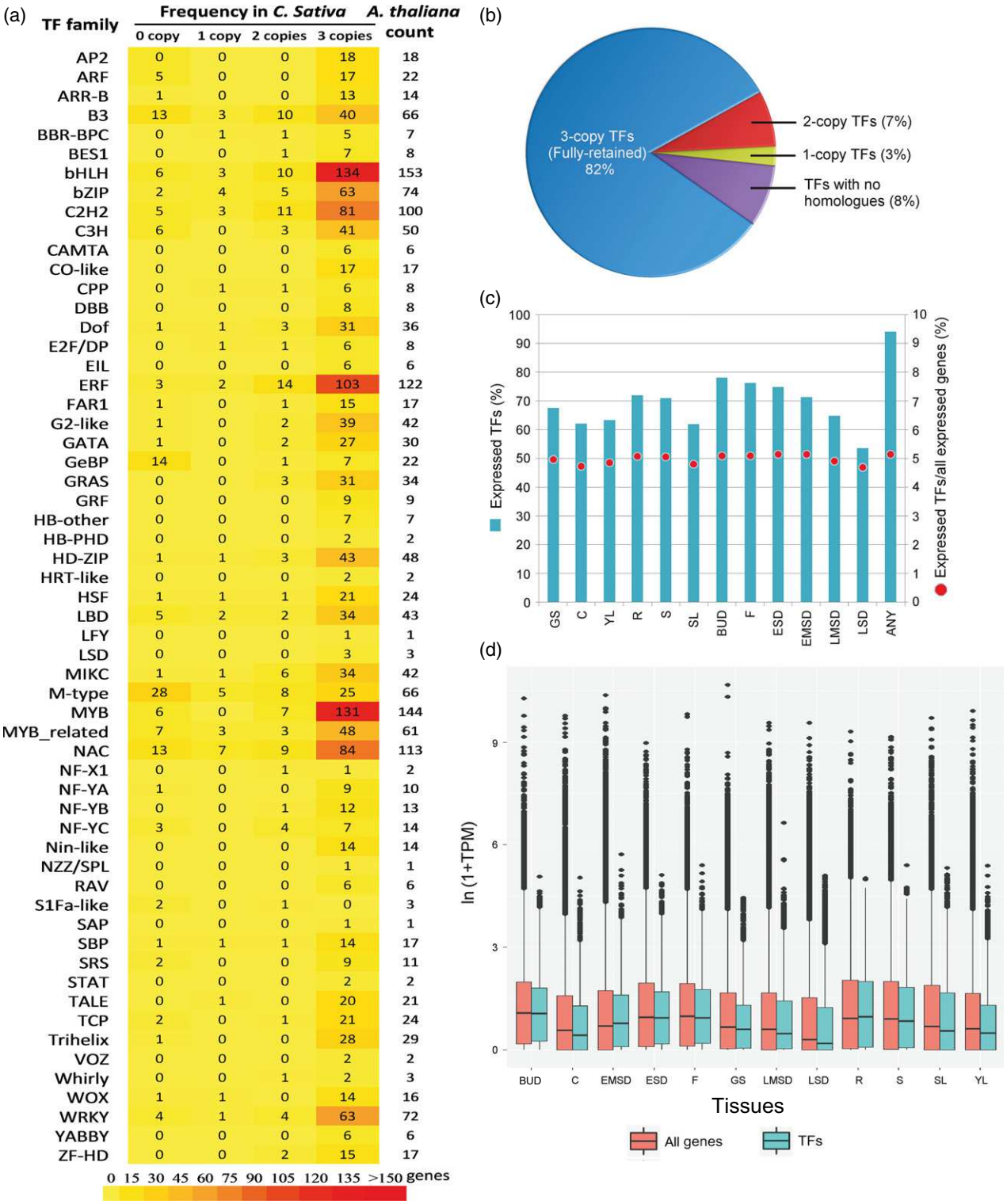


Figure 6. Copy number and expression of transcription factor genes in *Camelina sativa*

(a) Copy number and retention pattern of transcription factor (TF) genes in *C. sativa*. TFs were classified into families according to their DNA-binding domain composition. A family-wise count of TFs in *A. thaliana* and their frequency of occurrence in *C. sativa* are presented.

(b) Proportion of TFs that are fully-retained (three copies), fractionated (one or two copies) or completely lost (no homeologues).

(c) Proportion of TFs expressed in each tissue sample and the proportion of expressed TFs versus all expressed genes. For abbreviated tissue names see Table 1.

(d) Expression levels of TFs compared with the expression levels of all genes in *C. sativa*.

expression patterns of TFs between *A. thaliana* and the three sub-genomes of *C. sativa* (Figure S3; left panel). This distribution was distinct from that generated when comparing expression values for *A. thaliana* genes with no consideration for tissue type (Figure S3; right panel). These results suggest that a general similarity of the pattern of gene expression over the different tissue types is retained between *A. thaliana* and *C. sativa.*

The *C. sativa* eFP browser

To provide easy access and enable visualization of the expression levels of *C. sativa* genes, an eFP browser (located at http://bar.utoronto.ca/efp_camelina/cgi-bin/efpWeb.cgi) allowing basic queries by the annotated gene names was generated. The search interface produces an electronic fluorescent pictograph coloured on the basis of transcript abundance data for individual *C. sativa* genes in different tissues (exemplified in Figure 7). The eFP browsers generated for various other plant species, including Arabidopsis, poplar, *M. truncatula*, soybean, potato, tomato and *Eutrema salsugineum* (Winter *et al.*, 2007; Patel *et al.*, 2012; Champigny *et al.*, 2013) have proved to be highly useful in the display and interpretation of transcript abundance data. Similarly, the eFP browser for *C. sativa* is expected to be a valuable resource for *C. sativa* researchers in the functional characterization of gene copies. To demonstrate the utility of the eFP browser, we have presented eFP images (Figure S4) for a small subset of triplicated TFs that exhibited conservation of function (a similar expression pattern in all three homeologues), subfunctionalization (shared and/or partitioned function evident from expression of Csa17g016750.1), neofunctionalization (functional diversification as evident from expression of Csa03 g014170.1) and non-functionalization (silencing as evident from the loss of expression of Csa05 g020540.1 in all tissues).

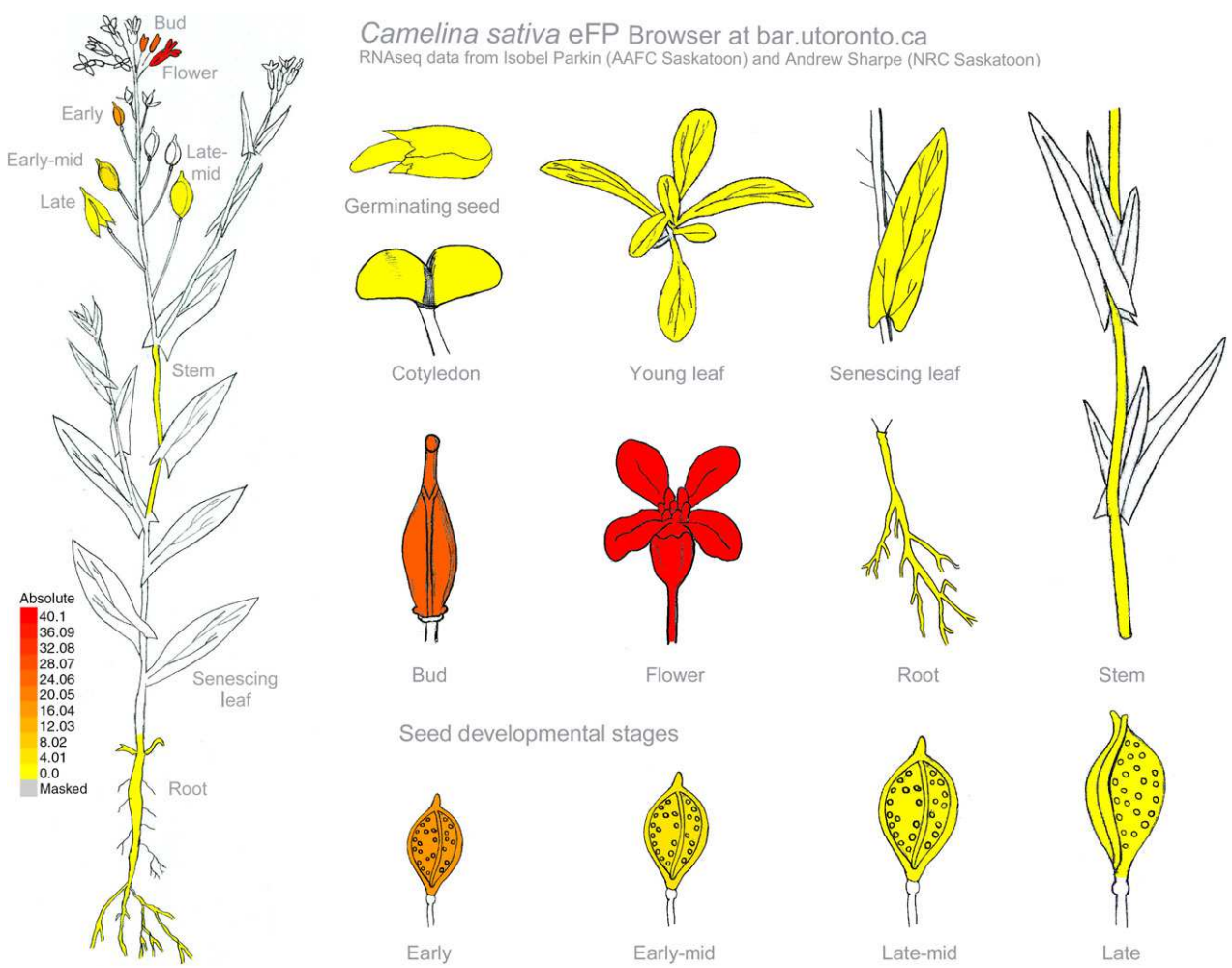


Figure 7. eFP browser view of expression changes during *Camelina sativa* development.
An exemplary *C. sativa* eFP browser output image showing expression values for the CsSEPALLATA3 gene (Csa17 g035980.1) is shown. In all cases, red indicates higher levels of transcript accumulation and yellow indicates a lower level of transcript accumulation.

CONCLUSIONS

In summary, we have generated an extensive and high-quality expression map that covers a wide range of tissues and developmental stages in *C. sativa*. Our results show that at least 88% of the entire gene complement in *C. sativa*, including 4259 TFs, was expressed, probably contributing to the underlying growth and development in *C. sativa*. It is evident that quantitative rather than qualitative changes in global gene expression effect tissue identity and developmental transitions in *C. sativa*. A combination of gene-centric and network-based systems approaches allowed us to uncover transcriptional relationships between genes and tissues. It has further helped in identification of tissue-specific expression signatures that highlight dynamic reprogramming of the *C. sativa* transcriptome and reveal functional transitions occurring during vegetative and reproductive tissue development. For example, about 3.1% of all genes detected during seed development were identified as seed-specific. This set of genes has fundamental implications for studying seed development and improving the quality of seed oil and protein in *C. sativa.* The analyses of multiple tissue types allowed further elucidation of the relationship between the orthologues found across the three sub-genomes of *C. sativa*. Further to the previously identified dominance of Cs-G3 over the two related sub-genomes (Kagale *et al.*, 2014a), Cs-G1 showed dominance of gene expression relative to Cs-G2, specifically in reproductive tissues.

Most *C. sativa* genes are yet to be functionally characterized, but by constructing tissue-specific gene co-expression networks we have created a framework for interpretation of the functions of uncharacterized genes. Furthermore, the analysis of timing and quantitative estimates of gene expression combined with prediction of gene function based on orthologous relationships with *A. thaliana* genes has established a platform for future experimental work aimed at understanding the functions of uncharacterized genes in *C. sativa*.

In conclusion, combined with the reference genome sequence, the comprehensive transcriptome atlas presented in this study provides a valuable resource for functional genomics and gene discovery, and elevates *C. sativa* to the status of a model crop for the predominantly polyploid Brassicaceae lineages (Kagale *et al.*, 2014b). An important next step will be to expand this resource by generating expression data from *C. sativa* plants exposed to various abiotic and biotic stresses. Furthermore, the highly undifferentiated hexaploid genome of *C. sativa* with limited evidence of functional diversification of genes has established a compelling need for generating similar proteomic, epigenomic and small RNA resources. Together these resources will help advance our understanding of the complex architecture of the regulatory mechanisms that govern biological processes in the undifferentiated polyploid genome of *C. sativa*.

EXPERIMENTAL PROCEDURES

Plant material

The doubled haploid genotype DH55 that was used for generating the reference genome sequence (Kagale *et al.*, 2014a) was selected for developmental transcriptome sequencing. Plants were grown in soil-less potting mixture in a controlled greenhouse environment (16-h light/8-h dark, 20/17°C). Twelve different tissue samples were collected at various developmental stages during the life cycle: germinating seed (36 h after imbibition), cotyledon (3 days after germination), young leaf (2 weeks after germination, WAG), senescing leaf (8 WAG), root and internodal stem (4 WAG), floral buds and flower (6–7 WAG) and four seed developmental stages (4–40 days post anthesis, DPA; Figure S5). Siliques were collected at regular intervals between 4 and 40 DPA. Seed developmental stages (Figure S5) were classified as early (siliques collected between 4 and 12 DPA were combined), early mid (16 and 21 DPA), late mid (25 and 29 DPA) and late (35 and 39 DPA). Whole siliques were used for RNA extraction. At least three independent biological replicates were analysed for each tissue type.

RNA isolation and transcriptome sequencing

RNA isolation, generation of Illumina libraries and sequencing were performed as detailed in Kagale *et al.* (2014a). In brief, total RNA from all tissues except whole siliques was isolated using the RNeasy Plant Mini Kit (Qiagen, http://www.qiagen.com/) according to the manufacturer's instructions, while total RNA was isolated from whole siliques using a method described by Suzuki *et al.* (2004). TruSeq RNA sequencing libraries were constructed following the standard preparation guide (Illumina, http://www.illumina.com/), 12 samples were multiplexed per lane of a flow cell and paired-end sequencing (100 cycles) was completed using the Illumina HiSeq 2000 platform.

Expression profiling

Before read mapping, expression quantification and assembly, all reads were filtered using Trimmomatic (version 0.17; Bolger *et al.*, 2014) by (i) removing adapter sequences, (ii) trimming leading and trailing low-quality sequences (below 15), (iii) removing sequences when the average quality per base dropped below 15 within a 4-base-wide sliding window, and (iv) keeping only those pairs where both reads were longer than 55 bp. Additionally, PCR duplicates were removed by finding read pairs with 100% identical sequences and gap sizes in a single sample, and culling all but one pair. Furthermore, trimmed reads with ambiguous residues (Ns; one or more) were removed entirely.

Clean reads were aligned end-to-end to the *C. sativa* reference genome (*Cs* genome sequence build V2.0; http://camelinadb.ca/downloads.html) with Tophat v.2.0.4 (Kim *et al.*, 2013), allowing a maximum of two mismatches in the alignment. Read alignments were reported only if both reads in a pair were mapped concordantly. The maximum and minimum intron lengths were set at 2500 and 20 bp, respectively. The expected inner distance between paired reads was set to be 100, with a standard deviation for the distribution set at 250. Default settings were used for all other parameters. Alignment results were then processed using the Cufflinks executable program within the Cufflinks v.2.0.2 suite of tools (Trapnell *et al.*, 2012) for gene and transcript quantification. The *C. sativa* reference annotation GFF file

(*Cs* gene annotation GFF; http://camelinadb.ca/downloads.html) was supplied, and all reference transcripts as well as any novel genes and isoforms were assembled. Multi-mapping read correction (-u/–multi-read-correct) was enabled. Again the maximum and minimum intron lengths were set at 2500 and 20 bp, respectively. Default settings were used for all other parameters. Transcript abundance was measured as an FPKM value.

As an alternative approach, estimation of transcript abundance was done using the novel program KALLISTO (Bray *et al.*, 2016). This program uses pseudoalignment for rapidly determining the compatibility of reads with targets. The KALLISTO index was built from *C. sativa* reference transcripts (*Cs* annotated CDS; http://camelinadb.ca/downloads.html) and the quantification algorithm was run with default parameters enabling sequence-based bias correction and plain text output of TPM and normalized count estimates.

As described in Kagale *et al.* (2014a) the relative expression values (TPM) were transformed to allow the use of statistical tests that assume a normal data distribution. By adding one and taking the natural logarithm of the TPM values the range of the data is reduced and the residuals more frequently follow a normal distribution (determined by the distribution of *P*-values for the Shapiro–Wilk test for normality). All subsequent calculations used the transformed values.

To assess the similarity among the biological replicates, pairwise Pearson correlation coefficients (*R*) between every two of the three replicates (using the TPM values) were calculated over all the 94 495 gene models, for each tissue type separately.

Differential expression analysis and tissue specificity

The significance of differences in the means of gene expression between tissues was assessed by ANOVA. The false discovery rate was calculated using the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995) and *P*-values were adjusted accordingly. Differences were considered statistically significant when the FDR-adjusted *P*-value was less than 0.01. Only differentially expressed genes were included in tissue specificity analysis. For each tissue sample, genes were classified as tissue specific if the mean expression for a particular gene across all other tissue samples was less than 0.1 TPM.

WGCNA analysis

Coexpression networks were constructed using the WGCNA package in R (Langfelder and Horvath, 2008). A signed weighted correlation network was constructed by creating a matrix of pairwise correlations between the 15 000 most differentially expressed (ANOVA *P*-value $< 2.88 \times 10^{-13}$) genes. WGCNA analysis, being a computationally intensive process, allows analysis of up to 15 000 genes at a time; hence the analysis was restricted to the 15 000 most variable genes (ANOVA; Data S4). The adjacency matrix was constructed by raising the co-expression measure to the power $\beta = 12$. The power of 12, which is the default value, is interpreted as a soft threshold of the correlation matrix. Modules were obtained using the automatic network construction function blockwiseModules with minModuleSize set to 30 and mergeCutHeight of 0.25. The Dynamic Hybrid Tree Cut algorithm was used to cut the hierarchal clustering tree.

Analysis of tissue-specific genome dominance

For each tissue type, ANOVAs were carried out for each gene triplet to test the significance of differences in the means of gene expression between the sub-genomes. This test was not possible for between 3.3 and 15.3% of gene triplets due to the presence of genes with zero expression, hence zero variance. For each gene triplet and tissue for which this could be done, the mean of ln(1 + TPM) was taken over the replicates to determine the sub-genome-representative gene having the highest expression. For each tissue, frequencies of the highest expressing genes (Figure 4) in each sub-genome were compared by the Pearson χ^2 test assuming in each case a null hypothesis in which two frequencies are the same, with the other fixed.

Identification of AS events

For the global AS analysis, transcript isoforms assembled using Tuxedo package (as described above in 'Expression profiling') from all tissue samples (including all three biological replicates) were mapped to corresponding gene models using CUFFCOMPARE (Trapnell *et al.*, 2012) and a GTF file containing the union of all Transfrags in replicate samples was generated. Identification of AS was performed using ASTALAVISTA v.3.2 (Foissac and Sammeth, 2007) run on a local server with command line parameters allowing retrieval of AS events in their complete resolution.

Hierarchical clustering of TF genes

Gene hierarchical clustering, as described by Kagale *et al.* (2014a), was completed for the TFs that showed variance for the 12 tissue means. In brief, one minus the Pearson sample correlation was applied as the distance measure and the average linkage method was used for clustering. The heatmap of the resultant clustering was drawn using the statistical package R.

Analysis of correlation of expression of TF genes in *A. thaliana* and their syntelogs in *C. sativa*

Arabidopsis TF expression data for nine different tissue types that were comparable to *C. sativa* tissue samples, including cotyledon, leaf, senescing leaf, root, flower and seed developmental stages 3, 5, 8 and 10 were accessed through the expression browser tool of the Bio-Array Resource for Arabidopsis functional genomics database (Toufighi *et al.*, 2005). The expression values were transformed by adding one and taking the natural logarithm. The mean transformed expression values for *A. thaliana* TFs were compared with the mean transformed expression values for corresponding syntelogs from each sub-genome (Cs-G1, Cs-G2 and Cs-G3) of *C. sativa.* For each gene, the means for the nine tissues were used in comparing any two genomes, i.e. the correlation test was done using the nine (x,y) data points. The analysis was done with the Spearman correlation because it is not sensitive to the distribution of the data (it is far from normal with the huge number of zeros), and the two-sided test was used that makes no assumption about the direction of the correlation (positive or negative).

Accession numbers

Sequence data from this paper can be found in the GenBank/EMBL data libraries under the following accession numbers: germinating seed (SRX472932), cotyledons (SRX472934), young leaf (SRX472935), senescing leaf (SRX472936), stem (SRX472937), root (SRX472938), buds (SRX472939), flower (SRX472941), early seed development (SRX472942), early mid seed development (SRX472943), late mid seed development (SRX472945) and late seed development (SRX472946).

ACKNOWLEDGEMENTS

This work was completed as part of the Prairie Gold Project that was funded through the Western Economic Partnership Agreement (WEPA) between the Government of Canada (Western

Economic Diversification) and the Government of Saskatchewan (Enterprise Saskatchewan). We would like to thank Raju Datla, National Research Council Canada, for critical reading of the manuscript. The authors declare no conflicts of interest.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Correlation between fragments per kilobase of exon per million fragments mapped (FPKM) and transcripts per kilobase million (TPM) expression profiles of *Camelina sativa* genes across the 12 tissue samples.

Figure S2. Hierarchical clustering dendrogram and heatmap generated based on expression of transcription factor genes in *Camelina sativa*

Figure S3. Correlation between expression profiles of orthologous transcription factor genes in *Arabidopsis thaliana* and *Camelina sativa*.

Figure S4. Functional conservation and diversification in *Camelina sativa* homeologues.

Figure S5. Seed developmental stages selected for RNA sequencing analysis.

Table S1. Characteristics of syntenic and non-syntenic genes in *Camelina sativa.*

Table S2. Lipid metabolism genes with seed-specific expression profiles.

Table S3. Pearson correlation test for analysing the differences between sub-genomes of *Camelina sativa* in each tissue.

Data S1. Expression levels (fragments per kilobase of exon per million fragments mapped values) of *Camelina sativa* genes across 12 different tissue samples collected at major developmental stages.

Data S2. Expression levels (transcripts per kilobase million values) of *Camelina sativa* genes across 12 different tissue samples collected at major developmental stages.

Data S3. Assessment of the replicate quality.

Data S4. ANOVA test to assess the differences in the means of gene expression between tissue types.

Data S5. Tissue-specific uniquely expressed genes in *Camelina sativa.*

Data S6. ANOVA test to assess the differences in gene expression of splicing factor genes between various tissue types.

Data S7. A syntelog matrix representing individual *Arabidopsis thaliana* transcription factor genes and the corresponding triplicated *Camelina sativa* homeologues.

REFERENCES

Abramovic, H. and Abram, V. (2005) Physio-chemical properties, composition and oxidative stability of *Camelina sativa* oil. *Food Technol. Biotechnol.* **43**, 63–70.

Benedito, V.A., Torres-Jerez, I., Murray, J.D. *et al.* (2008) A gene expression atlas of the model legume *Medicago truncatula*. *Plant J.* **55**, 504–513.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, **57**, 289–300.

Berretta, J. and Morillon, A. (2009) Pervasive transcription constitutes a new level of eukaryotic genome regulation. *EMBO Rep.* **10**, 973–982.

Black, D.L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.* **72**, 291–336.

Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Bray, N.L., Pimentel, H., Melsted, P. and Pachter, L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527.

Budin, J., Breene, W. and Putnam, D. (1995) Some compositional properties of camelina (*Camelina sativa* L. Crantz) seeds and oils. *J. Am. Oil Chem. Soc.* **72**, 309–315.

Champigny, M.J., Sung, W.W., Catana, V., Salwan, R., Summers, P.S., Dudley, S.A., Provart, N.J., Cameron, R.K., Golding, G.B. and Weretilnyk, E.A. (2013) RNA-Seq effectively monitors gene expression in *Eutrema salsugineum* plants growing in an extreme natural habitat and in controlled growth cabinet conditions. *BMC Genom.* **14**, 578.

Chen, M. and Manley, J.L. (2009) Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat. Rev. Mol. Cell Biol.* **10**, 741–754.

Chen, Z.J. and Ni, Z. (2006) Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. *BioEssays*, **28**, 240–252.

Cheng, F., Wu, J., Fang, L., Sun, S., Liu, B., Lin, K., Bonnema, G. and Wang, X. (2012) Biased gene fractionation and dominant gene expression among the subgenomes of *Brassica rapa*. *PLoS ONE*, **7**, e36442.

Clark, M.B., Amaral, P.P., Schlesinger, F.J. *et al.* (2011) The reality of pervasive transcription. *PLoS Biol.* **9**, e1000625; discussion e1001102.

Consortium, T.E.P (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

Eynck, C. and Falk, K.C. (2013) Camelina (*Camelina sativa*). In *Biofuel Crops: Production, Physiology and Genetics* (Singh, B.P. ed.). Wallingford, UK: CABI Publishing, pp. 369–391.

Filichkin, S.A., Priest, H.D., Givan, S.A., Shen, R., Bryant, D.W., Fox, S.E., Wong, W.K. and Mockler, T.C. (2010) Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Res.* **20**, 45–58.

Foissac, S. and Sammeth, M. (2007) ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. *Nucleic Acids Res.* **35**, W297–W299.

Guo, S., Zheng, Y., Joung, J.G., Liu, S., Zhang, Z., Crasta, O.R., Sobral, B.W., Xu, Y., Huang, S. and Fei, Z. (2010) Transcriptome sequencing and comparative analysis of cucumber flowers with different sex types. *BMC Genom.* **11**, 384.

Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y. and Barkai, N. (2002) Revealing modular organization in the yeast transcriptional network. *Nat. Genet.* **31**, 370–377.

Iven, T., Hornung, E., Heilmann, M. and Feussner, I. (2015) Synthesis of oleyl oleate wax esters in *Arabidopsis thaliana* and *Camelina sativa* seed oil. *Plant Biotechnol. J.* **14**, 252–259.

Jin, J., Zhang, H., Kong, L., Gao, G. and Luo, J. (2014) PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Res.* **42**, D1182–D1187.

Kagale, S., Koh, C., Nixon, J. *et al.* (2014a) The emerging biofuel crop *Camelina sativa* retains a highly undifferentiated hexaploid genome structure. *Nat. Commun.* **5**, 3706.

Kagale, S., Robinson, S.J., Nixon, J. *et al.* (2014b) Polyploid evolution of the Brassicaceae during the Cenozoic era. *Plant Cell*, **26**, 2777–2791.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36.

Langfelder, P. and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

Levy, A.A. and Feldman, M. (2002) The impact of polyploidy on grass genome evolution. *Plant Physiol.* **130**, 1587–1593.

Liang, C., Liu, X., Yiu, S.M. and Lim, B.L. (2013) De novo assembly and characterization of *Camelina sativa* transcriptome by paired-end sequencing. *BMC Genom.* **14**, 146.

Libault, M., Joshi, T., Benedito, V.A., Xu, D., Udvardi, M.K. and Stacey, G. (2009) Legume transcription factor genes: what makes legumes so special? *Plant Physiol.* **151**, 991–1001.

Li-Beisson, Y., Shorrosh, B., Beisson, F. *et al.* (2013) Acyl-lipid metabolism. *The Arabidopsis Book*, **11**, e0161.

Liu, X. and Clarke, N.D. (2002) Rationalization of gene regulation by a eukaryotic transcription factor: calculation of regulatory region occupancy from predicted binding affinities. *J. Mol. Biol.* **323**, 1–8.

Liu, J., Rice, A., McGlew, K., Shaw, V., Park, H., Clemente, T., Pollard, M., Ohlrogge, J. and Durrett, T.P. (2015) Metabolic engineering of oilseed crops to produce high levels of novel acetyl glyceride oils with reduced viscosity, freezing point and calorific value. *Plant Biotechnol. J.* **13**, 858–865.

Lu, C. and Kang, J. (2008) Generation of transgenic plants of a potential oilseed crop *Camelina sativa* by Agrobacterium-mediated transformation. *Plant Cell Rep.* **27**, 273–278.

Malik, M.R., Yang, W., Patterson, N. *et al.* (2015) Production of high levels of poly-3-hydroxybutyrate in plastids of *Camelina sativa* seeds. *Plant Biotechnol. J.* **13**, 675–688.

Marquez, Y., Brown, J.W., Simpson, C., Barta, A. and Kalyna, M. (2012) Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome Res.* **22**, 1184–1195.

Massie, C.E. and Mills, I.G. (2008) ChIPping away at gene regulation. *EMBO Rep.* **9**, 337–343.

Moser, B.R. (2012) Biodiesel from alternative oilseed feedstocks: camelina and field pennycress. *Biofuels*, **3**, 193–209.

Movahedi, S., Van de Peer, Y. and Vandepoele, K. (2011) Comparative network analysis reveals that tissue specificity and gene function are important factors influencing the mode of expression evolution in Arabidopsis and rice. *Plant Physiol.* **156**, 1316–1330.

Mudalkar, S., Golla, R., Ghatty, S. and Reddy, A.R. (2014) De novo transcriptome analysis of an imminent biofuel crop, *Camelina sativa* L. using Illumina GAIIX sequencing platform and identification of SSR markers. *Plant Mol. Biol.* **84**, 159–171.

Nguyen, H.T., Silva, J.E., Podicheti, R. *et al.* (2013) Camelina seed transcriptome: a tool for meal and oil improvement and translational research. *Plant Biotechnol. J.* **11**, 759–769.

Nguyen, H.T., Park, H., Koster, K.L., Cahoon, R.E., Nguyen, H.T., Shanklin, J., Clemente, T.E. and Cahoon, E.B. (2015) Redirection of metabolic flux for high levels of omega-7 monounsaturated fatty acid accumulation in camelina seeds. *Plant Biotechnol. J.* **13**, 38–50.

Parkin, I.A., Koh, C., Tang, H. *et al.* (2014) Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid *Brassica oleracea*. *Genome Biol.* **15**, R77.

Patel, R.V., Nahal, H.K., Breit, R. and Provart, N.J. (2012) BAR expressolog identification: expression profile similarity ranking of homologous genes in plant species. *Plant J.* **71**, 1038–1050.

Petrie, J.R., Shrestha, P., Belide, S. *et al.* (2014) Metabolic engineering *Camelina sativa* with fish oil-like levels of DHA. *PLoS ONE*, **9**, e85061.

Ponomarev, I., Wang, S., Zhang, L., Harris, R.A. and Mayfield, R.D. (2012) Gene coexpression networks in human brain identify epigenetic modifications in alcohol dependence. *J. Neurosci.* **32**, 1884–1897.

Reddy, A.S., Marquez, Y., Kalyna, M. and Barta, A. (2013) Complexity of the alternative splicing landscape in plants. *Plant Cell*, **25**, 3657–3683.

Rhee, S.Y. and Mutwil, M. (2014) Towards revealing the functions of all genes in plants. *Trends Plant Sci.* **19**, 212–221.

Ruiz-Lopez, N., Haslam, R.P., Napier, J.A. and Sayanova, O. (2014) Successful high-level accumulation of fish oil omega-3 long-chain polyunsaturated fatty acids in a transgenic oilseed crop. *Plant J.* **77**, 198–208.

Sammeth, M., Foissac, S. and Guigo, R. (2008) A general definition and nomenclature for alternative splicing events. *PLoS Comput. Biol.* **4**, e1000147.

Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Scholkopf, B., Weigel, D. and Lohmann, J.U. (2005) A gene expression map of *Arabidopsis thaliana* development. *Nat. Genet.* **37**, 501–506.

Schnable, J.C., Springer, N.M. and Freeling, M. (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc Natl Acad Sci U S A*, **108**, 4069–4074.

Shen, Y., Zhou, Z., Wang, Z. *et al.* (2014) Global dissection of alternative splicing in paleopolyploid soybean. *Plant Cell*, **26**, 996–1008.

Staiger, D. and Brown, J.W. (2013) Alternative splicing at the intersection of biological timing, development, and stress responses. *Plant Cell*, **25**, 3640–3656.

Stuart, J.M., Segal, E., Koller, D. and Kim, S.K. (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.

Suzuki, Y., Kawazu, T. and Koyama, H. (2004) RNA isolation from siliques, dry seeds, and other tissues of *Arabidopsis thaliana*. *Biotechniques*, **37**, 544.

Syed, N.H., Kalyna, M., Marquez, Y., Barta, A. and Brown, J.W. (2012) Alternative splicing in plants–coming of age. *Trends Plant Sci.* **17**, 616–623.

Tang, H., Woodhouse, M.R., Cheng, F., Schnable, J.C., Pedersen, B.S., Conant, G., Wang, X., Freeling, M. and Pires, J.C. (2012) Altered patterns of fractionation and exon deletions in Brassica rapa support a two-step model of paleohexaploidy. *Genetics*, **190**, 1563–1574.

Thatcher, S.R., Zhou, W., Leonard, A., Wang, B.B., Beatty, M., Zastrow-Hayes, G., Zhao, X., Baumgarten, A. and Li, B. (2014) Genome-wide analysis of alternative splicing in *Zea mays*: landscape and genetic regulation. *Plant Cell*, **26**, 3472–3487.

Toufighi, K., Brady, S.M., Austin, R., Ly, E. and Provart, N.J. (2005) The botany array resource: e-Northerns, expression angling, and promoter analyses. *Plant J.* **43**, 153–163.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L. and Pachter, L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578.

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A. and Luscombe, N.M. (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* **10**, 252–263.

Vitulo, N., Forcato, C., Carpinelli, E.C. *et al.* (2014) A deep survey of alternative splicing in grape reveals changes in the splicing machinery related to tissue, stress condition and genotype. *BMC Plant Biol.* **14**, 99.

Wagner, G.P., Kin, K. and Lynch, V.J. (2012) Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **131**, 281–285.

Walters, B., Lum, G., Sablok, G. and Min, X.J. (2013) Genome-wide landscape of alternative splicing events in Brachypodium distachyon. *DNA Res.* **20**, 163–171.

Wang, B.B. and Brendel, V. (2004) The ASRG database: identification and survey of Arabidopsis thaliana genes involved in pre-mRNA splicing. *Genome Biol.* **5**, R102.

Winter, D., Vinegar, B., Nahal, H., Ammar, R., Wilson, G.V. and Provart, N.J. (2007) An "Electronic Fluorescent Pictograph" browser for exploring and analyzing large-scale biological data sets. *PLoS ONE*, **2**, e718.

Yang, L., Takuno, S., Waters, E.R. and Gaut, B.S. (2011) Lowly expressed genes in *Arabidopsis thaliana* bear the signature of possible pseudogenization by promoter degradation. *Mol. Biol. Evol.* **28**, 1193–1203.

Zenoni, S., Ferrarini, A., Giacomelli, E., Xumerle, L., Fasoli, M., Malerba, G., Bellin, D., Pezzotti, M. and Delledonne, M. (2010) Characterization of transcriptional complexity during berry development in *Vitis vinifera* using RNA-Seq. *Plant Physiol.* **152**, 1787–1795.

Zhang, G., Guo, G., Hu, X. *et al.* (2010) Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Res.* **20**, 646–654.

Zou, C., Lehti-Shiu, M.D., Thibaud-Nissen, F., Prakash, T., Buell, C.R. and Shiu, S.H. (2009) Evolutionary and expression signatures of pseudogenes in Arabidopsis and rice. *Plant Physiol.* **151**, 3–15.