# Gene expression profiling identifies pathways involved in seeds maturation of Jatropha curcas

**Fatemeh Maghuly** ( fatemeh.maghuly@boku.ac.at )
Universitat fur Bodenkultur Wien

**Tamas Deak**
Szent Istvan Egyetem Kerteszettudomanyi Kar

**Klemens Vierlinger**
Austrian Institute of Technology GmbH

**Stephan Pabinger**
Austrian Institute of Technology GmbH

**Hakim Tafer**
Universitat fur Bodenkultur Wien

**Margit Laimer**
Universitat fur Bodenkultur Wien

**Research article**

**Keywords:** biofuel, gene expression, high-throughput quantitative real-time PCR, metabolic pathways, microarray, next generation sequencing

**Posted Date:** August 28th, 2019

**DOI:** https://doi.org/10.21203/rs.2.12916/v1



**Version of Record:** A version of this preprint was published on April 9th, 2020. See the published version at https://doi.org/10.1186/s12864-020-6666-1.

# Abstract

Background: Jatropha curcas, a tropical shrub, is a promising biofuel crop, which produces seeds with a high content of oil and protein. To better understand the development of its seeds to improve Jatropha`s agronomic performance, a two-step approach was performed: 1) generation of the entire transcriptome of six different maturation stages of J. curcas seeds using 454-Roche sequencing of a cDNA library, 2) comparison of transcriptional expression levels in six different developmental stages of seeds using a custom Agilent 8x60K oligonucleotide microarray. Results: A total of 793,875 high-quality reads were assembled into 19,841 unique full-length contigs, of which 13,705 could be annotated with Gene Ontology (GO) terms. Microarray data analysis identified 9,111 probes (out of 57,842 probes), which were differentially expressed between the six developmental stages. The expression results were validated for 70 randomly selected putative genes. Result from cluster analyses showed that transcripts related to sucrose, fatty acid, flavonoid, phenylpropanoid, lignin, hormone biosynthesis were over-represented in the early stage, while lipid storage, seed dormancy and maturation in the late stage. Generally, the expression of the most over-represented transcripts decrease in the last stage of seed maturation. Further, expression analyses of different maturation stages of J. curcas seed showed that most changes in transcript abundance occurred between the two last stages, suggesting that the timing of metabolic pathways during seed maturation in J. curcas is in late stages. The co-expression result showed a high degree of connectivity between genes that play essential role in fatty acid biosynthesis and nutrient mobilization. Furthermore, seed development and hormone pathways are significantly well connected. Conclusion: The obtained results revealed DESs regulating important pathways related to seed maturation, which could contribute to understanding the complex regulatory network during seed development. This study provides detailed information on transcription changes during J. curcas seed development and provides a starting point for a genomic survey of seed quality traits. The current results highlighted specific genes and processes relevant to the molecular mechanisms involved in Jatropha seed development, and it is anticipated that this data can be delivered to other Euphorbiaceae species of economic value.


# Background

Environmental protection and proper land use are some of the main concerns of mankind. Land degradation and dust are important environmental problems increased by climate change and drought occurring with urbanization and industrialization, which threaten the survival of more than 250 million people living in arid lands [1-2]. They affecting many problems created regarding health, social and economic impact of sustainable development. In addition, significant emissions of carbon dioxide ($CO_2$) and other greenhouse gasses into the atmosphere by the ignition of petroleum for various human activities and its reflection on global climate is quite obvious [3]. Actions to reduce the effects of climate change (https://www.apha.org/topics-and-issues/climate-change) offer an excellent opportunity to deliver further contributions to the control of both air pollution and greenhouse gas emission. Therefore, fuel derived from organic material (e.g. biofuel crops) receive more attention by the discussion in the shift from crude fossil oil to more sustainable resources [4]. Using agricultural crops including conventional

food plants as a first generation of biofuel overlapped with the food prices increase, leading to a worldwide food crisis. Therefore, the focus of biofuels production changed from first to second-generation, which use only non-food crops. Among the second-generation biofuel, *J. curcas* is an appealing crop plant, which is frequently mentioned as the best option for marginal soils with low nutrient levels, considered unsuitable for agriculture, and environments with low water reserves.

*J. curcas* grows in tropical and subtropical climates [5], between sea level and 1800 m, and is well adapted to semi-arid, arid conditions and regions with an annual rainfall range between 250 and 3000 mm [6]. *J. curcas*, a rapidly growing and easily propagated tree, is a multi-purpose plant for biodiesel supply, medicinal uses, veterinary purposes and livestock feed [7-8]. The oil quality obtained from the Jatropha oil methyl ester is very similar to the values of conventional diesel fuel and can be used without any modification in the existing diesel engine [9].

The *J. curcas* seed cake, which is a waste by-product of the biodiesel trans-esterification process, can be used for the production of various supplies such as organic fertilizer, high-quality paper, energy pellets, soap, cosmetics, toothpaste, embalming fluid, pipe joint cement and cough medicine [10]. *J. curcas* seed contains no edible oils, which are traditionally used for soap production and medicinal usage [11]. In addition, because of its therapeutic characteristics, people with various skin diseases and sensitivity to regular soap [12] use its solvents. All traits mentioned above make *J. curcas* one of the best candidate as a profitable biofuel crop species for restoring wastelands and improving employment chances and subsistence in rural areas [1, 13].

*J. curcas* seed kernels are rich in oil (54–58%), and protein (20–28%) compare to the shell, and several efforts were made to use seed cake or kernel meal remaining after extraction of oil [8]. However, a range of toxins and anti-nutritional compounds render the seedcake and oil unsuitable for use as animal feed and human consumption [14]. While the agronomic traits of *J. curcas* are already good, efforts are required to increase oil yield, content and composition by improving the ability of the plant to produce favorable fruits/seeds. Breeding efforts in this biofuel crop will be accelerated by the in-depth knowledge of seed transcripts of *J. curcas* for obtaining functional genomics information to discover genes that encode enzymes involved in the biosynthesis of oil and toxins precursors and to describe their relevant metabolic pathways. Therefore, it is necessary to establish a reliable method to characterize the temporal shifts in gene expression that underlie the biochemical and metabolic processes occurring during seed maturation. Furthermore, such data could help to identify, characterize and if necessary modify the possible transcripts of interest.

Among different high throughput analysis tools for large-scale comparisons of gene expression data, a microarray is a robust method used in many plant species. One of its main advantages is the comparison of a constant large number of genes for the expression evaluation of different genotypes, organs, and tissues [15]. Thus, we performed the following analyses: 1) generation of the entire transcriptome of six stages of *J. curcas* seed maturation using 454-Roche sequencing from pooled samples; 2) comparison

of transcriptome expression in seeds of six stages of seed maturation using custom Agilent 8x60K oligonucleotide gene expression microarrays.

In Jatropha, transcriptome studies generated data describing seed development and seed germination from manually pollinated plants [16-19]. Considering that Jatropha flowering and fruiting are fairly continuing, which results in the presence of mature and immature fruit at the same time on a plant, the purpose of this study was to provide for the first time an overview of transcripts that are potentially involved in seed maturation of open pollinated plants. These data show that most changes in transcript abundance occur between the last stages (brown and black epicarp), suggesting that the timing of metabolic pathways during seed maturation in *J. curcas* is in late stages.

# Methods

### Plant material

Flowering and fruiting of Jatropha are fairly continuous, which results in the presence of mature and immature fruits at the same time on the same plant. Seeds of a selected *J. curcas* accession in Kamisse, Ethiopia, Agricultural Research Center Adama, were collected at six maturation stages (I-VI) and characterized according to the color of epicarp and endocarp [green-white (I), green-brown (II), green-black (III), yellow-black (IV), brown-black (V), dry-black (VI)] [20] at the same time. Three biological replicates were used for each sample, immediately frozen in liquid nitrogen and stored at - 80 °C.

### Total RNA extraction

Total RNA was extracted from six stage of seed maturation of *J. curcas* using plant RNA purification reagents (Invitrogen) according to the supplier's instructions. The quality and concentration of total RNAs were determined using NanoVue Spectrophotometer (GE Healthcare Life Sciences) and gel electrophoresis. All RNA samples showing A260/280 ratios between 2.0 and 2.15 were selected and analyzed for RNA integrity using an Agilent 2100 Bioanalyzer (Agilent Technologies). RNA samples with an integrity number above 7.0 were used for further analyses.

### cDNA synthesis for sequencing

Equal amounts of extracted RNA from different seed maturation stages were pooled and used for cDNA library construction. To purified mRNA from 5 µg total RNA, the mRNA-Only Eukaryotic mRNA Isolation Kit (Epicentre) was used by exonuclease digestion followed by LiCl precipitation. One µg mRNA was used for the synthesis of the first-strand cDNA by the Mint-Universal cDNA Synthesis Kit (Evrogen). The Trimmer Kit (Evrogen) was used for normalization reaction using 800 ng amplified cDNA, which was re-amplified by 18 cycles.

## Size selection and cloning of cDNA

Two µg of normalized cDNA were digested by ten units of the *SfiI* restriction enzyme (New England Biolabs) for 2 hours at 48°C. Fragments (>800 bp) isolated from an LMP Agarose Gel were purified using the MinElute Gel Extraction Kit (Qiagen). The Fast Ligation Kit (New England Biolabs) was used for ligation of 200 ng purified cDNA fragments to 100 ng *SfiI* using dephosphorylated pDNR-lib Vector (Clontech). The product was desalted by ethanol precipitation and re-dissolved in 10 µl water. Out of them, 1.5 µl were used to transform NEB10b competent cells (New England Biolabs). To verify the success of normalization, 96 clones were randomly selected and sequenced.

## cDNA library preparation and sequencing using Roche 454 FLX

One million clones were plated on LB-Cm agar plates, collected and stored in glycerol stocks at -70°C. One-half of the cells were inoculated to a 300 ml Terrific Broth/Cm culture and were grown for 5 hours at 30°C. 100 Units SfiI digested 200 µg of purified plasmid DNA (Qiagen) for 2 hours at 48°C. LMP-Agarose/MinElute Gel Extraction Kit was used to purify inserts, which were ligated to high-molecular-weight DNA using a Sfi-linker.

The library for the Roche 454 FLX sequencing was generated according to the manufacturer's protocols (Roche/454 Life Sciences). The concatenated inserts were shared to fragments ranging from 400 to 900 bp. The two 454 A and B adaptors were ligated to the ends of the emulsion PCR and sequencing. The library was sequenced on one picotiter-plate of the GS FLX using the Roche/454 Titanium chemistry.

## Assembly of the sequence reads to transcripts

At first, the reads were screened for the Sfi-linker used for concatenation and linker sequences were removed. The Roche/454 Newbler software (454 Life Sciences Corporation, Software Release 2.3) at default setting was used to assemble clean reads to individual transcripts. All unique sequences with an average length of > 100 bp were used for oligonucleotide microarray design.

## Annotation of whole seed transcripts

The obtained sequences were annotated using the pipeline version of Blast2Go v2.5.0 [21]. Additional information was added to the annotation database from an InterProScan v5RC6 analysis of the sequences [22]. For protein-based similarity search, a protein sequence database of the reads was set up.

Amino acid sequences for each read were defined as the most extended open reading frame of the sequence.

Specific homology searches were carried out for three distinct molecular functions of special interest: transcription factors, transporters and resistance gene analogues.

To identify transcription factors, DNA binding domain alignments were fetched from the Plant Transcription Factor (TF) Database [23], Hidden Markov Models (HMMs) were built based on the alignment and sequence reads were searched for these DNA binding domain models using HMMER3 [24].

Transporters were predicted based on sequence homology search using BLAST [25] against sequence entries for the Transporter Classification (TC) Database [26]. Sequence hits with an E-value lower than 1e-100 were considered as transporters of the respective class.

Reference R-gene sequences (112 genes) were acquired from the Plant Resistance Genes (PRG) database [27] and InterPro (IPR) domains were identified for the reference sequences with InterProScan. To predict Resistance Gene Analogue (RGA) sequences, Blast2GO and InterProScan annotation tables were filtered for these IPR domains.

Further genes or functions of interest were analyzed using text-based searches (curcins, storage proteins) in the Blast2Go/InterProScan annotation or based on the enzyme codes also included in a Blast2Go annotation table.

## GO annotation

Blast2GO was used to obtain the GO information. Further, the functional annotation was used to refine annotation, and specific GO terms were labeled with their putative Biological Process (BP), Molecular Function (MF), and Cellular Component (CC). Furthermore, GO IDs were used to assign enzyme commission (EC) numbers and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [28] to contigs.

## Microarray oligo probe design

The probes were designed (Genotypic Technology LTD.) for an 8 x 60K oligonucleotide gene expression microarray (Agilent Technologies) using all unique sequence of whole seed transcript (contigs) from the transcriptome of different maturation stages, both in sense and antisense orientation using the Agilent's eArray software (https://earray.chem.agilent.com/earray/). A set of unique sequences was established as a database, and the probes were designed by tiling the contig sequences against the database. Probes specific to each transcript were selected for cross-hybridization when showing a hit with at least 30 bp and > 84% identity. Best probes were considered those showing single hits in the BLAST results.

## Probe labeling, hybridization, and detection

RNA labeling, hybridization onto Agilent 8x60K oligonucleotide microarrays as well as scanning and raw data analysis was carried out according to the One-Color Microarray-Based Gene Expression Analysis Protocol provided by Agilent Technologies. Total RNA (200 ng) from each sample was used to synthesize cyanine-3 labeled cRNA using the QuickAmp Labeling kit, one Color and RNA Spike-In kit one Color (Agilent Technologies). The cyanine labeled cRNA was transcribed and purified by a T7 polymerase and RNeasy mini kits (Qiagen), respectively. Samples labeled with Cy3 (825 ng) were hybridized for 17 hours at 65°C and 10 rpm in the hybridization oven using the Gene Expression Hybridization kit (Agilent Technologies). The arrays were washed according to supplier´s instructions and scanned on an Agilent G2505C scanner at 3 μm resolution. Data were acquired using Agilent Feature Extraction software version 10.5.1.1. Each sample was hybridized using three biological replicates, onto microarrays.

## Microarray data analyses

The R statistical (http://www.Rproject.org) and Bioconductor software [29] were used to perform the pre-processing analyses of Agilent 8x60K oligonucleotide gene expression microarrays data. Hybridized microarray slides were imaged with a high-resolution array scanner, and fluorescence signal intensities from each spot were quantified. Background correction was performed using Agilent spatial detrending background estimate, followed by averaging of replicate spots, $\log_2$-transformation, KNN (K nearest neighbor) imputation of missing values and quantile normalization. The linear modeling functions of the LIMMA package were used for inference statistics [30]. Statistical significance was determined by *t*-statistic for seeds and corresponding *P*-values. Genes with Benjamini-Hochberg false discovery rate (FDR) corrected *P*-value < 1e-8, were considered as significantly differentially expressed in different stages of seed maturation and leaf samples. Clustering was performed using normalized and filtered data. The differentially expressed sequences (DESs) were clustered according to their expression patterns across seeds. Normal mixture modeling for model-based clustering (expectation-maximization) was performed with p < 1e-8 [31]. The obtained microarray data in this study were stored in the Gene Expression Omnibus (GEO) (GSE109931).

## GO set enrichment analyses

Gene set enrichment analyses were carried out according to the GO/KEGG terms using the Bioconductor GOstats package [32]. Since *Jatropha* is not a supported model organism, the complete GO/KEGG categories for the differentially expressed genes were identified, using the Blast2GO annotation file of the whole seed transcriptome. After building the gene-set collection, the corresponding parameter object was

created followed by hyper-geometric testing. The differentially expressed genes of different maturation stages of seeds or clusters were analyzed for both over- and under-representation of GO terms, whereby KEGG and each GO category (BP, CC, MF) were analyzed separately.

Results of the GOstats analyses were plotted as flipped bar charts displaying each identified term using the negative log10 *P*-value for the top 15 terms. Both over- and under-represented GO terms were combined in one graph, showing the *P*-value of the over-represented term on the right side and the under-represented term on the left side. In addition, heatmaps were created using the negative log10 *P*-value.

## Co-expression network analysis

Co-expression networks based on partial correlations were calculated using the DESs with a *P*-value <1e-8 (f-statistics of the model) as described earlier [33-34]. Co-expression network was constructed from the 300 most significant edges. Nodes were colored according to the cluster membership from cluster analyses of DESs, and the number of connections for the top 20 nodes with the highest number of connections to other nodes was constructed as bar plots.

Further, the genes that were expected to be involved in seed development, seed storage, and hormone cross-talking were extracted by searching the available GO annotations of each contig. A partial correlation network was constructed, and a network of the top 50 most significant edges was extracted.

## Quantitative Real-Time (qRT)-PCR using BioMark

The BioMark qRT-PCR system (Fluidigm) in combination with a 48.48 Dynamic array was used according to the manufacturer's instructions. Primers for selected contigs from the microarray and housekeeping genes were designed using Primer3 software [35]. cDNA synthesis was performed on a total of 48 samples, including 42 test samples, four standard control samples, and two nucleases free water (negative control) samples. For standard control, a reference sample was prepared, consisting of an equivalent pool of all test samples. In each 20 µl reaction 100 ng of total RNA per test sample as well as 800, 200, 50 and 12.5 ng of reference RNA sample (as standard control) and only water in the two negative control samples were reverse transcribed, using the SuperScript III First-Strand Synthesis System Kit (Invitrogen) according to the manufacturer`s protocol. The RT reactions were diluted 1:3, and 1.25 µl of each dilution was applied to 4 different 5 µl pre-amplification reactions, each containing 1x Qiagen PCR buffer, 0.8 mM of dNTPs, 0.25 µl of DMSO, 0.15 Unit of HotStar Taq DNA polymerase (Qiagen) and a pool of 48 different primer pairs (200nM each). Cycling conditions for pre-amplification were 15 min at 95°C and 14 cycles of 40 s at 95°C, 40 s at 60°C and 80 s at 72°C. The cycle ended with a final step of 7 min at 72°C. After pre-amplification, products were diluted 1:5 in nuclease-free water. QPCR amplification was performed using the BioMark system (Fluidigm) and 48.48 Dynamic Arrays. For each qPCR run 6 µl

sample mix were prepared, consisting of 1x Qiagen PCR Buffer (including 1.5mM $MgCl_2$), 0.4 mM $MgCl_2$, 0.96 mM dNTPs, 0.3 µl DMSO, 1x EvaGreen Binding dye, 0.18 units HotStarTaq Polymerase (QIAGEN HotStarTaq$^{TM}$ PCR), 0.004 µl ROX, 1x DNA binding dye (Fluidigm) and 1.5 µl of pre-amplified and 1:5 diluted samples. In parallel, six µl of assay mix were prepared, including three µl of 2x Assay Loading reagent (Fluidigm), 0.3 µl nuclease free water and 2.7 µl of 200nM primer pair pool used for pre-amplification of test samples. 48.48 Dynamic Arrays were primed, sample mix, as well as assay mix, were loaded with the integrated fluidic circuit (IFC) controller MX (Fluidigm) and qPCR was performed using the BioMark system (Fluidigm) according to the manufacturer's instructions. Cycling conditions were 15 min at 95°C and 40 cycles of 40 s at 95°C, 40 s at 60°C and 80 s at 72°C. A final step of 7 min at 72°C ended the cycle. Ct values were calculated using the Fluidigm Real-Time PCR Analysis Software 4.1.2. Similar to microarray data analyses, qPCR data were normalized using quantile normalization, and linear models were calculated using the LIMMA package.

# Results

### Whole seed transcriptome sequencing

To cover the entire *J. curcas* seed transcriptome, total RNA was extracted from six stages of seed maturation, and equal amounts of total RNA from each sample were pooled together. From this pool, mRNA was isolated and reverse transcribed into cDNA. Normalized cDNA libraries were generated and sequenced using the GS FLX Titanium. Sequencing of cDNA libraries yielded a total of 793,875 high quality (HQ) reads with an average read length of 358 nucleotides.

After trimming and cleaning, a total of 603,459 HQ reads were assembled into 19,841 contigs (unique transcripts) containing 13,171,840 bases using the cDNA assembly feature of Newbler software v.2.3. (Roche). Out of them, 48,978 reads were identified as singletons. The size of contigs ranged from 100 to 4,088 bases, with 1,035 bases as N50 contig size.

### Functional annotation of whole transcript sequencing data

All 19,841 unique contigs were analyzed by Blast2GO [21] and aligned using BLASTX [36] search in the NCBI non-redundant nucleotide database using an E-value threshold of 1e-6. In total, 14,994 unique contigs were identified through BLAST search. The taxonomic distribution was examined, and over 8,000 transcripts had top hits to *Ricinus communis* (Figure S1), a species closely related to *Jatropha*. Based on text-annotation from the BLAST hits, approximately 3% (618 contigs) of the transcripts showed top BLAST hits with uncharacterized proteins, and 23% (4,847 contigs) had no significant similarity to any sequence in the public dataset. Generally, sequences without BLAST hits could not be annotated. Generally, sequences without BLAST hits could not be annotated. The average size of annotated and unannotated contigs was 800 and 400 bp, respectively.

Transcripts that show match with a BLAST were further classified into GO terms for biological process (BP), molecular function (MF), and cellular component (CC) (Figure 1). A total number of 13,705 *Jatropha* unique transcripts received at least one GO annotation (Table S1). Functional annotation of genes of the *J. curcas* library indicated that the highest percentage of GO terms was found in the category BP, containing 2,469 GO terms, followed by 1,891 in MF and 521 in CC (Table S1, Figure 1). The most abundance GO terms in the BP category were genes involved in oxidation-reduction processes (GO:0055114, 1,002 contigs, 5.1%), DNA-templated regulation of transcription (GO:0006355, 611 contigs, 3,1%) and response to cadmium (GO:0046686, 437 contigs, 2.2%). In CC, the most representative categories were nucleus (GO:0005634, 2,761 contigs, 13.9%), plasma membrane (GO:0005886, 1,796 contigs, 9.1%) and chloroplast (GO:0009507, 1,442 contigs, 7.3%). Within MF, the largest content of functionally assigned ESTs were related to ATP binding (GO:0005524, 1,285 contigs, 6.5%), zinc ion binding (GO:0008270, 975 contigs, 4.9%) and DNA binding (GO:0003677, 725 contigs, 3.7%).

InterProScan search was also used during the annotation process, using InterPro [22]. From the 19,841 contigs, 37.6% were also annotated based on homology to sequences in the InterPro database (Table S2). Out of the 13,705 sequences annotated with GO terms, 4,399 contigs were assigned with 5,703 EC numbers representing 1,040 unique enzymes, 799 of which are assigned to one or more KEGG pathways (Table S2). In addition, the data were compared with the Jatropha genomic sequences of Kazusa DNA Research Institute (JAT_r4.5, ftp://ftp.kazusa.or.jp/pub/Jatropha/)[37-38] and Chinese Academy of Sciences (JatCur_1.0, ftp://ftp.ncbi.nih.gov/ genomes/Jatropha_curcas/)[39-40] (Table S2). In total 73.7% (14,616 contigs), 1.3% (251 contigs), 13.4 % (2,671 contigs) showed sequence similarity to both, only to JAT_r4.5 or only to JatCur_1.0 databases, respectively. However, 11.61% (2,303 contigs) were found additionally in the transcript data set of the current study.

Additionally, 1001 different contigs were identified belonging to 20 transporter classes (Table S2). Out of them 233 and 246 contigs belong to the transporter classes 2A (porters) and 3A (p-p-bond-hydrolysis-driven transporters), respectively.

Protein domain characteristics for Resistance Gene Analogs (RGAs) have been identified in 85 contigs, with 70 carrying a kinase domain, 35 of which also harbored an additional serine/threonine (Ser-Thr) site. RGAs that contain Ser-Thr domain can phosphorylate serine and thereonine residues, which are involved in plant development, signaling and defense [41]. However, some RGAs like the *Pto* gene from tomato encode only Ser-Thr protein kinase. Four contigs with a nucleotide-binding site (NBS-ARC) domain and eight contigs with a leucine-rich repeat (LRR) domain were found. Both domains are abundantly present in plants and have an ATPase activity [42].

Moreover, 600 contigs from 52 different classes could be identified as TFs (Table S2). The most abundant TF families were MYB-related, MYB, bZIP, AP2, ERF and RAV, represented by 66, 54, 49, 49, 49 and 46 contigs, respectively.

Since a major goal of seed oil crop research is focused on oil quality and quantity, it is necessary to understand the processes involved in seed metabolism, especially in the accumulation of storage

compounds like carbohydrates, triacylglycerols (TAGs) and proteins. To reconstruct the metabolic pathways active during *J. curcas* seed development, annotated sequences were mapped to KEGG pathways using the Blast2GO platform [25,28] in search of their biological function. As shown in Figure 2, in total, 2,660 contigs were located on 144 KEGG pathways (Table S3). Using the KEGG classifications allowed us to identify that the most highly represented pathways were purine metabolism (315 contigs), followed by starch and sucrose metabolism (226), pyrimidine metabolism (154) and phenylalanine metabolism (138). Further, glycolysis (129), pyruvate metabolism (111), flavonoid biosynthesis (111), glycerolipid metabolism (103) and phenylpropanoid biosynthesis (96) also were found in the top 20 highly represented pathways.

## Array design and Microarray analyses

The examination of the changes in mRNA expression at selected points should allow the identification of genes associated with maturation and secondary metabolite biosynthesis.

An 8x60k oligonucleotide microarray containing 57,842 unique probes was produced from 19,841 transcriptome contigs (as described above). The probes were designed in sense and antisense direction with an average probe spacing of 250 bp (500bp sense + 500bp antisense) using the server based eArray platform from Agilent Technologies (https://earray.chem.agilent.com/earray/). In total 31,875 specific probes and 2,604 cross-hybridizing probes (Xhyb) in sense direction, as well as 21,680 specific probes and 1,683 Xhyb in antisense direction, were designed. The oligo microarrays were hybridized with probes of six stages of seeds maturation, each with three biological replicates.

## Genome-wide variation in transcript expression during seed maturation

The microarray data were normalized and, differential expression patterns were identified using statistical analyses. Specific transcriptional profiles were established for each evaluated interaction. Further, genes exhibiting differential expression among seed maturation stages were submitted to *in silico* evaluations and were classified and categorized by their possible molecular function and involvement in metabolic pathways. Principle components analysis (PCA) on transcript expression (abundance) of 57,842 probes showed a clear separation of the six different stages related to seed maturation along the first principle component (PC1); which explained 53% of total variation, and was associated mostly with variation in transcript expression over the maturation stages, where expression from stage IV and V were closer to each other (Figure 3). Significant changes in transcript expression (abundance) were observed in the first and last stages, suggesting a higher physiological differentiation in these stages. Further, the values of probes on PC1 showed that probes with very low negative value were related to the early stage of seed maturation (I and II), while the probes with high positive value were in response to the later stage (IV-V)

[43]. Furthermore, biological replicates of each stage clustered together, suggesting a minimal variation between replicates.

To identify changes in gene expression patterns during seed maturation, average $\log_2$ expression levels across all microarray data (greater than two-fold changes) with a cut-off of *P*-value < 1e-8 was performed (Table S4 and Figure S2). A total of 9,111 differentially expressed probes (16% of the total probes) from 7,299 contigs (38% of the total contigs) within the six stages of seed maturation were identified.

The co-expression patterns from the significant cut-off showed two intermodular hubs with around 40 edges and a broad range of nodes displaying between 15 and 10 edges (Figure 4A-B). The CB5-D hub with the highest number of edges belongs to a cytochrome B5 isoforms. HVA22J, an abscisic acid (ABA)- and a stress-inducible gene [44] from cluster 3 connected CB5-D from cluster 5 to aspartic proteinase (CDR1) from cluster 3, Dihydroflavonol reductase (DFR) from cluster 3 and Transparent Testa 8 (TT8) from cluster 6 (Figure 4A-B).

To focus on processes expected to be involved in seed storage and seed developments as well as hormone pathway, the GO terms from BP category were extracted. The co-expression results showed a high degree of connectivity between seed development and hormone pathways, while seed storage is less well connected to the other two pathways (Figure 4C).

Based on pairwise comparison between seed developmental, 135 and 231 contigs (300 and 144 probes) were up- and down-regulated between stage I and II, 426 and 535 contigs (479 and 659 probes) between stage II and III, 275 and 188 contigs (321 and 188 probes) between III and IV, 40 and 106 contigs (40 and 124 probes) between IV and V and finally the most contigs 889 and 272 (1,122 and 1,673 probes) were found between stages V and VI (Table S5). Data showed that most changes in transcript abundance occurred between last stages, suggesting that the timing of metabolic pathways during seed maturation in *J. curcas* is in late stages.

## Cluster analysis of differentially expressed sequences (DESs)

To understand patterns of co-expression of the DESs during seed development, their expression profiles were clustered using hierarchical clustering analysis and visualized as a heatmap. The cluster analysis showed that different expression patterns could be classified into ten major clusters (1-10) (Figure 5-6). Up-regulated transcripts, whose expression was increased during seed maturation, were enriched in clusters 2, 4, 8, 9 and 10 (group A), while down-regulated transcripts were enriched in clusters 1, 3, 5, 6 and 7 (group B). The biological replicates of each stage clustered together, which underlines their high similarity. Further, each maturation stage of *Jatropha* seeds could be clearly separated from the others.

Cluster 1 (1,136 probes corresponding to 892 contigs) represented contigs with higher expression at earlier stages (green-half) which decreased over time. Cluster 2 (1,061 probes corresponding to 894

contigs) and 4 (1,069 probes corresponding to 882 contigs) showed slightly increased expression during maturation. Cluster 3 (409 probes corresponding to 336 contigs) and 5 (525 probes corresponding to 393 contigs) contained contigs with higher expression at earlier stages (I and II, in clusters 3 and 5, respectively), which decreased over time. Cluster 6 contained most probes (1,478 probes corresponding to 1,180 contigs) and revealed the highest expression in stage I, and decreased expression levels during maturation, while in cluster 7 (929 probes corresponding to 732 contigs) expression decreased from stage III (Figure 5). Both clusters showed a slightly higher expression in stage V. Cluster 8 (642 probes corresponding to 468 contigs) on the other hand, showed increased expression levels from early to later stages. Both cluster 9 (1,004 probes corresponding to 810 contigs) and 10 (858 probes corresponding to 712 contigs) showed increased expression from stage I to V and a decrease in stage VI (Table S4).

Based on the Transporter Classification Database, 431 transcripts (699 probes) were assigned to 94 transporter subfamilies, 15 subclasses and 7 classes. These transcripts are distributed among the 10 clusters, representing the intense activities during maturation process, which requires transport of metabolites within the cell and between different parts of the seed. The highest number (151) of transcripts related to transport activities were identified in class 3 (Primary active transporters), followed by class 2 (Electrochemical potential-driven transporters) with 151 transcripts. 13 transcripts were classified to be transporter subfamily 1.A.33, which is related to heat shock protein (Hsp) 70 (Table S4, Figure S2). Furthermore, different kinds of sugar transporters (2.A.1) and ATP/ADP transmembrane transport (2.A.29) represent the role of transporters to provide necessary energy metabolism during seed maturation. Among 24 transcripts that were classified as ABC transporters (3.A.1), subfamilies A, C, E, F, G and I were identified (Table S4).

In addition, 58 families of TFs showed differential expression between the six seed development stages, involving all expression pattern clusters (Table S4). The highest number of transcripts related to TFs were found in cluster 1 (54) flowed with cluster 6 (48), while the least number of TFs were found in cluster 3 (11). The most abundant TF families were identified as AP2/ERF-RAV (29).

## GO enrichment analyses of DESs

To better understand the biological function of DESs during seed maturation, a GO set enrichment for each cluster was performed to identify the primary functional categories and pathways of these transcript clusters. GO enrichment analyses in BP categories for each cluster indicated that the most significantly over-represented DESs were found in cluster 6 containing 656 contigs with 314 GO terms, followed by cluster 4, with 509 contigs, while the most significantly under-represented DEs were also found in cluster 6, followed by cluster 2, with 570 contigs (detailed information and TOP 15 GO terms for each cluster are shown in Figure 7, S3, Table 1 and S6).

Visualization of enriched GO terms related to BP category for cluster group A, with higher DES expression level during late stage of seed maturation showed, that in cluster 8 GOs related to fatty acid metabolism

(e.g. unsaturated fatty acid, linoleic acid), lipid storage, dormancy process, response to alcohol and UDP-glucose metabolism were significantly over-represented, while transports were under-represented. Further, GOs related to aromatic acid transports were over-represented in clusters 9 and 10, while GOs related to monoterpenoid metabolism were only enriched in cluster 10. Raffinose family oligossacharides (RFOs), which are associated with late maturation in *Arabidopsis*, *Brasica napus* and *Medicago trunculata* [45-47] were over represented in cluster 10. Transcripts related to biosynthesis of serine and glycine were over-represented in cluster 9. Embryo sac development, RNA modification and methylation were significantly over-represented in cluster 4, whereas, transcripts related to maintenance of seed dormancy, protein folding and RNA modification in cluster 2.

In contrast, GOs related to phenylpropanoid and flavonoid metabolism and biosynthesis, as well as cell wall modification and carbohydrate metabolism in clusters 3 and 5 (cluster group B) were significantly enriched with high expression levels during early stage of seed maturation (Figure S2). Transcripts involved in hormone transporters were significantly enriched in cluster 6, while sequences related to signaling were enriched in both clusters 6 and 7. Genes associated with ATP hydrolysis coupled protein transport and purine ribonuclease metabolism were enriched in cluster 7, however, RNA metabolic process, RNA processing, and cellular lipid metabolic processes were significantly under-represented in cluster 7.

Glucan and beta-glucan biosynthesis, which plays a key role in regulating seed coat-imposed dormancy [48] were over-represented in two clusters (3 and 8, respectively). GOs, involved in translation, RNA processing, and gene expression were under-represented in two different clusters (6 and 9), with different pattern of gene expression during maturation, indicating the involvement of two different groups of genes.

## KEGG enrichment analysis of DESs

To further understand the biological function of DESs during seed development, enriched KEGG pathways in the set of DESs was assessed (Table 2 and S6, Figure S4). The KEGG pathways with *P*-value < 0.05 were considered as significantly enriched.

## Pathway enrichment in DESs related to carbohydrate metabolism

Clusters from different stages of seeds share an enrichment of starch and sucrose metabolism. According to KEGG pathway analysis, starch and sucrose metabolisms contributed to clusters 5 with 14 enzymes.

Data revealed that carbohydrate metabolism was the most over-represented pathway in developing seeds of *J. curcas*, containing 90 enzymes involved in 11 pathways, belonging to cluster group B, which

expression were decreased during maturation. The glycolysis/gluconeogenesis pathway was found to be significantly enriched in cluster 5 and 7. As shown in cluster 7 (Figure 5), over-represented contigs related to phosphoglucomutase (EC:5.4.2.2) catalyzing the reversible interconversion between glucose-6P and glucose-1P, the latter serving as a substrate for ADP-Glc pyrophosphorylase, the first deposited step in the starch biosynthesis pathway, indicate the reduction of starch biosynthesis during developmental stages [49-50]. The KEGG enrichment analysis of over-represented GO terms in the different clusters showed a reduced expression during seed maturation in 6 clusters with 158 contigs and 80 enzymes in 12 out of 15 pathways related to carbohydrate metabolism (ko00010, ko00020, ko00030, ko00040, ko00051, ko00053, ko00500, ko00520, ko00620, ko00630, ko00640, ko00660, Figure S5), mainly in clusters 3, 6 and 7. It is well known, that seed maturation is associated with a significant reduction of most sugars, organic acids, and amino acids, which relates to reserve accumulation and using the power of carbon compounds for the synthesis of fatty acids [51].

### Pathway enrichment in DESs related to lipid metabolism

Pathways contributing to lipid biosynthesis can be divided into three steps and cell compartments: the first step of fatty acid biosynthesis occurs in the plastids, triacylglycerol (TAG) biosynthesis in the endoplasmic reticulum (ER) and oil body formation released into the cytoplasm.

From the first step, we could identify 3-oxoacyl-ACP reductase (KAR, EC:1.1.1.100, contig01408, contig07257, contig14836, in cluster 1 and contig02217, contig21412 in clusters 8). This enzyme produces beta-hydroxyacyl-AC, and following various reaction butyryl-ACP with four carbon is subsequently generated. Beta-ketoacyl-ACP synthase I is involved in the synthesis of palmitolyo-ACP with 16 carbon (KASI, EC:2.3.1.41, contig02910 in clusters 5 and contig01218 in cluster 8) which in the current study shows two different regulation patterns. The elongation from 16:0-ACP or 18:1-ACP [52] occurs in plastid by acyl-ACP desaturase (AAD, EC:1.14.19.2, contig03293, contig16368, contig18144 in cluster 5 and contig19191 in cluster 8). Oleoyl-ACP hydrolase (OAH, EC:3.1.2.14, contig04873 in cluster 1 and contig14233 in cluster 5) removes acyl group from ACP. Acyl-CoA synthetase (EC:6.2.1.3, contig05825, contig13897) enriched in cluster 8 is engaged in glycerophospholipid metabolism and fatty acid elongation.

Conversion of mono-unsaturated fatty acids to poly-unsaturated fatty acids by certain desaturases occur in the ER and are transferred to diacylglycerol (DAG) to produce triacylglycerol (TAG) by phospholipid: diacylglycerol acyltransferase (PDAT1, EC:2.3.1.158, contig02496, contig13416, in cluster 8). In the ER, glycerol kinase catalyzes glycerol to glycerol-3-phosphate (G3P), an initial substrate in the Kennedy pathway. Acylation of G3P, the first step of the Kennedy pathway, occurs by the G3P acyltransferase (GPAT) and lysophosphatidic acid (LPA) is formed, which is then converted to phosphatidic acid (PA) by LPA acyltransferase (LPAAT, EC:2.3.1.51, contig15092, contig19182 in cluster 2 and contig00099, contig02040, contig15096 in cluster 4). These products can be dephosphorylated by PA phosphatase (EC:3.1.3.4, contig03775 in cluster 2) to DAG, an essential intermediate in the biosynthesis of

phosphatidylcholine (PC). Two contigs (contig04042, contig07701), belonging to cluster 10, encode diacylglycerol O-acyltransferase (DGAT, EC:2.3.1.20), contig02496 and contig13416 correspond to phospholipid diacylglycerol acyltransferase (PDAT, EC:2.3.1.158, in cluster 8), which are then acylating DAG to produce TAG. TAG is modified by triacylglycerol lipase (EC:3.1.1.3, contig02004, contig04368, contig12631, contig12633, contig13268 in cluster 4 and contig12671 in clusters 8) into fatty acids. Expression patterns of clusters 2, 4 and 8, showed an increase during the last stages of *J. curcas* seed development.

Contigs encoding enzymes like diacylglycerol kinase (ATP, EC:2.7.1.107, contig12203, contig16160 in cluster 4), aldehyde dehydrogenase (NAD+, EC:1.2.1.3, contig03833 in cluster 2, contig12312 in cluster 4 and contig03833 in cluster 8) and glycerate 3-kinase (EC:2.7.1.31, contig06990, contig 19842 in cluster 2), were found in cluster 2, 4 and 8. As expected, the synthesis of fatty acids requires a high amount of energy during seed maturation, which results in increased expression of enzymes related to photosynthesis as an energy supply [53].

Further, two important enzymes involved in alpha-linolenic acid metabolism and biosynthesis of unsaturated fatty acids were identified in cluster 1: acyl-CoA oxidase (EC:1.3.3.6, contig08139) and enoyl-CoA hydratase/3-hydroxyacyl-CoA dehydrogenase (EC:4.2.1.17, contig05058, contig09979).

Most of the enzymes involved in lipid biosynthesis in *J. curcas* were identified in the current study based on the annotation of the seed transcripts. Altogether, 97 contigs and 55 enzymes were annotated as being involved in lipid metabolism (Figure S6).

### Pathway enrichment in DESs related to phenylpropanoid biosynthesis or monolignol biosynthesis

Among the significantly enriched pathways, the phenylpropanoid biosynthesis pathway contains 30 over-represented contigs and 11 enzymes located in clusters 3, 5 and 8 (Figure S7). Lignin, a complex of phenylpropanoid polymers, is the second most abundant polymer after cellulose, mainly located in the cell wall supporting the plant with structural stability [54]. The lignin deposition and lignification happen in seed coat cells, siliques cells, tracheary elements, sclerenchyma cells and endodermal cells [55]. The first step begins with the deamination of phenylalanine to cinnamic acid by phenylalanine ammonia-lyase (PAL, EC:4.3.1.24) [56] and phenylalanine/tyrosine ammonia-lyase (EC:4.3.1.25, contig05064, contig05269 in cluster 3 and contig05269 in cluster 5) in the phenylalanine, tyrosine, and tryptophan biosynthesis pathways. Next, one contig corresponding to trans-cinnamate 4-monooxygenase (EC:1.14.13.11, contig00044 in cluster 5) converts cinnamic acid to P-coumaric acid. P-coumaric acid can be conjugated by 4-coumarate: CoA ligase (4CL, EC 6.2.1.12, contig07594, contig07611, contig08856 in cluster 3 and contig07611, contig09522 in cluster 5), and enriched to coenzyme A to form p-coumaroyl-CoA, which is the precursor for the synthesis of flavonoids, stilbenes, and other phenylpropanoids [57].

Alternatively, P-coumaric acid and p-coumaroyl-CoA can be hydroxylated by p-coumarate 3-hydroxylase (EC:1.14.13.-) to produce caffeic acid and Cafeoyl-CoA. The newly added hydroxyl group can be methylated by caffeic acid 3-O-methyltransferase (COMT, EC:2.1.1.68) and caffeoyl-CoA O-methyltransferase (CCoA-OMT; EC:2.1.1.104, contig12200 in clusters 3 and 5, and contig11569, contig12199, contig21555 in cluster 8), ferulic acid and feruloyl-CoA. Ferulic acid is hydroxylated by ferulate-5-hydroxylase (EC:1.14.-.-) to form 5-hydroxyferulic acid; however, so far the enzyme responsible for hydroxylation of feruloyl-COA to 5-hydroxyferuloyl-COA is unknown.

Cinnamoyl-CoA reductase (CCR, EC:1.2.1.44) can catalyze the CoA thioesters to the corresponding aldehydes, which are reduced to monolignols by cinnamyl alcohol dehydrogenase (EC:1.1.1.195) and peroxidase (EC:1.11.1.7, contig02903, contig08820, contig14072, contig14904, in cluster 3 and contig00763, contig02903, contig09344, contig11593, contig1160, contig14503, contig21004 in cluster 5). Eleven contigs for peroxidase were significantly enriched and over-represented. The presence of different patterns of gene expression for an enzyme, e.g. peroxidase, may indicate the presence of different isoenzymes. Interestingly, the expression of peroxidase was twice as high in the last stage compared to the first stage of seed development (Figure 5). A previous study by Gijzen et al. [58] suggested that the highly expressed soybean peroxidase (Prx2) in the seed coat may be subjected to degradation during seed maturation.

## Pathway enrichment in DESs of flavonoid biosynthesis-related pathways

p-Coumaroyl-CoA, which is synthesized in the monolignol biosynthetic pathway, can also lead to flavonoid biosynthesis. The decrease of carbon flow results in a limitation for the flavonoid pathway, leading to an increase of monolignols by p-coumaroyl-CoA [59]. However, flavonoid biosynthesis pathway is well conserved among plants [60].

After oil extraction, the *Jatropha* seed cake contains high amounts of polyphenols and pigments that result from flavonoid biosynthesis. In the current study were annotated and enriched in clusters 1, 3 and five as encoding 16 enzymes involved in flavonoid, flavone and flavonol biosynthesis and isoflavonoid biosynthesis based on KEGG pathways (Figure S8). Two contigs (contig11261, contig19461, cluster 1) were identified as 6'-deoxychalcone synthase (EC:2.3.1.170) and three contigs (contig16238, contig17619, contig19806 in clusters 3 and 5) as naringenin-chalcone synthase (EC:2.3.1.74), an important enzyme catalyzing the conversion of cinnamoyl-CoA to pinocembrin chalcone. Chalcone synthase is the first committed enzyme in this pathway and is a major regulator of flavonol biosynthesis, also playing an important role in flower development [61]. One contig was annotated for chalcone isomers (EC:5.5.1.6, contig04004, cluster 3) that catalyzes the conversion of pinocembrin chalcone to pinocembrin, a substrate for galangin synthesis [61]. Furthermore, it catalyzes 4,2´,4´,6´-tetrahydroxychalchone into naringenin. Four contigs (contig00198, contig01835, contig10674, contig14763) were identified as flavanone 3-dioxygenase or naringenin 3-dioxygenase (EC:1.14.11.9, contig00198, contig01835, contig14763 in cluster 5), which are involved in highly conserved pathways in

plants to catalyze naringenin into dihydrokaempferol, an important intermediate product, that can be converted to kaempferol by flavonol synthase (EC:1.14.11.23, contig00445, contig00715, contig14188, contig19609, contig19806 in cluster 1, contig08632, contig12686 in cluster 3 and contig12095, contig12100, contig12686 in cluster 5). Thirty-seven contigs of flavonoid biosynthesis pathway were significantly enriched in *J. curcas* seeds. The presence of different expression patterns (from different clusters) of one enzyme could be explained by the existence of different isoenzymes and possibly by the interaction with other genes involved in flavonoid biosynthesis at multiple loci [62].

Kaempferol can be hydroxylated and glycosylated at its position six by flavonol 3-O-glucosyltransferase (EC:2.4.1.91, contig16343, contig21456 in cluster 1) to 6-hydroxykaempferol-3-O- β-D-glucoside or 6-hydroxykaempferol-3,6-di-O-β-D-glucoside. However, kaempferol can also be hydroxylated by flavonoid 3'-monooxygenase (EC:1.14.13.21, contig06562 in cluster 1 and contig05713, contig06160 in cluster 3) or flavonoid 3', 5'-hydroxylase (EC: 1.14.13.88) to quercetin in *J. curcas*. Both enzymes can also convert dihydrokaempferol to dihydrocyanidin, which can be deoxidized by dihydroflavonol-4-reductase to leucoanthocyanidin, and further by leucoanthocyanidin reductase to catechin.

### DESs related to hormone pathways and seed development

Several transcripts related to plant hormones like brassinosteroids (BR), abscisic acid (ABA), ethylene, and jasmonic acid (JA) were identified in different clusters. Cytochrome P450 family 1 subfamily A polypeptide 1 (EC:1.14.14.1, contig04060, contig16501 in cluster 8 and contig01390, contig04280, contig05390, contig16501 in cluster 10), corticosteroid 11-beta-dehydrogenase isozyme 1 (EC:1.1.1.146, contig02785, contig06502, contig16374 in cluster 8 and contig11252 in cluster 10), 17beta-estradiol 17-dehydrogenase (EC:1.1.1.62, contig11561) and 3beta-hydroxy-Delta5-steroid dehydrogenase / steroid Delta-isomerase (EC:1.1.1.145, contig00630, contig00631) are all involved in brassinosteroid (BR) biosynthesis via mevalonic acid (MVA) and steroid biosynthesis pathways. Interestingly, most enzymes were found only in cluster 10. The enzymes involved in this pathway were slightly up-regulated during seed maturation and showed a higher expression level in the last stages compared to early stages (Figure 5). BR is a steroid hormone, which acts throughout the plant growth, early fruit development, and regulates seed size and seed number [63]. The functions of BR in seed development were demonstrated using insensitive mutants of *Arabidopsis*, *Oryza sativa*, *Pisum sativum*, and *Vicia faba* [64]. Over-expression of a P450 monooxygenase family gene CYP72C1 in *Arabidopsis* dwarf mutant shk1-D produced a lower endogenous BR level and limited growth and smaller seeds [65]. Also, the characterization of a cytochrome P450 gene in a dwarf rice mutant implicated in BR biosynthesis led to a reduction of seed length [66]. Further, the effect of cytochrome P450 of *J. curcas* on seed size were examined in transgenic tobacco [67]. Over-expression of a BR-biosynthetic gene in rice could increase seed filling and yield. A BR-deficient mutant of *Vicia faba* and pea resulted in smaller seeds [68-69]. Although molecular mechanisms of BR involved in seed development are unclear, these facts show that BR is required for normal seed development and determines seed size [64].

Abscisic acid (ABA), on the other hand, is known to influence seed development, plant growth and dormancy, it regulates the synthesis of seed storage proteins, starch, and lipids as well as the withstanding of environmental stresses. Carotenoid biosynthesis pathway drives ABA synthesis from C40 epoxycarotenoid precursors in plastids. From the carotenoid biosynthesis pathway, four enzymes with four contigs were over-represented in clusters 8 and 10, and their expression increased during seed maturation. Among them, 9-cis-epoxycarotenoid dioxygenase (NECD, EC:1.13.11.51; contig02415) and (+)-abscisic acid 8'-hydroxylase (EC:1.14.13.93; contig17785) involved in ABA synthesis were identified to be differentially expressed in our dataset. Over two-fold up-regulation of these enzymes in stage IV may indicate the significant increase in ABA biosynthesis. Basnet et al. [43] also showed that ethylene and ABA from hormonal metabolism were over-represented during seed development in different *Brassica rapa* accessions, which suggests that their abundance increased during seed development. These data are in agreement with other studies [43,70-73], where ABA and ethylene were up-regulated during late stages of seed development, underpinning the role of both hormones in growth and development of seed, accumulation of seed reserves and their beneficial impact on germination, maturation, desiccation tolerance and induction of seed dormancy.

The biosynthesis of ethylene, known to be generally involved in plant development, fruit maturation and senescence, occurs through two amino acid pathways: alanine, aspartate and glutamate metabolism and cysteine and methionine metabolism. Both pathways produce l-aspartate, which is the substrate for the formation of aminocyclopropanecarboxylate (ACC) an immediate precursor for ethylene using methionine adenosyltransferase (SAM-synthase, EC:2.5.1.6, contig00728, contig03867, contig07733, contig15674 in cluster 6 and contig07733 in cluster 7), and for the last step aminocyclopropanecarboxylate oxidase (EC:1.14.17.4, contig07599 in cluster 6, contig02109 in cluster 7) which form ethylene. Both enzymes were found in clusters 6 and 7, which also contained 15 enzymes corresponding to 24 contigs in the cysteine and methionine metabolism pathway. Among them, three enzymes involved in ethylene biosynthesis were found, i.e., methionine synthase (EC:2.1.1.13, contig03731) in cluster 7, aspartate kinase (EC:2.7.2.4, contig03307, contig13782) and homocysteine S-methyltransferase (EC:2.1.1.10, contig04946) both in cluster 6. Finally, in stage VI, the enzymes were down-regulated compared to stage I (Figure 5). Ketring and Morgan [74] found that low levels of ethylene are produced by dormant peanut seeds.

Jasmonic acid (JA) acts in fruit ripening, seed germination, protein storage and resistance or response to biotic and abiotic stresses as well as plant development, and is synthesized from the alpha-linolenic acid pathway. It is the only hormone that is biosynthesized from the fatty acid or alpha-linolenic acid metabolism pathway, which was found to be significantly enriched with five enzymes and nine contigs in cluster 1. It is well known that JA biosynthesis requires the action of acyl-CoA oxidase (EC:1.3.3.6, contig08139), which was found in cluster 1. However, the expression of this enzyme is reduced during seed maturation. Soybean lines with the acyl-CoA oxidase antisense construct were used to reduce the level of acyl-CoA oxidase, which increased germination [75]. Treatment of acx1/5 JA deficient plants with JA restored normal seed set [76].

### Validation of microarray data using qRT-PCR

A total of 70 contigs from the DESs represented transcripts in seeds, and three housekeeping genes were selected (Table S7) and used for independent validation using a 48.48 chip (Fluidigm) to confirm that the changes in expressions as indicated by microarray data were authentic and reliable. Candidates for qPCR were chosen based on expression levels, known function, clusters and length of contigs. Additionally, some contigs of unknown function were selected. The corresponding primers are listed in Table S7.

The expression patterns obtained by qRT-PCR correlate strongly to moderately with data from the microarray analyses (about half of the contigs correlate with the microarray data at a Pearson correlation < -0.8), thus confirming the reliability of the chosen approach (Figure S9).

## Discussion

The understanding of transcriptional variation during different seed development stages is of utmost importance for breeding strategies in *J. curcas*, especially for high-quality oil suitable for biodiesel and low toxin levels, which could make it appropriate for animal feed. In the current study, genome-wide transcriptome analysis was used to identify the global gene expression pattern of *J. curcas* seeds at six-time points of its development. Even though *J. curcas* is an important oil crop, this is the first study profiling genome-wide transcript expression during seed development and maturation of open pollinated plants.

The sequencing of the whole seed transcriptome of *J. curcas* allowed the design of an 80 x 60k oligonucleotide microarray platform containing 57,842 unique probes created from 19,841 contigs as templates. Comparison of contigs with different Jatropha genome sequences [37-40] revealed 2,303 contigs additionally in the transcript data set of the current study. However, 4,847 contigs had no significant similarity to any sequence in the public dataset, and therefore could not be annotated. The high number of unannotated transcripts might be an indication for potential limitations in transcriptome assembly and annotation. The unannotated sequences could include both novel transcripts and technical artifacts from the sequencing technology (library preparation and/or sequencing machine). Additionally, the applied BLAST parameters are optimized for complex full-length RNA sequences, which does not favor BLAST searches of short (150-200 bp) contigs of low complexity.

Different approaches were used to identify sets of genes with various transcript abundance during seed development. *First*, to obtain an overview of the variation in seed developmental stages, PCA was carried out, using all transcripts present in the microarray (Figure 3). The first principle component (PC1: 53% explained variation) captured mostly temporal variation in transcript abundance, which is supported by previous studies in *Brassica rapa* [43], and *Arabidopsis thaliana* [51,77], where seed developmental stages are the major source of transcriptional and metabolic variation.

*Second*, the center of our attention was on transcripts related to Jatropha seed maturation to correlate co-expression patterns within pathways and to anticipate putative regulatory elements of the metabolisms of interest (Figure 4). The co-expression result showed a high degree of connectivity between seed development and hormone pathways, while seed storage is less well connected to the other two pathways. Furthermore, The co-expression patterns from the significant cut-off showed the CB5-D hub has the highest number of edges, followed by unannotated contigs, CDR1,TT8, and HVA22 (Figure 4A-B). CB5 a small tail-anchored membrane proteins play an essential role in many cellular processes, including lipid biosynthesis. It is well known that Genome-wide analysis of coordinated transcript abundance during seed development in different *Brassica rapa* morphotypes provide electrons for various enzymes located in the endoplasmic reticulum (ER), including fatty acid desaturase (FAD) and FAD-like proteins. They also are physiological important for p450 protein family [78-81]. Hwang et al. [78], combined *in vivo* and *in vitro* assays to show that CB5-D are targeted exclusively to mitochondrial outer membranes, while the other isoforms of CB5 (A, B, and C) are targeted to the ER. On the other hand, the contig02686 (unannotation) from cluster 6 also contains a high number of edges, which connected to some hypothetical proteins. There is a need for functional annotation of this contig, which might support important biological cell functions and could potentially serve as targets for further studies. TT8 has an essential role in the regulation of flavonoid biosynthesis and the formation of seed coat color; however, recently Chen et al.[82] reported that TT8 also affect FA biosynthesis in seed of Arabidopsis maternally, which also inhibits FA accumulation by down-regulating of the expression of a carboxylase biotin carboxylase subunit (CAC2), beta-ketoacyl-acp synthetaseII (KASII), mosaic death1 (MOD1), fatty acid biosynthesis2 (FAB2), fata acyl-acp thioesterase (FatA), fatty acid elongation1 (FAE1), FAD2 and FAD3, playing important for FA biosynthesis during seed maturation. The TT8 also repressed the expression of leafy cotyledon1 (LEC1), LEC2, FUSCA3 (FUS3), and cytidinediphosphate diacylglycerol synthase2 (CDS2), which are critical to embryonic development [82]. On the other hand, TT8 impact DFR expression, which commits phenolics to proanthocyanidins synthesis responsible for seed coat and quality germplasm of canola [83]. Although proteins with hydrolase activity like CDR1 do not imply the production of seed oil, overexpression of microsomal diacylglycerol acyltransferase 1 (DGAT1, EC 2.3.1.20) – a key enzyme of triacylglycerol production – resulted in differential regulation of CDR1 expression in transgenic and untransformed control in Brassica [84]. Eventually, HVA22 was identified to be an ER- and Golgi-localized protein and is capable of regulating GA-mediated vacuolation negatively [44]. They also suggested that ABA induces the accumulation of HVA22 proteins to inhibit vesicular trafficking involved in the nutrient mobilization, which could delay coalescence of protein storage vacuoles, resulting in regulating seed germination and seedling growth.

*Third*, a subset of probes with variation in transcript abundance patterns between developmental stages based on *P*-value of < 1e-8 was selected for further analyses. This subset of genes were present in different clusters, which were enriched in various metabolic pathways such as fatty acid biosynthesis, flavonoid biosynthesis, glucan metabolic biosynthesis, seed maturation and dormancy, sucrose and hormone metabolic processes. In addition, pairwise transcript expression analyses of different maturation stages of *J. curcas* seeds showed that most changes in transcript abundance occurred

between stages V and VI (Table S5), suggesting that the timing of metabolic pathways during seed maturation in *J. curcas* is in late stages. The expression results were validated for 70 putative transcripts.

*Fourth*, considering the intense metabolic activity during seed maturation, which requires the regulation of target genes and the exchange of metabolites and proteins between different locations in seed and within the cell, it is important to identify transcripts related to TF and transport machinery.

Among the transcripts that were classified as a transporter, subfamily 1.A.33 were related to Hsps, which perform diverse biological functions in collaboration with chaperons either in stress or non-stress conditions. In the absence of heat stress, *Hsp* genes are accumulated during the late stage of seed maturation [85]. Several plant cytosolic Hsp70 were identified during development, maturation, and germination of seeds of pea and Arabidopsis [86-87]. In this study, three transcripts were identified as a homolog to BiP1 of plant Hsp70 family, appeared to be highly expressed in the early stage of seed maturation (cluster 6). Previous reports suggest that BiP1 has a role in protein quality and in seed maturation, which also appears to be highly expressed in all tissues [88-89].

In the group of ABC transporters (3.A.1), one transcript showed homology to AtABCG14 (cluster 5 and 10), that is involved in translocation of cytokinins between the root and the shoots in *Arabidopsis* [90-91]. This transcript could be essential for long-distance communication between root- shoot-fruit as well. We also identified one ABCG reporter transcript in cluster 10 homologous to *Arabidopsis* AtABCG25 reported to act as a carrier to export ABA from the vascular tissues, where it is mainly produced [92]. In addition, one homolog of AtABCD1 in cluster 4 and cluster 9 were found. AtABCD1 facilitates the transport of lipidic metabolites in *Arabidopsis* [93]. In addition, transcripts associated with transporters AtABCC2 and AtABCI17 expressed in cluster 7 and 9, respectively, which were reported to be involved in transport of toxic compounds in Arabidopsis. AtABCC2 are tolerated to metal and act as chlorophyll catabolic transporter, while AtABCI17 is expressed in roots and is highly sensitive to Aluminium [94]. It is clear that plant ABC transporters play an important role for survival of plant and seed maturation, however many questions remain to be answered since only a few of the plant ABC transporters were functionally analyzed (22 out of 130 in Arabidopsis) [93-94].

Transcripts related to TF are distributed among all clusters with different expression patterns. The most abundance one, AP2/ERF-RAV, has been studied extensively in the regulation of seed maturation and ABA-regulated gene expression in the Arabidopsis [95]. Additionally, bHLH, bZIP, MYB family and the other motifs, which represented in all clusters probably, play an essential role in regulating of gene expression during maturation of *J. curcas* seeds. In soybean, bHLH, ARF and MYB were related to cell division and cell expansion, while in Arabidopsis they could lead to cell expansion and extra cell division, which increase seed size and weight [96]. A bZIP motif was reported for triacylglycerol biosynthesis genes in *Arabidopsis* and *Brassica rapa* seeds [43,77]. Also, as reported in various studies [97-98], DOF TF families activate genes related to seed storage protein during seed maturation and germination, which explain the presence of five transcripts with highest expression level at the late stage of seed maturation (clusters 2, 4, 8, and 9), and 2 transcripts with highest expression in early stages (cluster 1 and 5).

Transcripts with the highest expression level in the early stage of seed maturation (clusters 1, 5 and 6) belong to FAR1 and WRKY TF families. Several homologues of WRKY were found to be related to the induction of longevity in *Medicago trunculata* and soybean [99-100]. WRKY transcription factors are also involved in germination and growth by regulating ABA signaling in soybean and *Arabidopsis* [100-101], and activated during late seed maturation [102]. FAR1, a positive regulator of chlorophyll biosynthesis, plays a key role in plant growth and development by being involved in a wide range of cellular process such as light signal transduction, flowering time regulation, and abscisic acid response [103].

*Finally,* cluster analyses were used to discover particular seed maturation-dependent patterns of gene expression. Transcripts related to fatty acid, flavonoid, lignin and sucrose biosynthesis were over-represented in the early stage, while lipid storage, seed dormancy and maturation in the late stage.

Generally, the expression of the most over-represented transcripts decrease in the last stage of seed maturation. The transcripts involved in the expression of the triacylglycerol and FA desaturation biosynthesis processes increased during middle to late stages of seed development, something previously reported in *Brassica rapa* [43]. This could be explained by the rise of storage lipid production [104], which is also confirmed by previous research studies [16-19,105-107].

Considering that the expression level of most contigs related to flavonoid biosynthesis was more than two times down-regulated in the last stage compared to early stages in all three clusters (1, 3, 5) (Figure 5) suggests that the genes involved in flavonoid biosynthesis may be essential during the early stages of seed development [62]. The expression pattern of major flavonoid biosynthesis genes was also down-regulated during seed development in *Arabidopsis thaliana* [108]. However, the expression of each contig did not follow a similar pattern during later stages, and even contigs of the enzyme EC: 1.14.13.88, present in clusters 1 (contig04825 and contig04272) and 3 (contig03370), which may indicate the presence of different isoenzymes and functions.

This study adds to the previous research and shows that temporal transcriptional variation is more dominant than phenotypic variation (toxic and none toxic of *J. curcas* accessions, data not shown).

The obtained results revealed DESs regulating important pathways related to seed maturation, which could contribute to understanding the complex regulatory network during seed development.

This study also provides detailed information on transcription changes during *J. curcas* seed development and provides a starting point for a genetically genomic survey of seed quality traits. The current results highlighted specific genes and processes relevant to the molecular mechanisms involved in Jatropha seed development, and it is anticipated that this data can be delivered to other *Euphorbiaceae* species of economic value.

# Abbreviations

**PBU**: Plant Biotechnology Unit **GO**: Gene Ontology **DESs**: Differentially Expressed Sequences **CO2**: carbon dioxide ***J. curcas***: *Jatropha curcas* **TF**: Transcription Factor **HMM**s: Hidden Markov Models **IPR**: InterPro **RGA**: Resistance Gene Analogue **BP**: Biological Process **MF**: Molecular Function **CC**: Cellular component **EC**: enzyme commission **KEGG:** Kyoto Encyclopedia of Genes and Genomes **KNN**: K nearest neighbor **FDR:** False Discovery Rate **GEO**: Gene Expression Omnibus HQ: High Quality **Xhyb**: cross-hybridizing probes **PCA**: Principle Components Analysis **DFR**: Dihydroflavonol reductase **CDR1**: aspartic proteinase **DFR**: Dihydroflavonol reductase **TT8**: Transparent Testa 8 **ER**: endoplasmic reticulum **CAC2**: Carboxylase biotin carboxylase subunit **KASII**: beta-ketoacyl-acp synthetaseII **MOD1**: mosaic death1 **FAB2**: fatty acid biosynthesis2 **FatA**: fata acyl-acp thioesterase **FAE1**: fatty acid elongation1 **FAD2**: fatty acid desaturase2 **LEC1**: leafy cotyledon1 **DGAT1**: diacylglycerol acyltransferase1 **FUS3**: FUSCA3 **CDS2**: cytidinediphosphate diacylglycerol synthase2 **ACC**: acetyl-CoA carboxylase **MCMT:** malonyl-CoA acyl carrier protein (ACP) transacylase **KAS III**: beta-ketoacyl-ACP synthase III **KAR**: 3-oxoacyl-ACP reductase **KASI**: palmitolyo-ACP **KASII**: beta-ketoacyl-ACP synthase II **AAD**: acyl-ACP desaturase, **OAH**: Oleoyl-ACP hydrolase **FAT**: acyl-ACP thioesterase **DAG**: diacylglycerol **TAG**: triacylglycerol **PDAT1**: phospholipid: diacylglycerol acyltransferase **G3P**: glycerol-3-phosphate **GPAT**: G3P acyltransferase **LPA**: lysophosphatidic acid **PA**: phosphatidic acid **LPAAT**: LPA acyltransferase **PC**: phosphatidylcholine **DGAT**: diacylglycerol O-acyltransferase **PDAT**: phospholipid diacylglycerol acyltransferase **ATP**: diacylglycerol kinase **NAD**: aldehyde dehydrogenase **PAL**: phenylalanine ammonia-lyase **4CL**: 4-coumarate: CoA ligase **COMT:** caffeic acid 3-O-methyltransferase **CCoA-OMT**: caffeoyl-CoA O-methyltransferase **CCR**: Cinnamoyl-CoA reductase **ACC**: aminocyclopropanecarboxylate

# Declarations

### *Ethical approval and consent to participate*

Not applicable.

### *Consent for publication*

Not applicable.

### *Availability of supporting data*

The whole dataset generated and analyzed during the current study is available from the corresponding authors on request.

### *Competing interests*

The authors declare that they have no competing interests.

### *Funding*

This research was financially supported by the Austrian Science Fund (FWF, P 23836). Authors thank the Austrian Research Promotion Agency (FFG), and Bioplant R&D, Vienna for the financial support.

### *Authors' contributions*

FM designed and carried out all experiments, participated in data analyses and wrote the final manuscript. TD, KV, SP, HT participated in data analyses and manuscript editing. All authors read and approved the final manuscript.

### Acknowledgements

We would like to thanks Debesaye Senbeto (Agricultural Research Center Adama, Melkasa, Ethiopia) and Yoseph Tewodros for their help in providing Jatropha samples.

# References

1. Bayen P, Sop T K, Lykke A M, Thiombiano Does *Jatropha curcas* L show resistance to drought in the Sahelian zone of West Africa? A case study from Burkina Faso Solid Earth, 2015; 6: 525-531.
2. Sop T K, Kagambèga F W, Bellefontaine R, Schmiedel U, Thiombiano A. Effects of organic amendment on early growth performance of *Jatropha curcas* L on a severely degraded site in the Sub-Sahel of Burkina Faso. Agroforest Syst, 2011; 86: 387-399.
3. Ovando-Medina I, Espinosa-García F, Núñez-Farfán J, Salvador-Figueroa M. Does Biodiesel from *Jatropha curcas* Represent a Sustainable Alternative Energy Source? Sustainability, 2009; 1: 1035-1041.
4. Blesgraaf RAR. Water use of Jatropha, Hydrological impacts of *Jatropha curcas* L. MSC Thesis, 2009; Delft University of Technology.
5. Grass M. *Jatropha curcas* L: Visions and Realities. Journal of Agriculture and Rural Development in the Tropics and Subtropics, 2009; 110: 29-38.
6. Rajaona A M, Sutterer N, Asch F. Potential of waste water use for Jatropha cultivation in arid environments. Agriculture, 2012; 2: 376-392.
7. Maghuly, F. and Laimer, M. Jatropha curcas, a biofuel crop: Functional genomics for understanding metabolic pathways and genetic improvement. Biotechnology Journal, 2013; 8: 1172-1182.
8. Chhetri A B, Tango M S, Budge S M, Watts K C, Islam M R. Non-Edible Plant Oils as New Sources for Biodiesel Production International journal of molecular sciences. 2008; 9: 169-180.
9. Rao P V, Rao G S. Production and characterization of Jatropha oil methyl ester international. Journal of Engineering Research, 2013; 2: 141-145.
10. Canu F A, Lyndrup M, Riiser N M, Jensen T Y H W. Off-Grid Electrification in Mozambique Based on Renewable Energy Sources - A Contribution to the EDENR Strategy. 2012; https://booksgoogleca/books?id=FwpZnQAACAAJ.

11. Agbogidi, O M, Akparobi, S O, Eruotor, P G. Health and environmental benefits of *Jatropha curcas* linn. App Sci Re, 2013; 1: 36-39.
12. Warra A A. Cosmetic potentials of physic nut (*Jatropha curcas* Linn) seed oil: A review. American Journal of Scientific and Industrial Research, 2012; 3: 358-366.
13. Achten W, Nielsen L, Aerts R, Lengkeek A, Kjær E D, Trabucco A, Hansen J K, Maes W H, Graudal L, Akinnifesi F K, Muys B. Towards domestication of *Jatropha curcas.* Biofuel, 2010; 1: 91-107.
14. Sabandar C W, Ahmat N, Jaafar F M, Sahidin I. Medicinal property, phytochemistry and pharmacology of several *Jatropha* species (Euphorbiaceae): a review. Phytochemistry, 2013; 85: 7-29.
15. Cardoso D, Martinati J, Giachetto P, Vidal R O, Carazzolle M F, Padilha L, Guerreiro-Filho O, Maluf M P. Large-scale analysis of differential gene expression in coffee genotypes resistant and susceptible to leaf miner-toward the identification of candidate genes for marker assisted-selection. BMC Genomics, 2014; 15: 66.
16. Costa G, Cardoso K Del Bem, L Lima, A Cunha, et al. Transcriptome analysis of the oil-rich seed of the bioenergy crop *Jatropha curcas* L. BMC Genomics, 2010; 11: 462.
17. King A J, Li Y, Graham I A. Profiling the developing *Jatropha curcas* L Seed transcriptome by pyrosequencing. Bioenerg Res, 2011; 4: 211-221.
18. King A J, Montes L R, Clarke J G, Affleck J, Li Y Witsenboer, H van der Vossen, E van der Linde, P Tripathi, Y Tavares, E Shukla, P Rajasekaran, T van Loo, EN Graham I A. Linkage mapping in the oilseed crop *Jatropha curcas* L reveals a locus controlling the biosynthesis of phorbol esters which cause seed toxicity Plant. Biotechnol J, 2013; 11: 986-996.
19. Jiang H, Wu P, Zhang S, Song C, Chen Y, Li M, Jia Y, Fang X, Chen F, Wu G. Global analysis of gene expression profiles in developing physic nut (*Jatropha curcas* L) seeds. PLoS ONE, 2012; 7: e36522.
20. Silva LJ, Dias DCFS; Milagres CC, Dias LAS. Relationship between fruit maturation stage and physiological quality of physic nut (*Jatropha curcas* L.) seeds. Revista Ciência e Agrotecnologia, 2012; 36: 39-44.
21. Conesa A, Götz S, García-Gómez J M, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization, and analysis in functional genomics research. Bioinformatics, 2005; 21: 3674-3676.
22. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R. InterProScan: protein domains identifier. Nucleic Acids Res, 2005; 33: W116-W120.
23. Jin J, Zhang H, Kong L, Gao G, Luo J. PlantTFDB 30: a portal for the functional and evolutionary study of plant transcription factors. Nucleic Acids Research, 2014; 42: 1182-1187
24. Eddy S R. Accelerated Profile HMM Searches. PLoS Comput Biol, 2011; 7: e1002195.
25. Altshuler D M, Gibbs R A, Peltonen L, Altshuler D M, Gibbs R A, et al. Integrating common and rare genetic variation in diverse human populations. Nature, 2010, 467: 52-58.

26. Saier M H, Yen M R, Noto K, Tamang D G, Elkan C. The Transporter Classification Database: recent advances. Nucleic Acids Res, 2009; 37: 274-278.
27. Sanseverino W, Hermoso A, D'Alessandro R, Vlasova A, Andolfo G, Frusciante L, Lowy E, Roma G, Ercolano M R. PRGdb 20: towards a community-based database model for the analysis of R-genes in plants. Nucleic Acids Res, 2013; 41: 1167-1171
28. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res, 1999; 27: 29-34.
29. R Core Team. R: A language and environment for statistical computing R Foundation for Statistical Computing, Vienna, Austria ISBN 3-900051-07-0, 2012, URL http://wwwR-projectorg/.
30. Smyth G. Limma: linear models for microarray data In: Bioinformatics and Computational Biology Solutions Using R and Bioconductor Statistics for Biology and Health (Gentleman R, Carey VJ, Huber W, Irizarry RA, Dudoit S eds), Springer, 2005; New York, NY, PP 397-420.
31. Fraley C, Raftery A E. Model-Based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association, 2002; 97: 611-631.
32. Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. Bioinformatics, 2007; 23: 257-258.
33. Schaefer J, Opgen-Rhein R, Strimmer K. GeneNet. Modeling and Inferring Gene Networks R package version 12132015, https://CRANR-projectorg/package=GeneNet.
34. Opgen-Rhein R, Strimmer K. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. BMC Syst Biol, 2007; 1: 37.
35. Rozen S, Skaletsky H. Primer3 on the WWW for General Users and for Biologist Programmers In: Bioinformatics Methods and Protocols, (Misener, S, Krawetz, S, eds), Humana Press, 1999; pp 365-386
36. Altschul S F, Madden T L, Schäffer A A, Zhang J, Miller W, Lipman D J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research, 1997; 25: 3389-3402.
37. Hirakawa H, Tsuchimoto S, Sakai H, Nakayama S, Fujishiro T, Kishida Y, Kohara M, Watanabe A, Yamada M, Aizu T, Toyoda A, Fujiyama A, Tabata S, Fukui K, Sato S. Upgraded genomic information of *Jatropha curcas* L.Plant Biotechnol, 2012; 29: 123-130.
38. Sato S, Hirakawa H, Isobe S, Fukai E, Watanabe A, et al. Sequence analysis of the genome of an oil-bearing tree, *Jatropha curcas* L. DNA Res, 2011; 18: 65-76.
39. Zhang, L, Zhang, C, Wu, P, Chen, Y, Li, M, Jiang, H, Wu, G. Global analysis of gene expression profiles in physic nut (*Jatropha curcas* L) seedlings exposed to salt stress. PloS one, 2014a; 9: e97878.
40. Wu P, Zhou C, Cheng S, Wu Z, Lu W, Han J, et al. Integrated genome sequence and linkage map of physic nut (*Jatropha curcas* L), a biodiesel plant. The Plant Journal 2015, 81: 810-821.

41. Afzal A J, Wood A J, Lightfoot D A. Plant receptor-like serine threonine kinases. roles in signalling and plant defense. Molecular plant-microbe interactions. MPMI, 2008; 21: 507-517
42. Sekhwal MK, Li P, Lam I, Wang X, Cloutier S, You FM. Disease Resistance Gene Analogs (RGAs) in Plants. International journal of molecular sciences, 2015; 16: 19248-19290.
43. Basnet R, Moreno-Pachon N, Lin K, Bucher J, Visser R G F, Maliepaard C, Bonnema G. Genome-wide analysis of coordinated transcript abundance during seed development in different *Brassica rapa* morphotypes. BMC Genomics, 2013; 14: 840.
44. Guo WJ, Ho TH. An abscisic acid-induced protein, HVA22, inhibits gibberellin-mediated programmed cell death in cereal aleurone cells. Plant physiology, 2008; 147: 1710-1722.
45. Baud S, Boutin J-P, Miquel M, Lepiniec L, Rocha C. Integrated overview of seed development in *Arabidopsis thaliana* ecotype. Plant Physiology and Biochemistry, 2002; 40: 151-160.
46. Jolivet P, Boulard C, Bellamy A, Valot B, d'Andréa S, Zivy M, Nesi N, Chardot T. Oil body proteins sequentially accumulate throughout seed development in *Brasica napus.* Journal of Plant Physiology, 2011; 168: 2015-2020.
47. Righetti K Vu, J Pelletier, S et al. Inference of longevity-related genes from a robust co-expression network of seed maturation identifies regulators linking seed storability to biotic defense-related pathways. The Plant Cell, 2015; 27: 2692-2708.
48. Leubner-Metzger G. Functions and regulation of b-1,3-glucanases during seed germination, dormancy release and after-ripening. Seed Science Research, 2003; 13: 17-34.
49. Firon N, LaBonte D, Villordon A, Kfir Y, Solis J, Lapis E, Perlman T S, Doron-Faigenboim A, Hetzroni A, Althan L, A Nadir L A. Transcriptional profiling of sweetpotato (*Ipomoea batatas*) roots indicates down-regulation of lignin biosynthesis and up-regulation of starch biosynthesis at an early stage of storage root formation. BMC Genomics, 2013; 14: 1471-2164.
50. Egli B, Kölling K, Köhler C, Zeeman SC, Streb S. Loss of cytosolic phosphoglucomutase compromises gametophyte development in Arabidopsis. Plant Physiol, 2010; 154: 1659-1671.
51. Fait A, Angelovici R, Less H, Ohad I, Urbanczyk-Wochniak E, Fernie A R, Galili G. Arabidopsis seed development and germination is associated with temporally distinct metabolic switches. Plant Physiology, 2006; 142: 839-854.
52. Liu H, Wang C, Komatsu S, He M, Liu G, Shen S. Proteomic analysis of the seed development in *Jatropha curcas*: From carbon flux to the lipid accumulation. Journal of Proteomics, 2013; 91: 23-40.
53. Borisjuk L, Nguyen T H, Neuberger T, Rutten T, Tschiersch H, et al. Gradients of lipid storage, photosynthesis and plastid differentiation in developing soybean seeds. The New Phytologist, 2005; 167: 761-776.
54. Ong W D, Voo L Y, Kumar V S. De novo assembly, characterization and functional annotation of pineapple fruit transcriptome through massively parallel sequencing. PloS one, 2012; 7: e46937.
55. Barros J, Serk H, Granlund I, Pesquet E. The cell biology of lignification in higher plants. Annals of Botany, 2015; 1: 1053-74.

56. Ma J, Kanakala S, He Y, Zhang J, Zhong X. Transcriptome tequence analysis of an ornamental plant, *Ananas comosus* var bracteatus, revealed the potential unigenes involved in terpenoid and phenylpropanoid biosynthesis. PLOS ONE, 2015; 10: e0119153.
57. Whetten R, Sederoff R. Lignin Biosynthesis. Plant Cell, 1995; 7: 1001-1013.
58. Gijzen M, Miller S S, Bowman L-A, Batchelor A K, Boutilier K, Miki B L A. Localization of peroxidase mRNAs in soybean seeds by in situ hybridization. Plant Mol Biol, 1999; 41: 57-63.
59. Lunkenbein S, Salentijn E M, Coiner H A, Boone M J, Krens F A, Schwab W. Up- and down-regulation of *Fragaria x ananassa* O-methyltransferase: impacts on furanone and phenylpropanoid metabolism. J Exp Bot, 2006; 57: 2445-2453.
60. Lulin H, Xiao Y, Pei S, Wen T, Shangqin H. The first Illumina-based de novo transcriptome sequencing and analysis of safflower flowers. PloS one, 2012; 7: e38653.
61. Li H, Dong Y, Yang J, Liu X, Wang Y, Yao N, Guan L, Wang N, Wu J, Li X. De novo transcriptome of safflower and the identification of putative genes for oleosin and the biosynthesis of flavonoids. PloS one, 2012; 7: e30987.
62. Qu C, Zhao H, Fu F, Wang Z, K Zhang, Zhou Y, Wang X, Wang R, Xu X, Tang Z, Lu K, Zhang K, Zhou Y, Wang X, Wang R Xu X, Tang Z, Lu K. Genome-Wide Survey of Flavonoid Biosynthesis Genes and Gene Expression Analysis between Black- and Yellow-Seeded *Brasica napus.* Frontiers in plant science, 2016; 7: 1755.
63. Gupta A, Singh M. Interaction between glucose and brassinosteroid during the regulation of lateral root development in Arabidopsis. Plant Physiol, 2015; 168: 307-320-
64. Jiang W-B, Lin W-H. Brassinosteroid functions in Arabidopsis seed development. Plant Signaling & Behavior, 2013; 8: e25928.
65. Takahashi N, Nakazawa M, Shibata K, Yokota T, Ishikawa A, Suzuki K, Kawashima M, Ichikawa T, Shimada H, Matsui M. shk1-D a dwarf Arabidopsis mutant caused by activation of the CYP72C1 gene, has altered brassinosteroid levels.The Plant Journal, 2005; 42: 13-22.
66. Tanabe S, Ashikari M, Fujioka S, Takatsuto S, Yoshida S, Yano M, Yoshimura A, Kitano H, Matsuoka M, Fujisawa Y, Kato H, Iwasaki Y A. novel cytochrome P450 is implicated in brassinosteroid biosynthesis via the characterization of a rice dwarf mutant, dwarf11, with reduced seed length. Plant Cell, 2005; 17: 776-790.
67. Tian Y, Zhang M, Hu X, Wang L, Dai J, Xu Y, Chen F. Over-expression of CYP78A98, a cytochrome P450 gene from *Jatropha curcas* L, increases seed size of transgenic tobacco. Electronic Journal of Biotechnology, 2016; 19: 15-22.
68. Wu C Y, Trieu A, Radhakrishnan P, Kwok S F, Harris S, Zhang K, Wang J, Wan J, Zhai H, Takatsuto S, Matsumoto S, Fujioka S, Feldmann K A, Pennell R I. Brassinosteroids regulate grain filling in rice. Plant Cell 2008; 20: 2130-2145.
69. Fukuta N, Fukuzono K, Kawaide H, Abe H, Nakayama M. Physical Restriction of Pods Causes Seed Size Reduction of a Brassinosteroid-deficient Faba Bean (*Vicia faba*). Annals of Botany, 2006; 97: 65-

69.

70. Yan, A, Chen, Z. The pivotal role of abscisic acid signaling during transition from seed maturation to germination. Plant cell reports, 2017; 36: 689-703.
71. Bogatek R, Gniazdowska A. Ethylene in seed development, dormancy and germination Annual plant reviews. The plant hormone ethylene. 2012; 44: 189-218.
72. Walton, L J, Kurepin, L V, Yeung, E C, Shah, S, Emery, R J N, Reid, D M, Pharis R P. Ethylene involvement in silique and seed development of canola, *Brasica napus* L. Plant physiology and biochemistry, 2012; 58: 142-150.
73. Yu B, Gruber M, Khachatourians G G, Hegedus D D, Hannoufa A. Gene expression profiling of developing *Brasica napus* seed about changes in major storage compounds. Plant science, 2010; 178: 381-389.
74. Ketring D L, Morgan P W. Physiology of Oil Seeds: IV Role of Endogenous Ethylene and Inhibitory Regulators during Natural and Induced Afterripening of Dormant Virginia-type Peanut Seeds. Plant Physiology, 1972; 50: 382-387.
75. Agarwal A K, Qi Y, Bhat D G, Woerner B M, Brown S M. Gene isolation and characterization of two acyl CoA oxidases from soybean with broad substrate specificities and enhanced expression in the growing seedling axis. Plant Mol Biol, 2001; 47: 519-531.
76. Schilmiller A L, Koo A J, Howe G A. Functional diversification of acyl-coenzyme A oxidases in jasmonic acid biosynthesis and action. Plant Physiology, 2007; 143: 812-824.
77. Peng F, Weselake R. Gene coexpression clusters and putative regulatory elements underlying seed storage reserve accumulation in Arabidopsis. BMC Genomics, 2011; 12: 286.
78. Hwang YT, Pelitire SM, Henderson MP, Andrews DW, Dyer JM, Mullen RT. Novel targeting signals mediate the sorting of different isoforms of the tail-anchored membrane protein cytochrome b5 to either endoplasmic reticulum or mitochondria. Plant Cell, 2004; 16: 3002-3019.
79. Smith MA, Jonsson L, Stymne S, Stobart K. Evidence for cytochrome b5 as an electron donor in ricinoleic acid biosynthesis in microsomal preparations from developing castor bean (*Ricinus communis* L). Biochem J, 1992; 287: 141-144.
80. Bafor M, Smith M A, Jonsson L, Stobart K, Stymne S. Biosynthesis of vernoleate (cis-12-epoxyoctadeca-cis-9-enoate) in microsomal preparations from developing endosperm of *Euphorbia lagascae* Arch. Biochem Biophys, 1993; 303: 145-151.
81. Napier J A, Michaelson L V, Sayanova O. The role of cytochrome b5 fusion desaturases in the synthesis of polyunsaturated fatty acids. Prostaglandins Leukot Essent Fatty Acids, 2003; 68: 135-143.
82. Chen M, Xuan L, Wang Z, Zhou L, Li Z, Du X, Ali E, Zhang G, Jiang L. TRANSPARENT TESTA8 Inhibits Seed Fatty Acid Accumulation by Targeting Several Seed Development Regulators in Arabidopsis. Plant Physiology, 2014; 165: 905-916.
83. Akhov L, Ashe, P, Tan, Y, Datla, R, Selvaraj G. Proanthocyanidin biosynthesis in the seed coat of yellow-seeded, canola quality *Brasica napus* YN01-429 is constrained at the committed step

catalyzed by dihydroflavonol 4-reductase. Botany, 2009; 87: 616-625.

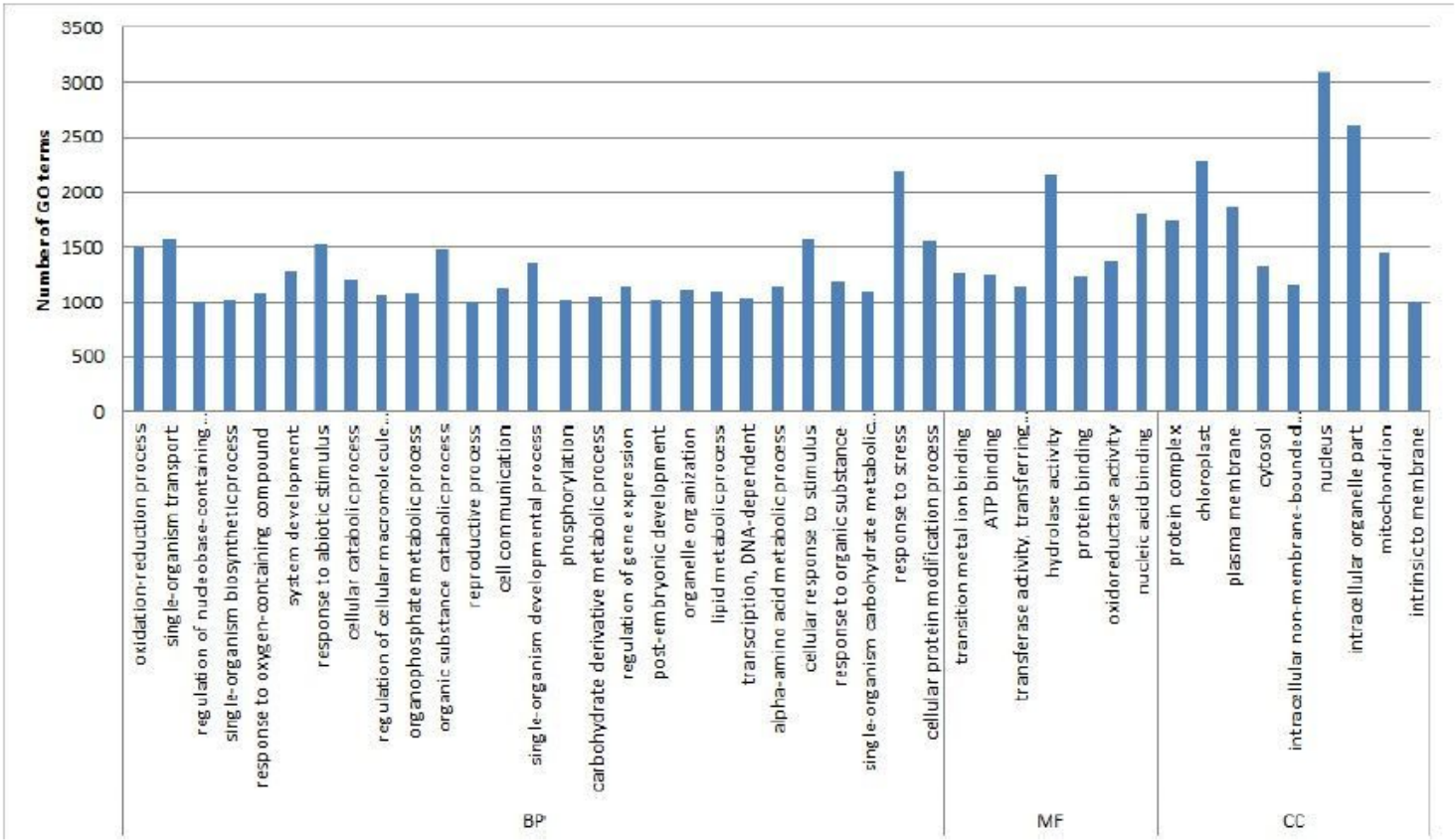
84. Sharma N, Anderson M, Kumar A, Zhang Y, Giblin EM, Abrams SR, Zaharia LI, Taylor DC, Fobert PR. Transgenic increases in seed oil content are associated with the differential expression of novel Brassica-specific transcripts. BMC genomics, 2008; 9: 619.
85. Kotak S, Vierling E, Bäumlein H, Koskull-Döring Pv. A Novel Transcriptional Cascade Regulating Expression of Heat Stress Proteins during Seed Development of Arabidopsis. The plant Cell, 2007; 19: 182-195.
86. DeRocher A, Vierling E. Cytoplasmic HSP70 homologues of pea: differential expression in vegetative and embryonic organs. Plant Mol Biol, 1995; 27: 441-56.
87. Sung, DY, Vierling, E, Guy, CL. Comprehensive expression profile analysis of the Arabidopsis Hsp70 gene family Plant Physiol 2001, 126: 789-800.
88. Wakasa Y, Yasuda H, Oono Y, Kawakatsu T, Hirose S, Takahashi H, Hayashi S Yang L, Takaiwa F. Expression of ER quality control-related genes in response to changes in BiP1 levels in developing rice endosperm. Plant J, 2011; 65: 675-89.
89. Sarkar NK, Kundnani P, Grover A. Functional analysis of Hsp70 superfamily proteins of rice (*Oryza sativa*). Cell Stress Chaperones, 2013; 18: 427-437.
90. Ko D, Kang J, Kiba T, Park J, Kojima M, Do J, Kim KY, Kwon M, Endler A, Song W-Y, et al. Arabidopsis ABCG14 is essential for the root-to-shoot translocation of cytokinin. Proc Natl Acad Sci USA, 2014; 111: 7150-7155
91. Zhang H, Zhu H, Pan Y, Yu Y, Luan S, Li L. A DTX/MATE-type transporter facilitates abscisic acid efflux and modulates ABA sensitivity and drought tolerance in Arabidopsis. Mol Plant, 2014b; 7: 1522-1532.
92. Kuromori T, Miyaji T, Yabuuchi H, Shimizu H, Sugimoto E, Kamiya A, Moriyama Y, Shinozaki K. ABC transporter AtABCG25 is involved in abscisic acid transport and responses. Proc Natl Acad Sci USA, 2010; 107: 2361-2366.
93. Hwang JU, Song WY, Hong D, Ko D, Yamaoka Y, Jang S, Yim S, Lee E, Khare D, Kim K et al. Plant ABC transporters enable many unique aspects of a terrestrial plant's lifestyle. Molecular plant, 2016; 9: 338-355.
94. Kang J, Park J, Choi H, Burla B, Kretzschmar T, Lee Y, Martinoia E. Plant ABC Transporters Arabidopsis Book, 2011; 9: e0153-e0153.
95. Gutierrez L, Van Wuytswinkel O, Castelain M, Bellini C. Combined networks regulating seed maturation. Trends Plant Sci, 2007; 12: 294-300-
96. Gao H, Wang Y, Li W, Gu Y, Lai Y, Bi Y, He C. Transcriptomic comparison reveals genetic variation potentially underlying seed developmental evolution of soybeans. Journal of Experimental Botany, 2018; 69: 5089-5104.
97. Gaur V, Singh U, Kumar A. Transcriptional profiling and in silico analysis of Dof transcription factor gene family for understanding their regulation during seed development of rice *Oryza sativa* L. Mol Biol Rep, 2011; 38: 2827- 2848.

98. Mena M, Vicente-Carbajosa J, Schmidt Robert J, Carbonero P. An endosperm-specific DOF protein from barley, highly conserved in wheat, binds to and activates transcription from the prolamin-box of a native B-hordein promoter in barley endosperm. The Plant, 1998; 16: 53-62.
99. Verdier J, Lalanne D, Pelletier S, Torres-Jerez I, Righetti K, Bandyopadhyay K, et al. A regulatory network-based approach dissects late maturation processes related to the acquisition of desiccation tolerance and longevity of *Medicago truncatula* seeds. Plant Physiol, 2013; 163: 757-74.
100. Pereira Lima JJ, Buitink J, Lalanne D, Rossi RF, Pelletier S, et al. Molecular characterization of the acquisition of longevity during seed maturation in soybean. PLOS ONE, 2017; 12: e0180282.
101. Huang Y, Feng C-Z, Ye Q, Wu W-H, Chen Y-F. Arabidopsis WRKY6 transcription factor acts as a positive regulator of abscisic acid signaling during seed germination and early seedling development. PLoS Genet, 2016; 2: e1005833.
102. Wang Z, Shu Y, Wang L, et al. A WRKY transcription factor participates in dehydration tolerance in *Boea hydrometrica* by binding to the W-box elements of the galactinol synthase (BhGolS1) promoter. Planta, 2009; 230: 1155-1166.
103. Ma L, Li G. FAR1-RELATED SEQUENCE (FRS) and FRS-RELATED FACTOR (FRF) family proteins in Arabidopsis growth and development. Front Plant Sci, 2018; 9: 692.
104. Gu K, Yi C, Tian D, Sangha J, Hong Y, Yin Z. Expression of fatty acid and lipid biosynthetic genes in developing endosperm of *Jatropha curcas.* Biotechnology for Biofuels, 2012; 5: 47.
105. Chen M-S, Wang G-J, Wang R-L, et al. Analysis of expressed sequence tags from biodiesel plant *Jatropha curcas* embryos at different developmental stages. Plant Sci, 2011; 181: 696-700.
106. Xu R, Wang R, Liu A. Expression profiles of genes involved in fatty acid and triacylglycerol synthesis in developing seeds of Jatropha (*Jatropha curcas* L). Biomass Bioenergy 2011, 35; 1683-1692.
107. Chandran D, Sankararamasubramanian HM, Kumar MA, Parida, A. Differential expression analysis of transcripts related to oil metabolism in maturing seeds of *Jatropha curcas* L. Physiology and molecular biology of plants: an international journal of functional plant biology, 2014; 20: 181-190.
108. Kleindt C K, Stracke R, Mehrtens F, Weisshaar B. Expression analysis of flavonoid biosynthesis genes during *Arabidopsis thaliana* silique and seed development with a primary focus on the proanthocyanidin biosynthetic pathway. BMC Research Notes, 2010; 3: 255.

# Tables

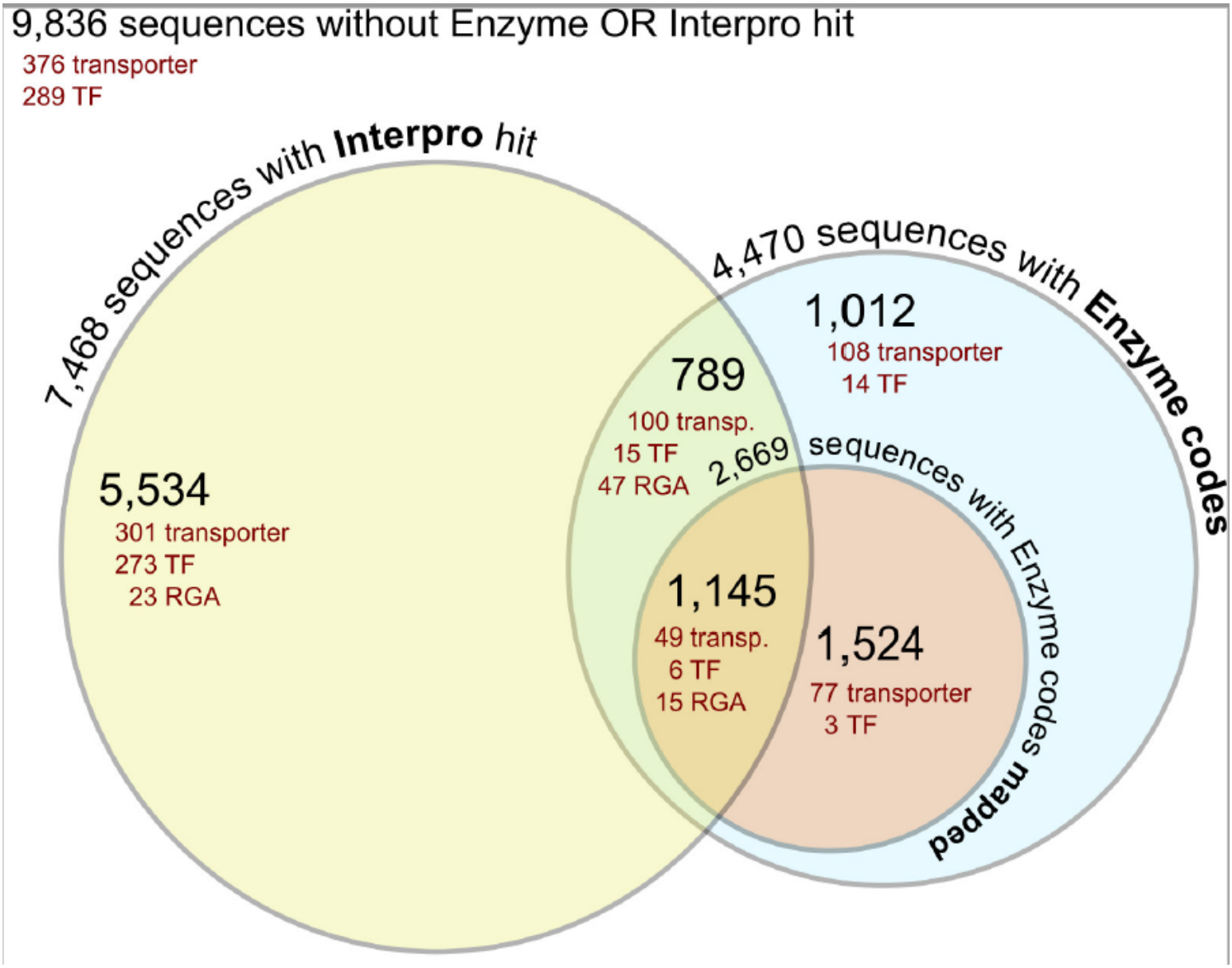
Due to technical limitations, the tables have been placed in the Supplementary Files section.

# Figures



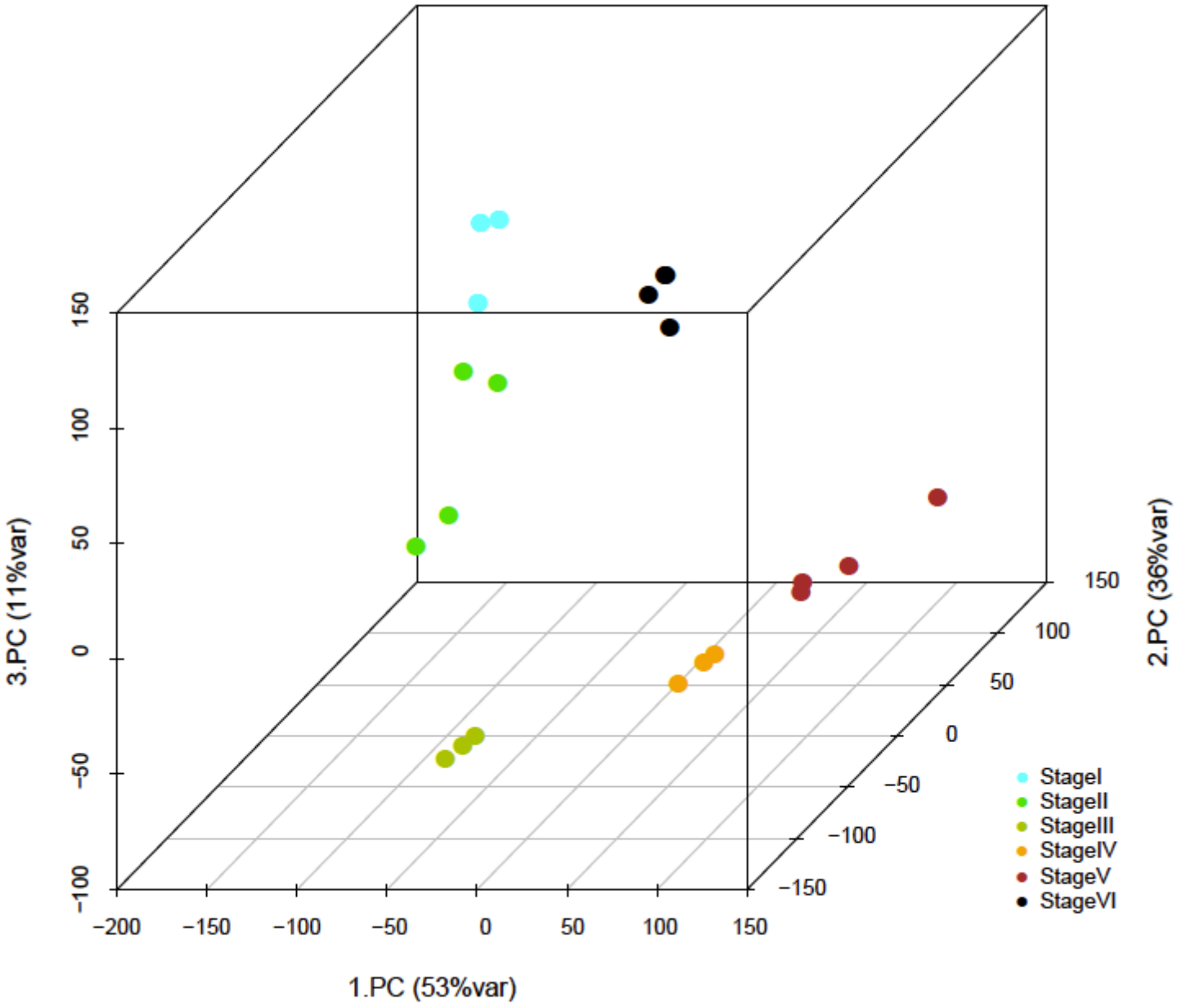

## Figure 1

GO annotation classification of whole seed transcript sequencing data Results are summarized for three main GO categories (BP, MF, CC) The x-axis indicates the names and the number of each GO term



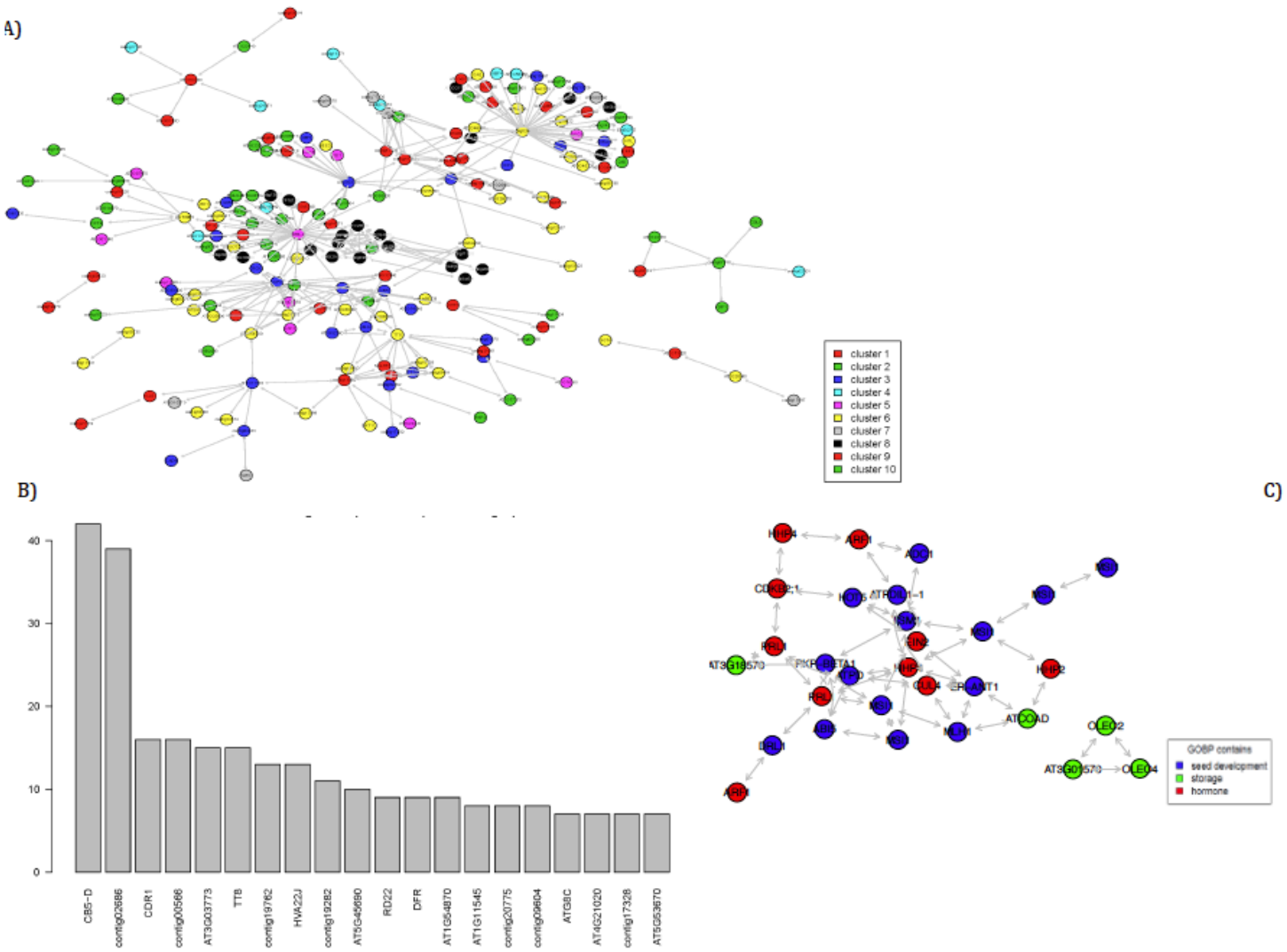

## Figure 2

Venn diagram summarizing the functional annotation process of whole seed transcriptome sequence data Enzyme codes originate from the Blast2Go annotation, while Interpro hits are resulting from InterProScan (With Interpro hit) Enzymes codes have been checked for mapping on the pathways of the KEGG (With Enzyme codes mapped) Results of the manual annotation for transporters, TFs and RGAs in all classes are also indicated



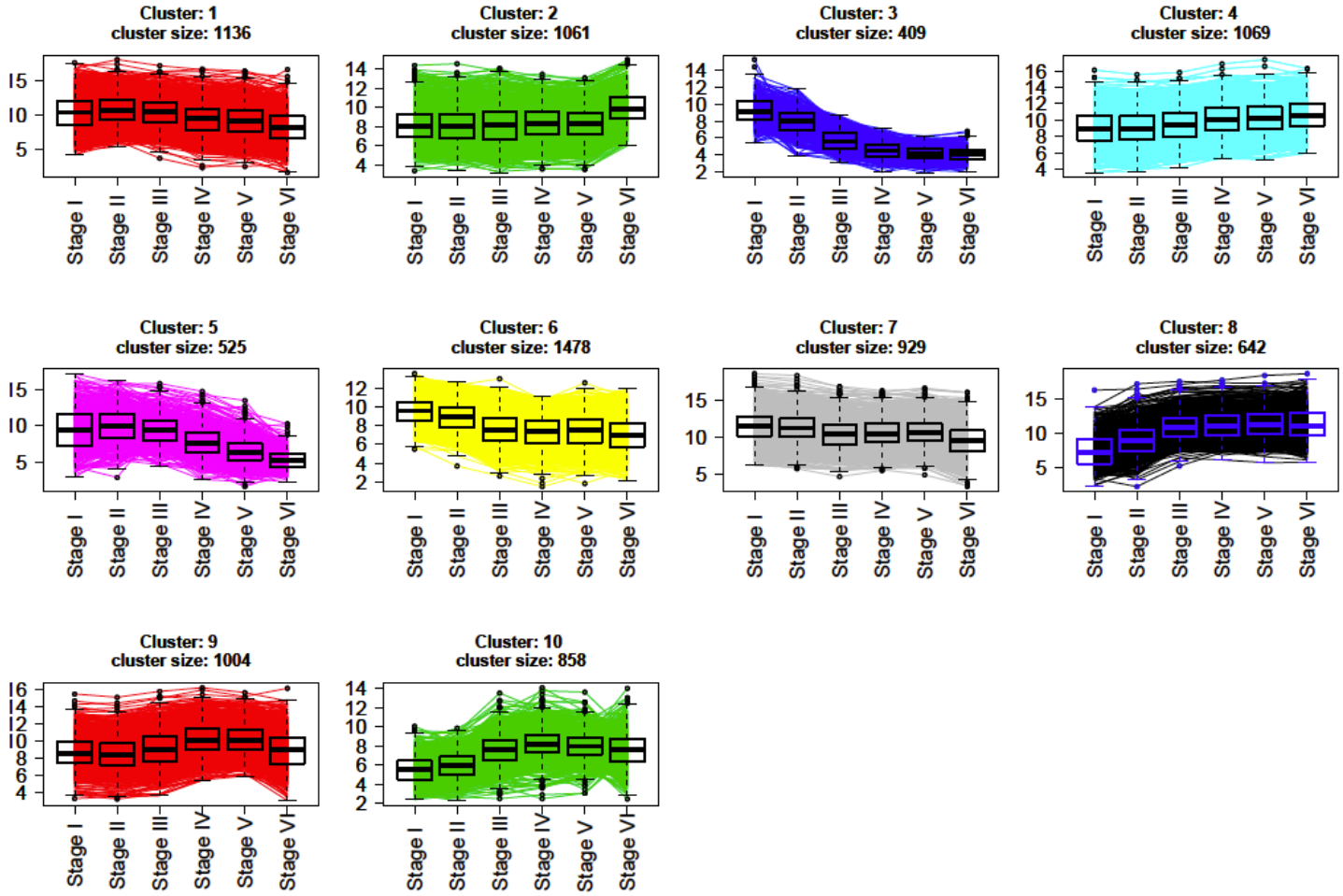

## Figure 3

PCA analysis explains the variance in gene expression of the six different seed maturation stages (I-VI) with their biological replications



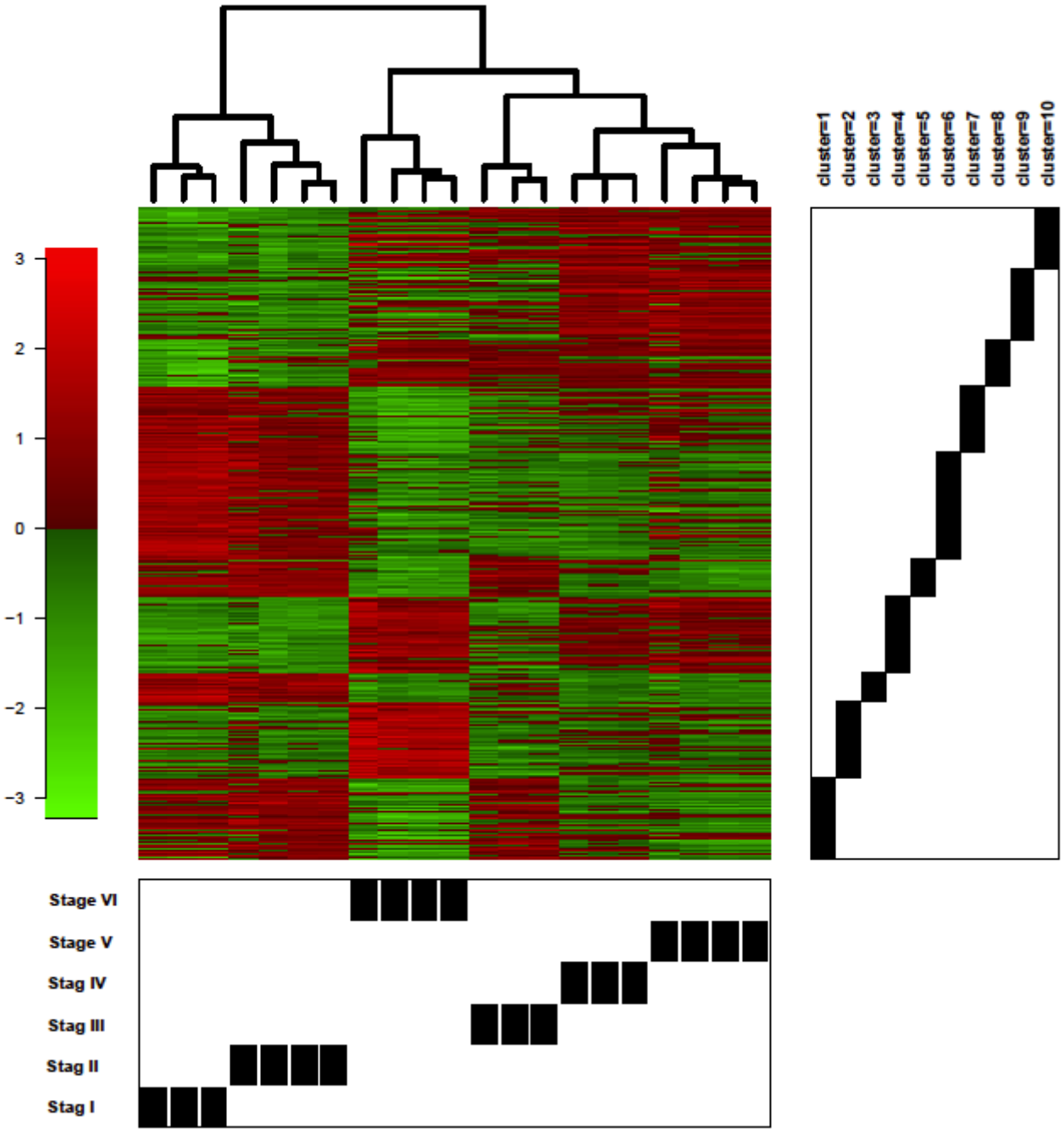

## Figure 4

Co-expression networks based on partial correlations A) A network of the 300 most significant edges between genes differentially expressed during seed development stages Node colouring represents cluster membership (see also Fig S3) B) Two major hubs could be identified with ~40 edges and a broad range of nodes display between 10 and 15 edges C) Co-expression networks based on GOs of biological processes related to seed storage, seed development, and hormones cross-talking



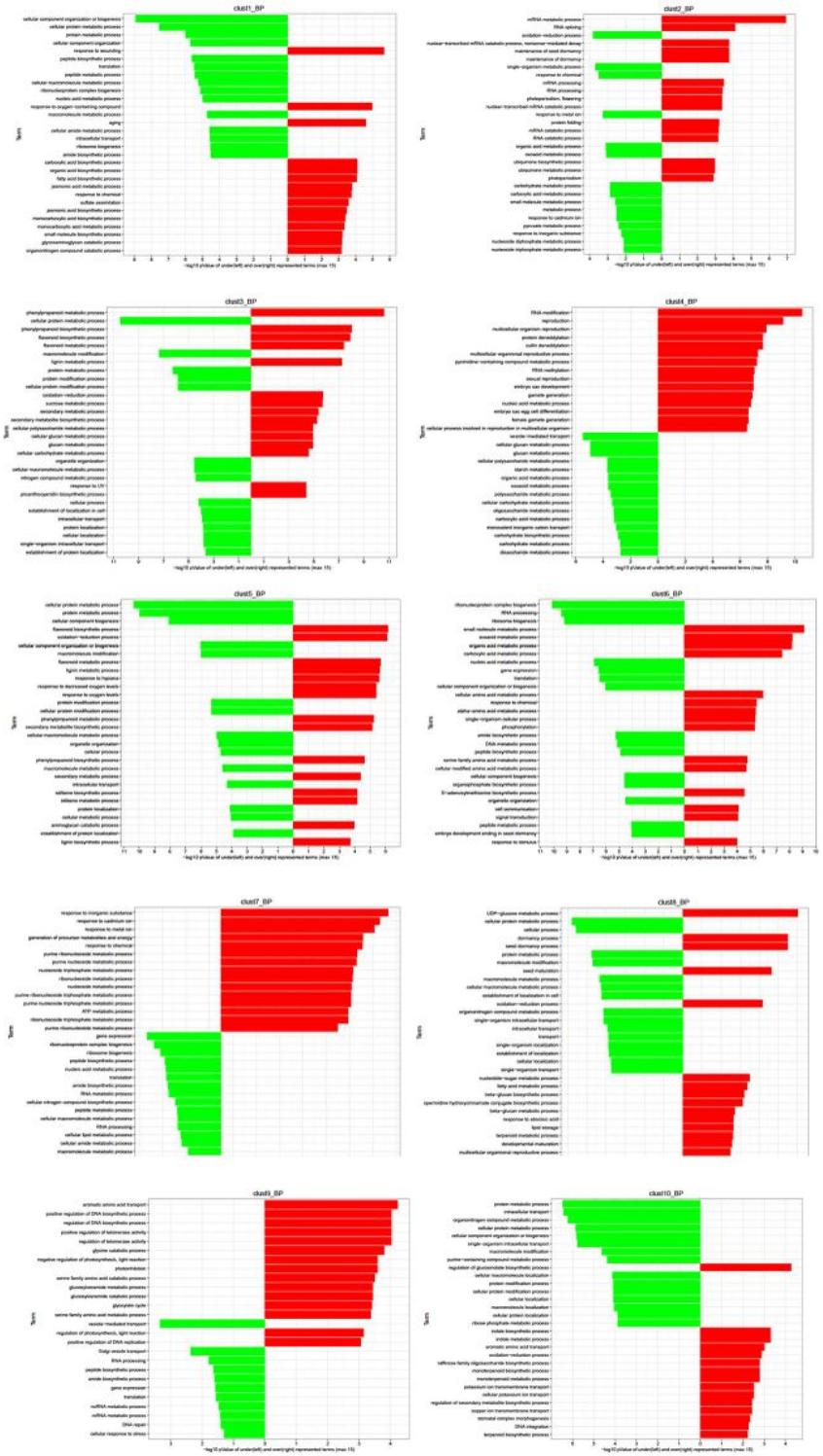

## Figure 5

Clustering analysis of DESs according to their expression profiles between seed maturation stages




## Figure 6

Global gene expression heatmap and cluster analysis of the DESs during seed development Cluster analysis on Y-axis represents similar expression patterns among the expressed sequences (ESs), while cluster analysis on the x-axis indicates the relatedness of DESs profiles among the developmental stages and samples



## Figure 7

Histogram representation the top 15 of the significantly enriched category of GO annotation for all DESs BP categories (y-axis) were grouped into two colors The green histogram, representing the under-represented GOs and the red histogram, representing over-represented GOs The x-axis indicates the statistical significance of the enrichment

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplement1.xlsx
- supplement2.xlsx
- supplement3.pdf
- supplement4.pdf
- supplement5.pdf
- supplement6.pdf
- supplement7.xlsx
- supplement8.xlsx
- supplement9.xlsx
- supplement10.xlsx
- supplement11.pdf
- supplement11.pdf
- supplement13.png
- supplement13.png
- supplement15.pdf
- supplement16.xlsx
- supplement17.pdf
- supplement18.pdf