scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Specialized metabolome and transcriptome atlas of developing *Arabidopsis thaliana* seed under warm temperatures

Léa Barreda, Stéphanie Boutet, Delphine De Vos, Céline Boulard, Damaris Grain, Loïc Lepiniec & Massimiliano Corso ✉

Seed development, which depends on parent plants genetic background and mother plant environmental conditions, is a major component determining seed composition. Seed quality is a main agricultural concern, impacting both food and non-food applications, while also playing a central role in biodiversity conservation and environment protection. Climate change, in particular the emergence of extremely high temperatures, constitute a critical global threat to agriculture. Specialized metabolites (SMs) play crucial roles in the interactions of plants and seeds with their environments. Several SMs are known to be protective compounds involved in seed stress responses, thus impacting their quality. In this study, we performed untargeted metabolomic (LC-MS/MS) and transcriptomic (RNA-Seq) analyses of *Arabidopsis thaliana* seeds harvested at six developmental stages (Globular, Transition, Torpedo, Bent cotyledon, Mature green and Dry seed), and developed under control and warm temperature conditions. Those data provide an original and valuable resource that could be used to identify SMs and genes involved in seed heat stress responses and for the study of their regulation and functions during seed development.

Background & Summary

Seeds constitute unique biological entities that have the ability to disperse new generations of plants. Seeds have undoubtedly participated to the success of spermaphytes on the earth allowing them to deal with unfavourable environmental conditions by suspending their life cycle and pursuing their growth when favourable conditions are reunited[1,2]. Moreover, seeds are crucial for humankind being the main source of human nutrition and animal feed. In particular, they constitute a major source of proteins, oils, starch, fibers, vitamins and minerals, and it is estimated that around 70% of calories consumed by humans derive from seeds[2,3]. Furthermore, starch and oils from seeds constitute valuable resources to produce alternatives to fossil energies (e.g. biofuels) or product derived from petrochemical industry (e.g. bioplastics)[1]. In addition to storage compounds, seeds also accumulate specialized metabolites, also known as secondary metabolites[4], (SMs) which are small organic compounds that play crucial roles in plants and seeds interactions with the environment[5–7]. Moreover, they constitute high added-value molecules with a broad range of applications in food, health, medicine, cosmetics and agroecology industry[1,5,7,8]. Finally, by storing plant genetic information, seeds are the major vector for improving agronomic practices and managing genetic resources, which are essential components for sustainable agricultural development and biodiversity and environment preservation[1]. Hence, knowing the factors impacting seed quality is a major agricultural concern, impacting both food and non-food applications, as well as playing a pivotal role in biodiversity conservation and environmental preservation. Seed "quality" include both their composition (SMs, lipids, proteins, etc) and other intrinsic and extrinsic factors influencing the seed capacity to germinate and to develop into a healthy robust growing plant when placed under ideal conditions (e.g. germination rate, vigour, genetic purity, etc)[9].

Seed quality is determined by the genetic and environmental controls of seed development, from the embryogenesis after the double fertilization of the ovule to the dehydration of the mature seed[3,10]. Environmental

Université Paris-Saclay, INRAE, AgroParisTech, Institute Jean-Pierre Bourgin for Plant Sciences (IJPB), 78026, Versailles, France. ✉e-mail: massimiliano.corso@inrae.fr

stresses such as drought, salinity, and cold or heat stresses have impacts on seed quality and yield[11,12]. In a context of climate change and global population growth, new studies in seed development under constraining environmental conditions are of paramount importance to understand the mechanisms and identify the genes involved in determining seed quality for the development of improved seed crops. SMs play crucial roles in the interactions of seeds with their environments[7,11]. SMs are highly accumulated in plant and seeds of a wide range of Brassicaceae species, including the wild species *Arabidopsis thaliana*[7]. Phenolic compounds (e.g. flavonoids, cinnamic acids, and coumarins), nitrogen-containing compounds (e.g. alkaloids, cyanogenic glycosides, and glucosinolates), and terpenoids constitute the three main classes of SMs[7,8,11]. Seed quality is widely determined by the accumulation of several storage compounds and SMs[8]. Moreover, seed SM accumulation is highly affected by environmental conditions[13]. However, while most studies have focused on SMs pathways regulation by stresses in vegetative tissues, the study of seed ability to produce these protective compounds in response to environmental stress has been neglected. This is also true for the effect of heat stress on specialized metabolome during Arabidopsis seed development, for which few information is available.

In this work we carried out untargeted metabolomics (LC-MS/MS) and transcriptomic analyses (RNA-Seq) on *A. thaliana* Col-0 seeds at different developmental stages harvested from plants grown under control and warm temperatures. First, to produce those datasets we had to synchronise both control and warm stress seed developments. Indeed, under warm temperature the seed development is faster than under control conditions (Fig. 1). The experimental procedure followed for seed developmental stage synchronisation is summarised in Fig. 1. Second, *A. thaliana* plants were grown in control and warm temperature conditions since the onset of flowering. Seeds were harvested at 6 developmental stages (Globular, Transition, Torpedo, Bent cotyledon, Mature green and Dry seed), and subjected to multiomic analyses to generate specialized metabolome and transcriptome datasets. Subsequently, metabolomic and transcriptomic data were integrated through a correlation network analysis. The experimental procedures and different steps followed for data production are summarised in Fig. 2.

Methods

Synchronisation of seed developmental stages. Under warm temperatures, seeds achieve their development faster[14,15]. A first plant culture was performed to find a correspondence between day(s) after flowering (DAF) and seed developmental stage (e.g. Globular, Transition, Heart, Torpedo, Bent cotyledon, Mature green, Post-mature green and Dry seed) under both control and warm temperature conditions (Fig. 1).

Plant growing conditions. *Arabidopsis thaliana* Col-0 (Columbia) seeds (20) were sown in fertilized soil in a controlled condition growth chamber with long day photoperiod (16 h of light – 21 °C/ 8 h of dark – 19 °C). Watering was made with water. Warm stress treatment was applied to concerned plants upon the onset of flowering (occurring between 5 and 6 weeks after seed sowing). Half of the plants (10/20) were therefore transferred to a second growth chamber (SANYO™, Thermo Fisher Scientific) and subjected to warm temperature (16 h of light – 27 °C/ 8 h of dark – 24 °C), whereas the other half, corresponding to control plants, remained in the initial growth chamber with control temperature conditions. The temperatures from the different growth chambers were monitored throughout the experiment with thermometers (KIMO®).

Flower bud tags and seeds harvests during seed-development synchronisation. Flower buds from both control and warm temperature condition plants were tagged (Fig. 1a) and corresponding growing siliques were harvested at 11–12 time-points (from 2 to 10 DAF and 12, 14, 17 and 20 DAF; Fig. 1b). Harvested siliques were collected for microscopy and real-time quantitative polymerase chain reaction (RT-qPCR) analyses. At harvest, siliques for RT-qPCR analysis were immediately frozen in liquid nitrogen and stored at −80 °C until RNA extraction. The other siliques were kept on ice until the microscopic observation that is described as follows.

Microscopic analyses during seed-development synchronisation. Siliques from each time-point and condition were carefully dissected by cutting through the dehiscence zone with a scalpel to harvest the seeds. Chloral hydrate solution was used to bleach seed coats and fix the samples that were placed between slide and slip cover[16]. Images were acquired on Axioplan 2 imaging epifluorescence microscope (Zeiss, Göttingen, Germany).

Gene expression analyses during seed-development synchronisation. RNA extraction. RNAs were extracted from siliques at each time point. Briefly, 25 mg of siliques were grinded in liquid nitrogen with some polyvinylpolypyrrolidone (PVPP; phenolic compound absorbing agent). Subsequently, 250 µL of extraction buffer (0.1 M LiCl, 0.1 M Tris pH8, 10 mM Ethylenediaminetetraacetic acid (EDTA), 1% Sodium Dodecyl Sulfate (SDS), 1.5% β-mercaptoethanol) were added and the mixtures were vortexed, then shaked gently (500 rpm) for 20 min at room temperature using a ThermoMixer™ C (Eppendorf). 250 µL of acidic phenol (pH 4, Sigma-Aldrich) were added, followed by a rigorous vortex, prior to the addition of 200 µL of chloroform (Sigma-Aldrich), vigorous vortex and centrifugation (15 min, 13 000 g). Supernatants were transferred to new Eppendorf tubes containing 500 µL of acidic phenol/chloroform (5:1) before vortexing and centrifuging (5 min, 13 000 g). Supernatants (around 400 µL) were transferred to new Eppendorf tubes and the equivalent volume of 8 M LiCl was added prior to mixing by inverting the tubes several times. Samples were precipitated at −80 °C for 1 h before centrifugation (30 min, 13 000 g, 4 °C). Subsequently, 1 mL of ethanol (70%) was added to each sample, followed by centrifugation (5 min, 13 000 g, 4 °C). Supernatants were discarded and RNAs dried prior to their resuspension in 40 µL of RNAse-free water. RNAs were purified by using Micro Bio Spin Chromatography columns (BioRad) previously pre-conditioned by adding 5 mg of PVPP and two consecutive washings with 400 µL of RNAse-free water (2 min centrifugations at 1 000 g). Subsequently, the 40 µL of RNA were placed into the pre-conditioned column, incubated on ice for 5 min and centrifuged three consecutive times to elute RNAs (2 min, 1 000 g). DNase treatment

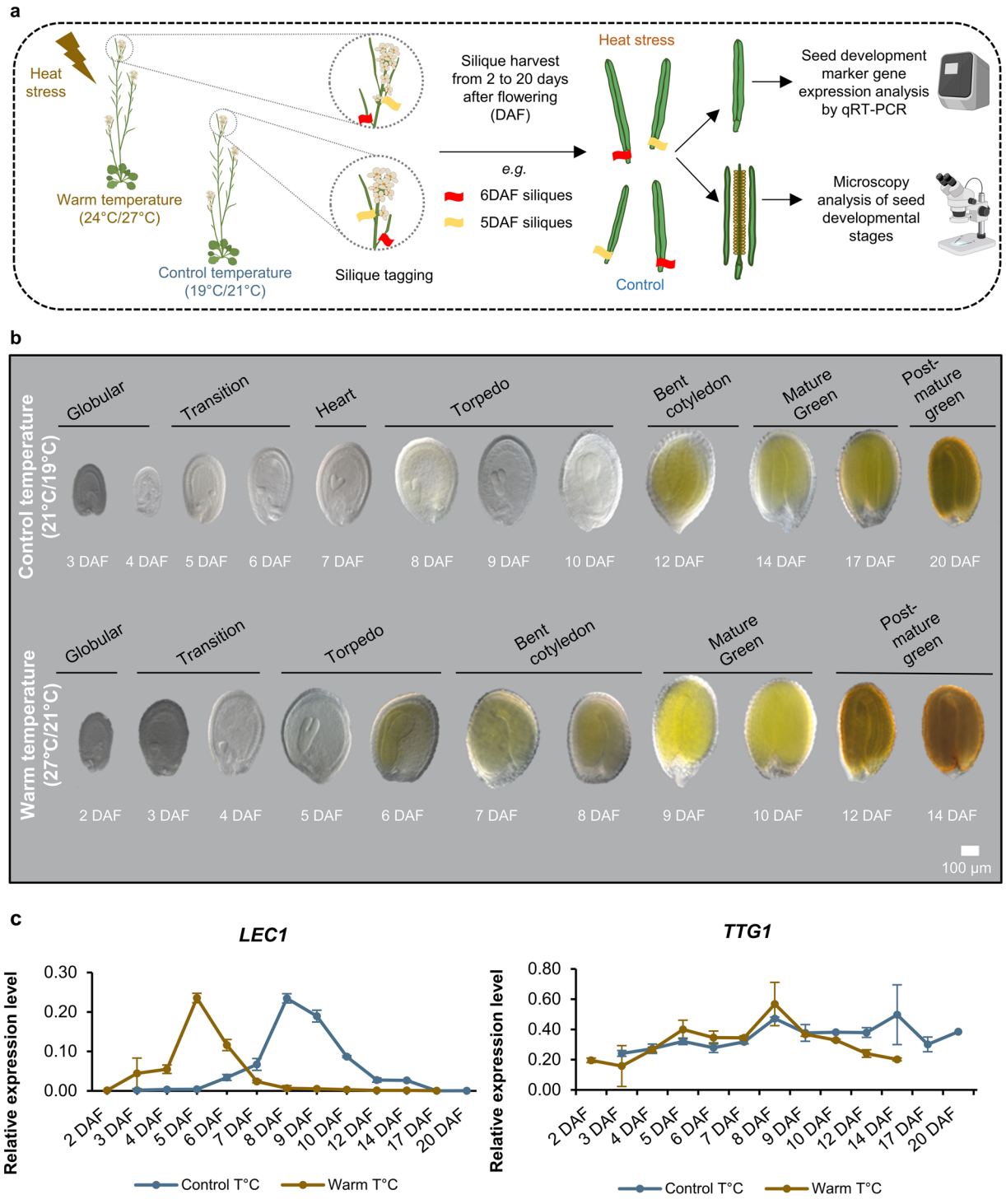

Fig. 1 Synchronisation of seed developmental stages between normal and warm temperature growing conditions. (**a**) Experimental procedure followed for the synchronisation of control and warm temperature seed developmental stages. Whole siliques were used to perform qRT-PCR analysis whereas siliques were dissected for seed microscopic analyses. (**b**) Seed developmental stages microscopically observed consecutive days after flowering (DAF) upon their development under normal and warm temperature conditions. (**c**) Expression of *LEC1* (*LEAFY COTYLEDON 1*; AT1G21970) and *TTG1* (*TESTA GLABRA 1*; AT5G24520) genes during seed development under normal and warm temperature developing conditions. Average expressions of technical replicates are represented. The software BioRender (biorender.com) was used to create this figure.

was made by adding 40 µL of DNase solution mix (10 µl DNase I Stock solution + 30 µl Buffer RDD; RNase-Free DNase Set, Qiagen) to each sample, followed by 15 min of incubation (ambient temperature) and washing by

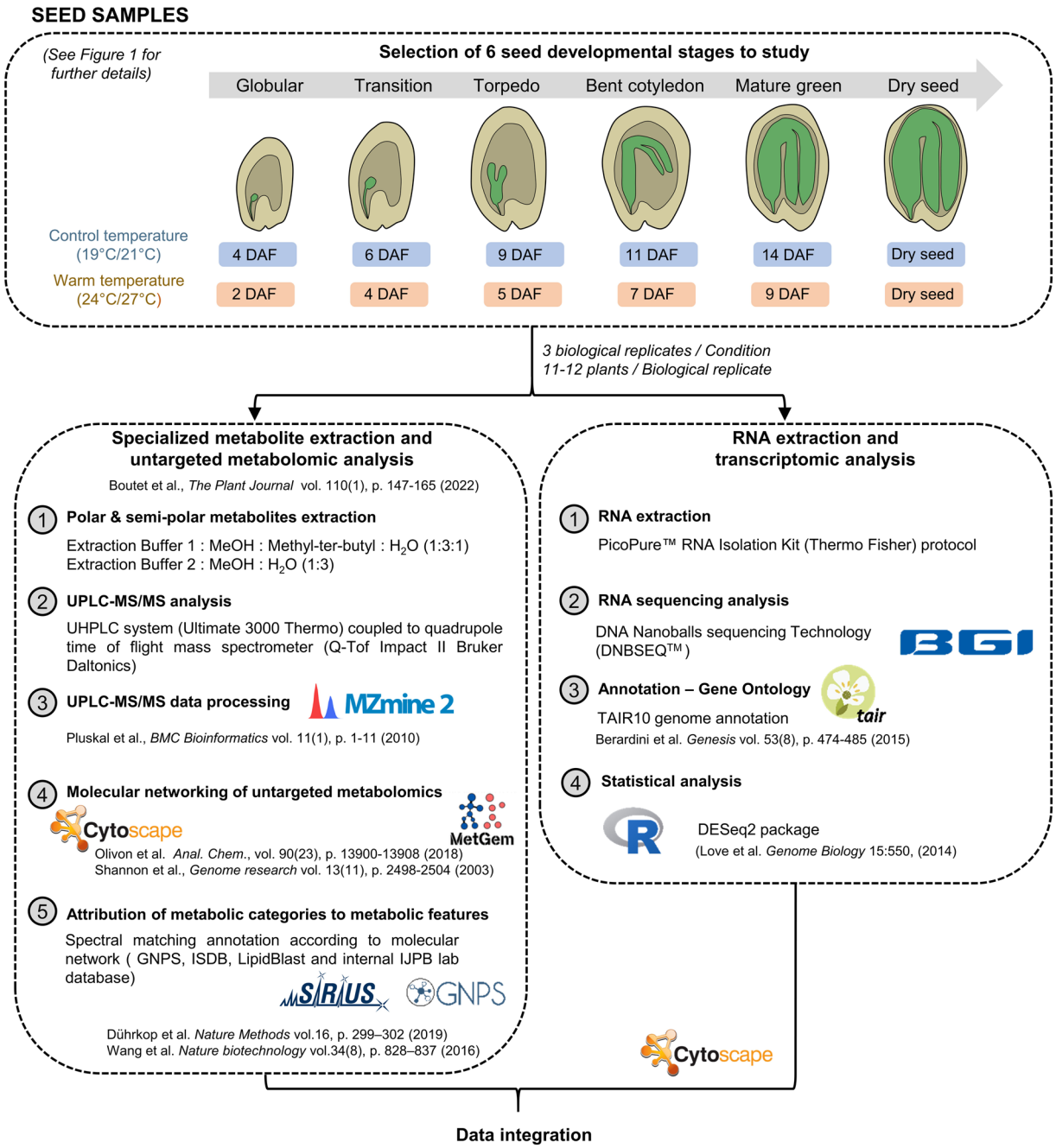

Fig. 2 Schematic representation of the experimental procedure followed to obtain untargeted metabolomics (LC-MS/MS) and transcriptomic (RNASeq) data from seeds produced under control or heat stress conditions, at six developmental stages (Globular, Transition, Heart, Torpedo, Bent cotyledon, Mature green and Dry seed). Three biological replicates, consisting in bulk seeds, from 11-12 plants, were used for each condition. DAF: days after flowering.

following the RNA Cleanup protocol (Quiagen). Briefly, RNA samples were adjusted to 100 µL with RNA-free water and 350 µL of Buffer RLT and 250 µL of ethanol (96%) were added. The mixtures were mixed well by pipetting, placed into RNeasy Mini spin columns (Quiagen), centrifuged (15 sec, 8 000 g) and flow-through was discarded. Spin columns were washed twice by adding 500 µL of Buffer RPE, followed by centrifugation (2 min, 8 000 g) with flow-through discard. Finally, the columns were placed in new collection tubes, 30 µL of RNase-free water were added to the columns which were centrifuged (1 min, 8 000 g) to elute RNAs. RNA concentrations and qualities were quantified with a NanoDrop (NanoDrop 2000C Spetrcophotometer) and stored at −80 °C prior to reverse transcription.

cDNA synthesis. 500 ng of total RNAs were used to perform cDNA retrotranscription. Briefly, 1 µL of 100 µM oligo$(dT)_{22}$, 2 µL of 5 mM dNTP and RNase-free water (q.s. to 14,75 µL) were added to RNA extracts. The

	Gene	Forward	Reverse
Reference genes	**Ubiquitine**	5′-AACAATTGGAGGATGGTCGT-3′	5′-AGCTGCCCAGCGAAGAT-3′
	EF-α	5′-ATGCCCCAGGACATCGTGATTTCAT-3′	5′-TTGGCGGCACCCTTAGCTGGATCA-3′
	Actin	5′-GGAAGGATCTGTACGGTAAC-3′	5′-GGACCTGCCTCATCATACT-3′
Targeted genes	**FUSC3**	QuantiTect Primer Assay (Qiagen) QT00780311	
	LEC1	QuantiTect Primer Assay (Qiagen) QT00863352	

Table 1. Primer sequences used for the real-time quantitative PCR analysis performed for seed development synchronisation.

reactions were incubated 5 min at 65 °C in a SimpliAmp thermal cycler (Applied biosystems by Thermo Fisher Scientific) and then kept on ice prior to the addition of 4 µL of the transcriptase inverse buffer 5X buffer (Thermo Fisher Scientific) and 1 µL of RiboLock (Thermo Fisher Scientific). The reactions were incubated 2 min at 42 °C in a SimpliAmp thermal cycler (Appliedbiosystems by Thermo Fisher Scientific. Subsequently, 0.25 µL of transcriptase inverse RevertAid H Minus (Thermo Fisher Scientific) were added. Samples were incubated at 42 °C for 50 min followed by 15 min at 70 °C in SimpliAmp thermal cycler (Appliedbiosystems by Thermo Fisher Scientific real-time thermal cycling system to perform the reverse transcription. The cDNAs obtained were stored at −20 °C prior to real-time quantitative polymerase chain reaction (PCR) quantification.

Real-time quantitative PCR. Real-time quantitative PCR (qRT-PCR) reactions were perfomed with a SimpliAmp thermal cycler (Appliedbiosystems by Thermo Fisher Scientific) using SYBR green to detect mRNA abundances. Briefly, reaction mixtures containing 5 µL of SsoAdvanced™ Universal SYBR® Green Supermix (Bio-Rad), 0,3 µL of each primer (10 µM) and 4,4 µL of cDNA were incubated for 8 min at 95 °C followed by 40 cycles at 95 °C for 10 s and 60 °C for 30 s and a last step of 15 sec at 95 °C. All reactions were carried out using three technical replicates. Information and sequences of the primers used for qRT-PCR analyses are included in Table 1.

Primer efficiencies were determined by quantifying cDNA concentrations in serial dilution of a mix of cDNA samples. Bio-Rad CFX Maestro™ Software was used to determine the primer efficiencies, corresponding to the slopes of the standard curves obtained (displaying the threshold cycle plotted against the log of starting quantity). Gene expressions were calculated with the following formula: E^{-Cq} with E being the primer efficiency and C_q being the average C_q (quantification cycle measured) of the three technical replicates of a cDNA sample[17]. The geometric means of the expressions of the reference genes *ACTIN2* (AT3G18780), *UBQ10* (AT4G05320) and *EF1-α* (AT5G60390) were used to normalize gene expressions by calculating relative gene expressions as follows: $Rg = (E_{GOI})^{-CqGOI}/GeoMean[(E_{REF})^{-CqREF}]$[18].

Heat stress experiment for multiomic analyses. *Plant culture and seeds harvests.* *Arabidopsis thaliana* Col-0 seeds were sown in fertilized soil. Plants (72 in total) were grown in optimal conditions ("control") in a growth chamber with long day photoperiod (16 h of light – 21 °C/ 8 h of dark – 19 °C). Watering was made with water. Heat stress was applied to concerned plants following the beginning of flowering (occurring between 5 and 6 weeks after seed sowing). Half of the plants were therefore transferred to another growth chamber (SANYO™, Thermo Fisher Scientific)(Warm temperature conditions: 16 h of light – 27 °C/ 8 h of dark – 24 °C), whereas the other half, corresponding to control plants, remained in the initial growth chamber with control temperature conditions. The temperatures from the different growth chambers were monitored throughout the experiment with thermometers (KIMO®). Three biological replicates (eleven to twelve plants per replica), were used for each condition. Upon the beginning of heat stress treatment, water content of each pot was carefully followed for both control and stressed plants. Plants were watered each day to maintain plant pot water capacities to 85–90%. Tagging of flower buds allowed to identify precisely the developmental stages of siliques and seeds. Control and heat stress developed seeds were harvested from the plants providing the biological replicates at five developmental stages according to the developmental stage-DAF correspondence determined during the synchronisation experiment. Hence to obtain seeds from Globular, Transition, Torpedo, Bent-cotyledon and Mature green stages, control developed seeds were harvested at 4, 6, 9, 11 and 14 DAFs respectively, whereas warm stress developed seeds were harvested at 2, 4, 5, 7 and 9 DAFs. Upon silique dissection, harvested seeds were directly frozen in Eppendorf tubes placed in liquid nitrogen and stored at −80 °C.

Once the majority of the siliques had reached mature stages, plant watering was progressively reduced to maintain plant pot water capacities to 65%, and then 50%, in order to favour the drying of silique and plants and to avoid potential pathogen infections related to high levels of humidity. Dry seeds were also harvested for both control and warm temperature conditions. Homogenization of the seed samples was made by grinding with a pylon in liquid nitrogen. Seed grinded samples were stored at −80 °C prior metabolomic and transcriptomic analysis.

Untargeted metabolomic analysis. Extraction of polar and semi-polar specialized metabolites. Polar and semi-polar metabolite fraction was extracted as previously described[13], with some modifications and adaptations to Arabidopsis seeds. Samples consisted of 5–7 mg of *A. thaliana* grinded seeds of each developmental stage studied (Globular, Transition, Torpedo, Bent cotyledon, Mature green and Dry seeds). Briefly, 1.5 mL of Methanol:Methyl-tert-butyl:Water (1:3:1) (stored at 4 °C) were added to each sample, that were subsequently homogenized using a FastPrep instrument (1 min, 14 000 rpm). The mixtures were then shaken for 30 min at 4 °C

using a ThermoMixer™ C (Eppendorf), placed in an ice-cooled ultrasonication bath for 15 min, and centrifuged for debris removal (1 min, 14 000 rpm). The extracted samples were then transferred to new Eppendorf tubes containing 650 mL of Methanol:Water (1:3), previously placed at - 20 °C. The mixtures were centrifuged for 1 min at 14 000 rpm, leading to a phase separation providing an upper organic phase (containing the non-polar lipid compounds), a lower aqueous phase (containing the polar and semi-polar metabolites), and a pellet of starch and proteins. The phases containing the polar/semi-polar metabolites were evaporated using a SpeedVac vacuum concentrator and stored at −80 °C. Shortly prior UPLC-MS/MS injection, extracted samples were resuspended in 150 µL of Acetonitrile:Water (1:9) of ULC/MS grade. 500 ng of Apigenin (stock solution of 50 ng/µL) was added to each extract to be used as internal standard. A pooled Quality Control (QC) sample was prepared by mixing equivalent aliquots of each extracted sample in order to assess the LC-MS/MS run quality.

UPLC-MS/MS analysis of polar and semi-polar specialized metabolites. As described by Boutet *et al.*[13], untargeted metabolomic data were acquired using a UHPLC system (Ultimate 3000 Thermo) coupled to quadrupole time of flight mass spectrometer (Q-Tof Impact II Bruker Daltonics, Bremen, Germany). A Nucleoshell RP 18 plus reversed-phase column (2 × 100 mm, 2.7 µm; Macherey-Nagel) was used for chromatographic separation with the mobile phases consisting in: (A) 0.1% formic acid in water; and (B) 0.1% formic acid in acetonitrile. The flow rate was 400 µL min^{-1}, and the following gradient was used: 95% (A) for 1 min, followed by a linear gradient from 95% (A) to 80% (A) from 1 to 3 min, then a linear gradient from 80% (A) to 75% (A) from 3 to 8 min, a linear gradient from 75% (A) to 40% (A) from 8 to 20 min; then, 0% of (A) was held until 24 min, followed by a linear gradient from 0% (A) to 95% (A) from 24 to 27 min. Finally, the column was washed by 30% (A) at for 3.5 min then re-equilibrated for 3.5 min (35 min total run time). Mass spectrometer data were obtained with data-dependent acquisition (DDA) method in both positive and negative electrospray ionisation (ESI) modes using the following parameters: capillary voltage, 4.5 kV; nebulizer gas flow, 2.1 bar; dry gas flow, 6.L.min^{-1}; drying gas in the heated electrospray source temperature, 140 °C. Samples were analysed at 8 Hz with a mass range of 100–1500 m/z. Stepping acquisition parameters were created to improve the fragmentation profile with a collision RF from 200 to 700 Vpp, a transfer time from 20 to 70 µsec, and collision energy from 20 to 40 eV. Each cycle included a MS fullscan and 5 MS/MS CID on the 5 main ions of the previous MS spectrum.

UPLC-MS/MS data processing. The UPLC-MS/MS data processing pipeline was performed as in Boutet *et al.*[13] with modifications. Briefly,.d data files (Bruker Daltonics, Bremen, Germany) were converted to.mzXML format using the MSConvert software (ProteoWizard package 3.0)[19]. mzXML data processing, mass detection, chromatogram building, deconvolution, samples alignment and data export were performed using MZmine 2.53 software (http://mzmine.github.io/) for both positive and negative data files. The ADAP chromatogram builder[20] method was used with a minimum group size of scan 4, a group intensity threshold of 1000, a minimum highest intensity of 1500 and m/z tolerance of 8 ppm. Deconvolution was performed with the ADAP wavelets algorithm using the following setting: S/N threshold 8, peak duration range = 0.01–2 min RT wavelet range 0.01–0.1 min, MS2 scan were paired using a m/z tolerance range of 0.01 Da and RT tolerance of 0.15 min. Then, isotopic peak grouper algorithm was used with a m/z tolerance of 8 ppm and RT tolerance of 0.15. All the peaks were filtered using feature list row filter keeping only peaks with MS2 scan. The alignment of samples was performed using the join aligner with an m/z tolerance of 8 ppm, a weight for m/z and RT at 1, a retention time tolerance of 0.15 min. Metabolite accumulation was normalized according to the internal standard (Apigenin) and weight of seeds used for the extraction. Pooled QC sample injections across LC-MS/MS run were used to evaluate the quality of the run and untargeted metabolomic dataset. The coefficient of variation ([QC average / QC standard deviation] × 100) across the QC samples was calculated for each metabolic feature detected (Table "Annotation and information of *Arabidopsis thaliana* seed specialized metabolome" [https://doi.org/10.57745/XSDTVE])[21]. The metabolic features were filtered according to their QC variation and only metabolic features with QC variation <35% were kept.

Molecular network generation of untargeted metabolomic data. As described previously, molecular networks were generated with MetGem software[22] (https://metgem.github.io) using the.mgf and .csv files obtained with MZmine2 analysis[13]. The molecular network was optimized for the ESI+ and ESI-datasets, and different cosine similarity score thresholds were tested in order to determine which value allowed to create several clusters of ions, allowing to expand the metabolic categories annotations of annotated ions (i.e. ions matching with databases annotations) to non-annotated ions, while not regrouping all the ions in a restricted number of clusters which would reduce the precision of annotation. ESI− and ESI+ molecular networks were both generated using a cosine score threshold of 0.8. Molecular networks were exported to Cytoscape software (v.3.10.1)[23] (https://cytoscape.org/) to format the metabolic categories (i.e. associate shapes and colours to the nodes [i.e. SMs] according to the metabolic categories to which they belong with the "Style" tool) (Fig. 3a).

Annotation of untargeted metabolomic data. Metabolite annotation was performed in four consecutive steps. First, the 'custom database search' module from Mzmine was used to compare the obtained LC-MS/MS data with the IJPB chemistry and metabolomic platform homemade experimental (m/z absolute tolerance of 0.0025 and RT tolerance of 0.3 min) and exact mass (m/z absolute tolerance of 0.0025 Da or 6 ppm) libraries containing respectively 166 standards or experimental common features (RT, m/z) and 1112 ion known m/z. Second, the ESI− and ESI+ metabolomic data used for molecular network analyses were searched against the available MS2 spectral libraries (Massbank NA, GNPS Public Spectral Library, NIST14 Tandem, NIH Natural Product and MS-Dial), with absolute m/z tolerance of 0.02, 4 minimum matched peaks and minimal cosine score of 0.8 (Fig. 3a). Third, not-annotated metabolites that belong to molecular network clusters containing annotated metabolites from steps 1 and 2 were assigned to the same metabolic category. Moreover, Sirius software was used to visualize the MS/MS spectra and confirm the putative

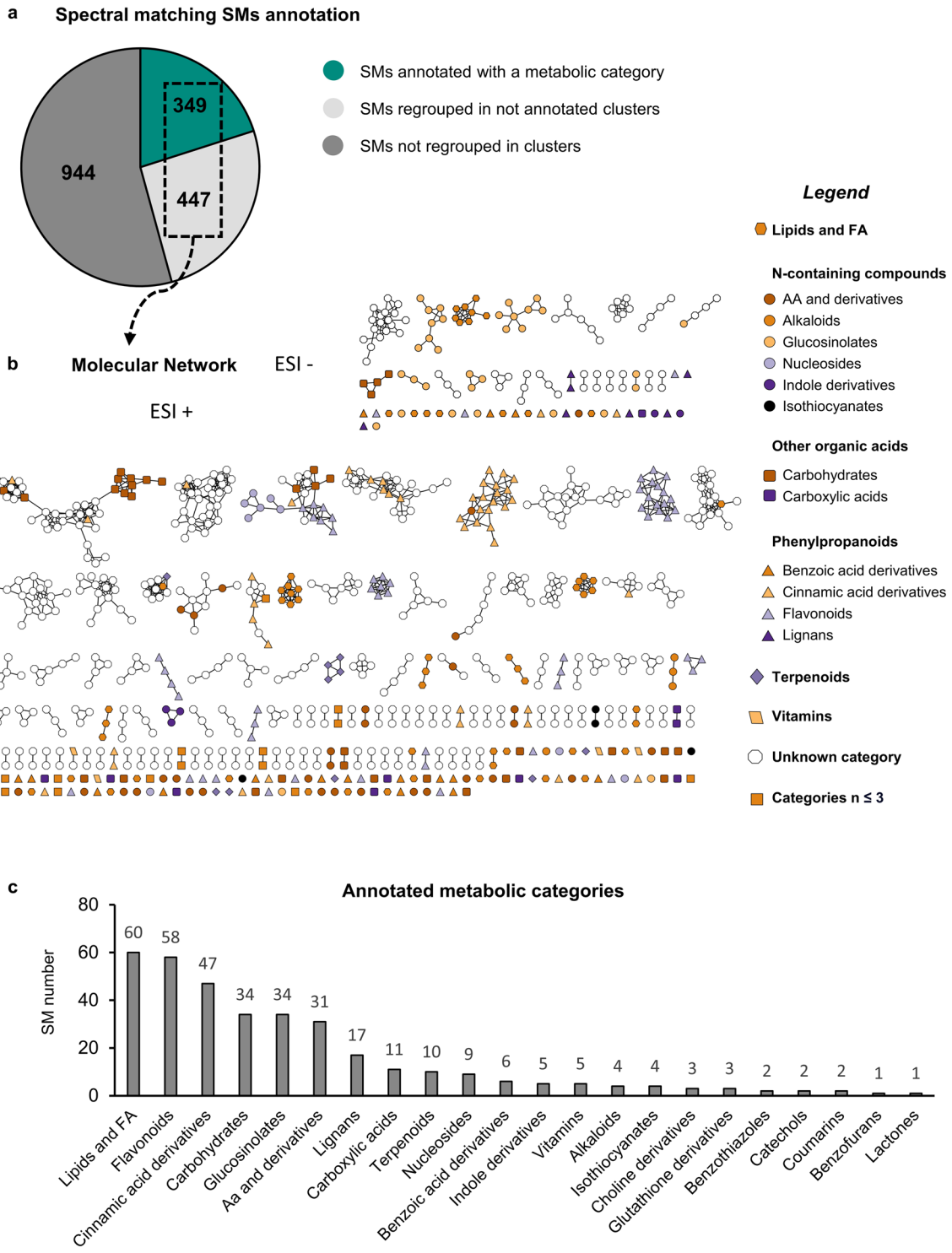

Fig. 3 Untargeted metabolomic analysis. (**a**) Spectral matching specialized metabolite annotation. (**b**) Molecular network. (**c**) Annotated metabolic categories. Metabolites belonging to lipids and FA, N-containing compounds, other organic acids, phenylpropanoids, terpenoids and vitamins groups are represented with different shapes. Different colors further precise the different metabolic categories. SMs: Specialized Metabolites; ESI: electrospray ionisation; AA: Amino acids; FA: Fatty acids.

annotation of some metabolic features[24]. The levels of metabolite annotation were settled up as recommended by the Chemical Analysis Working Group MSI[25,26]: (A) Compounds identified with chemical standards, Putatively annotated compounds: (Bi) based on m/z + RT and/or spectral matching with public m/z + RT and/or spectral libraries and (Bii) based on manual curation/annotation with Sirius software[24], (D) Putatively annotated compounds based

Sample ID	Sample	Stage	Day of harvest	Condition	Biological replicate	Metabolomic data		Transcriptomic data	
						ESI+	ESI−	File 1	File 2
1a_S	Globular_Ctr_1	Globular	4DAF	Control	1	LB-MC_Temp-II_1A-S_Pos.mzXML	LB-MC_Temp-II_1A-S_Neg.mzXML	1-1_ctr_1.fq.gz	1-1_ctr_2.fq.gz
1b_S	Globular_Ctr_2	Globular	4DAF	Control	2	LB-MC_Temp-II_1B-S_Pos.mzXML	LB-MC_Temp-II_1B-S_Neg.mzXML	1-2_ctr_1.fq.gz	1-2_ctr_2.fq.gz
1c_S	Globular_Ctr_3	Globular	4DAF	Control	3	LB-MC_Temp-II_1C-S_Pos.mzXML	LB-MC_Temp-II_1C-S_Neg.mzXML	1-3_ctr_1.fq.gz	1-3_ctr_2.fq.gz
2a_S	Transition_Ctr_1	Transition	6DAF	Control	1	LB-MC_Temp-II_2A-S_Pos.mzXML	LB-MC_Temp-II_2A-S_Neg.mzXML	2-1_ctr_1.fq.gz	2-1_ctr_2.fq.gz
2b_S	Transition_Ctr_2	Transition	6DAF	Control	2	LB-MC_Temp-II_2B-S_Pos.mzXML	LB-MC_Temp-II_2B-S_Neg.mzXML	2-2_ctr_1.fq.gz	2-2_ctr_2.fq.gz
2c_S	Transition_Ctr_3	Transition	6DAF	Control	3	LB-MC_Temp-II_2C-S_Pos.mzXML	LB-MC_Temp-II_2C-S_Neg.mzXML	2-3_ctr_1.fq.gz	2-3_ctr_2.fq.gz
3a_S	Torpedo_Ctr_1	Torpedo	9DAF	Control	1	LB-MC_Temp-II_3A-S_Pos.mzXML	LB-MC_Temp-II_3A-S_Neg.mzXML	3-1_ctr_1.fq.gz	3-1_ctr_2.fq.gz
3b_S	Torpedo_Ctr_2	Torpedo	9DAF	Control	2	LB-MC_Temp-II_3B-S_Pos.mzXML	LB-MC_Temp-II_3B-S_Neg.mzXML	3-2_ctr_1.fq.gz	3-2_ctr_2.fq.gz
3c_S	Torpedo_Ctr_3	Torpedo	9DAF	Control	3	LB-MC_Temp-II_3C-S_Pos.mzXML	LB-MC_Temp-II_3C-S_Neg.mzXML	3-3_ctr_1.fq.gz	3-3_ctr_2.fq.gz
4a_S	Bent_cotyledon_Ctr_1	Bent_cotyledon	11DAF	Control	1	LB-MC_Temp-II_4A-S_Pos.mzXML	LB-MC_Temp-II_4A-S_Neg.mzXML	4-1_ctr_1.fq.gz	4-1_ctr_2.fq.gz
4b_S	Bent_cotyledon_Ctr_2	Bent_cotyledon	11DAF	Control	2	LB-MC_Temp-II_4B-S_Pos.mzXML	LB-MC_Temp-II_4B-S_Neg.mzXML	4-2_ctr_1.fq.gz	4-2_ctr_2.fq.gz
4c_S	Bent_cotyledon_Ctr_3	Bent_cotyledon	11DAF	Control	3	LB-MC_Temp-II_4C-S_Pos.mzXML	LB-MC_Temp-II_4C-S_Neg.mzXML	4-3_ctr_1.fq.gz	4-3_ctr_2.fq.gz
5a_S	Mature_green_Ctr_1	Mature_green	14DAF	Control	1	LB-MC_Temp-II_5A-S_Pos.mzXML	LB-MC_Temp-II_5A-S_Neg.mzXML	5-1_ctr_1.fq.gz	5-1_ctr_2.fq.gz
5b_S	Mature_green_Ctr_2	Mature_green	14DAF	Control	2	LB-MC_Temp-II_5B-S_Pos.mzXML	LB-MC_Temp-II_5B-S_Neg.mzXML	5-2_ctr_1.fq.gz	5-2_ctr_2.fq.gz
5c_S	Mature_green_Ctr_3	Mature_green	14DAF	Control	3	LB-MC_Temp-II_5C-S_Pos.mzXML	LB-MC_Temp-II_5C-S_Neg.mzXML	5-3_ctr_1.fq.gz	5-3_ctr_2.fq.gz
6a_S	Dry_seed_Ctr_1	Dry_seed	Dry_seed	Control	1	LB-MC_Temp-II_6A-S_Pos.mzXML	LB-MC_Temp-II_6A-S_Neg.mzXML		
6b_S	Dry_seed_Ctr_2	Dry_seed	Dry_seed	Control	2	LB-MC_Temp-II_6B-S_Pos.mzXML	LB-MC_Temp-II_6B-S_Neg.mzXML		
6c_S	Dry_seed_Ctr_3	Dry_seed	Dry_seed	Control	3	LB-MC_Temp-II_6C-S_Pos.mzXML	LB-MC_Temp-II_6C-S_Neg.mzXML		
7a_S	Globular_Str_1	Globular	2DAF	Heat stress	1	LB-MC_Temp-II_7A-S_Pos.mzXML	LB-MC_Temp-II_7A-S_Neg.mzXML	1-1_stress_1.fq.gz	1-1_stress_2.fq.gz
7b_S	Globular_Str_2	Globular	2DAF	Heat stress	2	LB-MC_Temp-II_7B-S_Pos.mzXML	LB-MC_Temp-II_7B-S_Neg.mzXML	1-2_stress_1.fq.gz	1-2_stress_2.fq.gz
7c_S	Globular_Str_3	Globular	2DAF	Heat stress	3	LB-MC_Temp-II_7C-S_Pos.mzXML	LB-MC_Temp-II_7C-S_Neg.mzXML	1-3_stress_1.fq.gz	1-3_stress_2.fq.gz
8a_S	Transition_Str_1	Transition	4DAF	Heat stress	1	LB-MC_Temp-II_8A-S_Pos.mzXML	LB-MC_Temp-II_8A-S_Neg.mzXML	2-1_stress_1.fq.gz	2-1_stress_2.fq.gz
8b_S	Transition_Str_2	Transition	4DAF	Heat stress	2	LB-MC_Temp-II_8B-S_Pos.mzXML	LB-MC_Temp-II_8B-S_Neg.mzXML	2-2_stress_1.fq.gz	2-2_stress_2.fq.gz
8c_S	Transition_Str_3	Transition	4DAF	Heat stress	3	LB-MC_Temp-II_8C-S_Pos.mzXML	LB-MC_Temp-II_8C-S_Neg.mzXML	2-3_stress_1.fq.gz	2-3_stress_2.fq.gz
9a_S	Torpedo_Str_1	Torpedo	5DAF	Heat stress	1	LB-MC_Temp-II_9A-S_Pos.mzXML	LB-MC_Temp-II_9A-S_Neg.mzXML	3-1_stress_1.fq.gz	3-1_stress_2.fq.gz
9b_S	Torpedo_Str_2	Torpedo	5DAF	Heat stress	2	LB-MC_Temp-II_9B-S_Pos.mzXML	LB-MC_Temp-II_9B-S_Neg.mzXML	3-2_stress_1.fq.gz	3-2_stress_2.fq.gz
9c_S	Torpedo_Str_3	Torpedo	5DAF	Heat stress	3	LB-MC_Temp-II_9C-S_Pos.mzXML	LB-MC_Temp-II_9C-S_Neg.mzXML	3-3_stress_1.fq.gz	3-3_stress_2.fq.gz
10a_S	Bent_cotyledon_Str_1	Bent_cotyledon	7DAF	Heat stress	1	LB-MC_Temp-II_10A-S_Pos.mzXML	LB-MC_Temp-II_10A-S_Neg.mzXML	4-1_stress_1.fq.gz	4-1_stress_2.fq.gz
10b_S	Bent_cotyledon_Str_2	Bent_cotyledon	7DAF	Heat stress	2	LB-MC_Temp-II_10B-S_Pos.mzXML	LB-MC_Temp-II_10B-S_Neg.mzXML	4-2_stress_1.fq.gz	4-2_stress_2.fq.gz
10c_S	Bent_cotyledon_Str_3	Bent_cotyledon	7DAF	Heat stress	3	LB-MC_Temp-II_10C-S_Pos.mzXML	LB-MC_Temp-II_10C-S_Neg.mzXML	4-3_stress_1.fq.gz	4-3_stress_2.fq.gz
11a_S	Mature_green_Str_1	Mature_green	9DAF	Heat stress	1	LB-MC_Temp-II_11A-S_Pos.mzXML	LB-MC_Temp-II_11A-S_Neg.mzXML	5-1_stress_1.fq.gz	5-1_stress_2.fq.gz

Continued

Sample ID	Sample	Stage	Day of harvest	Condition	Biological replicate	Metabolomic data		Transcriptomic data	
						ESI+	ESI−	File 1	File 2
11b_S	Mature_green_Str_2	Mature_green	9DAF	Heat stress	2	LB-MC_Temp-II_11B-S_Pos.mzXML	LB-MC_Temp-II_11B-S_Neg.mzXML	5-2_stress_1.fq.gz	5-2_stress_2.fq.gz
11c_S	Mature_green_Str_3	Mature_green	9DAF	Heat stress	3	LB-MC_Temp-II_11C-S_Pos.mzXML	LB-MC_Temp-II_11C-S_Neg.mzXML	5-3_stress_1.fq.gz	5-3_stress_2.fq.gz
12a_S	Dry_seed_Str_1	Dry_seed	Dry_seed	Heat stress	1	LB-MC_Temp-II_12A-S_Pos.mzXML	LB-MC_Temp-II_12A-S_Neg.mzXML		
12b_S	Dry_seed_Str_2	Dry_seed	Dry_seed	Heat stress	2	LB-MC_Temp-II_12B-S_Pos.mzXML	LB-MC_Temp-II_12B-S_Neg.mzXML		
12c_S	Dry_seed_Str_3	Dry_seed	Dry_seed	Heat stress	3	LB-MC_Temp-II_12C-S_Pos.mzXML	LB-MC_Temp-II_12C-S_Neg.mzXML		

Table 2. Overview of the dataset comprising untargeted metabolomic (LC-MS/MS) and transcriptomic data (RNA-Seq) for six seed developmental stages Globular, Transition, Torpedo, Bent cotyledon, Mature green and Dry seed), and two seed developing conditions (Control and heat stress).

on m/z matching with m/z library. (M) Putatively characterized compound metabolic category based on molecular network [i.e. based on spectral similarity to known compounds of a chemical class]).

Following this annotation approach, 20% of the metabolic features of the LC-MS/MS dataset were assigned to one of the 22 metabolic categories identified (Lipids and Fatty Acids [FA], Flavonoids, Cinnamic acid derivatives, Carbohydrates, Glucosinolates, Amino Acids (AA) and derivatives, Lignans, Carboxylic acids, Terpenoids, Nucleosides, Benzoic acid derivatives, Indole derivatives, Vitamins, Alkaloids, Isothiocyanates, Choline derivatives, Glutathione derivatives, Benzothiazoles, Catechols, Coumarins, Benzofurans and Lactones (Fig. 3b, Table "Annotation and information of *Arabidopsis thaliana* seed specialized metabolome" [https://doi.org/10.57745/XSDTVE])[21]. Moreover, 25.6% of the metabolic features that were not annotated but still regrouped in clusters (and therefore sharing similar chemical structures and most probably belonging to the same or close-related metabolic categories) were classed according to the clusters to which they belong (Fig. 3a, Table "Annotation and information of *Arabidopsis thaliana* seed specialized metabolome" [https://doi.org/10.57745/XSDTVE])[21].

Metabolites table and redundancy in mass spectrometry. Redundancy cleaning due to the mass spectrometry technique was performed using a local application (see Code availability) built using the freely available Shiny R package (https://cran.r-project.org/web/packages/shiny/index.html). It allows to identify the formation of adducts complementary to the protonated form in positive mode as: sodium potassium and ammonium and de-protonated in negative mode as: sodium, potassium, chlorine, formic acid, nitric acid, acetate, formic acid clusters associated with sodium by comparing the masses and retention time.

In the same way, the complementarity of positive and negative modes for certain metabolites was also eliminated by cross-referencing the modes of these complementary forms with those of their adducted forms. Both analyses were done using a mass tolerance parameter of 0.006 Da and a retention time of 0.2 minutes.

Statistical analysis of metabolomic data. Statistical analyses were performed using Metaboanalyst 6.0 software[27,28]. Austoscale and log10 transformation were applied to the normalized data to perform the statistical analyses. An ANOVA test has been conducted to identify differentially accumulated metabolites among seed developmental stages (globular, Transition, Torpedo, Mature green, Bent cotyledon and Dry seed) and seed developing conditions (control and HS).

Transcriptomic analysis. RNA extraction. RNAs were extracted with the PicoPure™ RNA Isolation Kit (ThermoFisher) following the corresponding protocol with slight modifications. Briefly, 2 to 5 mg of seeds were grinded with some polyvinylpolypyrrolidone (PVPP; phenolic compound absorbing agent) in liquid nitrogen. 100 µL of extraction buffer was added to each sample. Following, samples were vortexed, incubated at 42 °C for 30 min (500 rpm) using a ThermoMixer™ C (Eppendorf), and centrifuged (2 min, 3000 g). Supernatants were collected, without picking-up the pelleted material, and placed in the pre-conditioned RNA purification columns to be centrifuged (1 min, 16 000 g). The cell extracts obtained were mixed with 100 µL of 70% ethanol by pipetting up and down. The mixtures were placed into the pre-conditioned purification columns which were centrifuged 2 min at 100 g, to allow RNA binding to the columns, immediately followed by a centrifugation to remove flow-through (30 sec, 16 000 g). The columns were washed by adding 100 µL of wash buffer 1 and centrifugation (1 min, 8 000 g). DNase treatment was made by adding 40 µL of DNase solution mix (10 µl DNase I Stock solution + 30 µl Buffer RDD; RNase-Free DNase Set, Qiagen) to each column, followed by 15 min of incubation (ambient temperature) and washing of the column (addition of 40 µL of wash buffer 1 and centrifugation 30 sec at 8 000 g). Columns were washed twice with 100 µL of wash buffer 2 (2 min at 8 000 g) and a last centrifugation was made after flow-through waste discard to remove all traces of wash buffer (1 min, 16 000 g). 12 µL of elution buffer

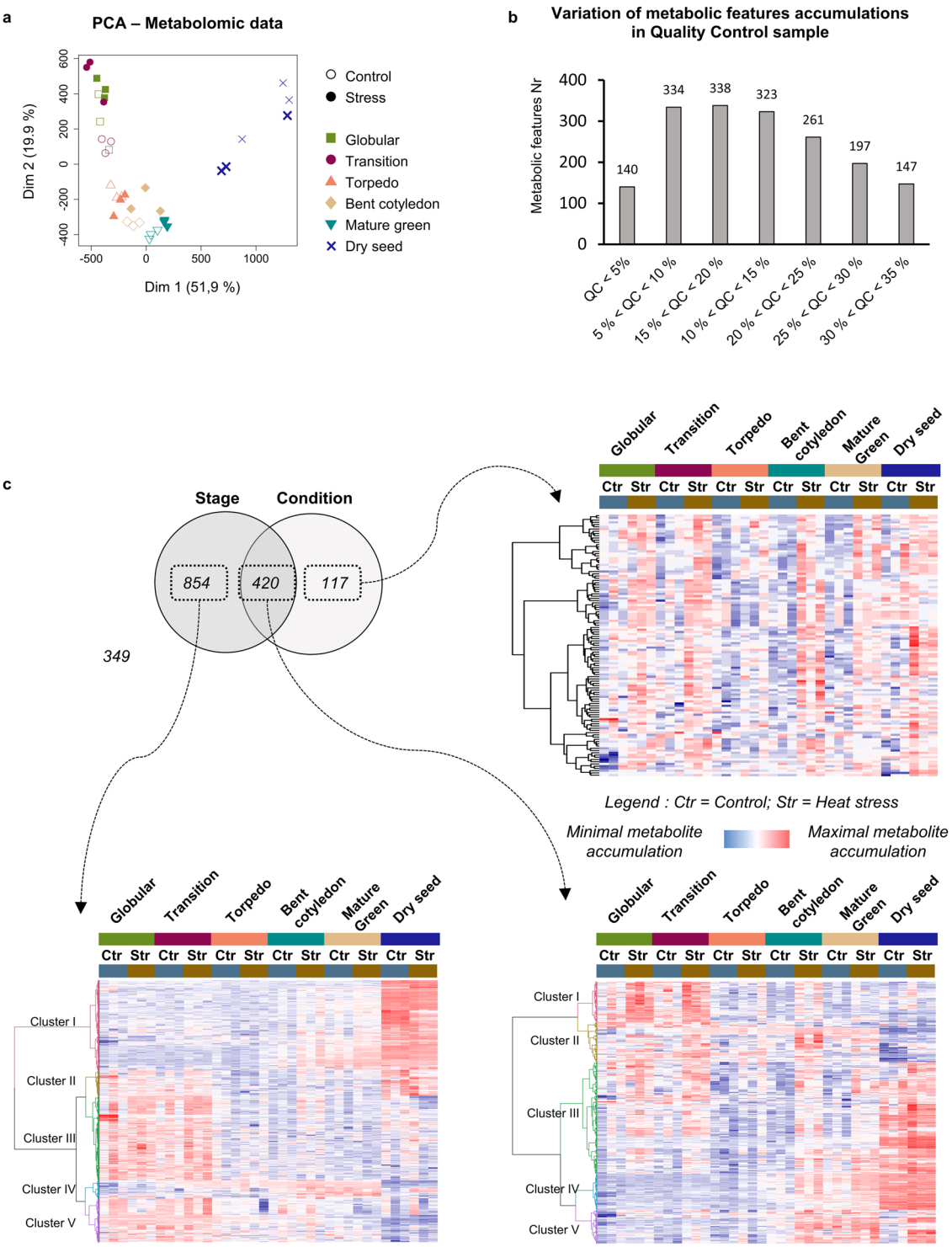

Fig. 4 Technical validation of untargeted metabolomic data. (**a**) Principal component analysis of untargeted metabolomic data. (**b**) Variation of metabolic features accumulations in quality control (QC) sample. (**c**) Two-way ANOVA analysis allowed to identify differentially accumulated metabolites (DAMs) affected by seed developmental stage (Globular, Transition, Heart, Torpedo, Bent cotyledon, Mature green and Dry seed) and/or seed development condition (control or heat stress). The analysis highlighted (854 DAMs specific to seed developmental stage, 117 DAMs specific to heat stress and 420 DAMs affected by both seed developmental stage and heat stress. 349 metabolites were not differentially accumulated (FDR > 0.05). Ctr: Control; Str: Stress.

were added to each column, which were incubated 1 min at ambient temperature before RNA elution by centrifugations (1 min at 1 000 g followed by 1 min at 16 000 g). Obtained RNAs were stored at −80 °C until sequencing.

	Ionisation mode acquisition	QC	Seed Samples
Internal standard (Apigenin) variation (%)	ESI+	6.12229803	14.05645164
	ESI−	12.5461795	16.79333219

Table 3. Metabolomic data validation: internal standard (apigenin) variation across quality control (QC) and biological samples.

RNA sequencing. Sequencing of messenger RNAs (mRNAs) was performed by BGI Genomics (Hong Kong, https://www.bgi.com). Prior to cDNA library construction (Eukaryotic Strand-specific Transcriptome), quality and concentrations of mRNAs were checked with the bioanalyzer from BGI platform. The sequencing (paired-end reads of 100 bp, 20 M clean reads per sample) was performed with DNA Nanoballs Technology (DNBSEQ™). After sequencing, raw reads were filtered by the BGI platform with the filtering software SOAPnuke (version 1.5.2, https://github.com/BGI-flexlab/SOAPnuke)[29,30] (Parameters used: l 15 -q 0.2 -n 0.05) to reach a Phred + 33 fastq quality score.

Annotation and statistical analysis of transcriptomic data. High quality raw reads were aligned to the *Arabidopsis thaliana* genome (TAIR10.1) using HISAT[31] (Parameters:–sensitive–no-discordant–no-mixed -I 1 -X 1000 -p 8). After that, Ericscript (0.5.5–5)[32] (Parameters: Default) and rMATS (v4.1.1)[33] (Parameters: Defaults) were used to detect fusion genes and differential splicing genes (DSGs), respectively. Bowtie2[34] (Parameters: -q–sensitive–dpad 0–gbar 99999999–mp 1,1–np 1–score-min L,0,-0.1 -I 1 -X 1000–no-mixed–no-discordant -p 1 -k 200) was applied to align the clean reads to the gene set. Reads were counted and normalized using the R package DESeq2[35]. In order to evaluate the individual effects of the seed developmental stage (Globular, Transition, Torpedo, Bent cotyledon, Mature green and Dry seed), and growing condition (Control and heat stress) on gene expression, a multifactorial analysis was conducted using the multifactor designs method of the DEseq2 R package[35] (https://bioconductor.org/packages/release/bioc/html/DESeq2.html). This method evaluates the weight of each factor considered in the analysis and its impact on differentially expressed genes (DEGs), according to a false discovery rate (FDR)-adjusted P-value < 0.05.

Principal Component Analysis, heatmap and hierarchical clustering analysis. The Principal Component Analysis (PCA) of both metabolomic and transcriptomic data were performed on R Studio using the 'PCA' and 'fviz_pca_ind' functions from 'FactoMineR' (https://cran.r-project.org/web/packages/FactoMineR/), 'Factoextra' (https://cran.r-project.org/web/packages/factoextra/) and 'dplyr' (https://cran.r-project.org/web/packages/dplyr/) R packages. The heatmap and hierarchical clustering analysis of both metabolomic and transcriptomic data were performed on R studio using 'pheatmap' and 'as.dendogram" functions from 'pheatmap' (https://cran.r-project.org/web/packages/pheatmap/), 'dentextend' (https://cran.r-project.org/web/packages/dendextend/) and 'tidyverse' (https://cran.r-project.org/web/packages/tidyverse/) R packages.

Metabolomic and transcriptomic data integration. Untargeted metabolomic and transcriptomic data from five seed developmental stages (Globular, Transition, Torpedo, Bent cotyledon and Mature green) developed under both control and heat stress conditions were used for data integration. The 'ExpressionCorrelation' app from Cytoscape (v.3.10.1)[23] software was used with a High Cuttoff of 0.7 to calculate the correlation between the expressions of the differentially expressed genes and the accumulation of all the SMs. According to the number of samples included in our correlation analysis (30), a correlation coefficient of 0.7 is strongly statistically significant (p-value ≤ 0,0001), ensuring that only strong and reliable correlations were retained.

Data Records

The data record contains the annotation and relative accumulations of specialized metabolites (Relative Unit [RU] normalized according to the intensity of Apigenin) from six seed developmental stages (Globular, Transition, Torpedo, Bent cotyledon, Mature Green and Dry seed) of *A. thaliana* developed either in control or heat stress conditions as well as the gene expression levels (normalized counts) of the corresponding five first developmental stages (Fig. 2). The data were obtained from three biological replicates (11-12 plants/replicate).

An overview of the dataset is presented in Table 2 with the following 8 columns:

1. Sample ID: sample identification
2. Sample: sample full name
3. Stage: seed developmental stage (Globular, Transition, Torpedo, Bent cotyledon, Mature green, Dry seed)
4. Day of harvest: seed sample harvest day (2, 4, 5, 6, 7, 9, 11, 14 DAF or dry seed)
5. Condition: condition of seed development (control or heat stress)
6. Biological replicate: the biological replicate (1,2,3)
7. Metabolomic data: ESI+ and ESI− raw data file names
8. Transcriptomic data: transcriptomic raw data file names (two files per sample: File 1 and File 2)

Raw metabolomic data. Untargeted metabolomic raw data (.mzXML) for both negative and positive ESI modes have been deposited at the MassiVE data repository portal (https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp) with the identifier MSV000095846 (https://doi.org/10.25345/C5S17T468)[36].

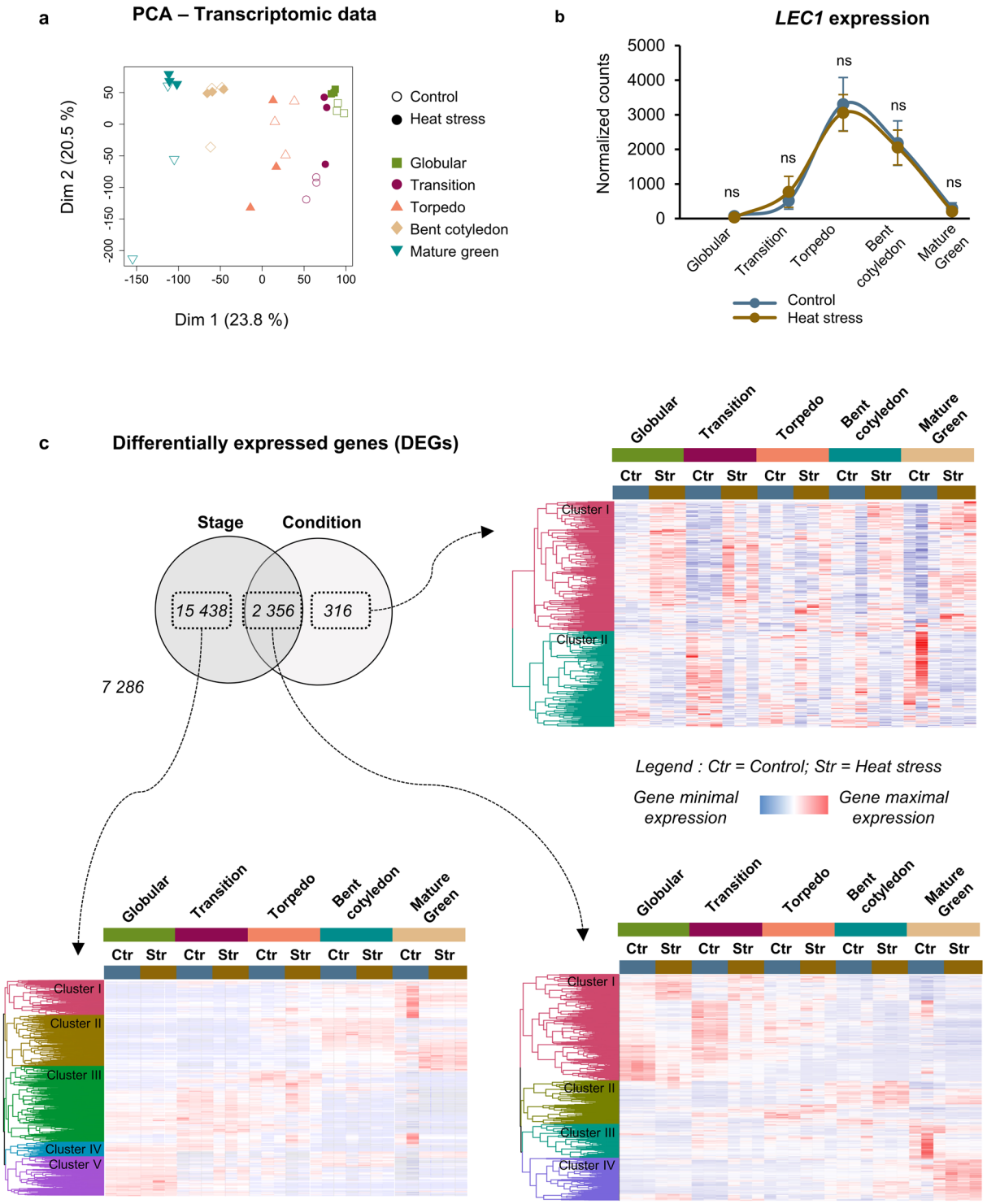

Fig. 5 Technical validation of transcriptomic data. (**a**) Principal component analysis of transcriptomic data. (**b**) Expression of *LEC1* (*LEAFY COTYLEDON 1*; AT1G21970) gene during seed development under normal and warm temperature growing conditions. Wilcoxon signed rank test (P-value < 0.05). ns, non-significant difference. (**c**) Multifactorial analysis allowed to identify differentially expressed genes (DEGs) affected by seed developmental stage (Globular, Transition, Heart, Torpedo, Bent cotyledon and Mature green) and/or seed development condition (control or heat stress). The analysis highlighted 15 438 DEGs specific to seed developmental stage, 316 DEGs specific to heat stress and 2 356 DEGs affected by both seed developmental stage and heat stress. 7 286 genes were not differentially expressed (FDR > 0.05). Ctr: Control; Str: Stress.

Raw transcriptomic data. The transcriptomic RNA-Seq raw data (FASTQ) have been deposited at the National Center for Biotechnology Information (NCBI) Transcriptome Shotgun Assembly Sequence Database (TSA) (https://www.ncbi.nlm.nih.gov/genbank/tsa/) with the SRP accession number: SRX24995807[37].

sample_name	total_clean_read	total_mapping_gene_ratio	uniquely_mapping_gene_ratio
1-1_ctr	47.52	92.84	86.90
1-1_stress	45.04	92.98	86.96
1-2_ctr	47.44	92.61	86.65
1-2_stress	44.95	92.41	86.14
1-3_ctr	44.92	92.59	86.79
1-3_stress	44.90	92.82	86.71
2-1_ctr	44.83	92.13	85.12
2-1_stress	50.58	92.35	86.15
2-2_ctr	44.94	93.07	86.54
2-2_stress	44.86	93.13	86.64
2-3_ctr	44.83	90.35	83.70
2-3_stress	44.66	93.02	86.95
3-1_ctr	47.42	93.46	87.20
3-1_stress	46.92	93.14	84.01
3-2_ctr	47.50	93.21	88.00
3-2_stress	44.83	93.12	85.87
3-3_ctr	47.48	93.57	87.47
3-3_stress	50.19	93.10	85.96
4-1_ctr	44.66	93.29	84.09
4-1_stress	44.62	92.85	84.26
4-2_ctr	44.76	93.26	85.20
4-2_stress	49.69	90.26	81.54
4-3_ctr	44.46	90.83	81.98
4-3_stress	44.05	93.10	84.49
5-1_ctr	44.21	91.55	80.45
5-1_stress	45.07	93.41	81.58
5-2_ctr	43.98	91.72	78.59
5-2_stress	44.26	94.03	81.76
5-3_ctr	44.39	94.20	83.22
5-3_stress	44.54	94.43	81.90

Table 4. Transcriptomic validation: reads filtering.

Tables including the processed metabolomic and transcriptomic data, statistical analysis and data integration analysis. The processed metabolomic data (including SM annotation, metabolic category, annotation level, m/z, RT, QC variation (Table "Annotation and information of *Arabidopsis thaliana* seed specialized metabolome" [https://doi.org/10.57745/XSDTVE])[21], the SM relative accumulations (Table "Relative accumulation of *Arabidopsis thaliana* specialized metabolites during seed development under heat stress and control conditions" [LC-MS/MS] [https://doi.org/10.57745/XSDTVE])[21], the metabolomic data statistical analyses results (Table "Statistical analysis on metabolomic data" [https://doi.org/10.57745/XSDTVE])[21], the RNA sequencing processed data (including transcripts annotation and expression table (Table "Normalized read counts of gene expression" [https://doi.org/10.57745/XSDTVE])[21], the transcriptomic data statistical analyses results (Table "Statistical analysis on transcriptomic data" [https://doi.org/10.57745/XSDTVE])[21], references of seed tissues gene expression (Seed tissue gene expression [https://doi.org/10.57745/XSDTVE])[21] and the multi-omic data integration analysis results (Table "Metabolomic and transcriptomic data integration" [https://doi.org/10.57745/XSDTVE])[21] have been deposited at the Data Gouv (Data INRAE) repository portal (https://doi.org/10.57745/XSDTVE).

Technical Validation

Under warm temperatures, seeds achieve their development faster[14,15]. A first plant culture was performed to find a correspondence between day(s) after flowering (DAF) and seed developmental stages (e.g. Globular, Transition, Heart, Torpedo, Bent cotyledon, Mature green, Post-mature green and Dry seed) under control and warm temperature conditions (Fig. 1). Siliques from both control and heat stress conditions were harvested at 11-12 time-points (from 2 to 10 DAF and 12, 14, 17 and 20 DAF) and used for microscopic and gene expression analyses (Fig. 1a). These data revealed that seeds reached dry seed mature stage almost one week earlier under warm temperature than in control conditions (Fig. 1b). Moreover, the faster development of seeds under stress was also highlighted by the expressions of developmental-stage specific marker genes (Fig. 1c). The expression of *LEC1* (*LEAFY COTYLEDON 1*; AT1G21970) peaked at 8 DAF for control grown siliques whereas it showed a peak 3 days earlier (8 DAF) for stress developed siliques (Fig. 1c). According to microscopic phenotypical analysis both of these time points correspond to the Torpedo stage, for each of the conditions considered (Fig. 1b). *LEC1*, is a transcription factor controlling embryo and endosperm development (including embryo morphogenesis, photosynthesis, and accumulation of storage compounds) and, as confirmed by our results, its

sample_name	total_clean_read	total_mapping_genome_ratio	uniquely_mapping_genome_ratio
1-1_ctr	47.52	96.46	94.97
1-1_stress	45.04	97.98	96.28
1-2_ctr	47.44	95.88	94.36
1-2_stress	44.95	98.07	96.24
1-3_ctr	44.92	97.54	95.93
1-3_stress	44.90	97.79	96.10
2-1_ctr	44.83	96.63	94.69
2-1_stress	50.58	96.48	94.90
2-2_ctr	44.94	97.18	95.37
2-2_stress	44.86	97.05	95.29
2-3_ctr	44.83	94.54	92.77
2-3_stress	44.66	97.60	96.02
3-1_ctr	47.42	96.81	95.27
3-1_stress	46.92	97.18	94.92
3-2_ctr	47.50	96.64	95.19
3-2_stress	44.83	97.60	95.73
3-3_ctr	47.48	96.69	95.33
3-3_stress	50.19	97.40	95.81
4-1_ctr	44.66	97.96	95.86
4-1_stress	44.62	97.53	95.19
4-2_ctr	44.76	98.16	96.21
4-2_stress	49.69	92.71	90.67
4-3_ctr	44.46	95.36	93.34
4-3_stress	44.05	97.58	95.47
5-1_ctr	44.21	94.96	92.98
5-1_stress	45.07	96.38	93.71
5-2_ctr	43.98	95.35	93.42
5-2_stress	44.26	97.47	95.54
5-3_ctr	44.39	97.62	95.69
5-3_stress	44.54	97.69	95.95

Table 5. Transcriptomic validation: genome mapping summary.

expression peaks at Torpedo stage, at 8 days after pollination, under control conditions[38,39]. The expression of *TTG1* (*TRANSPARENT TESTA GLABRA 1*; AT5G24520) is characterized by two peaks of expression, occurring at 8 DAF and 14 DAF in siliques grown in control conditions, and at 5 DAF and 8 DAF for siliques developed under stress (Fig. 1c). According to microscopic phenotypical analysis, both of these time points correspond to the Torpedo stage and the transition from Bent cotyledon to Mature Green stage, for each of the conditions considered (Fig. 1b). TTG1 is involved in the biosynthesis of proanthocyanidins (PAs) in the seed coat during seed development and it has been reported to have a maximal expression at 14 DAFs[16,40]. These data were essential to synchronise the developmental stages of seeds harvested from plants grown in control (optimal) or warmer conditions, and to appropriately harvest seeds used for multi-omic analyses.

The Principal Component Analysis (PCA) performed on untargeted metabolomic data allowed to separate the samples according to the seed developmental stage (Fig. 4a), indicating that seed developmental synchronisation (Fig. 1) between control and stress conditions was properly done. In addition, important differences between control and stressed seeds can also be observed (Fig. 4a). Moreover, the three biological replicates of a given sample (i.e. seeds of a given developmental stage and a given temperature condition) clustered together (Fig. 4a). Pooled quality control (QC) sample injections across LC-MS/MS run were used to evaluate the quality of the run and untargeted metabolomic dataset. The coefficient of variation across the QC samples was calculated for each metabolic feature detected (Fig. 4b and Table "Annotation and information of *Arabidopsis thaliana* seed specialized metabolome" [https://doi.org/10.57745/XSDTVE])[21]. More than 65% of the metabolic features of the dataset present a QC variation < 15% (Fig. 4b). QC variation of apigenin, which was used as internal standard, was of 6,122% and 12,546% for ESI+ and ESI− ionisation modes, respectively (Table 3). Moreover, apigenin variation across samples was also inferior to 20% (Table 3). Finally, the two-way analysis of variance (ANOVA) conducted allowed to identify a wide range of differentially accumulated metabolites (DAMs) impacted by the developmental stage, the growing condition (control vs heat stress treatment) or both factors (854 DAMs specific to seed developmental stage, 117 DAMs specific to heat stress and 420 DAMs affected by both seed developmental stage and heat stress). Hence, our data allowed the identification of a large number of SMs that were up or down–regulated by the heat stress, highlighting the strong impact of this stress on seed specialized metabolome (Fig. 4c).

Regarding transcriptomic data, the PCA analysis performed allowed to clearly separate the samples according to the seed developmental stage (Fig. 5a). Furthermore, the expression pattern of *LEC1*, peaking at Torpedo

sample_name	total_clean_read	total_mapping_gene_ratio	uniquely_mapping_gene_ratio
1-1_ctr	47.52	92.84	86.90
1-1_stress	45.04	92.98	86.96
1-2_ctr	47.44	92.61	86.65
1-2_stress	44.95	92.41	86.14
1-3_ctr	44.92	92.59	86.79
1-3_stress	44.90	92.82	86.71
2-1_ctr	44.83	92.13	85.12
2-1_stress	50.58	92.35	86.15
2-2_ctr	44.94	93.07	86.54
2-2_stress	44.86	93.13	86.64
2-3_ctr	44.83	90.35	83.70
2-3_stress	44.66	93.02	86.95
3-1_ctr	47.42	93.46	87.20
3-1_stress	46.92	93.14	84.01
3-2_ctr	47.50	93.21	88.00
3-2_stress	44.83	93.12	85.87
3-3_ctr	47.48	93.57	87.47
3-3_stress	50.19	93.10	85.96
4-1_ctr	44.66	93.29	84.09
4-1_stress	44.62	92.85	84.26
4-2_ctr	44.76	93.26	85.20
4-2_stress	49.69	90.26	81.54
4-3_ctr	44.46	90.83	81.98
4-3_stress	44.05	93.10	84.49
5-1_ctr	44.21	91.55	80.45
5-1_stress	45.07	93.41	81.58
5-2_ctr	43.98	91.72	78.59
5-2_stress	44.26	94.03	81.76
5-3_ctr	44.39	94.20	83.22
5-3_stress	44.54	94.43	81.90

Table 6. Transcriptomic validation: gene mapping summary.

stage for both control and warm temperature condition seeds is consistent with previously reported expression of *LEC1* (Fig. 5b)[38].

Those results indicate that the seed samples used for this whole study were properly synchronised. Moreover, as presented in the material and method section, RNA sequencing raw reads were filtered to keep reads with high quality scores. The clean reads ratios are presented in Table 4. Besides, each transcriptomic data sample specifically maps to more than 90% and 80% of respective *A. thaliana* genome and genes (Tables 5, 6). Finally, the multifactorial statistical analysis conducted allowed to identify a wide range of differentially expressed genes (DEGs) impacted by the developmental stage, the growing condition (control vs heat stress treatment) or both factors (15 438 DEGs specific to seed developmental stage, 316 DEGs specific to heat stress and 2 356 DEGs affected by both seed developmental stage and heat stress). These results indicated that our data are robust and allowed the identification of a large number of genes that were up or down–regulated by the heat stress, which drastically impact the transcriptome at several seed developmental stages (Fig. 5c).

Finally, integration of metabolomic and transcriptomic data was performed to identify correlations between the expressions of DEGs affected by heat stress and the accumulations of SMs (Fig. 6a). As an example, the expression of *MYB111* (*MYB PROTEIN DOMAIN 111*) gene correlates with the accumulation of different metabolic categories, including several flavonoids. In accordance with this result, *MYB111* transcription factor is known to regulate flavonol accumulation regulation in plants and seeds together with *MYB11* and *MYB12* transcription factors[41,42], highlighting therefore the strength and vigour of the performed multi-omic data integration analysis.

Usage Notes

These original data provide a valuable resource for further research into the metabolites and genes pathways that govern *A. thaliana* seed responses to heat stress during their development. In addition, it provides a large dataset to identify metabolites with a putative function in stress responses. Furthermore, information published in this work provide useful tools to identify gene targets for breeding, to improve seed quality and develop crops adapted to heat stress.

Specialized metabolome and transcriptome data obtained on seed developed under control or warm temperature may be used separately or combined under their integrated form to carry out comprehensive and robust analyses. DEGs-SMs correlation analysis results (correlation coefficient ≥ 0.7; p-value $\leq 0{,}0001$) are provide in the table "Metabolomic and transcriptomic data integration [https://doi.org/10.57745/XSDTVE]"[21].

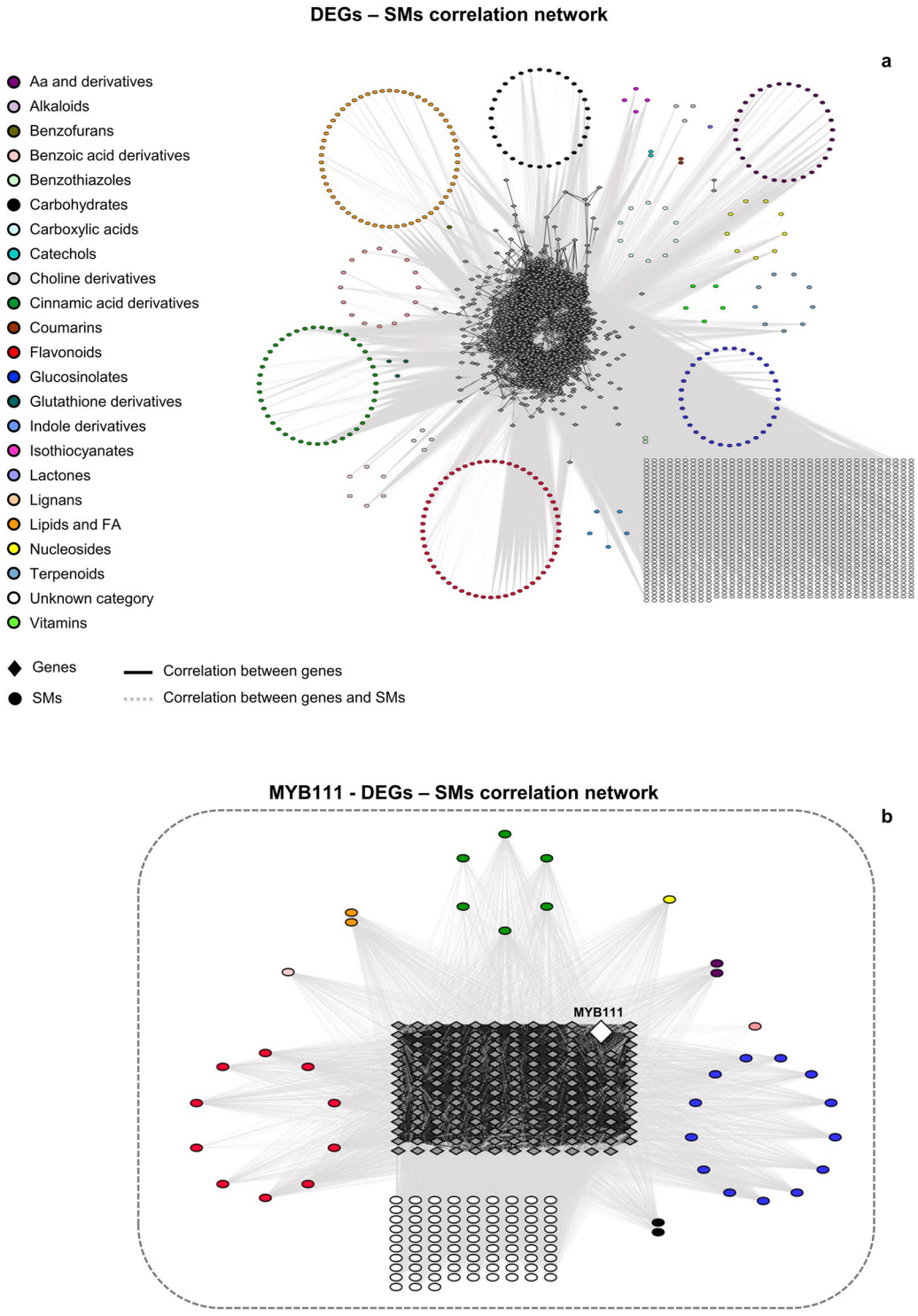

Fig. 6 Metabolomic and transcriptomic data integration. (**a**) Correlation network integration highlighting the interactions between differentially expressed genes (DEGs) and specialized metabolites (SMs). The correlation analysis was performed using the DEG expression and SM accumulation values of all samples (including the three biological replicates. (**b**) An example of correlation sub-network: DEGs and SMs correlating with *MYB111* flavonol transcription factor. Correlations between gene expressions are represented with black solid line edges whereas correlations between gene expressions and SM accumulations are represented with grey dashed line edges. The edges related to correlations between SM accumulations are not represented for network clarity, but corresponding correlation values are included in the supplementary table presenting related information.

Furthermore, the gene expression data from seed developed either under control or warm temperature may be combined with spatial transcriptomic data on specific seed tissues recorded during *A. thaliana* seed development and published by Belmonte *et al.*, Schneider *et al.* and Hofmann *et al.*[43–45]. The gene expression averages of

	Software version	Reference
Metabolomic data processing	Mzmine V2.53	Pluskal *et al*., *BMC Bioinformatics* vol. 11, p. 1–11 (2010)
Molecular network generation	MetGem V1.3.6	Olivon *et al. Anal. Chem*. vol. 90, p. 13900–13908[22]
Molecular network visualization	Cytoscape V3.10.1	Shannon *et al*., *Genome research* vol. 13, p. 2498–2504[23]
Metabolite annotation	Data analysis 4.4 (Bruker Daltonik GmbH)	
Metabolite annotation	Sirius V5.8.5	Dührkop *et al*. Nature Methods vol.16, p. 299–302[24]
Statistical analysis on metabolomic data	MetaboAnalyst V6.0	Pang *et al*., *Nucleic acids research* vol. 49, p. W388-W396[27]
Principal component analysis (PCA) and Heatmap generation	R Studio V2023.09.1+494	
qRT-PCR data treatment	Bio-Rad CFX Maestro V1.0	https://www.bio-rad.com/webroot/web/pdf/lsr/literature/10000068703.pdf
Transcriptomic data processing	HISAT	Kim *et al*., *Nat. Methods* vol. 12, p. 357–360[31]
Transcriptomic data processing	Ericscript V0.5.5-5	Benelli *et al*., *Bioinformatics* vol. 28, p. 3232–3239[32]
Transcriptomic data processing	rMATS V4.1.1	Shen *et al*., *PNAS* vol.111, p. E5593–E5601[33]
Transcriptomic data processing	Bowtie 2	Langmead *et al*., *Bried Commun*. vol. 9, p. 357–360[34]

Table 7. Software versions used to acquire and process data.

the biological replicate averages of the different seed developmental stages available in those published datasets are presented in the table "Seed tissue gene expression" [https://doi.org/10.57745/XSDTVE][21] .

Code availability

The different available software and the versions used to acquire and process data presented in the dataset are summarized in Table 7. The scripts for metabolomic data processing and DESeq2 multifactorial analysis of transcriptomic data have been deposited at the Data Gouv (Data INRAE) repository portal (https://doi.org/10.57745/XSDTVE)[21].

Received: 3 July 2024; Accepted: 30 January 2025;
Published online: 20 February 2025

References

1. Baud, S. *et al*. Recent progress in molecular genetics and omics-driven research in seed biology. *C R Biol* **354**, 61–110 (2023).
2. Harada, J. Seed biology. *J Integr Plant Biol* **61**, 530–532 (2019).
3. Sreenivasulu, N. & Wobus, U. Seed-development programs: A systems biology-based comparison between dicots and monocots. *Annu Rev Plant Biol* **64**, 189–217 (2013).
4. Pichersky, E. & Lewinsohn, E. Convergent evolution in plant specialized metabolism. *Annu Rev Plant Biol* **62**, 549–566 (2011).
5. Corso, M., Perreau, F., Mouille, G. & Lepiniec, L. Specialized phenolic compounds in seeds: structures, functions, and regulations. *Plant Science* **296**, 110471 (2020).
6. Abbas, F. *et al*. Volatile terpenoids: multiple functions, biosynthesis, modulation and manipulation by genetic engineering. *Planta* **246**, 803–816 (2017).
7. Barreda, L. *et al*. Specialized metabolite modifications in Brassicaceae seeds and plants: diversity, functions and related enzymes. *Nat Prod Rep* **41**, 834–859 (2024).
8. Corso, M. *et al*. *Chapter Two: Specialized Metabolites in Seeds*. *Advances in Botanical Research*. https://doi.org/10.1016/bs.abr.2020.11.001 (2021).
9. Hanif, K. *et al*. Biotechnological Approaches in Seed Quality Enhancement: Current Status and Future Prospects. *Trends in Animal and Plant Sciences* **1**, 37–48 (2023).
10. Verma, S., Attuluri, V. P. S. & Robert, H. S. Transcriptional control of Arabidopsis seed development. *Planta* **255**, 90 (2022).
11. Wu, M., Northen, T. R. & Ding, Y. Stressing the importance of plant specialized metabolites: omics-based approaches for discovering specialized metabolism in plant stress responses. *Front Plant Sci* **14**, 1272363 (2023).
12. Zhao, C. *et al*. Temperature increase reduces global yields of major crops in four independent estimates. *Proc Natl Acad Sci USA* **114**, 9326–9331 (2017).
13. Boutet, S. *et al*. Untargeted metabolomic analyses reveal the diversity and plasticity of the specialized metabolome in seeds of different Camelina sativa genotypes. *The Plant Journal* **110**, 147–165 (2022).
14. Mácová, K. *et al*. Long-Term High-Temperature Stress Impacts on Embryo and Seed Development in Brassica napus. *Front Plant Sci* **13**, 844292 (2022).
15. Nadakuduti, S. S. *et al*. Heat stress during seed development leads to impaired physiological function and plasticity in seed oil accumulation in Camelina sativa. *Front Plant Sci* **14**, 1284573 (2023).
16. Debeaujon, I. *et al*. Proanthocyanidin-Accumulating Cells in Arabidopsis Testa: Regulation of Differentiation and Role in Seed Development. *Plant Cell* **15**, 2514–2531 (2003).
17. Pfaffl, M. W. A new mathematical model for relative quantification in real-time RT–PCR. *Nucleic Acids Res* **29**, e45–e45 (2001).
18. Vandesompele, J. *et al*. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* **3**, 1–7 (2002).
19. Chambers, M. C., MacLean, B. & Burke, R. A Cross-platform Toolkit for Mass Spectrometry and ProteomicsZwane's mining charter lunacy. *Nat. Biotechnology* **30**, 1 (2012).
20. Myers, O. D., Sumner, S. J., Li, S., Barnes, S. & Du, X. One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks. *Anal Chem* **89**, 8696–8703 (2017).
21. Barreda, L. *et al*. Processed metabolomic (LC-MS/MS) and transcriptomic (RNAseq) data of Arabidopsis thaliana developing seeds under control and warm temperatures. *Data Gouv* https://doi.org/10.57745/XSDTVE (2025).
22. Olivon, F. *et al*. MetGem Software for the Generation of Molecular Networks Based on the t - SNE Algorithm. *Anal Chem* **90**, 13900–13908 (2018).
23. Shannon, P. *et al*. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res* **13**, 2498–2504 (2003).

24. Dührkop, K. *et al*. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat Methods* **16**, 299–302 (2019).
25. Alseekh, S. *et al*. Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. *Nat Methods* **18**, 747–756 (2021).
26. Sumner, L. W. *et al*. Proposed minimum reporting standards for chemical analysis: Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* **3**, 211–221 (2007).
27. Pang, Z. *et al*. MetaboAnalyst 5.0: Narrowing the gap between raw spectra and functional insights. *Nucleic Acids Res* **49**, W388–W396 (2021).
28. Pang, Z. *et al*. MetaboAnalyst 6.0: towards a unified platform for metabolomics data processing, analysis and interpretation. *Nucleic Acids Res* **52**, 398–406 (2024).
29. Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. & Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res* **38**, 1767–1771 (2009).
30. Chen, Y. *et al*. SOAPnuke: A MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *Gigascience* **7**, 1–6 (2018).
31. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357–360 (2015).
32. Benelli, M. *et al*. Sequence analysis Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript. *Bioinformatics* **28**, 3232–3239 (2012).
33. Shen, S. *et al*. rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *PNAS* **111**, E5593–E5601 (2014).
34. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Bried communications* **9**, 357–360 (2012).
35. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 1–21 (2014).
36. Barreda, L., Boutet, S. & Corso, M. Specialized metabolome from developing *Arabidopsis thaliana* seeds under warm temperatures. *MassIVE* https://doi.org/10.25345/C5S17T468 (2025).
37. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP515207 (2025).
38. Boulard, C., Fatihi, A., Lepiniec, L. & Dubreucq, B. Regulation and evolution of the interaction of the seed B3 transcription factors with NF-Y subunits. *Biochim Biophys Acta Gene Regul Mech* **1860**, 1069–1078 (2017).
39. Jo, L., Pelletier, J. M. & Harada, J. J. Central role of the LEAFY COTYLEDON1 transcription factor in seed development. *J Integr Plant Biol* **61**, 564–580 (2019).
40. Chen, M. *et al*. TRANSPARENT TESTA GLABRA1 regulates the accumulation of seed storage reserves in Arabidopsis. *Plant Physiol* **169**, 391–402 (2015).
41. Stracke, R. *et al*. Differential regulation of closely related R2R3-MYB transcription factors controls flavonol accumulation in different parts of the Arabidopsis thaliana seedling. *Plant Journal* **50**, 660–677 (2007).
42. Stracke, R. *et al*. Analysis of production of flavonol glycosides-dependent flavonol glycoside accumulation in Arabidopsis thaliana plants reveals MYB11-, MYB12- and MYB111-independent flavonol glycoside accumulation. *New Phytologist* **188**, 985–1000 (2010).
43. Schneider, A. *et al*. Potential targets of VIVIPAROUS1/ABI3-LIKE1 (VAL1) repression in developing Arabidopsis thaliana embryos. *The Plant Journal* **85**, 305–319 (2016).
44. Belmonte, M. F. *et al*. Comprehensive developmental profiles of gene activity in regions and subregions of the Arabidopsis seed. *Proceedings of the National Academy of Sciences* **110**, E435–E444 (2013).
45. Hofmann, F., Schon, M. A. & Nodine, M. D. The embryonic transcriptome of Arabidopsis thaliana. *Plant Reprod* **32**, 77–91 (2019).

Acknowledgements

The authors thank Anne-Solenn Valadon and Chandrodhay Saccaram (Institute Jean-Pierre Bourgin for Plant Sciences, FR) for their help and advice for PCA and heatmap analysis. They also thank François Perreau (Institute Jean-Pierre Bourgin for Plant Sciences, FR) for his advice on specialized metabolite annotation. They also thank Gwendal Cueff (Institute Jean-Pierre Bourgin for Plant Sciences; new address: UNH, Université de Clermont Auvergne, FR) and Gabrièle Adam (Institute Jean-Pierre Bourgin for Plant Sciences; new address: IPS2, Université Paris-Saclay, FR) for the development, help and advice about the R script for metabolomic data processing. The PhD scholarship of L.B. is funded by the 'Plant biology and breeding' (BAP) department of INRAE ("Appel à projets scientifiques BAP 2021" to M.C.) and by the project 'Seed MetSpe' of the Labex Saclay Plant Sciences-SPS (ANR-10-LABX-0040-SPS to L.L. and M.C.). This work was supported by the 'Plant biology and breeding' (BAP) of INRAE (grant "Appel à projets scientifiques BAP 2020" to M.C.), by the project 'Seed MetSpe' of the Labex Saclay Plant Sciences-SPS (ANR-10-LABX-0040-SPS to L.L. and M.C.), by a project financed by Plant2Pro Carnot institute ('BrassiMet' to M.C.) and by the project "MetAnnotation" of the Labex Saclay Plant Sciences-SPS ("SPS Bio-analyses") to M.C. The IJPB benefits from the support of Saclay Plant Sciences-SPS (ANR-17-EUR-0007). This work has benefited from the support of IJPB's Plant Observatory platform P0-Chemistry/Metabolism.

Author contributions

M.C. directed the research. M.C. designed the research, with the help of L.L. and L.B. L.B., D.D.V., D.G., C.B. and M.C. grown the plants and harvested seeds. L.B. and D.D.V. performed microscopic and gene expression analysis for seed developmental stage synchronisation. L.B. performed metabolite and RNA extractions from seeds. S.B. and L.B. generated the LC-MS/MS data and performed the initial quality analyses on the metabolomic data. L.B. and S.B. performed the metabolites annotation with MetGem, SIRIUS and using MS/MS spectra. L.B. and M.C. analysed and interpreted the untargeted metabolomic and transcriptomic data. L.B. generated all Fig.s and supplemental material. L.B. and M.C. wrote the paper, which was edited by L.L. All the authors commented on and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.C.