



**HAL**  
open science

## RNA sequencing data for heat stress response in isolated *Medicago truncatula* seed tissues

Zhijuan Chen, Benoit Ly Vu, Olivier Leprince, Jerome Verdier

### ► To cite this version:

Zhijuan Chen, Benoit Ly Vu, Olivier Leprince, Jerome Verdier. RNA sequencing data for heat stress response in isolated *Medicago truncatula* seed tissues. *Data in Brief*, 2021, pp.106726. 10.1016/j.dib.2021.106726 . hal-03123816

**HAL Id: hal-03123816**

**<https://hal.inrae.fr/hal-03123816>**

Submitted on 13 Feb 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

## Article Title

**RNA sequencing data for heat stress response in isolated *Medicago truncatula* seed tissues**

## Authors

Zhijuan Chen, Benoit Ly Vu, Olivier Leprince, Jerome Verdier

## Affiliations

Institut de Recherche en Horticulture et Semences-UMR1345, Université d'Angers, INRAE, Institut Agro, SFR 4207 QuaSaV, 49071, Beaucouzé, France

## Corresponding author(s)

Jerome Verdier - [jerome.verdier@inrae.fr](mailto:jerome.verdier@inrae.fr)

## Abstract

Legumes are important crop species as they produce highly nutritious seeds for human food and animal feed. In grain legumes, sub-optimal conditions affect seed developmental timing leading to impairment of seed quality traits acquired during seed maturation. To understand the molecular mechanisms of heat stress response in legume seeds, we analysed transcriptome changes of three seed tissues (i.e. embryo, endosperm and seed coat) at four developmental stages, during seed maturation, from seed filling to mature dry seeds, collected under optimal and heat stress conditions in the model legume, *Medicago truncatula* (reference genotype A17). The total RNA sequencing generated a dataset of 48 samples, representing more than 57 Gb fastq raw data. Mapping, quantification and annotation of the data were based on fifth release of *Medicago truncatula* genome and provided expression profiles of 44,473 transcripts in seed tissues at different developmental stages and under optimal and stress conditions. Time-course and pairwise comparisons between optimal and stress conditions showed that 9182, 8315 and 3481 genes were differentially expressed due to heat stress in embryo, endosperm and seed coat respectively. Moreover, it highlighted a common set of 975 genes that were differentially expressed in all the seed tissues.

## Keywords

RNA sequencing; Seed maturation; Heat stress; Seed quality; Embryo; Endosperm; Seed coat

## Specifications Table

<b>Subject</b>	Agricultural and Biological Sciences
<b>Specific subject area</b>	Omics: Transcriptomics Plant Science: Plant Physiology
<b>Type of data</b>	Tables Figures
<b>How data were acquired</b>	<b>Total RNA</b> samples were sent to BGI, Hong Kong, for library preparation. Libraries were constructed following a custom protocol from samples that passed quality controls (mass > 2µg, concentration>80ng/microl, OD260/280 ~ = 2.00, OD260/230 ~ = 2.20, RIN>6.5, 28S/18S<1.0, baseline smooth). After mRNA enrichment by <b>rRNA depletion and oligo dT selection</b> , RNA was fragmented and reverse transcribed to double-strand cDNA (dscDNA) by N6 random primer. The synthesized cDNA was subjected to end-repair and then was 3' adenylated. Adaptors were ligated to the ends of these 3' adenylated cDNA fragments. The ligation products were purified and many rounds of PCR amplification were performed to enrich the purified cDNA template using PCR primer, splint oligo and DNA ligase, followed by sequencing on BGISEQ-500 platform, generating an average 24M reads of 50bp per sample.
<b>Data format</b>	Filtered raw reads (FASTQ) Analysed RNA-seq data files (counts and DEG lists)
<b>Parameters for data collection</b>	Total RNAs were extracted from isolated embryo, endosperm and seed coat of four developmental stages of <i>Medicago truncatula</i> (A17) seeds that were harvested during maturation phase before and after acquisition of desiccation tolerance (respectively S1 and S2) and at the onset and after longevity acquisition (respectively S3 and S4) under standard temperature (20°C day/ 18°C night) and under high temperature (26°C/24°C).

<b>Description of data collection</b>	RNA-seq dataset was collected from single-end sequencing of cDNA libraries using BGISEq500 platform with 50bp reads. Raw reads were filtered to remove adapters and low-quality reads, then mapped to <i>Medicago truncatula</i> reference transcriptome (version 5). Total mapped reads and number of transcripts (counts and TPM) were estimated using Salmon algorithm. Differential expressions of genes between standard and heat stress conditions were calculated using ImpulseDE2 and DEseq2 algorithms.
<b>Data source location</b>	Institution: Growth chambers from the Institut de Recherche en Horticulture et Semences, INRAE City: Beaucauzé Country: France
<b>Data accessibility</b>	<u>Public Repository:</u>  Repository name: NCBI GEO Data identification number: GSE160725 Direct URL to data: <a href="https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE160725">https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE160725</a>

### Value of the Data

- These data represent valuable seed transcriptome dataset of heat stress in the model legume *Medicago truncatula* because it has been generated from isolated seed tissues (embryo, endosperm and seed coat) along the whole seed maturation.
- These data are useful resources for scientific communities working on seed and legume seed quality but also on plant stress biology to understand specific and common stress response pathways.
- These data provide new insights about molecular processes affected during heat stress in different seed tissues and candidate genes and molecular markers to predict seed quality in sub-optimal conditions.

### Data Description

This manuscript presents a transcriptomic dataset obtained from dissected seed tissues (embryo, endosperm and seed coat) at four developmental stages of *Medicago truncatula* (A17) seeds during maturation phase [i.e. before and after desiccation tolerance (respectively S1 and S2) and at the onset and after longevity acquisition (respectively S3 and S4)] produced under standard temperature (20°C day/ 18°C night) and heat stress (26°C/24°C). **Table 1** displays all sample names with information regarding seed tissues, treatments, developmental stages and numbers of 50bp reads sequenced per sample. **Figure S1** provides sequence quality histograms obtained from FastQC [1] and merged into a single graphic using MultiQC [2]. All 48 samples displayed Phred quality scores about 35, which corresponds to a base calling accuracy of about 99.95%. Salmon algorithm [3] was used to map and quantify raw reads to the reference *Medicago truncatula* transcriptome version 5 and corresponding count table is provided as **Table S1**. After mapping and quantification, these 48 samples were, then, hierarchically clustered based on their dissimilarity scores to validate reproducibility of replicates (Figure 1). From this count table, differentially expressed genes were identified using time-course comparisons of the four developmental stages for embryo (**Table S2**) and endosperm (**Table S3**); and using pairwise comparisons of the two developmental stages for seed coat (**Table S4**). **Figure 2** summarizes the differentially expressed genes (DEGs) in the three seed tissues between optimal and heat stress conditions (False Discovery Rate, FDR<1%) and highlights a common set of 975 DEGs within the three seed tissues.

## Experimental Design, Materials and Methods

### 1. Plant growth conditions and seed tissue sampling

*Medicago* plants were grown under standard conditions (20°C/18°C, 16h photoperiod) in growth chamber. At flowering time, half of plants were kept at same optimal conditions (20°C/18°C, 16h photoperiod) and half were grown under heat stress condition (26°C/24°C, 16h photoperiod). Developing and mature seeds were collected from standard and stress conditions, seed tissues were quickly dissected then immediately frozen in liquid nitrogen before RNA extraction. According to our previous study [4], four seed developmental stages were collected under standard conditions at 17 days after pollination (DAP), 26 DAP, 36DAP and 44 DAP and under heat stress conditions at corresponding developmental stages at 14 DAP, 17 DAP, 22 DAP and 28 DAP. These four developmental stages were shown to be the beginning of seed maturation and before acquisition of desiccation tolerance (S1), after the acquisition of desiccation tolerance and during seed filling (S2), at the onset of the acquisition of longevity and at the end of seed filling (S3) and mature seed (S4) according to our previous study [4].

### 2. RNA isolation and sequencing

Total RNAs were extracted from embryo and endosperm that were collected at four stages during seed maturation under standard and heat stress conditions. RNAs were only extracted from S1 and S2 for seed coat, as during late maturation stage seed coat is dying and does not provide good quality RNA. All RNA extractions were performed using MACHEREY-NAGEL NucleoSpin® RNA Plus kit. RNA quality was

checked using a nanodrop spectrophotometer ND-1000 (NanoDrop Technologies) and a 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). All samples with good qualities (260/280 and 260/230 absorbance ratio >2; RNA Integrity Number, RIN>8; 28S/18S>1.7) were sent to Beijing Genomics Institute (<https://www.bgi.com>) (Hong Kong) for library preparation and sequencing on BGISEQ-500 platform, generating an average of 20M reads of 50bp per sample.

### 3. RNA-seq data analyses

After quality control of fastq files using FastQC [1], high-quality reads were mapped onto the reference Medicago A17 transcriptome version 5 (<https://medicago.toulouse.inra.fr/MtrunA17r5.0-ANR/downloads/1.7/MtrunA17r5.0-ANR-EGN-r1.7.fastaFiles.zip>) (Table S5) [5] and transcript abundances were quantified with Salmon algorithm (version 0.14.1) [3] using the quasi-mapping mode and the '--validateMappings', '--useVBOpt' and '--seqBias' options. Reproducibility of replicates was validated by a hierarchical cluster analysis based on dissimilarity scores obtained from the normalized raw count values of the 48 samples using the 'cpm', 'dis' and 'hclust' functions in R. Differentially expressed genes (DEGs) were identified by time-course comparisons of the four developmental stages of embryo and endosperm using ImpulseDE2 [6] and by pair-wise comparisons of the two developmental stages of seed coat using DESeq2 [7]. All transcripts with change in expression showing adjusted p-values below 1% (no fold change cutoff) were considered as differentially expressed between standard and heat stress conditions and displayed on a Venn Diagram performed from the website: <http://bioinformatics.psb.ugent.be/webtools/Venn/>.

### **Ethics Statement**

This work does not contain any studies with human or animal subjects.

### **CRedit author statement**

Zhijuan Chen: Investigation, Methodology, Visualization, Writing- Original draft preparation. Benoit Ly Vu: Methodology. Olivier Leprince: Conceptualization, Reviewing and Editing. Jerome Verdier: Supervision, Conceptualization, Data curation, Writing- Reviewing and Editing.

### **Acknowledgments**

This research was conducted in the framework of the regional programme "Objectif Végétal, Research, Education and innovation in Pays de la Loire", supported by the French Region Pays de la Loire, Angers Loire Métropole.

### **Declaration of Competing Interest**

The authors declare that they have no known competing financial interests or personal relationships which have or could be perceived to have influenced the work reported in this article.

### **Supplementary materials**

Table S1: Count table of mapped reads from the 48 samples using Salmon algorithm.

Table S2: List of differentially expressed genes in Embryo using time-course comparisons of the four developmental stages obtained from ImpulseDE2 algorithm.

Table S3: List of differentially expressed genes in Endosperm using time-course comparisons of the four developmental stages obtained from ImpulseDE2 algorithm.

Table S4: List of differentially expressed genes in Seed Coat using pairwise comparisons of the two developmental stages obtained from DESeq2 algorithm.

Table S5: Reference transcriptome, from *Medicago truncatula* genome version 5, used in this study and available at <https://medicago.toulouse.inra.fr/MtrunA17r5.0-ANR/downloads/1.7/MtrunA17r5.0-ANR-EGN-r1.7.fastaFiles.zip>

Figure S1: Overview of the Phred quality values across all bases at each position in the fastq files from the 48 samples obtained from FastQC and MultiQC analyses.

## References

- [1] S. Andrews, FastQC - A quality control tool for high throughput sequence data. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>, Babraham Bioinforma. (2010).
- [2] P. Ewels, M. Magnusson, S. Lundin, M. Källér, MultiQC: Summarize analysis results for multiple tools and samples in a single report, Bioinformatics. (2016). <https://doi.org/10.1093/bioinformatics/btw354>.
- [3] R. Patro, G. Duggal, M.I. Love, R.A. Irizarry, C. Kingsford, Salmon provides fast and bias-aware quantification of transcript expression, Nat. Methods. (2017). <https://doi.org/10.1038/nmeth.4197>.
- [4] J. Verdier, D. Lalanne, S. Pelletier, I. Torres-Jerez, K. Righetti, K. Bandyopadhyay, O. Leprince, E. Chatelain, B.L. Vu, J. Gouzy, P. Gamas, M.K. Udvardi, J. Buitink, A Regulatory Network-Based Approach Dissects Late Maturation Processes Related to the Acquisition of Desiccation Tolerance and Longevity of *Medicago truncatula* Seeds, PLANT Physiol. 163 (2013) 757–774. <https://doi.org/10.1104/pp.113.222380>.
- [5] Y. Pecrix, S.E. Staton, E. Sallet, C. Lelandais-Brière, S. Moreau, S. Carrère, T. Blein, M.-F. Jardinaud, D. Latrasse, M. Zouine, M. Zahm, J. Kreplak, B. Mayjonade, C. Satgé, M. Perez, S. Cauet, W. Marande, C. Chantry-Darmon, C. Lopez-Roques, O. Bouchez, A. Bérard, F. Debelle, S. Muños, A. Bendahmane, H. Bergès, A. Niebel, J. Buitink, F. Frugier, M. Benhamed, M. Crespi, J. Gouzy, P. Gamas, Whole-genome landscape of *Medicago truncatula* symbiotic genes, Nat. Plants. 4 (2018) 1017–1025. <https://doi.org/10.1038/s41477-018-0286-7>.
- [6] D.S. Fischer, F.J. Theis, N. Yosef, Impulse model-based differential expression analysis of time course sequencing data, Nucleic Acids Res. (2018). <https://doi.org/10.1093/nar/gky675>.
- [7] M.I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biol. 15 (2014) 550. <https://doi.org/10.1186/s13059-014-0550-8>.

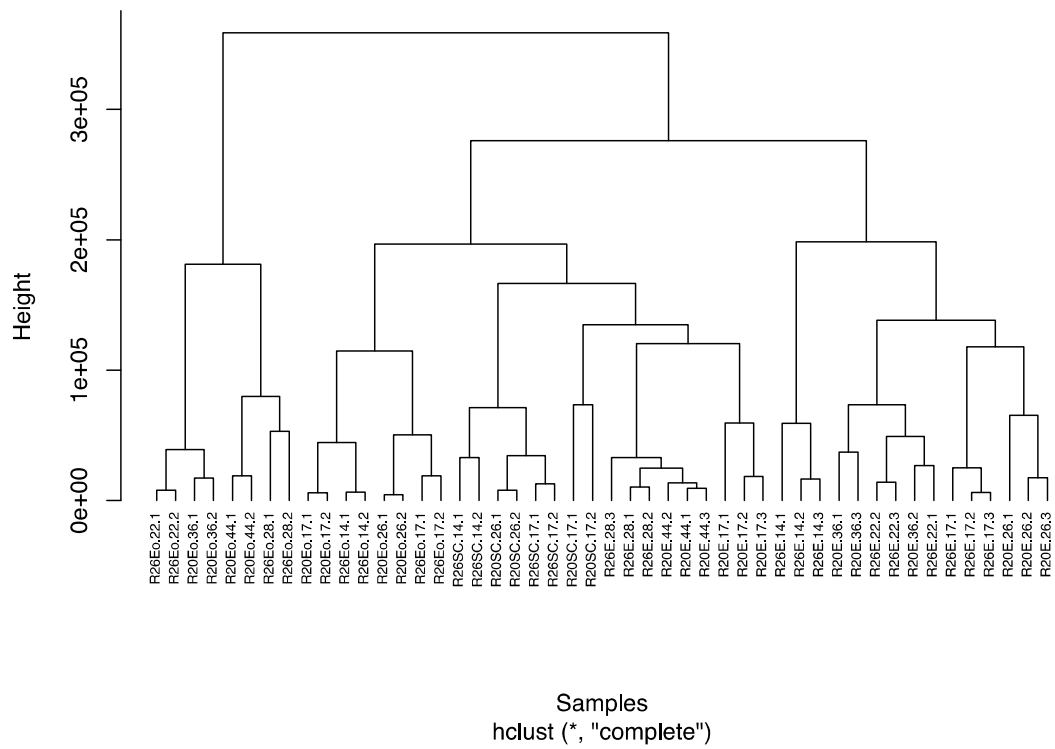
## Figures & Tables

**Table 1.** Summary of sample files with corresponding information related to seed tissues, seed developmental stages, growth conditions and numbers of cleaned 50bp reads used for RNA-seq mapping.

Tissues	Sample	Numbers of 50bp reads	Sample	Numbers of 50bp reads	Stages
Embryo	R20E-17-1	20,452,270	R26E-14-1	20,822,668	S1
Embryo	R20E-17-2	20,465,064	R26E-14-2	20,786,562	S1
Embryo	R20E-17-3	20,364,545	R26E-14-3	20,718,145	S1
Embryo	R20E-26-1	20,452,569	R26E-17-1	20,666,800	S2
Embryo	R20E-26-2	20,137,018	R26E-17-2	20,704,789	S2
Embryo	R20E-26-3	21,967,973	R26E-17-3	20,753,459	S2
Embryo	R20E-36-1	22,003,911	R26E-22-1	20,717,753	S3
Embryo	R20E-36-2	22,156,003	R26E-22-2	20,786,906	S3
Embryo	R20E-36-3	20,183,128	R26E-22-3	20,644,844	S3
Embryo	R20E-44-1	20,760,225	R26E-28-1	20,649,142	S4
Embryo	R20E-44-2	20,897,210	R26E-28-2	20,749,797	S4
Embryo	R20E-44-3	20,744,207	R26E-28-3	20,778,364	S4
Embryo	R20Eo-17-1	20,677,185	R26Eo-14-1	20,303,572	S1
Endosperm	R20Eo-17-2	20,659,667	R26Eo-14-2	20,257,351	S1
Endosperm	R20Eo-26-1	20,815,605	R26Eo-17-1	20,342,065	S2
Endosperm	R20Eo-26-2	20,174,987	R26Eo-17-2	20,221,552	S2
Endosperm	R20Eo-36-1	20,332,236	R26Eo-22-1	20,288,004	S3
Endosperm	R20Eo-36-2	20,129,499	R26Eo-22-2	20,463,956	S3
Endosperm	R20Eo-44-1	20,829,827	R26Eo-28-1	20,371,220	S4
Endosperm	R20Eo-44-2	20,810,635	R26Eo-28-2	20,405,449	S4
Seed Coat	R20SC-17-1	40,072,723	R26SC-14-1	41,944,942	S1
Seed Coat	R20SC-17-2	56,846,997	R26SC-14-2	42,596,392	S1
Seed Coat	R20SC-26-1	39,483,027	R26SC-17-1	22,920,874	S2
Seed Coat	R20SC-26-2	48,482,527	R26SC-17-2	34,721,666	S2
	CONTROL CONDITION		STRESS CONDITION		

**Figure 1.** Hierarchical cluster analysis based on dissimilarity scores obtained from the TPM values of the 48 samples to validate reproducibility of replicates.





**Figure 2.** Venn diagram with differentially expressed genes (DEGs, adjusted p-values below 1%) in the three seed tissues between optimal and heat stress conditions.

