OCL 2020, 27, 36
© C. Penouilh-Suzette *et al.*, Hosted by EDP Sciences, 2020
https://doi.org/10.1051/ocl/2020027

OCL Oilseeds & fats Crops and Lipids

Available online at:
www.ocl-journal.org

Data Paper Open Access

# RNA expression dataset of 384 sunflower hybrids in field condition

Charlotte Penouilh-Suzette[1,a], Lise Pomiès[2,a], Harold Duruflé[1,a], Nicolas Blanchet[1], Fanny Bonnafous[1], Romain Dinis[1], Céline Brouard[2], Louise Gody[1], Christopher Grassa[1], Xavier Heudelot[3], Marion Laporte[4], Marion Larroque[1], Gwenola Marage[1], Baptiste Mayjonade[1], Brigitte Mangin[1], Simon de Givry[2] and Nicolas B. Langlade[1,*]

[1] LIPM, Université de Toulouse, INRAE, CNRS, Castanet-Tolosan, France
[2] MIAT, Université de Toulouse, INRAE, Castanet-Tolosan, France
[3] Innolea, Domaine de Sandreau, Mondonville, 31700 Blagnac, France
[4] RAGT 2n,BP 3336, 12033 Rodez, France

Received 11 March 2020 – Accepted 5 June 2020

**Abstract –** This article describes how RNA expression data of 173 genes were produced on 384 sunflower hybrids grown in field conditions. Sunflower hybrids were selected to represent genetic diversity within cultivated sunflower. The RNA was extracted from mature leaves at one time seven days after anthesis. These data allow to differentiate the different genotype behaviours and constitute a valuable resource to the community to study the adaptation of crops to field conditions and the molecular basis of heterosis. It is available on data.inra.fr repository.

**Keywords:** sunflower / genetics / gene expression / drought

**Résumé – Données d'expression d'ARN issues de 384 hybrides de tournesol cultivés en champ.** Cet article décrit la production des niveaux d'expression de 173 gènes dans 384 hybrides de tournesol cultivés en conditions de champ. Les hybrides sont issus de parents choisis pour représenter la diversité génétique dans le tournesol cultivé. Les ARN ont été extraits à partir de feuilles matures environ sept jours après la floraison. Ces données permettent de différencier les comportements des différents génotypes et constituent une ressource importante pour les chercheurs intéressés dans l'adaptation des espèces cultivées aux conditions agronomiques et aux bases moléculaires de l'hétérosis. Elles sont disponibles sur le portail Data INRAE : data.inra.fr.

**Mots clés :** tournesol / génétique / génomique / sécheresse

## 1 Data

Domesticated sunflower, *Helianthus annuus L.*, is the fourth most important oilseed crop in the world (USDA, 2019). It can maintain stable yields across a wide range of environmental conditions, especially during drought stress (Hussain *et al.*, 2018). In the context of climate change, a major interest in crop science is to better understand the adaptation of those plants to this phenomenon. Response to drought stress involves a large number of molecular pathways and subsequent physiological processes. Cultivated sunflowers in the world are mostly hybrid genotypes produced using the heterosis phenomenon to improve the yield and the stress tolerance of plants in field. In this data article, we are sharing the RNA expression profile of 173 genes (linked to drought stress and heterosis) on 384 hybrid genotypes of sunflower grown in fields. These datasets are part of a larger project that integrates other phenotypic data collected on this trial. They include the development of harvested plants and of course agronomic and yield traits at the plot level. They are available upon request from the authors according intellectual property limitations. The raw data associated with this article can be found at https://doi.org/10.15454/HESVA0.

## 2 Experimental design, plant material and growth conditions

The experiment (code 15EX05) was performed in 2015 from May 2nd to September 29th, in the Anais

*Correspondence: nicolas.langlade@inrae.fr
[a] Co-first authors.

(Charente-Maritime, France) field station. The field station is divided into 540 plots including 475 plots with hybrids obtained by crossing 36 sterile (CMS PET1) to 36 restorer lines as previously described (Bonnafous *et al.*, 2018). The other plots (around 11% of the field) contained one of the four check-hybrids (Extrasol, ES Akustic, NK Kondi and LG5450HO). These four hybrids are used as environmental control. The harvest was performed on July 22nd 2015 between 11:00 and 12:30 local time (sunrise at 6:36 and sun zenith at 14:00). This harvest was performed seven days after anthesis as the studied hybrids flowered on July 15th on average ±3 days (standard deviation). On four plants per plot, we collected one leaf per plant positioned at $n-4$ rank (with $n$ the total number of leaves detached of the flower head) for this molecular analysis. Leaves were cut without their petiole and immediately frozen in dry ice and then, transferred to −80 °C freezer before grinding.

Field arrangement (file ..MAP.pdf), technical management (file ..ITK.pdf), meteorological (file ..METEO.xls) and soil humidity (files ..SOIL.pdf and ..METEO.xls) data can be found on the SUNRISE Phenotype Archive at the following link https://sunrise-archive.toulouse.inra.fr/ws/phenotype/docid/692a2e01e7d2fe0c51d1ae817b49bf8f.html.

## 3 Transcriptome analysis

### 3.1 RNA extraction

Leaf grinding was performed using a ZM200 grinder (Retsch, Haan, Germany) with a 0.5-mm sieve cooled with liquid nitrogen. Total RNA was extracted with the Nucleospin 96 RNA Tissue Core Kit (Macherey-Nagel, Düren, Germany) following the manufacturer's instructions. RNA quality controls were done with a Nanodrop Lite Spectrophotometer (Thermo Scientific, Waltham, USA) and the quantity was assessed using the Agilent RNA 6000 Nano Kit (Agilent, Santa Clara, CA, USA). DNAse treatment was performed with the TURBO DNAfree™ Kit from Invitrogen (Thermo Fisher Scientific, Vilnius, Lithuania) and RNA was finally subjected to reverse transcription using the Transcriptor Reverse Transcriptase Kit from Roche (Mannheim, Germany).

### 3.2 Primers design

Specific primer pairs for qPCR were designed according to the sunflower genome (Badouin *et al.*, 2017) using Primer3-web software (v.0.4.0) (Untergasser *et al.*, 2012). The following parameters were used:

- product size between 50 and 100 bp;
- primer size between 18–22 bp;
- primer Tm between 58–62 °C.

The specificity of each pair of primers was verified by BLAST on the whole sunflower genome and tested with the LightCycler 480 Real-Time PCR Instrument (Roche Diagnostics, Mannheim, Germany) on a pool of cDNA representative of the genetic diversity of our samples. Finally, the primers were selected according to their amplification and melting curves, their cycle detection thresholds, and after verification of their specificity (absence of dimer).

### 3.3 Expression measurements

qPCR analyses were conducted with the BioMark HD System using 96.96 dynamic arrays chips (Fluidigm Corporation, San Francisco, CA, USA) (Spurgeon *et al.*, 2008). Ten different Fluidigm plates were designed to measure a total of 180 genes against 435 experimental samples (including 353 hybrid genotypes and check hybrids). Plates are described in Figure 1.

For the sample part, plates were separated into five groups of two plates. Samples are the same on the two plates of the same group. Each plate always contained five points of dilution of pool representative of the genetic diversity of our samples; two genotype controls (LG5450HO and SF092_SF342); one point of water; one intern control of the Fluidigm Biomark HD platform and 87 genotype samples. For the 87 genotype samples, 11% of them correspond to field control genotypes (corresponding to the proportion of environmental control genotypes in the field). All genotype samples are repeated on the two different plates of the group. In total, 435 genotype samples were used, including 82 environment checks and 353 hybrids genotypes.

For the gene part, two batches of five plates were constituted. For the first batch, each plate contains two reference genes (HanXRQChr05g0131911 and HanXRQChr01g0029571) and three biomarker genes (HanXRQChr01g0021351, HanXRQChr06g0175391 and HanXRQChr04g0120731) as in Marchand *et al.* (2013). On the second batch, the two previous reference genes are present with three other reference genes (HanXRQChr04g0115631, HanXRQChr03g0090171 and HanXRQChr01g0021131). In both batches, an intern control, specific of the platform was added. Sunflower reference genes were chosen among the genes presenting no modulation under a range of drought stress intensities in eight genotypes obtained from the Affymetrix hybridizations performed in Rengel *et al.* (2012).

Results of qRT-PCR are analyzed following the $2-\Delta\Delta Ct$ method described by Livak and Schmittgen (2001). Threshold cycles ($Ct$) (number of cycles performed when the signal reaches the detection threshold) were calculated using Fluidigm Biomark software with the following parameters: quality threshold = 0.65, baseline correction = linear (derivative), $Ct$ threshold method = auto detector, peak sensitivity = 10, peak ratio threshold = 0.8. In the end, only four reaction chambers (of 92160) were empty due to defective valves on the Fluidigm plates:

- batch 1, plate 3, hybrid SF160_SF292 *versus* gene HanXRQChr05g0137501;
- batch 1, plate 3, hybrid SF160_SF292 *versus* gene HanXRQChr05g0131121;
- batch 1, plate 3, hybrid CSF2_SF278 *versus* gene HanXRQChr11g0336031;
- batch 2, plate 1, sample pool dilution 0,25 *versus* gene HanXRQChr12g0369571.

## 4 Data curation

Variation of $Ct$ of each gene measured on the different hybrid genotypes can be due to environmental and experimen-

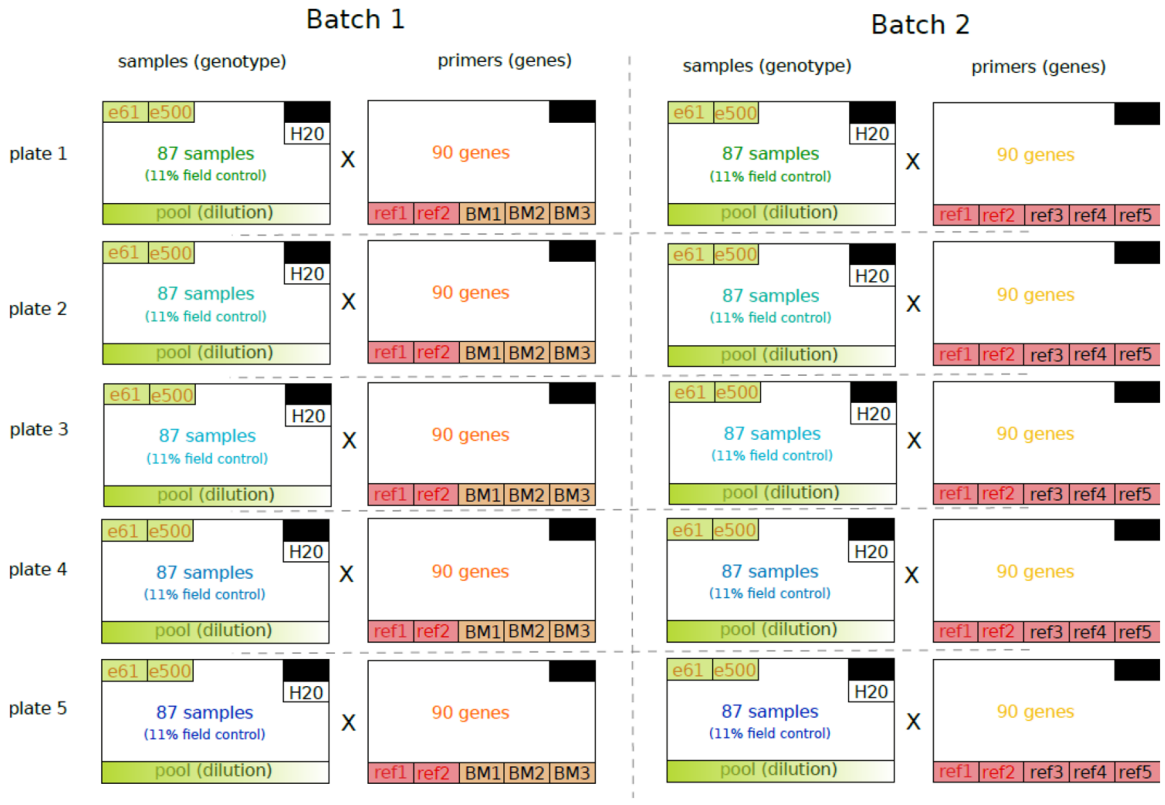


**Fig. 1.** Design of the 10 Fluidigm plates. The 10 plates are separated in two batches of five plates. Measured genes are the same between plates of the same batch. Samples (genotypes) are different between the five plates of a batch but are the same between the two batches.

**Table 1.** MSE of the reference genes.

| Gene ID | MSE |
|---|---|
| HanXRQChr05g0131911 | 2.7223 |
| HanXRQChr01g0029581 | 1.6324 |
| HanXRQChr04g0115631 | 13.9149 |
| HanXRQChr03g0090171 | 2.9755 |
| HanXRQChr01g0021131 | 3.1769 |

tal perturbations. Consequently, normalization and correction of *Ct* values are necessary.

### 4.1 Amplification efficiency

The amplification efficiency of each gene was estimated from its values on the range of dilution, using the robustfit function in Matlab (version 7.11) Statistics Toolbox (version 7.4), as follows:

$$X_t^i = X_0^i \times (1 + Eff_i)^{Ct}, \quad (1)$$

where *Effi* is the amplification efficiency of gene $i$, $X_0^i$ is the initial amount of target molecules, $X_t^i$ is the amount of target molecules when the threshold cycle *Ct* is reached. For 32 genes, efficiency was manually curated. The aberrant values deleted were determined as plus or minus three times the standard deviation.

### 4.2 Ct normalization

Expression levels of each gene are normalized by their amplification efficiency and the expression levels of reference genes and their amplification efficiency as follows:

$$Ct_i^s = \frac{(1 + Eff_i)^{-C_i^t}}{\frac{1}{R}\sum_{r=1}^{1}\left((1 + Eff_r)^{-C_r^s}\right)}, \quad (2)$$

where $Ct_i^s$ is the cycle threshold of gene $i$ on sample $s$, $\Delta Ct_i^s$ the corresponding normalized expression level and *Effi* its amplification efficiency. $r$ is one of the $R$ reference genes, *Effr* its amplification efficiency and $Ct_r^s$ its cycle threshold on sample s. Five different reference genes were measured on Fluidigm plates. The dispersion evaluation of those genes reveals a high variance of gene HanXRQChr04g0115631 expression (Tab. 1). Consequently, this gene was removed from the analysis, with also the gene HanXRQChr03g0090171 that shows low quality too. In total, only three reference genes were kept for the analysis: HanXRQChr05g0131911, HanXRQChr01g0029581, HanXRQChr01g0021131.

### 4.3 Plate effect

Each gene is measured on five different Fluidigm chips (hereafter named plate), as part of the expression variation can be due to the use of different plates, referred as plate effect. The range of dilution was constituted with a pool of genotypes and

**Table 2.** Hybrid genotypes with high number of NA measures, three genotypes had more than 10% of the measured genes on the Fluidigm plates with NA values.

| Hybrid genotype | Number of NA | NA percentage |
|---|---|---|
| SF005_CSR1 | 46 | 26.6 |
| SF017_SF268 | 34 | 19.7 |
| SF031_SF280 | 24 | 13.9 |

is present on each plate. The plate effect is estimated on the range of dilution for each gene with a linear model. In this model, expression level variation on the range of dilution is explained by the plate where the gene is measured and the dilution level. The $\Delta Ct_i^s$ of each gene are then adjusted depending on this plate effect.

### 4.4 Field effect

Sunflowers are field cultivated, local environmental variations in the field can influence expression levels referred as field effect. The field effect was estimated on field control genotypes (11% of total samples) using a mixed model including two spatial fixed factors (line and column numbers in the field), a replicate fixed factor if necessary, an independent random genetic factor, and the residual error (Bonnafous *et al.*, 2018). For each hybrid genotype, the $\Delta Ct$ value of each gene is corrected by the field effect measured on the closest control genotype in the field.

### 4.5 Missing values: curation and imputation

Among the 66,951 points measured using Fluidigm, 684 had NA value, which corresponds to 1% of missing data on the experiment. For the hybrids, 259 of them have one or more missing value. The average number of missing values per genotype is 1.8. Three genotypes had more than 10% of value missing (Tab. 2), we decided to remove those genotypes.

Over the 180 studied genes, 110 genes had missing values. The average number of missing values is four per gene. Five genes had 10% or more of missing values (see Tab. 3). We decided to keep those genes for future analysis.

Missing values were imputed gene per gene. For a gene $i$, its missing value on a hybrid genotype $h$ is replaced by the mean of the expression level of gene $i$ on the other hybrids, as follows:

$$Ct_i^h = \frac{1}{S}\sum_{s=1}^{S} Ct_i^s, \tag{3}$$

where $S$ is the total number of hybrid genotype, for which expression level is available for the gene $i$.

## 5 Data records

### 5.1 List of genotypes file

**15EX05_genotype.tsv.** This file contains the list of the different genotypes used. Each row corresponds to a genotype.

**Table 3.** Genes with high number of NA measures, five genes had NA expression values on more than 10% of hybrid genotypes.

| Gene id | Number of NA | NA percentage |
|---|---|---|
| HanXRQChr04g0109421 | 81 | 20.9 |
| HanXRQChr07g0207111 | 84 | 21.7 |
| HanXRQChr08g0216831 | 61 | 15.8 |
| HanXRQChr08g0227171 | 120 | 31.0 |
| HanXRQChr15g0481941 | 41 | 10.6 |

The first column (CROSS_GENOTYPE) contains the name of the hybrid genotype. Genotype name is composed of its parents name (first female genotype and second male genotype). The second column (STATUS_GENOTYPE) contains supplemental information about the genotype (*e.g.* control genotype or removed from the analysis)

### 5.2 List of genes file

**15EX05_geneList.tsv.** This file contains the list of genes measured on the Fluidigm plates. The first column (GENE_ID) contains the gene ID and the second column (STATUS_GENE) the nature of the gene (*e.g.* reference, biomarker or measured).

### 5.3 Field file

**15EX05_field.tsv.** This file contains the composition of each plot in the field (in rows). 'SAMPLE_NAME' is a number associated to the parcel, later used to name the sample providing from specific plot. 'XTRIAL' and 'YTRIAL' are the coordinate X and Y of the plot in the field. 'CROSS_GENOTYPE' is the name of the genotype. 'STATUS_EXP' is the type of sample ('exp' for studied hybrid or 'check' check-hybrids). The names of the other columns are the following: 'SOWING_DATE', 'EMERGENCE_DATE', 'PLANT_DENSITY', F1_DATE.

### 5.4 Raw data files

**15EX05_Raw_Data_Fluidigm_PlateX-BatchX.gz.** These 10 files contain all the raw data files obtained by the BioMark HD Fluidigm System depending to the plate and the batch.

### 5.5 Fluidigm results files

**15EX05_Fluidigm_results.gz.** This file contains the 10 result data sets (in CSV format) of the Fluidigm measurements obtained by the program, one file per Fluidigm plate. The structure of the file is the same for each file. No correction was applied to the data. Each row corresponds to a spot/chamber on the Fluidigm plate. The first column corresponds to the id of the chamber in the Fluidigm plate. The next three columns correspond to 'Sample' information (1) 'Name' contain a number that corresponds to the sample name of the hybrid genotype, (2) the 'Type' of sample, (3) 'rConc' the

concentration of the sample. The correspondence between the sample number and the hybrid name is given in the file: 15EX05_field.tsv. The next two columns correspond to the gene tested in the chamber plate (1) 'Name' its identifier and (2) its 'Type' (test or reference gene). The five next columns contain 'Ct' information (1) 'Value', (2) 'Calibrated rConc', (3) 'Quality', (4) 'Call', (5) 'Threshold'. The last three columns contain information about the melting temperature 'Tm', (1) 'In Range', (2) 'Out Range', (3) 'Peak Ratio'

### 5.6 Gene expressions file

**15EX05_expression_noNA.tsv.** This file contains the $\Delta Ct$ of each gene (in rows) on each genotype (in columns) corrected by plate and field effect where the missing data were imputed.

*Acknowledgements.* The authors are grateful to Sébastien Carrère and Ludovic Legrand for the storage help. These data were produced with the funding of the French National Research Agency (ANR SUNRISE ANR-11-BTBR-0005). This work was part of the "Laboratoire d'Excellence (LABEX)" TULIP (ANR-10-LABX-41).

## Supplementary material

Supplementary Table 1
The Supplementary material is available at https://www.ocljournal.org/10.1051/ocl/2020027/olm.

## References

Badouin H, Gouzy J, Grassa CJ, *et al*. 2017. The sunflower genome provides insights into oil metabolism, flowering and asterid evolution. *Nature* 546(7656): 148–152.

Bonnafous F, Fievet G, Blanchet N, *et al*. 2018. Comparison of gwas models to identify non-additive genetic control of flowering time in sunflower hybrids. *Theor Appl Genet* 131(2): 319–332.

Hussain M, Farooq S, Hasan W, *et al*. 2018. Drought stress in sunflower: physiological effects and its management through breeding and agronomic alternatives. *Agric Water Manage* 201: 152–166.

Livak KJ, Schmittgen TD. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2-*δ δct method*. *Methods* 25(4): 402–408.

Marchand G, Mayjonade B, Varès D, *et al*. 2013. A biomarker based on gene expression indicates plant water status in controlled and natural environments. *Plant Cell Environ* 36 (12): 2175–2189.

Rengel D, Arribat S, Maury P, *et al*. 2012. A gene-phenotype network based on genetic variability for drought responses reveals key physiological processes in controlled and natural environments. *PLoS One* 7(10): e45249.

Spurgeon SL, Jones RC, Ramakrishnan R. 2008. High throughput gene expression measurement with real time PCR in a microfluidic dynamic array. *PLoS One* 3(2): e1662.

Untergasser A, Cutcutache I, Koressaar T, *et al*. 2012. Primer3—new capabilities and interfaces. *Nucleic Acids Res* 40(15): e115–e115.

USDA. 2019. Oilseeds: world markets and trade. Technical report. USDA.

**Cite this article as**: Penouilh-Suzette C, Pomiès L, Duruflé H, Blanchet N, Bonnafous F, Dinis R, Brouard C, Gody L, Grassa C, Heudelot X, Laporte M, Larroque M, Marage G, Mayjonade B, Mangin B, de Givry S, Langlade NB. 2020. RNA expression dataset of 384 sunflower hybrids in field condition. *OCL* 27: 36.