TomExpress, a unified tomato RNA-Seq platform for visualization of expression data, clustering and correlation networks

Mohammed Zouine, Elie Maza, Anis Djari, Mattieu Lauvernier, Pierre Frasse, Abdelaziz Smouni, Julien Pirrello, Mondher Bouzayen

TomExpress, a unified tomato RNA-Seq platform for visualization of expression data, clustering and correlation networks

Mohamed Zouine[1,2,*], Elie Maza[1,2], Anis Djari[1,2], Mattieu Lauvernier[1,2], Pierre Frasse[1,2], Abdelaziz Smouni[3], Julien Pirrello[1,2] and Mondher Bouzayen[1,2]

[1]University of Toulouse, INPT, Laboratory of Genomics and Biotechnology of Fruit, Avenue de l'Agrobiopole BP 32607, Castanet-Tolosan F-31326,France,

[2]INRA, UMR990 Génomique et Biotechnologie des Fruits, Chemin de Borde Rouge, Castanet-Tolosan F-31326,France, and

[3]Laboratoire de Biotechnologie et Physiologie Végétales, Centre de recherche BioBio, Faculté des Sciences – Université Mohammed V de Rabat 4, Av. Ibn Batouta – BP. 1014 R.P., 10000, Rabat, Morocco

SUMMARY

The TomExpress platform was developed to provide the tomato research community with a browser and integrated web tools for public RNA-Seq data visualization and data mining. To avoid major biases that can result from the use of different mapping and statistical processing methods, RNA-Seq raw sequence data available in public databases were mapped de novo on a unique tomato reference genome sequence and post-processed using the same pipeline with accurate parameters. Following the calculation of the number of counts per gene in each RNA-Seq sample, a communal global normalization method was applied to all expression values. This unifies the whole set of expression data and makes them comparable. A database was designed where each expression value is associated with corresponding experimental annotations. Sample details were manually curated to be easily understandable by biologists. To make the data easily searchable, a user-friendly web interface was developed that provides versatile data mining web tools via on-the-fly generation of output graphics, such as expression bar plots, comprehensive *in planta* representations and heatmaps of hierarchically clustered expression data. In addition, it allows for the identification of co-expressed genes and the visualization of correlation networks of co-regulated gene groups. TomExpress provides one of the most complete free resources of publicly available tomato RNA-Seq data, and allows for the immediate interrogation of transcriptional programs that regulate vegetative and reproductive development in tomato under diverse conditions. The design of the pipeline developed in this project enables easy updating of the database with newly published RNA-Seq data, thereby allowing for continuous enrichment of the resource.

Keywords: tomato, RNA-Seq, database, platform, web tools, gene expression, data mining.

INTRODUCTION

The completion of the whole tomato genome sequence has opened new perspectives for the development of innovative tools that reinforce the position of the tomato as a reference species for Solanaceae and fleshy fruits (Tomato Genome, 2012). This achievement represents a key step in tomato genetics and genomic research. It facilitates the prediction of whole sets of genes and the assignment of their functional annotation, and allows researchers to take advantage of high-throughput technologies such as next-generation sequencing, the so-called NGS. Thanks to this resource and to the use of NGS technologies, RNA sequencing (RNA-Seq) is becoming the preferred method to study global gene expression patterns at the transcript level, aiming to gain insight into the mechanisms underlying plant developmental processes and the way they are affected by environmental and physiological conditions

(Wang *et al.*, 2009). Before the genome era, other technologies, such as microarrays, quantitative reverse transcription-polymerase chain reaction (qRT-PCR) and ESTs library screening, were widely used for gene expression analysis. However, these approaches can only address a limited number of genes (qRT-PCR) or with only a fraction of the whole transcriptome at best (EST screening or microarrays).

RNA-Seq is a powerful high-throughput approach that allows for the identification of global RNA transcripts and the deep quantification of their abundance (Wang *et al.*, 2009). RNA-Seq can also reveal important information on alternative splicing, allele-specific expression, unannotated exons and novel transcripts. RNA-Seq is now the method of choice for transcriptomic studies, leading to the generation of extensive amounts of publicly available data, the exploitation of which is instrumental to improve our understanding of tomato vegetative and reproductive growth. The high-quality reference genome sequences can be used as a backbone to guide RNA-Seq read mapping, thus helping to establish a precise and accurate gene expression value derived from the read counts mapped on each annotated gene.

However, the production and processing of the huge RNA-Seq data in different laboratories generates non-intentional undesirable effects. In particular, bioinformatics tools and the statistical methods used to calculate and normalize expression counts for each gene are not standardized. Indeed, different mapping parameters, different genome sequences, or different gene models and gene IDs are used in the published RNA-Seq projects. In addition, each study focuses on a limited set of genes that are linked to the specific scientific question addressed. To our knowledge, a tomato database allowing for the integration of the whole set of expression data arising from different studies does not yet exist. Another shortcoming relates to the description of the RNA-Seq experiments and the corresponding raw sequence data on the public databases SRA/ENA. The annotations of the samples are often incomplete and need manual curation to be more comprehensive.

To fill this gap and unify the whole tomato expression data set, the TomExpress platform (http://gbf.toulouse.inra.fr/tomexpress) has been developed with the aim of providing: (i) a standard pipeline to process and normalize all the tomato RNA-Seq raw sequence data; (ii) manually curated sample annotations; (iii) a dynamic database for sequence and experiment data handling; (iv) a user-friendly web interface to instantly access the whole set of data; and (v) several web tools to visualize and explore expression data (see Figure 1 for TomExpress screen

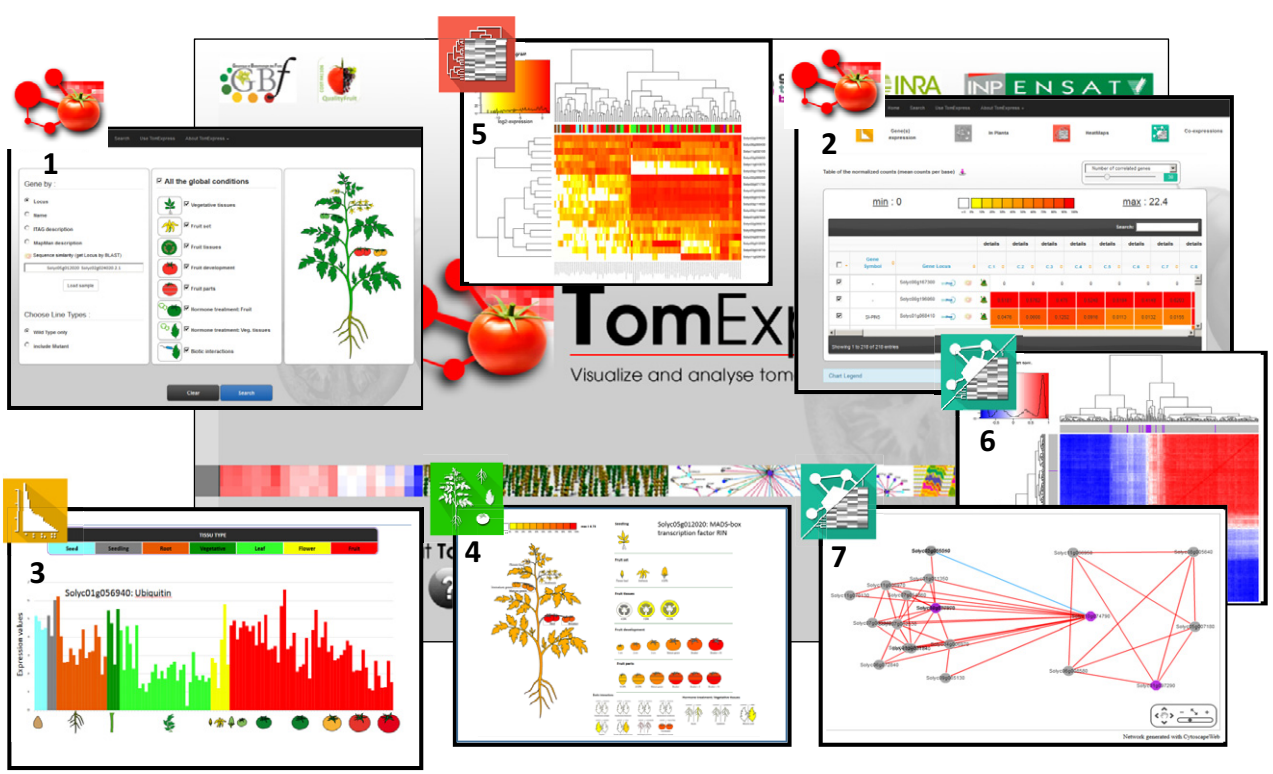

Figure 1. Screen shots of the TomExpress platform.
(1) Web browser; (2) data table; (3) bar plots of expression data; (4) *in planta* visualization; (5) heatmap of hierarchical clustering of normalized expression data; (6) heatmap of hierarchical clustering of Spearman's coefficients of correlations; (7) correlation networks graphics drawn by *Cytoscape*. [Colour figure can be viewed at wileyonlinelibrary.com].

shots). Importantly, while the pipeline has been developed for the tomato, it is designed to easily gather expression data from other plant species.

RESULTS AND DISCUSSION

RNA-Seq data source and manual curation of sample details

SRA and ENA public databases dedicated to NGS were probed to identify the tomato RNA-Seq projects. In total, 29 studies were found, and the corresponding raw data and descriptions were downloaded (http://gbf.toulouse.inra.fr/tomexpress/www/projectsTomExpress.php). These studies provide sequences derived from 349 RNA samples corresponding to 222 conditions. On average, three biological replicates per condition and 10 million reads per sample were performed. The transcriptomic sequencing studies cover most tomato organs, including roots, stems, leaves, flowers, and an exhaustive list of fruit development and ripening stages. In addition to a large panel of wild-type tomato cultivars, these transcriptomic data also cover several mutant lines as well as responses to biotic interactions with viruses, pathogenic bacteria and fungal and mycorrhizal interactions. Some of the studies focus on fruit or vegetative treatments with different hormones, such as auxin, gibberellins and ethylene. Overall, the transcriptomic data used to generate the TomExpress platform are comprehensive, being representative of the main types of tissues and organ development. To be easily accessible to all biologists, the annotation of the samples has been subjected to a manual curation and a unique nomenclature has been adopted.

Unified data processing

To unify the whole set of expression data and to make them suitable for comparable studies, all RNA-Seq raw sequences used in the TomExpress pipeline were mapped de novo on the latest version of the tomato reference genome sequence (SL2.50) using TopHat2-Bowtie2 mapping tools and accurate mapping parameters (Trapnell *et al.*, 2012). Following calculation of the number of counts for each iTAG2.40 annotated gene in each RNA-Seq project, the same normalization method was applied to the whole expression matrix as described in the Materials and methods. The summary of the pipeline is shown in Figure 2.

The normalized values in the different biological repeats corresponding to one global condition were averaged for each gene, resulting in an expression matrix of the whole Sl2.50/iTAG2.40 annotated tomato genes in the 222 global conditions.

Consistency and robustness of the TomExpress platform for the expression of ripening-associated genes

To experimentally validate the data delivered by the TomExpress platform, the expression patterns of 12 ERF genes assessed either by qRT-PCR or obtained by in silico mining using the TomExpress tool were compared. These ERF genes were previously shown to be associated with ripening (Liu *et al.*, 2016); their transcript levels were therefore assessed in tomato fruit tissues at four different ripening stages: namely, Mature Green (MG); Breaker (B); Orange (Or); and Red (R). Two methods were applied to quantify the correlation between two paired samples: the Spearman and Pearson correlation analyses. Pearson's coefficient quantifies a supposed linear relationship between given measures, while the less restrictive Spearman coefficient quantifies the intensity of a monotonic relationship. The two coefficients provide complementary information; for instance, if the Pearson coefficient is lower than the Spearman one for two given gene profiles, this indicates the existence of a link between observations even if the relationship is not linear. Figure 3 indicates that for 10 out of the 12 studied genes (ERF. B1, B2, B3, E1, E2, E4, E5, F1, F2 and F5), the expression patterns obtained by RT-PCR or extracted by TomExpress are highly correlated, whereas the correlation is weak for two ERFs (ERF.C1 and ERF.F4). These data show that the expression profiles obtained by qPCR and by TomExpress pipeline are correlated for more than 80% of the genes tested, which assures the consistency of the data extracted with this expression platform. Notably, even though the calculated correlation

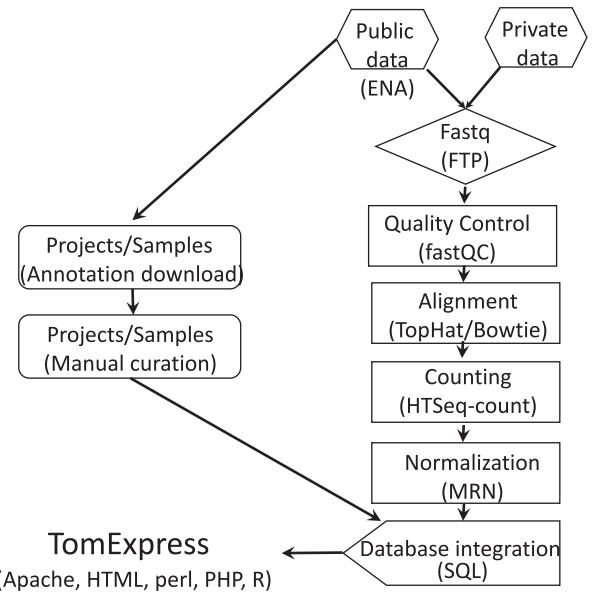

Figure 2. The TomExpress data processing and integration workflow.
Public or in-house private data are downloaded and subjected to a quality control step. Reads are then aligned to the same reference sequence using TopHat2/Bowtie2 alignment tools. For each gene, the number of read counts mapped on the exons part are calculated using HTSeq-count tool. The whole counts are normalized between the RNA-Seq projects. Sample details downloaded from ENA are manually curated and integrated in the database together with the normalized expression data.

coefficient is low for ERF.C1 and ERF.F4, the expression profiles provided by in silico mining are similar to those obtained experimentally. It is worth mentioning here that the TomExpress data correspond to the mean value of normalized counts of expression data achieved with different tomato cultivars; by contrast, the qPCR data refer to only the Microtom cultivar. Therefore, the differences may reflect a potential specificity of this cultivar.

Database features and data mining tools

To handle, explore and make searchable the whole set of processed data, the TomExpress platform was developed with integrated and versatile data visualization and mining web tools via a friendly dedicated web interface.

Data search browser. The TomExpress home page provides access to the whole expression data set through a comprehensive form page. Users can target specific genes by specifying their iTAG gene IDs or search genes by key words in iTAG gene descriptions or gene names. They can check expressions by selecting specific developmental stages or treatments that are easily understandable on this form. Finally, they can check expressions in wild-type cultivars only or can include mutant lines if desired.

Data visualization table. When users submit their desired search options and parameters to the TomExpress platform, the expression results corresponding to the specified genes and experimental conditions are first shown in an interactive web table. The first column presents the list of genes in study; from this column, the users can be directed to the Sol Genomics Network (SGN) web page of the corresponding gene by clicking on the SGN logo that gives all details concerning the sequence information and the structural annotation of the gene(s). In the same way, a second external link can direct users to the MapMan functional annotation.

The first line of the table presents the list of global conditions chosen by the user; for each row of this line, a dedicated link offers access to the sample details, the

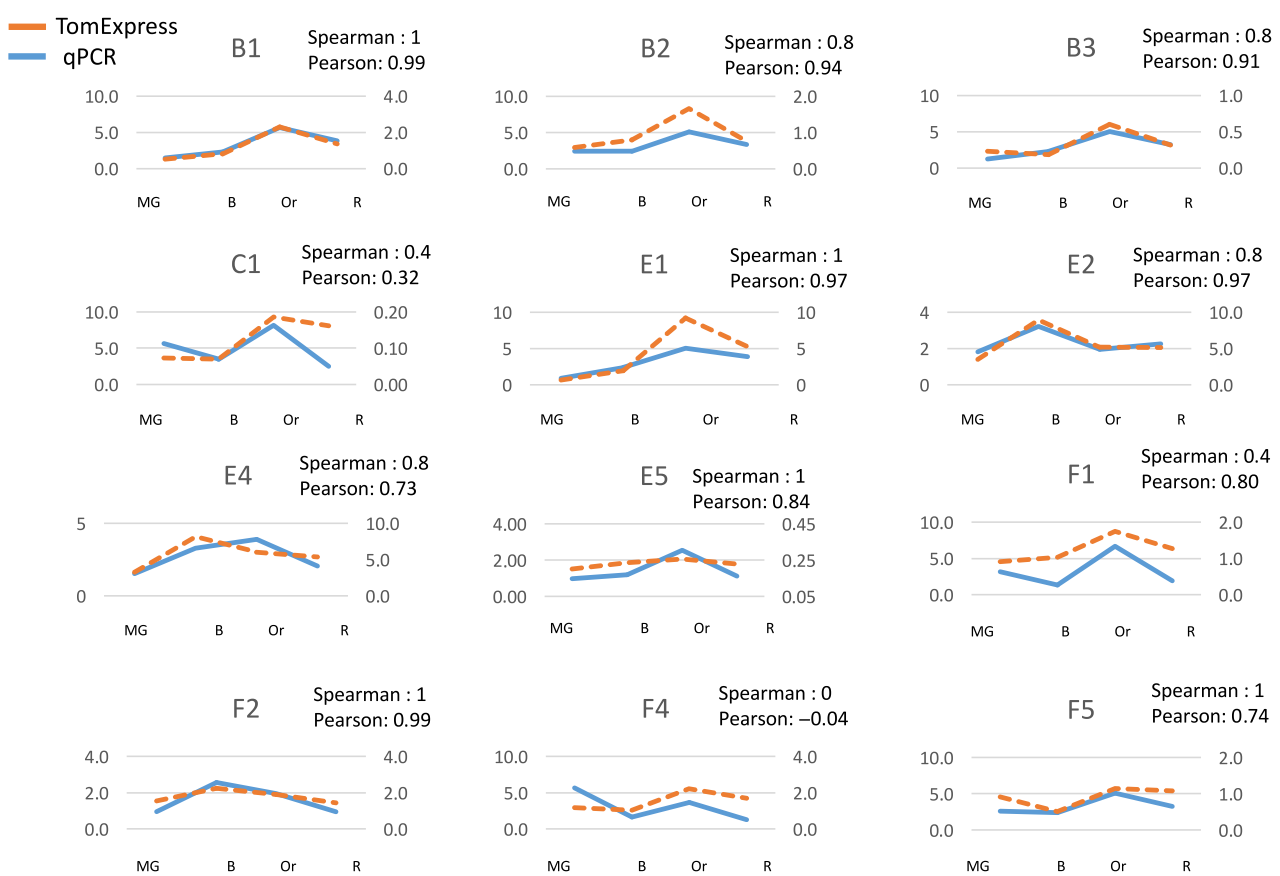

Figure 3. Expression pattern of ripening-associated SlERFs obtained by quantitative reverse transcriptase-polymerase chain reaction (qRT-PCR) or through mining the TomExpress platform. For each ERF gene, the expression pattern during four ripening stages in wild-type tomato fruit was assessed by qPCR (continuous lines, left axis) or determined by TomExpress data mining (dotted lines, right axis). For qRT-PCR data, the relative expression of each ERF gene is standardized to 1, referring to Sl-Actin gene as an internal control. In the case of TomExpress data, the expression of each gene represents the mean value of normalized counts of all expression data contained in TomExpress and corresponding to various tomato cultivars. The correlation coefficients determined by Spearman or Pearson methods are given for each gene. MG, Mature Green; Br, Breaker; Or, Orange; R, Red. The left *y*-axis is relative expression obtained by qPCR; the right *y*-axis is expression level obtained by TomExpress. [Colour figure can be viewed at wileyonlinelibrary.com].

published work related to the gene(s) when available, and to the row data table in the SRA or ENA databases. Users can also use intuitive icons that allow for sorting the expression data by a given condition. To make the expression data visually comparable among genes and conditions, specific background colors for each cell in the table correspond to the expression levels among the whole expression matrix. Specifically, all expression values are divided into 10 percentile groups; a different color is assigned to each subgroup based on the level of expression of the corresponding gene. The color scale is white (low)–yellow–red (high).

Starting from this table, the TomExpress platform offers the users several web tools to visualize the expression data through interactive graphics or through *in planta* representations. Additional data mining web tools can be used for hierarchical clustering approaches with custom parameters or for identification of co-expressed genes. If needed, users have the capacity to download the whole data set tables in a tabular format.

Data visualization. To visualize the expression data, two web tools were developed. The first tool allows users to draw the expression data in bar plot graphics, while the second tool generates pictures of tomato organs with artificial background colors highlighting the expression levels of the studied gene. To do so, the first step is to activate the check box beside each gene on the expression table to select one or more genes. The second step is to launch the dedicated web tool to visualize expression values in bar plots. When dealing with more than one gene, a multicurve graphic is generated instead of a bar plot. To enable easy visualization of comparative expressions across the global conditions, a color code specifies each bar plot corresponding to different plant organs/tissues. This color code allows the users to know at a glance where the gene of interest is being expressed. All graphics are interactive; the mouse moving over specific areas of the graphics allows for display of the details related to the samples, gene IDs and expression values. Codification adopted for sample annotation is accessible by clicking on the Chart Legend button located at the bottom of the graphic representation.

Alternative to the expression table and bar plot graphics, the normalized expression data of a gene of interest can be visualized on virtual *in planta* pictures that include several growth stages, different vegetative and reproductive tissues, and whole or parts of fruit organs (Figure 4). To easily depict the expression signals using the *in planta* visualization tool, the same color scale used for the expression table is also applied for the organ background.

Clustering gene expression data. Clustering is often an important step in gene expression analysis. The TomExpress clustering tool implements the most commonly used clustering methods for gene expression data analysis. Three classical distances can be used to cluster gene expressions: (i) the Euclidean distance focuses on expression levels and will cluster genes together that have the same level of expression; (ii) the Spearman distance (based on the Spearman correlation coefficient) will cluster genes together that are similar in a pattern point of view; (iii) and finally, the squared Spearman distance will cluster genes together that are highly correlated in absolute value (i.e. either positively or negatively). The output of one of these clustering methods is a heatmap showing both the clusters and the expression with an appropriate color palette. Moreover, genes alone or both genes and global conditions can be clustered in the same heatmap output.

Co-expression and correlation networks analysis. Co-expression methods are widely used to analyze gene expression data and other high-dimensional 'omics' data. The TomExpress co-expression tool allows users to identify genes that display similar or opposite expression profiles across either all samples or only user-selected ones. The main feature of this tool is that it allows for comprehensive visualization of the co-expression results based upon the calculation of the correlation values. Users can specify the thresholds of Spearman correlation coefficients or simply limit the results to a chosen number of top correlated genes. The network visualization is automatically generated through a Cytoscape plugin (Lotia *et al.*, 2013). The correlation map is interactive, allowing users to view additional information on gene annotations and on correlation values. Correlation data are also visualized as a heatmap after a hierarchical clustering to highlight the positively and negatively correlated groups. Finally, a web table with the pairs of co-expressed genes is displayed. This table displays the correlation values and the functional annotation details of each of the co-expressed genes.

Export of processed expression data and figures. All data tables and figures processed by the TomExpress platform built-in web tools are freely downloadable for any further custom analysis or direct integration in scientific presentations or other types of documents. Of note, unlike the web version of the pictures that is limited to a resolution adapted for screen visualization, the resolution of the downloadable version of pictures was increased to meet requirements for hard-copy printing.

TomExpress a versatile tool: case study and application

As shown below, in addition to the mining and visualization of expression data in a holistic way, TomExpress provides versatile tools that allow for gaining insight into a particular biological process just by in silico search.

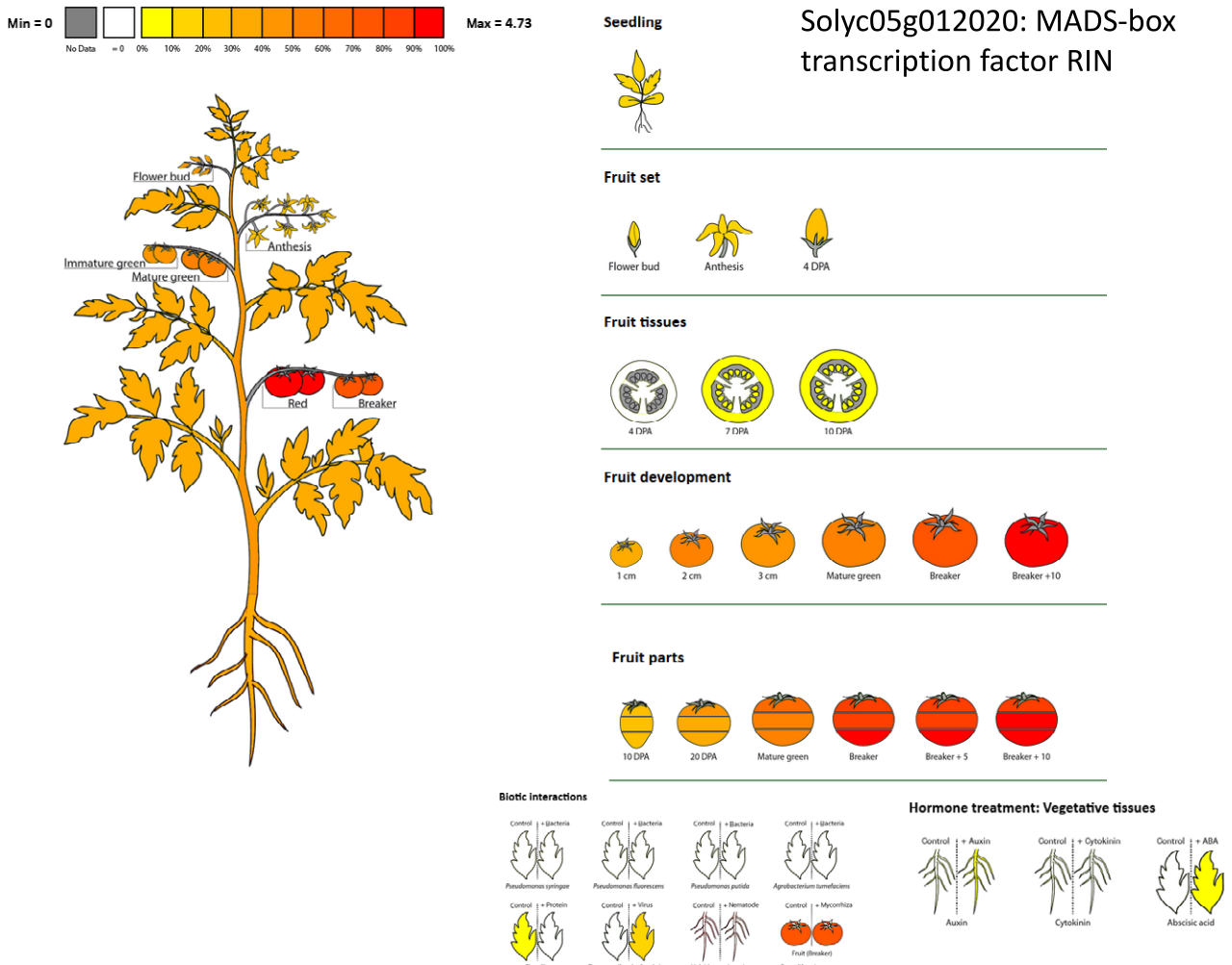

Figure 4. *In planta* visualization of RIN expressions.
To depict the expression signals, each organ is colored according to RIN expression levels. [Colour figure can be viewed at wileyonlinelibrary.com].

Furthermore, the TomExpress platform offers powerful tools to define the most appropriate reference genes to be used in quantitative RT-PCR experiments aiming to assess transcript accumulation in specific plant tissues or organs. Finally, the expression and clustering analysis tools offered by TomExpress can also reveal potential redundancy among gene family members.

TomExpress is a powerful tool to find appropriate RT-PCR reference genes for different plant tissues. Data normalization is always instrumental for accurate expression monitoring of individual genes by real-time RT-PCR at the transcript level. To correct for sample-to-sample variations, invariant endogenous controls called reference genes must be included in the RT-PCR experiment to ensure the reliability of the results. In most cases, housekeeping genes, such as ubiquitin and actin, are used as reference genes for the RT-PCR experiments in tomato. It is widely accepted that the expression of these genes is invariant in different plant tissues or organs. The TomExpress platform has been used to assess the quality of these reference genes by monitoring their expression in various tissues and tomato organs. The TomExpress pipeline, which processes hundreds of samples, reveals that the ubiquitin gene (Solyc01g056940), commonly used as a reference gene in expression studies, shows a non-uniform pattern of expression, indicating that this gene is not the ideal candidate for qRT-PCR reference during fruit development and ripening stages (Figure 5). Interestingly, TomExpress can be used to identify and validate new candidates, such as the GAGA-binding transcriptional activator (Solyc04g008380), which displays a nearly constant expression in various fruit tissues (Figure 5), clearly showing that this gene is much more suitable as an internal reference in qRT-PCR studies dealing with tomato fruit, flower and seeds. Indeed, while ubiquitin transcript accumulation

varies up to six times in reproductive tomato tissues, the abundance of GAGA-binding transcriptional activator transcript levels vary two times at most. Moreover, when comparing tomato vegetative and reproductive tissues, the variation in transcript levels exceeds 10 times in extreme cases. Altogether, TomExpress proves to be effective in finding more suitable reference genes than ubiquitin for gene expression profiling in tomato.

To extend this analysis, we included other popular reference genes published by Expósito-Rodríguez *et al.* (2008) in this comparison. We calculated the coefficient of variation as the mean of normalized expression values in all TomExpress conditions divided by the mean of normalized expression to reduce the effect of the global expression of each of these genes and better evaluate the expression variations across all conditions. As shown in Table S1, SAND (SGN-U316474, Solyc03g115810), CAC (SGN-U314153, Solyc08g006960) and GAGA (Solyc04g008380) genes show the least variation (coefficient of variation below 0.5), while TIP41 (SGN-U321250, Solyc10g049850), UBI (Solyc01g056940) and 'Expressed' (SGN-U346908, Solyc07g025390) change more across all TomExpress conditions.

To give users more possibilities for choosing the appropriate reference genes and to better fit within the expected expression level of their studied genes, we divided the whole set of tomato genes into three groups based on their global expression level (25–50%, 50–75%, 75–100%, most highly expressed genes). For each group, the top 10 candidates that represent genes with the least variance in

Figure 5. Bar plots of expression profiles of ubiquitin and GAGA-binding transcription factors. [Colour figure can be viewed at wileyonlinelibrary.com].

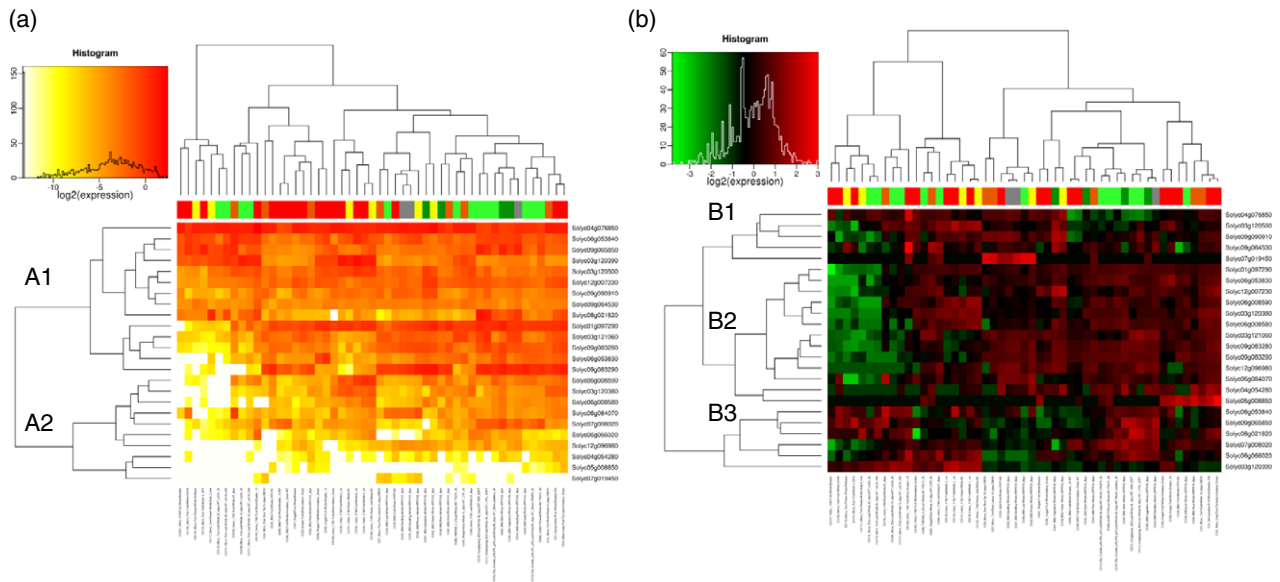

Figure 6. Heatmaps of hierarchical clustering of AuxIAA gene family expression data.
(a) Using the Euclidian-based clustering method, the AuxIAA genes can be divided into two major groups, A1 and A2, based on the levels of expression in the different conditions.
(b) Using Spearman's correlation, the AuxIAA genes can be divided into three major groups, B1, B2 and B3, depending on the expression profile of each gene. [Colour figure can be viewed at wileyonlinelibrary.com].

vegetative or reproductive tissues were identified and are listed in Table S2.

The current version of TomExpress platform allows the user to choose a desired cultivar and type of tissues to tune their analysis toward selecting the best set of reference genes for their qPCR experiment.

Potential redundancy in gene families revealed by expression and clustering analysis tools. Another advantage of TomExpress is the ability to rapidly assess potential redundancy among gene family members by comparing the expression profile of all family members simultaneously in diverse organs, tissues and conditions. This can be achieved by drawing the interactive expression curves in the desired tissues using the 'plot' web tool or using diverse data clustering methods followed by heatmap representations of the clusters. This is well illustrated by the comparative analysis of the tomato *AuxIAA* gene family members known to be required for mediating auxin-responses (Audran-Delalande *et al.*, 2012). Using the Euclidian-based clustering method, *AuxIAAs* are divided into two major clades, A1 and A2, with genes grouped based on their level of expression in different tissues or conditions (Figure 6a). From these results, we can observe that the A1 group contains *AuxIAA* genes that are more expressed overall than their paralogs in the A2 group. By contrast, when using the Pearson correlation, which clusters the expression data based on the pattern of each gene, *AuxIAA* gene family members are divided into three main clades: B1, B2 and B3 (Figure 6b). This classification method leads to the identification of tissue-preferential expression profiles displayed by these distinct clusters of genes.

Improvement perspectives and discussion

The bioinformatics and statistics pipeline used in the TomExpress platform allows for unification of the whole set of RNA-Seq expression data. The intuitive and user-friendly web interface offers a highly valuable tool to biologists to immediately access several transcriptomes covering many tomato organs, developmental stages and relevant treatments. The didactic web form provides a powerful means to query these data. The manual curation of the sample details allows the users to easily understand the kind of biological material used.

Thanks to the built-in web tools, biologists can instantly compare expressions and mine data with popular and approved methods, such as hierarchical clustering and co-expression identification algorithms. These tools generate easy-to-interpret graphics that can be downloaded for further processing. Taken together, the data and tools offer an opportunity to biologists to discover new stories, hoping to gain insight from functional genomics in an accelerated way. Indeed, the functional annotation of many tomato genes has not yet been established. The identification of co-expressed groups can help to assign a biological process to these genes. To validate the predicted functional annotation of a particular gene of interest and

promising candidate genes in its co-expressed group, biologists can use a reverse genetic approach.

The current version of TomExpress holds approximately 349 RNA-Seq samples from public and private data. The pipeline developed in this project allows the database to be easily updated with newly published RNA-Seq data, allowing this resource to be enriched and up-to-date with supplemental samples.

As TomExpress is fully dedicated to processing tomato RNA-Seq data, it makes this database complementary to other databases more specialized on microarray technology or covering a subset of public tomato RNA-Seq data, such as TED (Fei *et al.*, 2006) and the Botany Array Resource (BAR; Toufighi *et al.*, 2005) databases.

Other omics data generated using NGS technologies are publicly available. Part of these omics data, such as epigenomics or ChiP-Seq, can help biologists to understand how transcriptomes change. Integration of these data into TomExpress would help to define combinatory epigenetic and/or transcription factors that may orchestrate the different transcriptional programs that define cells, tissues and organs in different conditions.

EXPERIMENTAL PROCEDURES

RNA-Seq data processing

Sequence reads were cleaned using fastQC and mapped to a unique reference tomato genome (SOLY2.50) using TopHat/Bowties tools. The mapping parameters consider each library size when paired-end sequencing was performed. The mapping was guided using the gene model annotation file iTAG2.50. HTSeq-count was then used to calculate the read counts for each gene from the accepted-hits.bam mapping file.

Normalization procedure

A normalization was performed to obtain comparable expression values between genes and between global conditions. For this purpose, our pipeline takes into account the relative size of studied transcriptomes, the library sizes and the gene lengths, as described in Maza *et al.* (2013). The pipeline is described below.

(i) First, raw counts of technical replicates are summed to work only with biological replicates. These obtained biological replicates are also hereafter referred to as libraries or samples.

(ii) A reference global condition is needed to proceed to the normalization procedure described in Maza *et al.* (2013). For this, we normalized each sample by its library size (the sum of all obtained raw counts). We then chose the global condition containing the biological replicate closest to the mean expression profile of all biological replicates (in the Euclidean sense) as the reference.

(iii) Normalization of libraries is then carried out by the procedure developed in Maza *et al.* (2013). This method shares the same normalization goals as LRE and TMM normalization methods from *DESeq* and *edgeR* R packages, respectively: it removes the biases due to the sizes of libraries and transcriptomes. Normalized counts are then obtained by dividing each library count by its normalization factor.

(iv) Normalized biological replicates are then averaged to obtain normalized values for global conditions.

(v) Finally, gene length bias is considered by dividing each normalized global condition mean value by the length (number of bases) of each corresponding gene. We thus obtain mean normalized counts per base for each global condition.

ACKNOWLEDGEMENTS

This research was supported by the 'Laboratoire d'Excellence' (LABEX) entitled TULIP (ANR-10-LABX-41), by the ANR TomEpiSet project and the TomGEM H2020 project, and benefited from networking activities within the European COST Action FA1106. The authors are grateful to Christophe Klopp from GenoToul bioinformatics for his help.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

SUPPORTING INFORMATION

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

Table S1 Expression variation in different tissues of putative reference genes for qPCR studies.

Table S2 Candidate reference genes that show low variances from groups ranging from 25% to 50%, 50% to 75%, 75% to 100%, most highly expressed genes.

REFERENCES

Audran-Delalande, C., Bassa, C., Mila, I., Regad, F., Zouine, M. and Bouzayen, M. (2012) Genome-wide identification, functional analysis and expression profiling of the Aux/IAA gene family in tomato. *Plant Cell Physiol.* **53**, 659–672.

Expósito-Rodríguez, M., Borges, A.A., Borges-Pérez, A. and Pérez, J.A. (2008) Selection of internal control genes for quantitative real-time RT-PCR studies during tomato development process. *BMC Plant Biol.* **8**, 131.

Fei, Z., Tang, X., Alba, R. and Giovannoni, J. (2006) Tomato Expression Database (TED): a suite of data presentation and analysis tools. *Nucleic Acids Res.* **34**, D766–D770.

Liu, M., Gomes, B.L., Mila, I. *et al.* (2016) Comprehensive profiling of ethylene response factor expression identifies ripening-associated ERF genes and their link to key regulators of fruit ripening in tomato. *Plant Physiol.* **170**, 1732–1744.

Lotia, S., Montojo, J., Dong, Y., Bader, G.D. and Pico, A.R. (2013) Cytoscape app store. *Bioinformatics*, **29**, 1350–1351.

Maza, E., Frasse, P., Senin, P., Bouzayen, M. and Zouine, M. (2013) Comparison of normalization methods for differential gene expression analysis in RNA-Seq experiments: a matter of relative size of studied transcriptomes. *Commun. Integr. Biol.* **6**, e25849.

Tomato Genome, C. (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, **485**, 635–641.

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

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

Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63.