Transcriptome profiling of sorted endoreduplicated nuclei from tomato fruits: how the global shift in expression ascribed to DNA ploidy influences RNA-Seq data normalization and interpretation

Julien Pirrello, Cynthia Deluche, Nathalie Frangne, Frederic Gevaudant, Elie Maza, Anis Djari, Mickael Bourge, Jean-Pierre Renaudin, Spencer Brown, Chris Bowler, et al.

DR CHRISTIAN CHEVALIER (Orcid ID : 0000-0002-5727-6206)

Article type : Technical Advance

Handling Editor: Dr Alisdair Fernie

Transcriptome profiling of sorted endoreduplicated nuclei from tomato fruits: how global shift in expression ascribed to DNA ploidy influences RNA-Seq data normalization and interpretation

Pirrello Julien[1,2,#], Deluche Cynthia[1], Frangne Nathalie[1], Gévaudant Frédéric[1], Maza Elie[2], Djari Anis[2], Bourge Mickaël[3], Renaudin Jean-Pierre[1], Brown Spencer[3], Bowler Chris[4], Zouine Mohamed[2], Chevalier Christian[1,*] and Gonzalez Nathalie[1]

[1] *UMR1332 BFP, INRA, Univ. Bordeaux, 33882 Villenave d'Ornon Cedex, France*

[2] *UMR990 GBF, INRA, INP-ENSAT, 31326 Castanet-Tolosan Cedex, France*

[3] *Institute of Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91198, Gif-sur-Yvette, France*

[4] *IBENS, Département de Biologie, Ecole Normale Supérieure, CNRS, Inserm, PSL Research University, F-75005, Paris, France*

* For correspondence (email christian.chevalier@inra.fr)

[#] Present Address: UMR990 GBF, INRA, INP-ENSAT, 31326 Castanet-Tolosan Cedex,

Running Head: RNA-Seq analysis of sorted endoreduplicated nuclei

Keywords: endoreduplication; fruit development; RNA-Seq profiling; data interpretation; sorted nuclei; tomato; Solanum lycopersicum

SUMMARY

As part of normal development most eukaryotic organisms ranging from insects to mammals and plants display variations in nuclear ploidy levels resulting from somatic endopolyploidy. Endoreduplication is the major source of endopolyploidy in higher plants. Endoreduplication is a remarkable characteristic of the fleshy pericarp tissue of developing tomato fruits, where it establishes a highly integrated cellular system that acts as a morphogenetic factor supporting cell growth. However, the functional significance of endoreduplication is not fully understood. Although endoreduplication is thought to increase metabolic activity due to a global increase in transcription, the issue of gene-specific ploidy-regulated transcription remains opened. To investigate the influence of endoreduplication on transcription in tomato fruit, we tested the feasibility of a RNA-Seq approach using total nuclear RNA extracted from purified populations of flow cytometry-sorted nuclei based on their DNA content. Here we show that cell-based approaches to study RNA-Seq profiles need to take into account the putative global shift in expression between samples for correct analysis and interpretation of the data. From ploidy-specific expression profiles we found that the activity of cells inside the pericarp is related both to the ploidy level and their tissue location.

INTRODUCTION

Most multicellular organisms display variations in cell ploidy levels resulting from somatic endopolyploidy as part of their normal development (Brodsky and Uryvaeva, 1977; Edgar and Orr-Weaver, 2001; Frawley and Orr-Weaver, 2015). In higher plants, endoreduplication is the major source of endopolyploidy (D'Amato, 1984) and is widespread in angiosperms where it takes place in various vegetative and reproductive tissues (D'Amato, 1984; Joubès and Chevalier, 2000; Nagl, 1976). During endoreduplication, nuclear DNA duplication occurs in the absence of any obvious condensation and decondensation steps. Chromosomes with multivalent (n=2, 4, 8, 16...) chromatids are thus produced without any change in the karyotype. This typical structure in polytene chromosomes is observed in various plant species, especially in the endosperm, the embryo suspensor and the anther tapetum (Cavalheira, 2000), and also in the fleshy pericarp tissue of developing tomato fruits (Bourdon *et al.*, 2011).

The functional significance of endoreduplication and its adaptive value during plant development are not fully understood and are still actively debated. Endoreduplication is thought to provide a means to sustain plant growth under adverse abiotic and biotic environmental conditions (Adachi *et al.*, 2011; Ceccarelli *et al.*, 2006; Cookson *et al.*, 2006; Hase *et al.*, 2006; Scholes and Paige, 2015). Endoreduplication is also clearly associated with cell differentiation and morphogenesis (Bramsiepe *et al.*, 2010; Hülskamp *et al.*, 1999; Roeder *et al.*, 2010). For example, in Arabidopsis, the developmental program of trichomes (epidermal hair cells) requires four rounds of endoreduplication to produce mature large-branched single cells (Hülskamp *et al.*, 1999). Positive correlations between ploidy level and cell size have also been observed in different plant species, organs and cell types, ascribing to endoreduplication a role in cell growth (Chevalier *et al.*, 2011). During tomato fruit

development, endoreduplication contributes to establish a highly structured and integrated cellular system that acts as a morphogenetic factor supporting cell growth (Bourdon *et al.*, 2012).

Endoreduplication has also been suggested to enhance metabolic activity in both the animal and plant kingdoms, for example during embryogenesis in drosophila or in the endosperm of cereal seeds (Inzé and De Veylder, 2006; Lee *et al.*, 2009). This proposed increase in metabolic activity could originate from a global increase of transcription (Lee *et al.*, 2009). This long-standing hypothesis has only been recently supported since we have shown that increased transcription of rRNA and mRNA on a per-nucleus basis is associated with higher ploidy levels in the pericarp tissue of tomato fruit (Bourdon *et al.*, 2012). Various transcriptomic studies have been performed in tomato to profile gene expression in the developing whole ovary and fruit, or in various dissected fruit tissues (Lemaire-Chamley *et al.*, 2005; Vriezen *et al.*, 2008; Wang *et al.*, 2009; Ye *et al.*, 2015; Zhong *et al.*, 2013). Laser-capture microdissection coupled to genome-wide transcriptomics has also been performed for expression analyses of individual cell and tissue types in tomato fruit (Matas *et al.*, 2011; Pattison *et al.*, 2015). Although these studies provide high resolution transcriptomic information, they do not allow the study of the potential influence that endoreduplication may confer on gene expression. Yet Galitski *et al.* (1999) showed the ploidy-dependent expression of 17 genes in isogenic yeast strains ranging from ploidy levels of n to 4n. This report prompted us to search for a ploidy-dependent gene expression during fruit development.

To study and compare genome-wide transcript expression, RNA-Seq has become a widely used procedure. After sequencing, an important step in RNA-Seq data analysis is the normalization step in which raw data are adjusted to correct for experimental or biological biases and technical variation that prevent direct comparison of expression measurements across samples. Classical normalization methods, such as the largely used LRE and TMM methods (Anders and Huber, 2010; Robinson and Oshlack, 2010), commonly assume comparable total expression between samples or a specific amount of significantly up- and down-regulated genes (Maza *et al.*, 2013). However, as highlighted by Aanes *et al.* (2014), Chen *et al.* (2015) and Evans *et al.* (2016), the performance of normalization methods can be affected because these assumptions may not always be valid. For example, in zebrafish embryos, due to delayed polyadenylation of maternal transcripts, a large difference in polyA+ mRNA amounts is found between two early developmental stages (Aanes *et al.*, 2011). Such a case of global shift in expression could be observed in samples presenting different levels of ploidy (Bourdon *et al.*, 2012; Lemaire-Chamley *et al.*, 2005), therefore indicating that care has to be taken for the analysis of ploidy-dependent transcriptomes.

To investigate the effects of DNA ploidy levels on gene expression in tomato fruit and to further decipher the functional role of endoreduplication, we here report the feasibility of a RNA-Seq approach using total nuclear RNA extracted from purified populations of FACS-sorted nuclei based on their DNA content. We then provide recommendations to analyze transcriptomic data under the influence of ploidy levels, and show ploidy-specific expression profiles suggesting a ploidy-dependent cell differentiation within the fruit pericarp tissue.

RESULTS

Genome-wide RNA sequencing using endoreduplicated nuclei from fruit pericarp: experimental setup and data preprocessing

To analyze the influence of endoreduplication on gene expression, RNA-Seq was performed using RNA extracted from purified nuclei sorted based on their DNA content by FACS. Nuclei were prepared from pericarp cells of 30 day-post-anthesis (dpa) fruits because at this developmental stage the immature green tomato fruits have completed their growth and display a wide range of ploidy levels (from 2C to 512C) (Figure 1) (Joubès *et al.*, 1999; Cheniclet *et al.*, 2005). Four populations of endoreduplicated nuclei, namely of 4C, 8C, 16C, and 32C DNA content, were sorted on a per-fixed number (ca. 100,000 nuclei) and total nuclear RNA was extracted for the construction of cDNA libraries.

Unlike typical RNA-Seq libraries that are prepared from enriched mRNA resulting from poly-A purification of total RNA, the libraries for sequencing in this study were constructed from nuclear non-polyadenylated RNA. In such samples, due to the high abundance of ribosomal RNA (rRNA), the correct detection of mRNA can be difficult. To estimate and maximize the access to coding information, two sets of libraries were prepared following cDNA synthesis, adapter ligation and PCR amplification steps: with or without a duplex-specific nuclease (DSN) treatment (Zhao *et al.*, 2014) applied to remove highly abundant sequences that reanneal quickly such as highly repetitive sequences and rRNA. In total, the samples analyzed corresponded to 24 RNA-Seq libraries resulting from four DNA ploidy levels in triplicate, with (DSN) or without (no-DSN) DSN treatment.

After sequencing, quality check and pre-processing (see Experimental Procedures and Supplemental Figure S1), the reads were mapped against the Sly2.40 tomato reference genome (Figure 2a). Without DSN treatment, the overall mapping rate ranged between 90 and 98% from 4C to 32C samples, respectively. The multi-mapping rate was relatively high, about 38% for all ploidy levels to be considered, and consisted mainly of rRNA sequences (Zhao *et al.*, 2014; O'Neill *et al.*, 2013). Uniquely mapped reads corresponding to reads aligned at only one locus represented more than 50% of total reads and reached 60% for 32C samples. For DSN-treated samples, the overall mapping rate fell to 40% in 4C nuclei and increased with ploidy level, reaching 85% for 32C samples. The multi-mapping rate fell around 11-12% for all the DSN-treated samples. This significant decrease was probably due to the removal of rRNAs by the treatment. The rate of sequences that were uniquely mapped was also reduced compared to non-treated samples, suggesting that the DSN treatment did not only target rRNA sequences. Unlike the multi-mapping rate, uniquely mapped rates seemed to increase with ploidy level since it increased from 31.6% to 69% at 4C and 32C, respectively. The proportion of uniquely mapped reads falling in intergenic, intronic, exonic and ribosomal regions were respectively 33%, 23%, 19% and 25% in 4C samples; 37%, 23%, 15% and 25% in 8C samples; 39, 26, 19 and 16% in 16C samples; 34%, 23%, 18% and 25%, in 32C samples (Figure 2b). Overall, these rates were found to be in accordance with those obtained in previous experiments applying DSN treatment (Zhao *et al.*, 2014).

Data normalization is a crucial step for an RNA-Seq analysis of samples displaying different ploidy levels

We have previously shown that the transcription of 5.8S rRNA in endoreduplicated nuclei was enhanced and that RNA polymerase II protein levels were increased proportionally with ploidy, suggesting that gene expression globally doubles between ploidy levels k and k+1 (Bourdon *et al.*, 2012). The classical normalization methods (Anders and Huber, 2010; Robinson and Oshlack, 2010), generally based on the assumption that less than half of the genes are down- or up-regulated in a sample (Maza *et al.*, 2013; Maza, 2016), might not be adequate for the comparison of samples with different ploidy levels, such as in our experimental setup.

To properly identify ploidy-dependent differentially expressed genes, we therefore first evaluated the adequacy of the use of a classical RNA-Seq normalization approach when a global shift in expression between two conditions exists, i.e., when all genes are proportionally expressed in the two given conditions.

Case of a global shift in expression. We hereafter demonstrated mathematically the inadequacy of classical normalization methods for a case study in which gene expression is globally proportional between two conditions. For the sake of simplicity, we considered an RNA-Seq experiment with only two conditions and one biological sample per condition, but a similar proof could a priori be done with more than two conditions and replicates. Let G be the total number of expressed genes in at least one of the two given conditions. Let X_{gk} be the raw count of gene g in condition k. Up to a centered random error, we can model the expression value X_{gk} by:

$$X_{gk} = \frac{\mu_{gk} L_g}{S_k} N_k \text{ with } S_k = \sum_{g=1}^{G} \mu_{gk} L_g$$

where μ_{gk} is the unknown number of transcripts of gene g in condition k, L_g is the known length of gene g, N_k is the known library size, and S_k is the size of the transcriptome in condition k which is obviously unknown (Maza *et al.*, 2013). Let us then assume that a global shift in expression exists, which means that, for all genes g, the following proportionality assumption holds: $\mu_{g2} = \alpha\mu_{g1}$ with α a given positive coefficient. First, by straightforward calculations, we have that $S_2 = \alpha S_1$. Then, given both the expression model and the proportionality assumption above, the following calculations show that the raw fold change of gene g expressions between the two conditions, denoted by RFC_g, does not depend on the proportionality coefficient:

$$\mathrm{RFC}_g = \frac{X_{g2}/N_2}{X_{g1}/N_1} = \frac{\mu_{g2}}{\mu_{g1}} \frac{S_1}{S_2} = \frac{\alpha\mu_{g1}}{\mu_{g1}} \frac{S_1}{\alpha S_1} = 1.$$

Hence, with a classical RNA-Seq normalization approach, such as those defined above, we are not able to estimate the proportionality coefficient α of the global shift assumption. In other words, when the global shift in expression assumption holds, genes with a fold change equal to α will be declared as not differentially expressed (not DE).

Case of a few genes outside of the assumption of a global shift in expression. We now develop theoretically the case of a few genes that do not hold the global shift assumption. We then wanted to know what happens if a few genes do not hold the proportionality assumption but rather the following one: $\mu_{g2} = \alpha\mu_{g1} + \partial_{g2}$ with $\partial_{g2} \neq 0$. In that case, we have that $S_2 = \alpha S_1 + \sum_{g=1}^{G} \partial_{g2} L_g$ and then, a priori, contrary to what has been seen above, we have this $S_2 \neq \alpha S_1$. Then, it follows that, for all genes *g*, the raw fold change of gene expressions in both conditions can be written as follows:

$$\mathrm{RFC}_g = \frac{X_{g2}/N_2}{X_{g1}/N_1} = \frac{\mu_{g2}}{\mu_{g1}}\frac{S_1}{S_2} = \frac{\alpha\mu_{g1} + \partial_{g2}}{\mu_{g1}}\frac{S_1}{S_2}.$$

Then, an RNA-Seq normalization procedure (such as TMM, RLE or MRN that were compared as in Scholes and Paige (2015)) leads to the following normalized fold change, denoted by NFC_g:

$$\mathrm{NFC}_g = \frac{\mathrm{RFC}_g}{\underset{g\in\{1,\ldots,G\}}{\mathrm{median}}\left(\mathrm{RFC}_g\right)} = \frac{\mu_{g2}}{\mu_{g1}}\frac{S_1}{S_2}\frac{S_2}{\alpha S_1} = \frac{1}{\alpha}\frac{\mu_{g2}}{\mu_{g1}}.$$

Finally, we obtain the estimation of the true fold change for a given gene g by the following simple formula:

$$\boxed{\frac{\mu_{g2}}{\mu_{g1}} = \alpha \times \mathrm{NFC}_g}. \qquad \text{Equation (1)}$$

In conclusion, the formula obtained above implies that if we know a priori the value of the global shift in expression α, we can still perform a classical differential expression (DE) analysis (with one of the classical normalization methods defined above) to estimate the fold change of each gene by multiplying the corresponding fold change by α. This means also that, by using the classical DE analysis method, a gene declared as not being DE will in fact have a fold change equal to α.

Testing the assumption of a global shift in expression for samples with different ploidy levels

In the case study presented here, in which ploidy level doubles from 4C to 8C, then from 8C to 16C, and from 16C to 32C, we have assumed a global shift in expression α = 2 following Bourdon *et al*. (2012). This shift was then validated following two approaches.

We first tested the assumption of a global shift in expression in the samples by comparing the concentration of total nuclear RNA obtained per 100,000 nuclei of different ploidy levels (Table 1A). With a null hypothesis stating that a ploidy level k+1 is twice more concentrated than a ploidy level k, none of the six tests performed was found to be significant, suggesting a global increase of whole transcription in relation to the ploidy level (Table 1B).

Second, we performed RT-qPCR analyses using a subset of genes to validate the assumption of the global shift in expression. In a first step, to identify non-DE and DE genes from the RNA-Seq dataset, based on our assumption of a global shift, we used the classical normalization method from DESeq2 R package (see Experimental Procedures). According to our assumption, non-DE genes

would correspond to genes having transcript levels proportional to the level of ploidy, while DE genes would not follow this proportionality. This analysis identified 1211 genes as being DE, between at least two samples, with a FDR of 0.01 (see details below). In a second step, we compared the fold changes obtained experimentally by the RT-qPCR method for a selected set of 10 genes declared as non-DE to the theoretical fold change obtained with Equation (1) : $\alpha^k = 2^k$ with k $\in \{1,2,3\}$ (Figure 3 and Supplemental Table S1). For this RT-qPCR analysis, a fifth DNA ploidy level, namely 64C, was included in the test. For proper estimation of transcript levels and to ensure unbiased RT-qPCR analysis, spike-in RNAs were added to each sample prior to RNA extraction (Risso *et al.*, 2014; see Experimental Procedures). We found that: (i) all genes tested displayed increasing fold change dynamics (grey lines) following the theoretical values (red lines); (ii) the means of these dynamics (black lines) were very close to the theoretical ones (red lines).

We then carried out statistical one-sample mean tests to compare these fold changes. Of all tests carried out for all ploidy level comparisons, only the test corresponding to the 64C vs 32C ploidy levels (Figure 3d) was statistically significant (with a p-value around 0.01). We therefore have nine tests out of ten that do not reject the null hypothesis stating that the mean of the fold changes agrees with the theoretical fold changes $\alpha^k = 2^k$ with k $\in \{1,2,3\}$ (Figures 3a, 3b and 3c). This analysis thus showed that the global shift in gene expression goes even beyond the DNA ploidy level of 32C.

We also compared the fold changes obtained by RT-qPCR of 25 genes declared as DE (12 up-regulated and 13 down-regulated genes) to their fold changes estimated according to Equation (1) (Figure 4 and Supplemental Table S1). Almost all comparisons were aligned near the first bisector (diagonal dashed line) which represents the fold changes for which both fold changes are identical. Among all 150 statistical two sample mean tests carried out to compare these fold changes, only 23 tests were significant (10 for down-regulated genes and 13 for up-regulated ones). We therefore have about 85% of the tests (87% for down-regulated and 82% for up-regulated genes) that did not reject the null hypothesis stating that the fold changes estimated by qPCR and by Equation (1) are equal.

In conclusion, these analyses indicated that there is probably a global proportional shift in expression depending on the ploidy level suggesting a positive effect of ploidy level on whole transcription, corresponding to a doubling in transcripts. Therefore, we further analyzed the RNA-Seq data, assuming a global shift in expression with a factor $\alpha = 2$, implying that non-DE genes will correspond to genes having an expression proportional to ploidy level.

Differentially expressed genes: distinct functional categories expressed as a function of ploidy

By analyzing the RNA-Seq data with the assumption of a global shift in expression between the samples, we found 24919 genes to be expressed in at least one sample across the samples with different ploidy levels. By pairwise comparison of the expression values, we found that 1211 genes were DE in at least one comparison with a FDR of 0.01 (Supplemental Table S2). The number of genes up- or down-regulated increased by comparing more divergent ploidy levels (Figure 5). After removing genes not expressed in all samples and genes for which variability was too high, 4694 genes were declared as being non-DE, i.e., displaying an expression that was proportional to the ploidy level.

To detect potential specific expression patterns within the DE genes and therefore identify the major transcriptional dynamics associated with changes in ploidy levels, a k-means clustering approach was used on the 1211 DE genes. Genes with similar expression profiles were grouped into six clusters (Figure 6 and Supplemental Table S3). Clusters 1 to 4 contained genes showing higher expression in the nuclei of low ploidy levels (4C and 8C). In Cluster 1, the expression was high in cells with 4C ploidy levels and decreased progressively when ploidy levels increased. In Cluster 2, the expression of genes was high only in cells with 4C ploidy and in Cluster 4 only in cells with 8C ploidy. In Cluster 3, genes showed high expression levels in both 4C and 8C nuclei. In contrast, in Clusters 5 and 6 the expression of genes increased with ploidy: in Cluster 5, genes had low expression in 4C nuclei and then higher and similar expression in 8, 16 and 32C nuclei, while in Cluster 6 genes had high expression in 16 and 32C nuclei.

For each cluster, we tested the overrepresentation of functional categories by using the GO enrichment tool from PLAZA (Proost *et al.*, 2015) (Figure 6 and Supplemental Table S4). To simplify the functional annotation, we also classified the DE genes into one of the 31 functional categories of tomato genes provided by Pattison *et al.* (2015) (Supplemental Table S3). In Cluster 1, we found genes mainly involved in photosynthesis, lipid and organic acid metabolic processes and cell wall related processes. In Cluster 2, genes involved in photosynthesis were also found together with genes involved in ATP metabolic processes and transport. In Clusters 3, 4 and 5, due to the small numbers of genes, fewer categories were significantly overrepresented. Finally, the genes in Cluster 6 mainly belonged to categories related to metabolic processes such as genes encoding UDP-glycosyltransferases involved in the glycosylation process, a prevalent modification of plant secondary metabolites, or genes involved in flavone metabolism. In conclusion, we found that DE genes display different expression profiles depending on the ploidy level and that for each pattern of expression, genes belong to distinct functional categories.

DISCUSSION

RNA-Seq from endoreduplicated nuclei

The tomato fruit provides an excellent model system to study the role of endoreduplication in the regulation of transcription and in the determination of organ size because the tissues in this fleshy fruit, such as the pericarp and the locular jelly-like tissue, display an extraordinary extent of endoreduplication (Joubès *et al.*, 1999; Cheniclet *et al.*, 2005). This increase in nuclear ploidy levels leading to DNA content as high as 512C induces an important increase in the nuclear volume and has been correlated to very large cells that can reach spectacular sizes such as thousands of times their initial volume (Cheniclet *et al.*, 2005). Because cells displaying high ploidy levels can reach such huge sizes, their isolation by fluorescence-activated cell sorting using flow cytometry is impossible. To circumvent this bottleneck, we decided to prepare RNA from sorted nuclei in order to generate transcript profiles for ranges of ploidy levels. The use of nuclear RNA for genome-wide transcriptomic analyses has been largely used for animal samples (Grinberg *et al.*, 2013; Krishnaswami *et al.*, 2016; Trask *et al.*, 2009), but to a lesser extent in plants (Zhang *et al.*, 2008; Deal and Henikoff, 2010). In

these studies, the comparison of the expression data between nuclei RNA and total cell RNA showed that although some differences exist between these two pools of RNA, the majority of the transcripts is equally represented. These findings indicate that the use of nuclar RNAs does not introduce major bias for gene expression studies. Therefore, using sorted nuclei based on their ploidy level provides probably a good representation of the transcriptional activity within the cells.

Performing RNA-Seq on samples displaying a global shift in expression

The purpose of our present work was to describe genome-wide gene expression profiles from tomato fruit cells with different ploidy levels, based on an RNA-Seq experiment. We first showed that, when a global shift in expression between two conditions exists, i.e. when all genes are proportionally expressed in the two given conditions, a classical RNA-Seq normalization approach, such as the widely used RLE and TMM normalization methods (Anders and Huber, 2010; Robinson and Oshlack, 2010; Maza *et al.*, 2013), is inadequate. Indeed, the normalization methods cited above are generally based on the assumption that less than half of the genes are down-regulated and less than half of the genes are up-regulated (Maza *et al.*, 2013; Maza, 2016) which is a priori not the case with ploidy-dependent conditions (Bourdon *et al.*, 2012). Such an issue, encountered whatever the organisms, tissues or conditions studied, has been tackled by Aanes *et al.* (2014) and Chen *et al.* (2015), as well as by Evans *et al.* (2016) in a more general framework. Nevertheless, these publications underline the issue of using RNA-Seq experiments for the analysis of transcriptomes that are, one from another, globally over- or under-expressed. A first solution to solve this issue is the use of external control genes named spike-ins (Chen *et al.*, 2015; Risso *et al.*, 2014) or sequins (Hardwick *et al.*, 2016). This method consists in adding a known amount of artificial sequences to the RNA samples to provide a scaling factor for between-samples normalization after quantification. Another solution for the analysis of a global shift in expression is introduced in the present work and consists in the use of Equation (1). Indeed, on the one hand, if we know a priori the value of the global shift in expression α, as is the case during endoreduplication, we can perform a classical DE analysis and then estimate the fold change of each gene by multiplying the corresponding fold change by α as in Equation (1). On the other hand, if the global shift in expression α is unknown, we can estimate it by RT-qPCR analysis of a subset of genes that have been declared non-DE using a classical approach. Obviously, the more genes tested, the more accurate the estimation of α.

Ploidy-dependent transcriptome may reflect different cellular functions within the pericarp

To approach the issue of ploidy-dependent gene expression inside the pericarp of immature green tomato fruits that have completed their growth, and as a basis to dissect the genetic regulation of endoreduplication in this tissue, we analyzed the expression profiles obtained from nuclei sorted based on their ploidy. At 30 dpa, the ploidy map representing the distribution of ploidy levels within the pericarp tissue of immature green fruits showed that cells with lower ploidy levels (2C-16C) are located mainly in the exocarp, while cells with high ploidy levels (32C-256C) locate mainly within the mesocarp, at the centre of the pericarp (Bourdon *et al.*, 2011). The 32C ploidy level is the most prominent one, present in all cell layers within the mesocarp and endocarp (Bourdon *et al.*, 2011).

The four levels of ploidy studied here (4C, 8C, 16C, 32C) may thus represent cells with various locations, and thus with various gene expression profiles. Interestingly, cells with the lowest levels of ploidy (4C and 8C) express genes related to photosynthesis and lipid synthesis, while cells with higher ploidy levels (16C and 32C) express genes related to carbohydrate metabolism (Figure 5).

At 30 dpa the fruit has stopped its growth, but has not yet started to ripen. Interestingly, our data suggest that the transcriptome of the largest cells (of high ploidy levels) at this developmental stage has already changed when compared to smaller cells (of low ploidy levels). From this ploidy distribution and the ploidy-specific expression profiles it is suggested that the activity of the cells is related to their ploidy level and tissue location. Therefore not only endoreduplication acts as a morphogenetic factor supporting cell growth during tomato fruit development according to the "karyoplasmic ratio" theory (Bourdon *et al.*, 2012; Chevalier *et al.*, 2014), but it is also likely to be an important determinant of cell identity, as demonstrated to be crucial for cell fate acquisition and maintenance (Bramsiepe *et al.*, 2010; Roeder *et al.*, 2010), thus contributing to the formation of specialized cell type patterns. The present transcriptomic analysis thus provides new information about the functional role of endoreduplication in orienting cell metabolic activities during specialization in the course of tomato fruit growth. Determining the ploidy distribution in the pericarp of growing fruits and overlapping it with ploidy-specific transcriptional information will be essential for understanding the transcriptional dynamics of fruit growth, for understanding the spatial and temporal specialization of cells.

Conclusion

We here showed that cell-based approaches to study expression profiles by RNA-Seq need to take into account the possible global shift in expression between samples for correct analysis of the data and for valid interpretation of the observed expression measurements. We indeed demonstrated that endoreduplication leads to a global (doubling) shift in gene expression. However, more than a thousand genes were found to be differentially expressed during the endoreduplication process, suggesting a ploidy- and/or cell size-specific gene expression programme. Since endoreduplication is a major source of polyploidization in Eukaryotes (Edgar and Orr-Weaver, 2001), our data thus provide recommendations and new information that may be useful for analysing transcriptomic data under the influence of ploidy levels not only in plants but in all types of Eukaryotic cells, tissues or organs harbouring developmentally programmed endoreduplication. As far as tomato fruit development is concerned, our ongoing project now intends to apply this methodology for a broader approach, analysing the transcriptome of nuclei-sorted samples based on a whole set of ploidy levels and prepared from pericarp tissue harvested at different developmental stages, to perform a full kinematic study of endoreduplication-influenced gene expression. This methodology now also opens the way to analyse such transcriptomic data from tomato fruit dissected tissues or selected parts of tissues.

EXPERIMENTAL PROCEDURES

Plant material and growing conditions

Cherry tomato (*Solanum lycopersicum* Mill. cv Wva106) plants were grown in a greenhouse under a thermoperiod of 25°C/20°C and a photoperiod of 14/10 hours (day/night). Fruits were harvested at immature green stage (30 days post anthesis, dpa), and 6 to 12 fruits served for nuclei sorting.

Nuclei sorting and RNA extraction

Tomato pericarp nuclei were prepared as described in Bourdon *et al.* (2011), and sorted by flow cytometry using with a MoFlo Astrios cytometer (Beckman Coulter, Roissy, France; www.beckmancoulter.com) using standard isotonic sheath at 25 psi (pound per square inch), 3 drop sorting and coincidence abort.

The flow cytometer-sorted nuclei were collected in 2 mL tubes containing TRIzol® reagent (Ambion, Life Technology) for subsequent RNA extraction with a ratio of 1:4 (vol. of nuclei suspension/vol. of TRIzol®). Since the MoFlo Astrios cytometer allowed obtain 320 nuclei per µL, we sorted 100,000 nuclei in a total volume of 312 µL, which were collected in tubes containing 937 µL of TRIzol® reagent. Tubes containing the nuclei suspension were mixed and frozen in liquid nitrogen and stored at -80°C.

Prior to RNA extraction, the nuclei suspension in TRIzol® was allowed to melt at room temperature. For the extraction of RNA used for qRT-PCR, a mix of four artificial poly(A^+) RNAs (ArrayControl™ Spots and Spikes, Ambion, Life Technologies) at 1.10^9 copy number was added in each nuclei suspension. Two hundred µL of chloroform were added; tubes were gently shaken and left at room temperature for 3 min. After a centrifugation at 12 000 *g* and 4°C for 15 min, the aqueous phase was transferred to 0.5 volume of EtOH 100%. After precipitation, the RNA was re-suspended and transferred to a RNeasy Mini Spin column Mini Kit (Qiagen) and purified following the manufacturer's protocol. The purified RNAs were then treated with 0.06 units.μL^{-1} of DNA-free TURBO DNase (Ambion, Life Technology) to remove genomic DNA. The quantification of purified nuclear RNA was accessed using a Qubit 3.0 Fluorometer and its integrity was analyzed using an Agilent 6000 Pico Kit on an Agilent 2100 Bioanalyser.

Library construction and Illumina RNA-Seq runs

To perform Illumina RNA-seq experiments as to investigate the effects of endoreduplication on gene expression, we extracted total nuclear RNA from a fixed number of nuclei (ca. 100,000 nuclei). According to ploidy levels, total nuclear RNA concentrations ranged from 0.44 to 7.27 ng.μL^{-1} (Table 1A). A volume of 22 µL for each sample was sent out to Fasteris Life Sciences SA (Plan-les-Ouates, Switzerland) for construction of mRNA-seq libraries following the TruSeq RNA Sample Prep Kit.

After cDNA synthesis, PCR amplification for 15 cycles (4C and 8C) or 18 cycles (16C and 32C) was performed prior to library construction. A total of 24 cDNA libraries were then constructed corresponding to 4 ploidy levels (4C, 8C, 16C and 32C) in 3 replicates, with duplex-specific nuclease (DSN) or without (no-DSN) treatment. The multiplex sequencing reaction was performed on the Illumina HiSeq 2000 machine.

RT-qPCR

Complementary DNAs of total nuclear RNAs were synthetized using SuperScript IV Reverse Transcriptase (Invitrogen Life Technologies) in the presence of random nonamers. The cDNA obtained were diluted (1/100) in distilled water. For real time PCR, a master mix for each PCR run was prepared with GoTaq® qPCR Master Mix (Promega, France). In a total volume of 20 µL, 0.2 µM of each PCR primer and 10 µL SYBR Green PCR Buffer were added to 5 µL of diluted cDNA. The following amplification program was used: 95°C 3 min, 40 cycles at 95 °C for 10 s followed by 60°C for 15 s. Three biological replicates corresponding to three different nuclei preparations for each level of ploidy were used; all samples were amplified in duplicate from the same RNA preparation and the mean value was considered. PCR amplifications were carried out using the CFX96 System (Bio-Rad). The real-time PCR efficiency was determined for each gene with the slope of a linear regression model using the spikein as reference genes and the Bio-Rad CFX Manager™ Software version 3.1. The list of primers used to amplify the genes for RNA-Seq confirmation and the spikein is provided in Supplemental Table S5.

Mapping and normalization of RNAseq data

Sequence quality check and pre-processing. Twenty-four samples were sequenced, 6 per ploidy level. Half of them were treated with duplex specific nuclease (DSN), in order to remove overrepresented RNA. A total of 259 045 386 100bp raw reads were processed through the quantification pipeline (Supplemental Figure S1). Briefly, sequence quality was verified using FASTQC (Andrews, 2010) and sequences were cleaned with TrimGalore (Krueger, 2012) that aims at removing remaining sequencing adaptors and trimming bad quality base pairs. This step removed about 2% of the sequences for a useful total of 258 972 491 reads.

Mapping. The cleaned reads were then mapped against Sly2.40 tomato reference genome with STAR (Dobin *et al.*, 2013), a splice aware aligner, fixing read length to 100 bp (corresponding to our reads) and leaving other parameters to default. The alignments were finally sorted by genome coordinates and merged by ploidy level with Samtools (Li *et al.*, 2009).

Mapping distribution. A more in depth analysis of uniquely mapped sequences has been performed with RNA-SeQC (DeLuca *et al.*, 2012). This quality control tool for RNA-seq data provides us with the read mapping distribution between genomic regions. This tool can also produce a BWA mapping against rRNA sequences to estimate the rRNA rate inside the intergenic regions (Supplemental figure 3).

Raw quantification. Merged alignments obtained from DSN treated sample were then processed with featureCounts (Liao *et al.*, 2014), an efficient read summarization program that counts mapped reads for genomic features such as genes or exons. This program requires an alignment file and an annotation file and we used ITAG2.30 annotation file.

DE analyses.

Differential Expression (DE) analyses have been performed with the DESeq2 package (Love *et al.*, 2014) in the R environment (Team R Core, 2014) using data obtained from DSN treated samples. The DESeq function from DESeq2 has been carried out with default parameters. It has been shown in Maza *et al.* (2013) and Maza (2016) that the LRE normalization method (used as default in DESeq2) is very similar to the TMM method used in the edgeR package (Robinson and Osklack, 2010; Robinson *et al.*, 2010).

Clustering of the differentially expressed genes and GO enrichment analysis

The differentially expressed genes were clustered according to the k-means clustering method using the R base stats package (Team R Core, 2014). All cluster plots were generated with the ggplot2 R-package (Wickham, 2009).

The GO enrichment analysis to determine the over-representation of certain GO terms in the DE gene set was done using the PLAZA tool with a cutoff of 0.05 (Proost *et al.*, 2015). This tool compares the occurrence of a certain GO label in a gene set with the occurrence in the genome. The significance of over-representation is determined using the hypergeometric distribution and the Bonferroni method is applied to correct for multiple testing.

Data access

The Illumina short reads generated in this study have been submitted to the European Nucleotide Archive (ENA) with study accession number PRJEB21753, and are available at http://www.ebi.ac.uk/ena.

ACKNOWLEDGEMENTS

This work was carried out with the financial support of the French Agence Nationale de la Recherche (grant no. ANR-09-GENM-105). The authors declare no conflict of interest.

SHORT SUPPORTING INFORMATION LEGENDS

The following materials are available in the online version of this article.

Supplemental Figure S1. RNA-Seq data processing pipeline for expression quantification.

Supplemental Table S1. Set of genes non-DE and Differentially Expressed (DE) genes selected for RT-qPCR.

Supplemental Table S2. List of differentially expressed genes in at least one comparison.

Supplemental Table S3. Classification into six expression clusters and functional category of the Differentially Expressed (DE) genes.

Supplemental Table S4. Overrepresentation of functional categories for each cluster of DE genes.

Supplemental Table S5. List of primers used for the RT-qPCR.

REFERENCES

Aanes, H., Winata, C.L., Lin, C.H., Chen, J.P., Srinivasan, K.G., Lee, S.G.P., Lim, A.Y.M., Hajan, H.S., Collas, P., Bourque, G. *et al.* (2011) Zebrafish mRNA sequencing deciphers novelties in transcriptome dynamics during maternal to zygotic transition. *Genome Res*. **21**, 1328–1338.

Aanes, H., Winata, C.L., Moen, L.F., Østrup, O., Mathavan, S., Collas, P., Rognes, T. and Aleström, P. (2014) Normalization of RNA-sequencing data from samples with varying mRNA levels. *PloS One* **9**, e89158.

Adachi, S., Minamisawa, K., Okushima, Y., Inagaki, S., Yoshiyama, K., Kondou, Y., Kaminuma, E., Kawashima, M., Toyoda, T., Matsui, M. *et al.* (2011) Programmed induction of endoreduplication by DNA double-strand breaks in Arabidopsis. *Proc. Natl Acad. Sci. USA* **108**, 10004–10009.

Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106.

Andrews, S. (2010) FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Bourdon, M., Coriton, O., Pirrello, J., Cheniclet, C., Brown, S.C., Poujol, C., Chevalier, C., Renaudin, J.-P. and Frangne, N. (2011) In planta quantification of endoreduplication using fluorescent in situ hybridization (FISH). *Plant J*. **66,** 1089–1099.

Bourdon, M., Pirrello, J., Cheniclet, C., Coriton, O., Bourge, M., Brown, S., Moïse, A., Peypelut, M., Rouyère, V., Renaudin, J.-P., Chevalier, C. and Frangne, F. (2012) Evidence for karyoplasmic homeostasis during endoreduplication and a ploidy-dependent increase in gene transcription during tomato fruit growth. *Development*, **139**, 3817–3826.

Bramsiepe, J., Wester, K., Weinl, C., Roodbarkelari, F., Kasili, R., Larkin, J.C., Hülskamp, M. and Schnittger, A. (2010) Endoreplication controls cell fate maintenance. *PLoS Genet*. **6**, e1000996.

Brodsky, W.Y. and Uryvaeva, I.V. (1977) Cell polyploidy: its relation to tissue growth and function. *Int. Rev. Cytol.* **50**, 275–332.

Carvalheira, G.M.G. (2000) Plant polytene chromosomes. *Genet. Mol. Biol*. **23**, 1043–1050.

Ceccarelli, M., Santantonio, E., Marmottini, F., Amzallag, G.N. and Cionini, P.G. (2006) Chromosome endoreduplication as a factor of salt adaptation in Sorghum bicolor. *Protoplasma***, 227**, 113–118.

Chen, K., Hu, Z., Xia, Z., Zhao, D., Li, W. and Tyler, J.K. (2015) The overlooked fact: fundamental need for spike-in control forvirtually all genome-wide analyses. *Mol. Cell Biol*. **3**, 662–667.

Cheniclet, C., Rong, W.Y., Causse, M., Frangne, N., Bolling, L., Carde, J.-P. and Renaudin, J.-P. (2005) Cell expansion and endoreduplication show a large genetic variability in pericarp and contribute strongly to tomato fruit growth. *Plant Physiol*. **139**, 1984–1994.

Chevalier, C., Nafati, M., Mathieu-Rivet, E., Bourdon, M., Frangne, N., Cheniclet, C., Renaudin, J.-P., Gévaudant, F., Hernould, M. (2011) Elucidating the functional role of endoreduplication in tomato fruit development. *Ann. Bot*. **107**, 1159–1169.

Chevalier, C., Bourdon, M., Pirrello, J., Cheniclet, C., Gévaudant, F. and Frangne, N. (2014) Endoreduplication and fruit growth in tomato: evidence in favour of the karyoplasmic ratio theory. *J. Exp. Bot.* **65,** 2731–2746.

Cookson, S.J., Radziejwoski, A. and Granier, C. (2006) Cell and leaf size plasticity in Arabidopsis: what is the role of endoreduplication? *Plant Cell Environ*. **29,** 1273–1283.

D'Amato, F. (1984) Role of Polyploidy in Reproductive Organs and Tissues. In Johri, P.B.M. (ed), Embryology of Angiosperms. Springer Berlin Heidelberg, pp. 519–566.

Deal, R.B. and Henikoff, S. (2010) A simple method for gene expression and chromatin profiling of individual cell types within a tissue. *Dev. Cell*, **18**, 1030–1040.

DeLuca, D.S., Levin, J.Z., Sivachenko, A., Fennell, T., Nazaire, M.-D., Williams, C., Reich, M., Winckler, W. and Getz, G. (2012) RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics*, **28**, 1530–1532.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics,* **29**, 15–21.

Edgar, B.A. and Orr-Weaver, T.L. (2001) Endoreplication cell cycles: more for less. *Cell*, **105**, 297–306.

Evans, C., Hardin, J. and Stoebel, D. (2016) Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. Cornell University Library. arXiv:1609.00959v1 [q-bio.GN.

Frawley, L.E. and Orr-Weaver, T.L. (2015) Polyploidy. *Curr. Biol.* **25**, R353-358.

Galitski, T., Saldanha, A.J., Styles, C.A., Lander, E.S. and Fink, G.R. (1999) Ploidy regulation of gene expression. *Science*, **285**, 251–254.

Grindberg, R.V., Yee-Greenbaum, J.L., McConnell, M.J., Novotny, M., O'Shaughnessy, A.L., Lambert, G.M., Araúzo-Bravo, M.J., Lee, J., Fishman, M., Robbins, E. *et al.* (2013) RNA-sequencing from single nuclei. *Proc. Natl Acad. Sci. USA,* **110**, 19802–19807.

Hardwick, S.A., Chen, W.Y., Wong, T., Deveson, I.W., Blackburn, J., Andersen, S.B., Nielsen, L.K., Mattick, J.S. and Mercer, T.R. (2016) Spliced synthetic genes as internal controls in RNA sequencing experiments. *Nat. Methods,* **13**, 792–798.

Hase, Y., Trung, K.H., Matsunaga, T. and Tanaka, A. (2006) A mutation in the uvi4 gene promotes progression of endo-reduplication and confers increased tolerance towards ultraviolet B light. *Plant J.* **46**, 317–326.

Hülskamp, M., Schnittger, A. and Folkers, U. (1999) Pattern formation and cell differentiation: trichomes in Arabidopsis as a genetic model system. *Int. Rev. Cytol*. **186**, 147–178.

Inzé, D. and De Veylder, L. (2006) Cell cycle regulation in plant development. Annu Rev Genet **40**: 77–105

Joubès, J., Phan, T.H., Just, D., Rothan, C., Bergounioux, C., Raymond, P. and Chevalier, C. (1999) Molecular and biochemical characterization of the involvement of cyclin-dependent kinase A during the early development of tomato fruit. *Plant Physiol.* **121**, 857–869.

Joubès, J. and Chevalier, C. (2000) Endoreduplication in higher plants. *Plant Mol. Biol.***43**, 735–745.

Krishnaswami, S.R., Grindberg, R.V., Novotny, M., Venepally, P., Lacar, B., Bhutani, K., Linker, S.B., Pham, S., Erwin, J.A., Miller, J.A. *et al.* (2016) Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons. *Nat. Protoc*. **11**, 499–524.

Krueger, F. (2012) Trim Galore! : A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/trim_galore.

Lee, H.O., Davidson, J.M. and Duronio, R.J. (2009) Endoreplication: polyploidy with purpose. *Genes Dev.* **23,** 2461–2477.

Lemaire-Chamley, M., Petit, J., Garcia, V., Just, D., Baldet, P., Germain, V., Fagard, M., Mouassite, M., Cheniclet, C. and Rothan, C. (2005) Changes in Transcriptional Profiles Are Associated with Early Fruit Tissue Specialization in Tomato. *Plant Physiol*. **139**, 750–769.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Liao, Y., Smyth, G.K. and Shi, W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.

Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. **15**, 550.

Matas, A.J., Yeats, T.H., Buda, G.J., Zheng, Y., Chatterjee, S., Tohge, T., Ponnala, L., Adato, A., Aharoni, A., Stark, R. *et al.* (2011) Tissue- and cell-type specific transcriptome profiling of expanding tomato fFruit provides insights into metabolic and regulatory specialization and cuticle formation. *Plant Cell*, **23**, 3893–3910.

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. Integ. Biol*. **6**, e25849.

Maza, E. (2016) In Papyro Comparison of TMM (edgeR), RLE (DESeq2), and MRN normalization methods for a simple two-conditions-without-replicates RNA-Seq experimental design. *Front. Genet.* **7**, 164.

Nagl, W. (1976) DNA endoreduplication and polyteny understood as evolutionary strategies. *Nature*, **261**, 614–615.

O'Neil, D., Glowatz, H. and Schlumpberger, M. (2013) Ribosomal RNA depletion for efficient use of RNA-seq capacity. *Curr. Protoc. Mol. Biol.* **Chapter 4**, Unit 4.19.

Pattison, R.J., Csukasi, F., Zheng, Y., Fei, Z., van der Knaap, E. and Catalá, C. (2015) Comprehensive Tissue-Specific Transcriptome Analysis Reveals Distinct Regulatory Programs during Early Tomato Fruit Development. *Plant Physiol*. **168**, 1684–1701.

Proost, S., Van Bel, M., Vaneechoutte, D., Van de Peer, Y., Inzé, D., Mueller-Roeber, B. and Vandepoele, K. (2015) PLAZA 3.0: an access point for plant comparative genomics. *Nucl. Acids Res.* **43**, D974-981.

Risso, D., Ngai, J., Speed, T.P. and Dudoit, S. (2014) Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol*. **32**, 896–902.

Robinson, M.D. and Oshlack, A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. **11**, R25.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Roeder, A.H.K., Chickarmane, V., Cunha, A., Obara, B., Manjunath, B.S. and Meyerowitz, E.M. (2010) Variability in the control of cell division underlies sepal epidermal patterning in Arabidopsis thaliana. *PLoS Biol.* **8**, e1000367.

Scholes, D.R. and Paige, K.N. (2015) Plasticity in ploidy: a generalized response to stress. *Trends Plant Sci.* **20**, 165–175.

Team R Core (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2015.

Trask, H.W., Cowper-Sal-lari, R., Sartor, M.A., Gui, J., Heath, C.V., Renuka, J., Higgins, A.-J., Andrews, P., Korc, M., Moore, J.H. *et al.* (2009) Microarray analysis of cytoplasmic versus whole cell RNA reveals a considerable number of missed and false positive mRNAs. *RNA*, **15**, 1917–1928.

Vriezen, W.H., Feron, R., Maretto, F., Keijman, J. and Mariani, C. (2008) Changes in tomato ovary transcriptome demonstrate complex hormonal regulation of fruit set. *New Phyto*l. **177**, 60–76.

Wang, H., Schauer, N., Usadel, B., Frasse, P., Zouine, M., Hernould, M., Latché, A., Pech, J.-.C, Fernie, A.R. and Bouzayen, M. (2009) Regulatory Features Underlying Pollination-Dependent and -Independent Tomato Fruit Set Revealed by Transcript and Primary Metabolite Profiling. *Plant Cell*, **21**, 1428–1452.

Wickham, H. (2009) ggplot2: Elegant graphics for data analysis. Springer-Verlag New York.

Ye, J., Hu, T., Yang, C., Li, H., Yang, M., Ijaz, R., Ye, Z. and Zhang, Y. (2015) Transcriptome profiling of tomato fruit development reveals transcription factors associated with ascorbic acid, carotenoid and flavonoid biosynthesis. *PloS One* **10**, e0130885.

Zhao, W., He, X., Hoadley, K.A., Parker, J.S., Hayes, D.N. and Perou, C.M. (2014) Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genomics*, **15**, 419.

Zhang, C., Barthelson, R.A., Lambert, G.M. and Galbraith, D.W. (2008) Global characerization of cell-specific gene expression through fluorescence-activated sorting of nuclei. *Plant Physiol*. **147**, 30–40.

Zhong, S., Fei, Z., Chen, Y.-R., Zheng, Y., Huang, M., Vrebalov, J., McQuinn, R., Gapper, N., Liu, B., Xiang J. *et al.* (2013) Single-base resolution methylomes of tomato fruit development reveal epigenome modifications associated with ripening. *Nat. Biotechnol*. **31**, 154–159.

Table 1. Statistical one-sample t-tests on total nuclear RNA concentrations. (A) Total nuclear RNA concentrations from four sample replicates (in ng.µL^{-1}) using a constant number of nuclei (ca. 100,000). (B) T-tests on concentration ratios with a null hypothesis that a ploidy level k+1 is twice more concentrate than a ploidy level k.

A.

Ploidy level	Replicate 1	Replicate 2	Replicate 3	Replicate 4
4C	0.91	0.62	0.50	0.44
8C	1.31	1.10	2.28	2.12
16C	3.29	2.08	1.64	2.84
32C	7.27	4.98	3.89	4.60

B.

Ploidy level comparison	Ratio 1	Ratio 2	Ratio 3	Ratio 4	Null Hypothesis	p-value
8C / 4C	1.43	1.77	4.56	4.82	2.00	0.29
16C / 8C	2.51	1.89	0.72	1.34	2.00	0.39
32C / 16C	2.51	2.39	2.37	1.62	2.00	0.47
16 / 4C	3.60	3.35	3.28	6.45	4.00	0.84
32C / 8C	5.55	4.53	1.71	2.17	4.00	0.62
32C / 4C	7.96	8.03	7.78	10.45	8.00	0.45

Cite this article as:
Pirrello, J., Deluche, C., Frangne, N., Gevaudant, F., Maza, E., Djari, A., Bourge, M., Renaudin, J.-P., Brown, Bowler, C., Zouine, M., Chevalier, C., Gonzalez, N. (2018). Transcriptome profiling of sorted endoreduplicated nuclei from tomato fruits: how global shift in expression ascribed to DNA ploidy influences RNA-Seq data normalization and interpretation. Plant Journal, 93

FIGURES LEGENDS

Figure 1. Description of the biological material used for nuclei sorting.
(a) Cross-section of a tomato fruit at the Mature Green (MG) stage (Bar = 5 mm). (b) Cross-section of 30 dpa-fruit pericarp illustrating the cellular composition (Bar = 200 µm). (c) Bivariate analysis of endoreduplicated nuclei isolated from pericarp tissue of 30 dpa-fruits using flow cytometry. Nuclei were stained with 5 $\mu g.mL^{-1}$ DAPI to measure DNA content. Side Scatter (SSC) expresses granularity of events.

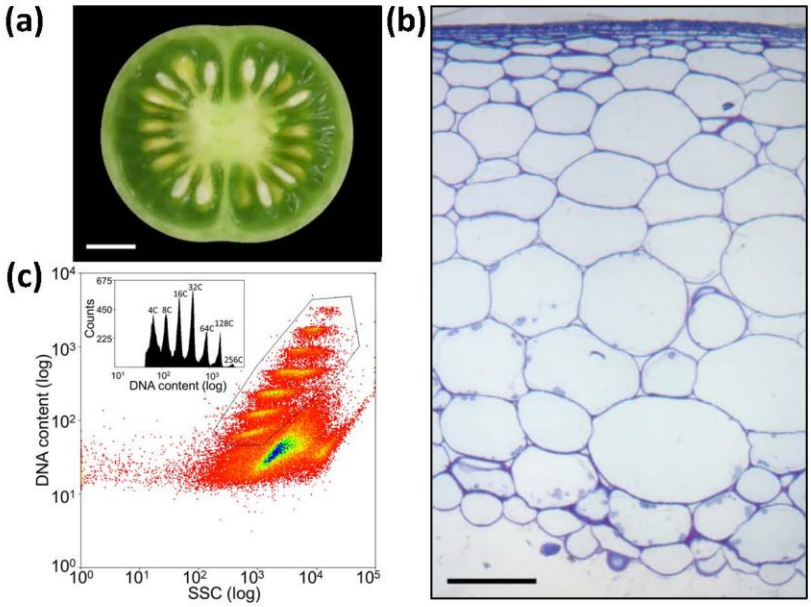
Figure 2. RNASeq general statistic for non-DSN and DSN-treated samples and mapping distribution.
(a) Distribution of uniquely mapped, multi-mapped and unmapped reads for non-DSN and DSN-treated samples and for each ploidy level. (b) Distribution of reads, using ITAG2.30 annotation, into intergenic, intronic, exonic and ribosomal regions for each ploidy level of DSN-treated samples.

Figure 3. Expression obtained by RT-qPCR of selected genes declared as being not Differentially Expressed (non-DE).
Grey dots represent the (log_2) fold changes of the qPCR expressions of 10 genes that have been declared non DE with DESeq2 (y-axis) depending on the ploidy level (x-axis). Panels (a), (b), (c), and (d) represent respectively the comparisons of levels 8C, 16C, 32C and 64C vs 4C (panel (a)), 16C, 32C and 64C vs 8C (panel (b)), 32C and 64C vs 16C (panel (c)), and 64C vs 32C (panel (d)). Black plus and red multiplication symbols represent respectively the observed (log_2) mean and the theoretical ploidy level (1, 2, 3 or 4). A blue star is added between plus and multiplication symbols that are statistically different (by a t-test with a type I error of 5%).

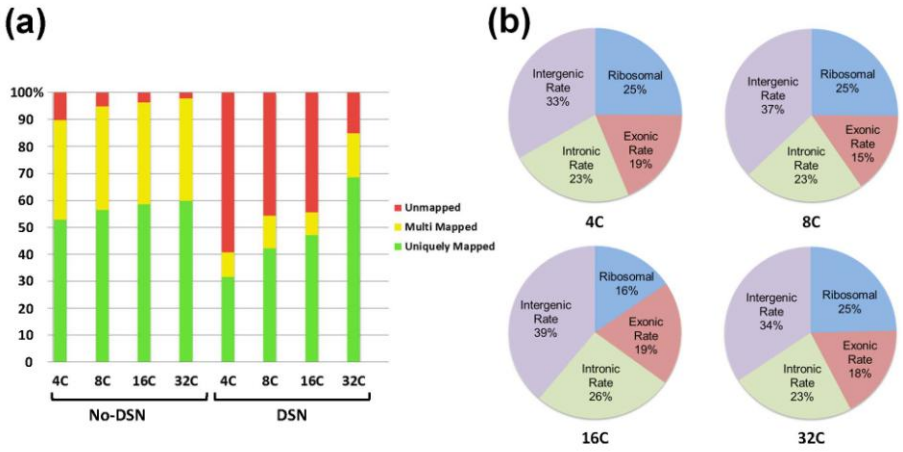
Figure 4. Expression obtained by RT-qPCR of selected genes declared as being up- or down-regulated as a function of ploidy.
Each triangle represents a gene declared as Differentially Expressed (DE) with DESeq2. The x-axis and y-axis represent, respectively, the (log_2) fold changes of gene expressions measured by qPCR and by DESeq2 multiplied by $\alpha^k = 2^k$ with $k \in \{1,2,3\}$ the ploidy level in keeping with Equation (1). Down-pointing and up-pointing triangles represent, respectively, genes that are declared down-regulated and up-regulated by DESeq2. Black, green and blue colors represent DE fold changes associated respectively with the three ploidy levels $k \in \{1,2,3\}$. The diagonal dashed line is the first bisector representing the genes for which both fold changes are identical.

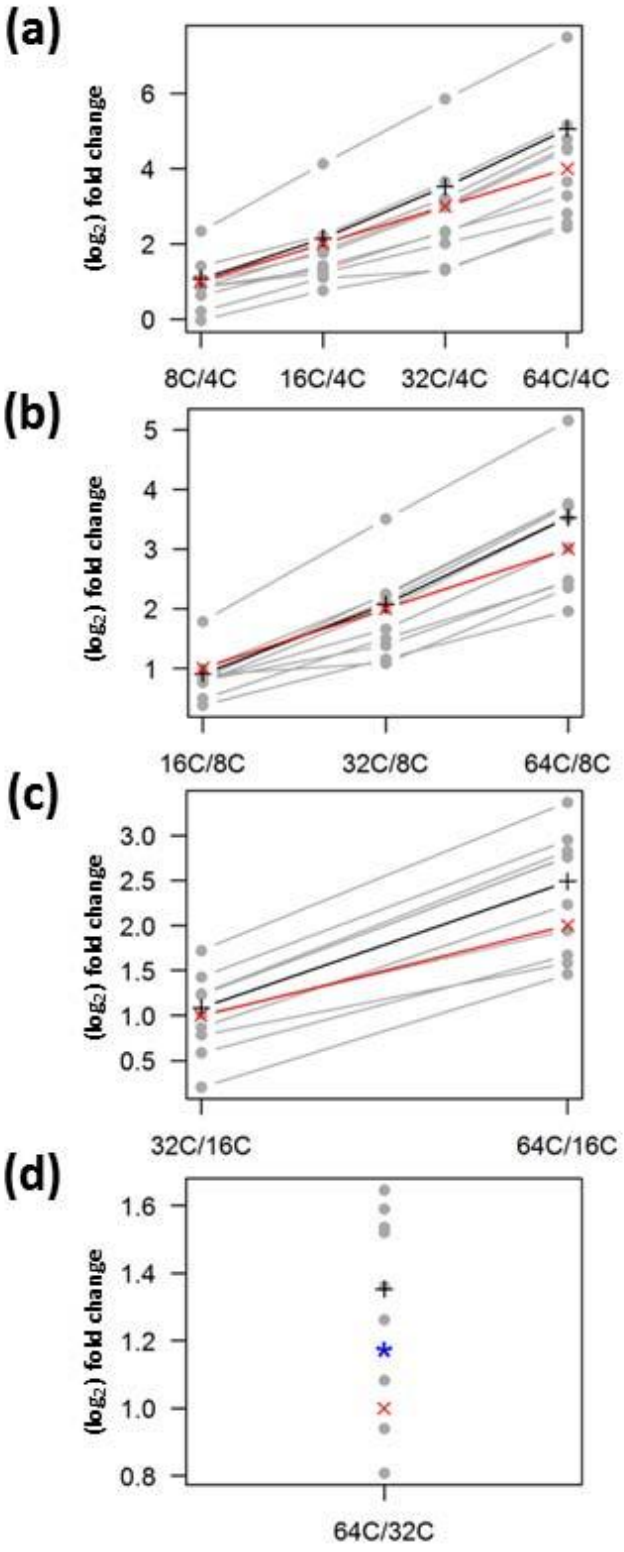
Figure 5. Number of genes up- (yellow) and down- (blue) regulated in pairwise comparisons of the expression values (FDR < 0.01).

Figure 6. Clustered expression profiles of DEGs and functional categories. Six clustered expression profiles of DEGs obtained with a k-means clustering approach.
The major overrepresented functional categories, identified by using the GO enrichment tool from PLAZA (Zhao *et al.*, 2014), are indicated on the right side of each cluster.

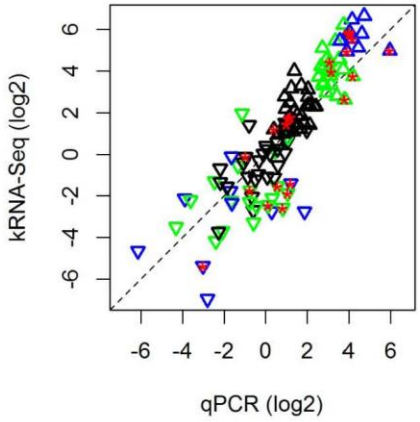


Pirrello, J., Deluche, C., Frangne, N., Gevaudant, F., Maza, E., Djari, A., Bourge, M., Renaudin, J.-P., Brown, Bowler, C., Zouine, M., Chevalier, C., Gonzalez, N. (2018). Transcriptome profiling of sorted endoreduplicated nuclei from tomato fruits: how global shift in expression ascribed to DNA ploidy influences RNA-Seq data normalization and interpretation. Plant Journal, 93

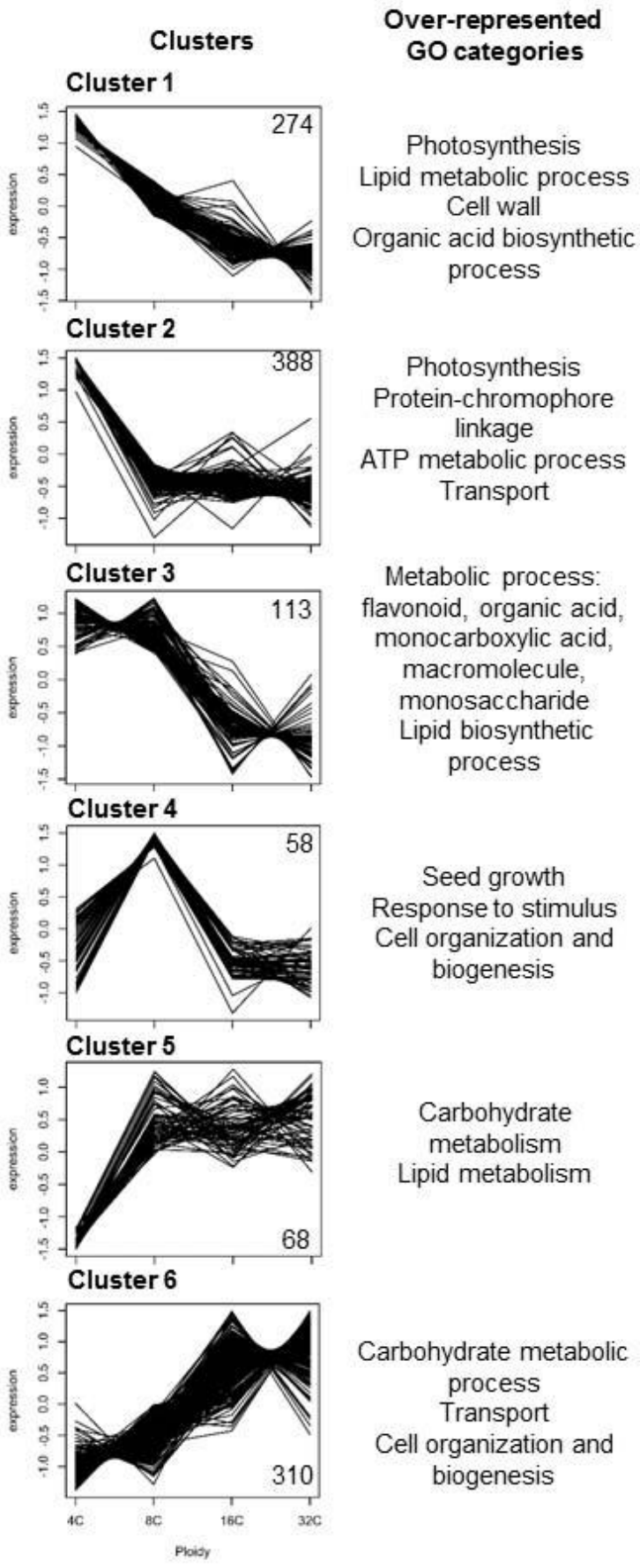




Pirrello, J., Deluche, C., Frangne, N., Gevaudant, F., Maza, E., Djari, A., Bourge, M., Renaudin, J.-P., Brown, Bowler, C., Zouine, M., Chevalier, C., Gonzalez, N. (2018). Transcriptome profiling of sorted endoreduplicated nuclei from tomato fruits: how global shift in expression ascribed to DNA ploidy influences RNA-Seq data normalization and interpretation. Plant Journal, 93



	4C	8C	16C	32C
4C	0	47	254	325
8C	45	0	26	71
16C	525	69	0	0
32C	715	270	12	0



Pirrello, J., Deluche, C., Frangne, N., Gevaudant, F., Maza, E., Djari, A., Bourge, M., Renaudin, J.-P., Brown, Bowler, C., Zouine, M., Chevalier, C., Gonzalez, N. (2018). Transcriptome profiling of sorted endoreduplicated nuclei from tomato fruits: how global shift in expression ascribed to DNA ploidy influences RNA-Seq data normalization and interpretation. Plant Journal, 93