New Phytologist Supporting Information

Article title: Histone Post Translational Modifications rather than DNA methylation underlie gene reprogramming in pollination-dependent and pollination-independent fruit set in tomato
Authors: Guojian Hu, Baowen Huang, Keke Wang, Pierre Frasse, Elie Maza, Anis Djari, Moussa Benhamed, Philippe Gallusci, Zhengguo Li, Mohamed Zouine, Mondher Bouzayen.
Article acceptance date: 10 August 2020

The following Supporting Information is available for this article:

**Fig. S1** Cluster dendrogram of DNA cytosine methylation level in 0DPA, 4DPA and 4IAA samples.

**Fig. S2** DNA methylation profiles in promoter of GA and auxin-related genes differentially expressed during fruit set.

**Fig. S3** Expression profile of genes involved in the regulation of DNA methylation during fruit set process.

**Fig. S4** ChIP-qPCR validation of ChIP-seq data.

**Fig. S5** Genomic distribution of H3K9ac, H3K4me3 and H3K27me3.

**Fig. S6** Identification of histone modified regions in 4IAA sample.

**Fig. S7** GO terms enriched in RNA-seq and ChIP-Seq datasets.

**Fig. S8** Mean normalized expression of the six histone modifier genes selected.

**Fig. S9** Generation of *SISDG16* knock-out lines by CRISPR/Cas9 technology.

**Table S1** Primers used in this study.

**Table S2** Read mapping summary for RNA-seq, ChIP-seq, BS-seq libraries.

**Table S3** DA regions identified by MAnorm method.

**Table S4** List of genes related to epigenetic regulation in tomato.

**Table S5** List of differentially expressed genes (DEGs)
Table S6 List of GO enrichment biological processes from DEGs (p-value<0.05).

Table S7 Identification of DMRs in CG, CHG and CHH sequence context.

Table S8 Size of associated region for each mark and percent of the tomato genome covered by the three histone marks.

Table S9 Differential expression of important fruit set-related genes and their differential association with histone marks.

Table S10 Gene Ontology analysis of DEGs and active mark associated-DA genes.

Table S11 Hormone-related DEGs and DA genes.

Table S12 Putative bivalent genes associated with both H3K9ac/H3K4me3 and H3K27me3.

Methods S1 Processing of RNA-seq, ChIP-seq and RNA-seq data.
Supporting Figures

Fig. S1 Cluster dendrogram of DNA cytosine methylation level in 0DPA, 4DPA and 4IAA samples. (a) Cluster dendrogram of DNA cytosine methylation level in 0DPA, 4DPA and 4IAA samples with two independent replicates for each sample. Dendrograms were generated by hierarchically clustering samples based on distance values with ‘canberra’ method. Fifty million cytosines were randomly selected for cluster analysis. (b) The distribution of identified DMRs in genomic regions.
Fig. S2 DNA methylation profiles in promoter of GA and auxin-related genes differentially expressed during fruit set. Vertical lines represent cytosine methylation level in 0DPA (dark green), 4DPA (light green) and 4IAA (turquoise) samples. Transcript read counts were profiled in red, and the corresponding gene locus were marked in blue. The DMR regions were zoomed individually in bottom panels.
Fig. S3 Expression profile of genes involved in the regulation of DNA methylation during fruit set process. Left panel shows log2Fold change in gene expression during pollination-dependent (4DPA vs 0DPA) and auxin-induced (4IAA vs 0DPA) as assessed in the present study. Right panel shows normalized expression level obtained using TomExpress platform (http://gbf.toulouse.inra.fr/tomexpress), in several developmental stages and tissue types during fruit set from wild species *S. pimpinellifolium* and *S. lycopersicum*. 
Fig. S4 ChIP-qPCR validation of ChIP-seq data. ChIP experiments were performed with Mock (no antibody), IgG, or H3K9ac, H3K4me3 and H3K27me3 antibodies. Nine genes showing differentially association with histone marks in the ChIP-seq experiment were selected for validation by Real-time PCR using independent replicates. The results of the ChIP-qPCR are presented in (a) for H3K9ac, (b) H3K4me3, and (c) for H3K27me3. Enrichments refer to the percentage of input in 0DPA (light green) and 4DPA (dark green) samples. The sequences of the primers used in the ChIP-qPCR experiment are provided in Supplemental Table 12. (d) Correlations between ChIP-seq and ChIP-qPCR results are provided for the three histone marks.
Fig. S5 Genomic distribution of H3K9ac, H3K4me3 and H3K27me3. (a) Distribution of histone marks in the 12 tomato chromosomes. Peak regions associated with each histone mark were positioned in the tomato genome (ITAG2.3). Pink shaded regions represent euchromatin, green shaded regions heterochromatin region, and blue bars represent centromeres. (b) Proportion (percent) of peak regions overlapping genes. Y-axis: percentage of peak regions covering a given number of genes.
Fig. S6 Identification of histone modified regions in 4IAA sample. (a) Number of identified regions for H3K9ac, H3K4me3 and H3K27me3. (b) Length distribution of the peak regions. The outliers are represented by black dots. (c) Frequency of peaks associated with genic and intergenic regions. A region spanning 1.5kb upstream of the annotated transcription start site (TSS) to 0.5kb downstream of transcription end site (TES) was designated as genic region. The regions between gene territories were designated as the intergenic regions. (d) Proportion (percent) of peak regions overlapping genes. Y-axis: percentage of peak regions covering a given number of genes. (e) Association profile of input (grey), H3K9ac (yellow), H3K4me3 (red) and H3K27me3 (blue) in genic regions within 4IAA samples. The gene set is adapted from publicly available RNA end-sequencing data (Zhong et al., 2013) which defines the TSS and TES. Mean counts within 100bp window covering a total of 2.5kb upstream to 2.5kb downstream the TSS (left panel) and 2.5kb upstream to 2.5kb downstream the TES (right panel) were extracted and plotted.
Fig. S7 GO terms enriched in RNA-seq and ChIP-Seq datasets. Plots showing the correlation between significant GO terms identified in RNA-seq and ChIP-Seq experiments (H3K9ac or H3K4me3 active marks considered only). BH-adjusted p-value<0.05 was used to filter the significantly enriched biological processes. Dash lines indicate the boundary of p-value=0.05. The enriched GO terms were dotted with different colors (blue for RNA-seq, green for ChIP-seq and red for both RNA-seq and ChIP-seq). Highly enriched GO terms common to the two fruit set processes were labeled and annotated aside the graph (the full list is provided in Supplemental Table 9).
Fig. S8 Mean normalized expression of the six histone modifier genes selected. The tissue types indicated below the chart are based on TomExpress annotation. The expression value of the selected genes in each tissue or organ corresponding to different experiments is averaged and plotted.
Fig. S9 Generation of *SISDG16* knock-out lines by CRISPR/Cas9 technology. Two guides (sgRNA1 and sgRNA2; green bars) were designed for editing the target gene. Protospacer-adjacent motifs (PAMs) are indicated in blue letters. Mutated sequences of *SISDG16* showing sequence gaps (red slashes). The corresponding changes in protein sequence are schematically illustrated in the middle panel. The mutant CR#SISDG16-530/1 corresponds to three base deletion leading to a loss of one amino acids without affecting the remaining part of the protein. The mutant CR#SISDG16-530/2 corresponding to 12 base deletion resulting in a frame shift with a predicted truncated protein lacking the SET domain. The corresponding changes in protein sequence are showed in the bottom panel. The SET domain is outlined by an orange frame.
Supporting Tables

Table S1 Primers used in this study.

Table S2 Read mapping summary for RNA-seq, ChIP-seq, BS-seq libraries

Table S3 DA regions identified by MAnorm method.

Table S4 List of genes related to epigenetic regulation in tomato.

Table S5 List of differentially expressed genes (DEGs).

Table S6 List of GO enrichment biological processes from DEGs (p-value<0.05).

Table S7 Identification of DMRs in CG, CHG and CHH sequence context.

Table S8 Size of associated region for each mark and percent of the tomato genome covered by the three histone marks.

| Stage | Mark     | Size (bp) | percentage of genome |
|-------|----------|-----------|----------------------|
| 0DPA  | H3K9ac   | 44,095,124| 5.28%                |
|       | H3K4me3  | 36,451,471| 4.36%                |
|       | H3K27me3 | 54,000,541| 6.46%                |
| 4DPA  | H3K9ac   | 55,976,284| 6.70%                |
|       | H3K4me3  | 39,698,620| 4.75%                |
|       | H3K27me3 | 52,805,348| 6.32%                |
| 4IAA  | H3K9ac   | 53,006,005| 6.34%                |
|       | H3K4me3  | 41,271,786| 4.94%                |
|       | H3K27me3 | 56,740,305| 6.79%                |

Table S9 Differential expression of important fruit set-related genes and their differential association with histone marks.

Table S10 Gene Ontology analysis of DEGs and active mark associated-DA genes.
Table S11 Hormone-related DE and DA genes. For DA genes only changes corresponding to H3K9ac or H3K4me3 active histone marks are taking into consideration in this table.

| Hormone   | DEG<sup>a</sup> | DA<sup>b</sup> | DEG & DA |
|-----------|-----------------|----------------|----------|
|           | Up   | Down | Gain | Loss | Up & Gain | Down & Loss | % of DE |
| 4DPAvs0DPA|       |      |      |      |           |             |         |
| Auxin     | 29   | 15   | 30   | 13   | 24        | 8            | 72.70%  |
| GA        | 8    | 8    | 6    | 7    | 5         | 6            | 68.80%  |
| Ethylene  | 11   | 45   | 19   | 40   | 8         | 36           | 78.60%  |
| CK        | 11   | 22   | 12   | 25   | 7         | 18           | 75.80%  |
| BR        | 10   | 1    | 11   | 2    | 10        | 1            | 100%    |
| ABA       | 10   | 14   | 13   | 15   | 9         | 12           | 87.50%  |
| JA        | 4    | 13   | 5    | 11   | 3         | 8            | 64.70%  |
| SA        | 14   | 5    | 13   | 4    | 12        | 3            | 78.70%  |
| 4IAAvs0DPA|       |      |      |      |           |             |         |
| GA        | 6    | 7    | 6    | 9    | 4         | 6            | 76.90%  |
| Ethylene  | 16   | 34   | 26   | 31   | 13        | 20           | 66.00%  |
| CK        | 7    | 26   | 11   | 25   | 6         | 21           | 81.80%  |
| BR        | 11   | 1    | 11   | 2    | 11        | 1            | 100%    |
| ABA       | 19   | 6    | 13   | 12   | 10        | 3            | 52.00%  |
| JA        | 6    | 7    | 10   | 9    | 4         | 4            | 61.50%  |
| SA        | 5    | 4    | 12   | 5    | 4         | 3            | 77.80%  |

<sup>a</sup> DEG differentially expressed genes with *p*-value<0.01, fold change>2;
<sup>b</sup> DA genes associated with gain/loss of either H3K9ac or H3K4me3 histone marks with *p*-value<0.01;

Table S12 Putative bivalent genes associated with both H3K9ac/H3K4me3 and H3K27me3.
Supporting Methods

Methods S1 RNA-seq, ChIP-seq and BS-seq data processing.

RNA-seq data processing

Raw paired-end RNA-seq sequences in FASTQ format were analyzed. Low quality reads were removed with FASTQ quality filter by FASTX toolkit version 0.0.13 (http://hannonlab.cshl.edu/fastx_toolkit/). Trimmed reads were then mapped to the *S. lycopersicum* reference genome and gene annotation (ITAG2.3 (Tomato_Genome_Consortium, 2012), https://solgenomics.net/) using TopHat-2.0.14 (Trapnell et al., 2009) calling bowtie 2.1.0 (Langmead & Salzberg, 2012). To perform differential gene analysis, HTSeq (Anders et al., 2015) was used to calculate raw counts. Raw counts of 34,727 tomato genes were normalized and mean counts per kilobase of transcript was set as gene expression. Differentially expressed genes between 0DPA and 4DPA or 4IAA tissues were identified with DEseq2 (Love et al., 2014). Raw p-value were adjusted as ‘padj’ by multiple tests using methods of Benjamini and Hochberg (1995). Genes with padj <0.01 were defined as significantly differentially expressed genes.

ChIP-seq data processing

ChIP-seq read alignment was performed using Bowtie2 with default parameters, and only uniquely aligned reads were retained. Enriched regions in the non-redundant mapped reads were identified by MACS2 v1.4.2 (Zhang et al., 2008) (effective genome size=770 Mb, pvalue cutoff = 1.00e-05). Heatmap representations of signal intensity (computeMatrix scale-regions followed by plotHeatmap) were generated using deepTools suite (Ramírez F, Dündar F, Diehl S, Grüning BA, 2014). BEDtools package (Quinlan & Hall, 2010) was used for detecting the tomato genes (ITAG2.3) overlapping with the detected peaks. A matrix of genes intersected with peaks for every sample was created for downstream analyses by R software (www.r-project.org/). Differentially associated peaks were normalized and identified using the ‘MAnorm’ method (Shao et al., 2012). The normalized M value \( M = \log_2 (\text{read density in 4DPA (or 4 IAA) samples / read density in 0DPA sample}) \) represents log2-transformed fold changes of enrichment intensities at
each peak region. Only those regions with \( p\text{-value} < 0.01 \) were defined as differentially associated (DA) (See Table S3).

**BS-seq data processing**

Pair-end BS-seq reads were first trimmed with trim_Galore, and aligned to the tomato genome (SL2.40) by calling Bowtie2 using Bismark software (v0.14.3) with minor changes to parameters \(--\text{score\_min} \ L,0,-0.6\) (Krueger & Andrews, 2011). After removing duplicated reads, methylated and unmethylated cytosines were identified by bismark methylation extractor. Methylation levels at each cytosine were calculated as \(#C/(#C+#T)\) in CG, CHG and CHH contexts separately. DMRs were then defined with R package DMRcaller (Catoni et al., 2018) using a “bins” approach, and incorporating biological replicates with beta regression test. DMRs were detected in 100bp bins with at least 4, 4 and 8 cytosines counted inside for CG, CHG and CHH sequence context, respectively, and each cytosine with at least 4 reads being detected. Nearby DMRs were merged if they were less than 50bp distant. Weighted methylation level in each DMR showing 30%, 20%, 10% difference for CG-, CHG- and CHH-DMRs, respectively, and with a \( p\text{-value} < 0.01 \) were considered for further analysis. P-values were corrected for false discovery rate (FDR) using Benjamini and Hochberg method (Benjamini and Hochberg, 1995). Finally, the CG/CHG/CHH-DMRs were overlapped with distinct gene tracks for further analysis.

**References**

Anders S, Pyl PT, Huber W. 2015. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31: 166–169.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J.R.Stat.Soc.Ser.B Stat.Methodol. 57*: 289–300.

Catoni M, Tsang JMF, Greco AP, Zabet NR. 2018. DMRcaller: a versatile R/Bioconductor package for detection and visualization of differentially methylated regions in CpG and non-CpG contexts. *Nucleic Acids Research* 46: e114.

Krueger F, Andrews SR. 2011. Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 27: 1571–1572.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9: 357–359.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* 15: 550.

Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26: 841–842.
Ramírez F, Dündar F, Diehl S, Grüning BA MT. 2014. DeepTools: A flexible platform for exploring deep-sequencing data. *Nucleic Acids Research* 42: W187–W191.

Shao Z, Zhang Y, Yuan G-C, Orkin SH, Waxman DJ. 2012. MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. *Genome Biology* 13: R16.

Tomato_Genome_Consortium. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485: 635–41.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* 25: 1105–1111.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* 9: R137.