Enhancing resolution of natural methylome reprogramming behavior in plants

Robersy Sanchez[1†], Xiaodong Yang[1†], Hardik Kundariya[2], Jose R Barreras[1], Yashitola Wamboldt[2], and Sally A. Mackenzie[1*]

[1]Departments of Biology and Plant Science, The Pennsylvania State University, University Park, PA 16802 and [2]Department of Agronomy and Horticulture, University of Nebraska, Lincoln, NE 68588

[†]Equal contributors

*Correspondence to: sam795@psu.edu

Sally Mackenzie
362 Frear North Bldg
Pennsylvania State University
University Park, PA 16802
Ph 814-863-8324, email sam795@psu.edu

Robersy Sanchez: rus547@psu.edu
Xiaodong Yang: xiaodongy86@gmail.com
Hardik Kundariya: kundariyahardik@gmail.com
Jose R Barreras: barreras@gmail.com
Yashitola Wamboldt: yashitola@yahoo.com
Sally Mackenzie: sam795@psu.edu

Abstract

We have developed a novel methylome analysis procedure, Methyl-IT, based on information thermodynamics and signal detection. Methylation analysis involves a signal detection problem, and the method was designed to discriminate methylation regulatory signal from background noise induced by thermal fluctuations. Comparison with three commonly used programs and various available datasets to furnish a comparative measure of resolution by each method is included. To confirm results, methylation analysis was integrated with RNAseq and network enrichment analyses. Methyl-IT enhances resolution of genome methylation behavior to reveal network-associated responses, offering resolution of gene pathway influences not attainable with previous methods.

Keywords

Epigenomics, DNA methylation, gene expression, information theory, Arabidopsis, tomato

Background

Most chromatin changes that are associated with epigenetic behavior are reprogrammed each generation, with the apparent exception of cytosine methylation, where parental patterns can be inherited through meiosis [1]. Genome-wide methylome analysis, therefore, provides one avenue for investigation of transgenerational and developmental epigenetic behavior. Complicating such investigations in plants is the dynamic nature of DNA methylation [2, 3] and a presently incomplete understanding of its association with gene expression. In plants, cytosine methylation is generally found in three contexts, CG, CHG and CHH (H=C, A or T), with CG most prominent within gene body regions [4]. Association of CG gene body methylation with changes in gene expression remains in question. There exist ample data associating chromatin behavior with plant response to environmental changes [5], yet, affiliation of genome-wide DNA methylation with these effects, or their inheritance, remains inconclusive [6, 7].

The epigenetic landscape is modulated by thermodynamic fluctuations that influence DNA stability. Most genome-wide methylome studies have relied predominantly on statistical approaches that ignore the subjacent biophysics of cytosine DNA methylation, offering limited resolution of those genomic regions with highest probability of having undergone epigenetic change. Jenkinson and colleagues [8] described the implementation of statistical physics and information theory to the analysis of whole genome methylome data to define sample-specific energy landscapes. Our group [9, 10] has proposed an information thermodynamics approach to investigate genome-wide methylation patterning based on the statistical mechanical effect of methylation on DNA molecules. The information thermodynamics-based

approach is postulated to provide greater sensitivity for resolving true signal from thermodynamic background within the methylome [9]. Because the biological signal created within the dynamic methylome environment characteristic of plants is not free from background noise, the approach, designated Methyl-IT, includes application of signal detection theory [11-14].

A basic requirement for the application of signal detection is a probability distribution of the background noise. Probability distribution, as a Weibull distribution model, can be deduced on a statistical mechanical/thermodynamics basis for DNA methylation induced by thermal fluctuations [9]. Assuming that this background methylation variation is consistent with a Poisson process, it can be distinguished from variation associated with methylation regulatory machinery, which is non-independent for all genomic regions [9]. An information-theoretic divergence to express the variation in methylation induced by background thermal fluctuations will follow a Weibull distribution model, provided that it is proportional to minimum energy dissipated per bit of information from methylation change.

The information thermodynamics model was previously verified with more than 150 Arabidopsis and more than 90 human methylome datasets [9]. To test application of the Methyl-IT method to methylome analysis, and to compare resolution of the Methyl-IT approach to publicly available programs DSS [15], BiSeq [16] and Methylpy [17], we used three Arabidopsis methylome datasets. Genome-wide methylation data from a Col-0 single-seed decent population [3], maintained over 30 generations under controlled growth conditions, provides a measure of thermodynamic properties within an unperturbed system. To assess resolution of methylation signal during plant development, we included previously reported datasets from various stages of seed development and germination in Arabidopsis ecotypes Col-0 and Ws [18]. Both of these systems have been described for methylome behavior with Methylpy, and direct comparison of the two datasets allowed estimation of developmental epigenetic signal above background. For more detailed study of methylation and gene expression, and to provide empirical testing of Methyl-IT predictions, we focused on the trans-generational 'memory' line derived by suppression of the *MSH1* (*MUTS HOMOLOG 1*) gene [19, 20], which has not been previously described for methylome features.

MSH1 is a plant-specific gene that encodes an organelle-localized protein [21, 22]. Plastid-depletion of MSH1 conditions 'developmental reprogramming' in the plant [23]. The *msh1* mutant is altered in expression of a broad array of environmental and stress response pathways [24], and the mutant phenotype is also produced by *MSH1* RNAi knockdown [20]. Differentially expressed gene (DEG) analysis of the *msh1* TDNA mutant identifies major components from numerous abiotic and biotic stress,

102 phytohormone, carbohydrate metabolism, protein translation and turnover, oxidative stress and
103 photosynthetic pathways [24]. Subsequent null segregation of the RNAi transgene restores *MSH1*
104 expression but leaves a heritably altered phenotype, with delayed flowering, reduced growth rate, delayed
105 maturity transition and pale leaves [20]. This condition is termed *msh1* 'memory', and provides for direct
106 investigation of transgenerational methylation variation and its association with altered gene expression.
107
108 Here, we report on Methyl-IT sensitivity relative to three commonly used methylome analysis programs.
109 We demonstrate resolution of methylome repatterning by Methyl-IT analysis, and empirical validation of
110 gene networks undergoing changes in methylation and gene expression as identified by the Methyl-IT
111 procedure.
112

113 ## Results

114 **The Methyl-IT method**

115 For resolution of DNA methylation signal, we employed Hellinger divergence (*H*) as a means of
116 quantifying dissimilarity between two probability distributions: that associated with a reference, defining
117 background changes, and that associated with treatment.
118
119 Signal detection is a critical step to increase sensitivity and resolution of methylation signal by reducing
120 the signal-to-noise ratio and objectively controlling the false positive rate and prediction accuracy/risk
121 (Fig. 1). Optimal detection of signals requires knowledge of the noise probability distribution that, from a
122 statistical mechanical basis, can be modeled for each individual sample by a Weibull distribution [9]. The
123 methylation regulatory signal does not hold Weibull distribution and, consequently, for a given level of
124 significance α (Type I error probability, eg. α = 0.05), cytosine positions with $H_{\alpha=0.05}$ can be selected as
125 sites carrying potential signals (shown as the blue region under the curve in Fig.1). Laws of statistical
126 physics can account for background methylation, a response to thermal fluctuations that presumably
127 function in DNA stability [9]. True signal is detected based on the optimal cutpoint [25], which can be
128 estimated from the area under the curve (AUC) of a receiver operating characteristic (ROC) built from a
129 logistic regression performed with the potential signals from controls and treatments. In this context, the
130 AUC is the probability to distinguish biological regulatory signal naturally generated in the control from
131 that induced by the treatment. In this context, the cytosine sites carrying a methylation signal are
132 designated *differentially informative methylated positions* (DIMPs). The probability that a DIMP is not
133 induced by the treatment is given by the probability of false alarm (P_{FA}, false positive). That is, the
134 biological signal is naturally present in the control as well as in the treatment.

Estimation of optimal cutoff from the AUC is an additional step to remove any remaining potential methylation background noise that still remains with probability $\alpha = 0.05 > 0$. We define as methylation signal (DIMP) each cytosine site with Hellinger divergence values above the cutoff ($H_{33}^{D_T}$), as shown in Fig. 1. Each DIMP is a cytosine position carrying a significant methylation signal, which may or may not be represented within a differentially methylated position (DMP) according to Fisher's exact test (or other current tests, Fig. 1). The difference in resolution by current methods versus Methyl-IT is illustrated by positioning H value sensitivity of the Fisher's exact test (FET) at greater than H_{min} for cytosine sites that are DMP and DIMPs simultaneously. For example, the ROC curve that corresponds to logistic regression for potential signals from the closest wild type control to *msh1* memory line (control 3 and treatment 1 in Fig. 1) has an AUC cutpoint of H =1.028052.

The probability of false alarm (estimated for best fit found for the Weibull cumulative distribution of H in the mentioned control) for DIMP detection based on the mentioned cutpoint is $P_{FA}=1.466\times10^{-6}$. Thus, in the *msh1* memory line dataset under study, any cytosine position k with $H_k \geq 1.028052$ is a DIMP. Although the probability P_{FA} =1.466 x 10^{-6} is small, there is still an average of 44844 CG-DIMPs per wild type sample. The average of CG-DIMPs in the memory line samples is 225835. We found that the strength of biological regulatory signal (evaluated in terms of AUC) was different for each methylation context. The strongest signal by Hellinger divergence found in our analyses was in CG context. A parsimony decision to reduce the rate of false positives used the cutpoint estimated for the AUC from the strongest signal. A flow chart of Methyl-IT analysis, with integration of these major procedures described above, is shown in Fig. 2.

Relative sensitivity of the Methyl-IT method versus other procedures

Table 1 provides a critical but nonunique example for the 2x2 contingency table with read counts $n_i^{mC_c} = 8$, $n_i^{C_c} = 2$, $n_i^{mC_t} = 350$, and $n_i^{C_t} = 20$. In this situation, and for any value $80 \leq n_i^{mC_t} \leq 350$, there exists strong methylation signal in the treatment, significantly stronger than in the control, but a 2x2 contingency independence test cannot detect it. Even small genomes like Arabidopsis contain millions of methylated cytosine sites, and situations analogous to the one presented in Table 1 are not rare. If this hypothetical cytosine site were to occur in the memory line, with $n_i^{mC_t} = 350$, then, according to its p-value estimate from the corresponding Weibull distribution, it would be a potential signal included in the logistic regression and, since H = 1.12 in this example and AUC cutpoint $H_{cutpoint}$ = 1.028, it would be a DIMP ($H_{cutpoint} < H$).

168

169 In the memory line, 100% of differentially methylated cytosines (*TVD* > 0.23) in all methylation contexts
170 found by root-mean-square test (RMST, bootstrap test of goodness-of-fit [26] implemented in methylpy
171 [17]), Fisher exact test (FET), and HDT (bootstrap test of goodness-of-fit based on Hellinger divergence,
172 see methods) are also detected by Methyl-IT (Fig. 3). RMST does not detect 17.7% of CG-DIMPs, 47.8%
173 CHG-DIMPs, and 59.7% CHH-DIMPs. HDT does not detect 19.7% of CG-DIMPs, 51.5% CHG-DIMPs,
174 and 66.1% CHH-DIMPs, while FET does not detect 46.2% of CG-DIMPs, 73.9% CHG-DIMPs, and 84%
175 CHH-DIMPs. Together, RMST, HDT and FET do not detect 13.5% of CG-DIMPs, 43.2% CHG-DIMPs,
176 and 52.5% CHH-DIMPs. The DIMPs not detected by these alternative approaches come from situations
177 analogous to that presented in Table 1. RMST is a robust test of goodness-of fit for 2x2 contingency
178 tables. The statistic used in RMST is an information divergence. Results obtained with RMST were very
179 close to those estimated based on Hellinger divergence [26, 27](see Table 1). Therefore, the differences in
180 outcome between Methyl-IT and Methylpy do not reside in RMST but, rather, in the signal detection
181 limitation, which requires knowledge of the null distribution for methylation background variation. The
182 null distribution of the control sample testing statistic must be taken into account.
183

184 Relative sensitivity and resolution of the Methyl-IT method can also be assessed by parallel analyses of
185 the three datasets, generational, seed development and *msh1* memory. Fig. 4 shows a single-scale, direct
186 comparison of differential methylation behavior in these datasets. Rather than total DIMP number, we
187 present relative. The absolute DIMP counts and DIMP counts per genomic region are provided in the
188 Additional File 2 Table.S1 for seed development and germination dataset. In Fig. 4, DIMP number is
189 normalized to the corresponding local cytosine context number. The signal detection step of Methyl-IT
190 discriminates signal unique to the sample from background patterning changes shared within the control
191 without regard to DMP density. Consistent with expectations, the generational dataset displays lowest
192 level variation across lineages, with greater inter-lineage variation than generational, and highest DIMP
193 signal in CG context. Direct comparison between the generational and seed development studies
194 estimated pattern and magnitude differences between the two datasets. Methylation signal in the seed
195 development dataset taken from the original study by Kawakatsu et al. [18] was greater than that of the
196 generational study, with DIMP signal in all three CG, CHG, CHH contexts. CHG and CHH changes were
197 associated predominantly with non-genic and TE regions, and CG DIMPs showed higher density within
198 gene regions (Fig. 4). Analysis of *msh1* memory, when compared to the generational and seed
199 development data, showed significantly greater magnitude change and prevalent methylation DIMP signal
200 within genic CG context. Genome-wide analysis of methylation in the memory line, enhanced by signal
201 detection, revealed considerable CG, CHG and CHH DIMPs across all chromosomes. Results are shown

202 for data before (Fig. 3) and after (Fig. 4 and Additional file 1: Figure S1) normalization to demonstrate
203 that while the vast majority of methylation resides in CHH context, normalized for density, changes in
204 CG context predominated on chromosome arms (Additional file 1: Figure S1).
205
206 A hierarchical cluster based on AUC criteria, and built on the set of 7006 selected DIMPs associated
207 genes, permitted the classification of seed developmental stages into two main groups: morphogenesis
208 and maturation phases (Additional File 1 Figure. S2a). In this case, the methylation signal was expressed
209 in terms of *log2*(*DIMP-counts on gene*). Within the 7006-dimensional metric space generated by 7006
210 AUC-selected genes, the linear cotyledon (*COT*) and mature green (*MG*) stages (morphogenesis-
211 maturation phase) grouped into a cluster quite distant from the cluster of post mature green (*PMG*) and
212 dry seed (*DRY*) stages (Dormancy phase). The latter cluster was closer to the leaf dataset derived from 4-
213 week-old plants. Similar analysis was performed for the seed germination experiment from the mentioned
214 study, and a hierarchical cluster built on the set of 3864 selected genes based on AUC criteria permitted
215 the classification of seed developmental stages into two main groups: 1) dormancy and 2) germination-
216 emerging phases (Additional File 1 Figure S2b).
217

218 **Differentially methylated genes (DMG)**

219 Here we propose the concept of differentially methylated genes (DMGs) based on the comparison of
220 group DIMP counts by applying generalized linear regression model (GLM). In particular, the use of
221 DMRs (clusters of DMPs within a specified region), can be tested in a group comparison by applying
222 GLM.
223
224 Genes displaying a statistically significant difference in the number of DIMPs relative to control were
225 defined as DMGs. Additional File 3 Table.S2 shows the number of DMGs observed in the seed
226 development data, based on Methyl-IT analysis. In this case, the analysis included DIMPs, regardless of
227 hypo or hyper methylation direction, and from all cytosine methylation contexts. Genes were defined as
228 the region covered by gene body plus 2kb upstream of the gene start site.
229
230 The number of DMGs (1068 genes) is considerably lower than the number of genes associated with
231 DMRs derived in the original study by Kawakatsu et al. (2017) [18]. Methylpy-derived DMR number
232 reflects genomic intervals with a given density of cytosine methylation changes, defined relative to a
233 control. Methyl-IT DMG number reflects gene regions with highest probability of differential methylation
234 distinct from background activity in the control. For example, after combining the embryogenesis CG,

235 CHG, and CHH DMRs reported in Kawakatsu et al. [18] (Table S5 from [18]) into a single set of DMRs,
236 only 468 from 6433 DMR-associated genes (after removing duplicated genes and updating annotation)
237 were Methyl-IT DMGs that met our GLM criteria in the group comparison of maturation phase versus
238 morphogenesis phase (Additional File 1 Figure S3a). DMR-associated gene analysis was also performed
239 with the set of DMRs detected in the germination experiment from the same study [18]. Similarly, 53
240 from 7638 DMR-associated genes were identified DMGs that met our GLM criteria in the group
241 comparison of germination-emerging versus dormancy phases (Additional File 1 Figure S3b). In this
242 case, 7638 DMR-associated genes comprise the resulting set from pooling germin-CHG and germin-CHH
243 DMRs (as reported in Table S5 from reference [18]). Analysis for the set of all genes yielded 136 DMGs
244 (Additional File 1 Figure S3c).
245
246 To more generally investigate the relative efficacy of commonly used methylation analysis programs, we
247 applied DSS, BiSeq and Methylpy to the *msh1* memory line and corresponding Col-0 control methylome
248 datasets. The control line was acquired as a transgene-null within the same transformation experiment that
249 produced MSH1-RNAi lines from which the memory line derives, and has been grown in parallel each
250 subsequent generation. The overlaps of DMR-associated genes from DMRs found in the memory line by
251 the methylome analysis pipelines DSS, BiSeq, and Methylpy is presented in Fig. 5a. What is striking is
252 the degree of data non-conformity from the three methods. Because the subjacent algorithms of these
253 programs are based not only on different statistical and computational approaches and do not define
254 DMRs uniformly, the data output differs in sensitivity and methylation change criteria. The application of
255 GLM to estimate the DMG set by Methyl-IT and its overlap with DMR-associated genes retrieved from
256 DMRs identified by the mentioned programs is shown in Fig. 5b. For the group comparison counting only
257 gene-body DIMPs, a total of 9271 loci (from the entire set of genes) were identified as DMGs in the *msh1*
258 memory line (Additional file 4: Table S3), while 8798 DMGs were identified for the group comparison
259 counting DIMPs within gene body plus 2kb upstream and downstream (with $TVD > 0.15$). The
260 application of GLM in estimating DMGs is not implemented to identify DMRs, but to evaluate whether
261 or not a statistically significant difference exists between methylation signals observed in two individual
262 groups for an already defined DMR.
263

264 **Methyl-IT identifies gene networks in seed development and germination dataset**

265 If heightened sensitivity in methylome signal detection imparts added biological information, this should
266 be evident in tests for association of methylome signal with gene expression changes. Observed CG and
267 CHG signal implies that changes in methylation during seed development relate to gene expression and/or

268 developmental transitioning. To investigate this possibility further, we conducted a network enrichment
269 analysis test (NEAT) of the Methyl-IT output from seed development and germination datasets.
270
271 Analysis of data from stages of seed development, including cotyledonary, mature green and post-mature
272 green, contrasted to globular as reference, suggested a methylome repatterning following the mature
273 green stage (Additional File 1 Figure. S2). Data indicate that methylome patterns are more similar
274 between cotyledonary and mature green stages, transitioning to a distinguishable state for post-mature
275 green and dry seed. This methylome transition may relate to the dessication and dormancy shift that also
276 occurs with this timing [28, 29]. Further analysis of differentially methylated loci with NEAT detected
277 statistically significant network enrichment of links between genes from the set of DMGs (Ws-0 seed)
278 and the set of GO-biological process terms associated with seed functions (Table 2). The list of genes
279 found in networks includes genes known to participate in seed development such as, For example,
280 transcription factors *DPBF2* (*AT3G44460*) from an abscisic acid-activated signaling pathway expressed
281 during seed maturation in the cotyledons, *ABSCISIC ACID BINDING FACTOR (ABF1, AT1G49720),* and
282 *WRKY22* (*AT4G01250*) a member of *WRKY* transcription factors involved mainly in seed development.
283 Other genes were found to be involved in seed dormancy, like *SLY1* (*SLEEPY1*), and seedling
284 development, like *EIN4* (AT3G04580), *CML16* (*AT3G25600*) (full gene list in Additional file 5: Table
285 S4). GeneMANIA (http://www.cytoscape.org/), identified interaction networks within the data, indicating
286 that many DMGs in the seed development dataset function together (Additional file 1: Figure S4).
287
288 Similar analysis of the seed germination and the Col-0 single-seed decent datasets did not detect DMGs
289 within networks. Results in the single-seed decent generational study are consistent with expectations,
290 since samples were grown under controlled conditions and sampled uniformly over generations. In the
291 case of the seed germination dataset, this outcome may be consistent with the fact that only CHG and
292 CHH DMRs were found in the original seed germination study by Kawakatsu et al. (2017) [18], while the
293 seed developmental experiment showed 60% of CG DMRs overlapping with protein-coding genes. These
294 data suggest that methylome signal may be more prominent under particular developmental transitions,
295 like seed preparation for dormancy and dessication, than during processes like germination.
296
297 **The memory line phenotype**
298 Transgene-null plants following segregation of the *MSH1*-RNAi transgene, termed *msh1* 'memory' lines,
299 display full penetrance and transgenerational inheritance of the altered phenotype, and the *msh1* memory
300 effect recapitulates in tomato [30]. Arabidopsis lines that have undergone silencing of *MSH1* segregate
301 for the *MSH1*-RNAi transgene by self-crossing to produce heritable phenotype changes in ca. 7-25% of

302 the resulting transgene-null progeny (Fig. 6a). The *msh1* memory phenotype is milder and more uniform
303 than that observed in *msh1* mutants derived by point mutation, T-DNA mutation or RNAi suppression [19,
304 20, 23] (Fig. 6b). Memory lines show normal *MSH1* transcript levels (Fig. 6c), but 100% penetrance and
305 heritability of the altered phenotype in subsequent self-crossed generations. Over 3,000 RNAi-null
306 memory line progeny under greenhouse conditions produced neither visible reversion to wild type nor
307 more severe *msh1* phenotypes (Additional file 1: Figure S5). In Arabidopsis, memory lines were stably
308 carried forward four generations and, in tomato, ten generations to date.
309

310 ***Memory line methylome changes detected by Methyl-IT associate with gene expression***

311 The derived transgene-null *msh1* memory lines display gene expression changes in ca. 955 genes
312 (Additional file 6: Table S5), approximately 67% of which are shared with the *msh1* mutant (Additional
313 file 7: Tables S6, Additional file 6: Tables S5).
314
315 The memory line DEG profile is distinctive. Unlike the mutant, which shows widespread gene ontology
316 enrichment in nearly every stress response pathway (Additional file 7: Table S6), memory line gene
317 ontology enrichment shows skewing toward integrated pathways for circadian clock, starch metabolism,
318 and ethylene and abscisic acid response (Fig. 6d). These studies use the *msh1* TDNA insertion mutant
319 rather than transgenic MSH1-RNAi for comparisons to ensure that each plant is *msh1*-depleted.
320 Transgenic RNAi knockdown lines are variable for *MSH1* suppression across plants (Fig. 6c), potentially
321 confounding interpretation, and MSH1-RNAi and *msh1* TDNA mutant appear identical in phenotype (Fig.
322 6b).
323
324 Application of Network-Based Enrichment Analysis (NBEA) to the set of 955 DEGs in the memory line
325 detected over-enrichment in five pathways: "*circadian rhythm*", "*response to red or far red light*",
326 "*regulation of circadian rhythm*", "*long-day photoperiodism/flowering*", and "*regulation of*
327 *transcription*". The permutation test applied to these data indicates that the observed simultaneous over-
328 enrichment of these pathways by chance holds a probability of lower than $4x10^{-5}$, reflecting a non-random
329 outcome (Additional file 8: Table S7).
330

331 ***The msh1 "memory" is a candidate system for non-genetic methylome reprogramming***

332 Similar to investigation of methylation changes during seed development and germination, we followed
333 Methyl-IT analysis of *msh1* memory line data with NEAT and network-based enrichment analysis
334 (NBEA) to assess biologically meaningful data based on DMGs alone. Additional file 9: Table S8 shows

335 results classifying methylation signal into networks for circadian clock, abscisic acid-activated signaling,
336 and defense response. Approximately 32% of identified DEGs overlap with DMGs in the memory line
337 (Fig 7a). These differentially methylated and expressed loci are over-enriched for genes contributing to
338 circadian rhythm, plant hormone signal transduction, and MAPK signaling pathway (Fig. 7b-7d).
339 Network analysis of expression, shown in (Fig. 7b-7d). , suggests dysregulation of these pathways in
340 *msh1* memory.
341
342 Integration of independently derived DEG, DMG and NBEA data from the memory lines converged on
343 16 loci (Fig. 7a and Table 3), of which 10 directly participate in circadian rhythm regulation and the
344 remainder, associated with light, ABA and ethylene response, are directly influenced by circadian clock
345 regulators (Table 3). Principal component (PC) analyses based on the mean of CG- Hellinger divergence
346 covering the gene regions delimited by DMGs (Fig. 8a), DMG/DEG intersection (Fig. 8b) and the
347 mentioned 16 loci (Fig. 8c) suggest a distinctive role of gene-associated CG methylation in *msh1*-memory
348 effect. For all analyses, more than 80% of variance among wild type, *msh1* memory and *msh1* TDNA
349 mutant was explained on the plane PC1-PC2, where *msh1* memory effect is clearly distinguishable from
350 control. Quantitative discriminatory power of CG methylation in the 16 signature loci is reflected in
351 hierarchical clustering based on their PC1-PC2 coordinates (Fig. 8d) and in their strong correlation with
352 the first two components (Fig. 8e). In particular, eight circadian rhythm genes strongly correlate with PC1,
353 which carries 65% of the whole sample variance. Thus, for these genes, CG methylation conveys enough
354 discriminatory power to distinguish individual wild type phenotypes from the *msh1* memory effect.
355
356 These observations are the first inference of association between CG methylation and gene expression
357 changes in the *msh1* memory line. DIMP distribution along the 16 signature loci showed most CG and
358 non-CG DIMPs located within exonic regions in memory lines with little individual CG-DIMP variation
359 (sometimes balanced with non-CG), suggesting that a programmed distribution pattern might exist
360 (Additional file 10: Table S9).
361
362 Predicted changes in methylation pattern at core circadian clock genes were subsequently confirmed by
363 sequence-specific bisulfite (BS) PCR analysis (Fig. 9a-9d). DIMPs were confirmed in the memory line at
364 *GI*, *TOC1, LHY* and *CCA1* genes. BS-PCR primer set BS-GI-P2, designed to bind to a predicted DIMP-
365 rich region, confirmed DIMPs within the region (Fig. 9e), while primer set BS-GI-P7, designed to bind to
366 a DIMP-free region, detected no changes (Fig. 9f). The DNA bisulfite conversion rate in this experiment
367 was confirmed by using *DDM1* as control, with a calculated bisulfite conversion rate of 99.47% for WT
368 and 100% for memory line sample (Additional file 1: Figure S6).

369

370 Germination of the memory line and isogenic Col-0 wild type on media containing 100 uM 5-azacytidine
371 alleviated the phenotype differences between the two lines, resulting in similar growth rates (Additional
372 file 1: Figure S7). Transfer to potting media to assess later growth showed wild type and memory lines to
373 be similar in phenotype following treatment (Additional file 1: Figure S7). Likewise, RNAseq analysis of
374 the treated and untreated memory and control lines showed 5-azacytidine treatment had genome-wide
375 effects on the gene expression pattern of both *msh1* memory line and wild type, and brought overall gene
376 expression patterns of treated *msh1* memory line and wild-type closer than before treatment (Additional
377 file 1: Figure S8). These observations reflect association between DNA methylation behavior and the
378 altered phenotype.

379

380 Wild type and memory line plants treated with 5-azacytidine were also tested for changes in expression of
381 the sixteen identified loci shown in Table 3. Quantitative RT-PCR assays confirmed previous RNAseq
382 results, showing significant differences in steady state transcript levels for 14 of the 16 loci in wild type
383 versus memory line plants growing under no treatment conditions (Additional file 1: Figure S9). Plants
384 germinated in 5-azacytidine prior to transfer to growth media, however, produced no significant
385 differences in gene expression for these loci in memory lines versus wild type (Additional file 1: Figure
386 S9). These data show a relationship between methylation state and gene expression changes in *msh1*-
387 induced memory, and provide evidence that altering methylation via chemical treatment can return gene
388 expression to nearly wild type steady state levels for these loci within the time period assayed.

389

390 **The *msh1* memory effect is related to circadian rhythm changes**

391 Both gene expression and methylome datasets, analyzed independently, indicated alteration in
392 components of the circadian clock. To test for modified circadian oscillation behavior in *msh1* memory,
393 gene expression levels for 4 core circadian clock genes in Arabidopsis and 2 genes in tomato were
394 evaluated over a 48-h time course under constant light (LL) and light-dark cycles (LD). Results confirmed
395 a degree of circadian rhythm dysregulation for all tested loci in both Arabidopsis memory lines, with
396 varying levels of altered expression (Fig 10). DEG analysis in Arabidopsis showed that the proportion of
397 genes regulated by *TOC1/CCA1* and altered in expression increased from 10.4% in the *msh1* T-DNA
398 mutant line to 33.1% in the *msh1* memory line (Fig 11a). Memory-associated processes identified in
399 Figure 6d, starch metabolism and cold, ethylene and abscisic acid response, are circadian clock output
400 pathways [31] (Fig 11b-e), again signifying that methylome repatterning influences genes that function
401 coordinately. The altered expression of three genes from these pathways was confirmed in Arabidopsis by

402 qRT-PCR (Additional file 1**:** Figure S10). Data to date suggest that circadian clock dysregulation
403 contributes to the memory line phenotype; it is not yet known whether clock dysregulation acts causally
404 in memory programming.
405

406 **Comparable memory effects are detected in tomato**

407 The *msh1* effect is recapitulated across plant species [23, 30]. We exploited this observation by
408 comparing *msh1* memory lines in Arabidopsis and tomato (cv 'Rutgers'). Genome-wide methylome
409 (BSseq) data were derived from Rutgers wild type and *MSH1*-RNAi transgene-null lines (fifth
410 generation). Similar to Arabidopsis, tomato memory lines are attenuated and more uniform in phenotype
411 relative to RNAi suppression lines, described by Yang et al. (2015)[30], and display reduced growth rate
412 and delayed flowering.
413

414 To test Methyl-IT analysis value in a dataset derived from another plant species, and to learn whether
415 signature pathways identified in Arabidopsis *msh1* memory line are shared in tomato *msh1* memory, we
416 conducted parallel analysis with the derived tomato memory line methylome dataset. Available gene
417 annotation in tomato is incomplete. Therefore, identified differentially methylated tomato loci were cross-
418 referenced to Arabidopsis orthologs. We identified 7802 tomato DMGs (Additional file 11: Table S10).
419 About 4277 of them were shared with Arabidopsis, accounting for ca. 55% of tomato DMGs and 46% of
420 Arabidopsis DMGs (Fig. 12a). With NBEA analysis, we identified 147 tomato genes predominantly
421 associated with phytohormone response, including auxin, salicylic acid, ethylene and ABA pathways,
422 together with circadian regulators, abiotic and biotic stress genes, and light response (Additional file 12:
423 Table S11). Arabidopsis homologs for 43% (63) of these 147 genes were found in Arabidopsis DMGs by
424 NBEA (Additional file 13: Table S12). Homologs for 6 of the 16 loci identified in Arabidopsis and listed
425 in Table 3 were present in the list of 147 tomato genes. Similar circadian clock dysregulation was
426 observed in tomato *msh1* memory as in its Arabidopsis counterparts. Gene expression levels for 2 core
427 circadian clock genes, *Sl_TOC1 (Solyc06g069690)* and *Sl_LHY (Solyc10g005080)* in tomato were
428 evaluated over a 48-h time course under light-dark cycles (LD) to confirm dysregulation (Fig. 12b), along
429 with downstream circadian clock-regulated genes (Fig. 12c). Together, these data reflect cross-species
430 conservation underlying *msh1* memory.

431 ## Discussion

432 Methyl-IT draws from the perspective that DNA methylation functions to stabilize DNA [32-34] and, as
433 such, may exist in "activated" versus "maintenance" states with regard to bioenergetics. We have begun

434 to investigate DNA methylation patterning as a "language" of sorts, identifying pattern changes that
435 comprise "signal" in response to treatment, without regard to density of methylation changes within a
436 given interval. While the theoretical premise underlying our approach, and based on Landauer's principle,
437 is detailed elsewhere [9, 10], the present study compares resolution of this methodology to current
438 methods for analysis of whole-genome methylation datasets.
439

440 Methyl-IT permits methylation analysis as a signal detection problem. Our model predicts that most
441 methylation changes detected, at least in Arabidopsis and tomato, represent methylation "background
442 noise" with respect to methylation regulatory signal, and are explainable within a statistical probability
443 distribution. Implicit in our approach is that DIMPs can be detected in the control sample as well. These
444 DIMPs are located within the region of false alarm in Fig. 1, and correspond to natural methylation signal
445 not induced by treatment. Thus, using the Methyl-IT procedure, methylation signal is not only
446 distinguished from background noise, but can be used to discern natural signal from that induced by the
447 treatment.

448

449 Whereas Methylpy, DSS and BiSeq provide essential information about methylation density, context and
450 positional changes on a genome-wide scale, Methyl-IT provides resolution of subtle methylation
451 repatterning signals distinct from background fluctuation. Data derived from analysis with Methylpy,
452 BiSeq or DSS alone could lead to an assumption that gene body methylation plays little or no role in gene
453 expression, or that transposable elements are the primary target of methylation repatterning. Yet ample
454 data suggest that this picture is incomplete [35]. Methyl-IT results show that these conclusions more
455 likely reflect inadequate resolution of the methylome system. GLM analysis applied to the identification
456 of DMR-associated genes by Methylpy, BiSeq and DSS indicates that DMRs (or DMR associated genes)
457 do not provide sufficient resolution to link them with gene expression.
458

459 Signal detected by Methyl-IT may reflect gene-associated methylation changes that occur in response to
460 local changes in gene transcriptional activity. Comparative analysis of the *msh1* memory line data with
461 *msh1* T-DNA mutant, a more extreme phenotype, showed 42.3% of memory line DMGs (3921 out of
462 5354) to overlap with *msh1* T-DNA DEGs. With the memory line DEGs estimated to number only 935,
463 it is possible that methylation repatterning within the memory line serves to stabilize or re-establish gene
464 expression following the extreme, stress-related changes that accompany *MSH1* silencing [24]. Similarly,
465 the pathway-associated methylome changes detected in seed development data may reflect participation

466 of methylation in gene expression stage transitions, particularly prominent between green mature and
467 post-green mature stages.
468
469 Methyl-IT analysis of various stages in seed development and germination showed evidence of
470 methylation changes. Previous Methylpy output [18] defined predominant changes in non-CG
471 methylation residing within TE-rich regions of the genome, whereas Methyl-IT data resolved statistically
472 significant methylation signal within gene regions. With the complementary resolution provided by
473 Methyl-IT, it becomes possible to investigate the nature of chromatin response within identified genes in
474 greater detail during the various stages of a seed's development. Several of the identified DMGs in this
475 study involved genes that interact within known development pathways.
476
477 There is little detail available in plants of local intragenic methylation behavior during transitions in gene
478 activation, but transcription factor-associated recruitment of methylation machinery has been postulated
479 [35], and supported by data in other systems [36]. A large proportion of the intervals identified by this
480 study are components of signal transduction, so expression effects may be below the detection limits of
481 the assay. Among the 1717 transcription factors reported in PlantTFDB, 340 are identified as DMGs in
482 our list for memory line. Effects of alternative splicing in memory changes, also known to respond to
483 local methylation [37], would similarly have escaped detection in our gene expression analysis. However,
484 for a better comprehension of which genes would be controlled by the regulatory methylation machinery
485 in processes like seed developmental or the induced *msh1* memory effect, the network enrichment
486 analysis of DMGs and DEGs can reduce the number of potential regulators to a minimal number of genes
487 testable under lab conditions, as presented in our study. Analysis produced evidence of a relationship
488 between *msh1* memory line gene expression and differential methylation data for at least 16 regulatory
489 loci, 10 of which comprise components of the circadian clock.
490
491 Plants have the capacity to respond to a wide array of abiotic and biotic stresses and developmental cues
492 through overlapping gene networks. It is increasingly evident that phytohormone, light response, abiotic
493 and biotic stress response, photosynthesis and carbohydrate metabolism are integrated output pathways of
494 the plant's circadian clock [31]. A significant proportion of the plant's gene expression profile is
495 influenced by circadian regulation [38], introducing the concept of a master regulator of adaptation.
496 Numerous reports underscore extensive pathway integration under circadian clock control, with starch
497 metabolism, cold response and abscisic acid-mediated stress response, for example, as particularly
498 prominent pathways altered by *msh1* memory. The link between plant response to cold and epigenetic
499 memory involves histone modifications of the FLC locus during vernalization [39]. Cold temperature also

500 influences alterative splicing patterns of clock genes to alter their function [40]. ABA, a stress hormone,
501 shows rhythmic diel levels in plants [41], and associates with *TOC1* and an ABA-related gene, *ABAR*, in
502 a highly regulated feedback loop [42]. Epigenetic modification of circadian clock genes effect changes in
503 starch metabolism [43], and can educe enhanced growth vigor in hybrids and allopolyploids [44]. Studies
504 of classical heterosis in Arabidopsis also show association with changes in circadian clock behavior [45].
505 Data from this study indicate that *MSH1* suppression includes circadian clock, ABA and ethylene
506 dysregulation as components of the associated *msh1* global stress condition. Segregation of the *MSH1*-
507 RNAi transgene only partially reverts the phenotype, revealing loci that have apparently sustained
508 cytosine methylation repatterning, and producing a phenotypic memory effect, presumably methylation-
509 based, that is reproducible and heritable. If correct, the *msh1* memory phenomenon comprises a robust
510 medium for addressing epiallelic stability.
511
512 Identification of gene networks in both seed development and *msh1* memory was based on DNA
513 methylation data analysis with the enhanced resolution of Methyl-IT. In the case of *msh1* memory, gene
514 expression, phenotype and cross-species comparison served to confirm the identified networks. While
515 early in the process, these outcomes argue compellingly for the feasibility of genome-wide methylome
516 decoding of the gene space.

517 Conclusions

518 Methyl-IT is an alternative and complementary approach to plant methylome analysis that discriminates
519 DNA methylation signal from background and enhances resolution. Analysis of publicly available
520 methylome datasets showed enhanced signal during seed development and germination within genes
521 belonging to related pathways, providing new evidence that DNA methylation changes occur within gene
522 networks. Similarly, *msh1* transgenerational memory phenomena in Arabidopsis and tomato identified
523 methylation-altered gene networks involving circadian clock components and linked stress response
524 pathways altered in expression and connected to phenotype. Whereas, previous methylome analysis
525 protocols identify changes in methylome density and landscape, predominantly non-CG, Methyl-IT
526 reveals effects within gene space, mostly CG and CHG, for elucidation of methylome linkage to gene
527 effects.

528 Methods

529 Methylome analysis

530 The alignment of BS-Seq sequence data from *Arabidopsis thaliana* was carried out with Bismark 0.15.0
531 [46]. BS-Seq sequence data from tomato experiment were aligned using ERNE 2.1.1 [47]. The basic

532 theoretical aspects of methylation analysis applied in the current work are based on previous published
533 results [9]. Details on Methyl-IT steps are provided in the next sections.
534

535 ***Methylation level estimation***

536 To estimate methylation levels at each cytosine position, we followed a Bayesian approach. In a Bayesian
537 framework assuming uniform priors, the methylation level p_i can be defined as:
538 $p_i = \left(n_i^{mC} + 1\right) / \left(n_i^{mC} + n_i^{C} + 2\right)$ (1), where n_i^{mC} and n_i^{C} represent the numbers of methylated and non-
539 methylated read counts observed at the genomic coordinate i, respectively. We estimate the shape
540 parameters α and β from the beta distribution $P\left(p \middle| \alpha, \beta\right) = \dfrac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)}$ (2) minimizing the
541 difference between the empirical and theoretical cumulative distribution functions (ECDF and CDF,
542 respectively), where $B(\alpha, \beta)$ is the beta function with shape parameters α and β. Since the beta
543 distribution is a prior conjugate of binomial distribution, we consider the *p* parameter (methylation level
544 p_i) in the binomial distribution as randomly drawn from a beta distribution. The hyper-parameters α and
545 β are interpreted as pseudo counts. Then, the mean $E\left[p_i \middle| D\right] = \hat{p}_i$ of methylation levels p_i, given the
546 data *D*, is expressed by $\hat{p}_i = \dfrac{\alpha + n_i^{mC}}{\alpha + \beta + n_i^{mC} + n_i^{C}}$ (3). The methylation levels at the cytosine with genomic
547 coordinate i are estimated according to this equation.
548

549 ***Hellinger and Total Variation divergences of the methylation levels***

550 The difference between methylation levels from reference and treatment experiments is expressed in
551 terms of information divergences of their corresponding methylation levels, $\hat{p}_i^r$ and $\hat{p}_i^t$, respectively. The
552 reference sample(s) can be additional experiment(s) fixed at specific conditions, or a virtual sample
553 created by pooling methylation data from a set of control experiments, e.g. wild type individual or group.

554 Hellinger divergence between the methylation levels from reference and treatment experiments is
555 defined as:

556 $$H\left(\hat{p}_i^r, \hat{p}_i^t\right) = w_i \left[\left(\sqrt{\hat{p}_i^r} - \sqrt{\hat{p}_i^t}\right)^2 + \left(\sqrt{1-\hat{p}_i^r} - \sqrt{1-\hat{p}_i^t}\right)^2 \right] \quad (4)$$

557 Where $w_i = 2\frac{m_i^t m_i^r}{m_i^t + m_i^r}$, $m_i^t = n_i^{mC_t} + n_i^{C_t} + 1$ and $m_r = n_i^{mC_r} + n_i^{C_r} + 1$. The total variation of the

558 methylation levels $TV\left(\hat{p}_i^r, \hat{p}_i^t\right) = \hat{p}_i^r - \hat{p}_i^t$ (5) indicates the direction of the methylation change in the

559 treatment, hypo-methylated $TV < 0$ or hyper-methylated $TV > 0$. *TV* is linked to a basic information

560 divergence, the total variation distance, defined as: $TVD\left(\hat{p}_i^r, \hat{p}_i^t\right) = \left|TV\left(\hat{p}_i^r, \hat{p}_i^t\right)\right|$ (6). Distance

561 $TVD\left(\hat{p}_i^r, \hat{p}_i^t\right)$ and Hellinger divergence hold the inequality: $TVD\left(\hat{p}_i^r, \hat{p}_i^t\right) \le \frac{2}{\sqrt{2w_i}}\sqrt{H\left(\hat{p}_i^r, \hat{p}_i^t\right)}$ (7)

562 [48]. Under the null hypothesis of non-difference between distributions $\hat{p}_i^r$ and $\hat{p}_i^t$, Eq. 4 asymptotically

563 has a chi-square distribution with one degree of freedom. The term w_i introduces a useful correction for

564 the Hellinger divergence, since the estimation of $\hat{p}_i^t$ and $\hat{p}_i^r$ are based on counts (see Table 1).

565

566 ***Non-linear fit of Weibull distribution***

567 The cumulative distribution functions (CDF) for $H_k\left(\hat{p}_k^r, \hat{p}_k^t\right)$ can be approached by a Weibull

568 distribution $P\left(H_k \le H^0 \middle| \alpha, \lambda, \mu\right) = 1 - e^{-\left(\frac{H_k - \mu}{\lambda(l)}\right)^{\alpha}}$ (8) [9]. Parameter $\hat{\alpha}, \hat{\lambda}$ and $\hat{\mu}$ were estimated by non-

569 linear regression analysis of the ECDF $\hat{F}_n\left(\hat{H}_k \le H^0\right)$ versus $H_k\left(\hat{p}_k^r, \hat{p}_k^t\right)$ [9]. The ECDF of the variable

570 $\hat{H}_k$ is defined as:

571 $$\hat{F}_n\left(\hat{H}_k \le H^0\right) = \frac{\text{number of CDMs in the samples with } \hat{H}_k \le H}{n} = \frac{1}{n}\sum_{k=1}^{n} 1_{\hat{H}_k \le H^0} \quad (9)$$

572 , where $1_{\hat{H}_k \le H^0} = \begin{cases} 1 \text{ if } \hat{H}_k \le H^0 \\ 0 \text{ if } \hat{H}_k > H^0 \end{cases}$ is the indicator function. Function $\hat{F}_n\left(\hat{H}_k \le H^0\right)$ is easily computed

573 (for example, by using function "*ecdf*" of the statistical computing program "R"[49]).

574

575 ***A statistical mechanics-based definition for a potential/putative methylation signal (PMS)***

576 Most methylation changes occurring within cells are likely induced by thermal fluctuations to ensure

577 thermal stability of the DNA molecule, conforming to laws of statistical mechanics [9]. These changes do

578 not constitute biological signals, but methylation background noise induced by thermal fluctuations, and
579 must be discriminated from changes induced by the treatment. Let $P\left(E_k^D \le E_k^{D_0}\right)$ be the probability that
580 energy E_k^D, dissipated to create an observed divergence D between the methylation levels from two
581 different samples at a given genomic position k, can be lesser than or equal to the amount of energy $E_k^{D_0}$.
582 Then, a single genomic position k shall be called a PMS at a level of significance α if, and only if, the
583 probability $P\left(E_k^D > E_k^{D_0}\right) = 1 - P\left(E_k^D \le E_k^{D_0}\right)$ to observe a methylation change with energy dissipation
584 higher than $E_k^{D_0}$ is lesser than α. The probability $P\left(E_k^D \le E_k^{D_0}\right)$ can be given by a member of the
585 generalized gamma distribution family and, in most cases, experimental data can be fixed by the Weibull
586 distribution [9]. Based on this dynamic nature of methylation, one cannot expect a genome-wide
587 relationship between methylation and gene expression. A practical definition of PMS based on Hellinger
588 divergence derives provided that H_k is proportional to E_k^H and using the estimated Weibull CDF for
589 H_k given by Eq. 8. That is, a single genomic position k shall be called a PMS at a level of significance
590 α if, and only if, the probability $\hat{P}\left(H_k > H^0 \middle| \hat{\alpha}, \hat{\lambda}, \hat{\mu}\right) = 1 - \hat{P}\left(H_k \le H^0 \middle| \hat{\alpha}, \hat{\lambda}, \hat{\mu}\right)$ to observe a
591 methylation change with Hellinger divergence higher than H_k is lesser than α.

592 The PMSs reflect cytosine methylation positions that undergo changes without discerning whether they
593 represent biological signal created by the methylation regulatory machinery. The application of signal
594 detection theory is required for robust discrimination of biological signal from physical noise-induced
595 thermal fluctuations, permitting a high signal-to-noise ratio.

596

597 ***Robust detection of differentially informative methylated positions (DIMPs)***

598 Application of signal detection theory is required to reach a high signal-to-noise ratio [50, 51]. To
599 enhance DIMP detection, the set of PMSs is reduced to the subset of cytosines with
600 $TVD\left(\hat{p}_i^r, \hat{p}_i^t\right) \le TVD_0$, where TVD_0 is a minimal total variation distance defined by the user, preferably
601 $TVD_0 > 0.1$. If we are interested not only in DIMPs but also in the full spectrum of biological signals,
602 this constraint is not required. Once potential DIMPs are estimated in the treatment and in the control
603 samples, a logistic regression analysis is performed with the prior binary classification of DIMPs, i.e., in
604 terms of PMSs (from treatment versus control), and a receiver operating curve (ROC) is built to estimate
605 the cutpoint of the Hellinger divergence at which an observed methylation level represents a true DIMP.

606 There are several criteria to estimate the optimal cutpoint, many of which are implemented in the R
607 package *OptimalCutpoints* [25]. The optimal cutpoint used in Methyl-IT corresponds to the *H* value that
608 maximizes Sensitivity and Specificity simultaneously [52, 53]. These analyses were performed with the R
609 package *Epi* [54].

610 Once all pairwise comparisons are done, a final decision of whether a DFMP is a DIMP is taken based on
611 the highest cutpoint detected in the ROC analyses (Fig. 1). That is, the decision is taken based on the
612 cutpoint estimated in the ROC analysis for the control sample with the closest distribution to treatment
613 samples. The position of the cutpoint will determine a final posterior classification for which we would
614 estimate the number of true positive, true negatives, false positives and false negatives. For each cutpoint
615 we would estimate, the accuracy and the risk of our predictions. We may wish to use different cutpoints
616 for different situations. For example, if our goal is the early detection of a terminal disease and high
617 values of the target variable indicates that a patient carries the disease, then to save lives we would prefer
618 the lowest meaningful cutpoint reducing the rate of false negative.

619

620 ***Estimation of differentially methylated genes (DMGs) using Methyl-IT***

621 Our degree of confidence in whether DIMP counts in both control and treatment represent true biological
622 signal was set out in the signal detection step. To estimate DMGs, we followed similar steps to those
623 proposed in Bioconductor R package DESeq2 [55], but the test looks for statistical difference between the
624 groups based on gene body DIMP counts rather than read counts. The regression analysis of the
625 generalized linear model (GLMs) with logarithmic link was applied to test the difference between group
626 counts. The fitting algorithmic approaches provided by *glm* and *glm.nb* functions from the R packages
627 *stat* and MASS were used for Poisson (PR), Quasi-Poisson (QPR) and Negative Binomial (NBR) linear
628 regression analyses, respectively.

629 Likewise for DESeq2 we used the linear regression model $\log_2\left(q_{ij}\right) = \sum_k x_{jk}\beta_{ik}$, with design matrix
630 elements x_{jk}, coefficients β_{ik}, and mean $\mu_{kj} = s_j q_{kj}$, where s_j normalization constants are considered
631 constant within a group. Only two groups were compared at a time. The design matrix elements indicate
632 whether a sample *j* is treated or not, and the GLM fit returns coefficients indicating the overall
633 methylation strength at the gene and the logarithm base 2 of the fold change (log2FC) between treatment
634 and control [55]. In particular, in the case of NBR, the inverse of the variance was used as prior weight

635 ($\sigma_{jk}^2 = \frac{1}{\mu_{ij} + \mu_{ij} disp}$, where *disp* is data dispersion computed by the *estimateDispersions* function from

636 DESeq2 R package).

637 To test difference between group counts we applied the fitting algorithmic approaches: PR and PQR if

638 $\rho < \frac{\mu_{ij}}{\sigma_{ij}} \leq 1 (0.9 < \rho < 1)$, NBR and NBR with '*prior weights*'. Next, best model based on Akaike

639 information criteria (AIC). The Wald test for significance of the independent variable coefficient indicates
640 whether or not the treatment effect is significant, while the coefficient sign (*log2FC*) will indicate the
641 direction of such an effect.

642

643 ***Bootstrap goodness-of-fit test for 2x2 contingency tables***

644 The goodness-of-fit RMST 2x2 contingency tables as implemented in methylpy [17] for the estimation of
645 DMSs (based on the root-mean-square (RMS) statistics) is explained in Perkins et al. in reference [26](a
646 complemental description is found at arXiv:1108.4126v2). The bootstrap heuristic to perform the test is
647 given in reference [56]. An analogous bootstrap goodness-of-fit test based on Hellinger divergence was
648 also applied to estimate DMPs (HDT). In this case, Hellinger divergence estimated according to the first
649 statistic given in Theorem 1 from reference [27].

650

651 ***Identification of differentially methylated regions by using BiSeq, DSS and MethyPy***

652 For BiSeq, raw sequence reads were trimmed to remove both poor-quality calls and adapters using Trim
653 galore! (version 0.4.1) with options --paired --trim1 --gzip --phred33 --fastqc and Cutadapt (version 1.9.1)
654 with cutoff 20. Remaining sequences were mapped to the Arabidopsis TAIR10 genome using Bismark
655 (version v0.15.0) [46] and Bowtie2 (Version 2.2.9) [57].Duplicates were removed using the Bismark
656 deduplicate function, and methylation calls were extracted with Bismark methylation extractor, reading
657 methylation calls of overlapping parts of the paired reads from the first read (–no_overlap parameter).
658 Differentially methylated regions were detected with BiSeq (version 1.18.0) [16, 58] with clusters at least
659 15 methylated sites with 100 bp between clusters.

660

661 For DSS, raw sequence reads were trimmed to remove both poor-quality calls and adapters using Trim
662 galore! (version 0.4.1) with options --paired --trim1 --gzip --phred33 --fastqc and cutadapt (version 1.9.1)
663 with cutoff 20. Remaining sequences were mapped to the Arabidopsis TAIR10 genome using Bismark

664 (version v0.15.0) [46] and Bowtie2 (Version 2.2.9)[57]. Duplicates were removed using the Bismark
665 deduplicate function and methylation calls were extracted with Bismark methylation extractor, reading
666 methylation calls of overlapping parts of the paired reads from the first read (–no_overlap parameter).
667 Differentially methylated regions were detected with DSS (Dispersion shrinkage for sequencing data,
668 version 2.26.0) using the default parameters.
669
670 For MethylPy, differentially methylated regions (DMR) were identified using the MethylPy pipeline
671 (version v0.1.0) [17] and Bowtie2 (Version 2.3.3)[57]. This pipeline used Cutadapt (version >=1.9) to
672 trim the raw sequence reads to remove both poor-quality calls and adapters. Picard (>=2.10.8) was used
673 for PCR duplicate removal. Chloroplast DNA sequence was used as the unmethylated control; the
674 conversion rate observed was between 0.3% - 0.4%. Cytosine sites with less than four reads were
675 discarded. Adjacent differential methylated sites closer to 100bp were collapsed into DMRs. CNN DMRs,
676 CGN DMRs, CHG DMRs, and CHH DMRs with fewer than four, eight, four, and four DMSs,
677 respectively, were discarded in following analyses, and CNN DMRs, CGN DMRs, CHG DMRs, and
678 CHH DMR candidate regions with less than 0.1, 0.4, 0.2, and 0.1 differences between maximum and
679 minimum methylation levels were also discarded.
680
681 For Methyl-IT, raw sequence reads were trimmed to remove both poor-quality calls and adapters using
682 Trim galore! (version 0.4.1) with options --paired --trim1 --gzip --phred33 --fastqc and Cutadapt (version
683 1.9.1) with cutoff 20. Remaining sequences were mapped to the Arabidopsis TAIR10 genome using
684 Bismark (version v0.15.0) [46]; and Bowtie2 (Version 2.2.9) [57]. Duplicates were removed using the
685 Bismark deduplicate function and methylation calls were extracted with Bismark methylation extractor,
686 reading methylation calls of overlapping parts of the paired reads from the first read (–no_overlap
687 parameter). Differentially methylated regions were detected with Methyl-IT, using cytosine sites with at
688 least 4 reads, and with default parameters.
689
690 Since methods DSS, BiSeq and Methylpy do not provide an equivalent concept to DMGs, we adopted the
691 concept of *DMR associated genes* (DAGs) introduced in reference [18]. Basically, *a gene and a DMR are*
692 *associated if the DMR is located within 2 kb of gene upstream regions, gene bodies and 2 kb of gene*
693 *downstream regions* [18].
694

695 ***Available methylome datasets used in this work***

696 Methylome datasets from Arabidopsis (Ws-0) major seed developmental phases, globular stage (GLOB),
697 linear cotyledon stage (COT), mature green stage (MG), post mature green stage (PMG) and dry seed,
698 and Arabidopsis (Col-0) germination datasets of dry seed and 0-4 days after imbibition were analyzed.
699 Ws-0 seed development and germination datasets were obtained from the Gene Expression Omnibus
700 (GEO) under accession numbers GSE68132 and GSE94710. Both dataset were original studied by
701 Kawakatsu et al. (2017) [18].
702

703 ***Network enrichment analysis***

704 Network based enrichment analysis (NBEA) was applied using the EnrichmentBrowser R package [59,
705 60] and the Network Enrichment Analysis Test (NEAT) was performed by using the R package "neat"
706 version 1.1.1[60].

707 These network enrichment approaches permitted identification of main network regulators involved in the
708 *msh1* memory transgenerational effect and in seed developmental and germination datasets.

709

710 ***Individual sample gene CG methylation principal component analysis (PCA) and***
711 ***classification***

712 Individual samples were represented as vectors of variables carrying the mean of CG Hellinger
713 divergence covering gene regions delimited by Arabidopsis *msh1*-memory DMGs. Principal component
714 analysis (PCA) was performed on the individual vector-spaces determined by the gene regions: 1) DMGs ,
715 2) intersection DEGs (*msh1*-memory)/DMGs, and 3) intersection NBEA-DMG/NBEA-DEG between the
716 subsets derived from independent NBEA on the subsets DMGs and DEGs, respectively. PCA and
717 hierarchical cluster analysis were applied by using *prcomp* and *hclust* functions, respectively, from the R
718 package *stats*.

719

720 **Specific locus bisulfite sequencing PCR**

721 To confirm our analysis for DIMP calling based on methylome sequencing, PCR-based bisulfite
722 sequencing was performed. Genomic DNA from leaf tissue of 4-week-old plants was isolated by the
723 DNeasy Plant Kit (Qiagen, Germany). 400 ng of genomic DNA was bisulfite-treated using EpiMark
724 Bisulfite Conversion Kit (New England Biolabs, USA). Bisulfite-treated DNA was used as template for
725 PCR in a 25 ul reaction system by using EpiMark Hot Start Taq DNA Polymerase (New England Biolabs,

726 USA), in the PCR program: Initial denaturation 30 sec at 95 °C, 40 cycles of 95°C for 15 sec ,45°C for 30
727 sec, 68°C for 1 min, and final extension 5 min at 68 °C. PCR product was gel-purified using kit (Qiagen,
728 Germany) and ligated to TOPO TA cloning kit (Life, USA) for sequencing. At least 25 independent
729 clones were sequenced. Bisulfite DNA sequence methylation status was analyzed by the online program
730 "Kismeth". Methylation at locus AT5G66750 was used as a control for bisulfite conversion. Primers used
731 in this experiment are listed in the Additional file 14 Table S13.
732

733 **Plant materials and growth conditions**

734 For Arabidopsis plants used in this study, clean seeds were sown on peat mix in square pots, with
735 stratification at 4 °C for 2 days before moving to growth chamber (22 °C, 120-150 μ mol·m^{-2}·s^{-1} light).
736 Tomato seeds were germinated on MetroMix 200 medium (SunGro, USA) in square pots and grown in a
737 reach-in chamber (26 °C, 300 μ mol·m^{-2}·s^{-1} light).

738

739 **5-azacytidine treatment**

740 The 5-azacytidine treatment protocol was adopted from Griffin et al [57] and Yang et al [30]. Col-0 wild
741 type and *msh1* memory line seeds were surface-sterilized in 10% (v/v) sodium hypochlorite, rinsed
742 thoroughly with sterile water, and sown in 8-oz clear cups (Fabri-Kal, USA) containing 30 mL 0.5 M
743 Murashige and Skoog medium (Sigma, USA) supplemented with 1% (w/v) agar and 0 (control) or 100
744 µM 5-azacytidine (Sigma, USA). The 100 µM concentration was derived from a concentration gradient
745 experiment of 4 concentrations (0 µM, 30 µM ,50 µM ,100 µM) where 100 µM showed visible impact on
746 plant growth for both wild type Col-0 and *msh1* memory line plants. Seeds were germinated and grown at
747 24°C, 18-h day length, and 120-150 μ mol·m^{-2}·s^{-1} light intensity for 14 days. 10 days old seedling on the
748 MS medium were collected for RNAseq experiment. For longer observation, the treated plants were
749 transferred to square pots with soil and grow under standard conditions in the growth chamber. The
750 experiment was repeated three times, with at least 18 replicates per treatment each experiment.
751

752 **Sample collection for circadian clock gene expression assays**

753 To assess the expression pattern of core circadian clock genes under clock-driven free running conditions,
754 we adopted the protocol of [38]. Plants were entrained at LD condition (12 hr light/ 12 hr dark) for 4
755 weeks, then moved to LL (24 hr constant light) for 48 hours before sample collection was initiated.

756 For expression of core circadian clock genes under life-like conditions, plants were entrained at LD (12 hr
757 light/12 hr dark) for 4 weeks before samples were collected. The entire above-ground plant was collected
758 and placed into liquid nitrogen. Samples were taken every 4 hr (ZT6, ZT10, ZT14, ZT18, ZT22,
759 ZT26.ZT30, ZT34, ZT38, ZT42, ZT46, ZT50) in both LD and LL conditions. For each genotype at each
760 time point, at least 3 plants were collected and used in qPCR experiments as biological replicates.
761 An identical sample collection strategy, and LD, LL entrainment conditions, were used for tomato
762 circadian clock gene expression experiments.
763

764 **Gene Expression Analysis by qPCR**

765 The MIQE [61] was used as standard protocol for the qPCR experiments. Briefly, total RNA from each
766 sample was extracted by NucleoSpin RNA Plant kit (Macherey-Nagel, Germany) following
767 manufacturer's protocol, including genomic DNA removal. First-strand cDNA was synthesized from
768 400ng total RNA with oligo primers using iScript Reverse Transcription Supermix for RT-PCR (Bio-Rad,
769 USA). The qPCR was performed on the CFX real-time system (Bio-Rad, USA) with 95 °C for 3 min, 40
770 cycles of 95 °C for 30 sec and 60 °C for 1 min. Three biological replicates were performed. RNA
771 abundance of target genes was calculated from the average of four technical replicates using $\Delta\Delta$Cq
772 method, where Cq is the cycle number at which amplification signal reaches saturation in each PCR run.
773 The Cq values of AT4G05320 and AT5G15710 were used as normalization controls in the calculation.
774

775 Real-time PCR primers used in this study and their reference are listed in Supplemental Primers Table.
776 The PCR amplification efficiency was calculated based on a calibration standard curve specific for each
777 primer set, and only primers having amplification efficiency greater than 0.97 were used in the study.
778

779 **Sample preparation and bisulfite DNA methylome sequencing**

780 For Arabidopsis genome-wide bisulfite methylome sequencing experiments, three individual plants of
781 wild type *Arabidopsis thaliana* ecotype Col-0 and three isogenic *msh1* memory line plants were used. All
782 wild type control plants selected from negative events of RNAi transformation and were maintained in
783 parallel with their msh1 memory counterparts. Whole plants at early bolting were flash frozen in liquid
784 nitrogen. Tissues were ground by motor and pestle in liquid nitrogen, and divided to two, with one half
785 processed by DNeasy Plant Kit (Qiagen, Germany) for genomic DNA (RNA removed) and subsequent
786 bisulfite sequencing. The other half was used for RNA extraction by NucleoSpin RNA Plant Kit
787 (Macherey-Nagel, Germany) following manufacturer's protocol, including genomic DNA removal, for
788 RNA-seq analysis.

789

790 For tomato bisulfite sequencing, wild type tomato (*Solanum lycopersicum* cv Rutgers) and the
791 corresponding MSH1-RNAi transgene-null segregant (*msh1* memory line) were used. Phenotype and line
792 generation details can be found in [30]. The top three leaves from each four-week-old tomato plant were
793 collected and frozen in liquid nitrogen, followed by genomic DNA extraction using DNeasy Plant Kit
794 (Qiagen, Germany). Genomic DNA from three individual plants for both WT and *msh1* memory line
795 were used for BSseq.
796
797 All BSseq experiments were conducted on the Hiseq 4000 analyzer (Illumina, USA) at BGI-Tech
798 (Shenzhen, China) according to manufacturer's instructions. Briefly, Genomic DNA was sonicated to
799 100-300 bp fragments and purified with MiniElute PCR Purification Kit (Qiagen, Germany), and
800 incubated at 20°C after adding End Repair Mix. DNA was purified, a single 'A' nucleotide added to the 3'
801 ends of blunt fragments, purified again and Methylated Adapter was added to 5' and 3' ends of each
802 fragment. Fragments of 300-400 bp size range were purified with QIAquick Gel Extraction Kit (Qiagen,
803 Germany) and subjected to bisulfite treatment by Methylation-Gold Kit (ZYMO). These steps were
804 followed by PCR and gel purification (350-400 bp fragments were selected). Qualified libraries were
805 paired-end sequenced on the HiSeq X-ten system.
806

807 **RNA sequencing and analysis**

808 RNA libraries were constructed as described in the TruSeq RNA Sample Preparation v2 Guide. These
809 libraries were sequenced with the 150-bp reads option, in Hi-Seq 4000 analyzer (Illumina, USA) at BGI-
810 Tech (Shenzhen, China). Alignments were performed using RUM 2.0.4 (default parameters) [62] keeping
811 only uniquely mapped reads. The read count data were generated from the SAM files by using QoRTs
812 software package[63]. DESeq2 [55] was used for gene count normalization and to identify differentially
813 expressed genes (FDR < 0.05, $|log2FC| > 0.5$.
814

815 ## Abbreviations

816 **AUC:** Area under the receiver operating characteristic curve
817 **MSH1:** MUTS HOMOLOG 1
818 **CDM:** Cytosine DNA methylation
819 **DAGs:** DMR associated genes
820 **DEG:** Differentially expressed gene

821 **DIMPs:** Differentially informative methylated positions
822 **DMGs:** Differentially methylated genes
823 **DMPs:** Differentially methylated positions
824 **DMRs:** differentially methylated regions
825 **DSS:** Dispersion Shrinkage for Sequencing
826 **FET:** Fisher's exact test
827 **GLM:** generalized linear regression model
828 **HD:** Hellinger divergence
829 **HDT:** goodness-of-fit test based on Hellinger divergence
830 **NEAT**: Network Enrichment Analysis Test
831 **NBEA**: Network based enrichment analysis
832 **RMST:** Root-mean-square test
833 **ROC:** Receiver operating characteristic curve
834 **SD:** Signal detection
835 **TVD:** total variation distance
836 **PMS:** Potential/putative methylation signal

837 Acknowledgments

838 We thank Ojus Jain and Kasim Hamo for technical assistance. We also thank Dr. Yingzhi Xu for valuable
839 conversations early in the study. The data presented in this manuscript are tabulated in the main text and
840 supplementary materials.

841 Funding

842 The work was supported by funding from NSF-SBIR (2015-33610-23428-UNL) and the Bill and Melinda
843 Gates Foundation (OPP1088661).

844 Availability of data and materials

845 The Methyl-IT pipeline source code is available at the GitLab: https://git.psu.edu/genomath/MethylIT
846 Seed development methylome data (accession number GSE68132) were obtained from the Gene
847 Expression Omnibus database.
848 All Next Generation Sequencing data generated by this study are deposited to Gene Expression Omnibus
849 database under accession numbers listed:

850 Arabidopsis methylome (GSE106309), Arabidopsis *msh1* memory 4 week old plant RNAseq
851 (GSE106536), Arabidopsis 10 days old seedling 5-azacytidine treatment RNAseq (GSE109164), Tomato
852 methylome (GSE105008).
853

854 ## Authors' contributions

855 R.S. developed the application of the information thermodynamic theory on cytosine DNA methylation
856 and conducted mathematical and computational biology analyses, XY, HK and YW designed and
857 conducted biological experiments, JRB conducted computation. SM designed experiments, participated in
858 data analysis and wrote manuscript.
859

860 ### Competing interests

861 S. Mackenzie has served as co-founder for a company that tests the MSH1 system for possible
862 agricultural commercial value.
863

864 ### Consent for publication

865 Not applicable
866

867 ### Ethics approval and consent to participate

868 Not applicable
869

Figures

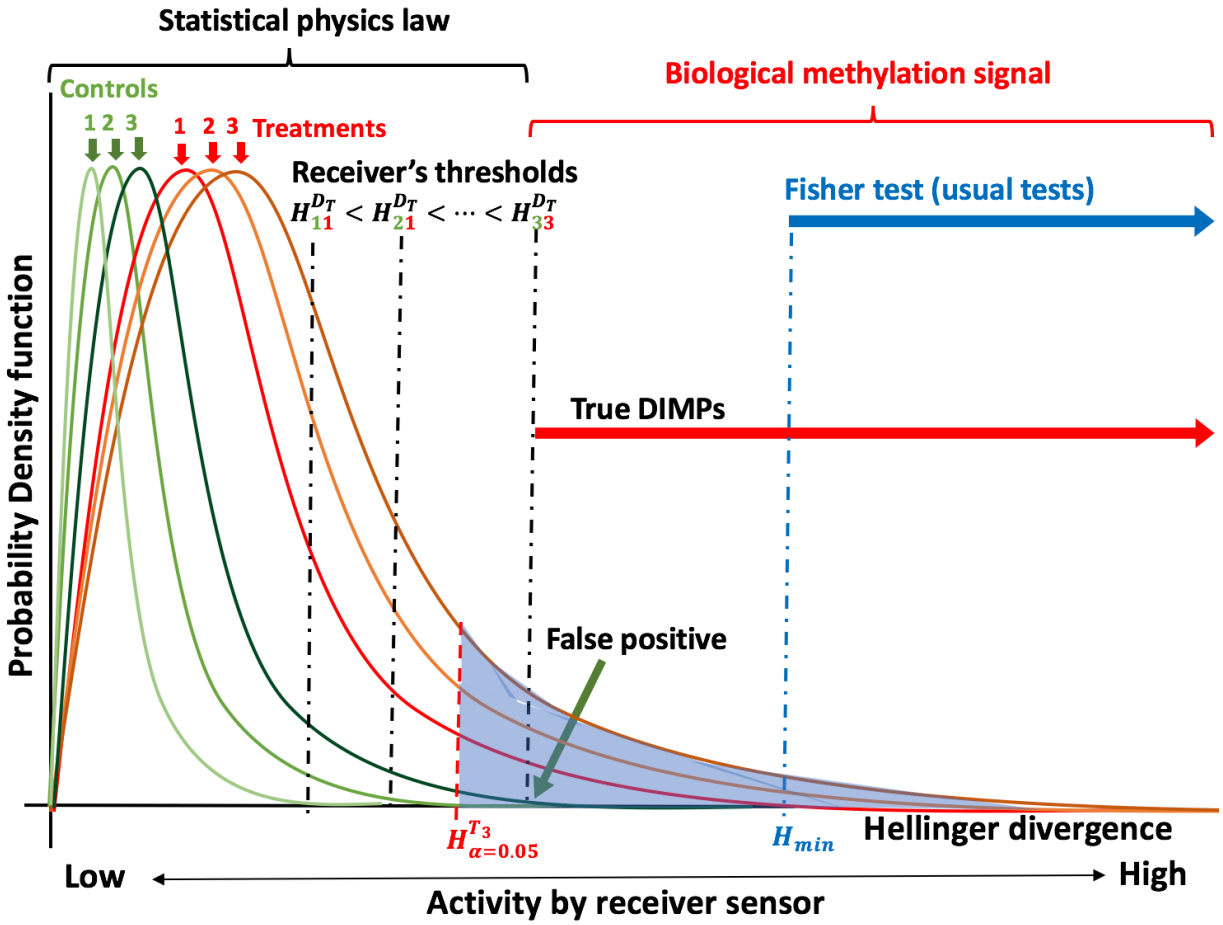


Fig. 1 Diagrammatic representation of the theoretical principle behind Methyl-IT. Methyl-IT is designed to identify a statistically significant cutoff between thermal system noise (conforming to laws of statistical physics) and treatment signal (biological methylation signal), based on Hellinger divergence (H), to identify "true" differentially informative methylation positions (DIMPs). Empirical comparisons allow the placement of Fisher's exact test for discrimination of DMPs.

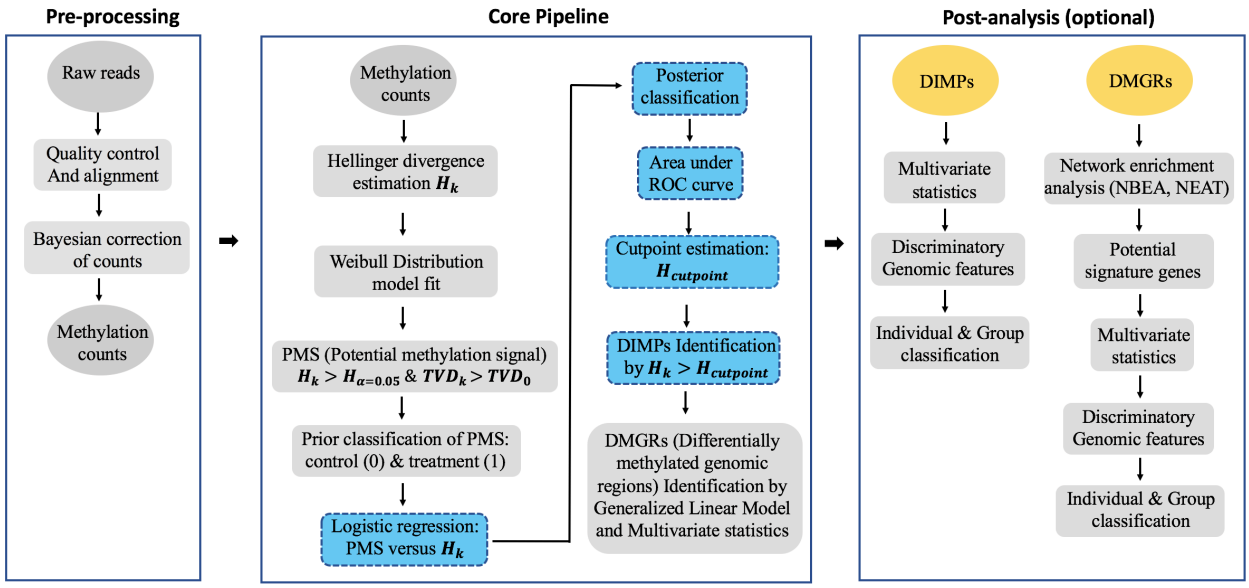


Fig. 2 Methyl-IT processing flowchart. Ovals represent input and output data, squares represent processing steps, with signal detection processing steps highlighted in blue and DIMPs and DMGRs, as main outputs of Methyl-IT, highlighted in yellow. The generalized linear model is incorporated for group comparison of genomic regions (GRs) based on the number of DIMPs in the treatment group relative to control group. DIMPs and DMGRs can be subjected to further statistical analyses to perform network enrichment analysis and to identify potential signature genes, multivariate statistical analysis (and machine learning applications) for individual and group classifications.

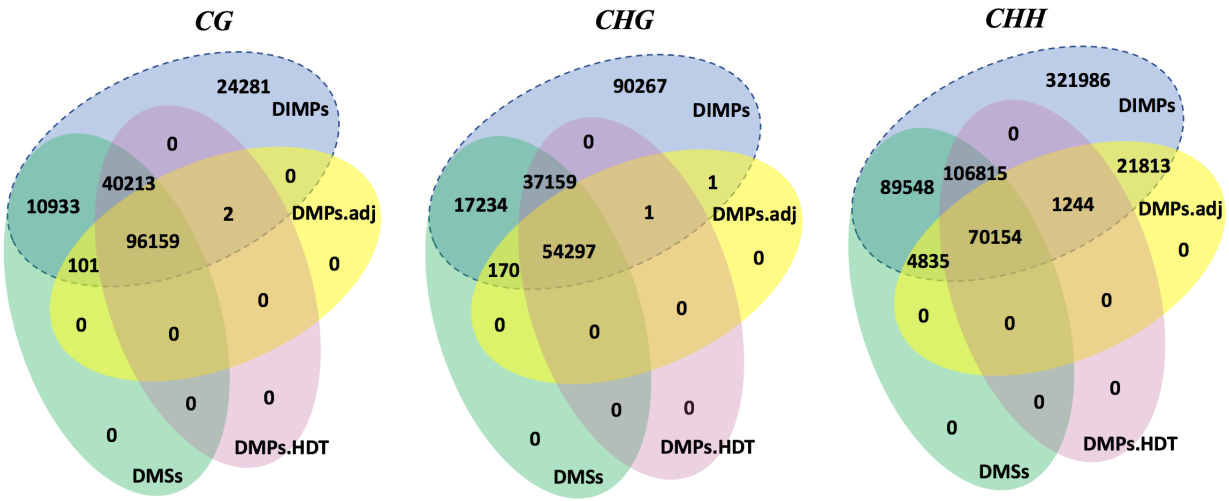


Fig. 3 Venn diagrams of overlapping DMSs (RMST implemented in Methylpy software), DMPs.adj (obtained with Fisher Exact Test), DMPs (DMPs.HDT, obtained with HDT, see methods) and DIMPs (obtained with Methyl-IT) in the memory line. Only methylated cytosine positions with total variation distance (TVD) greater than 0.23 (23% of methylation level difference) are shown for the three methylation contexts. Only DIMPs carry methylation signal (region within the dashed oval). Notice that any DMPs and DMSs outside the dashed oval (if any would be found in a different dataset or for TVD < 0.23) follow a Weibull distribution on a statistical mechanical basis as described in Fig.1. In such a case, with high probability, these DMPs and DMSs correspond to "background" methylation patterning and do not correspond to signal. This background effects can be discriminated by application of a signal detection step against a specific control (in this case, wild type Col-0 under the same experimental conditions).

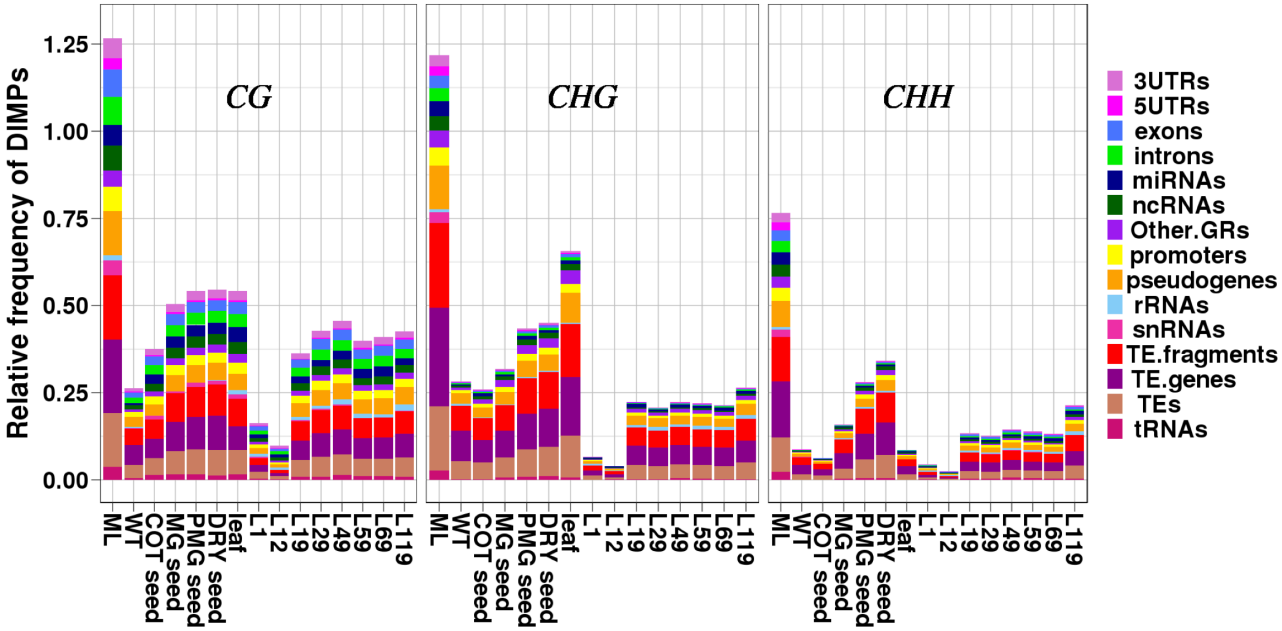


Fig. 4 Results of signal detection with Methyl-IT for genome-wide methylome data from the *msh1*-memory line (ML), a Col-0 wildtype pool (WT), seed development data from Kawakatsu et al [18] at five seed stages (GLOB, COT, MG, PMG, DRY) and leaf (globular (GLOB) stage used as control), and various Col-0 generational lineage samples (L1-L119) taken from Becker et al [3]. The experimental results provide a direct, scaled comparison of methylation signal between datasets. The relative frequency of DIMPs was estimated as the number of DIMPs divided by the number of cytosine positions.

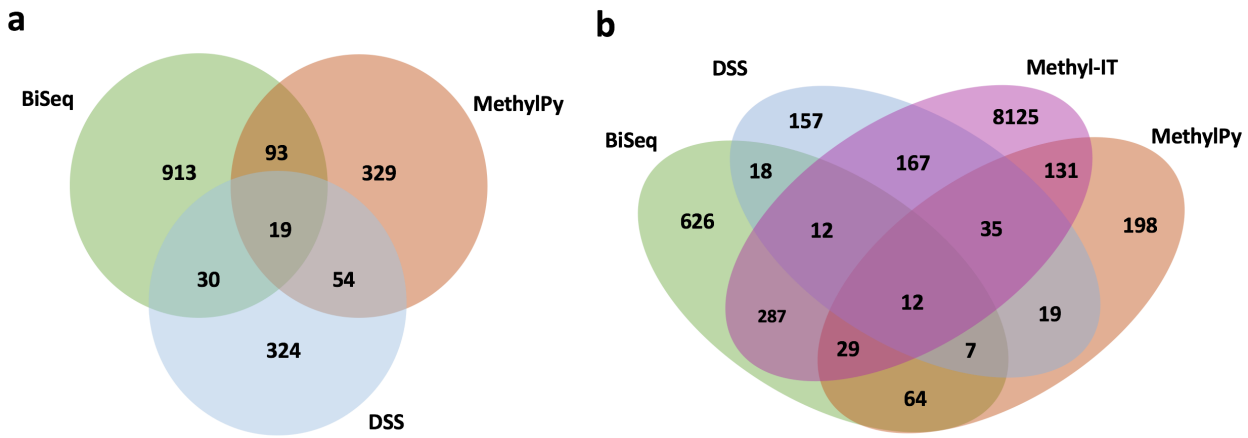


Fig. 5 Comparison of DMR associated genes identified by DSS, BiSeq, MethylPy and DMGs predicted by Methyl-IT for *msh1* memory dataset. (**a**) Venn Diagram showing a comparison of DMR associated genes (DAGs) identified with the three methylome analysis programs DSS, BiSeq and MethylPy. (**b**) Venn Diagram showing a comparison of differentially methylated genes (DMGs) identified with Methyl-IT and the DAGs with the methylome analysis programs DSS, BiSeq, MethylPy. DMGs for gene regions plus 2kb upstream and downstream are shown, and only DIMPs with *TVD* > 0.15 were counted for DMG estimations.

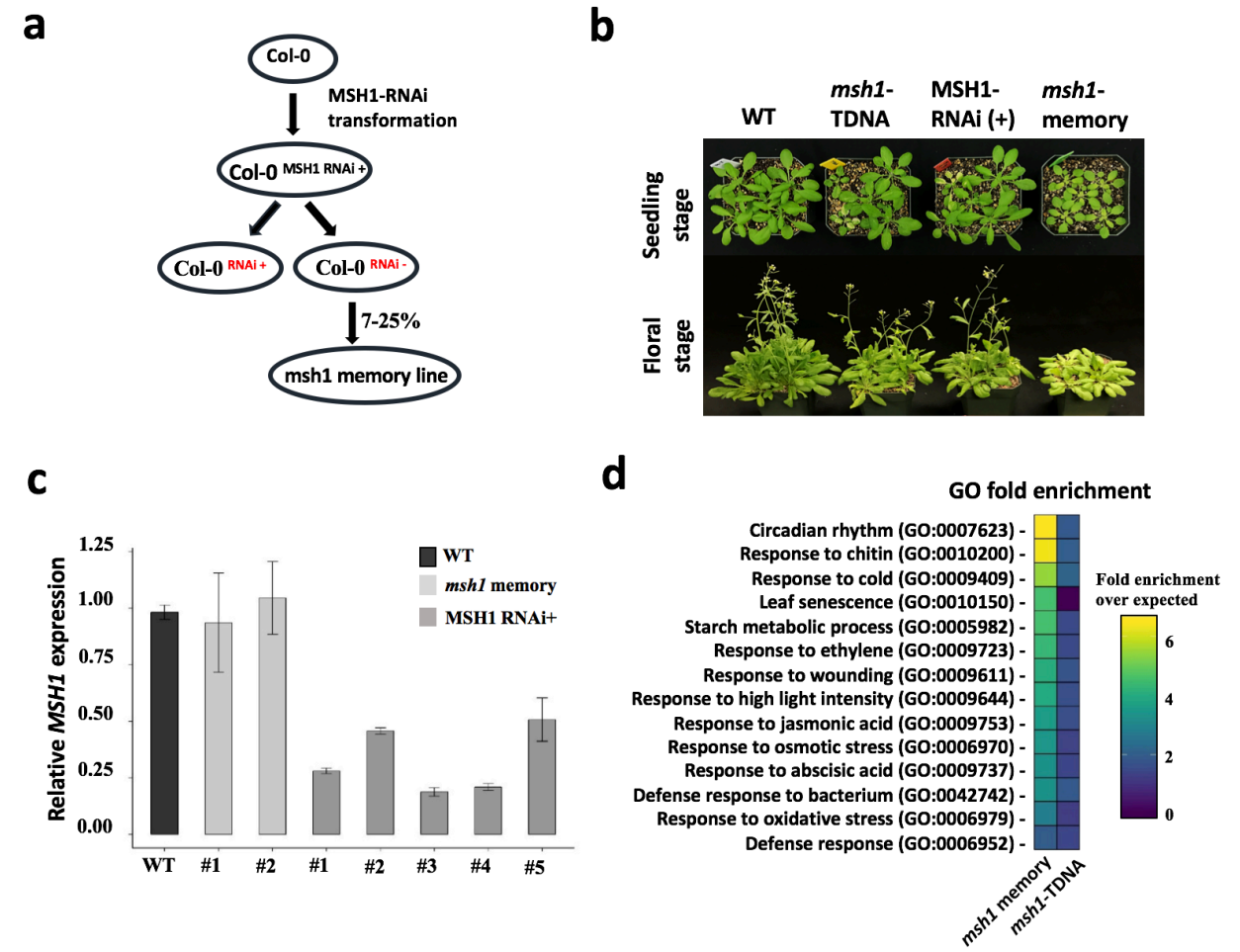


Fig. 6 *MSH1* disruption produces transgenerational memory. (**a**) Pedigree of *msh1* memory line. (**b**) Phenotypic range in different types of *MSH1*-derived development reprogramming, with *msh1* memory plants uniformly reduced in growth rate, delayed flowering and pale leaves. Seedling stage photo at 4 weeks and floral stage at 6 weeks. (**c**) *MSH1* expression levels in *msh1* memory and *MSH1*- RNAi line. Each column represents one individual plant, error bars represent ± SD of 9 technical replicates. (**d**) Functional enrichment analysis of differentially expressed genes in *msh1* memory line and *msh1* T-DNA mutant. GO enrichment categories (above cutoff FDR<0.01) are shown.

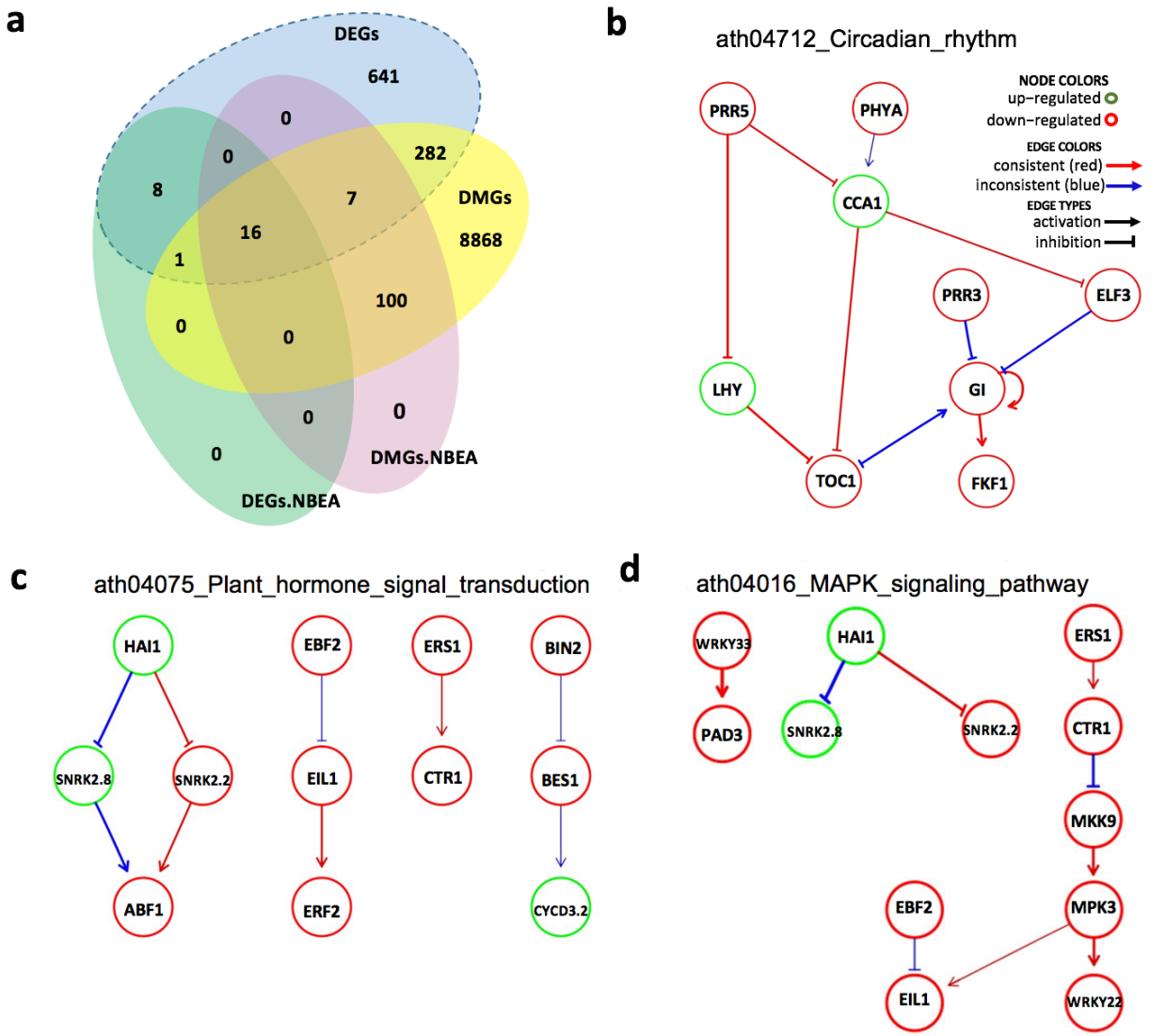
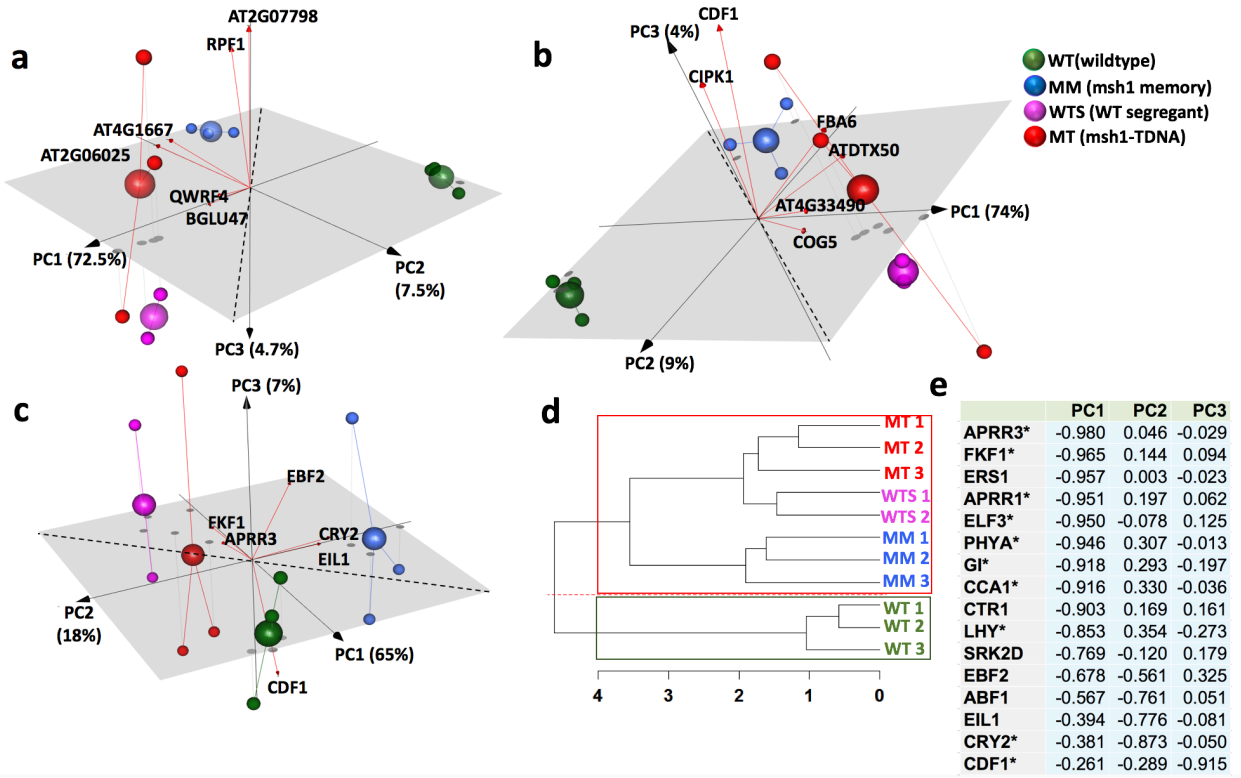


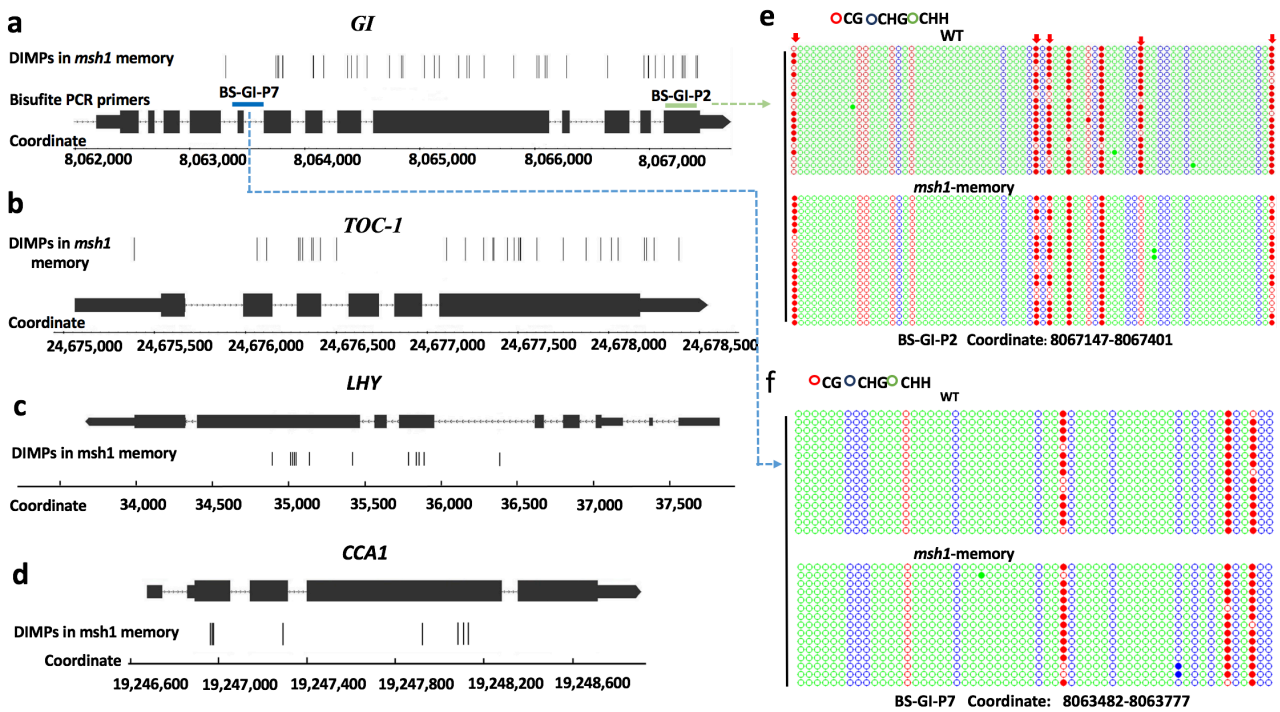
Fig. 7 Application of network-based enrichment analysis (NBEA) on Methyl-IT-based differentially methylated genes (DMGs) identifies signature pathways associated with *msh1* memory phenotype. (**a**) Venn diagram showing intersection between independent assays of *msh1* memory-associated gene expression and methylation changes. The main intersection from DMGs and DEGs datasets, and their corresponding result with the application of NBEA, identified 16 putative regulatory loci. (**b-d**) Examples of identified regulatory genes and the network in which they participate. The expression change (up, green or down, red) is indicated, as well as the inconsistent change trends, marked as blue lines.



	PC1	PC2	PC3
APRR3*	-0.980	0.046	-0.029
FKF1*	-0.965	0.144	0.094
ERS1	-0.957	0.003	-0.023
APRR1*	-0.951	0.197	0.062
ELF3*	-0.950	-0.078	0.125
PHYA*	-0.946	0.307	-0.013
GI*	-0.918	0.293	-0.197
CCA1*	-0.916	0.330	-0.036
CTR1	-0.903	0.169	0.161
LHY*	-0.853	0.354	-0.273
SRK2D	-0.769	-0.120	0.179
EBF2	-0.678	-0.561	0.325
ABF1	-0.567	-0.761	0.051
EIL1	-0.394	-0.776	-0.081
CRY2*	-0.381	-0.873	-0.050
CDF1*	-0.261	-0.289	-0.915

Fig. 8 Principal component analysis (PCA) and classification of individual samples based on genic CG methylation identifies primary contributors to the memory effect. (**a**) A three-dimensional representation of PCA outcomes with the set of all differentially methylated genes (DMGs). Samples are color-coded; "wild type segregant" (WTS) represents a wild type plant derived from crossing of the *msh1* T-DNA mutant with wild type Col-0, while "wild type" (WT) represents Col-0. The centroid from each group is represented by a large sphere connected by straight lines to smaller ones representing individual groups. Red arrows represent the magnitude and direction of the contributions to each PC by the first two genes with the greatest loadings. The square of the loadings reveals the proportion of variance of one variable explained by one principal component, while its sign gives the direction of gene contribution to a given component. (**b**) PCA performed at the intersection of DEGs and DMGs. (**c**) PCA performed at the intersection of the DMG and DEG subsets derived from independent network-based enrichment analyses (NBEA-DMG/NBEA-DEG) (see Fig. 7). These are genes involved in regulatory pathways. Since the two first PCs carry most of the total explained variance, panels A, B, and C, suggest that the weight of the sample classification rests on the planes defined by PC1 and PC2 (PC1-PC2), as observed in the projections of the spheres (shadows) on these planes. A straight-line can be drawn on the planes PC1-PC2 (black dashed-line) to clearly classify the samples into two groups, wild type versus msh1 effect (WTS, DW, and MM). Thus, there is a discriminant function or a support vector to accomplish the classification. (**d**) Hierarchical clustering with individual PC coordinates from the PCA on the intersection subset

959 NBEA-DMG/NBEA-DEG. (**e**) Correlation of genes from the subset NBEA-DMG/NBEA-DEG with the
960 first three principal components. All the genes reported in (e) carry a negative contribution to PC1 (which
961 carries a total explained variance of about 65%). The effect of these genes significantly separates the
962 msh1 effect (DW, WTS, and ML) from the wildtype control (WT). Asterisks indicate genes included in
963 the list of 16 signatures for msh1 memory.
964
965



966

967 **Fig. 9** Altered methylation was assayed at circadian clock loci. DIMP calling by Methyl-IT (only CG
968 shown) in the *msh1* memory line at *GI* (**a**), *TOC-1*(**b**), *LHY*(**c**) and *CCA1*(**d**) regions is represented by
969 black vertical bars. The gene structure and coordinates were adopted from TAIR10, with thickest bar for
970 exons, medium bar for UTRs, and dotted line for introns. DIMP calling was further confirmed by specific
971 bisulfite-PCR sequencing. The green bar represents the amplification interval designed to detect DIMPs
972 within the *GI* gene, and the blue bar represent the interval used as negative control (no DIMPs predicted).
973 The PCR result is presented in (**e**) for primer set BS-GI-P2 and (**f**) for primer set BS-GI-P7. Dot-plot
974 analysis was applied to bisulfite sequencing result. Red, blue, and green circles represent CG, CHG and
975 CHH respectively (methylation solid, no methylation blank). Each line represents one clone sequenced,
976 and at least 15 clones were sequenced for each PCR reaction.
977

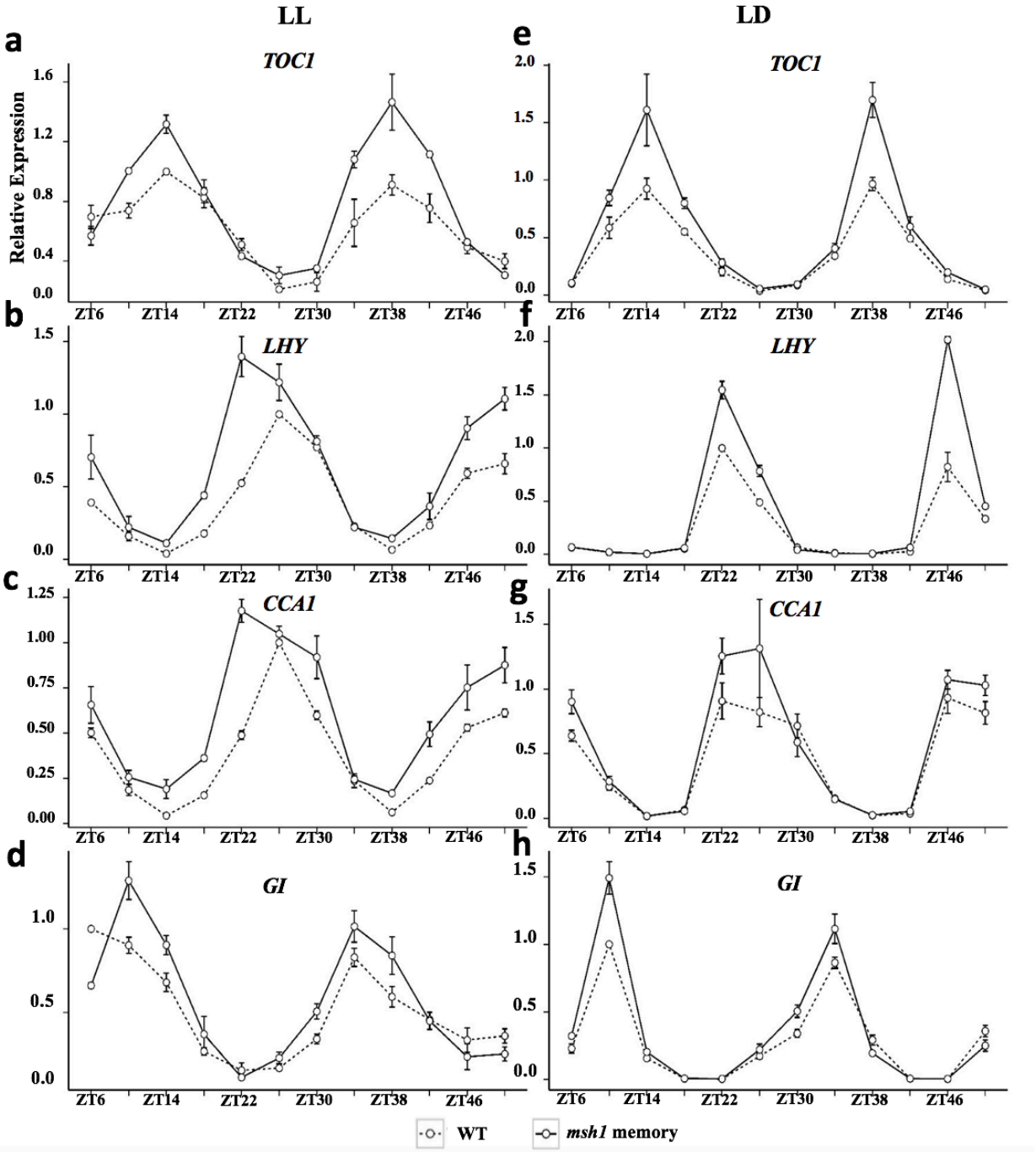


Fig. 10 Test of altered circadian behavior in the Arabidopsis *msh1* memory line. Relative transcript levels of indicated genes in wild type (dashed line) and *msh1* memory (solid line) grown under LL (24 hours light) following entrainment for 4 weeks under LD (12 hours light,12 hours dark) (a, b, c, d) or retained under LD (e, f, g, h). Zeitgeber time (ZT) indicates the sampling time (with ZT0 when light starts). Transcript levels were measured by qPCR, and expression levels were normalized to the highest peak of WT control. Error bars represent mean ± SD of three independent biological replicates.

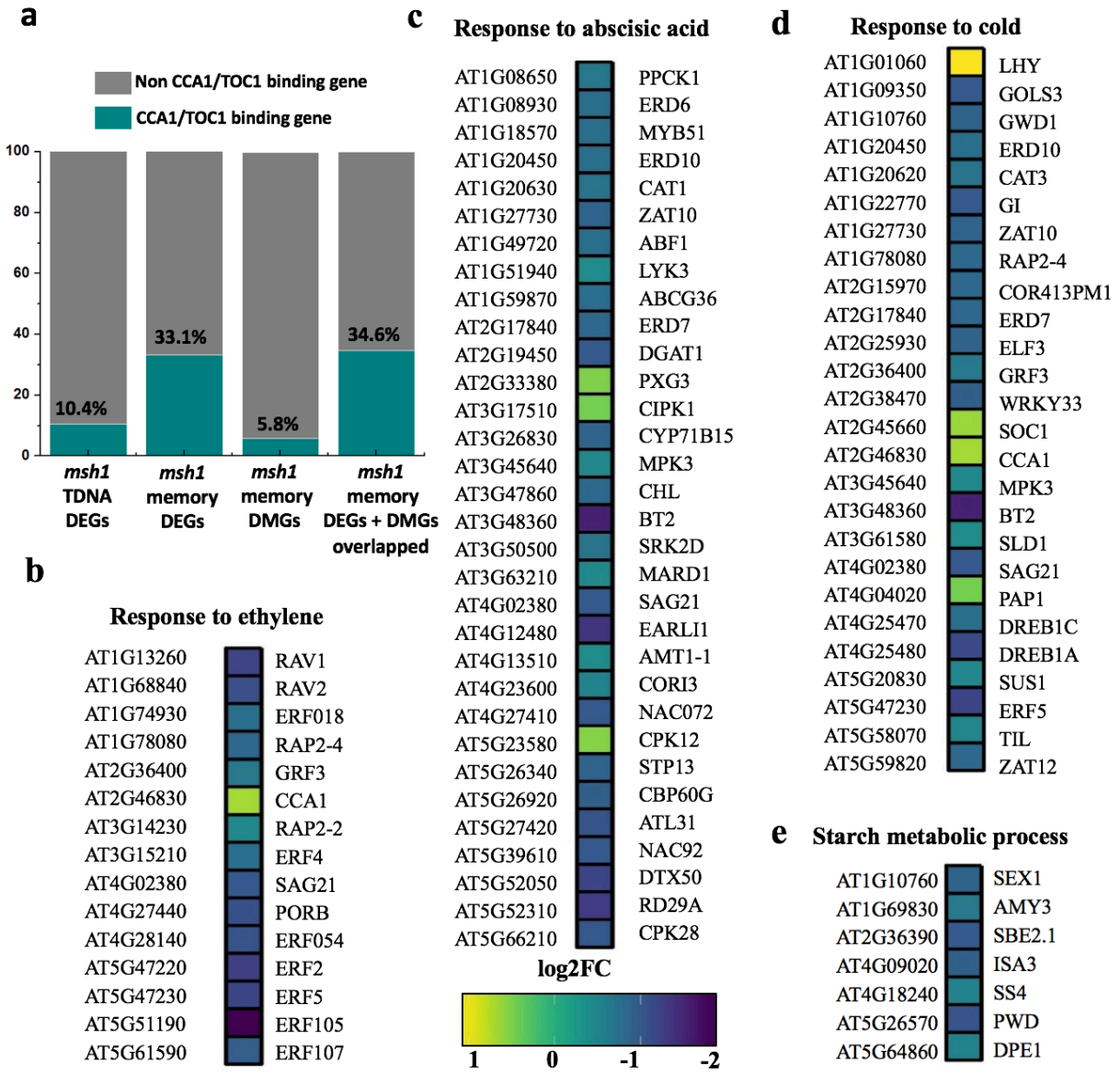


Fig. 11 RNA-seq analysis of expression in circadian clock-regulated genes in the Arabidopsis *msh1* memory line. (**a**) Genes under *TOC1* and *CCA1* regulation are represented at 10.4% in *msh1* TDNA DEGs, increasing to 33.1% in *msh1* memory line DEGs. The analysis used published *CCA1* and *TOC1* binding site CHIP-seq data [38, 64] and RNA-seq data from *msh1* TDNA and *msh1* memory line. Selected, significantly altered, circadian clock-regulated pathways in *msh1* memory line are shown as (**b**) response to ethylene, (**c**) response to abscisic acid, (**d**) response to cold, and (**e**) starch metabolic process. Important genes in each pathway are listed with expression level, compared to wild type, indicated by color boxes. The full list of DEGs can be found in Additional File 6: Table S5 for *msh1* memory, and Additional File 7: Table S8 for *msh1* TDNA.

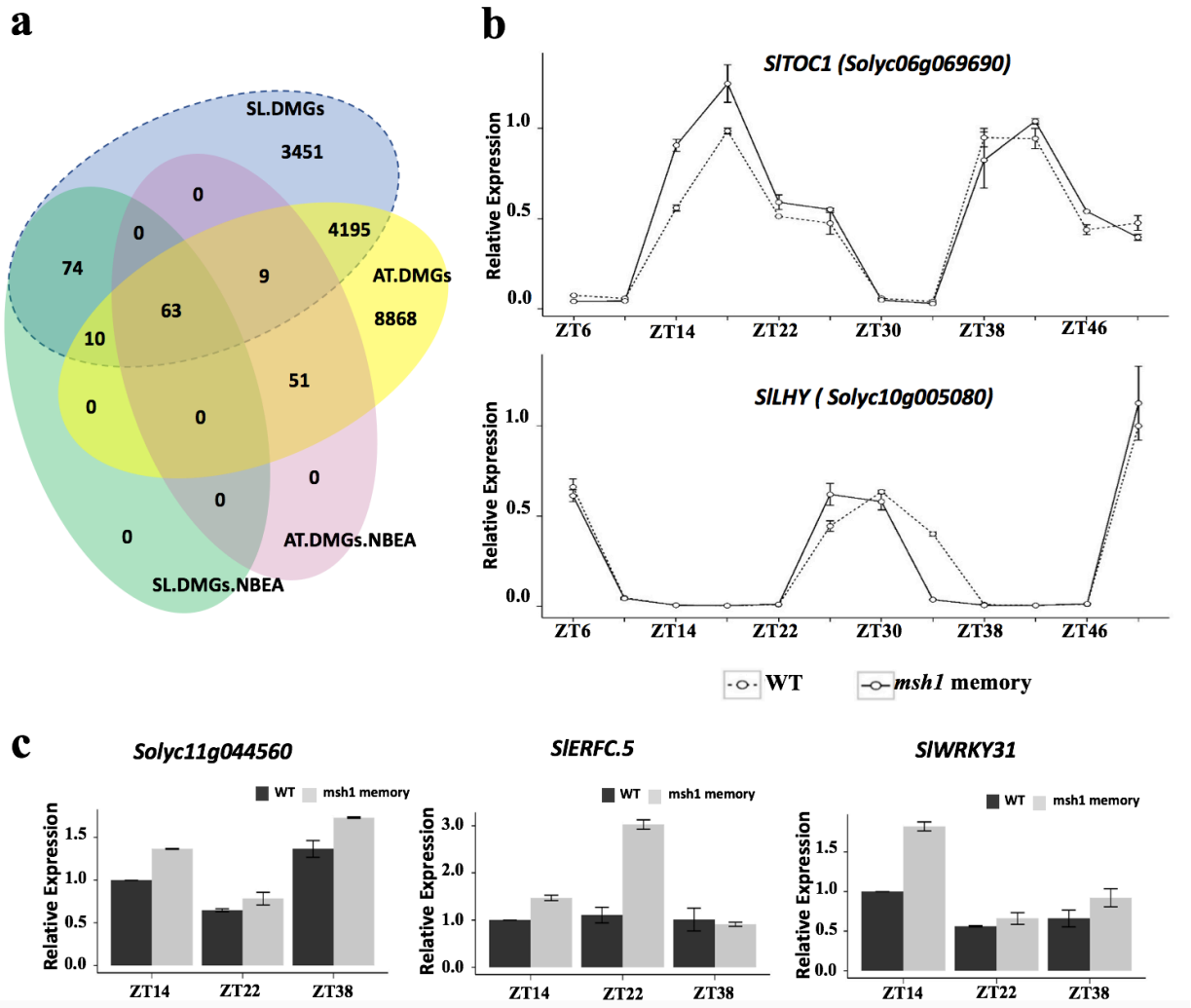


Fig. 12 Testing altered methylation pattern, circadian rhythm core genes and downstream gene expression behavior in the tomato *msh1* memory line. (**a**) Venn Diagram for DMGs in tomato *msh1* memory (SL. DMGs) versus Arabidopsis *msh1* memory (AT. DMGs) and their corresponding NBEA subsets. For SL. DMGs, only the best two mappings of each tomato gene to an Arabidopsis locus, obtained with BLAST aligner, were taken into account. (**b**) Expression patterns of tomato *TOC1* and *LHY* in wild type (circle, dashed line) and *msh1* memory (circle, solid line) grown under LD (12 hours light,12 hours dark) were assayed by quantitative real-time PCR. (**c**) The expression patterns of three circadian clock-regulated genes, *SlABF* (Solyc11g044560)*, SlERFC.5*(Solyc02g077370), and *SlWRKY31*(Solyc06g066370) were assayed by quantitative real-time PCR under LD (12 hours light, 12 hours day) conditions. For both (**b**) and (c), relative expression was calculated by normalizing to the highest value of corresponding wild type in each biological replicate. Error bars represent mean ± SD of three independent biological replicates.

1011
1012
1013
1014
1015
1016
1017
1018

Tables

Table 1. Relative sensitivity differences between several statistical tests applied to identify differentially methylated cytosines. P-values for the 2x2 contingency table with read counts $n_i^{mC_c} = 8$, $n_i^{C_c} = 2$, $n_i^{mC_t} = 350$, and $n_i^{C_t} = 20$.

Approach	*p*-value
FET	0.108615
FET one tail	0.108615
FET *p*.value MC 3k[(1)]	0.1086
RMST Boot 3k[(2)]	0.051
HT Boot 3k [(3)]	0.050667
Weibull ML1 CG[(4)]	0.000118
Weibull ML2 CG[(4)]	1.67E-05
Weibull ML3 CG[(4)]	2.21E-05

[1]*p*.value simulated with Monte Carlo (MC) simulation with 3000 resamplings (3k). [2]Bootstrap goodness-of-fit RMST as implemented in Methylpy [17]. [3]Bootstrap goodness-of-fit test based on Hellinger divergence estimated according to the first statistic given Theorem 1 from reference [27]. [3]*p*-value based on the Weibull distribution for memory lines (ML 1 to 3). n_i^{mCc} refers to methylated cytosine counts in control, n_i^{Cc} refers to non-methylated cytosine counts in control, n_i^{mCt} refers to methylated cytosine counts in treatment and n_i^{Ct} refers to non-methylated cytosine counts in treatment.The R script to compute RMST and *H* MC estimation is provided in GitLab: https://git.psu.edu/genomath/MethylIT

1033 **Table 2.** Network enrichment analysis test (NEAT) on the set of GO-biological process (BP-GO) for the
1034 differentially methylated genes in Ws-0 seed development dataset.
1035

BP-GO	NAB	Expected NAB	Adj. *p*-value
GO:0000902 cell morphogenesis	3	0.2492	0.00280
GO:0006623 protein targeting to vacuole	4	0.299	< 0.001
GO:0006891 intra-Golgi vesicle-mediated transport	4	0.3323	< 0.001
GO:0009723 response to ethylene	8	2.9072	0.00873
GO:0009740 gibberellic acid mediated signaling pathway	5	0.9802	0.00375
GO:0009845 seed germination	6	1.3456	0.00301
GO:0009938 negative regulation of gibberellic acid mediated signaling pathway	4	0.2658	< 0.001
GO:0010162 seed dormancy process	5	1.03	0.00434
GO:0010187 negative regulation of seed germination	3	0.4319	0.00916
GO:0010325 raffinose family oligosaccharide biosynthetic process	5	0.3323	0.00102
GO:0016049 cell growth	3	0.3655	0.00640
GO:0016192 vesicle-mediated transport	5	0.3987	< 0.001
GO:0016197 endosomal transport	2	0.0665	0.00280
GO:0048444 floral organ morphogenesis	5	0.3323	< 0.001
GO:2000033 regulation of seed dormancy process	3	0.1994	0.0017
GO:2000377 regulation of reactive oxygen species metabolic process	4	0.4153	0.00154

1036
1037 Only over-enriched pathways are included
1038 NAB: observed number of (network) links from DMG list to GO term gene list
1039 Expected NAB: expected number of links from DMG list to GO term gene list (in absence of enrichment)
1040 Enrichment Fold: the ratio of NAB (observed number of network links) / expected nab (expected number of links)
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054

1055 **Table 3.** Putative signature genes for *msh1* memory line

	ID	Name	Annotation
1*	AT1G01060	*LHY*	circadian rhythm myb factor
2*	AT1G22770	*GI*	circadian controlled flowering
3*	AT5G60100	*APRR3*	affects the period of the circadian clock
4*	AT1G04400	*CRY2*	positive regulator of constans
5*	AT5G61380	*TOC1*	generation of circadian rhythms
6*	AT2G46830	*CCA1*	overlaps LHY in circadian oscillator
7*	AT1G09570	*PHYA*	photomorphogenesis
8*	AT2G25930	*ELF3*	core component of cirdadian clock
9*	AT1G68050	*FKF1*	complex with GI in clock regulation
10*	AT5G62430	*CDF1*	a circadian regulator of flowering time
11	AT2G40940	*ERS1*	ethylene receptor
12	AT5G25350	*EBF2*	involved in the ethylene-response pathway
13	AT2G27050	*EIL1*	ethylene activated signaling
14	AT5G03730	*CTR1*	negative regulator in the ethylene signal transduction pathway
15	AT3G50500	*SRK2D*	ABA signaling during seed germination
16	AT1G49720	*ABF1*	transcriptional regulation ABA response

1056 Genes directly associated with plant circadian core component are indicated with "*"

1057

1058 ## Additional files

1059 **Additional file 1: Figures S1 to S10**

1060

1061 **AddItional file 2: Table S1** Absolute DIMPs counts and DIMPs counts per genomic region for seed
1062 development and germination datasets

1063

1064 **Additional file 3: Table S2** DMGs Arabidopsis (ws-0) seed development dataset

1065

1066 **Additional file 4: Table S3** DMGs from Arabidopis memory line

1067

1068 **Additional file 5 Table S4** List of seed development DMGs found in networks based on NEAT

1069

1070 **Additional file 6: Table S5** Total 955 of DEGs of Arabidopsis msh1-memory-line

1071

1072 **Additional file 7: Table S6** Total 9867 DMGs of Arabidopsis TDNA mutant

1073

1074 **Additional file 8: Table S7** NBEA analysis of DEGs in Arabidopsis msh1 memory line

1075

1076 **Additional file 9: Table S8** NEAT and NBEA analysis on DMGs from arabidopsis msh1 memory line
1077
1078 **Additional file 10: Table S9** DIMPs distribution in 16 regulatory genes in msh1 memory individual
1079 plants
1080
1081 **Additional file 11: Table S10** DMGs in tomato msh1 memory line
1082
1083 **Additional file 12: Table S11** NBEA analysis of DMGs in tomato msh1 memory line
1084
1085 **Additional file 13: Table S12** Main intersection between Arabidopsis and tomato DMGs NBEA list
1086
1087 **Additional file 14: Table S13** Primers used in this paper
1088

1089 **References**

1090 1. Calarco JP, Borges F, Donoghue MTA, Van Ex F, Jullien PE, Lopes T, Gardner R,
1091 Berger F, Feijo JA, Becker JD, Martienssen RA: Reprogramming of DNA Methylation
1092 in Pollen Guides Epigenetic Inheritance via Small RNA. Cell 2012. 151:194-205.
1093 2. Schmitz RJ, Schultz MD, Lewsey MG, O'Malley RC, Urich MA, Libiger O, Schork NJ,
1094 Ecker JR: Transgenerational Epigenetic Instability Is a Source of Novel Methylation
1095 Variants. Science 2011. 334:369-373.
1096 3. Becker C, Hagmann J, Muller J, Koenig D, Stegle O, Borgwardt K, Weigel D:
1097 Spontaneous epigenetic variation in the Arabidopsis thaliana methylome. Nature
1098 2011. 480:245-U127.
1099 4. Matzke MA, Mosher RA: RNA-directed DNA methylation: an epigenetic pathway of
1100 increasing complexity (vol 15, 394, 2014). Nature Reviews Genetics 2014. 15.
1101 5. Crisp PA, Ganguly D, Eichten SR, Borevitz JO, Pogson BJ: Reconsidering plant
1102 memory: Intersections between stress recovery, RNA turnover, and epigenetics.
1103 Science Advances 2016. 2.
1104 6. Kinoshita T, Seki M: Epigenetic Memory for Stress Response and Adaptation in
1105 Plants. Plant and Cell Physiology 2014. 55:1859-1863.

1106 7. Colaneri AC, Jones AM: Genome-Wide Quantitative Identification of DNA
1107 Differentially Methylated Sites in Arabidopsis Seedlings Growing at Different Water
1108 Potential. Plos One 2013. 8.
1109 8. Jenkinson G, Pujadas E, Goutsias J, Feinberg AP: Potential energy landscapes identify
1110 the information-theoretic nature of the epigenome. Nature Genetics 2017. 49:719-+.
1111 9. Sanchez R, Mackenzie SA: Information Thermodynamics of Cytosine DNA
1112 Methylation. Plos One 2016. 11.
1113 10. Sanchez R, Mackenzie SA: Genome-Wide Discriminatory Information Patterns of
1114 Cytosine DNA Methylation. International Journal of Molecular Sciences 2016. 17.
1115 11. Greiner M, Pfeiffer D, Smith RD: Principles and practical application of the receiver-
1116 operating characteristic analysis for diagnostic tests. Prev Vet Med 2000. 45:23-41.
1117 12. Carter JV, Pan J, Rai SN, Galandiuk S: ROC-ing along: Evaluation and interpretation of
1118 receiver operating characteristic curves. Surgery 2016. 159:1638-1645.
1119 13. Harpaz R, DuMouchel W, LePendu P, Bauer-Mehren A, Ryan P, Shah NH:
1120 Performance of pharmacovigilance signal-detection algorithms for the FDA adverse
1121 event reporting system. Clin Pharmacol Ther 2013. 93:539-546.
1122 14. Kruspe S, Dickey DD, Urak KT, Blanco GN, Miller MJ, Clark KC, Burghardt E, Gutierrez
1123 WR, Phadke SD, Kamboj S, et al: Rapid and Sensitive Detection of Breast Cancer Cells
1124 in Patient Blood with Nuclease-Activated Probe Technology. Mol Ther Nucleic Acids
1125 2017. 8:542-557.
1126 15. Wu H, Xu T, Feng H, Chen L, Li B, Yao B, Qin Z, Jin P, Conneely KN: Detection of
1127 differentially methylated regions from whole-genome bisulfite sequencing data
1128 without replicates. Nucleic Acids Res 2015. 43:e141.
1129 16. Hebestreit K, Dugas M, Klein HU: Detection of significantly differentially methylated
1130 regions in targeted bisulfite sequencing data. Bioinformatics 2013. 29:1647-1653.
1131 17. Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, Rajagopal N,
1132 Nery JR, Urich MA, Chen H, et al: Human body epigenome maps reveal noncanonical
1133 DNA methylation variation. Nature 2015. 523:212-216.
1134 18. Kawakatsu T, Nery JR, Castanon R, Ecker JR: Dynamic DNA methylation
1135 reconfiguration during seed development and germination. Genome Biol 2017.
1136 18:171.

1137 19. Virdi KS, Laurie JD, Xu YZ, Yu JT, Shao MR, Sanchez R, Kundariya H, Wang D,
1138 Riethoven JJM, Wamboldt Y, et al: Arabidopsis MSH1 mutation alters the epigenome
1139 and produces heritable changes in plant growth. Nature Communications 2015. 6.
1140 20. Virdi KS, Wamboldt Y, Kundariya H, Laurie JD, Keren I, Kumar KRS, Block A, Basset
1141 G, Luebker S, Elowsky C, et al: MSH1 Is a Plant Organellar DNA Binding and
1142 Thylakoid Protein under Precise Spatial Regulation to Alter Development. Molecular
1143 Plant 2016. 9:245-260.
1144 21. Davila JI, Arrieta-Montiel MP, Wamboldt Y, Cao J, Hagmann J, Shedge V, Xu YZ,
1145 Weigel D, Mackenzie SA: Double-strand break repair processes drive evolution of
1146 the mitochondrial genome in Arabidopsis. Bmc Biology 2011. 9.
1147 22. Xu YZ, Arrieta-Montiel MP, Virdi KS, de Paula WBM, Widhalm JR, Basset GJ, Davila JI,
1148 Elthon TE, Elowsky CG, Sato SJ, et al: MutS HOMOLOG1 Is a Nucleoid Protein That
1149 Alters Mitochondrial and Plastid Properties and Plant Response to High Light. Plant
1150 Cell 2011. 23:3428-3441.
1151 23. Xu YZ, Santamaria RD, Virdi KS, Arrieta-Montiel MP, Razvi F, Li SQ, Ren GD, Yu B,
1152 Alexander D, Guo LN, et al: The Chloroplast Triggers Developmental
1153 Reprogramming When MUTS HOMOLOG1 Is Suppressed in Plants. Plant Physiology
1154 2012. 159:710-+.
1155 24. Shao MR, Raju SKK, Laurie JD, Sanchez R, Mackenzie SA: Stress-responsive pathways
1156 and small RNA changes distinguish variable developmental phenotypes caused by
1157 MSH1 loss. Bmc Plant Biology 2017. 17.
1158 25. Mónica López-Ratón MXR-Á, Carmen Cadarso-Suárez, Francisco Gude-Sampedro:
1159 OptimalCutpoints: An R Package for Selecting Optimal Cutpoints in Diagnostic Tests.
1160 Journal of statistical software 2014.61:4896.
1161 26. William Perkins MT, Rachel Ward: Computing the confidence levels for a root-mean-
1162 square test of goodness-of-fit. Applied Mathematics and Computation 2011.
1163 217:9072-9084.
1164 27. F. Liese IV: On Divergences and Informations in Statistics and Information Theory.
1165 IEEE Transactions on Information Theory 2006. 52:4394 - 4412.
1166 28. Le BH, Cheng C, Bui AQ, Wagmaister JA, Henry KF, Pelletier J, Kwong L, Belmonte M,
1167 Kirkbride R, Horvath S, et al: Global analysis of gene activity during Arabidopsis

1168 seed development and identification of seed-specific transcription factors. Proc Natl
1169 Acad Sci U S A 2010. 107:8063-8070.
1170 29. Bassel GW, Lan H, Glaab E, Gibbs DJ, Gerjets T, Krasnogor N, Bonner AJ, Holdsworth
1171 MJ, Provart NJ: Genome-wide network model capturing seed germination reveals
1172 coordinated regulation of plant cellular phase transitions. Proc Natl Acad Sci U S A
1173 2011. 108:9709-9714.
1174 30. Yang XD, Kundariya H, Xu YZ, Sandhu A, Yu JT, Hutton SF, Zhang MF, Mackenzie SA:
1175 MutS HOMOLOG1-Derived Epigenetic Breeding Potential in Tomato. Plant
1176 Physiology 2015. 168:222-U390.
1177 31. Sanchez SE, Kay SA: The Plant Circadian Clock: From a Simple Timekeeper to a
1178 Complex Developmental Manager. Cold Spring Harbor Perspectives in Biology 2016,
1179 8.
1180 32. Lefebvre A, Mauffret O, el Antri S, Monnot M, Lescot E, Fermandjian S: Sequence
1181 dependent effects of CpG cytosine methylation. A joint 1H-NMR and 31P-NMR study.
1182 Eur J Biochem 1995. 229:445-454.
1183 33. Nathan D, Crothers DM: Bending and flexibility of methylated and unmethylated
1184 EcoRI DNA. J Mol Biol 2002. 316:7-17.
1185 34. Severin PM, Zou X, Gaub HE, Schulten K: Cytosine methylation alters DNA
1186 mechanical properties. Nucleic Acids Res 2011. 39:8740-8751.
1187 35. Huang SC, Ecker JR: Piecing together cis-regulatory networks: insights from
1188 epigenomics studies in plants. Wiley Interdiscip Rev Syst Biol Med 2017.
1189 36. Marchal C, Miotto B: Emerging concept in DNA methylation: role of transcription
1190 factors in shaping DNA methylation patterns. J Cell Physiol 2015. 230:743-751.
1191 37. Naftelberg S, Schor IE, Ast G, Kornblihtt AR: Regulation of Alternative Splicing
1192 Through Coupling with Transcription and Chromatin Structure. Annual Review of
1193 Biochemistry, Vol 84 2015. 84:165-198.
1194 38. Nagel DH, Doherty CJ, Pruneda-Paz JL, Schmitz RJ, Ecker JR, Kay SA: Genome-wide
1195 identification of CCA1 targets uncovers an expanded clock network in Arabidopsis.
1196 Proceedings of the National Academy of Sciences of the United States of America
1197 2015. 112:E4802-E4810.

1198 39. De Lucia F, Crevillen P, Jones AME, Greb T, Dean C: A PHD-Polycomb Repressive
1199 Complex 2 triggers the epigenetic silencing of FLC during vernalization. Proceedings
1200 of the National Academy of Sciences of the United States of America 2008.
1201 105:16831-16836.
1202 40. Grundy J, Stoker C, Carre IA: Circadian regulation of abiotic stress tolerance in
1203 plants. Frontiers in Plant Science 2015. 6.
1204 41. Lee KH, Piao HL, Kim HY, Choi SM, Jiang F, Hartung W, Hwang I, Kwak JM, Lee IJ,
1205 Hwang I: Activation of glucosidase via stress-induced polymerization rapidly
1206 increases active pools of abscisic acid. Cell 2006. 126:1109-1120.
1207 42. Legnaioli T, Cuevas J, Mas P: TOC1 functions as a molecular switch connecting the
1208 circadian clock with plant responses to drought. Embo Journal 2009. 28:3745-3757.
1209 43. Graf A, Schlereth A, Stitt M, Smith AM: Circadian control of carbohydrate availability
1210 for growth in Arabidopsis plants at night. Proceedings of the National Academy of
1211 Sciences of the United States of America 2010. 107:9458-9463.
1212 44. Ni ZF, Kim ED, Ha MS, Lackey E, Liu JX, Zhang YR, Sun QX, Chen ZJ: Altered circadian
1213 rhythms regulate growth vigour in hybrids and allopolyploids. Nature 2009.
1214 457:327-U327.
1215 45. Miller M, Song QX, Shi XL, Juenger TE, Chen ZJ: Natural variation in timing of stress-
1216 responsive gene expression predicts heterosis in intraspecific hybrids of
1217 Arabidopsis. Nature Communications 2015. 6.
1218 46. Krueger F, Andrews SR: Bismark: a flexible aligner and methylation caller for
1219 Bisulfite-Seq applications. Bioinformatics 2011. 27:1571-1572.
1220 47. Prezza N, Vezzi F, Kaller M, Policriti A: Fast, accurate, and lightweight analysis of BS-
1221 treated reads with ERNE 2. Bmc Bioinformatics 2016. 17.
1222 48. Steerneman T: On the Total Variation and Hellinger Distance between Signed
1223 Measures - an Application to Product Measures. Proceedings of the American
1224 Mathematical Society 1983. 88:684-688.
1225 49. R Core Team: A language and environment for statistical computing. 2016.
1226 50. Hippenstiel RD: Detection Theory: Applications and Digital Signal Processing. CRC
1227 Press 2001.

1228 51. Stanislaw H, Todorov N: Calculation of signal detection theory measures. Behavior
1229 Research Methods Instruments & Computers 1999. 31:137-149.
1230 52. Youden WJ: Index for rating diagnostic tests. Cancer 1950. 3:32-35.
1231 53. Perkins NJ, Schisterman EF: The inconsistency of "optimal" cutpoints obtained using
1232 two criteria based on the receiver operating characteristic curve. Am J Epidemiol
1233 2006. 163:670-675.
1234 54. Carstensen B, Plummer, M., Laara, E. & Hills, M.: Epi:A Package for Statistical
1235 Analysis in Epidemiology. R package version 27 2016.
1236 55. Love MI, Huber W, Anders S: Moderated estimation of fold change and dispersion
1237 for RNA-seq data with DESeq2. Genome Biology 2014. 15.
1238 56. He Y, Gorkin DU, Dickel DE, Nery JR, Castanon RG, Lee AY, Shen Y, Visel A,
1239 Pennacchio LA, Ren B, Ecker JR: Improved regulatory element prediction based on
1240 tissue-specific local epigenomic signatures. Proc Natl Acad Sci U S A 2017.
1241 114:E1633-E1640.
1242 57. Langmead B, Salzberg SL: Fast gapped-read alignment with Bowtie 2. Nature
1243 Methods 2012. 9:357-U354.
1244 58. Wang XF, Yu XQ, Zhu W, McCombie WR, Antoniou E, Powers RS, Davidson NO, Li E,
1245 Williams J: A trimming-and-retrieving alignment scheme for reduced representation
1246 bisulfite sequencing. Bioinformatics 2015. 31:2040-2042.
1247 59. Geistlinger L, Csaba G, Zimmer R: Bioconductor's EnrichmentBrowser: seamless
1248 navigation through combined results of set- & network-based enrichment analysis.
1249 Bmc Bioinformatics 2016. 17.
1250 60. Signorelli M, Vinciotti V, Wit EC: NEAT: an efficient network enrichment analysis
1251 test. Bmc Bioinformatics 2016. 17.
1252 61. Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, Mueller R, Nolan T,
1253 Pfaffl MW, Shipley GL, et al: The MIQE Guidelines: Minimum Information for
1254 Publication of Quantitative Real-Time PCR Experiments. Clinical Chemistry 2009.
1255 55:611-622.
1256 62. Grant GR, Farkas MH, Pizarro AD, Lahens NF, Schug J, Brunk BP, Stoeckert CJ,
1257 Hogenesch JB, Pierce EA: Comparative analysis of RNA-Seq alignment algorithms
1258 and the RNA-Seq unified mapper (RUM). Bioinformatics 2011. 27:2518-2528.

1259 63. Hartley SW, Mullikin JC: QoRTs: a comprehensive toolset for quality control and data
1260 processing of RNA-Seq experiments. Bmc Bioinformatics 2015. 16.
1261 64. Huang W, Perez-Garcia P, Pokhilko A, Millar AJ, Antoshechkin I, Riechmann JL, Mas
1262 P: Mapping the Core of the Arabidopsis Circadian Clock Defines the Network
1263 Structure of the Oscillator. Science 2012. 336:75-79.
1264
1265
1266