doi:10.1038/nature10555

Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome

Claude Becker[1]*, Jörg Hagmann[1]*, Jonas Müller[1], Daniel Koenig[1], Oliver Stegle[2], Karsten Borgwardt[2] & Detlef Weigel[1]

Heritable epigenetic polymorphisms, such as differential cytosine methylation, can underlie phenotypic variation[1,2]. Moreover, wild strains of the plant *Arabidopsis thaliana* differ in many epialleles[3,4], and these can influence the expression of nearby genes[1,2]. However, to understand their role in evolution[5], it is imperative to ascertain the emergence rate and stability of epialleles, including those that are not due to structural variation. We have compared genome-wide DNA methylation among 10 *A. thaliana* lines, derived 30 generations ago from a common ancestor[6]. Epimutations at individual positions were easily detected, and close to 30,000 cytosines in each strain were differentially methylated. In contrast, larger regions of contiguous methylation were much more stable, and the frequency of changes was in the same low range as that of DNA mutations[7]. Like individual positions, the same regions were often affected by differential methylation in independent lines, with evidence for recurrent cycles of forward and reverse mutations. Transposable elements and short interfering RNAs have been causally linked to DNA methylation[8]. In agreement, differentially methylated sites were farther from transposable elements and showed less association with short interfering RNA expression than invariant positions. The biased distribution and frequent reversion of epimutations have important implications for the potential contribution of sequence-independent epialleles to plant evolution.

Although there is no doubt that DNA sequence mutations are the primary raw material for evolutionary change, local DNA methylation variants with major effects on the expression of nearby genes can be inherited over many generations[1,2]. However, such epialleles are not always as stable as the primary DNA sequence[3,9–11]. New sequencing technologies have recently enabled the direct determination of spontaneous DNA mutation rates[12], and we have previously reported that *A. thaliana* experiences about one single-base-pair mutation per haploid genome and generation[7]. This analysis was based on a set of five mutation accumulation lines that had been derived from a single individual of the inbred strain used to produce the high-quality reference genome sequence for *A. thaliana*. These lines had been separately propagated in a common environment by single-seed descent for 30 generations[6]. We examined whole-genome cytosine methylation[13,14] in these five lines plus five additional lines of the same population by Illumina sequencing. We interrogated two siblings each of the 31st generation with an average strand-specific coverage depth of 20× per individual; changes shared within a line should predominantly reflect differences that had accumulated by the 30th generation. Because seeds from the founders were no longer available, we compared the 31st generation individuals to two independent lines that had been propagated for only three generations (Supplementary Fig. 1).

Out of all cytosine residues with high-quality sequencing support (see Supplementary Methods), on average 2.8 million were found to be methylated in each line (Supplementary Table 1). The higher genome-wide methylation rate in our analysis compared to previous studies[13,14] reflects the greater statistical power afforded by increased sequencing depth. We subsequently evaluated 13.9 million cytosines that had at least threefold coverage in all individuals, of which 3 million were methylated in at least one strain. Using Fisher's exact test, we identified about 186,000 (6.2%) positions with a significant change in methylation (false discovery rate <0.05) between at least one 31st generation and both 3rd generation lines. Almost all, 99.6%, of these differentially methylated positions (DMPs) were also detected with an entropy-based method[15]. Given the limited statistical power for weakly methylated or poorly covered sites (Supplementary Figs 2–4), our DMP estimate would almost certainly increase with higher sequencing depth. For further analyses, we considered sites that agreed between 31st generation siblings (on average, 99.8%) and between the two strains closest to the founder generation (99.7%).

CG sites were highly over-represented among DMPs (Fig. 1a). This is unlikely to reflect greater instability of CG compared to CHG and CHH positions (where H is A, T or C), but rather higher statistical power in detecting a change at CG sites, which are on average much more highly methylated[13,14] (Supplementary Fig. 4). Among CG sites in genic regions, including those producing non-coding RNAs, relative abundance of DMPs was two- to fourfold higher compared with non-differentially methylated positions (N-DMPs). The opposite was the case for CG positions in transposable elements and intergenic regions, with a similar, but less pronounced, bias for CHG and CHH sites (Fig. 1b). These observations were in agreement with CG-DMPs being found most often on chromosome arms, which have the highest gene density (Fig. 1c), even though cytosine methylation near the centromeres is the highest[13,14]. Gene body methylation gradually increases towards the 3′ end, before sharply decreasing at the end of the last exon[13,14,16,17], although genes 1 kb or less in length were generally only weakly methylated (Supplementary Fig. 5a). The profiles of DMPs and N-DMPs were similar across individual genes, exons, introns and transposable elements (Fig. 1d and Supplementary Fig. 5b, c), but DMPs were less frequent in promoter and downstream regions. Notably, CG-DMPs accounted for 42% of methylated sites in gene bodies, despite all CG-N-DMPs outnumbering CG-DMPs four to one.

Twenty-four-nucleotide-long small interfering RNAs (siRNAs) are important in maintaining DNA methylation[8], and N-DMPs coincided seven times more often than DMPs with sites to which 24-nt siRNAs mapped[18]. N-DMPs were also on average only half as far from such sites as DMPs ($P<2.2\times10^{-16}$) (Fig. 1e and Supplementary Fig. 6a). siRNAs are enriched in and around transposable elements[19]. In agreement, the average distance to the closest transposable element was much shorter for N-DMPs outside of transposable elements, compared to DMPs ($P<2.2\times10^{-16}$), even when only considering those in the centromere-distant regions of each chromosome, which contain relatively few transposable elements (Fig. 1e and Supplementary Fig. 6b, c).

A first major insight from our analyses is that transgenerational maintenance of CG methylation in transposable elements is apparently

[1] Department of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany. [2] Machine Learning and Computational Biology Research Group, Max Planck Institute for Developmental Biology and Max Planck Institute for Intelligent Systems, 72076 Tübingen, Germany.

*These authors contributed equally to this work.

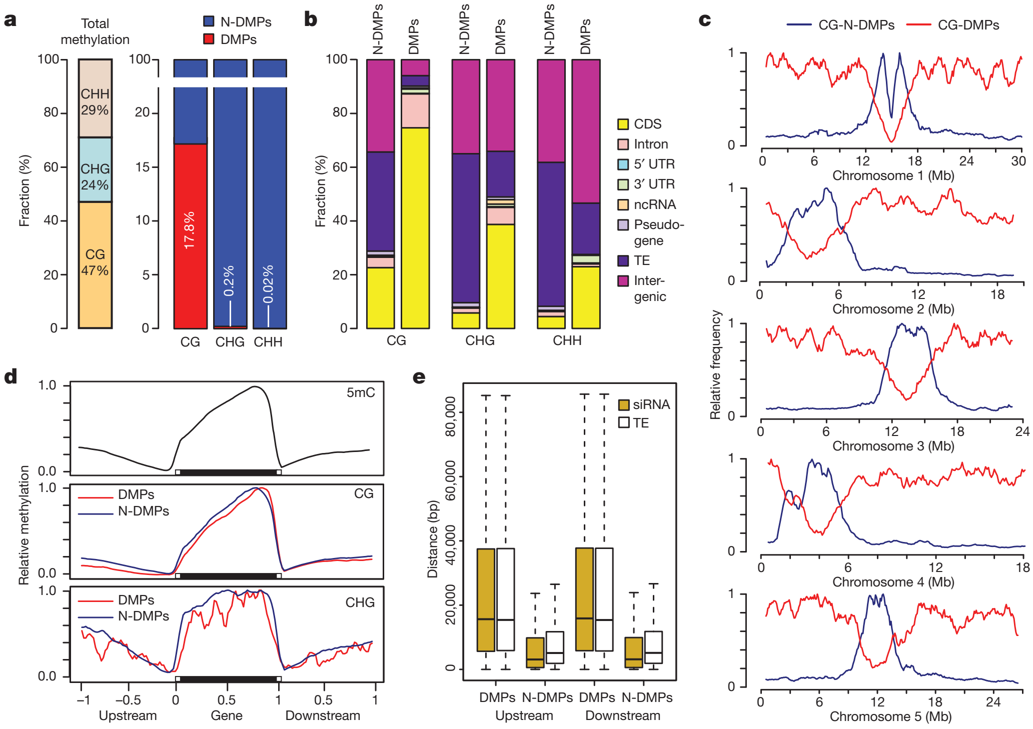

Figure 1 | Genome-wide distribution of methylation polymorphisms. **a**, Contribution of CG, CHH and CHG sites to total and differential cytosine methylation. 32.8% of all CG, 15.7% of CHG and 4.6% of CHH sites, adding up to 10.8% of all cytosines, showed evidence of methylation. **b**, Distribution of DMPs and N-DMPs according to local annotation. CDS, coding sequence; ncRNA, non-coding RNA; TE, transposable element. **c**, Distribution of CG-N-DMPs and CG-DMPs along each chromosome. Data were normalized to the highest value for each chromosome and class. **d**, Averaged distribution of all methylated sites (5mC) and methylated CG and CHG sites along genes. Data were normalized to the highest value for each sequence context and class. The coding region is indicated by a black bar. **e**, Distance of DMPs and N-DMPs to the closest upstream and downstream 24-nucleotide siRNA and transposable element. Horizontal bar corresponds to median, whiskers indicate entire 75th percentile.

much more stable than CG methylation of protein-coding genes, consistent with DNA methylation being more important for controlling the activity of transposable element compared to the latter[14,17,20,21]. This also agrees with a report that genic methylation is much more variable between wild strains of *A. thaliana* than methylation of transposable elements[3].

Hierarchical clustering based on DMPs grouped siblings as well as 3rd and 31st generation lines together. An arbitrary selection of methylated positions, which included about 6% DMPs, produced a similar pattern; however, with N-DMPs only, clusters became much more random (Fig. 2a). These observations indicate that our DMPs capture most of the methylation differences between lines. We next calculated the pairwise distance between strains based on DMPs (Fig. 2b). Correlation was highest between the two 3rd generation strains, and each individual of the 31st generation was more similar to these two lines, from which they were separated by 34 generations, than to the other lines from the 31st generation, from which they had diverged for 62 generations. Taken together, we conclude that whole-genome methylation patterns are largely stable and therefore heritable in *A. thaliana*, but that differences in methylation status accumulate gradually, similar to genetic mutations.

One strain, 69, was exceptional and had 40% more DMPs in comparison with the 3rd generation than the other 31st generation lines (Fig. 2b). To determine whether this strain might have a defect in the methylation machinery, we sequenced its genome with more and longer reads compared to our previous analysis[7]. We found a non-synonymous change in *MATERNAL EFFECT EMBRYO ARREST 57* (*MEE57*), which encodes a protein related to METHYLTRANSFERASE 1 (MET1) (Supplementary Fig. 7). *MEE57* has been reported as essential for endosperm development[22], although several *A. thaliana* strains lack functional *MEE57* copies[23]. Thus, whether the *MEE57* mutation contributes to the increased DMP number in line 69 remains unclear. The fact that the siblings of this line were as similar to each other as other sibling pairs (Fig. 2a) argues against a generally increased epimutation rate.

Compared to genetic mutations, the frequency of epimutations at single cytosine residues was many orders of magnitude higher, with an average of close to 30,000 DMPs in the analysed sequence space, compared with less than 30 DNA sequence mutations per strain[7]. Thirty-two per cent of DMPs between generations 3 and 31 occurred more than once, and 13% more than twice (Fig. 2c). If DMPs arose randomly, we would expect less than 1% of recurrent events. That we observe many more indicates that certain positions are particularly prone to increases or decreases in methylation rate. To investigate directly how many DMPs emerge from one generation to the next, we analysed the 32nd generation of lines 39 and 49. These individuals were progeny of siblings of the individuals interrogated in the 31st generation, and shared changes in the 32nd generation should reflect differences that arose between the 30th and 31st generation. We found on average over 3,300 between-generation DMPs. This is in the same range as DMPs between siblings (on average, about 5,000), but more than we would have expected from the 30,000 that had accumulated between the 3rd and each of the 31st generation lines. One explanation is that frequent transgenerational changes in methylation status occur at a limited number of sites, and that only a fraction of new DMPs is maintained over the longer term. This is corroborated by the observation that more than two-thirds of DMPs distinguishing the 32nd from the 31st generation in lines 39 and 49 had already been found in other 31st generation individuals.

DNA methylation is known to occur nonrandomly, and to cluster in specific segments of the genome[3,17,24]. We identified 249 differentially methylated regions (DMRs) that were at least 50 bp long (median 100 bp, maximum 650 bp) (Supplementary Table 2 and Supplementary Methods). Although probably a conservative estimate, the number of DMRs per line is in the same range as the DNA sequence mutations,

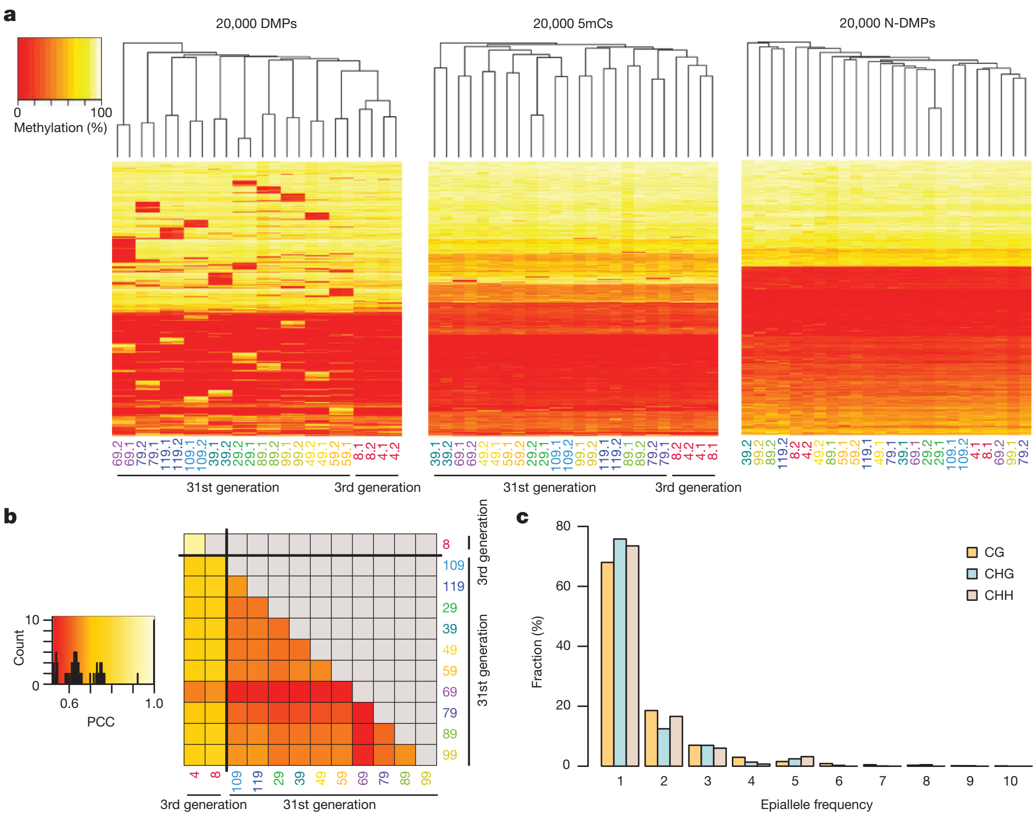

Figure 2 | Epigenetic diversity in the analysed population. **a**, Hierarchical clustering based on 20,000 sites each, drawn randomly from DMPs identified in pairwise comparison between all strains, cytosines methylated in at least one of the analysed strains (including about 6% DMPs), or N-DMPs. **b**, Heat map representing pairwise Pearson's correlation coefficient (PCC) between individuals, considering all 250,000 DMPs identified between all strains. PCCs between 3rd generation strains, 0.92; between 3rd and 31st generation, 0.63–0.77; between 31st generation lines, 0.52–0.66. The histogram on top of the colour key indicates counts of PCC bins. **c**, Epiallele frequency of DMPs in the 31st generation.

less than 30 per line[7]. As with CG-DMPs, DMRs preferentially localized to genes (Fig. 3a). DMRs did not overlap with known DNA mutations in these strains[7]. Similarly, structural variant discovery with established methods[23,25] did not reveal evidence for DMRs being due to gross DNA lesions. The frequency of DMRs along genes was similar to the overall distribution of methylated cytosines, and was reminiscent of the pattern of variation seen in wild strains of *A. thaliana*[3] (Supplementary Fig. 8). There were almost ten times as many DMRs in exons as in introns (Fig. 3a). Because exon-specific methylation may influence RNA splicing patterns[26,27], this could also be a source of variation in gene activity. Hierarchical clustering according to DMRs separated early- and late-generation strains into distinct groups (Fig. 3b). Notably, if we consider the methylation status in the 3rd generation individuals as largely reflecting the ancestral pattern, similar fractions of DMRs had lost or gained methylation by the 31st generation (Fig. 3c).

Similarly to DMPs, recurrent events constituted more than one-third of all DMRs, indicating that the affected genomic regions were privileged sites of change (Fig. 3d, e). In addition, comparison of generations 32 and 31 identified four short DMRs per line, with re-methylation of one segment that had become unmethylated in generation 31 (Supplementary Fig. 9). Together, these observations demonstrate that large changes in methylation, although rare, can occur even within a single generation.

Differences in promoter and genic DNA methylation can affect RNA levels[16]. We compared the transcriptomes of two randomly selected strains of the 31st generation with the 3rd generation strains by RNA-seq (Supplementary Fig. 10) and identified 320 differentially expressed genes in pairwise comparisons between strains (Fig. 3f and Supplementary Table 3). The two 31st generation lines were separated from each other by the most changes, and the two 3rd generation lines by the fewest. Seven differentially expressed genes overlapped with a DMR in these strains (Fig. 3g). For the three genes with the highest difference in expression level and overlapping with the most conspicuous DMRs, we observed a negative correlation between DNA methylation and gene expression. The remaining four genes overlapped with much shorter DMRs and no correlation was apparent (Fig. 3e, g and Supplementary Fig. 11).

We have presented a high-resolution analysis of transgenerational variation in DNA methylation of *A. thaliana*. The molecular mechanisms underlying these changes remain elusive, but siRNAs, which map very often in or near transposable elements[19], probably have a role in stabilizing DNA methylation, which is corroborated by our finding that DMPs tend to be farther from transposable elements and to be associated with lower local siRNA activity than N-DMPs. These observations indicate that the density and distribution of transposable elements, which can differ greatly even between closely related species[28], affect epigenetic variation throughout the genome. In the material analysed here, there was no evidence for DNA mutations acting *in cis* as an important cause of DMRs, although we cannot rule out that a non-synonymous mutation in a *MET1* homologue might contribute to increased variation in DNA methylation *in trans* in one of the lines.

In contrast to the high frequency of single-nucleotide methylation polymorphisms, larger regions appear to change methylation status at a rate that is comparable to genetic mutations. On the basis of previous work[7], it is conceivable that the emergence of DMRs requires specific structural features such as nearby repeats. Although DMRs are rare, we found evidence for DMRs affecting gene expression, indicating that natural, sequence-independent epialleles could potentially contribute to phenotypic diversity. There are subtle morphological differences between the mutation accumulation lines[6], and quantitative genetic approaches could be used to link specific DNA mutations or DMRs with such traits. How many of the methylation differences found

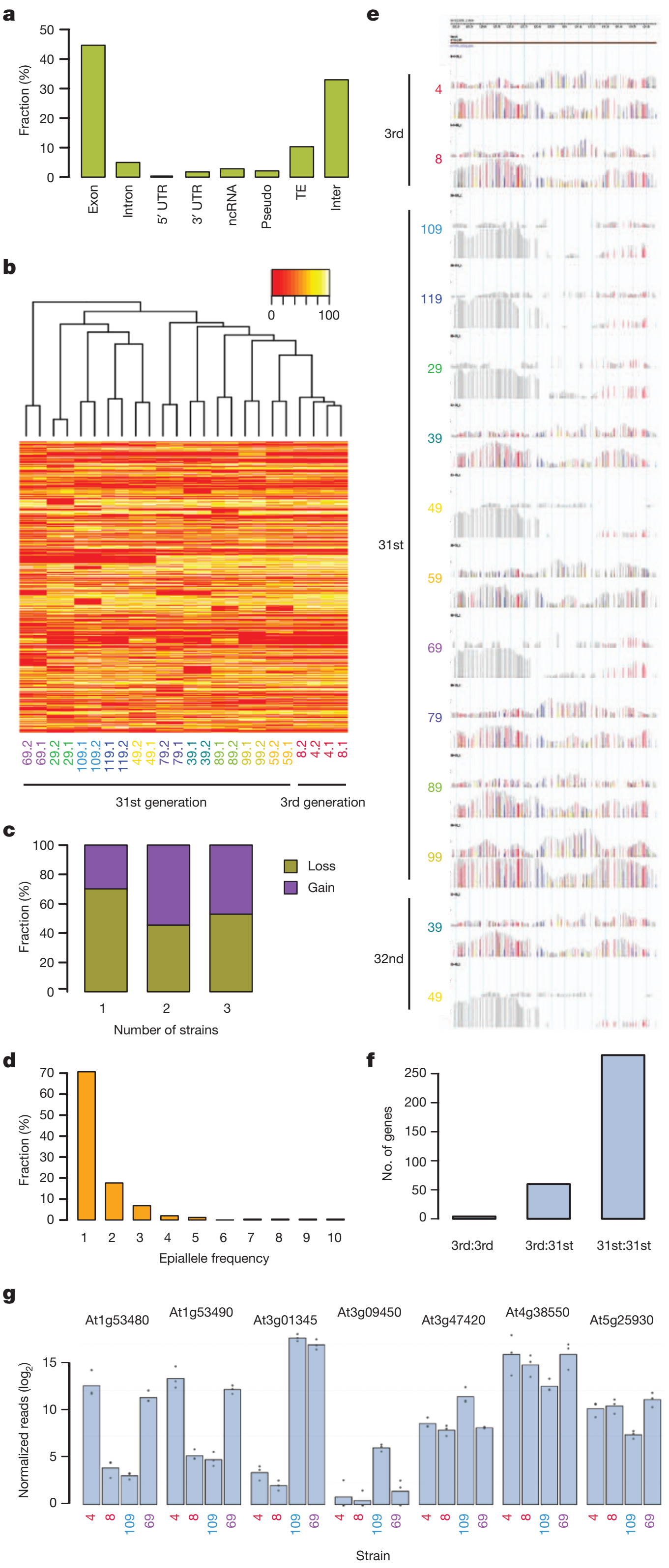

Figure 3 | Differentially methylated regions (DMRs). a, Distribution of DMRs according to local annotation. Inter, intergenic; pseudo, pseudogene. **b**, Hierarchical clustering of individuals from the 3rd and 31st generation based on methylated sites in DMRs, ranked according to their position in the genome. Note the shared methylation differences across lines, but strict pairing of siblings. **c**, Regions with losses and gains of methylation in 1, 2 or 3 strains in generation 31 compared to the 3rd generation strains. **d**, Epiallele frequency of DMRs. **e**, DMR at At3g01345 (Chr3:129,159–130,670) across all strains. Methylation on both strands is indicated for each strain. Colours indicate methylated reads (red, CG; blue, CHG; yellow, CHH). Grey indicates reads supporting non-methylation. For simplicity, only one sibling is shown per strain. **f**, Differentially expressed genes in comparisons between 3rd and 31st generation strains. **g**, Average RNA expression levels of the genes overlapping with the regions in **e** and in Supplementary Fig. 11. Dots indicate values of individual samples.

between wild strains[3] are due to sequence-independent changes versus ones driven by transposable elements and other structural variants is an important area for further investigation. In addition, it will be necessary to follow DNA methylation not only under benign greenhouse conditions, but also in the much more variable and stressful natural environment.

Perhaps our most important finding is that the number of epimutations does not increase linearly with time, indicating that many are not stably inherited over the long term. In addition to DMPs and DMRs that arose apparently independently in several strains, we even discovered a DMR that had become demethylated after 31 generations, but was re-methylated in the following generation. This suggests that DNA methylation in specific regions of the genome can fluctuate over relatively short timescales. Such sites can be considered as going through recurrent cycles of forward and reverse epimutation, which is very different from what is found at the level of the genome sequence, where reverse mutations are exceedingly rare. Importantly, reversion rates directly determine the ability of any type of allele to be subject to Darwinian selection. This needs to be taken into account when considering the potential of epialleles as a factor in evolution[5].

METHODS SUMMARY

Methylome sequencing. DNA was prepared from nuclei isolated from leaf tissue, bisulphite treated using a modification of a published protocol[14], and paired-end sequenced on the Illumina GAIIx platform. After image analysis and base calling with the Illumina pipeline, reads were processed using SHORE[7,29], and aligned to the Col-0 reference genome with GenomeMapper[30], adapted to the analysis of bisulphite sequencing data. Bisulphite conversion rates, as determined from unmethylated chloroplast and spiked-in lambda phage DNA, were 99.72% to 99.84%.

Analysis of methylated positions. Single sites were classified as methylated or unmethylated by fitting a binomial model based on reads falsely reporting methylation on the unmethylated plastid genome. We only considered cytosine residues covered by at least three independent high-quality base calls in all strains. For the determination of significant differences in methylation across strains on single sites or regions, we used Fisher's exact test.

Data visualization. A Gbrowse instance of the methylation profiles is available at http://gbrowse.weigelworld.org/fgb2/gbrowse/ath_methyl_ma.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 19 July; accepted 13 September 2011.
Published online 20 September 2011.

1. Richards, E. J. Population epigenetics. *Curr. Opin. Genet. Dev.* **18,** 221–226 (2008).
2. Paszkowski, J. & Grossniklaus, U. Selected aspects of transgenerational epigenetic inheritance and resetting in plants. *Curr. Opin. Plant Biol.* **14,** 195–203 (2011).
3. Vaughn, M. W. *et al.* Epigenetic natural variation in *Arabidopsis thaliana. PLoS Biol.* **5,** e174 (2007).
4. Rangwala, S. H. *et al.* Meiotically stable natural epialleles of Sadhu, a novel *Arabidopsis* retroposon. *PLoS Genet.* **2,** e36 (2006).
5. Slatkin, M. Epigenetic inheritance and the missing heritability problem. *Genetics* **182,** 845–850 (2009).
6. Shaw, R. G., Byers, D. L. & Darmo, E. Spontaneous mutational effects on reproductive traits of *Arabidopsis thaliana. Genetics* **155,** 369–378 (2000).
7. Ossowski, S. *et al.* The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana. Science* **327,** 92–94 (2010).
8. Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Rev. Genet.* **11,** 204–220 (2010).
9. Teixeira, F. K. *et al.* A role for RNAi in the selective correction of DNA methylation defects. *Science* **323,** 1600–1604 (2009).
10. Reinders, J. *et al.* Compromised stability of DNA methylation and transposon immobilization in mosaic *Arabidopsis* epigenomes. *Genes Dev.* **23,** 939–950 (2009).
11. Widman, N., Jacobsen, S. E. & Pellegrini, M. Determining the conservation of DNA methylation in *Arabidopsis. Epigenetics* **4,** 119–124 (2009).
12. Lynch, M. Evolution of the mutation rate. *Trends Genet.* **26,** 345–352 (2010).

13. Cokus, S. J. *et al.* Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452,** 215–219 (2008).
14. Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in *Arabidopsis. Cell* **133,** 523–536 (2008).
15. Zhang, Y. *et al.* QDMR: a quantitative method for identification of differentially methylated regions by entropy. *Nucleic Acids Res.* **39,** e58 (2011).
16. Zilberman, D., Gehring, M., Tran, R. K., Ballinger, T. & Henikoff, S. Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nature Genet.* **39,** 61–69 (2007).
17. Zhang, X. *et al.* Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis. Cell* **126,** 1189–1201 (2006).
18. Fahlgren, N. *et al.* MicroRNA gene evolution in *Arabidopsis lyrata* and *Arabidopsis thaliana. Plant Cell* **22,** 1074–1089 (2010).
19. Kasschau, K. D. *et al.* Genome-wide profiling and analysis of *Arabidopsis* siRNAs. *PLoS Biol.* **5,** e57 (2007).
20. Slotkin, R. K. *et al.* Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell* **136,** 461–472 (2009).
21. Lisch, D. Epigenetic regulation of transposable elements in plants. *Annu. Rev. Plant Biol.* **60,** 43–66 (2009).
22. Pagnussat, G. C. *et al.* Genetic and molecular identification of genes required for female gametophyte development and function in *Arabidopsis. Development* **132,** 603–614 (2005).
23. Cao, J. *et al.* Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nature Genet*. doi:10.1038/ng.911 (28 August 2011).
24. Tran, R. K. *et al.* DNA methylation profiling identifies CG methylation clusters in *Arabidopsis* genes. *Curr. Biol.* **15,** 154–159 (2005).
25. Mirouze, M. *et al.* Selective epigenetic control of retrotransposition in *Arabidopsis. Nature* **461,** 427–430 (2009).
26. Chodavarapu, R. K. *et al.* Relationship between nucleosome positioning and DNA methylation. *Nature* **466,** 388–392 (2010).
27. Laurent, L. *et al.* Dynamic changes in the human methylome during differentiation. *Genome Res.* **20,** 320–331 (2010).
28. Hollister, J. D. *et al.* Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata. Proc. Natl Acad. Sci. USA* **108,** 2322–2327 (2011).
29. Ossowski, S. *et al.* Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res.* **18,** 2024–2033 (2008).
30. Schneeberger, K. *et al.* Simultaneous alignment of short reads against multiple genomes. *Genome Biol.* **10,** R98 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to R. Shaw for providing seeds of the mutation accumulation lines. We thank C. Lanz for help with Illumina sequencing, J. Fitz for help with the Gbrowse implementation, K. Schneeberger for help with DNA mutation analysis, D. Zilberman for providing bisulphite sequencing protocols, and R. Clark, M. Pellegrini, R. Neher and J. Paszkowski for discussion. This work was supported by an HFSP Long-Term Fellowship (D.K.), the Volkswagen Foundation (O.S.), FP7 Collaborative Project AENEAS (contract KBBE-2009-226477), a Gottfried Wilhelm Leibniz Award of the DFG, and the Max Planck Society (D.W.).

Author Contributions C.B., J.H. and D.W. conceived the study; C.B. performed the experiments; C.B., J.H., J.M., D.K. and O.S. analysed the data; K.B. provided advice on statistical analysis; and C.B. and D.W. wrote the paper with contributions from all authors.

Author Information DNA sequencing data are deposited in the European Nucleotide Archive (http://www.ebi.ac.uk/ena/data/view/ERP000902) under accession number ERP000902. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to D.W. (weigel@weigelworld.org).

METHODS

Plant growth and material. Seeds were derived from *Arabidopsis thaliana* Columbia-0 lines in generation 3 (lines 4 and 8), generation 31 (lines 29, 39, 49, 59, 69, 79, 89, 99, 109 and 119) and generation 32 (lines 39 and 49), counting from the founders, as described by Shaw and colleagues[6] (Supplementary Fig. 1). Plants were grown on soil under long-day conditions (23 °C, 16 h light, 8 h dark) after seeds had been stratified in 150 nM GA-supplemented water at 4 °C for 6 days. Siblings were grown independently at different time points. Positions of the pots were randomized.

Nucleic acid extraction. DNA was extracted from rosettes of individual 21-day-old plants. Plant material was flash-frozen in liquid nitrogen and ground in a mortar. The ground tissue was re-suspended in nuclei extraction buffer (10 mM Tris-HCl pH 9.5, 10 mM EDTA, 100 mM KCl, 0.5 M sucrose, 0.1 mM spermine, 0.4 mM spermidine, 0.1% β-mercaptoethanol). After cell lysis in nuclei extraction buffer with 10% Triton X-100, nuclei were pelleted by centrifugation at 2,000*g* for 120 s. Genomic DNA was extracted using the Qiagen Plant DNeasy kit (Qiagen). Total RNA was extracted from rosette leaves of individual plants using the Trizol (Invitrogen) method according to the manufacturer's instructions. Residual DNA was eliminated by DNase I (Thermo Fisher Scientific) treatment.

Library preparation. Preparation of DNA libraries for genomic sequencing was done using the NEBNext DNA Sample Prep Reagent Set 1 (New England Biolabs), following the Illumina Genomic Sample Prep Guide (Illumina). 500–1,000 ng genomic DNA was fragmented to 300 bp average size with a Covaris S2 instrument using the following settings for 120 s in frequency sweeping mode: intensity 5, duty cycle 10%, 200 cycles per burst. DNA was purified on Qiaquick PCR purification columns. Preparation of DNA libraries for bisulphite sequencing was adapted from ref. 14. Input DNA was fragmented as described above. Libraries were constructed using the NEBNext DNA Sample Prep Reagent Set 1 (New England Biolabs) according to the Illumina Genomic Sample Prep Guide with the following modifications. We used the Illumina Early Access Methylation Adapter Oligo Mix (catalogue number ME-100-0010). After size selection, the non-methylated cytosine residues were converted to uracil using the EpiTect Plus DNA Bisulfite kit (Qiagen) according to the manufacturer's guidelines. For higher conversion efficiency the bisulphite incubation was repeated. Library enrichment was performed with Pfu Cx HotStart Polymerase (Agilent) and 18 PCR cycles. Libraries for RNA sequencing were prepared from 4 μg of total RNA using the Illumina Truseq RNA sample prep kit B according to the manufacturer's protocol.

Sequencing. All sequencing was performed on an Illumina GAIIx instrument. Genomic and bisulphite-converted libraries were sequenced with 2 × 101-bp paired-end reads. For bisulphite sequencing, conventional *A. thaliana* DNA genomic libraries were analysed in control lanes. Transcriptome libraries were sequenced with 101-bp single end reads, with three libraries with different indexing adapters pooled in one lane; no control lane was used. For image analysis and base calling, we used the Illumina OLB software version 1.8.

Processing and alignment of bisulphite-treated reads. The SHORE pipeline[29] was used to trim and quality-filter the reads. Its default parameters were applied for the filtering step: reads with more than 2 (or 5) bases in the first 12 (or 25) positions with a quality score less than 3 were discarded. Reads were trimmed to the rightmost occurrence of two adjacent bases with quality values equal to or greater than 5. Trimmed reads shorter than 50 bases were discarded. The remaining high quality sequences (on average 82% of raw reads across the sequenced strains) were aligned against the *Arabidopsis thaliana* genome sequence version TAIR9 (http://www.arabidopsis.org/portals/genAnnotation/gene_structural_annotation/annotation_data.jsp) using a modified version of the mapping tool GenomeMapper[30] that supports the alignment of bisulphite converted reads. Bisulphite converts non-methylated cytosines into uracils, which are propagated as adenine-thymine base pairs after PCR amplification. GenomeMapper tolerates asymmetrical T-to-C or A-to-G mismatches (read base against reference base) and can distinguish between reads from the bisulphite-converted strand of a DNA fragment and sequences from its complementary amplified strand, if the reads have been obtained by paired-end sequencing. Only the read from the strand with converted Cs is informative about the methylation status of the underlying cytosine site. We allowed for up to 10% single-base-pair substitutions along the read length in the alignment process for each read to retain most coverage. GenomeMapper reports all alignments with the least amount of mismatches for each read. However, only reads mapping uniquely to a single position were used for this study. Furthermore, all but one read were removed from further analysis if their 5′ ends aligned to the same genomic position, to account for amplification biases. A paired-end correction method[23] was used to discard repetitive reads by comparing the distance between reads and their partner to the average distance between all read pairs. Reads with abnormal distances were removed if there was at least one other alignment of this read in a concordant distance to its partner. Finally, read counts on all cytosine sites were obtained with SHORE. The 'scoring matrix approach' of SHORE[23] assigns a score to each site by testing against different sequence and alignment related features. The criteria and complete scoring matrix can be found in Supplementary Table 4. For comparisons across lines, cytosines were accepted if at most one intermediate penalty on its score was applicable to at least one strain (score ≥32). In this case, the threshold for the other strains was lowered, accepting at most one high penalty (score ≥ 15). In this way, information from other strains is used to assess sites from the focal strain under the assumption of mostly conserved methylation patterns, allowing the analysis of additional sites. The methylation statistics on each single strain assumed a quality score of 25 or higher, which means no more than two intermediate penalties.

Determination of methylated sites. Sequencing errors, noise and imperfections of the bisulphite conversion contribute to the occurrence of sites that appear to be weakly methylated. Reads mapping against the non-methylated chloroplast sequence allow for objective estimation of the effective background rate of false-positive methylation detection. For this purpose, we fitted an independent binomial model to the relative proportions of converted and unconverted reads that cover cytosines in the chloroplasts. We estimated the binomial rate of false-positive methylation from the maximum likelihood estimate, separately for each library and for different bins of total read coverage:

$$\hat{r} = \arg\max_r \prod_s \text{Binomial}(\overline{n}_s, (n_s + \overline{n}_s)|r).$$

Here, n_s and $\overline{n}_s$ denote the number of converted and unconverted reads from the considered cytosine sites. Supplementary Fig. 12a shows the obtained background methylation rate for a single strain, line 30-39, as a function of the total read coverage per site. The overall false methylation rate when combining read data across the range of read coverage was 0.22%, which deviates significantly from higher error estimates when considering low-coverage regions in isolation. To account for the variability in error rates in the downstream analyses, we used specific error models for each strain and for read-coverage bins of multiples of fivefold, yielding error rates between 0.2% and 5.0% (Supplementary Fig. 12b). For coverage bins with too few sites for robust statistical estimation (<50), we imputed the false methylation rate from the closest sufficiently populated coverage bin. Given the estimated rates for false methylation, we carried out a genome-wide test for significant methylation of cytosines. For each site, we calculated the *P* value under the background model. We then used Storey's method[31], an extension of the Benjamini–Hochberg stepdown procedure, to assess genome-wide significance using *q* values. At a joint false discovery rate (FDR) of 5% we found between 2,316,966 and 3,458,949 methylated sites in each strain (Supplementary Table 1). When reducing FDR to 0.1%, we still retained almost 85% of the methylated sites, showing that the number of sites with weak methylation evidence was low. For analysis of methylated sites reported in this study, an FDR of 5% was deemed to be acceptable.

Identification of differentially methylated positions. From the 13.9 million cytosines for which we had at least three independent high-quality reads in each strain, we selected sites that showed significant methylation in at least one strain, resulting in 3,067,017 positions. Sites with statistically significant methylation differences were identified with Fisher's exact test. *P* values from individual tests per site were combined into single *P* values via conservative Bonferroni correction. Genome-wide FDRs were then estimated using Storey's method[31]. To limit false-positive DMPs, we first identified 69,583 DMPs between siblings at a relaxed FDR of 10%. These sites were excluded as were 8,893 DMPs distinguishing the two 3rd generation strains. This left 2,988,541 positions as the final set to test for differential methylation between generations. Twenty pairwise tests of each of the ten 31st generation strains against both 3rd generation strains were conducted on sites consistently methylated between 31st generation siblings and in the 3rd generation. At an FDR of 5%, this yielded 186,248 DMPs. DMP allele frequency was obtained by progressively removing the strain with the lowest *q* value and correcting the remaining *P* values for multiple testing by the methods described above.

We applied the same strategy to identify DMPs that differed either between the 31st and 3rd generation, or between 31st generation strains. Count data from replicates were combined for each site, followed by pairwise Fisher's exact tests between all combinations of strains (66 tests). We estimated *P* values for at least one differential pair using a Bonferroni correction, followed by Storey's method[31] to assess genome-wide significance. At a joint FDR of 5%, this identified 253,546 DMPs across all 12 strains.

Assessing statistical power. Two main factors influence the power to detect methylation differences: the number of statistical tests and local read coverage. To assess the impact of multiple testing, we applied the approach described above to all sites with at least threefold coverage in at least 12 of 24 individuals examined. Of 25.3 million such positions, 4,547,568 were found to be methylated in at least one of the lines, compared to 3,067,017 out of 13.9 million positions when considering only sites with complete information. The number of sites assessed as methylated thus increased roughly linearly with the number of tested sites, as did the number of differentially methylated positions. Similarly, the fraction of DMPs shared in more than one 31st generation strain, ~31%, was very similar to the ~32% found among sites with complete information. We conclude that our method is largely insensitive to the number of tests performed.

To assess the effect of read coverage, we determined how many DMPs could be identified after subsampling at 25%, 50% and 75% of total coverage. We identified almost twice as many DMPs with 50% compared to 25% coverage, but only 13% additional DMPs were identified when increasing coverage from 75% to 100% (Supplementary Fig. 13). Although not yet asymptotic, we estimate that the false-negative rate is well below 50%, and most likely closer to 10%.

Identification of differentially methylated regions. The ~186,000 DMPs distinguishing 31st from both 3rd generation lines were consolidated into regions of adjacent DMPs for each strain, with a maximum distance of 50 bp between DMPs. We then used Fisher's exact test on the sum of methylated and unmethylated reads in both siblings, averaged across positions within the region. Resulting *P* values were corrected with Storey's method[31] and an FDR of 5% was accepted. Statistically significant regions from different strains were merged if they overlapped by at least 20% of their combined length and if the methylation change was in the same direction compared to the 3rd generation lines. Short regions containing only a small number of strongly differential sites were excluded by requiring DMRs to have a minimum length of 50 bp and to contain at least ten methylated positions and at least five DMPs.

Mapping of DMPs, N-DMPs and DMRs to genomic elements. We used the TAIR10 annotation (http://www.arabidopsis.org/portals/genAnnotation/gene_structural_annotation/annotation_data.jsp) to determine overlap of genes, pseudogenes and transposable elements with methylated positions. We defined intergenic regions as regions that did not correspond to any annotated feature. A DMR was considered mapping to a particular genomic element if it overlapped with such an element for more than 20% of its length.

Assessing the distance to the closest siRNA or transposable element. We determined the distance between a methylated position and the closest upstream and downstream siRNA using a published data set for *A. thaliana* aerial tissue[18] (NCBI GEO accession number GSM518432). For transposable elements, we used the TAIR10 annotation. Statistical significance was tested with a two-sided, unpaired Student's *t*-test on the measured distances. Pericentromeric regions were defined as described[23].

Analysis of gene expression. DNA sequences resulting from RNA library preparation were barcode-sorted and quality filtered in SHORE[29], and aligned using BWA[32] to the TAIR10 gene annotation. Reads were filtered for duplicates and required to have a mapping quality of at least 37. The remaining mappings were used to generate gene-level counts for expression analysis. We only considered genes for which the total counts in all samples combined exceeded 30. Between-sample expression correlations and strain-distribution plots were used for quality control to identify poor samples, and pairwise comparisons of expression were performed using the DEseq package[33] implemented in R. Differentially expressed genes were identified by a combination of per-gene variance ($P \leq 0.01$, with Benjamini and Yekutieli correction[34]) and common variance (≥2× change). Density of genes without expression support were plotted along chromosomes in sliding windows considering only genes with at least 50 methylated or unmethylated calls. Windows of gene methylation were calculated for entire gene bodies as the fraction of methylated positions (methylated in any sample) divided by the total number of called positions. Very similar results were obtained considering sites methylated in all samples. Data were visualized with the ggplots package in R[35].

Data visualization. A Gbrowse instance of the methylation profiles is available at http://gbrowse.weigelworld.org/fgb2/gbrowse/ath_methyl_ma.

31. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100,** 9440–9445 (2003).
32. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
33. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).
34. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29,** 1165–1188 (2001).
35. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2009).