Howard Hughes Medical Institute

Published as: *Cell*. 2016 July 14; 166(2): 492–505.

Epigenomic diversity in a global collection of *Arabidopsis thaliana* accessions

Taiji Kawakatsu[1,2,4,*], **Shao-shan Carol Huang**[1,2,*], **Florian Jupe**[1,2,*], **Eriko Sasaki**[6,*], **Robert J. Schmitz**[2,5], **Mark A. Urich**[2], **Rosa Castanon**[2], **Joseph R. Nery**[2], **Cesar Barragan**[2], **Yupeng He**[2], **Huaming Chen**[2], **Manu Dubin**[6], **Cheng-Ruei Lee**[6], **Congmao Wang**[7,8], **Felix Bemm**[7], **Claude Becker**[7], **Ryan O'Neil**[2], **Ronan C. O'Malley**[2], **Danjuma X. Quarless**[9], **The 1001 Genomes Consortium**, **Nicholas J. Schork**[9], **Detlef Weigel**[7], **Magnus Nordborg**[6], and **Joseph R. Ecker**[1,2,3]

[1] Plant Biology Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA

[2] Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA

[3] Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA

[4] Genetically Modified Organism Research Center, National Institute of Agrobiological Sciences, Tsukuba, Ibaraki 305-8602, Japan

[5] Department of Genetics, University of Georgia, Athens, GA 30602, USA

[6] Gregor Mendel Institute, Austrian Academy of Sciences, Vienna Biocenter, 1030 Vienna, Austria

[7] Department of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany

[8] Institute of Digital Agriculture, Zhejiang Academy of Agricultural Sciences, Hangzhou, Zhejiang, 310021, PR China

[9] Human Biology, J. Craig Venter Institute, La Jolla, CA 92037, USA

Corresponding author: Joseph R. Ecker (ecker@salk.edu).
[*]Co-first authors

The 1001 Genomes Consortium Participants
Carlos Alonso-Blanco, Jorge Andrade, Claude Becker, Felix Bemm, Joy Bergelson, Karsten Borgwardt, Eunyoung Chae, Todd Dezwaan, Wei Ding, Joseph R. Ecker, Moisés Expósito-Alonso, Ashley Farlow, Joffrey Fitz, Xiangchao Gan, Dominik G. Grimm, Angela Hancock, Stefan R. Henz, Svante Holm, Matthew Horton, Mike Jarsulic, Randall A. Kerstetter, Arthur Korte, Pamela Korte, Christa Lanz, Chen-Ruei Lee, Dazhe Meng, Todd P. Michael, Richard Mott, Ni Wayan Muliyati, Thomas Nägele, Matthias Nagler, Viktoria Nizhynska, Magnus Nordborg, Polina Novikova, F. Xavier Picó, Alexander Platzer, Fernando A. Rabanal, Alex Rodriguez, Beth A. Rowan, Patrice A. Salomé, Karl Schmid, Robert J. Schmitz, ümit Seren, Felice Gianluca Sperone, Mitchell Sudkamp, Hannes Svardal, Matt M. Tanzer, Donald Todd, Samuel L. Volchenboum, Congmao Wang, George Wang, Xi Wang, Wolfram Weckwerth, Detlef Weigel, Xuefeng Zhou

ACCESSION NUMBERS
All sequencing data are available in NCBI GEO with accession numbers GSE43857 for MethylC-seq, GSE43858 and GSE80744 for RNA-seq. Optical genome maps are available at http://signal.salk.edu/opticalmaps/.

AUTHOR CONTRIBUTIONS
Conceptualization methylome and transcriptome, R.J.S and J.R.E.; Conceptualization optical mapping, F.J. and J.R.E.; Methodology, Y.H., R.O'Neil, D.X.Q. and N.J.S.; Formal Analysis, T.K., S.C.H., F.J., H.C. and E.S.; Investigation, T.K., S.C.H., F.J., R.J.S., J.R.N., M.A.U., C. Barragan, and R.C.; Writing – Original Draft, T.K., S.C.H, F.J. and E.S; Writing – Review & Editing, T.K., S.C.H., F.J., E.S., R.J.S., M.N., D.W. and J.R.E.; Visualization, H.C.; Resources, M.D., C.L., C.W., F.B., R.O'Malley, C. Becker, The 1001 Genomes Consortium, M.N. and D.W.; Supervision, J.R.E.

SUMMARY

The epigenome orchestrates genome accessibility, functionality and three-dimensional structure. Because epigenetic variation can impact transcription and thus phenotypes, it may contribute to adaptation. Here we report 1,107 high-quality single-base resolution methylomes and 1,203 transcriptomes from the 1001 Genomes collection of *Arabidopsis thaliana*. Although the genetic basis of methylation variation is highly complex, geographic origin is a major predictor of genome-wide DNA methylation levels and of altered gene expression caused by epialleles. Comparison to cistrome and epicistrome datasets identifies associations between transcription factor binding sites, methylation, nucleotide variation and co-expression modules. Physical maps for nine of the most diverse genomes reveals how transposons and other structural variants shape the epigenome, with dramatic effects on immunity genes. The 1001 Epigenomes Project provides a comprehensive resource for understanding how variation in DNA methylation contributes to molecular and non-molecular phenotypes in natural populations of the most studied model plant.

INTRODUCTION

Cytosine methylation and histone modification are epigenomic marks with effects on activity of transposable elements (TEs; all abbreviations are listed in Table S1), transcription of genes and formation of heterochromatin. In plants, DNA methylation occurs in the symmetric contexts CG and CHG (H = C, A or T), and the asymmetric context CHH (Law and Jacobsen, 2010). CG methylation is propagated through a simple copy mechanism during DNA replication, whereas CHG and CHH methylation are maintained by self-reinforcing loops (Kawashima and Berger, 2014). Although changes in DNA methylation may arise spontaneously (Becker et al., 2011; Schmitz et al., 2011), genetic and environmental factors are almost certainly more important. The genetic basis of DNA methylation variation includes structural variations such as TE insertions/deletions (indels), chromosome rearrangements, and mutations in methylation factors (Pecinka et al., 2013), whereas important environmental conditions include temperature and other stresses (Dowen et al., 2012; Dubin et al., 2015; Secco et al., 2015).

It has been proposed that, as sessile organisms that can persist in the same location for a long time, plants may be particularly likely to exploit DNA methylation for rapid adaptation to changing environments. DNA methylation can affect gene expression, cause visible phenotypes (Pecinka et al., 2013; Schmitz and Ecker, 2012) and measurable variation in adaptive traits (Cortijo et al., 2014; Johannes et al., 2009; Kooke et al., 2015). Therefore, cataloging variation in DNA methylation, transcriptomes as well as genome structural variation in natural populations is a prerequisite for understanding the role of natural epigenetic variations in adaptation to local environments.

We have previously described base-resolution DNA methylomes of two medium-sized sets of *Arabidopsis thaliana* accessions, a global set of 144 accessions and a focused regional set of 150 Swedish accessions (Schmitz et al., 2013; Dubin et al., 2015). These and related studies (Hagmann et al., 2015; Pignatta et al., 2014; Shen et al., 2014; Vaughn et al., 2007) have provided initial evidence for the interplay of genetic and epigenetic variation in shaping molecular and non-molecular phenotypes. Leveraging the expanded analysis of sequence

variations in the genomes of 1,135 natural accessions (The 1001 Genomes Consortium, 2016), here we describe results from the accompanying 1001 Epigenomes Project, with 1,107 methylomes from 1,028 accessions and 1,203 transcriptomes from 998 accessions. Additionally, we analyzed optical genome maps from nine accessions to infer how structural variations in the genome shape the methylome and transcriptome. The full representation of epigenomic diversity in *A. thaliana* will accelerate studies in this model plant to provide insight into general principles of adaptive variation.

RESULTS AND DISCUSSION

The Dataset

The 1001 Epigenomes Project reports on 1,227 worldwide *A. thaliana* accessions selected based on their genetic and geographic diversity. We generated high-quality base-resolution methylomes for 1,028 and transcriptomes for 998 accessions (Fig. 1A). Of these, 866 accessions have both methylomes and transcriptomes from rosette leaves, as well as SNP and small indel data from the 1001 Genomes Project (The 1001 Genomes Consortium, 2016) (Fig. 1A). The methylomes for 745 accessions have not been reported before (Fig. 1B). Overall, the 1001 Epigenomes Project provides 1,107 methylomes and 1,203 transcriptomes (Fig. 1C and D).

The Methylomes

MethylC-seq bisulfite sequencing reads were mapped against individual pseudo-reference genomes generated for each accession by substituting SNPs and short deletions (up to 40 bp) in the Col-0 reference genome sequence (TAIR10). On average, 88% of each genome was covered by unique reads, with 8.4x strand-specific coverage (Fig. S1A and Table S2).

Over a third of all cytosines (14,799,349) were methylated in at least one accession (Fig. S1B). On average, the genome-wide weighted methylation level was 5.8% (Fig. S1E and Table S2). Seventy-eight percent (11,554,831) of methylated cytosines (mC) were differentially methylated across accessions (dmCs; Fig. S1C and D). Among dmCs epi-genotyped in at least 110 methylomes (10% of analyzed methylomes), singleton epi-alleles (where only one accession was methylated or unmethylated) accounted for 5.4% dmCs in CG context, 6.7% in CHG context and 17.0% in CHH context. In terms of chromosomal distribution, mC and dmC in all contexts were enriched in the pericentromere while mCG and dmCG have higher frequencies along chromosome arms, as expected for CG gene body methylation (gbM) (Schmitz et al., 2013).

We collapsed dmCs within 200 bp blocks and identified 22,060 differentially methylated regions (DMRs) that covered 45 Mb (38%) of the reference genome. We classified them into mutually exclusive categories: CG-DMRs (differentially methylated only in the CG context), CH-DMRs (in CHG and/or CHH context), and C-DMRs (in CG and CHG and/or CHH context) (Fig. S1G-J and Table S3). CG-DMRs generally overlapped with genes, reflecting variable CG gbM (Fig. S1K and L). About half of CH-DMRs overlapped with TEs and 35% did not overlap with any annotated regions (Fig. S1K and L). C-DMRs overlapped with genes and TEs (Fig. S1K and L). DMR distribution reflects the general chromosomal

distribution of the overlapping genomic features (Fig. S1F). Gene Ontology (GO) enrichment analysis revealed that genes for housekeeping processes, such as protein localization/transport related genes and metabolism, were enriched in CG-DMRs (Fig. S1M), whereas CH-DMRs showed no enriched GO terms. In line with previous results (Schmitz et al., 2013), genes that had particularly variable expression levels across tissues or environments in the reference accession, including disease resistance genes, were enriched in C-DMRs (Fig. S1N), suggesting that C-DMRs might be linked to environmental adaptation by regulating responsive gene expression.

Gene body methylation does not have a major role in shaping transcriptome variation

We examined gbM variation in our dataset, defined as CG-only methylation within gene bodies with a depletion of methylation at transcription start sites (TSS) and transcription termination sites (TTS). The numbers of genes with gbM were highly variable between accessions, and positively correlated with the average mCG levels of these genes (Fig. 2A and B; Pearson's $r = 0.62$, $p < 2e\text{-}16$). In relation to geographical origins, hypermethylated accessions were generally found in Sweden (Fisher exact test p = 4.0e-9), whereas hypomethylated accessions were found mainly in Spain (Fisher exact test p = 1.4e-3) (Fig. 2C).

gbM is associated with constitutive gene expression (Tran et al., 2005; Zhang et al., 2006; Zilberman et al., 2007), and the expression levels of gbM genes were indeed higher than those of unmethylated (UM) and TE-like methylated (teM; mCHG or mCHH and/or mCG) genes across all tested accessions (Fig. 2D; Wilcoxon rank sum test $p < 2.2e\text{-}16$ and $p < 2.2e\text{-}16$, respectively). To examine genome-wide relationship between gbM levels and transcription, we compared pairwise correlations for mCG within gene bodies and those for gene transcript levels (Fig. 2E). Transcriptomes among accessions were more similar to each other than mCG levels (Wilcoxon rank sum test $p < 2.2e\text{-}16$). Notably, although the hypomethylated accessions Cvi-0 and UKID116 exhibited greatly reduced gbM mCG levels, global gene expression levels were similar to the moderately methylated Col-0 and the hypermethylated Bak-5 (Fig. 2E and 2F). These results suggest that although gbM is correlated with constitutive gene expression in the Col-0 reference, it is largely dispensable under laboratory growth conditions, which is consistent with recent observations of a complete loss of gbM in some angiosperms (Bewick et al, 2016). Indeed it has been argued that gbM is either a direct or indirect consequence of transcription rather than a cause (Teixeira and Colot, 2009; Inagaki and Kakutani, 2012).

Establishment and reversal of TE-like methylation of genes

Our DMR analyses revealed that certain genes were poly-epiallelic (PE) with some accessions being unmethylated, some exhibiting gbM, and some teM. Examining the 846 accessions grown at Salk, we found 21,939 genes that had gbM in at least one accession, 8,889 genes that had teM in at least one accession, and 7,524 genes that were part of both sets (PE) (Fig. 2G and H). In general, teM epialleles were less frequent than gbM epialleles (Fig. 2I), which were typically shared by about 90% of the accessions, suggesting that the teM alleles are younger than the gbM alleles. Interestingly, teM of 2,053 PE genes (27%) was found in single accessions (teM singletons). So-called relict accessions (The 1001

Genomes Consortium, 2016), which occur at low frequency around the Mediterranean and are the product of ice age refugia, generally contained more teM singletons (Fig. 2J; Wilcoxon rank sum test p = 2.1e-7).

Next, we examined the functional relevance of gbM versus teM. Compared to non-PE genes, PE genes had more non-synonymous mutations (Fig. 2K; Wilcoxon rank sum test p = 4.0e-236), and were less likely to be duplicated (13% vs. 18%; Fisher exact test p=5.1e-31), but were more often members of multi-gene families (54% versus 45%; Fisher exact test: p=2.1e-36). GO analysis of PE genes identified enrichment for phosphorylation-related and, similar to C-DMRs, immune response-related terms (Fig. 2L), suggesting that PE genes are generally involved in signaling and metabolic processes.

Among the 1,934 genes that have gbM and teM epialleles in at least five accessions, 199 teM genes have significantly lower expression (FDR < 0.05) than their gbM epialleles. Notably, the teM epialleles of the temperature-dependent flowering repressor *MADS AFFECTING FLOWERING 3* (*MAF3*) (Ratcliffe et al., 2003) was associated with lower expression (Fig. 2M). Although we did not detect a significant association between flowering time at 10°C or 16°C (The 1001 Genomes Consortium, 2016) and the epialleles (Wilcoxon rank sum test p > 0.01), it is possible that teM associated reduction in *MAF3* expression is involved in flowering variation under natural conditions.

One possible explanation for the emergence of poly-epialleles is the spreading of RNA directed DNA methylation (RdDM) from nearby TEs. Consistent with this, TE annotations were enriched within 500 bp or inside PE genes that were teM in Col-0, compared to all protein-coding genes (Fisher exact test p = 0.015). The remaining 367 PE genes showed enrichment of mCHH in gene bodies (Fig. 2N). Other known potential triggers of teM include inverted repeats and RdDM triggered by unlinked loci, but it is also possible that aberrant mRNAs or gene-silencing associated RNAs are produced from gbM genes and processed into siRNAs, with the potential to promote non-canonical RdDM within these genes and their paralogs (Nuthikattu et al., 2013; Pecinka et al., 2013).

Multiple pathways contribute to methylation variation

We next examined overall methylation levels across the 1001 Epigenomes population, focusing in particular on the correlation between methylation in different contexts, and on the correlation with climate and geography. mCHH in TEs is separately catalyzed by two distinct DNA methyltransferases, DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2) in the RdDM pathway and CHROMOMETHYLASE 2 (CMT2), which recognizes heterochromatic regions via H3K9 di-methylation (Stroud et al., 2014; Zemach et al., 2013). To distinguish these pathways, we considered TEs affected in *drm1 drm2* and *cmt2* mutants, respectively (Stroud et al., 2013). Methylation was correlated between these two contexts, and comparison with data from plants grown at lower temperatures confirmed the positive correlation between temperature and mCHH (Fig. 3A; Dubin et al, 2015). For leaf samples from Salk-grown accessions, hypermethylated accessions were mainly found in Germany (Fisher exact test p = 2.0e-7), whereas hypomethylated accessions were almost randomly distributed (Fig. B). In summary, mCHH variation within TEs is likely due to differences in a combination of pathways, guided by environmental and developmental cues.

The pattern of correlation, across individuals, between methylation and environmental variables (Fig. 3C) revealed clear similarities between all types of mCHH, whether RdDM- or CMT2-targeted, and the same for mCHG. However, mCG behaved very differently in that mCG of TEs was correlated with mCHH, while mCG gbM was correlated with mCHG. This finding strongly suggests that not all mCG is created equally. It also supports the notion that gbM is connected to *CMT3* and mCHG (Miura et al, 2009; Bewick et al., 2016). mCG also stood out in terms of its genetic architecture (Fig. 3D). Viewed as a phenotype, the variation of mCG across lines was well explained by genome-wide SNP variation. It is thus heritable in the statistical sense, which is entirely consistent with it being heritable in the direct sense of being stably transmitted across generations through meiosis.

All types of methylation showed striking correlations with place of origin and its climate (Fig. 3C). Very broadly, methylation levels within TEs were positively correlated with latitude and precipitation, and negatively correlated with warmer temperatures. The correlation between TE methylation and temperature of origin is thus the opposite of the correlation between TE methylation and experimental growth temperature (Fig. 3A and C), suggesting that temperature compensation has evolved in the natural range (Shen et al., 2014; Dubin et al., 2015). gbM again behaved very differently, and showed strong correlation with colder winters (Dubin et al. 2015).

Genome-wide association reveals the genetic basis of methylation variation

To gain further insight into the genetics of methylation, we turned to genome-wide association studies (GWAS), using the dense SNP data available for our sample with minor allele frequency (MAF) 5% cutoff. For TE methylation, several peaks with genome-wide significance were detected (Fig. 4A-B), and there was massive enrichment for *a priori* candidates (Fig. 4D-E). Among the latter, our analysis confirmed the previously reported strong effect of *CMT2* itself on CMT2-dependent mCHH (Dubin et al., 2015), but the top SNP here is considerably closer to the gene (chr4:10,422,486, 1.2 kb downstream of *CMT2*, −log10 p = 7.88). Another striking candidate was *ARGONAUTE 9* (*AGO9*), which is involved in siRNA silencing, and for which natural variants have been connected to differences in epigenetic control of cell specification (Rodriguez-Leal et al., 2015). Interestingly, AGO9 appears to be associated both with RdDM- and CMT2-dependent mCHH, although the SNPs associated differ, suggesting that different alleles are involved. For RdDM-targeted mCHH, the top SNP was over 200 kb away from the coding region (chr5:7,344,821, −log10 p = 6.78), whereas for CMT2-dependent mCHH the top SNP, was located 16 kb downstream of *AGO9* (chr5:7,214,350, −log10 p = 6.13). However, multiple rare alleles may be responsible for both associations, because if we include rarer SNPs in the analysis (see below), we find several highly significant associations very close to *AGO9* (top SNP 3.6 kb downstream; chr5:7,201,933, −log10 p = 11.69 in CMT2-targeted mCHH, 8.15 in RdDM-targeted mCHH, minor allele count = 21) (Fig. 4G-H, S2A-B). The more distant top SNPs may thus be “synthetic” or “ghost” associations (Atwell et al., 2010).

GWAS for RdDM-dependent mCHH also identified another argonaute gene, *ARGONAUTE 1* (*AGO1*), with a crucial role in post-transcriptional gene silencing (Brodersen et al., 2008; the top SNP is found in the promoter region: chr1:17,895,231, −log10 p = 6.10), and

NUCLEAR RNA POLYMERASE D1B (*NRPD1B*), which encodes the largest subunit of nuclear DNA-dependent RNA polymerase V, and is an essential component of the RdDM pathway (Law and Jacobsen, 2010; top SNP 0.5 kb downstream: chr2:16,724,013, −log10 p = 6.74).

There was striking enrichment of *a priori* candidates even for p-value cutoffs well below genome-wide significance (Fig. 4D-E), demonstrating that many non-significant associations deserve further investigation. Strong enrichment was also found when we allowed associations with rarer alleles (Fig. 4G-H, S2A-B) or used a slightly less conservative correction for population structure (Fig. S2D-E), although in both cases produced clearly biased p-values (Fig. S3). Among the notable candidates identified this way was *METHYL-CPG-BINDING DOMAIN 3* (*MBD3*), for which several non-synonymous polymorphisms are associated with CMT2-dependent mCHH (Fig. S2B).

These less conservative approaches also identify a clear candidate for gbM, which otherwise has no clear associations (Fig. 4C and F). Although significance levels are clearly inflated (Fig. S3), we find a strong association at *DNA METHYLTRANSFERASE 1* (*MET1*), which responsible for replication of CG methylation (Kawashima and Berger 2014), and hence is an excellent candidate (Fig. S2F, chr5: 19,925,444, −log10 p = 9.02).

Natural variations of transcriptomes and transcriptional regulation

Because DNA methylation can modulate gene expression, we next analyzed the transcriptomes from 727 accessions grown at 22°C (Fig. 1C). These accessions express, on average, transcripts from 18,000 genes (Fig. 5A). Comparing groups of accessions defined by genetic distances (The 1001 Genomes Consortium, 2016), we found 5,725 differentially expressed genes (DEGs) between relict accessions, an ancestral diverse group, and non-relict accessions (Fig. 5B). These DEGs were a subset of the 22,085 DEGs between all admixture groups (Fig. 5B; Table S4), suggesting further diversification of the transcriptomes among geographic groups. The two sets of DEGs were enriched for distinct biological processes (Fig. 5C). The most variable genes were enriched in functions related to biotic and temperature responses, likely reflecting adaptation to their natural environments. DEGs between relict and non-relict groups were enriched in ribosomal biogenesis and translation processes, suggesting the regulation of this energy intensive process contributed to the successful expansion of non-relict groups.

Co-expression network analysis (Langfelder and Horvath, 2008) identified eight modules each for relict and non-relict accessions (Table S4). Seven of the eight relict modules had significant overlap with at least one non-relict module (Fig. 5D) and were enriched for distinct biological processes preserved in one of the overlapping non-relict modules (Fig. 5E): biotic responses (M4 and M5; Fig. 5F), abiotic responses (M1; Fig. 5G), development (M2; Fig. 5H), cell cycle (M3; Fig. S4A) and photosynthesis (M7 and M8; Fig. S4B). The non-relict modules showed no or weak correlation with flowering time (Atwell et al., 2010) (Fig. S4C), suggesting that coexpression is unlikely driven by developmental stage at the time of sample collection. Using transcription factor binding sites (TFBS) identified by DNA affinity purification sequencing (DAP-seq) for the non-relict accession Col-0 (O'Malley et al., 2016), we found that non-relict modules were targeted by distinct TF

families (Fig. 5I, S4D), including the expected WRKY TFs for the biotic response modules, bZIPs for abiotic response modules, NACs for the development module, as well as yet unknown connections. Further DAP-seq experiments using TF variants and DNA from relict accessions will provide evidence for the mechanism behind preservation and emergence of co-expression modules (Fig. 5E).

To link methylation and expression differences we mapped expression quantitative trait loci (eQTL) with the 1001 Genomes SNP data, which identified genetic loci associated with gene expression. We then used GWA of gene expression with differentially methylated bins (100 bp; DMB) to pinpoint methylation-dependent eQTL (eQTL^{epi}, where *epi* is CG-, CH-, C-DMB; Table S5). Both cis-eQTL and cis-eQTL^{epi} were enriched at the TSS, and the highest numbers of cis-eQTL^{epi} were found for CH- and C-DMB (Fig. 6A), consistent with the silencing effect of these methylation contexts. As TF binding provides a mechanism for how methylation may affect gene expression, we compared the genetic and methylation variants to the 2.7 million TFBS of 329 TFs identified on Col-0 leaf DNA with methylcytosines (Col-0 cistrome) and the additional ~180,000 TFBS identified on methylation-free DNA (Col-0 epicistrome) (O'Malley et al., 2016). Around 25% of CH-DMBs (73,366) and 22% (48,109) of C-DMBs overlapped with the Col-0 cistrome and epicistrome (Fig. 6B-C), regions that harbor binding sites that may become available or occluded depending on the methylation state. Merged binding profiles of TF families showed two patterns of enrichment in DMBs (Fig. S5A). Of 45 families, 13 were depleted in CG-DMB but slightly enriched in CH-DMB, and one, the E2FDP family, was specifically enriched in C-DMB. This family includes the cell cycle regulator E2F, and methylation-regulated transcription is a potential mechanism for cell cycle variations (Sterken et al., 2009).

Members of the same TF family that have similar binding motifs may differ in their genome-wide binding profiles (O'Malley et al., 2016). We therefore also performed enrichment analysis on individual TFs. Most TF binding sites were depleted at eQTL$^{CH\text{-}DMB}$ while the associations with eQTL were evenly distributed between enrichment and depletion (Fig. 6D). Ranking of the TFs by enrichment in eQTL or eQTL$^{CH\text{-}DMB}$ identified three groups (Fig. 6D dotted and dashed lines, Fig. 6E). Group 1, including the C2H2 zinc finger TF STZ, had binding sites enriched in both eQTL and eQTL$^{CH\text{-}DMB}$. Binding sites for group 2 and 3 were enriched in either eQTL or eQTL$^{CH\text{-}DMB}$, respectively. Group 2 TF included the heat shock response factor HSFA6B and the meristem formation TF CUC2. MYB-related family members were found in both Group 1 (AT1G74840) and Group 3 (EPR1, AT4G01280, AT3G10113). These results suggest that genome and methylome variation interact to regulate gene expression through distinct sets of TFs.

In mammals, methylation in both CG and non-CG contexts is absent in binding sites of selected TFs (Lister et al., 2009; Domcke et al., 2015), but the relationship between methylation variation in natural populations and TF binding has not been analyzed systematically. Binding inhibition by methylation (O'Malley et al., 2016) was predicted to be stronger for TFs depleted at eQTL$^{CH\text{-}DMB}$ compared to those that are enriched for such loci (Fig. 6D-E). This general trend held true for the entire set of 352 TFs with methylation inhibition data, i.e., the more strongly a TF was inhibited by mCH methylation, the more

depleted its binding sites were at eQTL$^{CH\text{-}DMB}$, while the level of methylation inhibition and enrichment at eQTL were not correlated (Fig. 6F). The depletion of TFBS in mCH regions may be due to the low CG content of the TF motifs (Fig. S5B), although the motif CG content also contributed to the methylation inhibition of binding (Fig. S5C). This suggests a complex interplay between evolution of genetic and methylation variation and TF binding: binding sites for methylation inhibited TFs are selected against in methylated regions, possibly by the elimination of CG dinucleotides, to avoid dramatic changes in binding in response to methylation changes. Consistent with this hypothesis, TFs for which binding is enriched in eQTL$^{CH\text{-}DMB}$ have moderate methylation sensitivity (Group 3, Fig. 6D-E), potentially allowing methylation changes to fine tune binding.

Epigenome variation is shaped by genome structural variation

Our methylome analyses were based on the Col-0 reference genome substituted with accession-specific SNPs and small deletions, but did not include structural variation (SV) information, which may also affect plant epigenomes (Lisch, 2013). To relate SVs to methylome variation, we created physical genome maps (*contigs*) for nine accessions that represent a high-diversity panel (The 1001 Genomes Consortium, 2016) including Col-0 as reference control (TAIR10; Fig. 7A). These contigs were built from images of ultra-long fluorescently labeled DNA molecules (Lam et al., 2012). These averaged 284 kb (max. 1.5 Mb), long enough to span very large repeat arrays. The nine genomes assembled into 86 (Lesno-4) to 239 (Cvi-0) contigs (N50 > 1.1 Mb; Fig. 7A).

Aligning the Col-0 contigs to the TAIR10 assembly identified 29 mis-assemblies in the original reference (2.5 - 59 kb, Table S6A). For the accessions in the diversity panel, alignments covered 76% (Cvi-0) to 94% (Lu4-2) of the reference (Fig. 7A and 7C and S6A), with most alignment gaps being pericentromeric (Table S6A and B). We found an average of 6.2 SVs per Mb (Fig. 7A), representing insertions, deletions (indels) or rearrangements relative to the reference. The German accessions Erg2-6 and Lu4-2 represented the lower (5.7 indels/Mb) and upper (6.8 indels) end of the range, although their collection sites were only 20 km apart. Indel size ranged between 2.5 kb (resolution cutoff) and over 110 kb (average 10.8 kb; Fig. 7B, Table S6A). Notably, each accession had on average 3.43 Mb unique sequences not present in the reference, and lacked 3.54 Mb of reference sequences. The nearly symmetrical “gains” and “losses” relative to the reference set the optical maps apart from previous efforts based on *de novo* assemblies of short reads, which suffered from reference bias and therefore always reported more “losses” than “gains” (e.g. Cao et al., 2011). Since the reference largely lacks centromere sequences, these statistics only reflect variation in the chromosome arms. The “deletion” or “absence” alleles were more likely to be the major alleles than “insertion” or “presence” alleles, which were also less frequently shared between accessions (46%) than deletion alleles (67%) (Fig. 7E). In fact, only 5% of all insertions, but 22% of all deletions were shared among six or more accessions. This is expected if *A. thaliana* genomes are continuing to shrink, as suggested before (Hu et al., 2011). Indels were dispersed along the chromosomes with increasing density of shared insertions towards the centromeres (Fig. 7C and S6A). Physical contigs also allowed us to observe large scale rearrangements and more complex SVs, such as a 1.2 Mb inversion on the short arm of chromosome 4 (Fransz et al., 2000), and a local translocation on

chromosome 1 where DNA fragments (289 kb Cvi-0 and 412 kb Lesno-4) swapped place with a neighboring fragment, without changing orientation. As another example, Yeg-8 chromosome 4 (Fig. 7D) harbored a local inverted translocation of 907 kb, including a 323 kb insertion.

As the physical contigs do not provide DNA sequence content, we analyzed reference annotations around the SVs. TEs were present in the vast majority of SV loci (92%)(Table S6C). Helitron-class TEs were enriched around insertions, potentially reflecting copy-number variation as Helitrons replicate as rolling circles (Kapitonov and Jurka, 2001). Genes, present in 86% of SVs, were functionally enriched for defense response with emphasis on NLR genes, independent of SV-type (in/del, shared/unique; Table S6D). Indeed, NLRs reside in highly syntenic and TE rich clusters (Meyers et al., 2003; Leister, 2004), and 37% of TEs within 10 kb of NLR genes inside SVs were Helitrons.

The nine accessions analyzed had together 1,317 PE genes, with 729 (55%) being in SV regions (Table S6E; Fisher exact test p = 4.3e-58). Insertion or deletion of TEs in combination with rapid silencing of recently inserted TEs, may change the propensity of genes to change epiallelic state. We speculate that a subset, if not all, of the remaining 588 PE genes were located in SVs smaller than 2.5 kb and thus undetected by our optical maps.

DMRs could only be analyzed at the borders of SVs, and possibly reflect gain or loss of spreading teM. In insertions, we observed hypermethylated DMRs in up to 11%, and hypomethylated DMRs in up to 17% (Table S6F). Over half of all deletion sites were hypomethylated, and up to 17% harbored hypermethylated DMRs (Table S6F). Up to eight SVs per accession harbored both types of DMRs. Overall, 22-50% of SVs were differentially methylated (Table S6F), suggesting SVs in natural populations are closely related to methylation variants.

Disease resistance loci are major targets of both structural and methylation variation

The predominant gene family linked to C-DMRs and PE loci were NLR type disease resistance genes (Fig. 2L and S1N), which represent one of the largest plant gene families with over 150 members in *A. thaliana*. Our physical contigs were particularly variable at NLR loci, consistent with previous, more limited analyses of individual NLR clusters (Chae et al., 2014; Leister, 2004; Meyers et al., 2003). To provide an example of such an extremely polymorphic region, we focused on a cluster of nine NLR genes in the reference Col-0, which includes the NLR pair *RRS1/RPS4* (chr5:18,150,000-18,352,500) (Gassmann et al., 1999). Indels, on average five (Table S7), expanded this region (Col-0 201 kb) by up to 9 kb (Yeg-8), or shrunk it by up to 11 kb relative to the reference (IP-Cum-1; Fig. S6B). *RRS1B* and *RPS4B* (Saucet et al., 2015) were present in all accessions, flanked by 12 differentially methylated TEs (Helitron and MuDR) (Fig. 7F and S6B). While transcriptome data revealed no effect of the variable proximal indel state, a close-by *F-box* gene (*AT5G44980)* had elevated expression levels in accessions with overlapping insertions, suggesting a duplication and dosage effect (Fig. 7G). The larger *RRS1/RPS4* sub-cluster encodes seven NLRs and 29 differentially methylated TEs (Col-0), seven within NLR introns (*AT5G45200,* 2 Helitron; *AT5G45230,*4 MuDR; *RRS1,* 1 MuDR), but without effects on expression levels.

The *RRS1/RPS4* pair was, in contrast to *RRS1B/RPS4B*, only expressed in indel-free lines (Fig. 7G; Table S7).

Importantly, while the lack of mapped short reads from genome and methylome sequencing had suggested deletions of three NLRs (*AT5G45220*, *AT5G45230* and *AT5G45240*) in three accessions (21 kb; Lu4-2, Nicas-1 and Yeg-8), and additionally of *RRS1/RPS4* (36 kb total) in IP-Cum-1 (Fig. S6B), optical map contigs provided clear evidence for insertions rather than deletions, indicating that these regions can be completely replaced by unknown sequence content.

CONCLUSION

The *A. thaliana* 1001 Epigenomes project provides evidence that methylation is correlated with geography and climate of origin. This supports the notion that methylation plays a role in adaptation (Fig. 3C; Dubin et al. 2015). Indeed, our study shows that epigenomic changes are associated with environmental responses, and especially immunity genes. This makes plants distinct from humans, where epigenomic changes in germ cells (Gkountela et al., 2015; Guo et al., 2015; Tang et al., 2015) or adult tissues (Schultz et al., 2015) are associated with developmental control genes.

TEs are responsible for most indels and are enriched at disease resistance loci, where Helitron and MuDR transposons shape gene arrangements, DNA methylation and gene expression. While we identify that gbM is not required for a functional transcriptome, epiallele conversion between gbM and teM, likely induced by TE movement, can be a part of the evolutionary toolbox to alter gene expression either directly on the gene, or its regulatory elements. Selection could also explain the existence of major alleles leading to striking GWAS results for TE methylation. Further exploration of these should provide insight into the evolution and function of this genomic immune system.

TF binding may provide a further mechanism for linking genome and epigenome variation to adaptation: binding sites for distinct sets of TFs may respond to changes in sequence and methylation to establish gene expression modules for major biological processes essential for adaptation.

Surprisingly, *AGO1* and *AGO9* were associated with genome-wide average mCHH levels, given that knockout of either locus does not affect average mCHH levels within RdDM-target regions (Stroud et al., 2013). Importantly, GWAS associations not only identified genes known to be involved in epigenetics, but also novel loci. Identifying these genes (which could be lethal when knocked out) would lead to significant insight into DNA methylation and gene silencing pathways.

Methylome studies for crops like rice, maize and soybean, which have larger genomes with expanded TE families, have higher mCG and mCHG levels but similar mCHH levels compared to *Arabidopsis* (Niederhuth et al., 2016; Seymour et al., 2014; Takuno et al., 2016). Since TE transposition greatly impacts epigenomic diversity among *A. thaliana* accessions, crops are likely to show much more local epigenomic diversity within a species. The high variability in average methylation levels between *A. thaliana* accessions is a

reminder that conclusions about species-specific DNA methylomes based on single accessions should be met with caution. Deeper understanding of epigenome evolution is thus a prerequisite for future inter- and intraspecific comparative epigenomic studies.

SV analyses not only revealed that sequence gains and losses in individual accessions are nearly symmetrical, but also suggested a tight interplay between genome and epigenome evolution. The next step will be to integrate these with high-quality sequence-based genome assemblies, as a prerequisite for identifying the specific DNA sequences that vary between accessions and that contribute to methylome and transcriptome variation.

EXPERIMENTAL PROCEDURES

Please see EXTENDED EXPERIMENTAL PROCEDURES for detailed experimental and analysis methods.

Plant materials

Seeds are available from the Arabidopsis Biological Resource Center (ABRC) under accession IDs CS76427, CS76636, CS78885 and CS78942.

MethylC-seq

MethylC-seq library preparation, read mapping, base calling was performed as described previously (Lister et al. 2011).

Identification of differentially methylated regions

Differentially methylated regions (DMRs) were identified using the methylpy pipeline (Schultz et al., 2015). Methylation levels of each region are calculated as the frequency of C base calls at C positions within the region divided by the frequency of C and T base calls at C positions within the region.

RNA-seq and identification of differentially expressed genes

RNA-seq libraries were prepared using Truseq RNA kit (Illumina, San Diego, CA) following manufacturer's instruction. Reads were mapped using STAR aligner (Dobin et al., 2012) to TAIR10 genome and annotation. Gene level expression was quantified for TAIR10 annotated genes and batch normalized by the RUVseq package (Risso et al., 2014). Differentially expressed genes were called by the DESeq2 package (Love et al., 2014).

Physical mapping and identification of structural variations

HMW DNA was extracted using the Fix'n'Chop protocol (BioNano Genomics, San Diego, CA), and then fluorescently nick-labeled (Nt.BspQI; New England Biolabs, Ipswich, MA) using IrysPrep kit. Single molecule physical mapping was performed using the BioNano Genomics Irys system following manufacturers recommendations. Molecule data was assembled using IrysView 2.3 and SVs were called using custom Python scripts.

Genome wide association studies

Genome wide association mapping was performed using EMMAX algorithm (Kang et al., 2010). eQTL and eQTL^{epi} analysis was performed by the LIMIX (Lippert et al., 2014).

Data release

Data can be visualized using the 1001 Epigenomes Project genome browser (http://neomorph.salk.edu/1001.php).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

We thank members of the Ecker laboratory for assistance of sample preparation, Matthew D. Schultz for assistance of methylome analyses. T.K. was supported by the Japan Society for the Promotion of Sciences Research Abroad Fellowship. F.J. is supported by a Human Frontier Science Program long-term fellowship. This research was supported by grants from the National Institutes of Health (R00GM100000 to R.J.S.), a collaborative grant from Austrian Science Fund and DFG (SPP ADAPTOMICS to M.N. and D.W.), the ERC (MAXMAP, M.N., IMMUNEMESIS, D.W.), the National Science Foundation (MCB 0929402 and MCB 1122246 to J.R.E.). J.R.E. is an investigator of the Howard Hughes Medical Institute and Gordon and Betty Moore Foundation (GBMF 3034). We acknowledge the Texas Advanced Computing Center at The University of Texas at Austin for providing computing resources.

REFERENCES

Atwell S, Huang YS, Vilhjalmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT, et al. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature. 2010; 465:627–631. [PubMed: 20336072]

Becker C, Hagmann J, Muller J, Koenig D, Stegle O, Borgwardt K, Weigel D. Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. Nature. 2011; 480:245–249. [PubMed: 22057020]

Bewick AJ, Ji L, Niederhuth CE, Willing E-M, Hofmeister BT, Shi X, Wang L, Lu Z, Rohr NA, Hartwig B, et al. On the Origin and Evolutionary Consequences of Gene Body DNA Methylation. bioRxiv. 2016

Brodersen P, Sakvarelidze-Achard L, Bruun-Rasmussen M, Dunoyer P, Yamamoto YY, Sieburth L, Voinnet O. Widespread translational inhibition by plant miRNAs and siRNAs. Science. 2008; 320:1185–1190. [PubMed: 18483398]

Cao J, Schneeberger K, Ossowski S, Gunther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, et al. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat Genet. 2011; 43:956–963. [PubMed: 21874002]

Chae E, Bomblies K, Kim ST, Karelina D, Zaidem M, Ossowski S, Martin-Pizarro C, Laitinen RA, Rowan BA, Tenenboim H, et al. Species-wide genetic incompatibility analysis identifies immune genes as hot spots of deleterious epistasis. Cell. 2014; 159:1341–1351. [PubMed: 25467443]

Cortijo S, Wardenaar R, Colome-Tatche M, Gilly A, Etcheverry M, Labadie K, Caillieux E, Hospital F, Aury JM, Wincker P, et al. Mapping the epigenetic basis of complex traits. Science. 2014; 343:1145–1148. [PubMed: 24505129]

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013; 29:15–21. [PubMed: 23104886]

Domcke S, Bardet AF, Adrian Ginno P, Hartl D, Burger L, Schubeler D. Competition between DNA methylation and transcription factors determines binding of NRF1. Nature. 2015; 528:575–579. [PubMed: 26675734]

Dowen RH, Pelizzola M, Schmitz RJ, Lister R, Dowen JM, Nery JR, Dixon JE, Ecker JR. Widespread dynamic DNA methylation in response to biotic stress. Proc Natl Acad Sci USA. 2012; 109:E2183–2191. [PubMed: 22733782]

Dubin MJ, Zhang P, Meng D, Remigereau MS, Osborne EJ, Paolo Casale F, Drewe P, Kahles A, Jean G, Vilhjalmsson B, et al. DNA methylation in *Arabidopsis* has a genetic basis and shows evidence of local adaptation. eLife. 2015; 4:e05255. [PubMed: 25939354]

Fransz PF, Armstrong S, de Jong JH, Parnell LD, van Drunen C, Dean C, Zabel P, Bisseling T, Jones GH. Integrated cytogenetic map of chromosome arm 4S of A. thaliana: structural organization of heterochromatic knob and centromere region. Cell. 2000; 100:367–376. [PubMed: 10676818]

Gassmann W, Hinsch ME, Staskawicz BJ. The *Arabidopsis RPS4* bacterial-resistance gene is a member of the TIR-NBS-LRR family of disease-resistance genes. Plant J. 1999; 20:265–277. [PubMed: 10571887]

Gkountela S, Zhang KX, Shafiq TA, Liao WW, Hargan-Calvopina J, Chen PY, Clark AT. DNA Demethylation Dynamics in the Human Prenatal Germline. Cell. 2015; 161:1425–1436. [PubMed: 26004067]

Guo F, Yan L, Guo H, Li L, Hu B, Zhao Y, Yong J, Hu Y, Wang X, Wei Y, et al. The Transcriptome and DNA Methylome Landscapes of Human Primordial Germ Cells. Cell. 2015; 161:1437–1452. [PubMed: 26046443]

Hagmann J, Becker C, Muller J, Stegle O, Meyer RC, Wang G, Schneeberger K, Fitz J, Altmann T, Bergelson J, et al. Century-scale methylome stability in a recently diverged *Arabidopsis thaliana* lineage. PLoS Genetics. 2015; 11:e1004920. [PubMed: 25569172]

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, et al. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet. 2011; 43:476–481. [PubMed: 21478890]

Johannes F, Porcher E, Teixeira FK, Saliba-Colombani V, Simon M, Agier N, Bulski A, Albuisson J, Heredia F, Audigier P, et al. Assessing the impact of transgenerational epigenetic variation on complex traits. PLoS Genetics. 2009; 5:e1000530. [PubMed: 19557164]

Inagaki S, Kakutani T. What triggers differential DNA methylation of genes and TEs: contribution of body methylation? Cold Spring Harb Sym. 2012; 77:155–160.

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010; 42:348–354. [PubMed: 20208533]

Kapitonov VV, Jurka J. Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci USA. 2001; 98:8714–8719. [PubMed: 11447285]

Kawashima T, Berger F. Epigenetic reprogramming in plant sexual reproduction. Nature Reviews Genetics. 2014; 15:613–624.

Kooke R, Johannes F, Wardenaar R, Becker F, Etcheverry M, Colot V, Vreugdenhil D, Keurentjes JJ. Epigenetic basis of morphological variation and phenotypic plasticity in *Arabidopsis thaliana*. Plant Cell. 2015; 27:337–348. [PubMed: 25670769]

Lam ET, Hastie A, Lin C, Ehrlich D, Das SK, Austin MD, Deshpande P, Cao H, Nagarajan N, Xiao M, et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nat Biotech. 2012; 30:771–776.

Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008; 9:559. [PubMed: 19114008]

Law JA, Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet. 2010; 11:204–220. [PubMed: 20142834]

Leister D. Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance gene. Trends in Genetics : TIG. 2004; 20:116–122. [PubMed: 15049302]

Lippert C, Casale FP, Rakitsch B, Stegle O. LIMIX: genetic analysis of multiple traits. bioRxiv. 2014

Lisch D. How important are transposons for plant evolution? Nat Rev Genet. 2013; 14:49–61. [PubMed: 23247435]

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009; 462:315–322. [PubMed: 19829295]

Lister R, Pelizzola M, Kida YS, Hawkins RD, Nery JR, Hon G, Antosiewicz-Bourget J, O'Malley R, Castanon R, Klugman S, et al. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. Nature. 2011; 471:68–73. [PubMed: 21289626]

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; 15:550. [PubMed: 25516281]

Meyers BC, Kozik A, Griego A, Kuang H, Michelmore RW. Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. Plant Cell. 2003; 15:809–834. [PubMed: 12671079]

Miura A, Nakamura M, Inagaki S, Kobayashi A, Saze H, Kakutani T. An *Arabidopsis* jmjC domain protein protects transcribed genes from DNA methylation at CHG sites. EMBO J. 2009; 28:1078–1086. [PubMed: 19262562]

Niederhuth CE, Bewick AJ, Ji L, Alabady M, Kim KD, Page JT, Li Q, Rohr NA, Rambani A, Burke JM, et al. Widespread natural variation of DNA methylation within angiosperms. bioRxiv. 2016

Nuthikattu S, McCue AD, Panda K, Fultz D, DeFraia C, Thomas EN, Slotkin RK. The initiation of epigenetic silencing of active transposable elements is triggered by RDR6 and 21-22 nucleotide small interfering RNAs. Plant Phys. 2013; 162:116–131.

O'Malley RC, Huang SC, Song L, Lewsey MG, Bartlett A, Nery JR, Galli M, Gallavotti A, Ecker JR. Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. Cell. 2016; 165:1280–1292. [PubMed: 27203113]

Pecinka A, Abdelsamad A, Vu GT. Hidden genetic nature of epigenetic natural variation in plants. Trends in Plant Science. 2013; 18:625–632. [PubMed: 23953885]

Pignatta D, Erdmann RM, Scheer E, Picard CL, Bell GW, Gehring M. Natural epigenetic polymorphisms lead to intraspecific variation in *Arabidopsis* gene imprinting. eLife. 2014; 3

Ratcliffe OJ, Kumimoto RW, Wong BJ, Riechmann JL. Analysis of the *Arabidopsis* MADS AFFECTING FLOWERING gene family: *MAF2* prevents vernalization by short periods of cold. Plant Cell. 2003; 15:1159–1169. [PubMed: 12724541]

Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotech. 2014; 32:896–902.

Rodriguez-Leal D, Leon-Martinez G, Abad-Vivero U, Vielle-Calzada JP. Natural variation in epigenetic pathways affects the specification of female gamete precursors in *Arabidopsis*. Plant Cell. 2015; 27:1034–1045. [PubMed: 25829442]

Saucet SB, Ma Y, Sarris PF, Furzer OJ, Sohn KH, Jones JD. Two linked pairs of *Arabidopsis* TNL resistance genes independently confer recognition of bacterial effector AvrRps4. Nat Comm. 2015; 6:6338.

Schmitz RJ, Ecker JR. Epigenetic and epigenomic variation in *Arabidopsis thaliana*. Trends in Plant Science. 2012; 17:149–154. [PubMed: 22342533]

Schmitz RJ, Schultz MD, Lewsey MG, O'Malley RC, Urich MA, Libiger O, Schork NJ, Ecker JR. Transgenerational epigenetic instability is a source of novel methylation variants. Science. 2011; 334:369–373. [PubMed: 21921155]

Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, Libiger O, Alix A, McCosh RB, Chen H, Schork NJ, et al. Patterns of population epigenomic diversity. Nature. 2013; 495:193–198. [PubMed: 23467092]

Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, Rajagopal N, Nery JR, Urich MA, Chen H, et al. Human body epigenome maps reveal noncanonical DNA methylation variation. Nature. 2015; 523:212–216. [PubMed: 26030523]

Secco D, Wang C, Shou H, Schultz MD, Chiarenza S, Nussaume L, Ecker JR, Whelan J, Lister R. Stress induced gene expression drives transient DNA methylation changes at adjacent repetitive elements. eLife. 2015; 4

Seymour DK, Koenig D, Hagmann J, Becker C, Weigel D. Evolution of DNA methylation patterns in the Brassicaceae is driven by differences in genome organization. PLoS Genetics. 2014; 10:e1004785. [PubMed: 25393550]

Shen X, De Jonge J, Forsberg SK, Pettersson ME, Sheng Z, Hennig L, Carlborg O. Natural CMT2 variation is associated with genome-wide methylation changes and temperature seasonality. PLoS Genetics. 2014; 10:e1004842. [PubMed: 25503602]

Sterken R, Kiekens R, Coppens E, Vercauteren I, Zabeau M, Inze D, Flowers J, Vuylsteke M. A population genomics study of the *Arabidopsis* core cell cycle genes shows the signature of natural selection. Plant Cell. 2009; 21:2987–2998. [PubMed: 19880799]

Stroud H, Do T, Du J, Zhong X, Feng S, Johnson L, Patel DJ, Jacobsen SE. Non-CG methylation patterns shape the epigenetic landscape in *Arabidopsis*. Nature Structural & Molecular Biology. 2014; 21:64–72.

Stroud H, Greenberg MV, Feng S, Bernatavichute YV, Jacobsen SE. Comprehensive analysis of silencing mutants reveals complex regulation of the *Arabidopsis* methylome. Cell. 2013; 152:352–364. [PubMed: 23313553]

Takuno S, Ran J-H, Gaut BS. Evolutionary patterns of genic DNA methylation vary across land plants. Nat Plants. 2016; 2:15222. [PubMed: 27249194]

Tang WW, Dietmann S, Irie N, Leitch HG, Floros VI, Bradshaw CR, Hackett JA, Chinnery PF, Surani MA. A Unique Gene Regulatory Network Resets the Human Germline Epigenome for Development. Cell. 2015; 161:1453–1467. [PubMed: 26046444]

Teixeira FK, Colot V. Gene body DNA methylation in plants: a means to an end or an end to a means? EMBO J. 2009; 28:997–998. [PubMed: 19384348]

The 1001 Genomes Consortium. 1135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. Cell. 2016 *in press*. DOI 10.1016/j.cell.2016.05.063.

Tran RK, Henikoff JG, Zilberman D, Ditt RF, Jacobsen SE, Henikoff S. DNA methylation profiling identifies CG methylation clusters in *Arabidopsis* genes. Curr Biol. 2005; 15:154–159. [PubMed: 15668172]

Vaughn MW, Tanurdzic M, Lippman Z, Jiang H, Carrasquillo R, Rabinowicz PD, Dedhia N, McCombie WR, Agier N, Bulski A, et al. Epigenetic natural variation in *Arabidopsis thaliana*. PLoS Biology. 2007; 5:e174. [PubMed: 17579518]

Zemach A, Kim MY, Hsieh PH, Coleman-Derr D, Eshed-Williams L, Thao K, Harmer SL, Zilberman D. The *Arabidopsis* nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. Cell. 2013; 153:193–205. [PubMed: 23540698]

Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW, Chen H, Henderson IR, Shinn P, Pellegrini M, Jacobsen SE, et al. Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*. Cell. 2006; 126:1189–1201. [PubMed: 16949657]

Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S. Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. Nat Genet. 2007; 39:61–69. [PubMed: 17128275]

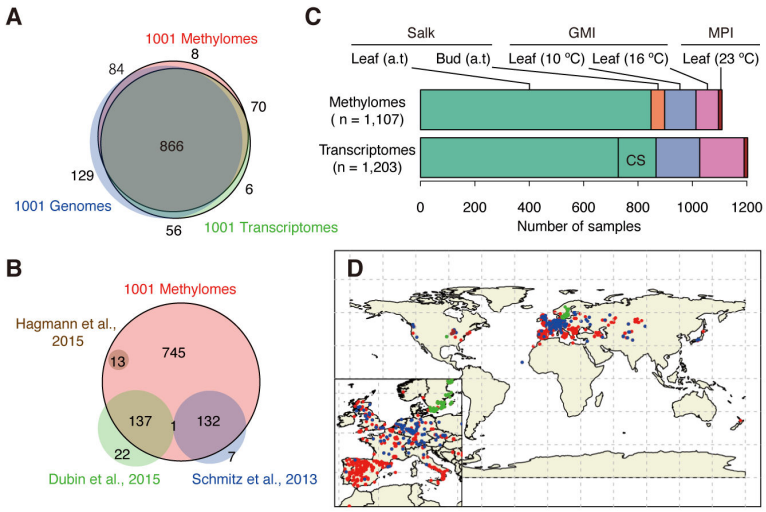

Figure 1. Origins of 1,028 accessions included in the 1001 Epigenomes project methylomes and transcriptomes

(A) Overlap between accessions used in the 1001 genomes, methylomes and transcriptomes projects. All are included in the initial selection of 1,227 accessions.

(B) Overlap with published population methylome studies (Dubin et al., 2015; Hagmann et al., 2015; Schmitz et al., 2013).

(C) Sample types for the 1,028 accessions. Plants were grown and sequenced at the Salk, GMI or MPI. Since more than one sample type was analyzed for some accessions, there were 1,107 methylomes from 1,028 accessions and 1,203 transcriptomes from 998 accessions. Transcriptomes were sequenced mainly on the Illumina platform, and partly with SOLiD platform (CS). Growth temperatures at in parentheses. a.t: ambient temperature 22°C.

(D) Original collection locations of accessions in the 1001 Epigenomes project. Colors correspond to (B). Dotted lines indicate longitude and latitude grids at 30° intervals.

See also Figure S1, Table S2 and S3.

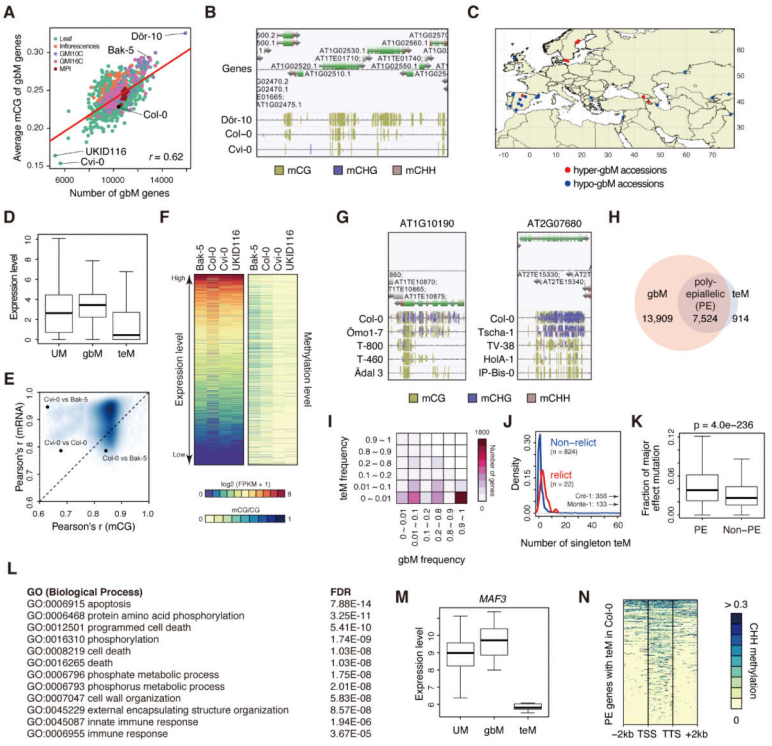

Figure 2. DNA methylation patterns within gene bodies are associated with expression

(A) Correlation between the number of gene body methylated (gbM) genes (x-axis) and their average CG methylation levels (y-axis). Each point is one accession, colored by data source in Fig. 1C. Cvi-0 and UKID116 are the most hypomethylated accessions, while Dör-10 is the most hypermethylated.

(B) A snapshot of the 1001 Epigenomes Anno-J browser (http://neomorph.salk.edu/1001.php) for an example region on chromosome 1, showing hyper-, average and hypo-gene body methylation in Dör-10, Col-0 and Cvi-0. Top track is gene model and yellow ticks in the bottom three tracks indicate CG methylation levels at each cytosine.

(C) Geographical distribution of hyper- and hypo- gbM accessions.

(D) Population-wide relation between epiallele and gene expression levels. Expression levels are shown as log2 (FPKM + 1). UM: unmethylated genes. gbM: gene body methylated genes. teM: TE-like methylated genes.

(E) Comparison of pairwise correlations for mCG within gene bodies (x-axis) and mRNA abundance across all accessions (y-axis), indicating positions for hypomethylated Cvi-0 vs. hypermethylated Bak-5, Cvi-0 vs. average methylated Col-0 and Col-0 vs. Bak-5.

(F) Transcript abundance (left) of hypermethylated (Bak-5), average (Col-0) and hypomethylated (Cvi-0, UKID116) accessions and mCG within gene bodies (right). Genes were sorted by average expression level.

(G) AnnoJ browser snapshots for representative poly-epiallelic (PE) genes AT1G10190 and AT2G07680 that show gbM (mainly mCG) or teM (all contexts) in selected accessions.

(H) Venn Diagram for the numbers of gbM genes, teM genes and their overlap (PE genes), based on Salk-grown samples.

(I) Binning of PE genes based on gbM frequency (the fraction of accessions with gbM epiallele among Salk-grown accessions) and teM frequency. Each tile on the heatmap indicates the number of PE genes in the corresponding bin.
(J) Density distribution of teM singletons in relict and non-relict accessions.
(K) Enrichment of PE genes for major effect mutations.
(L) Enrichment of PE genes for GO terms related to immunity and phosphorylation.
(M) Association of epiallele state and gene expression level at *MAF3*.
(N) Heatmap of CHH methylation around PE genes that have a teM epiallele but do not contain TEs within their gene bodies or within 500 bp up-/downstream in Col-0. TSS: transcription start site, TTS: transcription termination site.

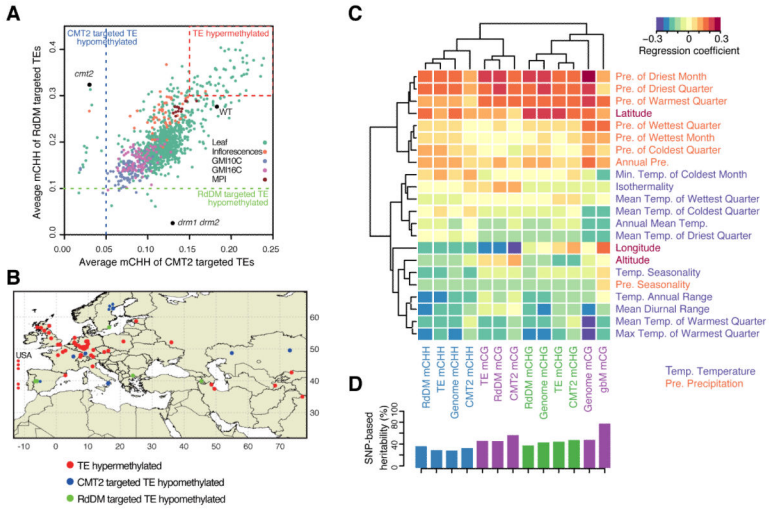

Figure 3. Global patterns of methylation variation

(A) Average CHH methylation levels of CMT2 targeted TEs (x-axis) and RdDM targeted TEs (y-axis) in worldwide accessions and mutants.

(B) Geographic distribution of Salk-grown accessions with hypermethylated TEs and hypomethylated CMT2/RdDM targeted TEs.

(C) Heatmap for kinship-corrected correlations between the genome-wide methylation level for a particular methylation context (in columns) and environmental/geographic variables (in the rows). Rows and columns were ordered by clustering by similarity in correlation. Pre.: Precipitation. Temp.: Temperature.

(D) The fraction of variation in genome-wide methylation (all contexts) across accessions that can be explained by genome-wide kinship, i.e., SNP heritability. See also Extended Experimental Procedures.

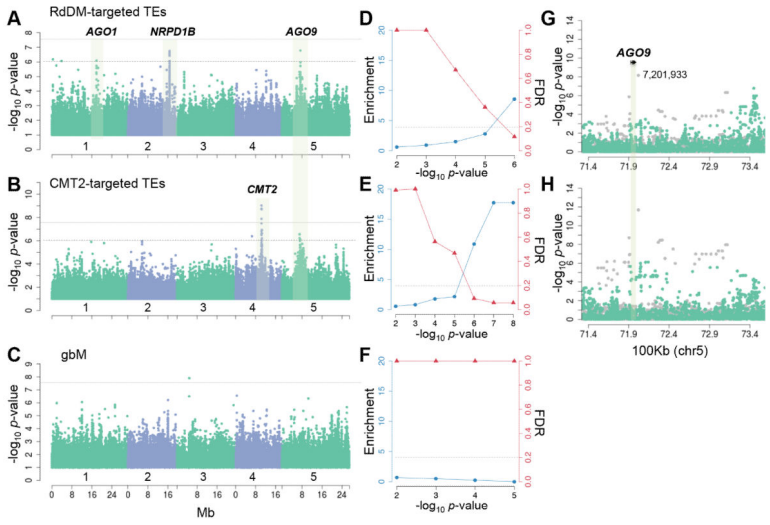

Figure 4. Genome-wide association study on methylation levels

(A-C) Manhattan plots of GWAS results for genome-wide average methylation phenotypes: (A) CHH methylation of RdDM-targeted TEs; (B) CHH methylation of CMT2-targeted TEs; (C) CG gbM. Highlights indicate peaks containing strong *a priori* candidates. Horizontal gray solid and dashed lines indicate genome-wide threshold p=0.05 with Bonferroni correction and FDR 20% defined by enrichment analysis, respectively. Only SNP with minor allele frequency (MAF) over 5% are included.

(D-F) Enrichment and FDR corresponding to (A-C) (based on enrichment of *a priori* candidates, see Extended Experimental Procedures). The horizontal dashed lines at 0.2 correspond to FDR 20%.

(G-H) Close-up of chromosome 5 peak around *AGO9* corresponding to (A-B). Green dots show non-reference SNPs with MAF > 5%, and and gray dots show rare SNPs (MAF 1 - 5%).

See also Figure S2 and S3

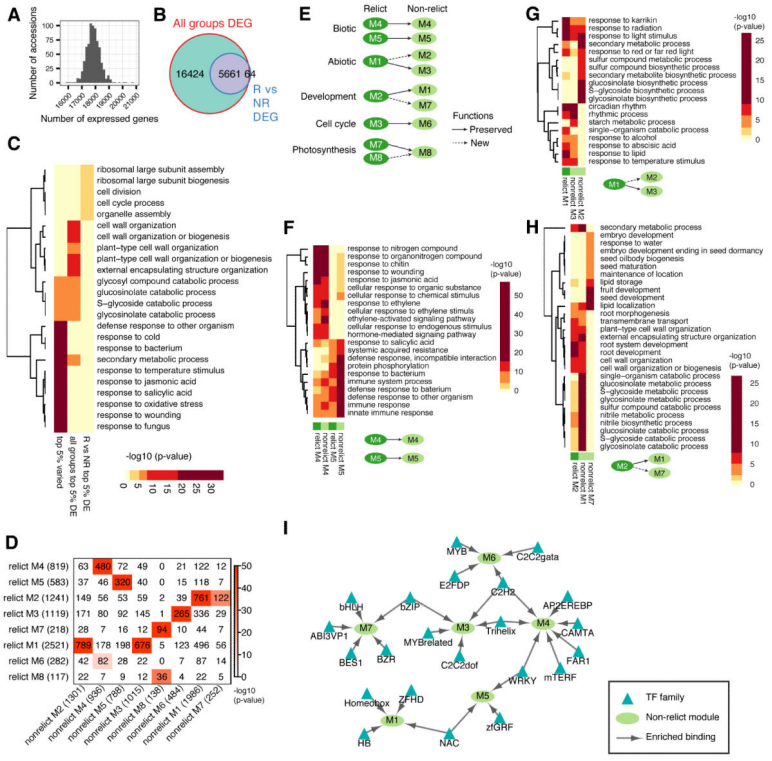

Figure 5. Differentially expressed genes among accessions and co-expression networks

(A) Histogram of number of expressed genes in the accessions.

(B) Differentially expressed genes (DEG) between relic and non-relict groups (“R vs. NR”) were a subset of DEGs between all admixture groups (“All groups”).

(C) Heatmap of −log10 enrichment p-values for the ten most enriched GO terms (rows) in top 5% varied genes and DEGs (columns). The row dendrogram was obtained by hierarchical clustering.

(D) Overlap of co-expression gene modules between relict and non-relict accessions. P-values from Fisher’s exact test.

(E) Shared and divergent functions between relict and non-relict modules.

(F-H) Heatmaps of −log10 enrichment p-values for the ten most enriched GO terms in relict modules M4, M5 and non-relict modules M4, M5 (F), relict module M1 and non-relict modules M2, M3 (G), and relict module M2 and non-relict modules M1, M7 (H). Row dendrograms were generated as in (C).

(I) Non-relict modules were enriched for binding sites from distinct TF families.

See also Figure S4 and Table S4.

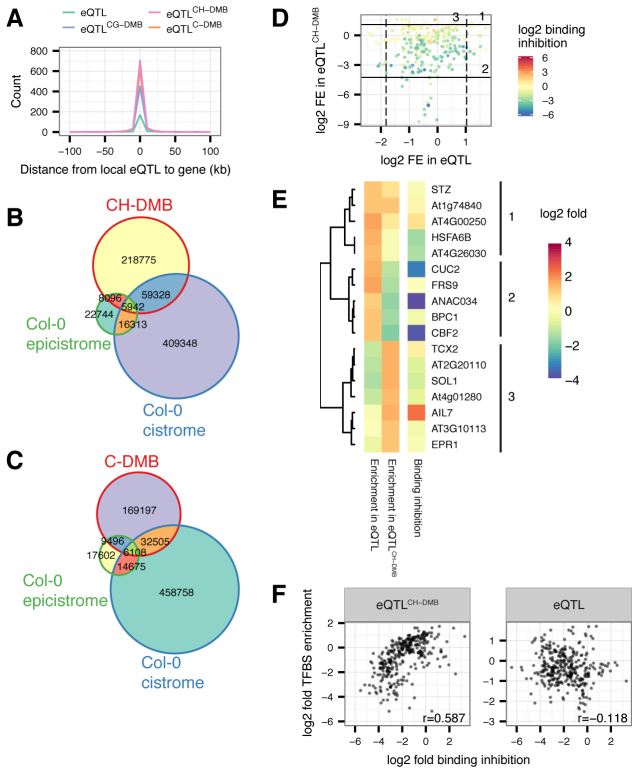

Figure 6. Relationship between eQTL, eQTL^{epi} and TFBSs

(A) Distribution of distances from cis-eQTL and cis-eQTL^{epi} to TSS (within 100kb), where *epi* is CG-, CH-, and C-DMB.

(B, C) Overlap of CH-DMB (B) and C-DMB (C) with Col-0 cistrome and epicistrome.

(D, E) Enrichment/depletion of TFBS at eQTL and eQTL$^{CH\text{-}DMB}$ identified three TF groups.

(F) TF methylation sensitivities (x-axis) were correlated with enrichment of binding sites (y-axis) at eQTL$^{CH\text{-}DMB}$ (left) but not at eQTL (right).

See also Figure S5 and Table S5.

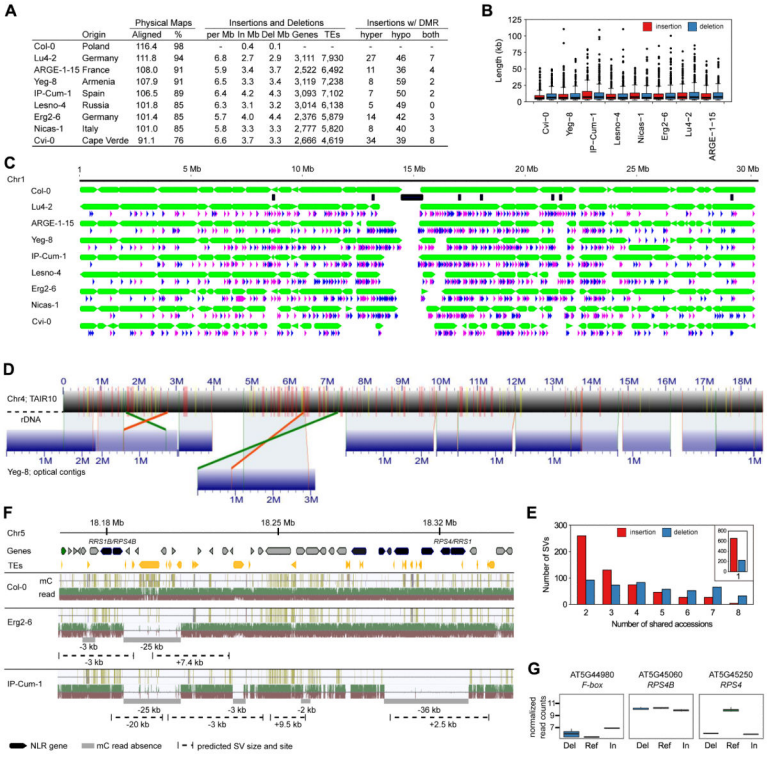

Figure 7. Genome structure is linked to differential methylation and transcription

(A) Summary of genome maps created using images of nick-labeled ultra-long DNA molecules for nine Arabidopsis accessions, including the reference accession Col-0. Columns are (from left): Accession ID, country of origin, total alignment length of optical maps against TAIR10 in Mb and percentage, counts for combined insertions and deletions (indels) per Mb of TAIR10, insertions per Mb, deletions per Mb, genes and TEs within indels, insertions with hyper-, hypo- or mixed DMRs.

(B) Boxplot for the length distribution of insertions (red) and deletions (blue) for all eight accessions in kb.

(C) Graphical representation of optical contigs aligned to chromosome 5 (green boxed arrows). Black boxes show TAIR10 mis-assemblies. Arrows in magenta represent regions not present in TAIR10 (insertion), and blue represents regions absent in that accession (deletion).

(D) Overview of Yeg-8 chromosome 4 optical contig alignments (blue) against TAIR10 (grey). Crossing green and red lines identify two inversions. Red and Yellow lines depict insertions and deletions against TAIR10. The dashed line represents 1.2 Mb of rDNA/ nucleolar organizer. Labels show size in Mb.

(E) Alignments were used to call insertions (red) and deletions (blue) relative to the TAIR10 reference. A large portion of SVs is shared amongst accessions.

(F) *RRS1-RPS4* NLR locus on chromosome 5, comparing Erg2-6 and IP-Cum-1 to Col-0. TAIR10 annotations are shown on top as non-NLR genes (grey), NLR genes (black), TEs (orange) and F-box gene (green; see 7G). Both methylated cytosines (mC) and WGS read coverage (read) tracks are shown per accession. Grey bars show mapping-free regions that overlap with predicted SV loci (dashed lines), and size differences are indicated.

(G) Transcript expression levels of three genes in accessions where the gene overlap with deletion (Del), as reference (Ref), and insertion (In) loci. Y-axis shows normalized RNA-seq read counts.

See also Figure S6, Table S6 and S7.