Supplementary Material

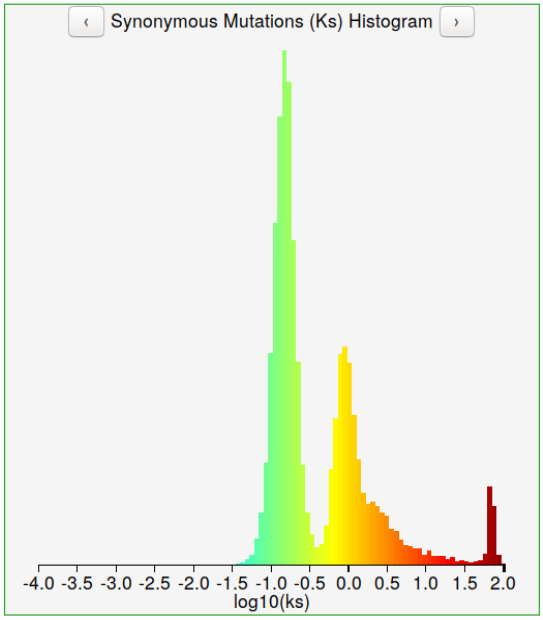
Gene body methylation is under selection in *Arabidopsis thaliana*

Aline Muyle[*], Jeffrey Ross-Ibarra[†], Danelle K. Seymour[‡], Brandon S. Gaut[*]

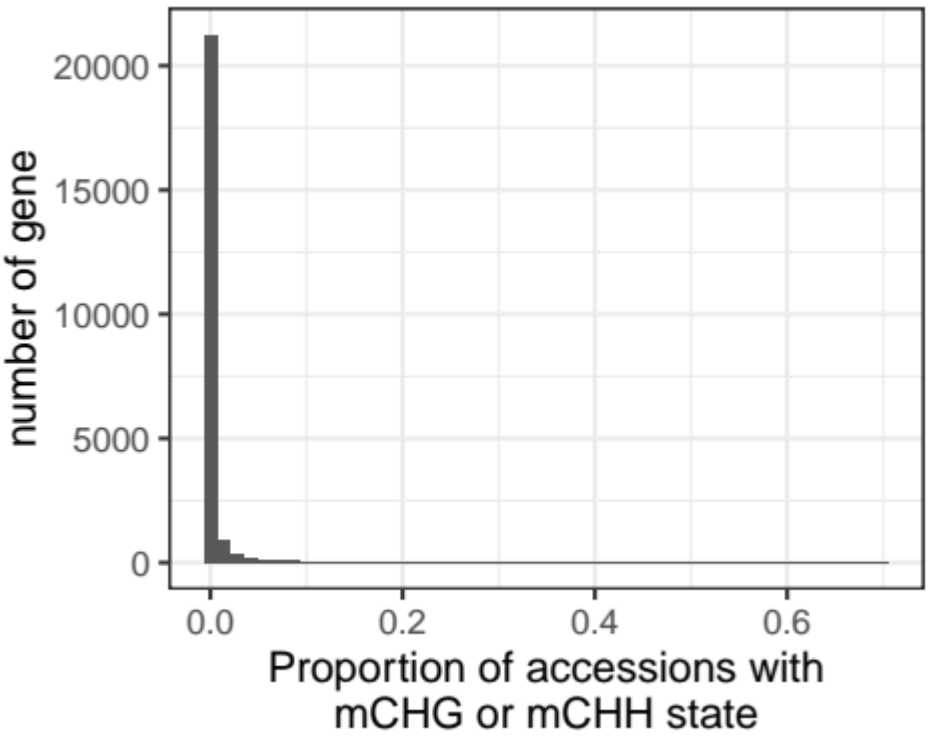
[*] Ecology and Evolutionary Biology, UC Irvine, Irvine.

[†] Evolution and Ecology, UC Davis, Davis.

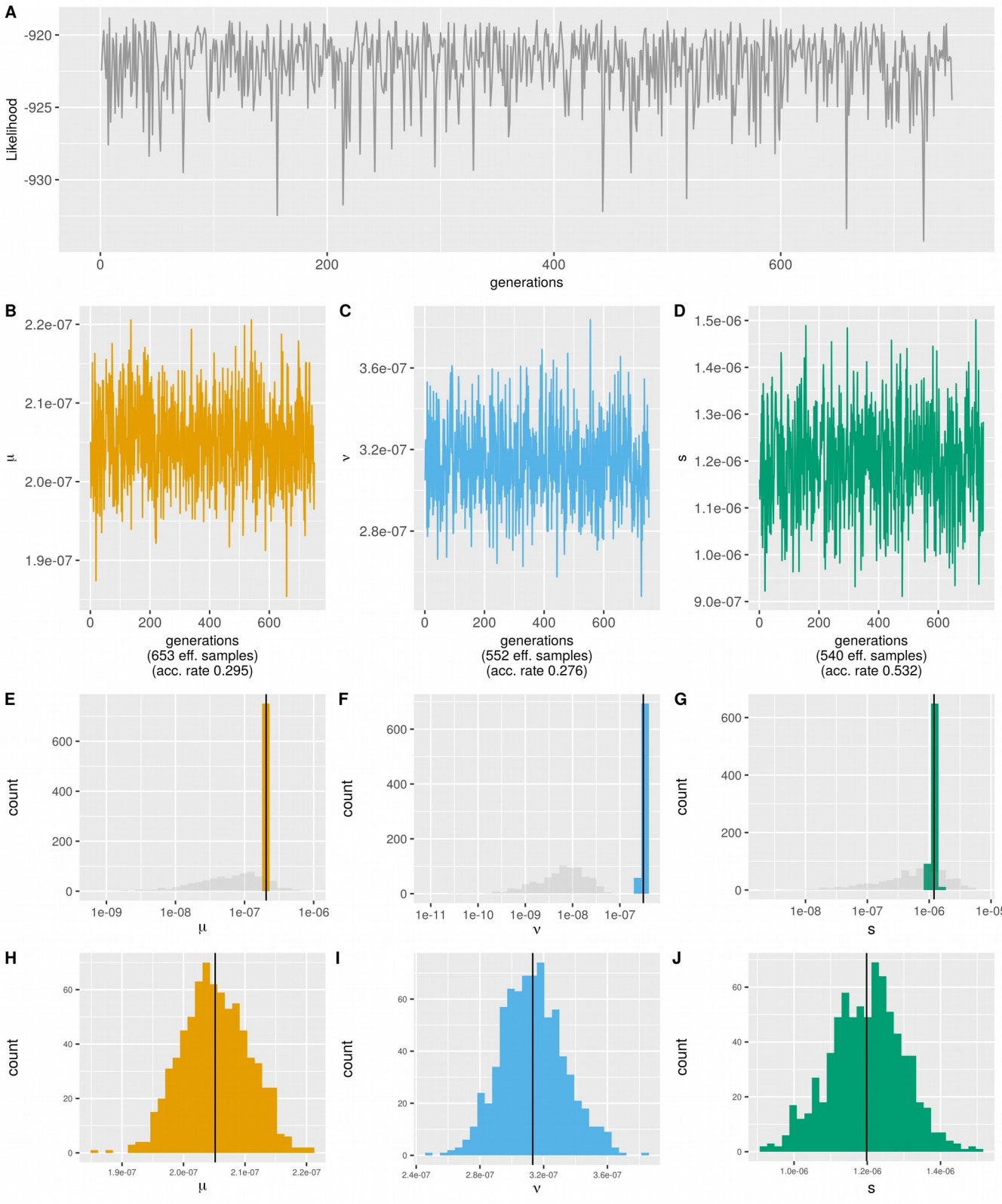
[‡] Botany & Plant Sciences, UC Riverside, Riverside, United States.



Supplementary Figure S1: CoGe tool SynMap3D distribution of dS values between *A. thaliana*, *A. lyrata* and *C. rubella*. The green peak corresponds to syntenic orthologs while the yellow-orange and red peaks correspond to out-paralogs (paralogs caused by duplications that predate speciation).

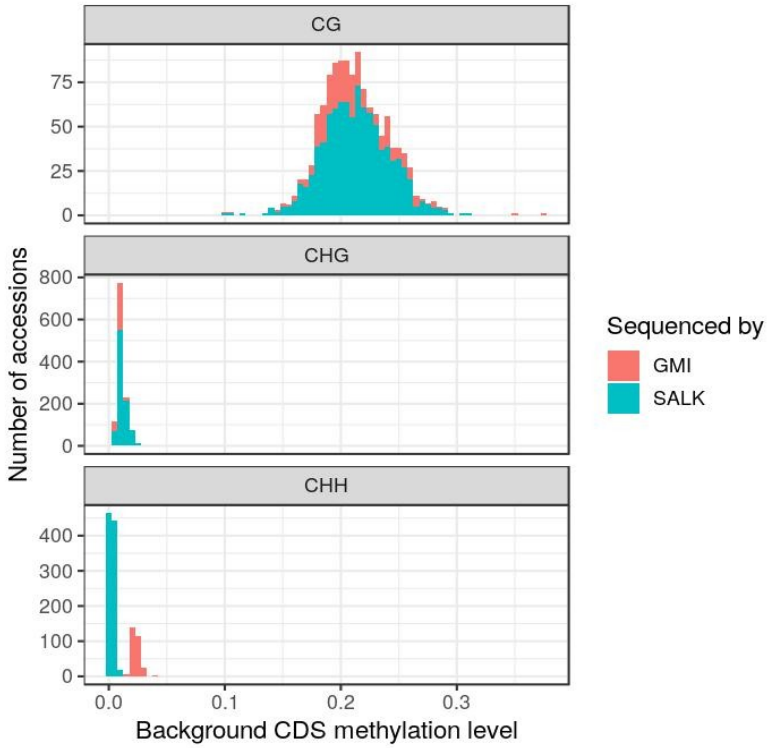


Supplementary Figure S2: Distribution of the proportion of accessions with either mCHG or mCHH methylation state among 22,609 genes with at least 600 accessions with UM or gbM methylation state in the Salk dataset.

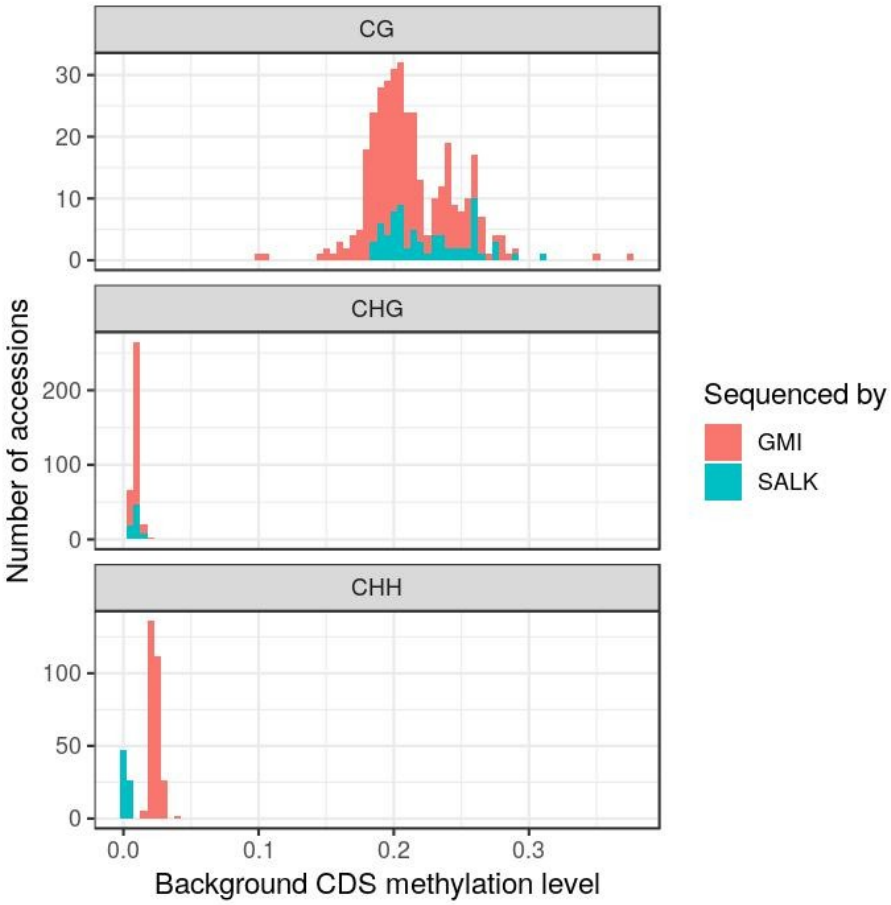


<u>Supplementary Figure S3:</u> Example of mcmc run diagnostics for the ancestrally gbM genes with selection on the gbM state for the Salk dataset. A. Traces for the likelihood over the sampled generations. **B.** Traces for the epimutation rate μ over the 653 sampled generations (acceptance rate 29.5%). **C.** Traces for the epimutation

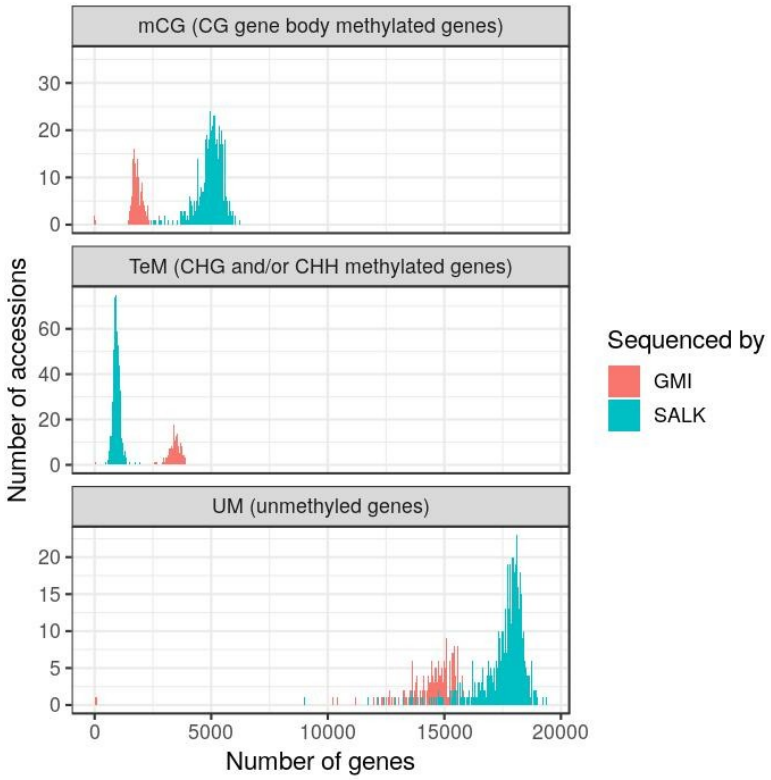
rate v over the 552 sampled generations (acceptance rate 27.6%). **D.** Traces for the selection coefficient s over the 540 sampled generations (acceptance rate 53.2%). **E.** Distribution of the prior (in grey) and the posterior (in orange) for the epimutation rate μ. **F.** Distribution of the prior (in grey) and the posterior (in blue) for the epimutation rate v. **G.** Distribution of the prior (in grey) and the posterior (in green) for the selection coefficient s. **H.** Zoom on the posterior distribution of the epimutation rate μ. **I.** Zoom on the posterior distribution of the epimutation rate v. **J.** Zoom on the posterior distribution of the selection coefficient s.



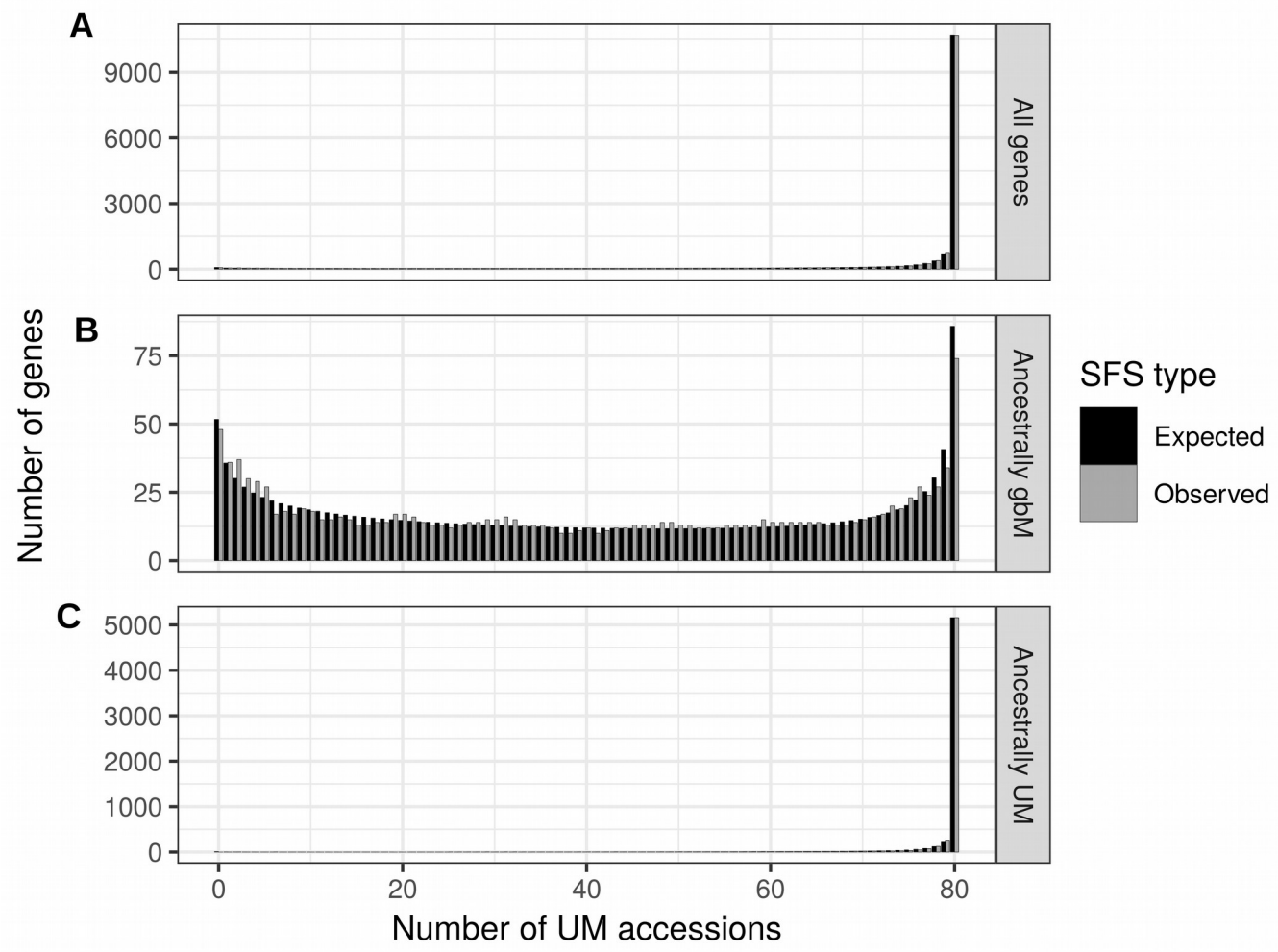
Supplementary Figure S4: Background methylation level of gene CDS in the 1001 methylome dataset of *A. thaliana* for the two sequencing Institutes (GMI stands for Gregor Mendel Institute). Background CHH methylation level is significantly higher in GMI accessions (Wilcoxon rank sum test one-sided p-value < 2.2×10^{-16}).



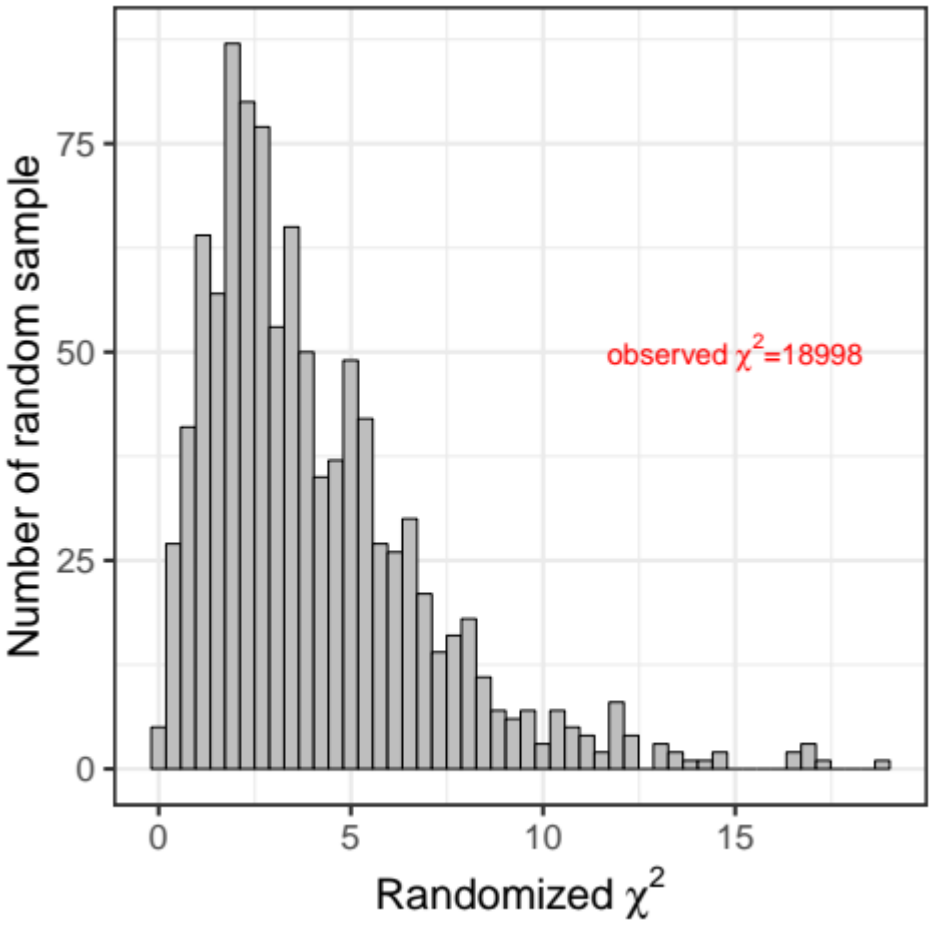
Supplementary Figure S5: Background methylation level of gene CDS in **Swedish** accessions of the 1001 methylome dataset of *A. thaliana* for the two sequencing Institutes (GMI stands for Gregor Mendel Institute). Background CHH methylation level is significantly higher in GMI accessions (Wilcoxon rank sum test one-sided p-value < 2.2e-16), suggesting that the different geographical origins of the accessions does not explain the difference in background CHH methylation between the two Institutes.



Supplementary Figure S6: Distribution of gene numbers for different methylation states in accessions of the 1001 methylome dataset of *A. thaliana* for the two sequencing Institutes (GMI stands for Gregor Mendel Institute). The difference in background CHH methylation between the two Institutes results in large differences in the inferred number of CHH methylated genes which results in lower numbers of gbM genes in accessions sequences by the GMI.



<u>Supplementary Figure S7:</u> Expected and Observed Site Frequency Spectra (SFS) of gene body methylation in the <u>GMI dataset</u>. The x-axis provides the number of UM accessions, out of a sample of 80. For these data, accessions that are not UM are gbM, meaning that genes with 80 UM accessions are fixed for the UM state in *A. thaliana* and genes with 0 UM accessions are fixed for the gbM state. The number of genes is provided on the y-axis. **A.** All genes (15,720 genes). **B.** Ancestrally gbM genes (1,383 genes). **C.** Ancestrally UM genes (6,078 genes). The expected SFS were drawn using the parameters estimated by the mcmc, using the best model in Supplementary Table S3. All three expected SFS fit the observed SFS well and did not differ significantly from the observed distribution (Pearson's χ-squared test $p>0.3$).



<u>Supplementary Figure S8:</u> Distribution of χ^2 values obtained after randomizing the data to test the association between gene methylation and expression level in the *A. thaliana*. For each permuted dataset, a linear model with mixed effects was used to assess the correlation between methylation and expression levels (Materials and Methods, equation 5). A χ^2 value > 7.8 represents a significant correlation between methylation level and expression level within genes (i.e., a *p*-value < 0.05 with 3 degrees of freedom), and such values were observed 5.4% of the time. The observed χ^2 (18,998) was higher than the highest χ^2 obtained on randomized data (18.94) by over a thousand-fold.