PNAS
www.pnas.org

Supplementary Information for

Natural variation in DNA methylation homeostasis and the emergence of epialleles

Yinwen Zhang[1,3], Jered M. Wendte[2,3], Lexiang Ji[1], Robert J. Schmitz[2,*]

Affiliations. [1]Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA,

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

[3]Contributed equally

*Correspondence: schmitz@uga.edu

This PDF file includes:

Supplementary Methods
Figures S1 to S5
SI References

Other supplementary materials for this manuscript include the following:

Datasets S1 to S11

www.pnas.org/cgi/doi/10.1073/pnas.1918172117

Supplemental Methods

Gene methylation status classification

To classify genes as gbM, teM, or UM, the total number of cytosines and the methylated cytosines were counted for cytosines in each context (CG, CHG, and CHH) for the coding sequences (CDS) of the primary transcript for each gene. The percentage of methylated sites for each sequence context in all coding regions from each accession were used as the background probability of having methylation on a single site. Given a background probability and the total number of cytosines and methylated cytosines, a p-value was calculated using a binomial distribution to show the cumulative probability of having higher number of methylated cytosines on a given gene. Then a q-values were calculated by adjusting p values by Benjamin–Hochberg FDR to control the false discovery rate.

Genes were classified as gbM if they had reads mapping to at least 20 CG sites and had a q-value < 0.05 for mCG and a q-value > 0.05 for mCHG and mCHH. Genes were classified as mCHG if they had reads mapping to at least 20 CHGs, a mCHG q-value < 0.05, and a mCHH q-value > 0.05. As mCG is commonly associated with mCHG, the q-value for mCG could be significant or insignificant in mCHG genes. Genes were classified as mCHH if they had reads mapping to at least 20 mCHH sites and a mCHH q-value < 0.05. Q-values for mCG and mCHG could be anything as both types of methylation are associated with mCHH. mCHG and mCHH genes were collectively referred to as teM genes. Genes were classified as unmethylated (UM) if they had reads mapping to at least 20 mCHH sites and had a q-value > 0.05 for all sequence contexts. For other cases, genes were classified into a low coverage category, since they did not meet the minimum requirement of number of reads mapping to the sequence context illustrated above.

To characterize the difference between gbM genes and UM genes, a gene's methylation status was further summarized based on its status among the 725 accessions. For each gene, the percentage of four methylation classes (UM, teM, gbM and low coverage) were summarized over 725 accessions. Genes were defined as UM genes if they are UM in over 90% of the accessions, without having been identified as gbM or teM in the remaining accessions. A gene was defined as gbM gene if it is gbM in at least one accession, unless there are more accessions with teM than gbM or the proportion of low coverage accessions was more than 50%. All gbM genes and UM genes defined in this way were used for the following gene feature comparisons and training of prediction models.

Gene features preparation

Number of CWG (CAG or CTG) and CG (CGG or CGT or CGC or CGA) sites were calculated by scanning a gene's primary transcript sequence with a three base window and step size of one base. Then, CWG and CG site frequencies were calculated by normalizing the number of each context to a gene's primary transcript length. Genes relative location to the pericentromere was obtained from calculating the absolute distance between a gene's location on the chromosome and the location of pericentromere, which was obtained from *A. thaliana* genome assembly annotation file (https://www.arabidopsis.org/download/). Gene expression levels used in the gene feature comparison between gbM genes and UM genes as shown in Figure 3 were the averaged FPKM values across all investigated 725 accessions for each gene.

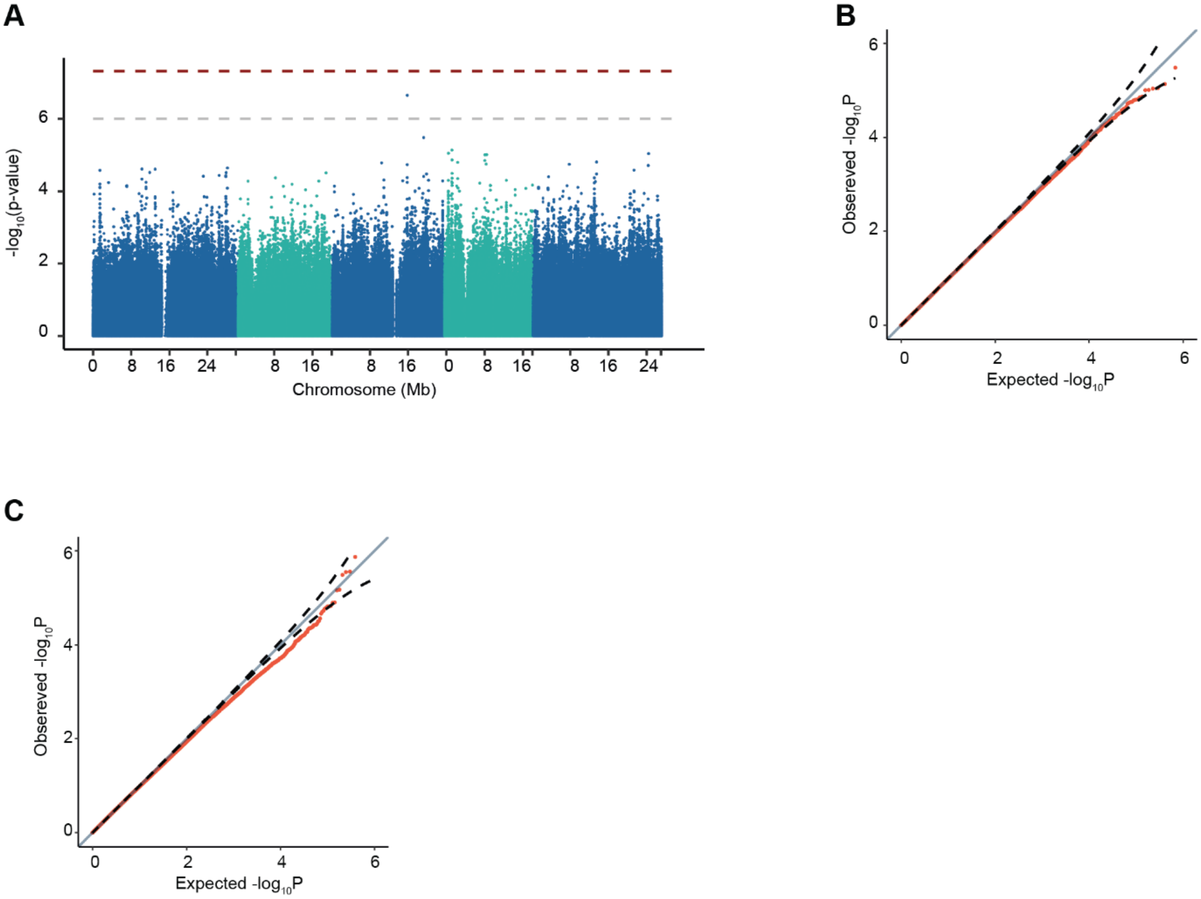
Substitution rates were calculated between CDS pairs between *A. thaliana* and *A. lyrata*. Reciprocal best BLAST with an e-value cutoff of <= 1e-08 was used to identify orthologs between *A. thaliana* and *A. lyrata*. Individual CDS pairs were aligned using PRANK (1), and Gblocks was applied to eliminate poorly aligned positions in an alignment with a cutoff of eight contiguous non-conserved positions and no gap positions allowed (2). The yn00 package in the program PAML for pairwise sequence comparison was used to estimate synonymous substitution rates, non-synonymous substitution rates, and adaptive evolution (dS, dN and w respectively) (3).

Prediction model for gene's methylation status

To build prediction models for a gene's methylation status, a correlation matrix was created from the 10 genic features listed in Dataset S6, and three of them (exon number, number of CWG/CG sites on the coding sequences) were removed before the model training process, since they showed an absolute correlation of 0.6 or higher with gene length. Before model training, the features were transformed to normalized data by subtracting the mean for each feature and dividing by the standard deviation. Among all gbM and UM genes, training and testing datasets were created by random sampling of 50% of the genes into a training set balancing the proportion of gbM genes and UM genes based on the original gene set. A dozen machine learning algorithms (Figure S4B) were used to train the prediction models. Five default values were tried for the main parameters of each algorithm during the training process and values with the best performance were chosen in the final evaluation. However, it should be noted that all algorithms performed well (Figure S4B).

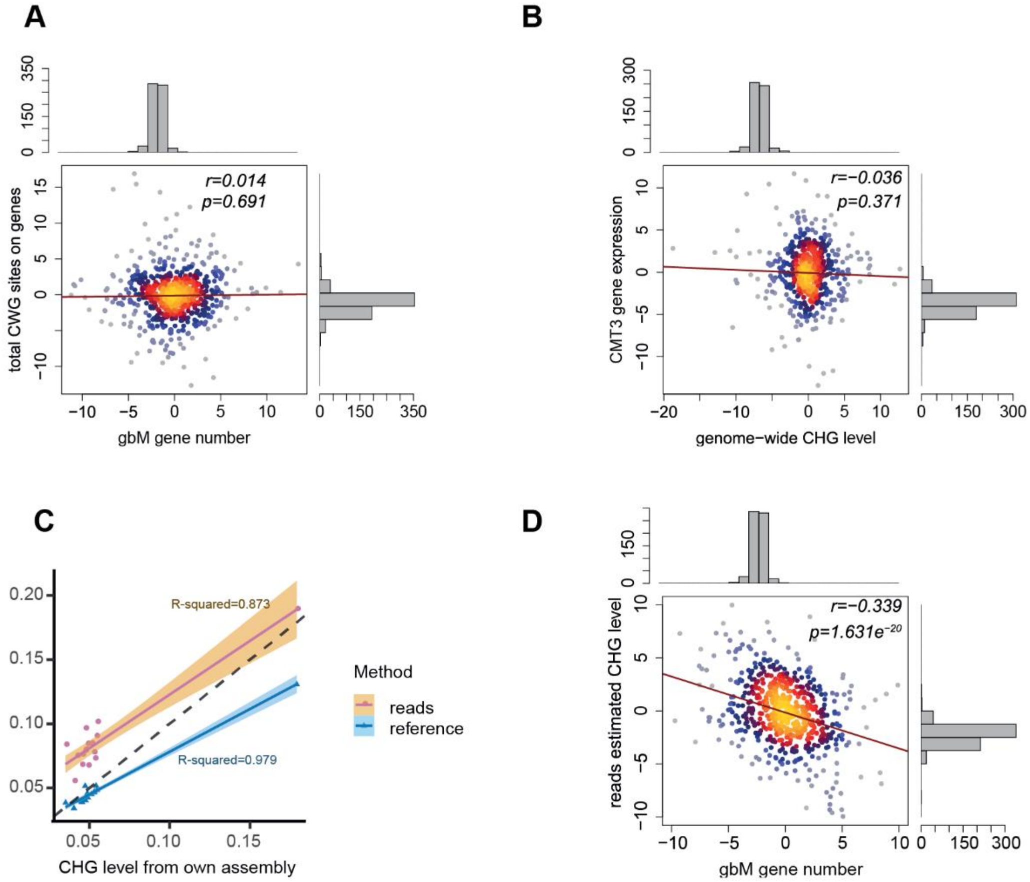
The accuracy of each algorithm is determined by a 10-fold cross-validation method. The 10-fold cross-validation method involves randomly splitting the training dataset into 10 equal sized subsets. Of the 10 subsets, a single subset was retained as the validation data for testing the model, and the remaining 9 subsets were used as training data. the cross-validation process was then repeated 10 times, with each of the 10 subsets used exactly once as the validation data. The process of randomly splitting the training dataset into 10 subsets was repeated 5 times. Together, the final model accuracy was taken as the mean of 50 validations. Since the model trained by the random forest algorithm showed the highest prediction accuracy, it was applied to predict a gene's methylation status in the testing set. The importance of features was estimated during model training. The prediction accuracy for all features was recorded, then testing was performed in the same way by permutating each feature. The difference between the prediction accuracy of the training dataset, including one permutated feature and that of the original dataset, were then averaged over all decision trees generated by the random forest algorithm and normalized by the standard deviation of the differences. The model training and performance evaluation process was done with the support of the caret package (https://github.com/topepo/caret/) in R.

FIG. S1.



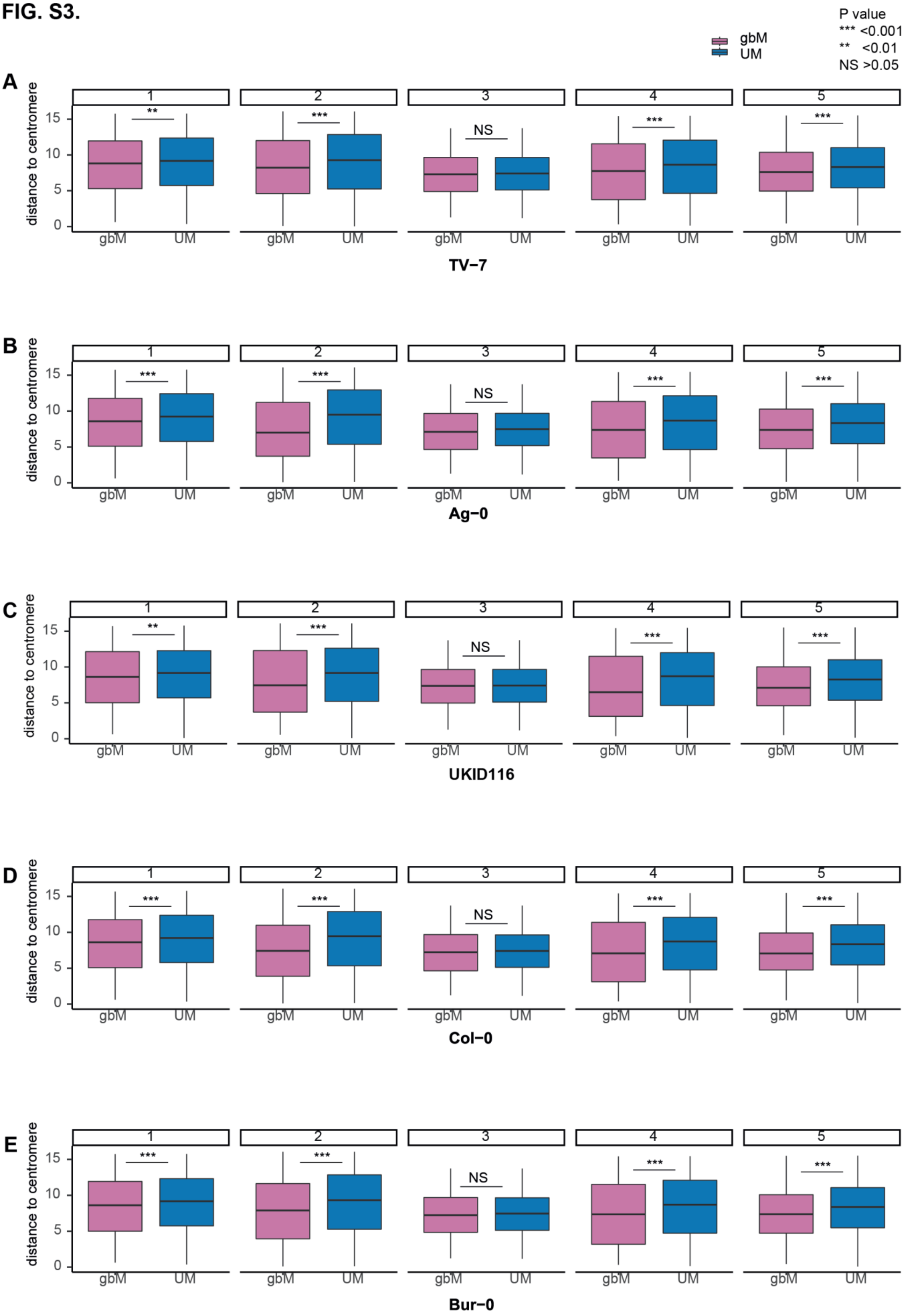
GWAS analysis for 620 accessions and corresponding QQ plots. (A) Manhattan plots of GWAS analysis for 620 accessions using gbM gene number as the phenotype. **(B)** QQ plot of GWAS analysis for 620 accessions. **(C)** QQ plot of GWAS analysis for 198 accessions in Figure 2B.

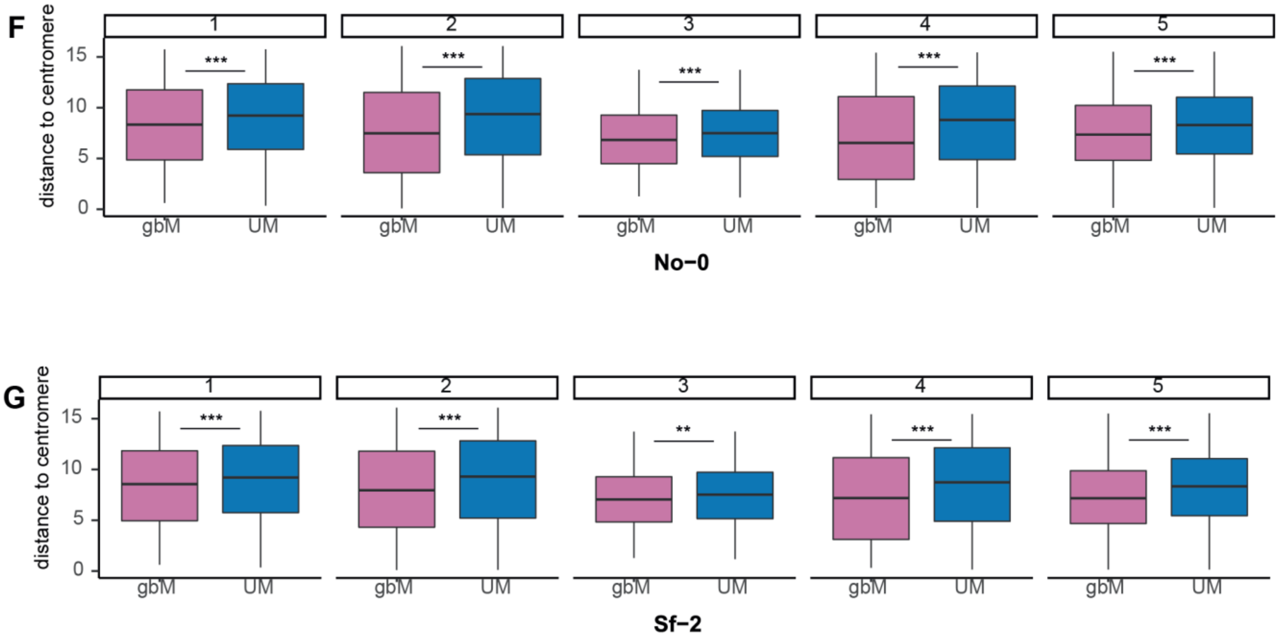
FIG. S2.



Correlations of gbM gene number with various genomic features. (A) The correlation between the number of gbM genes and total CMT3 sites number on coding genes. **(B)** The correlation between CMT3 gene expression and genome wide percent CHG methylation calculated by a nonreference-based DNA methylation predictive model. **(C)** The scatter plot shows whole genome CHG methylation levels estimated by the different methods. The x-axis is the whole genome CHG methylation level of 17 accessions estimated from mapping the methylomes to their own genome assemblies. The blue line shows the whole genome CHG methylation level estimated from mapping to the Col-0 reference. The yellow line shows the whole genome CHG methylation levels calculated by a nonreference-based DNA methylation predictive model, FASTmC. **(D)** The correlation between the number of gbM genes and genome-wide CHG methylation levels calculated by the nonreference-based method.

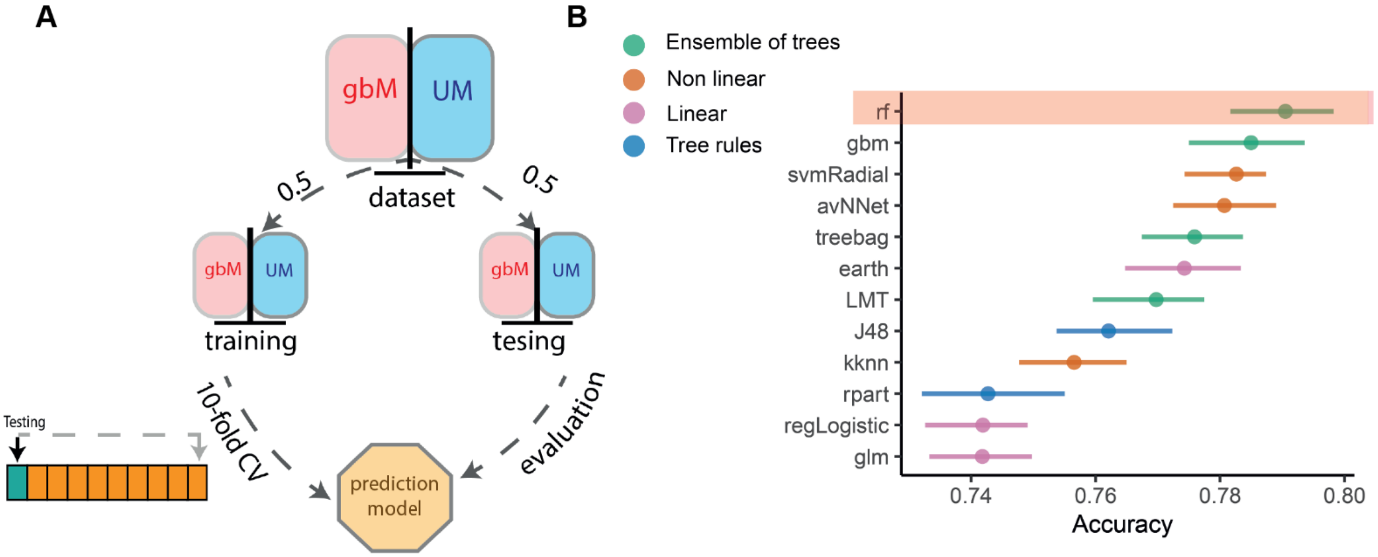
FIG. S3.


P value
*** <0.001
** <0.01
NS >0.05
gbM
UM
A
B
C
D
E
1
2
3
4
5
distance to centromere
0
5
10
15
gbM
UM
**
***
NS
TV−7
Ag−0
UKID116
Col−0
Bur−0



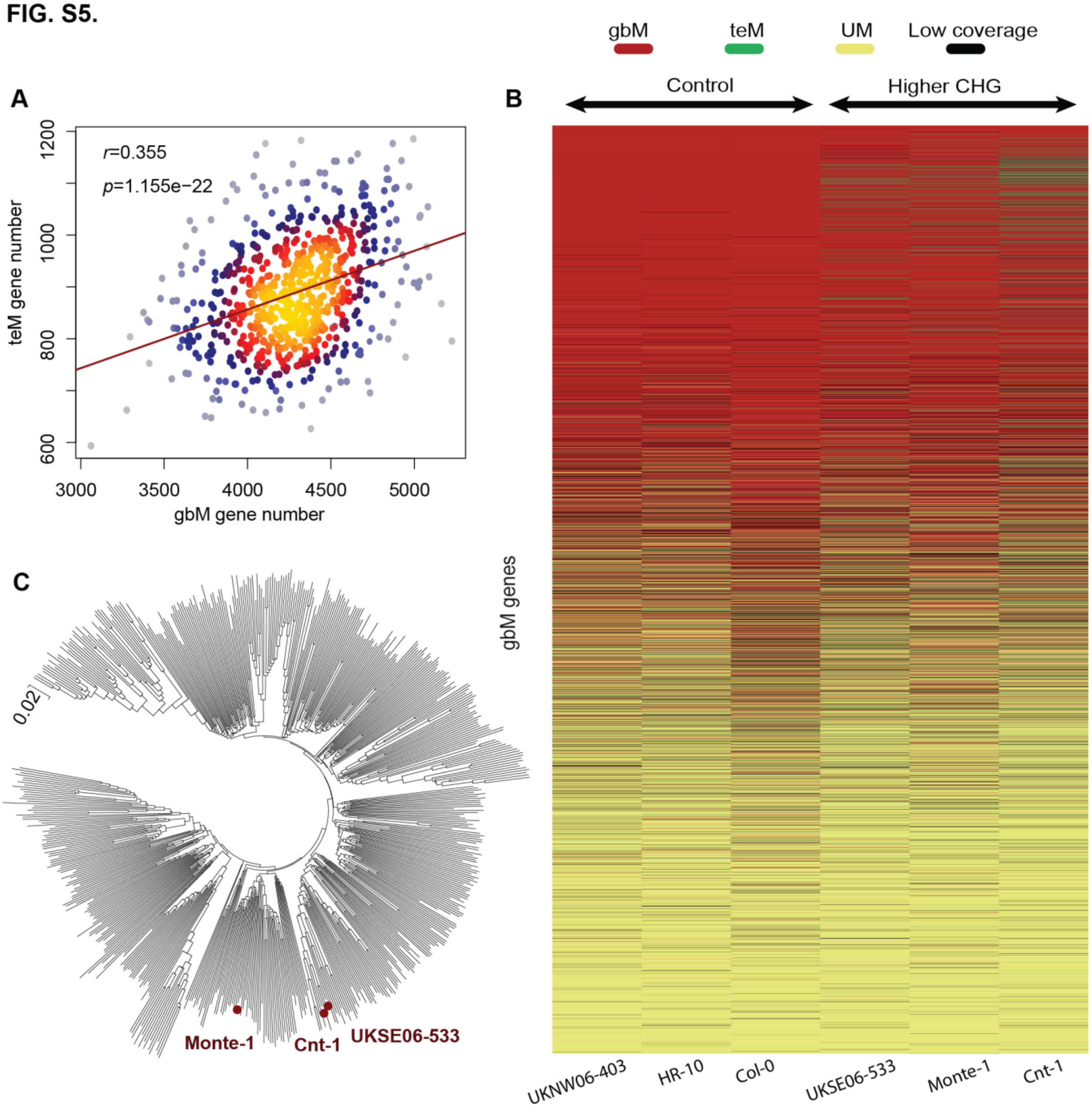
Gene location comparison between gbM and UM genes for seven accessions.

FIG. S4.



Gene methylation status prediction model construction and algorithm selection. (A) The pipeline for the machine learning model data sampling and model training process. **(B)** Rank of 12 machine learning algorithms based on the average prediction accuracy taken from 50 cross-validations during the model training process.

FIG. S5.



gbM and teM-status are on a continuous spectrum. (A) The number of gbM genes and TE-like genes are positively correlated in the population. **(B)** Heatmap shows the methylation status of gbM genes for six accessions including Cnt-1, Monte-1 and UKSE06-533 (top 3 accessions with highest number of genes gain CHG methylation within core gbM genes (with >=90% gbM frequency), and 3 more typical accessions. **(C)** Neighbor-joining tree of all 725 accessions, with the location of Cnt-1 and Monte-1 and UKSE06-533 noted.

Dataset S1 (separate file). Methylation status for each gene in each *A. thaliana* accession used in this study.

Dataset S2 (separate file). The number of gbM, teM and UM genes in each accession.

Dataset S3 (separate file). Distribution of the number of genes in each methylation category.

Dataset S4 (separate file). Average frequency of epiallelic states for gbM genes.

Dataset S5 (separate file). Values of genetic and epigenetic factors in each accession related to gbM gene number.

Dataset S6 (separate file). Epiallele frequencies and genic features of coding genes in *A. thaliana*.

Dataset S7 (separate file). Comparison of the distance to the centromere of gbM genes relative to UM genes in seven *A. thaliana* accessions.

Dataset S8 (separate file). True and predicted methylation status for genes in the testing dataset.

Dataset S9 (separate file). Distribution and enrichment p-value for mCHG-gain genes.

Dataset S10 (separate file). Genes that gain CHG methylation in CMT3-expressing *E. salsugineum* and their orthologs' methylation status in the *A. thaliana*.

Dataset S11 (separate file). The number of genes for each accession that are otherwise classified as core gbM genes in the population which exhibit high levels of CHG methylation

SI Reference

1. Loytynoja A (2014) Phylogeny-aware alignment with PRANK. Methods Mol Biol 1079:155-170.

2. Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular Biology and Evolution 17(4):540-552.

3. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution 24(8):1586-1591.