Supplementary Information

Expanded section and Discussion

Supplementary Fig. S1 to S30

Supplementary Tables 1-26

Expanded section and Discussion

RdDM targeted genes are re-expressed in pollen

The mQTL identified revealed that there is a strong association among some genetic variants and variation in DNA methylation, especially for C-DMRs. It is also well established that other genetic features, such as repeats, are important for guiding the RdDM machinery to target loci. For example, the intergenic sub-telomeric repeats 3' to the *MEDEA* locus and the repeated SINE elements and tandem repeats around the transcription start site of *FWA* are key regulatory sequences for controlling gene expression of these loci[1,2]. Although these loci are under transcriptional control by genetic elements, these specific elements are present and invariably methylated in every accession examined. Therefore, to understand the potential role of regions of the epigenome that are less prone to natural epigenetic variation we searched for loci that contained methylated alleles (methylation level ≥ 10%) in greater than 90% of the accessions sequenced and identified a total of 283 genes and 255 transposons. Essentially, these loci represent invariably methylated loci in this population similar to *FWA* and *MEDEA*. Furthermore, the expression of these loci was specifically activated

during pollen development (Fig. 5a and b). A previous study demonstrated that *DDM1* is not expressed in the pollen vegetative nucleus and that this results in release from epigenetic silencing and expression of transposons, providing a substrate to generate mobile small RNAs, which are transmitted to the germ line (sperm cells)[3]. This mechanism is not restricted to transposons as we identify protein-coding genes that are under control of the RdDM pathway and invariably methylated across this population are also expressed specifically in pollen (Fig. 5b). This re-activation of gene expression is not a general feature of pollen development as a control set of genes that show no evidence of being targeted by RdDM within this population or preference for pollen re-expression (Fig. 5c). In fact, these loci are expressed at lower levels in pollen development consistent with previous transcriptome analyses that revealed pollen has the fewest number of expressed loci in comparison to other vegetative tissues[4] (Supplementary Table 26). A closer examination of these invariably methylated genes with gene ontology[5] revealed a significant enrichment for two major categories, cell wall biology and translation (Table 1). These enriched GO terms are related to the major functions of pollen. For example, germination of pollen upon making contact with the stigma requires cell wall fusion and pollen tube elongation requires rapid cell wall expansion and protein synthesis to reach the central and egg cells for double fertilization. Therefore, these data indicate that the invariably methylated, pollen expressed genes are not only released from epigenetic silencing to reinforce their repressed state in the germ line, but also to function in pollen development.

Although these invariably methylated loci are under similar epigenetic control as transposons (Fig. 5a and b), it is likely that all RdDM-targeted loci are under control of

this mechanism regardless of their variability within this population. In fact, Col-0 genes targeted by RdDM and their corresponding expression levels are positively correlated (Spearman correlation; P Value $5.81e^{-27}$) specifically in pollen and seed development (Fig. 5d); whereas, all 55 other tissues tested revealed either a significant negative correlation or no correlation (Supplementary Table 18) between the loci targeted by RdDM and their corresponding expression levels (Fig. 5d). It is noteworthy that categories of genes showing positive correlations are stronger for loci that overlap transposon sequences (Fig. 5d). These data indicate that these loci have come under control of sequences that are evolutionarily silenced, which acts to restrict their expression to these specific stages of development (Fig. 5d).

Conclusion

Natural epigenomic variation is widespread within *Arabidopsis thaliana* and the population-based epigenomics approach used throughout this study has uncovered features of the DNA methylome that are not linked to underlying genetic variation such as all forms of SMPs and CG-DMRs. However, C-DMRs have positional association decay patterns similar to LD decay patterns for SNPs and in some cases are associated with local and distant genetic variants. Our combined analysis of genetic and methylation variation did not uncover a significant correlation between major effect mutations and genes silenced by the RdDM pathway suggesting that these genes may be targeted by this pathway for another purpose.

In fact, our study identifies protein-coding genes that are under control of the RdDM pathway, which likely serves to enact two possible fates. The first fate is permanent silencing of these loci in vegetative tissues similar to transposons. In fact,

ectopic expression of some of these genes in vegetative tissues results in pleiotropic phenotypes[6,7] indicating the importance of restricting their expression from vegetative development. A possible second fate of being targeted by the RdDM pathway could be the coordinated expression specifically to pollen and endosperm to ensure proper development. An example of this phenomenon is illustrated by the *FWA* locus, which reaches its peak expression in young carpels[4] where the central cell is developing and *FWA* is specifically expressed[8]. It reaches its second highest peak expression in pollen, a tissue in which it has no known functional role and thus its expression in this tissue likely serves to reinforce silencing in the sperm cells and ultimately vegetative tissues in a fashion analogous to transposon silencing[3]. Of the loci targeted by RdDM in Col-0, most are repressed in vegetative tissues and some have clear roles in pollen tube growth and elongation, such as RIC5 and PPME1[9], or seed development, such as the MADS box transcription factors (AGL23, 28, 34, 36, 42, 90, 92) and MEE15, 23, 38[1,10,11,12,13,14,15]. In fact, seven members of the ARF gene family (*ARF12, 14, 15, 20-23*) are targeted by the RdDM pathway and all seven are found specifically expressed in the endosperm and not detected in vegetative tissues[16]. Together, these data support a role for some RdDM-targeted loci in both pollen and endosperm development.

RNA silencing machinery is an evolutionarily conserved process found across eukaryotes, but the recent evolution of RNA polymerases, PolIV and PolV, is specific to flowering plants[17]. These two components have enabled the assembly of the RdDM pathway, which uses small RNAs to guide DNA methylation to target loci. This pathway targets transposon sequences to ensure their silenced state is maintained throughout vegetative tissues and the germline[3]. Animals also use small RNA directed DNA

methylation and heterochromatin formation mechanisms to maintain the epigenome of the germline through the use of Piwi-interacting RNAs[18,19,20]. In both plants and animals these small RNAs are derived from the genome of companion cells, which are terminal in nature and thus can afford widespread reactivation of gene expression of transposon and repeat sequences as they are not passed on to the next generation. Our study provides evidence that protein-coding genes have co-opted this transposon silencing mechanism to maintain their silenced state in vegetative tissues and transgenerationally as well as to ensure proper expression of genes important for pollen, seed, and germ line development.

References

1 Kinoshita, T. *et al.* One-way control of FWA imprinting in Arabidopsis endosperm by DNA methylation. *Science* **303**, 521-523 (2004).

2 Xiao, W. *et al.* Imprinting of the MEA Polycomb gene is controlled by antagonism between MET1 methyltransferase and DME glycosylase. *Dev. Cell* **5**, 891-901 (2003).

3 Slotkin, R. K. *et al.* Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell* **136**, 461-472 (2009).

4 Schmid, M. *et al.* A gene expression map of Arabidopsis thaliana development. *Nature Genetics* (2005).

5 Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).

6 Soppe, W. J. *et al.* The late flowering phenotype of fwa mutants is caused by gain-of-function epigenetic alleles of a homeodomain gene. *Mol. Cell* **6**, 791-802 (2000).

7 Yoo, S. K., Lee, J. S. & Ahn, J. H. Overexpression of AGAMOUS-LIKE 28 (AGL28) promotes flowering by upregulating expression of floral promoters within the autonomous pathway. *Biochem Biophys Res Commun* **348**, 929-936 (2006).

8 Kinoshita, T. *et al.* One-way control of FWA imprinting in Arabidopsis endosperm by DNA methylation. *Science* **303**, 521-523 (2003).

9 Tian, G. W., Chen, M. H., Zaltsman, A. & Citovsky, V. Pollen-specific pectin methylesterase involved in pollen tube growth. *Dev Biol* **294**, 83-91 (2006).

10 Colombo, M. *et al.* AGL23, a type I MADS-box gene that controls female gametophyte and embryo development in Arabidopsis. *Plant J* **54**, 1037-1048 (2008).

11 Dubreucq, B. *et al.* The Arabidopsis AtEPR1 extensin-like gene is specifically expressed in endosperm during seed germination. *Plant J* **23**, 643-652 (2000).

12 Nishimura, N. *et al.* ABA hypersensitive germination2-1 causes the activation of both abscisic acid and salicylic acid responses in Arabidopsis. *Plant Cell Physiol* **50**, 2112-2122 (2009).

13 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).

14 Shirzadi, R. *et al.* Genome-wide transcript profiling of endosperm without paternal contribution identifies parent-of-origin-dependent regulation of AGAMOUS-LIKE36. *PLoS Genet* **7**, e1001303 (2011).

15 You, W. *et al.* Atypical DNA methylation of genes encoding cysteine-rich peptides in Arabidopsis thaliana. *BMC Plant Biol* **12**, 51 (2012).

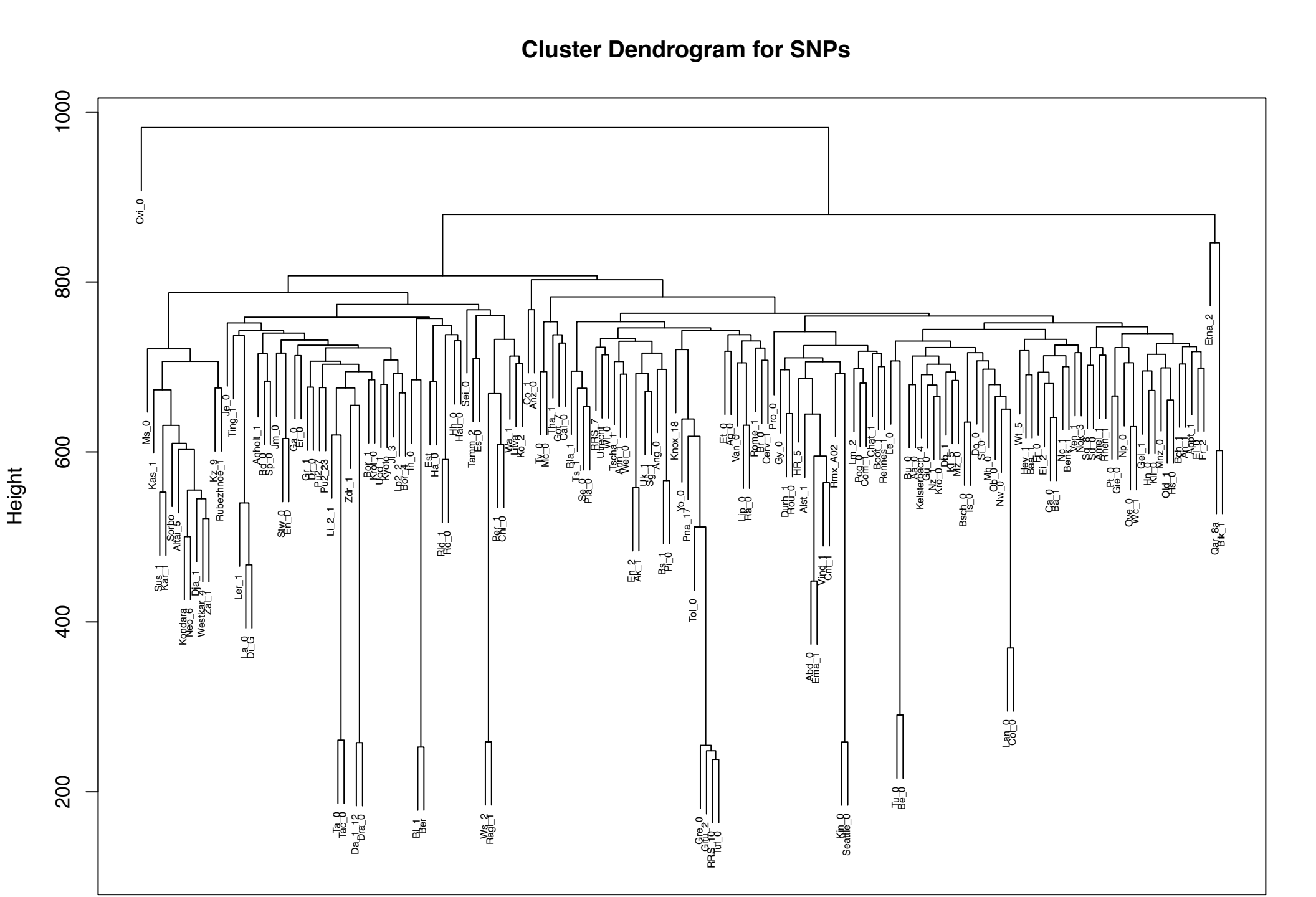
16 Rademacher, E. H. *et al.* A cellular expression map of the Arabidopsis AUXIN RESPONSE FACTOR gene family. *Plant J* **68**, 597-606 (2011).

17 Luo, J. & Hall, B. D. A multistep process gave rise to RNA polymerase IV of land plants. *J Mol Evol* **64**, 101-112 (2007).

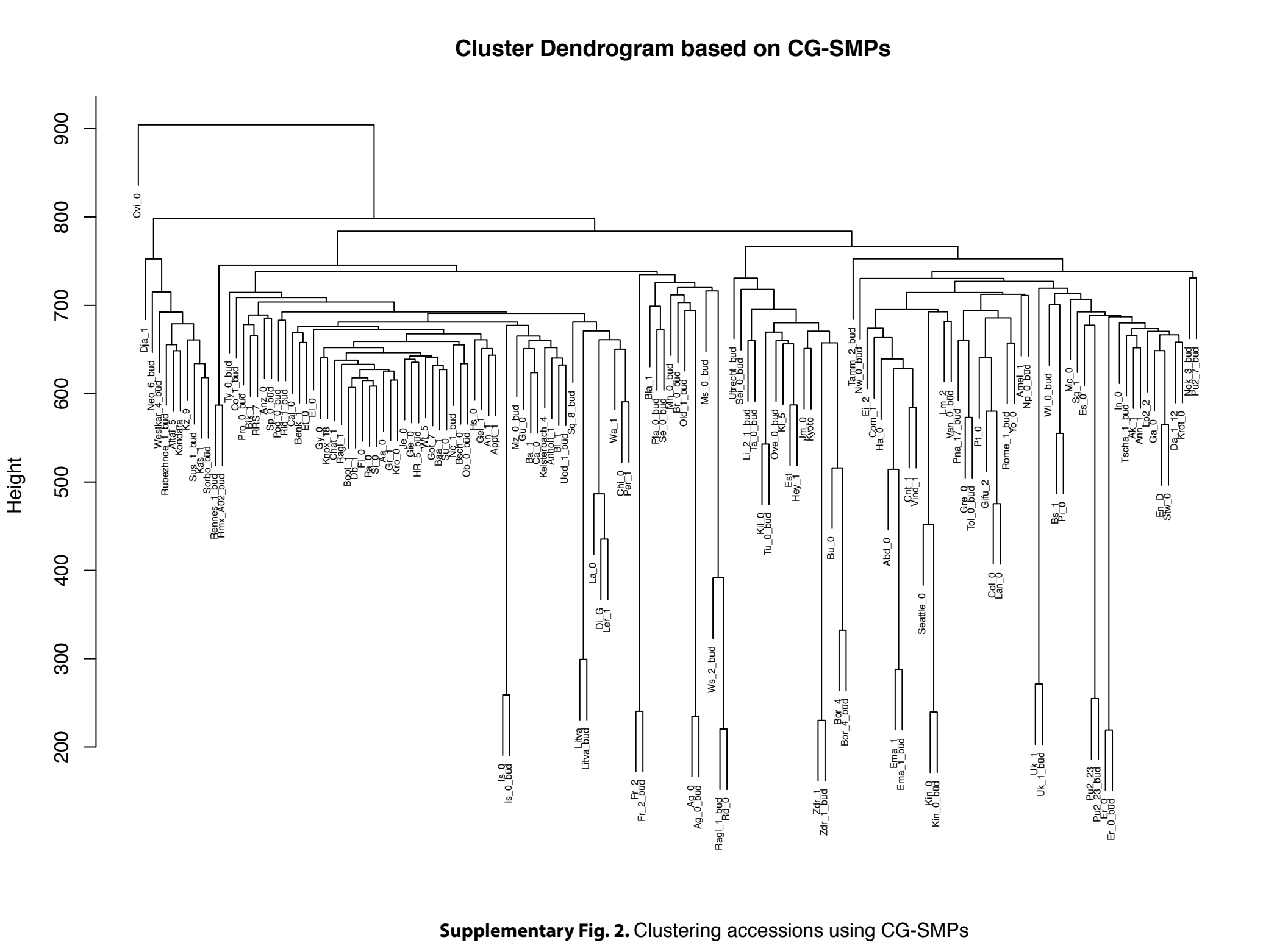
18 Carmell, M. A. *et al.* MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. *Dev Cell* **12**, 503-514 (2007).

19 Vagin, V. V. *et al.* A distinct small RNA pathway silences selfish genetic elements in the germline. *Science* **313**, 320-324 (2006).

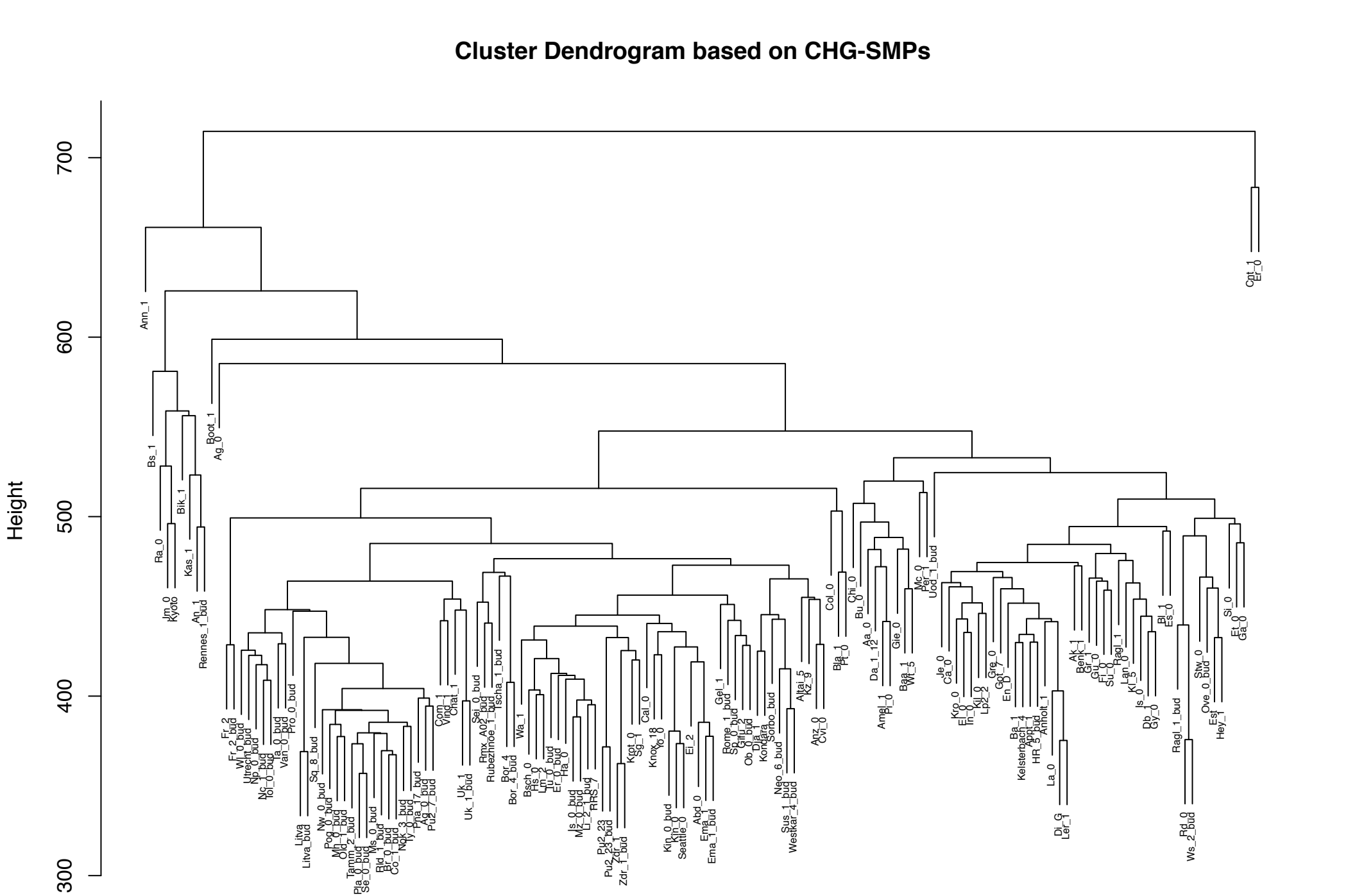
20 Watanabe, T. *et al.* Role for piRNAs and noncoding RNA in de novo DNA methylation of the imprinted mouse Rasgrf1 locus. *Science* **332**, 848-852 (2011).



Supplementary Fig. 1. Clustering accessions using SNPs

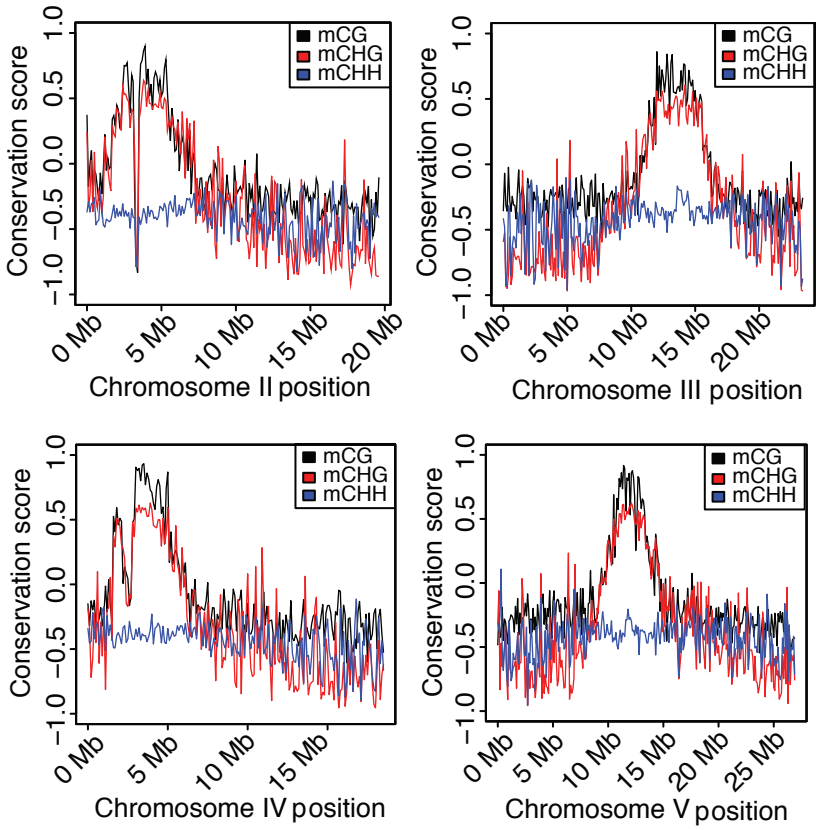


Supplementary Fig. 2. Clustering accessions using CG-SMPs

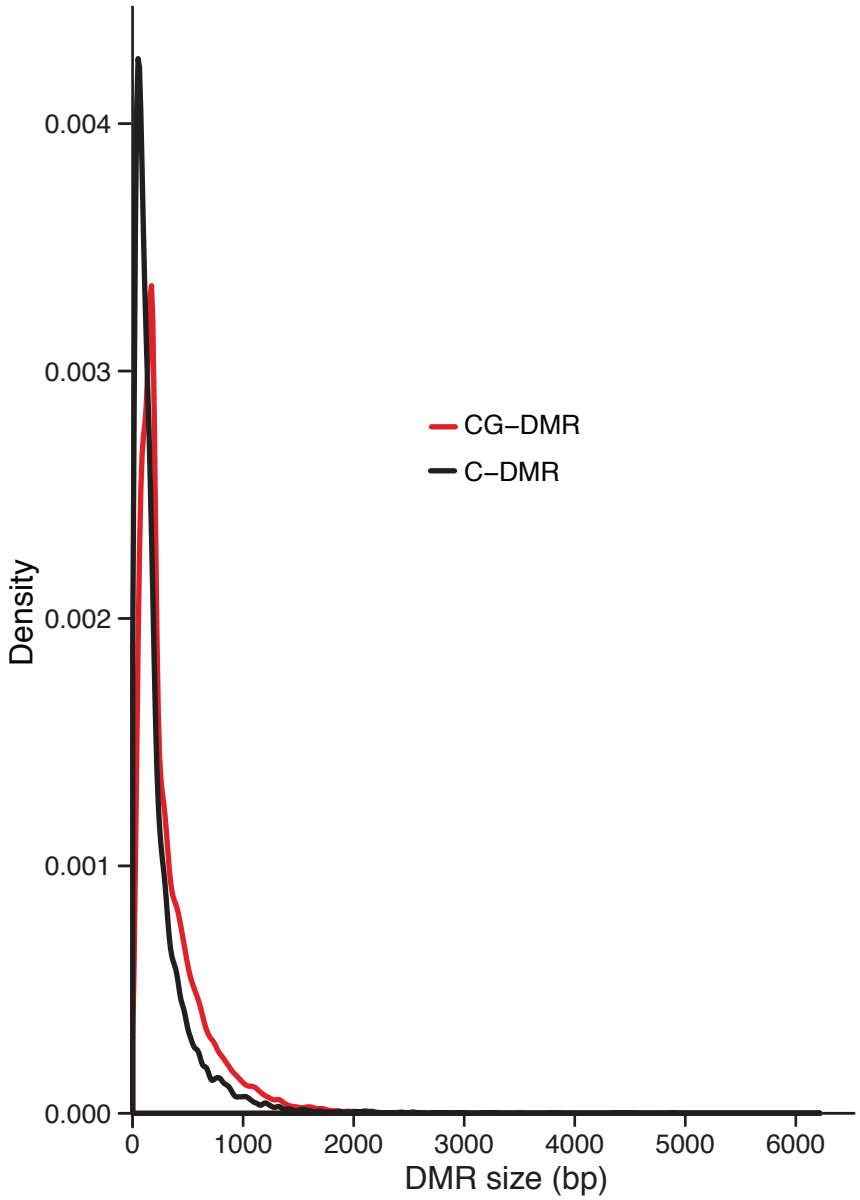


Supplementary Fig. 3. Clustering accessions using CHG-SMPs.

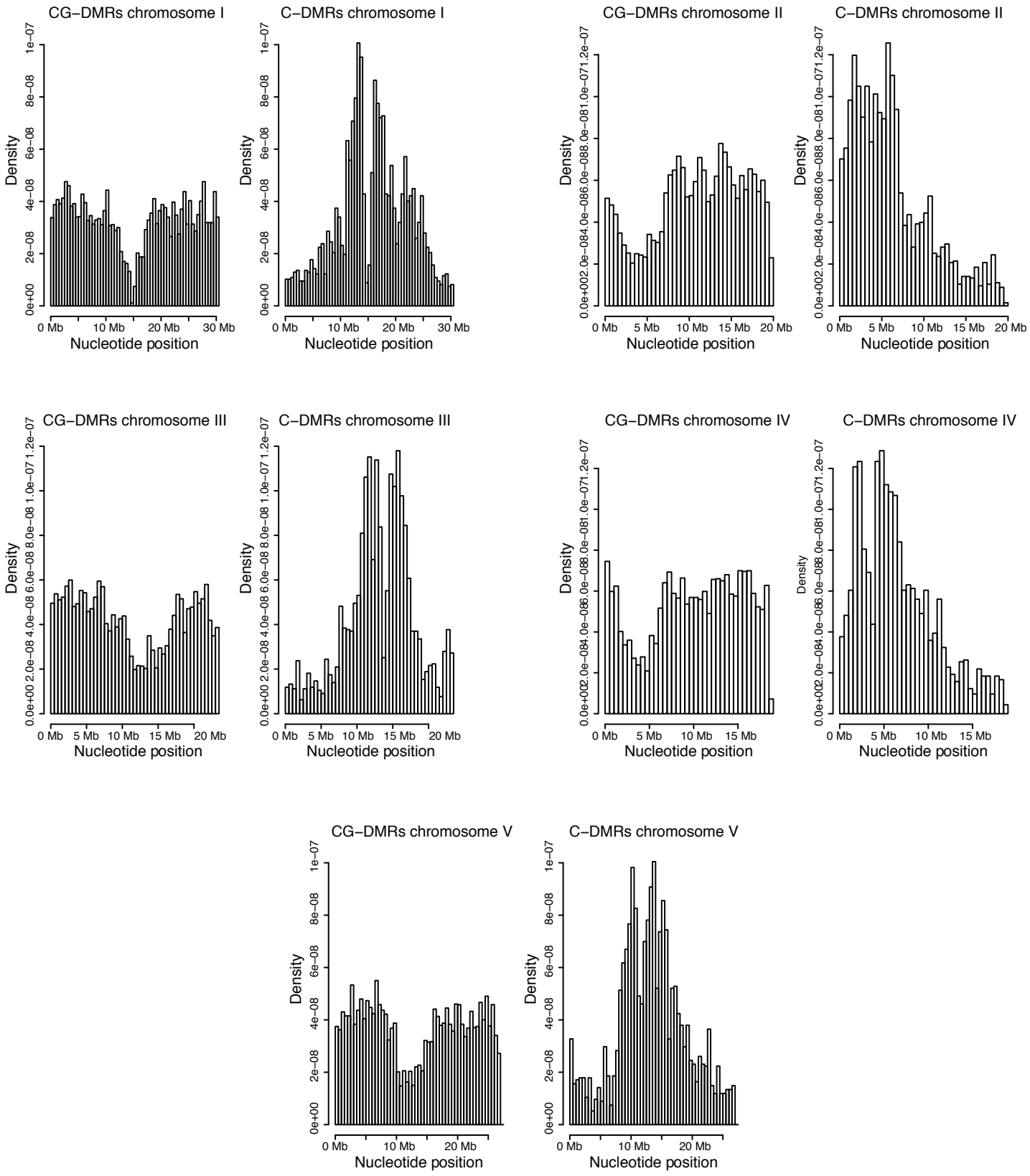
Supplementary Fig. 4. Clustering accessions using CHH-SMPs



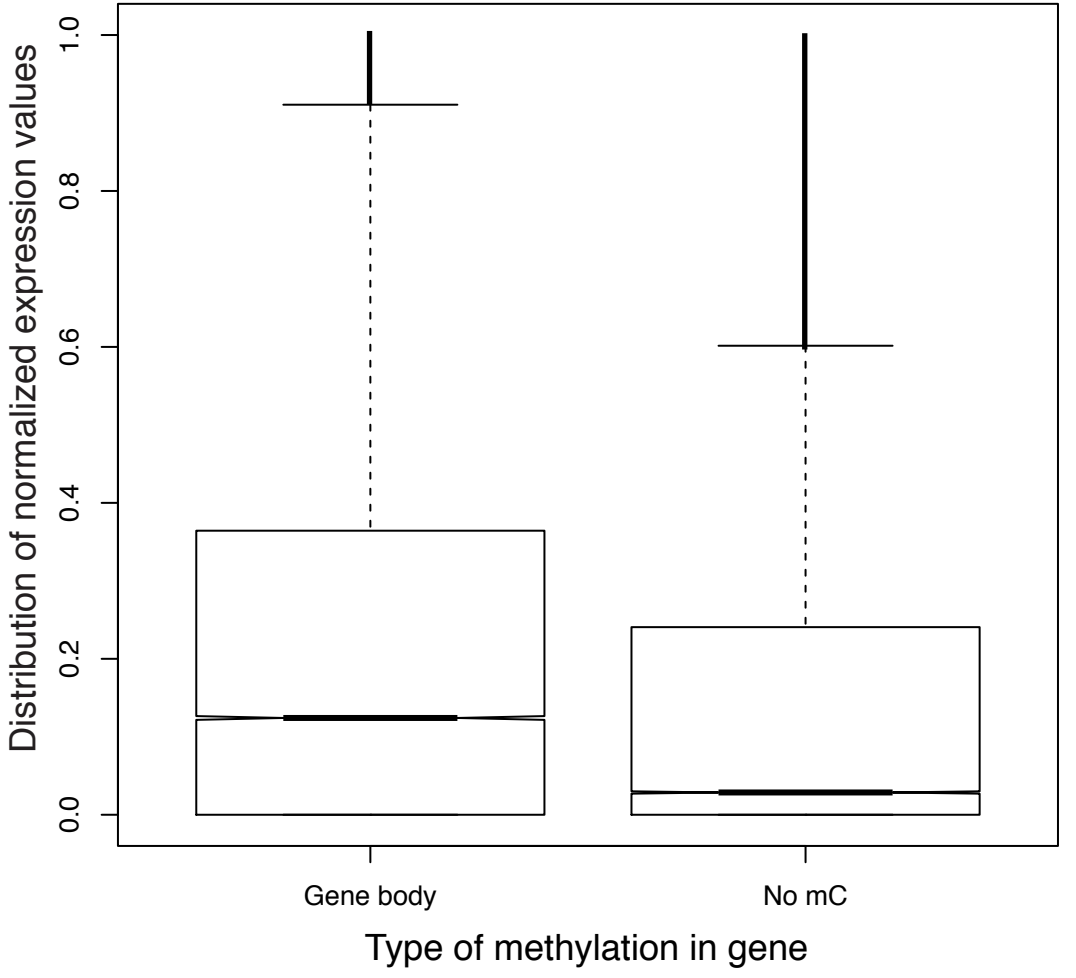
Supplementary Fig. 5. Plots of the genome-wide distribution of methylation conservation across chromosomes II, III, IV, V. A methylation state conservation score was calculated across all SMPs in each nucleotide context (CG, CHG, and CHH) and averaged together into 100kb windows. A structural variant is visible on the top arm of chromosome II and in the centromere of chromosome IV.



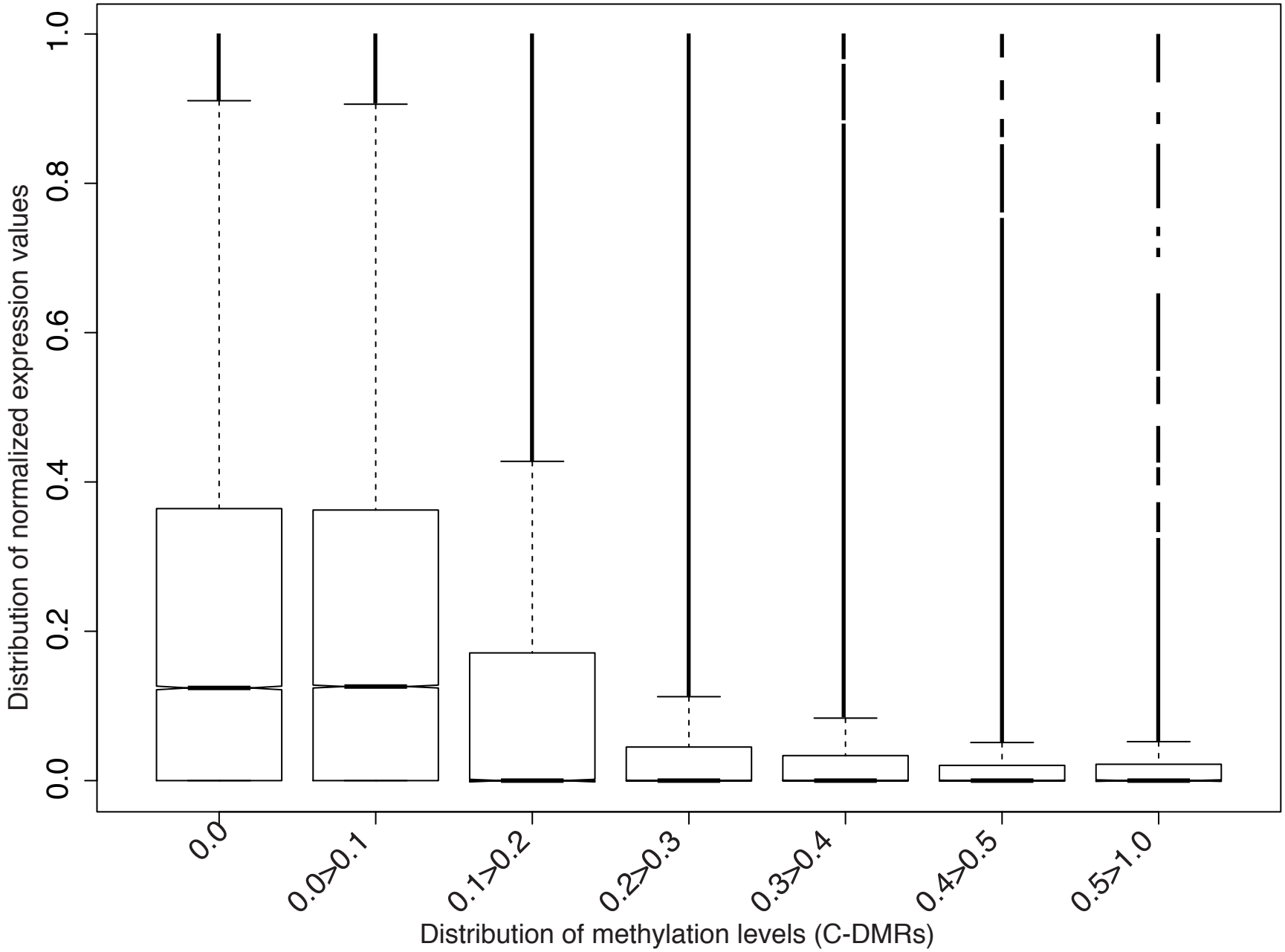
Supplementary Fig. 6. Size distribution of CG-DMRs (red line) and C-DMRs (black line).



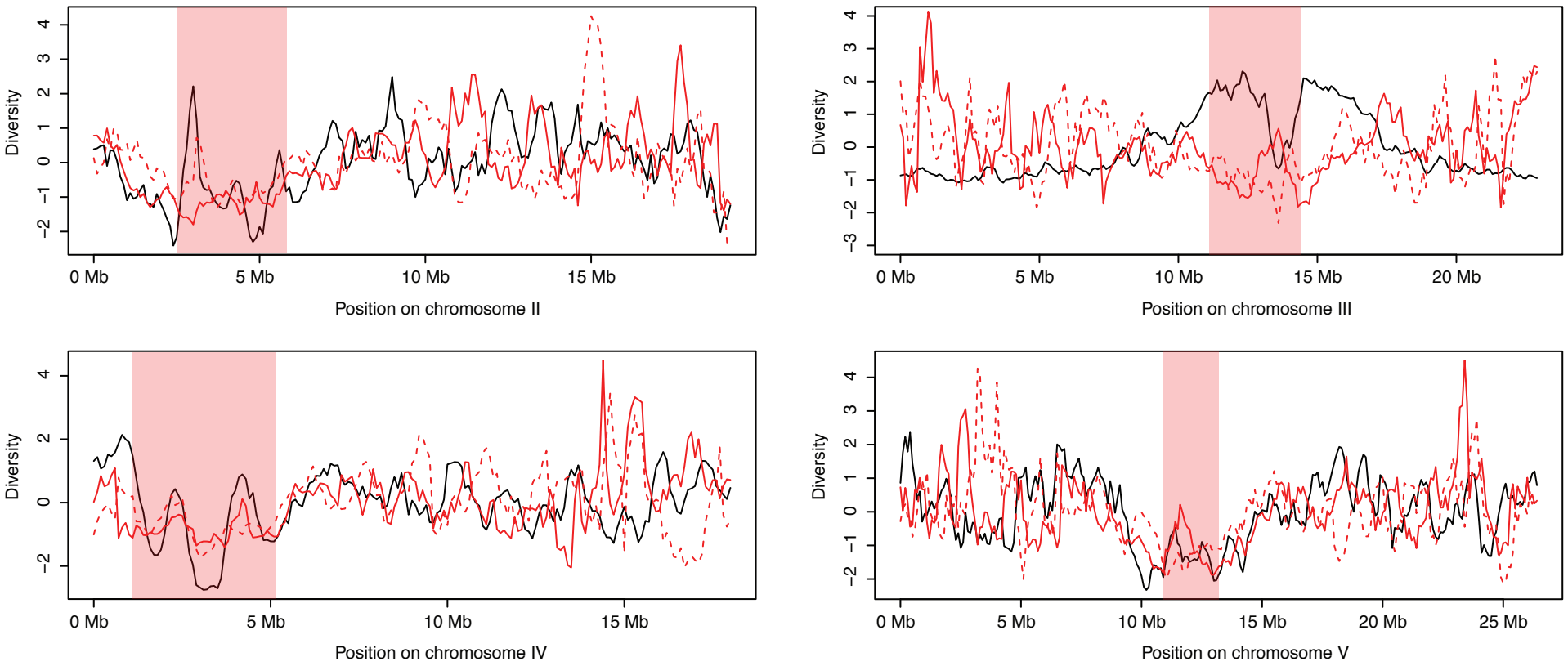
Supplementary Fig. 7. Genome-wide distribution of CG-DMRs and C-DMRs. CG-DMRs are enriched along the euchromatic gene-rich regions, whereas C-DMRs are enriched within the pericentromeric gene-poor, transposon-rich regions of the genome.



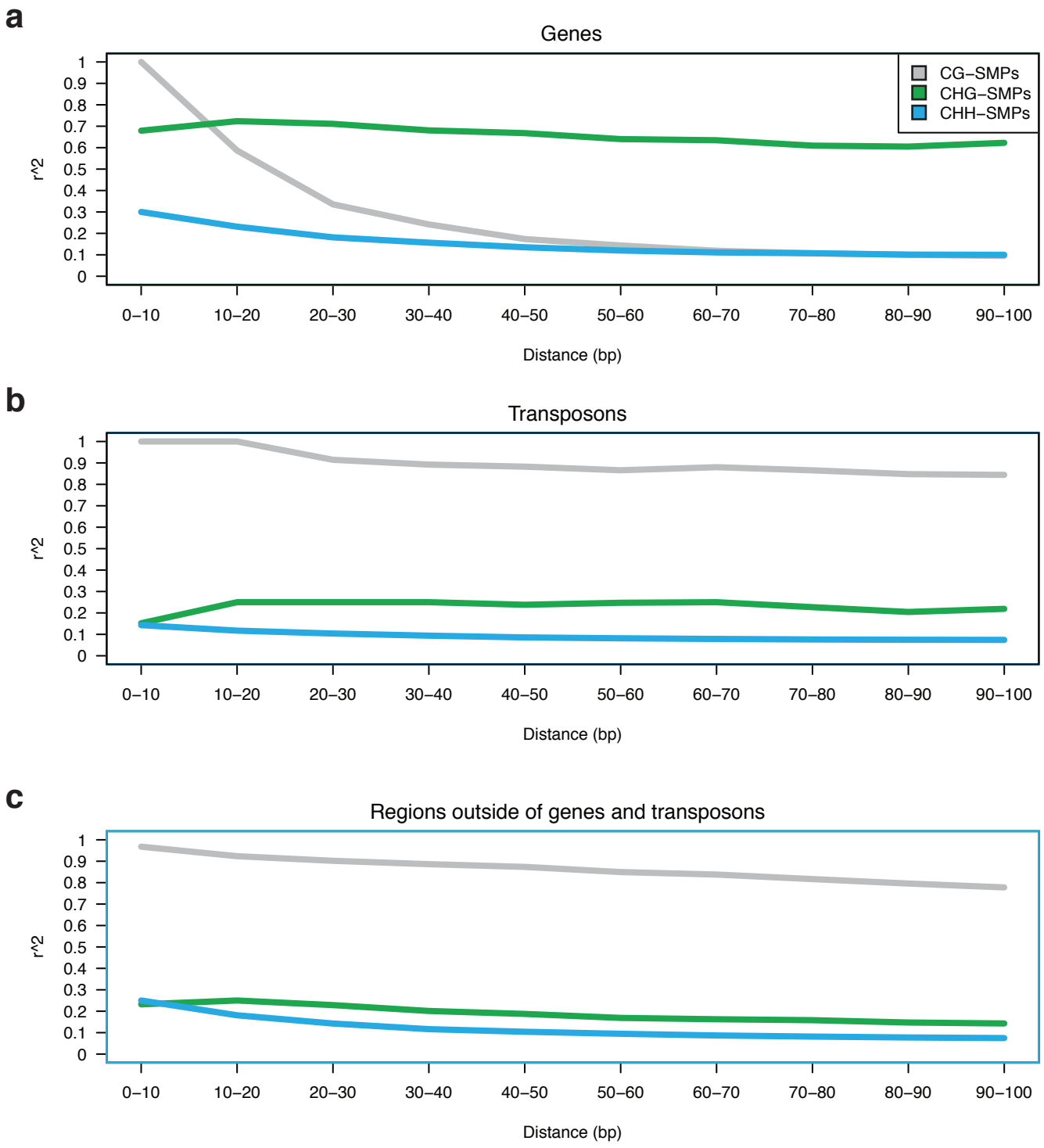
Supplementary Fig. 8. Distributions of the expression levels of genes with gene body methylation and those without methylation entirely. Gene body methylated genes in general are more highly expressed.



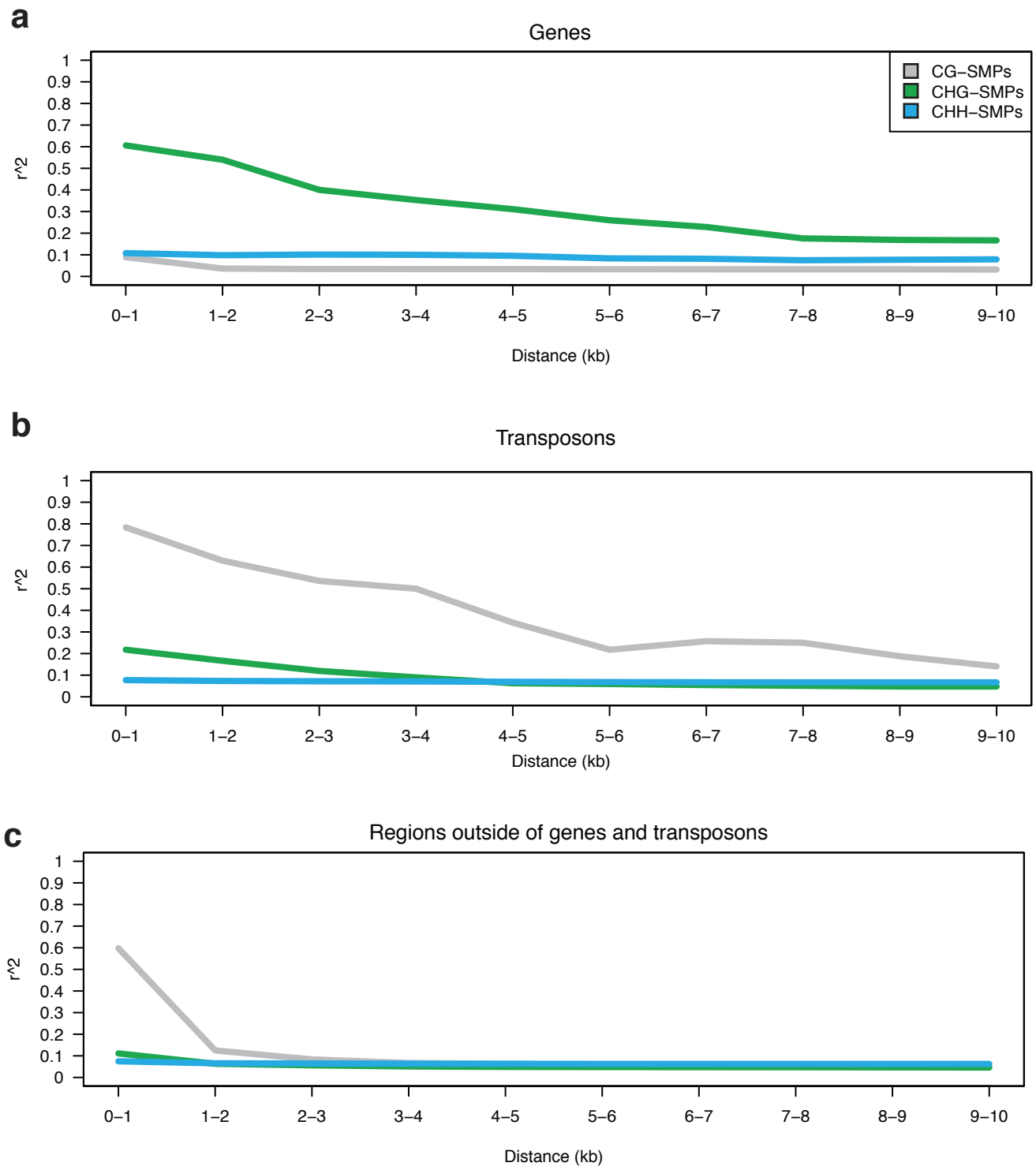
Supplementary Fig. 9. Boxplot representation of expression values for genes containing C-DMRs. In this plot unlike in Fig. 2l, the 0.0 category is comprised of genes with no nonCG methylation as genes with no methylation tend to be less expressed than those with gene body methylation.



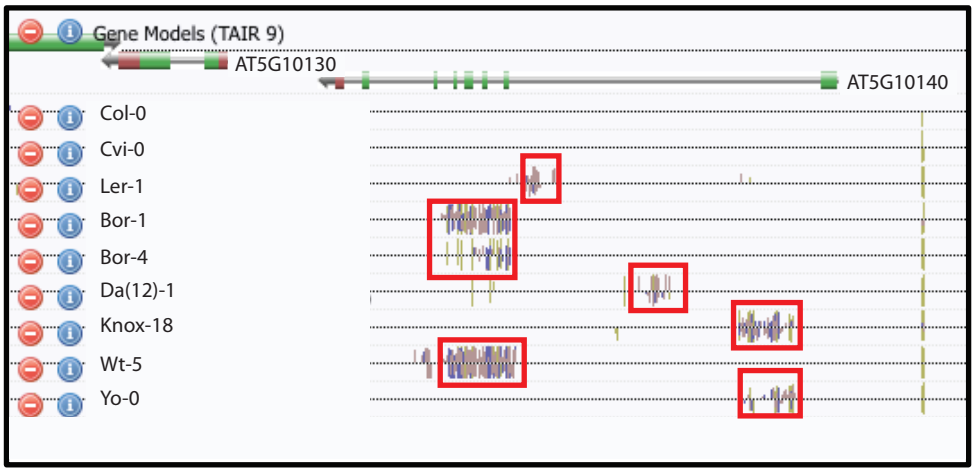
Supplementary Fig. 10. A plot of SNP, SMP, and C-DMR diversity across chromosomes II, III, IV, and V. The diversity for each data type was computed by calculating the Euclidean distances between the C-DMRs/SNPs/SMPs of all pairwise combinations of accessions in overlapping 500 kb windows offset by 100 kb (binary values for SNPs and SMPs, methylation levels for C-DMRs). Each diversity measure was normalized to have a mean of zero and a standard deviation of one. Pink shaded regions indicate the location of the pericentromere.



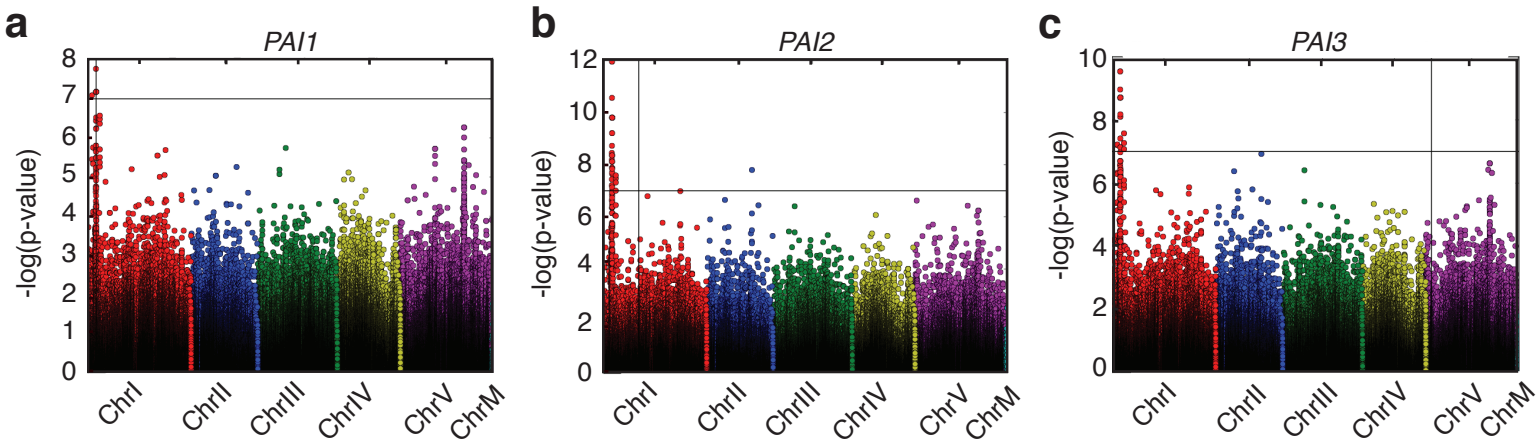
Supplementary Fig. 11. Assocation decay rates for CG-, CHG- and CHH-SMPs across genomic features. (a) The association decay rate for SMPs in genes. (b) the association decay rate of SMPs in transposons. (c) the association decay rate of SMPs in regions not containing genes or transposons. The rates of decay outside of genes for CG-SMPs indicate that these SMPs have different properties compared to CG-SMPs in transposons, which could be due to other factors such as nucleosome density, an important factor for maintenance of CG methylation.



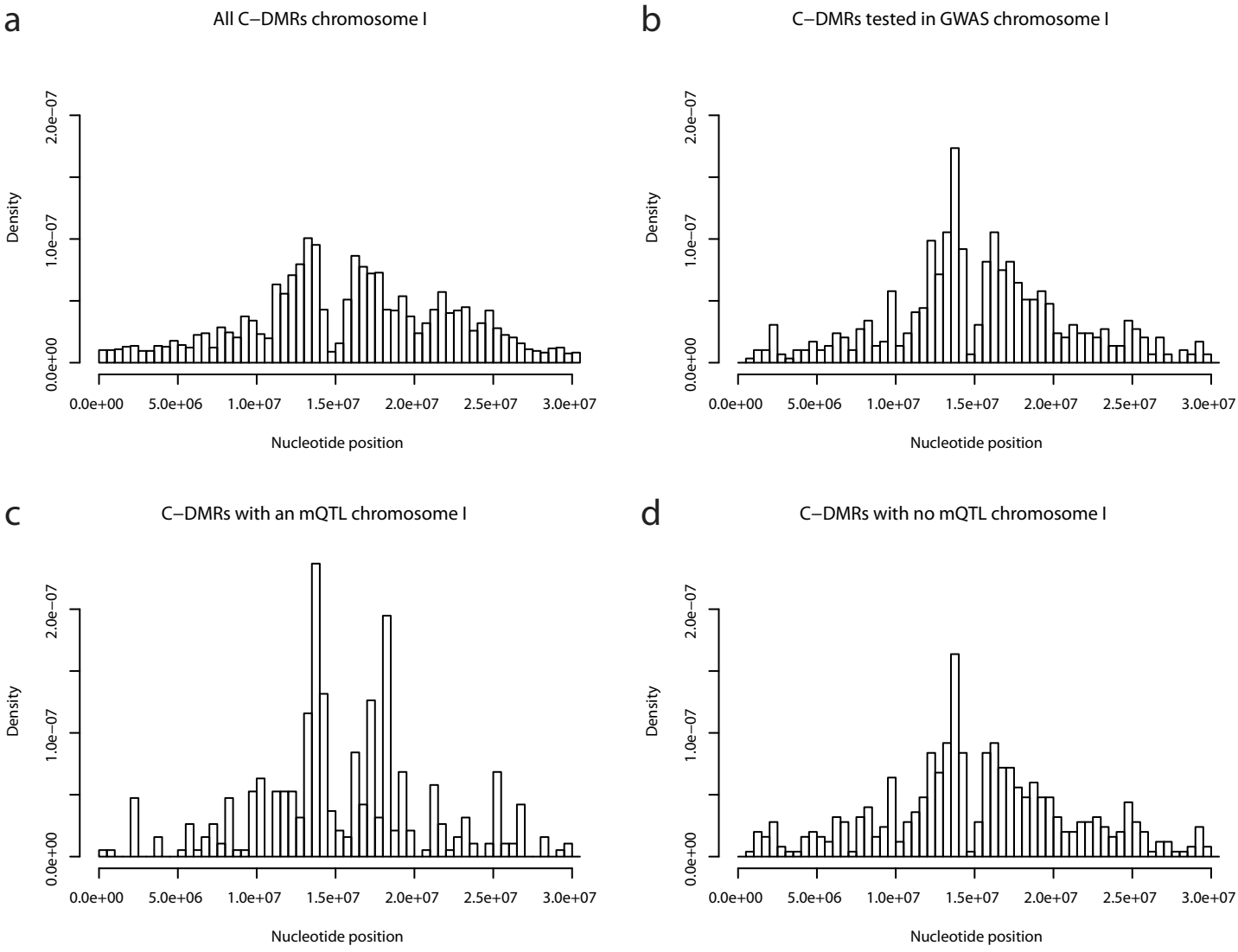
Supplementary Fig. 12. Assocation decay rates for CG-, CHG- and CHH-SMPs across genomic features over long ranges. (a) The association decay rate for SMPs in genes. (b) the association decay rate of SMPs in transposons. (c) the association decay rate of SMPs in regions not containing genes or transposons. The rate of decay outside of genes for CHG-SMPs and for CG-SMPs in transposons eventually decreases, but the distance of decay is far greater than the linkage disequilibrium deteremined by using SNPs from these same accessions. CG-SMPs in genes decay at rates faster than LD and CHG-SMPs in genes and CG-SMPs in transposons decay at rates greater than LD indicating that SMPs can be unliked from genotype.



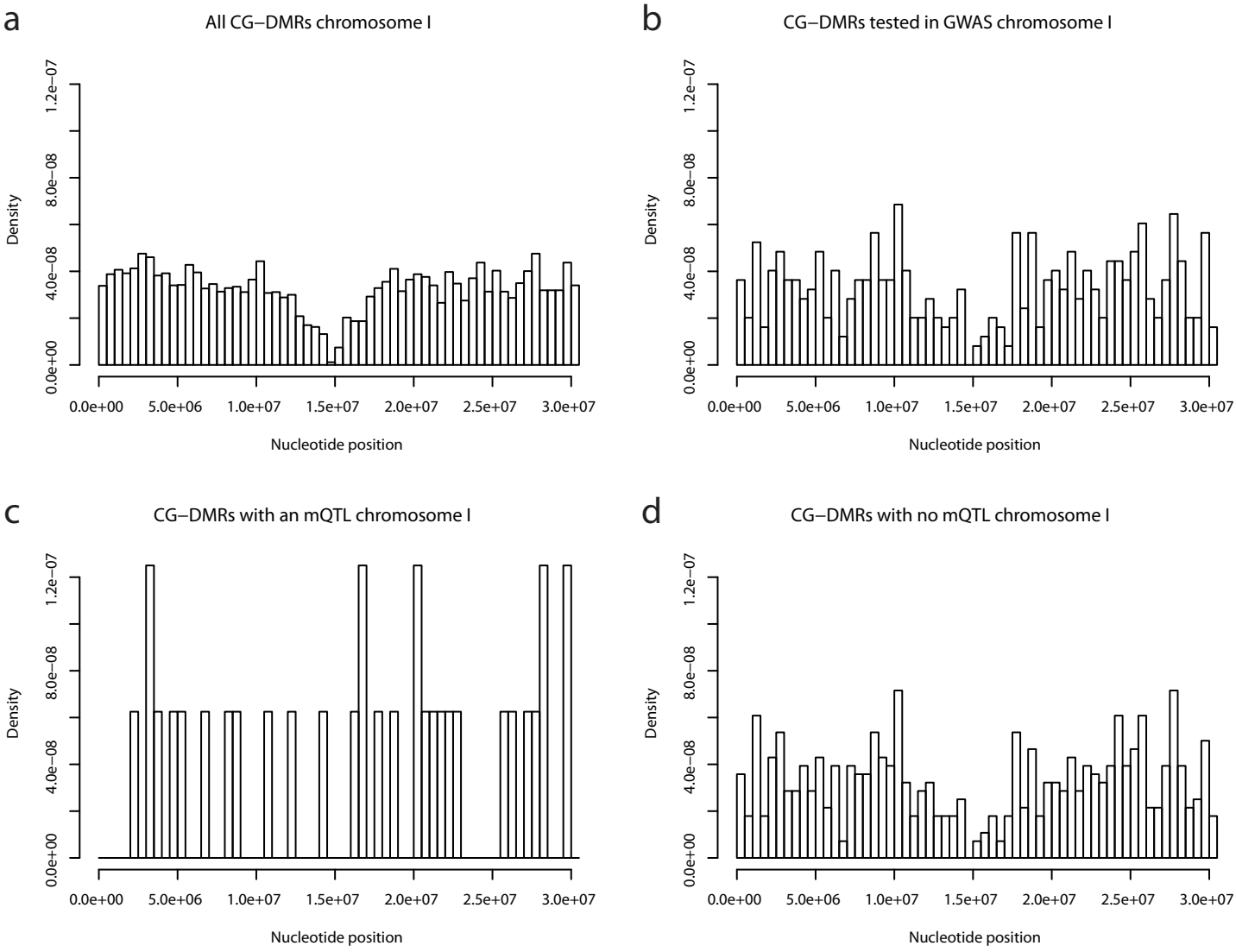
Supplementary Fig. 13. A selected screen shot of C-DMRs (red boxes) that overlap with the *FLC* locus. Within some of these C-DMRs there is a transposition event that has occurred and the methylation has spread from the transposon into the inserted locus.



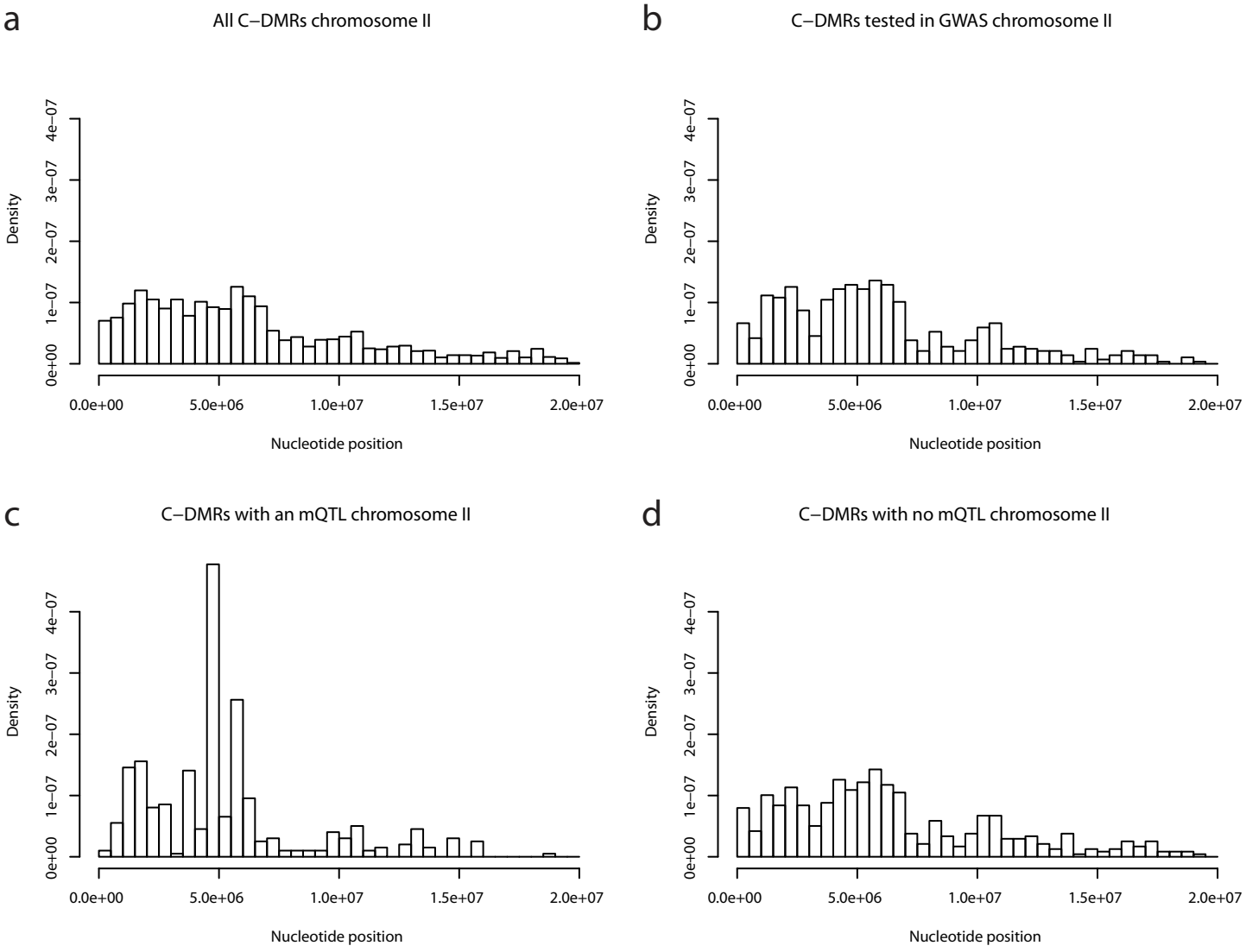
Supplementary Fig. 14. Manhattan plots for the PAI-gene family. Each point represents an association test between a particular C-DMR, the location of which is indicated by a vertical line, and a SNP. The y-axis indicates the level of significance, which increases from the bottom to the top of the plot, and the x-axis denotes the position of the SNP being tested. An association test is deemed significant if the dot is above the horizontal line, which indicates a 1% FDR.



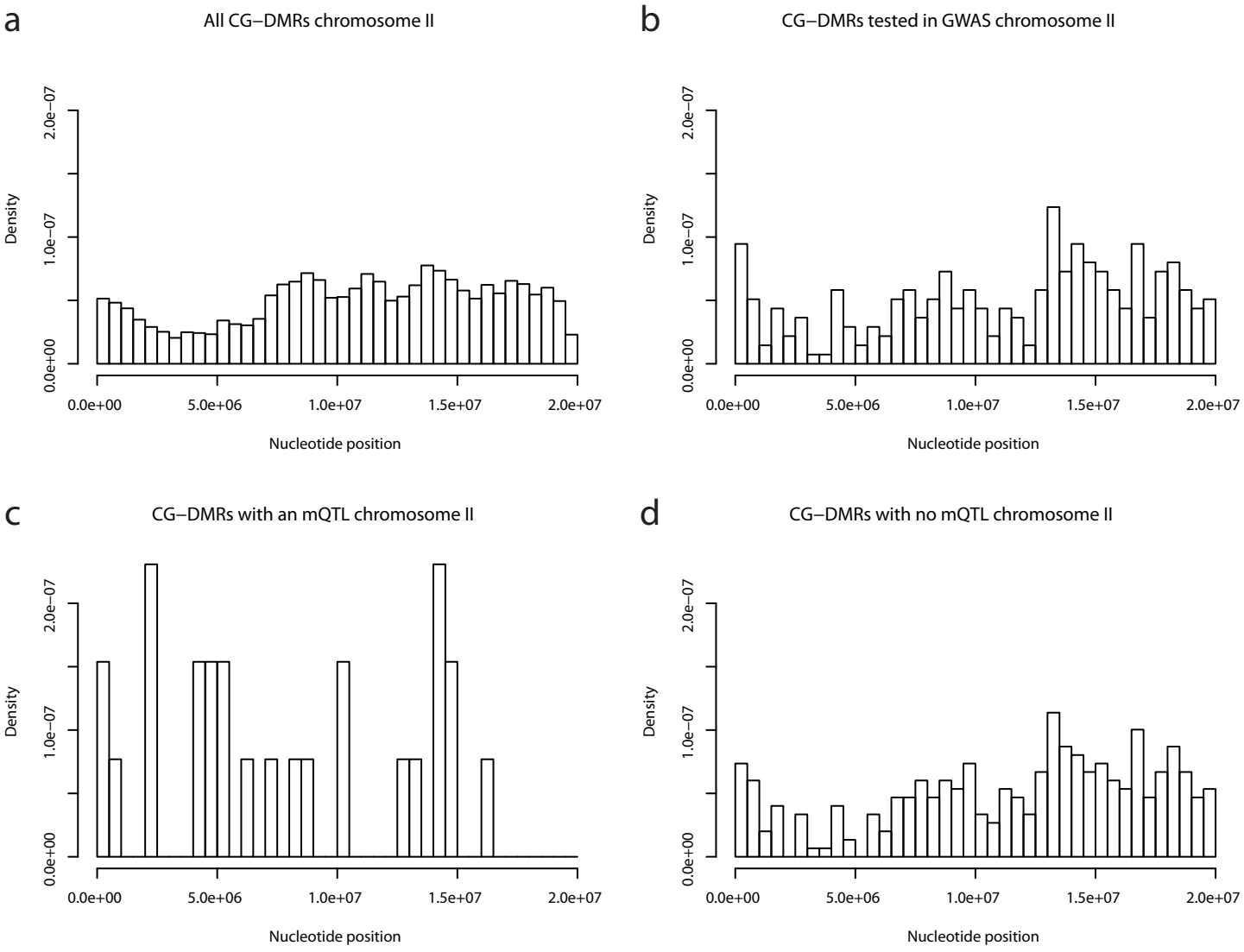
Supplementary Fig. 15. Chromosome I mQTL anlaysis. (a) a histogram representing the distribution of C-DMRs along the chromosome. (b) a histogram representing the distribution of C-DMRs that passed the required thresholds for association mapping.along the chromosome. (c) a histogram representing the distribution of C-DMRs with an mQTL along the chromosome. (d) a histogram representing the distribution of C-DMRs with no significantly associated mQTL along the chromosome.



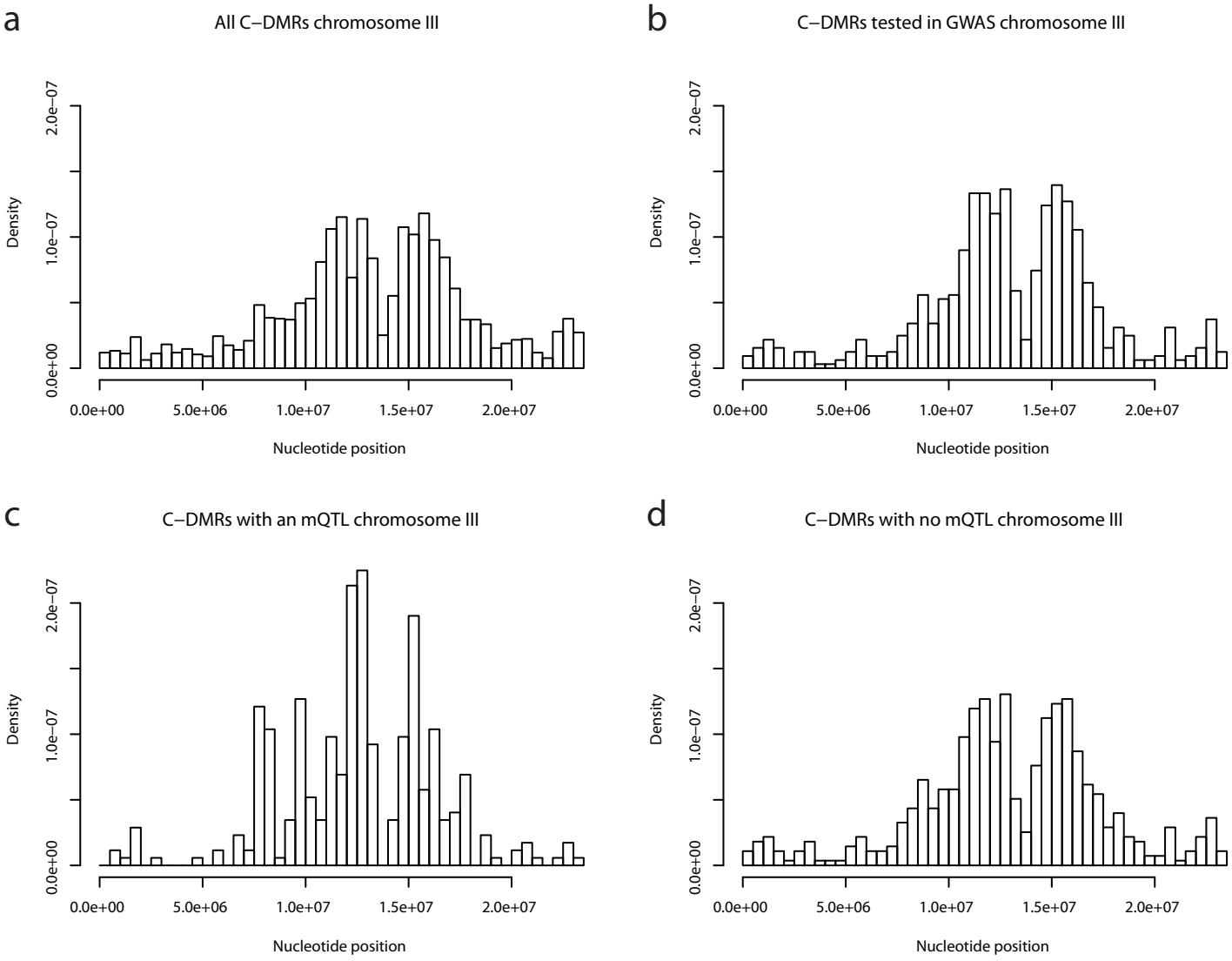
Supplementary Fig. 16. Chromosome I mQTL anlaysis. (a) a histogram representing the distribution of CG-DMRs along the chromosome. (b) a histogram representing the distribution of CG-DMRs that passed the required thresholds for association mapping.along the chromosome. (c) a histogram representing the distribution of CG-DMRs with an mQTL along the chromosome. (d) a histogram representing the distribution of CG-DMRs with no significantly associated mQTL along the chromosome.



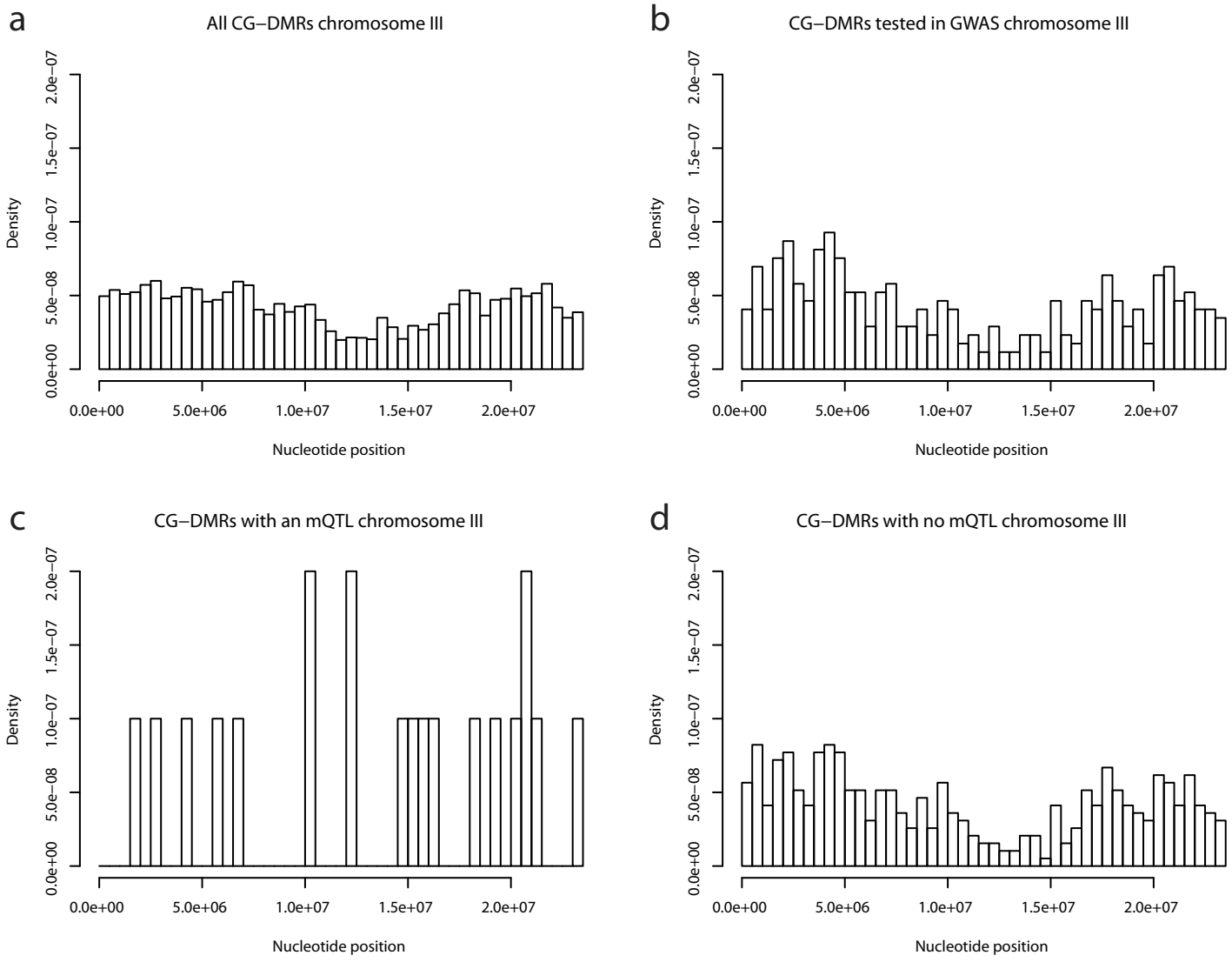
Supplementary Fig. 17. Chromosome II mQTL anlaysis. (a) a histogram representing the distribution of C-DMRs along the chromosome. (b) a histogram representing the distribution of C-DMRs that passed the required thresholds for association mapping.along the chromosome. (c) a histogram representing the distribution of C-DMRs with an mQTL along the chromosome. (d) a histogram representing the distribution of C-DMRs with no significantly associated mQTL along the chromosome.



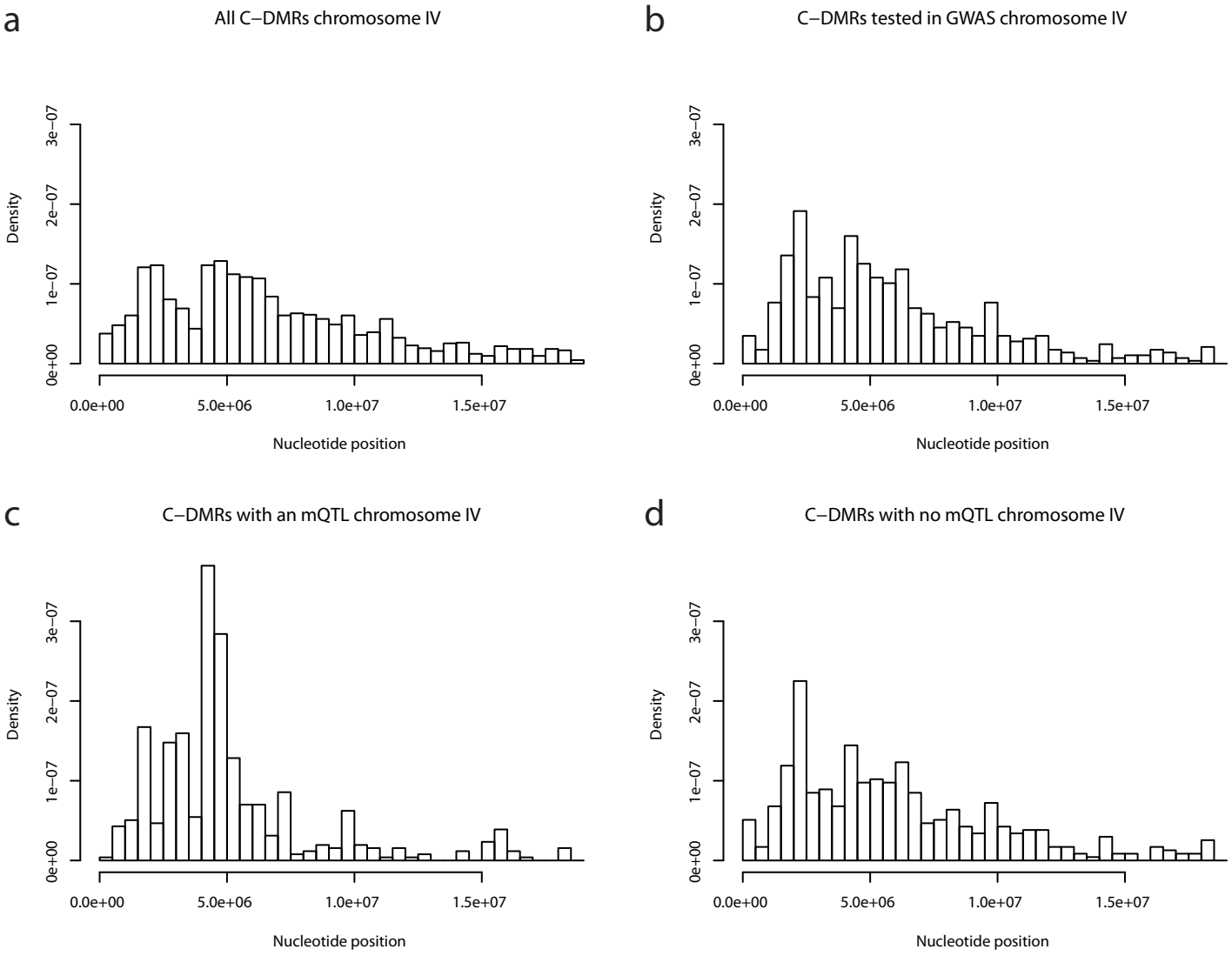
Supplementary Fig. 18. Chromosome II mQTL anlaysis. (a) a histogram representing the distribution of CG-DMRs along the chromosome. (b) a histogram representing the distribution of CG-DMRs that passed the required thresholds for association mapping.along the chromosome. (c) a histogram representing the distribution of CG-DMRs with an mQTL along the chromosome. (d) a histogram representing the distribution of CG-DMRs with no significantly associated mQTL along the chromosome.



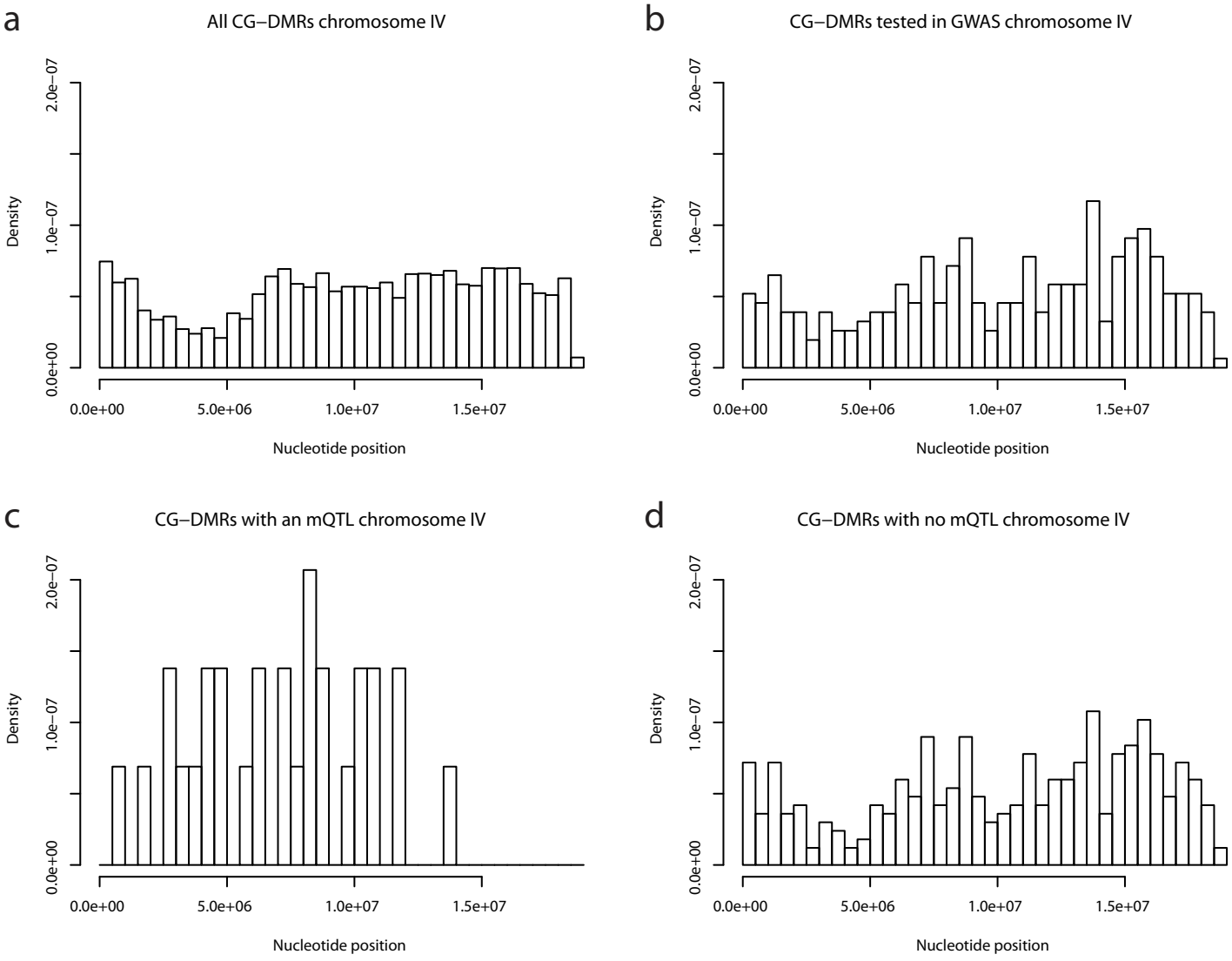
Supplementary Fig. 19. Chromosome III mQTL anlaysis. (a) a histogram representing the distribution of C-DMRs along the chromosome. (b) a histogram representing the distribution of C-DMRs that passed the required thresholds for association mapping.along the chromosome. (c) a histogram representing the distribution of C-DMRs with an mQTL along the chromosome. (d) a histogram representing the distribution of C-DMRs with no significantly associated mQTL along the chromosome.



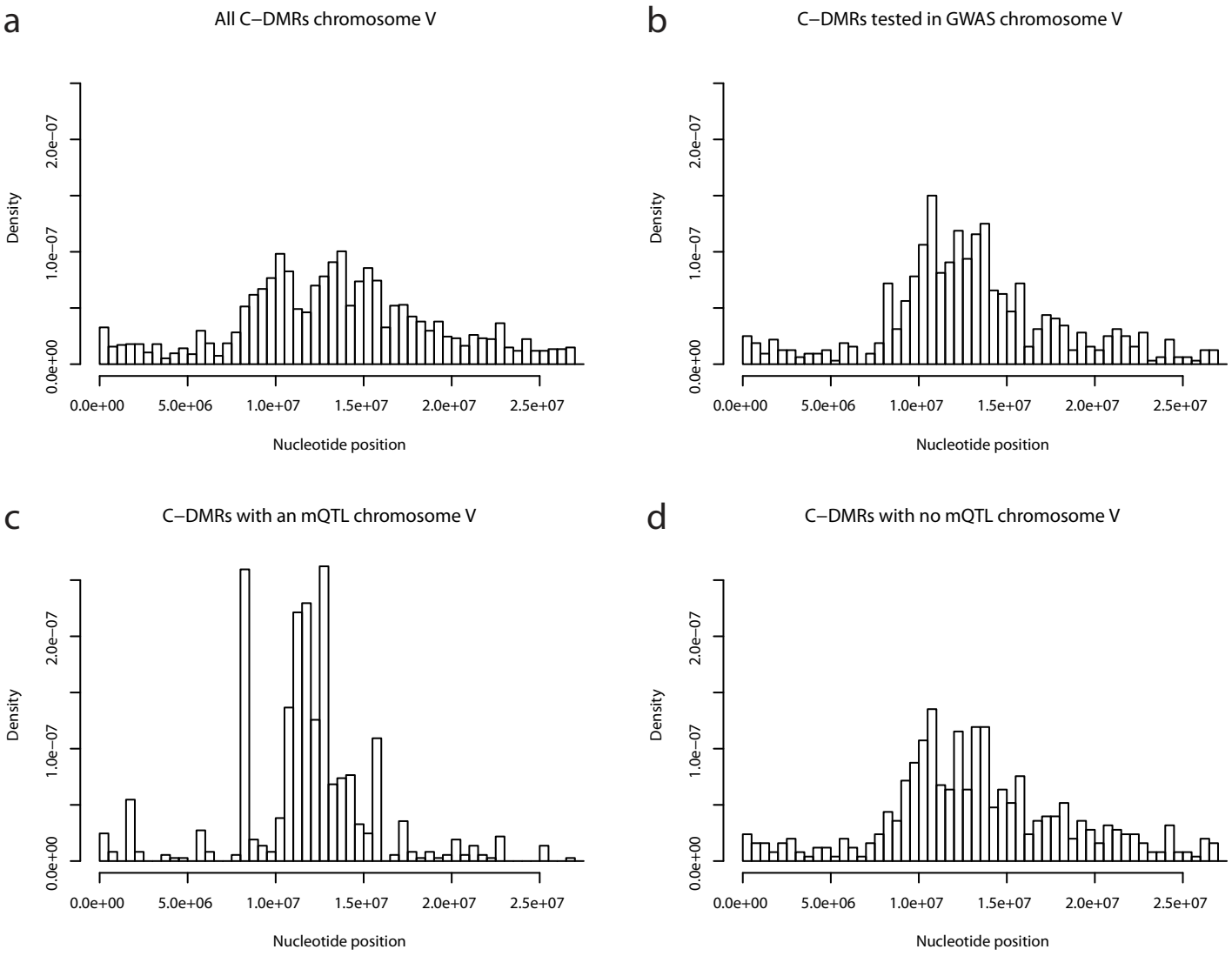
Supplementary Fig. 20. Chromosome III mQTL anlaysis. (a) a histogram representing the distribution of CG-DMRs along the chromosome. (b) a histogram representing the distribution of CG-DMRs that passed the required thresholds for association mapping.along the chromosome. (c) a histogram representing the distribution of CG-DMRs with an mQTL along the chromosome. (d) a histogram representing the distribution of CG-DMRs with no significantly associated mQTL along the chromosome.



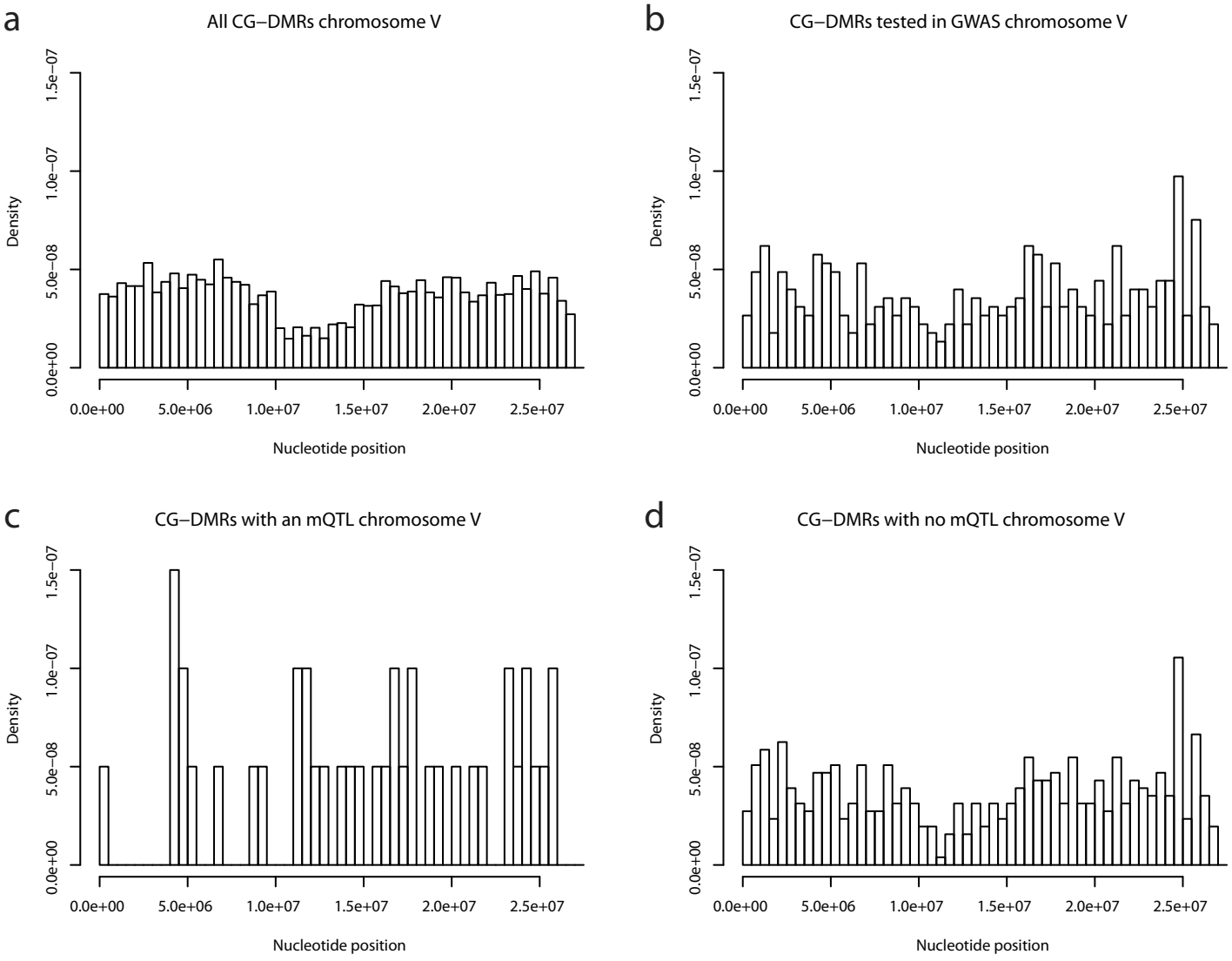
Supplementary Fig. 21. Chromosome IV mQTL anlaysis. (a) a histogram representing the distribution of C-DMRs along the chromosome. (b) a histogram representing the distribution of C-DMRs that passed the required thresholds for association mapping.along the chromosome. (c) a histogram representing the distribution of C-DMRs with an mQTL along the chromosome. (d) a histogram representing the distribution of C-DMRs with no significantly associated mQTL along the chromosome.



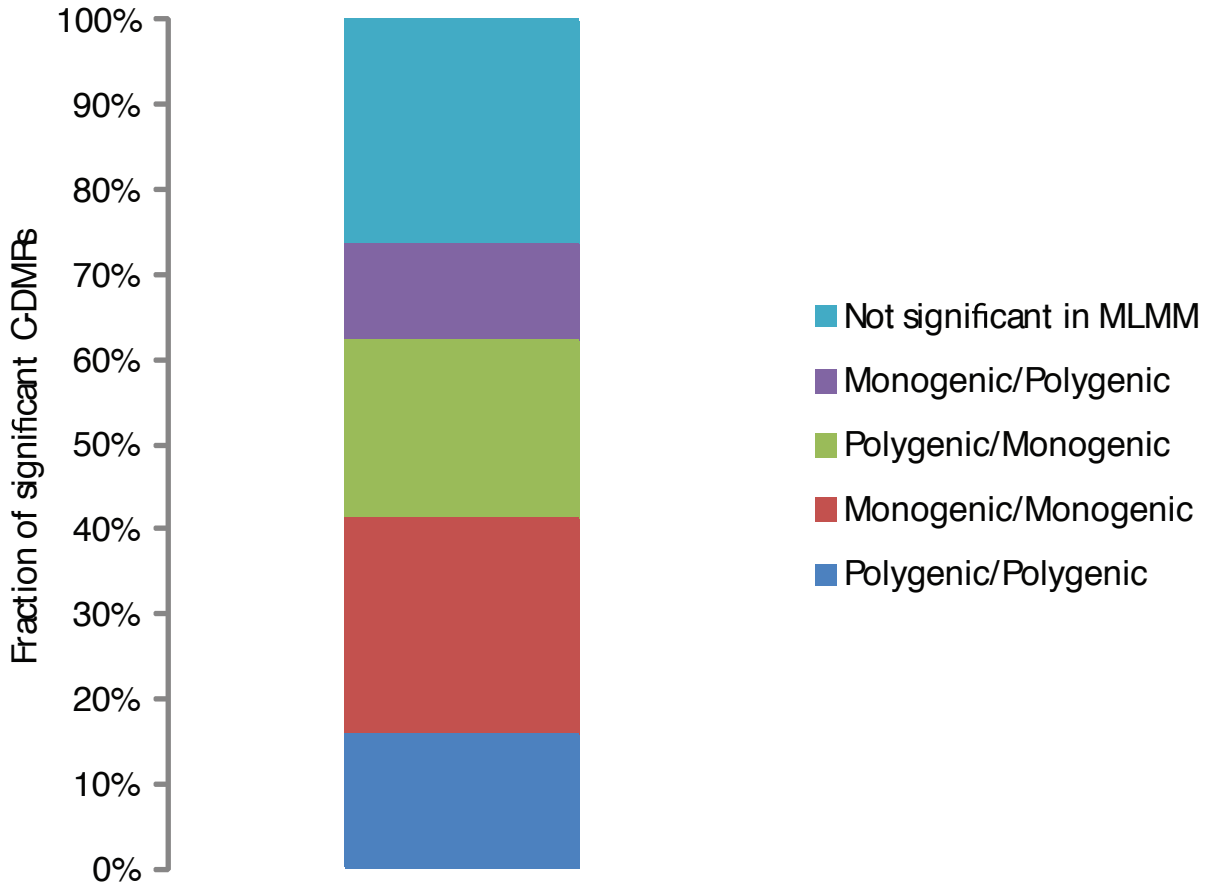
Supplementary Fig. 22. Chromosome IV mQTL anlaysis. (a) a histogram representing the distribution of CG-DMRs along the chromosome. (b) a histogram representing the distribution of CG-DMRs that passed the required thresholds for association mapping.along the chromosome. (c) a histogram representing the distribution of CG-DMRs with an mQTL along the chromosome. (d) a histogram representing the distribution of CG-DMRs with no significantly associated mQTL along the chromosome.



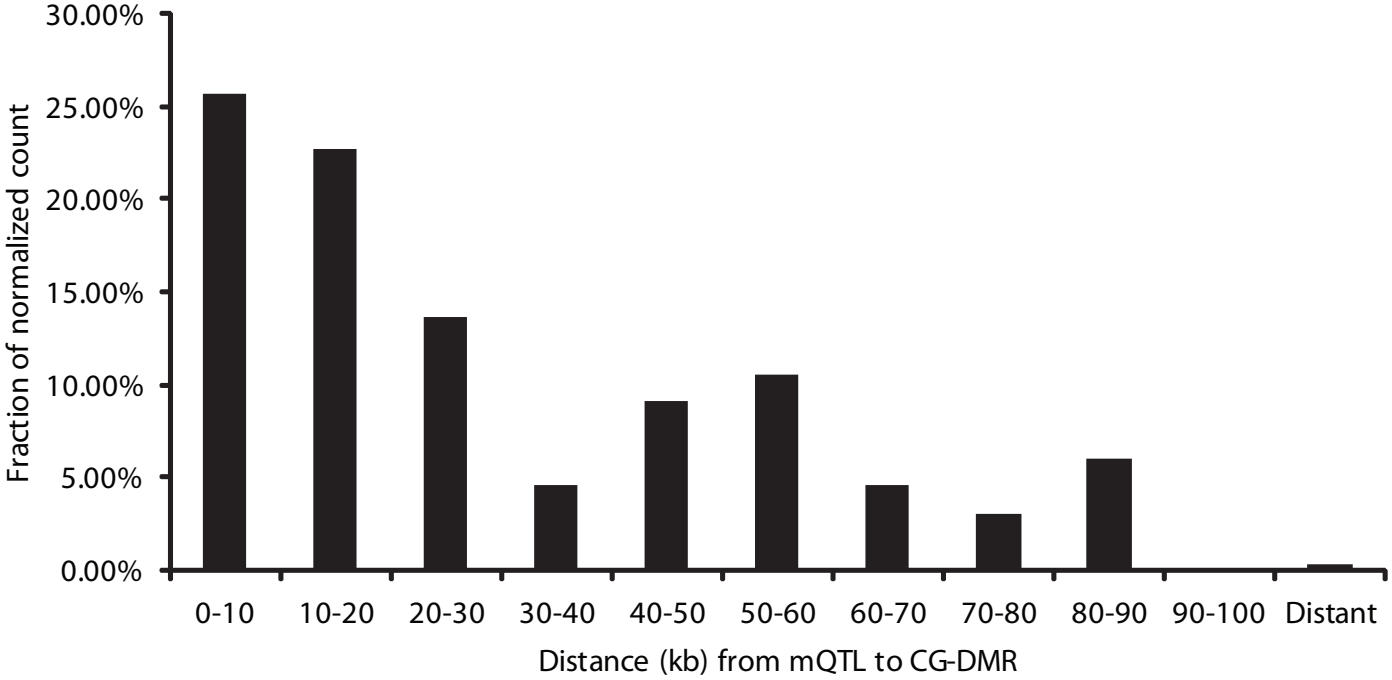
Supplementary Fig. 23. Chromosome V mQTL anlaysis. (a) a histogram representing the distribution of C-DMRs along the chromosome. (b) a histogram representing the distribution of C-DMRs that passed the required thresholds for association mapping.along the chromosome. (c) a histogram representing the distribution of C-DMRs with an mQTL along the chromosome. (d) a histogram representing the distribution of C-DMRs with no significantly associated mQTL along the chromosome.



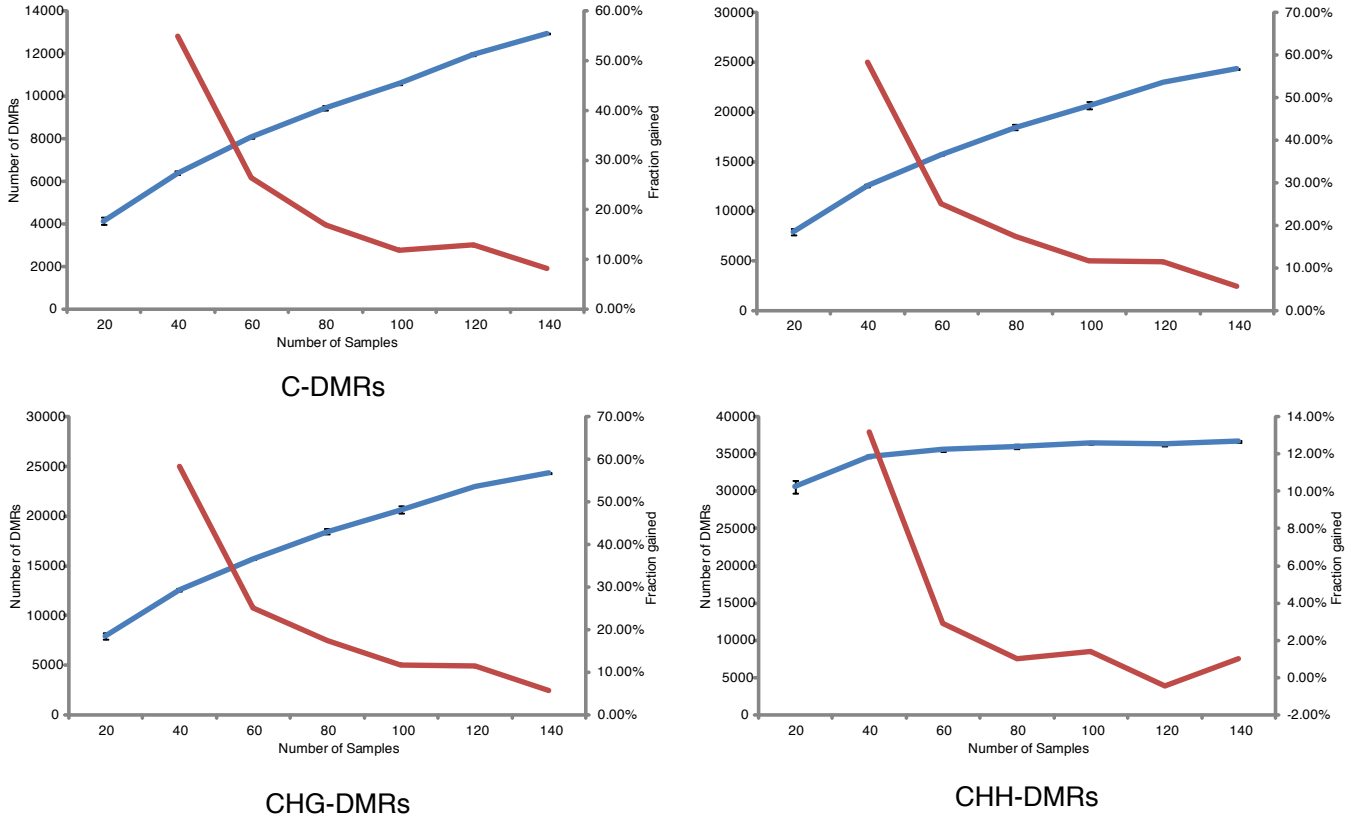
Supplementary Fig. 24. Chromosome V mQTL anlaysis. (a) a histogram representing the distribution of CG-DMRs along the chromosome. (b) a histogram representing the distribution of CG-DMRs that passed the required thresholds for association mapping.along the chromosome. (c) a histogram representing the distribution of CG-DMRs with an mQTL along the chromosome. (d) a histogram representing the distribution of CG-DMRs with no significantly associated mQTL along the chromosome.



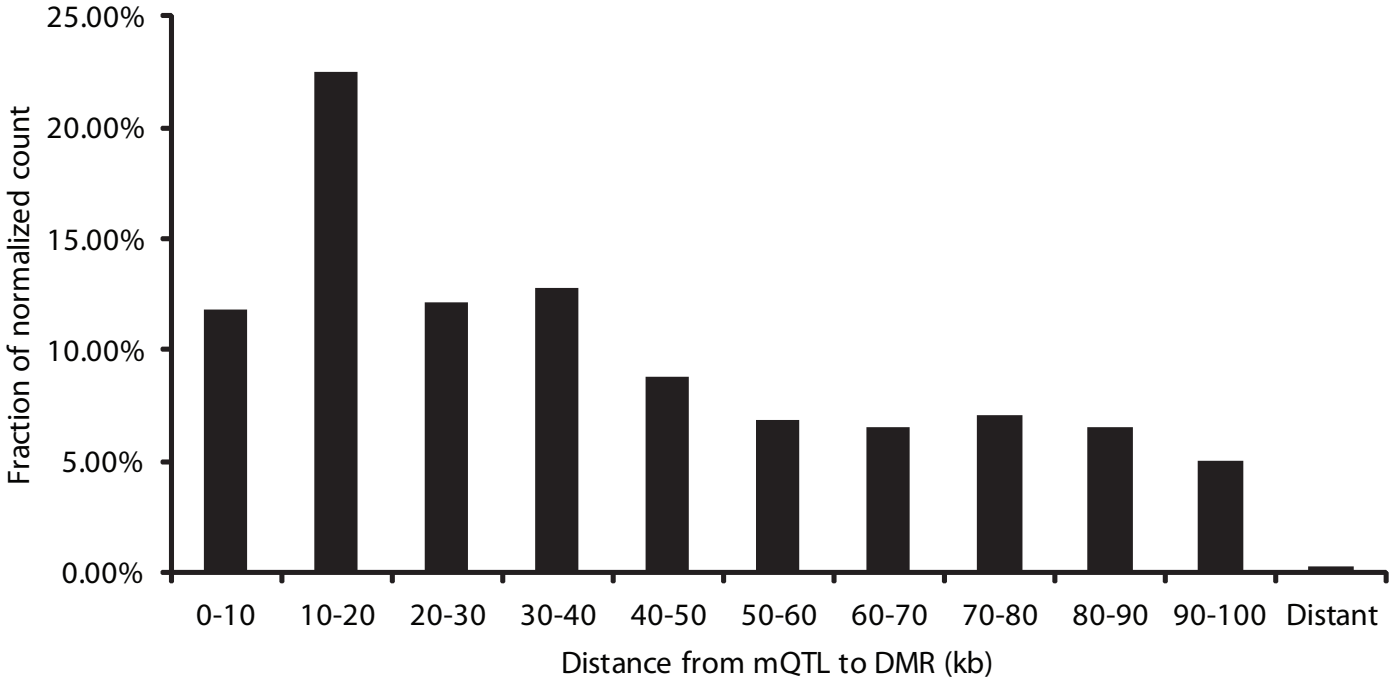
Supplementary Fig. 25. A summary of the comparison of the results from using EMMAX or MLMM for the C-DMR GWA. THe word before the slash indicates the state of the C-DMR when run with EMMAX, and the word after the slash indicates the state of the C-DMR when run with MLMM.



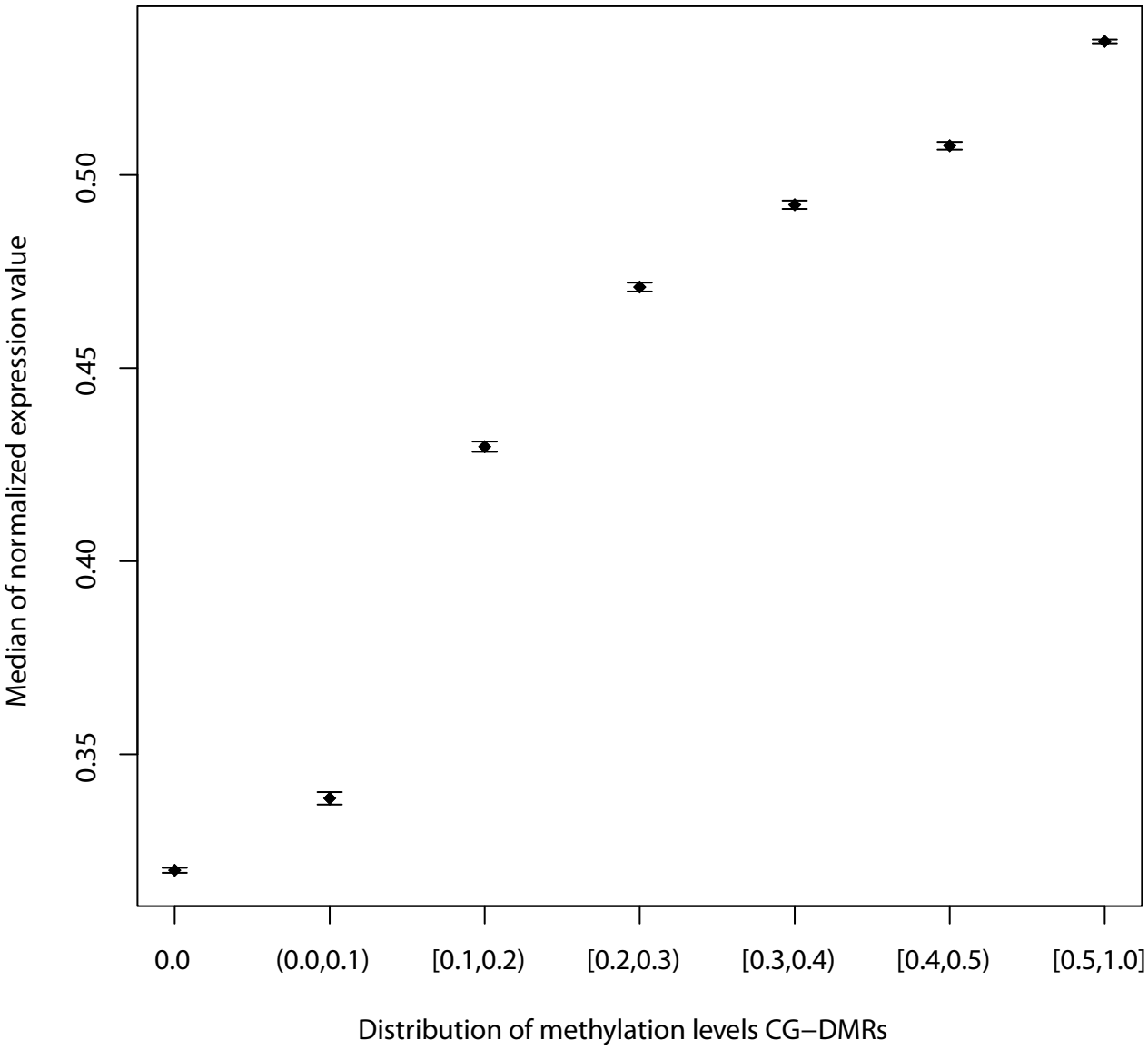
Supplementary Fig. 26. The distribution of distances of significant mQTL from their CG-DMRs normalized for the base pair space covered by each range of distances. A distant mQTL is defined as any mQTL that is more than 100 kb from its CG-DMR.



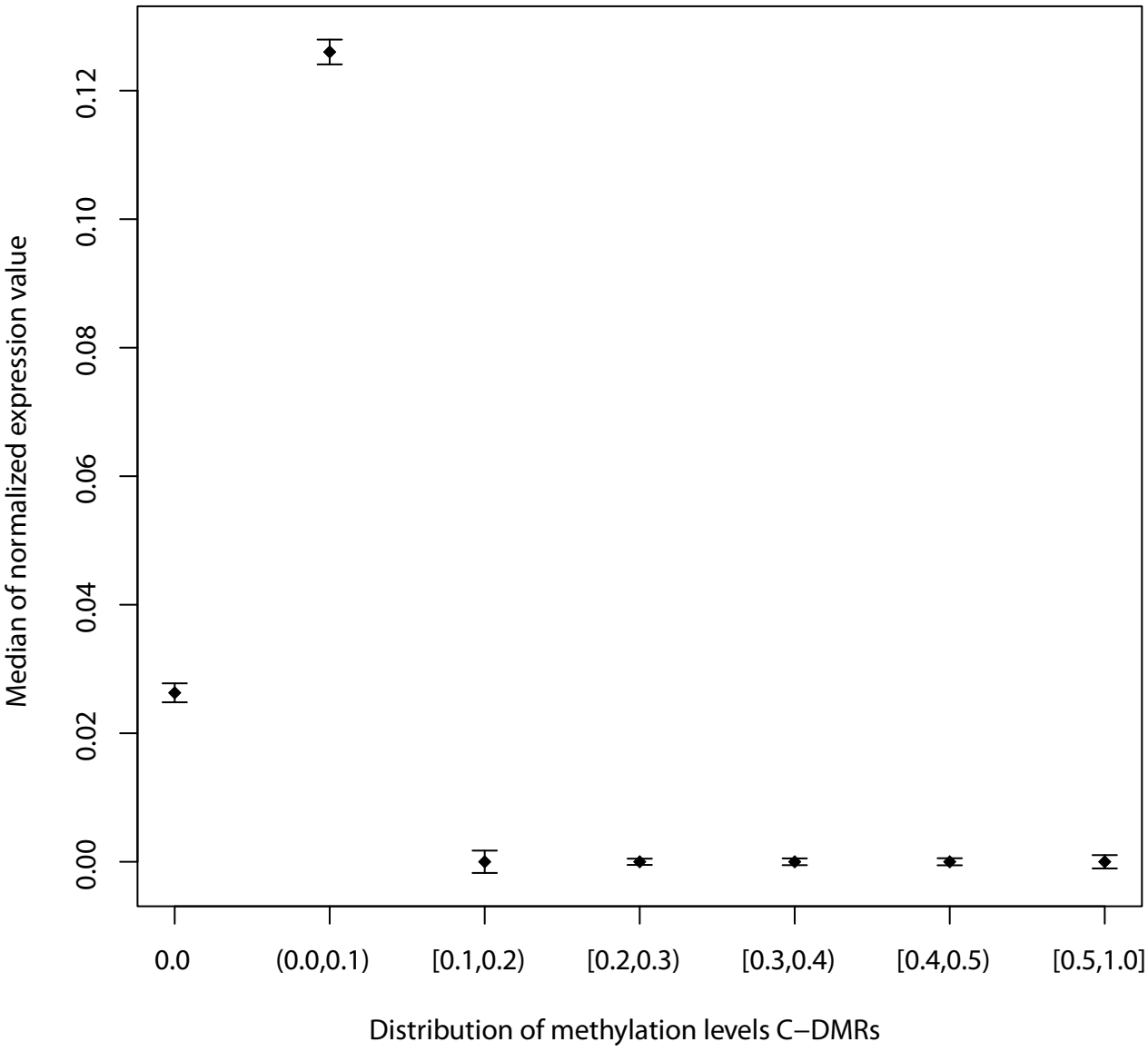
Supplementary Fig. 27. Plots of the number of DMRs in various nucleotide contexts found after randomly subsampling the data. The blue line represents the mean number of DMRs found in a particular context for a particular number of samples and the red line represents the precent of additional DMRs found over the last number of samples. Error bars represent the s.e.m. Although CHH DMRs seemed to have reaached a saturation point, this analysis indicates that there are still other kinds of DMRs to be found.



Supplementary Fig. 28. The distribution of distances of significant mQTL from the randomization method from their C-DMRs for the base pair space covered by each range of distances. A distant mQTL is defined as any mQTL that is more than 100 kb from its CG-DMR.



Supplementary Fig. 29. The median normalized expression values from Figure 2j. Error bars indicate the bootstrap confidence intervals around the median.



Supplementary Fig. 30. The median normalized expression values from Figure 2k. Error bars indicate the bootstrap confidence intervals around the median.