

FM pulsing can be implemented by a simple circuit of three genes (*rsbW*, *rsbV*, and *sigB*), with input from a phosphatase complex. This system provides a fundamental signal-processing capability to bacterial cells, enabling them to convert steady “DC” inputs into pulsatile, predominantly “AC” outputs. Noise plays a key functional role in this signal processing system (3). The  $\sigma^B$  circuit conserves its core architecture in diverse bacteria (7), and other alternative sigma factors similarly feature both posttranslational regulation by anti-sigma factors and autoregulatory feedback. Thus, related stochastic pulse modulation schemes are likely employed more generally in bacteria (10). The relatively slow time scale of  $\sigma^B$  pulses (Fig. 1E) could confer advantages in responding to unpredictable environments and maintaining a broad, but dynamic, distribution of states in the population through bet-hedging (25, 26). Given the negative effect of  $\sigma^B$  activation on growth rate in some conditions, even under energy stress (27), these results suggest that cells balance the benefits and costs of  $\sigma^B$  activation dynamically. It will be interesting to see whether other dynamic encoding schemes are similarly implemented by relatively simple circuit modules.

#### References and Notes

1. A. Raj, A. van Oudenaarden, *Cell* **135**, 216 (2008).
2. N. Rosenfeld, J. W. Young, U. Alon, P. S. Swain, M. B. Elowitz, *Science* **307**, 1962 (2005).
3. A. Eldar, M. B. Elowitz, *Nature* **467**, 167 (2010).
4. R. Losick, C. Desplan, *Science* **320**, 65 (2008).
5. L. Cai, C. K. Dalal, M. B. Elowitz, *Nature* **455**, 485 (2008).
6. E. Rotem *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 12541 (2010).
7. M. Hecker, J. Pané-Farré, U. Völker, *Annu. Rev. Microbiol.* **61**, 215 (2007).
8. W. G. Haldenwang, R. Losick, *Nature* **282**, 256 (1979).
9. O. A. Igoshin, M. S. Brody, C. W. Price, M. A. Savageau, *J. Mol. Biol.* **369**, 1333 (2007).
10. T. M. Gruber, C. A. Gross, *Annu. Rev. Microbiol.* **57**, 441 (2003).
11. M. J. Kazmierczak, S. C. Mithoe, K. J. Boor, M. Wiedmann, *J. Bacteriol.* **185**, 5722 (2003).
12. U. Lorenz *et al.*, *Microbes Infect.* **10**, 217 (2008).
13. M. Hecker, U. Völker, *Mol. Microbiol.* **29**, 1129 (1998).
14. J. C. Locke, M. B. Elowitz, *Nat. Rev. Microbiol.* **7**, 383 (2009).
15. S. Zhang, W. G. Haldenwang, *J. Bacteriol.* **187**, 7554 (2005).
16. A. L. Hodgkin, A. F. Huxley, *J. Physiol.* **117**, 500 (1952).
17. G. M. Süel, J. Garcia-Ojalvo, L. M. Liberman, M. B. Elowitz, *Nature* **440**, 545 (2006).
18. M. B. Elowitz, S. Leibler, *Nature* **403**, 335 (2000).
19. G. M. Süel, R. P. Kulkarni, J. Dworkin, J. Garcia-Ojalvo, M. B. Elowitz, *Science* **315**, 1716 (2007).
20. A. Goldbeter, D. E. Koshland Jr., *Proc. Natl. Acad. Sci. U.S.A.* **78**, 6840 (1981).
21. G. J. Melen, S. Levy, N. Barkai, B. Z. Shilo, *Mol. Syst. Biol.* **1**, 2005.0028 (2005).
22. Z. Cheng, F. Liu, X. P. Zhang, W. Wang, *Biophys. J.* **97**, 2867 (2009).
23. J. C. Ray, O. A. Igoshin, *PLOS Comput. Biol.* **6**, e1000676 (2010).

24. S. Alper, L. Duncan, R. Losick, *Cell* **77**, 195 (1994).
25. M. Acar, A. Becskei, A. van Oudenaarden, *Nature* **435**, 228 (2005).
26. E. Kussell, S. Leibler, *Science* **309**, 2075 (2005).
27. T. Schweder, A. Kolyschokow, U. Völker, M. Hecker, *Arch. Microbiol.* **171**, 439 (1999).
28. A. Dufour, W. G. Haldenwang, *J. Bacteriol.* **176**, 1813 (1994).
29. M. S. Brody, K. Vijay, C. W. Price, *J. Bacteriol.* **183**, 6422 (2001).
30. A. A. Wise, C. W. Price, *J. Bacteriol.* **177**, 123 (1995).

**Acknowledgments:** We thank C. Price and D. Rudner for providing strains. We thank A. Eldar, R. Kishony, C. Price, N. Wingreen, J. Levine, and other members of M.B.E.’s laboratory for helpful discussions. Work in M.B.E.’s laboratory was supported by NIH grants R01GM079771 and P50 GM068763, U.S. National Science Foundation CAREER Award 0644463, and the Packard Foundation. J.C.W.L. was supported by the International Human Frontier Science Program Organization and the European Molecular Biology Organization.

#### Supporting Online Material

www.sciencemag.org/cgi/content/full/science.1208144/DC1

Materials and Methods

SOM Text

Figs. S1 to S20

Table S1

References

Movies S1 and S2

10 May 2011; accepted 1 September 2011

Published online 6 October 2011;

10.1126/science.1208144

# Transgenerational Epigenetic Instability Is a Source of Novel Methylation Variants

Robert J. Schmitz,<sup>1,2</sup> Matthew D. Schultz,<sup>1,2,3</sup> Mathew G. Lewsey,<sup>1,2</sup> Ronan C. O’Malley,<sup>2</sup> Mark A. Urch,<sup>1,2</sup> Ondrej Libiger,<sup>4</sup> Nicholas J. Schork,<sup>4</sup> Joseph R. Ecker<sup>1,2,5\*</sup>

Epigenetic information, which may affect an organism’s phenotype, can be stored and stably inherited in the form of cytosine DNA methylation. Changes in DNA methylation can produce meiotically stable epialleles that affect transcription and morphology, but the rates of spontaneous gain or loss of DNA methylation are unknown. We examined spontaneously occurring variation in DNA methylation in *Arabidopsis thaliana* plants propagated by single-seed descent for 30 generations. We identified 114,287 CG single methylation polymorphisms and 2485 CG differentially methylated regions (DMRs), both of which show patterns of divergence compared with the ancestral state. Thus, transgenerational epigenetic variation in DNA methylation may generate new allelic states that alter transcription, providing a mechanism for phenotypic diversity in the absence of genetic mutation.

Cytosine methylation is a DNA base modification with roles in development and disease in animals as well as in silencing transposons and repetitive sequences in plants and fungi (1). In plants, CG methylation is commonly found within gene bodies (2–5), whereas non-CG methylation, CHG and CHH (where H is A, C, or T), is enriched in transposons and repetitive sequences (1). The RNA-directed DNA methylation (RdDM) pathway targets both CG and non-CG sites for methylation and is com-

monly associated with transcriptional silencing (6). This pathway can also target and silence protein-coding genes, giving rise to epigenetic alleles or so-called epialleles that can be heritable through mitosis and/or meiosis (7, 8) and can be dependent on the methylation of a single CG dinucleotide (9).

Two meiotically heritable epialleles resulting in morphological variation are the *peloric* (*Linaria vulgaris*) and *colorless non-ripening* (*Solanum lycopersicum*) loci (10, 11). Both show

spontaneous epigenetic silencing events within their respective populations (10, 12). However, the frequency at which such spontaneous meiotically heritable epialleles naturally arise in populations is unknown. Although epiallelic variation has been identified between genetically diverse populations within *Arabidopsis thaliana* (13), it is unclear whether these identified epialleles are due to underlying genetic variation. Epialleles have also been artificially generated after mutagenesis or because of mutations in the cellular components required for the maintenance of DNA methylation (14–16).

An *A. thaliana* (Columbia-0) population, the MA lines, derived by single-seed descent for 30 generations (17) was used to examine the extent of naturally occurring variation in DNA methylation and the frequency at which spontaneous epialleles emerge over time. We used the MethylC-Seq method (3) to determine the whole-genome base resolution DNA methylomes for three ancestral

<sup>1</sup>Plant Biology Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA. <sup>2</sup>Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA. <sup>3</sup>Bioinformatics Program, University of California at San Diego, La Jolla, CA 92093, USA. <sup>4</sup>The Scripps Translational Science Institute and the Department of Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla, CA 92037, USA. <sup>5</sup>Howard Hughes Medical Institute, The Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, CA 92037, USA.

\*To whom the correspondence should be addressed. E-mail: ecker@salk.edu

MA lines (numbers 1, 12, and 19) and five descendant MA lines (numbers 29, 49, 59, 69, and 119) (fig. S1). We refer to lines 1, 12, and 19 as ancestors throughout this study, although they are not direct ancestors because they are three generations removed from the original founder line (fig. S1). These specific descendant lines were selected because their genomes have been sequenced and they have a known level of spontaneous mutation (18). Biological replicates (sibling plants) for each leaf methylome were sequenced to an average of ~34-fold coverage, which allowed for an average per line examination of 39,897,093 (96.35%) uniquely mapped cytosines and 5,307,077 (98.39%) uniquely mapped CGs (table S1).

A total of 1,730,761 CGs were methylated (mCGs) in at least one MA line (Fig. 1A), and about 91% of the covered mCGs were invariably methylated across all eight lines (19). The variable mCGs revealed a set of 114,287 high-confidence CG single methylation polymorphisms (SMPs) that showed a consensus of the methylation status of CG dinucleotides between biological replicates (Fig. 1A). Next, a reference MA founder DNA methylome was created by pooling the completely conserved mCG site calls for all ancestral MA lines and used to determine the frequency of discordant CG-SMP sites within the descendant population (Fig. 1B). Within the descendant lines, ~1.62% of the CG methylome shows susceptibility to dynamic acquisitions and losses of mCGs over time (table S2). On average, ~66,000 methylated CG-SMPs (mCG-SMPs) were identified for each ancestral and descendant line (fig. S2). Although the total number of mCG-SMPs was similar between all lines, the conservation of these polymorphisms among and between ancestral and descendant populations was different (Fig. 1C and table S3). A pairwise comparison of both populations for methylation conservation, estimated by global similarity of mCG-SMP sites (19), revealed that all of the ancestral lines are highly similar (table S4). Descendant lines showed greater similarity in CG-SMPs methylation status to ancestral lines than to other descendant lines (table S4).

We calculated an estimate of the epimutation rate per generation in this population by using linear regression and TREE PUZZLE, which revealed 704 and 2876 methylation changes each generation, respectively (19). We estimated a lower bound of the epimutation rate with the linear regression results, which revealed  $4.46 \times 10^{-4}$  methylation polymorphisms per CG site per generation ( $P < 0.0000216$ ) (table S5). This finding contrasts with the previously reported spontaneous genetic mutation rate of  $7 \times 10^{-9}$  base substitutions per site per generation for these same MA lines (18). The TREE PUZZLE analysis revealed higher estimated epimutation rates in earlier generations (19). One possible source of this variation could be due to seed age, storage, and/or selection for seed survival. Therefore, although

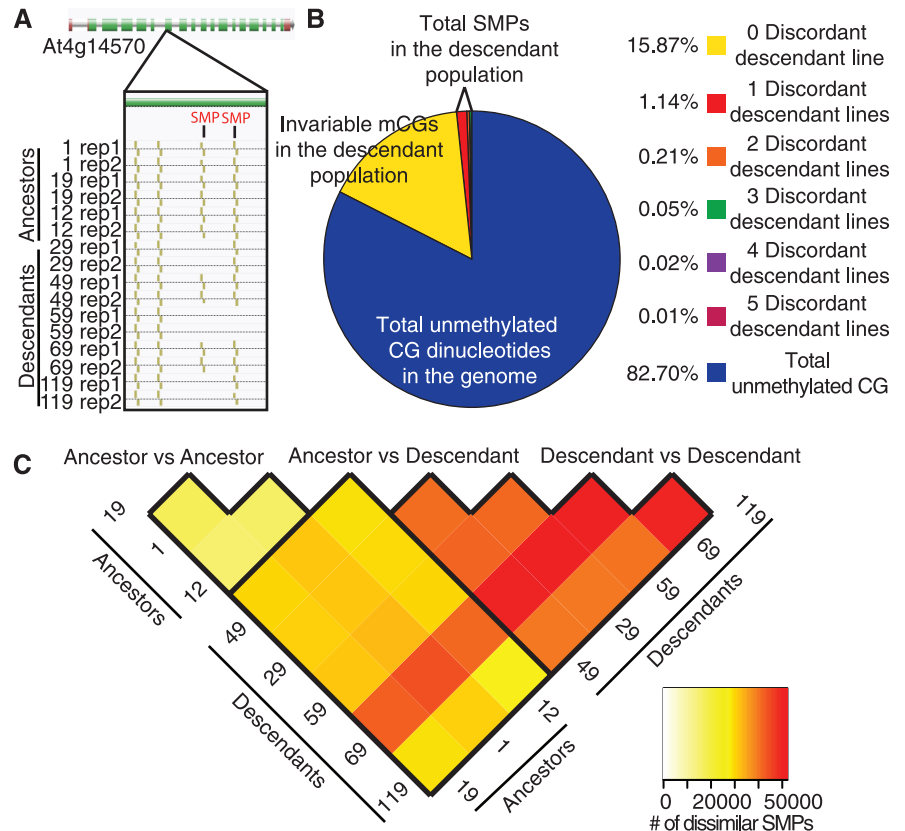
DNA methylation is predominantly static over relatively long periods of time, changes in cytosine methylation do occur and at a frequency greater than that of mutation observed at the DNA sequence level.

By using CG-SMPs derived from both ancestral and descendant populations, we carried out a genome-wide analysis of differentially methylated regions (DMRs) and identified 2485 CG-DMRs that ranged in size from 11 to 1110 base pairs (bp) (Fig. 2A and table S6). Hierarchical clustering of CG-DMRs in this population, calculated solely on the basis of the methylation density, revealed that the ancestral lines segregate as an independent cluster from the descendant lines (Fig. 2B and fig. S3). Multivariate distance-based regression (MDMR) (20, 21) confirmed this finding, indicating a statistically significant ( $P < 0.00005$ ) association between ancestor or descendant status and methylation density of the CG-DMR profiles. The ancestor or descendant status explained 47% of the variance in the dissimilarity in methylation density of CG-DMRs between pairs of samples, indicating that, over time, there is a divergence of DNA methylation patterns in both formation and elimination of CG-DMRs. Furthermore, the genome-wide locations of these CG-DMRs were not uniformly distributed ( $P < 2.20 \times 10^{-16}$ ), because 60.5% (1504/2485)

were found in genic regions compared with 3.3% (82/2485) and 36.2% (899/2485) located in intergenic regions and transposons, respectively (Fig. 2B).

Next, we performed a genome-wide survey for nonCG-DMRs and uncovered a total of 284 among all eight lines (table S7). In general, the nonCG-DMRs were largely localized to intergenic regions (141/284) of the genome, because only 57/284 overlapped with genes and 86/284 overlapped with transposons. The size ranges of the nonCG-DMRs were similar to those of the CG-DMRs because the vast majority occurred in smaller segments of the genome (10 to 682 bp). Therefore, variation in DNA methylation appears to occur in all three methylation sequence contexts.

CG methylation is present within gene bodies and is enriched toward the 3' end (2–5), whereas CG and nonCG methylation is associated with heterochromatin, transposons, and repetitive sequences (1). In agreement with these findings, we observed that the 3' portion of genes contained the greatest source of CG-DMRs and that the majority of nonCG-DMRs were enriched outside of the gene bodies (Fig. 2C). Furthermore, we observed a ~twofold depletion of CG-DMRs in exons compared with introns (Fig. 2D). The genome-wide distributions of CG-SMPs, CG-DMRs,



**Fig. 1.** Epigenetic variation of CG-SMPs. **(A)** An example of a CG-SMP. Gold lines indicate CG methylation, maroon rectangle indicates the untranslated regions, and green rectangles indicated exons. **(B)** A breakdown of the methylation distribution of CG dinucleotides among all samples. **(C)** A heatmap indicating the number of CG-SMPs that differ between two samples (table S3).

and nonCG-DMRs were depleted in heterochromatic regions in the genome (Fig. 2, E and F). These depletions were mostly observed at the pericentromeres and centromeres (Fig. 2, E and F, and figs. S4 and S5). CG-DMRs are enriched in transposons located in euchromatin but depleted in transposons present near the centromere. Because the centromeric regions of the genome contain the highest density of DNA methylation (Fig. 2, E and F), these observations combined with the observations that CG-DMRs are enriched in intron sequences may indicate that DNA methylation that is associated with nucleosomes (22) (i.e., exons or tightly packaged chromatin in the pericentromeres and centromeres) may be maintained at a higher fidelity and that DNA methylation not associated with nucleosomes may undergo greater epigenetic drift.

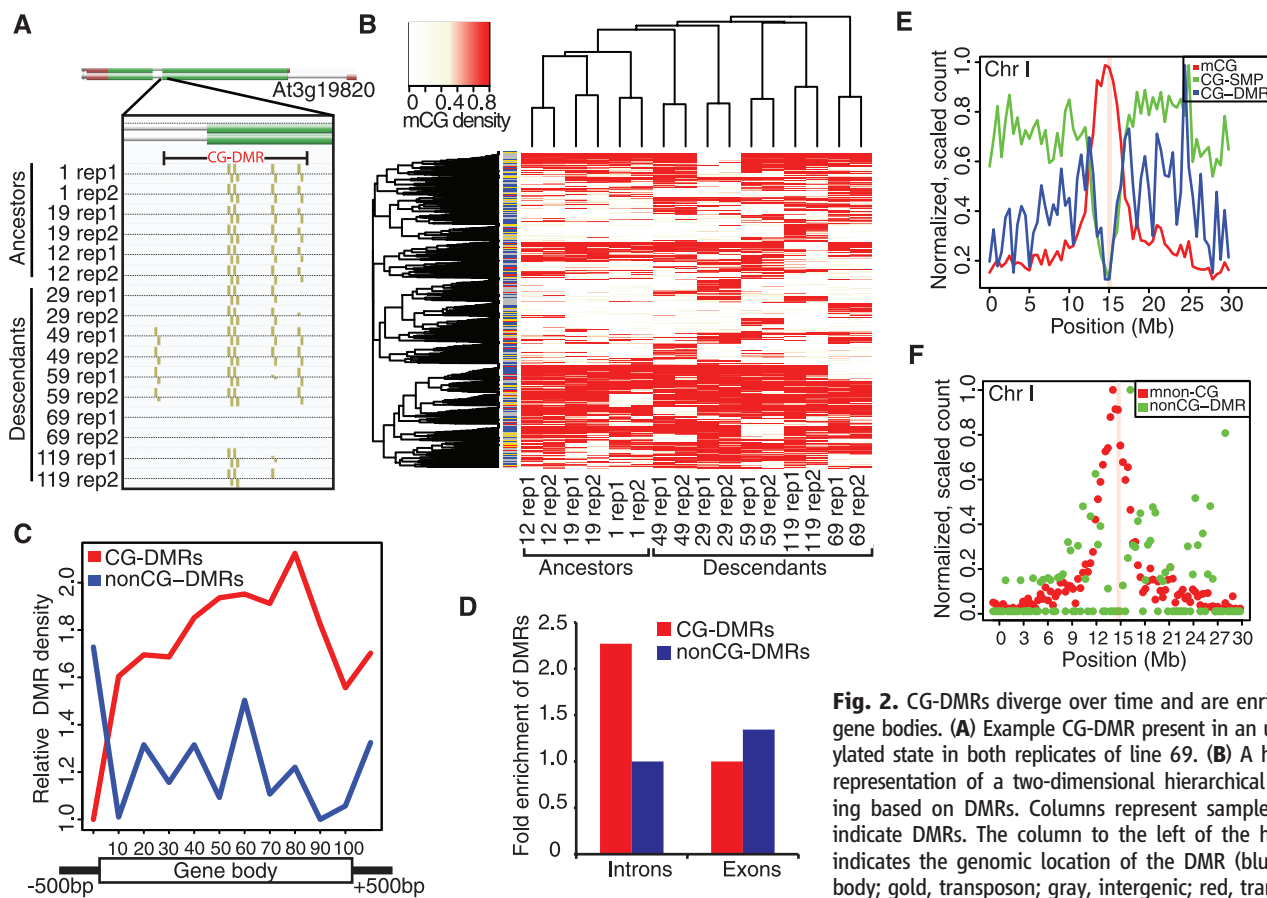
A genome-wide screen for DMRs simultaneously occurring in all three methylation sequence contexts (C-DMRs are CG, CHG, and CHH) was performed to assess the extent of epiallelic variation that is characteristic of RdDM across the MA population. In total, 72 C-DMRs were identified, of which functional categorization

revealed that two-thirds overlapped with transposon and intergenic sequences whereas about one-third overlapped with gene bodies and promoters (Fig. 3A and table S8). To determine whether transposition-induced methylation could potentially give rise to the methylated C-DMRs (mC-DMRs) (23), genomic DNA encompassing all C-DMRs was amplified and compared in all ancestral and descendant lines. In every case, the observed amplicon size was identical for all MA lines and was equal to the expected size of the locus (table S8), indicating that these C-DMRs are unlinked to cis-genetic variation located within 500 bp, a distance that would be expected to reveal methylation induced by transposon insertions at these loci (23). Additionally, none of the genetic variants identified by genome resequencing of this population (18) overlapped with any of these C-DMRs. Lastly, restriction enzyme digestion and Southern blot analyses were performed to rule out the possibility that copy number variants were the cause of spontaneous epiallele formation, as is the case for the *PAI* epialleles (24). In all cases examined, the observed hybridization pattern and gene copy number were identical for each of the MA lines

(fig. S6). Therefore, we conclude that the 72 C-DMRs represent a set of spontaneously occurring epialleles within the MA lines, because they were not associated with any genetic variation.

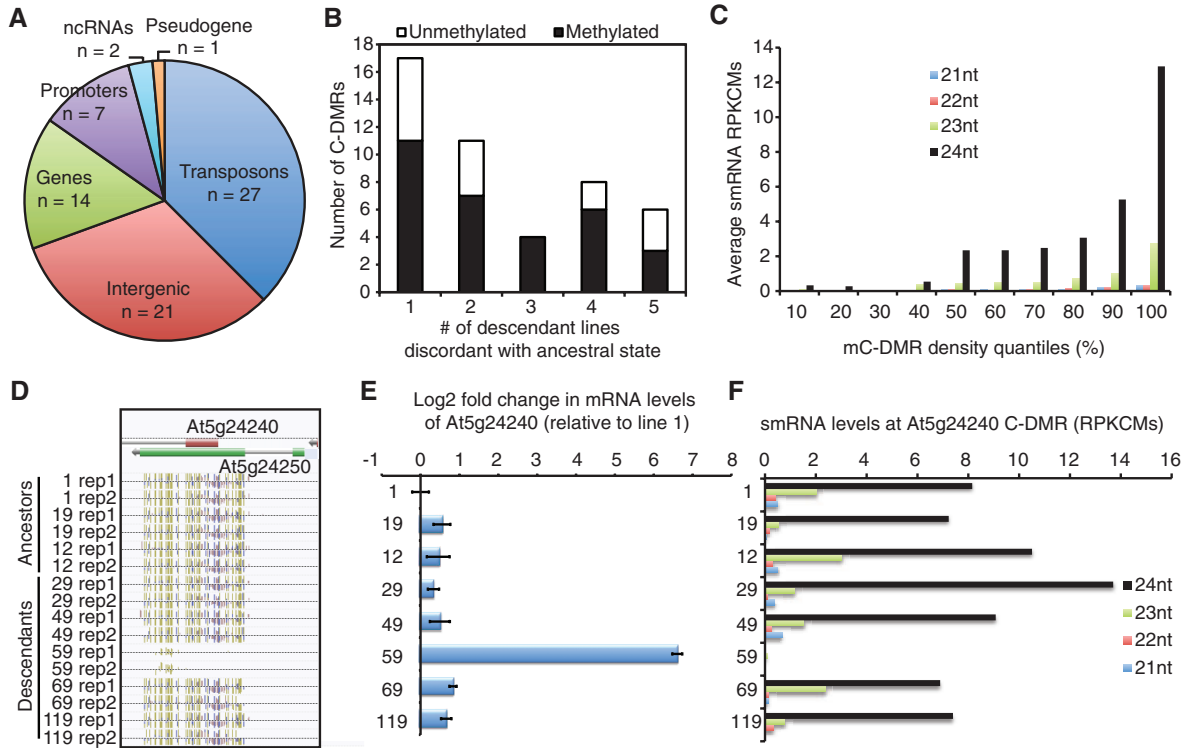
By using a set of C-DMRs that exhibited an identical methylation status (fig. S7), we determined the frequency of discordance of the ancestral state with the descendant lines and found that 29 of the C-DMRs were highly variable (>1 descendant line was discordant with the ancestral state) (Fig. 3B). C-DMRs discordant in only one of the five descendant lines were the most frequent class, but there was an unexpectedly high number of C-DMRs (63%) that were discordant in more than one descendant (Fig. 3B). Within the set of 576 C-DMRs identified (eight lines by 72 C-DMRs), 7 were discordant between the biological replicates (table S8). These data suggest that, although many C-DMRs represent the formation of spontaneous epialleles, a small subset may reflect the presence of “hotspots” (metastable epialleles).

We sequenced small RNA (smRNA) populations for all eight lines and found that smRNAs [represented as RPKCMs (reads per kilobase of each C-DMR per million reads) in



**Fig. 2.** CG-DMRs diverge over time and are enriched in gene bodies. **(A)** Example CG-DMR present in an unmethylated state in both replicates of line 69. **(B)** A heatmap representation of a two-dimensional hierarchical clustering based on DMRs. Columns represent samples. Rows indicate DMRs. The column to the left of the heatmap indicates the genomic location of the DMR (blue, gene body; gold, transposon; gray, intergenic; red, transposon in gene body). **(C)** The average distribution of CG-DMRs

(red) and nonCG-DMRs (blue) across gene bodies (from the start of the 5' UTR to the end of the 3' UTR, including 500 bp up- and downstream). **(D)** CG gene-body DMRs are specifically depleted in exons. **(E)** Genome-wide distributions of mCG (red), CG-SMPs (green), and CG-DMRs (blue) across chromosome I. **(F)** Genome-wide distributions of methylated nonCGs (mnonCG, red) and nonCG-DMRs (green) across chromosome I. The centromere is indicated by the pink vertical bar for (E) and (F).



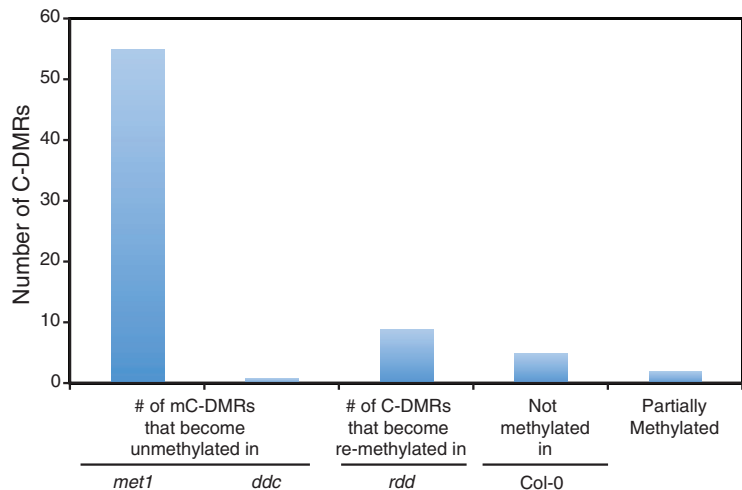
**Fig. 3.** Epiallelic variation at protein-coding loci is associated with transcriptional variation. **(A)** Classification of C-DMRs and their genomic locations. **(B)** The number of descendant lines discordant with the ancestral C-DMR state and the C-DMR methylation status. The black portions of the bar indicate the descendant C-DMRs that became methylated, whereas the white portions indicate regions that became unmethylated, compared with the ancestral population. **(C)** The 24-nt smRNA levels are associated with increasing methylation density. The 24-nt smRNA RPKCMs for all 576 C-DMRs (8 MA lines by

72 C-DMRs) were ranked and binned into 10% quantiles, and then the average mC densities were plotted. **(D)** A representative C-DMR at At5g24240 in which both biological replicates of descendant line 59 were unmethylated. **(E)** qRT-PCR analysis of At5g24240 reveals >50-fold increase in mRNA abundance in unmethylated line 59. Error bars indicate SEM. **(F)** The 24-nt smRNAs are enriched specifically in the MA lines that are transcriptionally silenced in **(E)** for the At5g24240 locus with the exception of line 59, which is abundantly expressed in **(E)**.

tables S9 to 12] were associated with an increase in the average methylation density of C-DMRs (Fig. 3C). Furthermore, this association resembled a binary switch, because the most densely methylated C-DMRs contained abundant 24-nucleotide (nt) smRNAs (Fig. 3C).

Of the eight previously documented plant epialleles resulting in phenotypic variation, all affected transcriptional output of the differentially methylated locus (*9-11*, *23-28*). mRNA abundance was measured in all eight lines with quantitative reverse transcription polymerase chain reaction (qRT-PCR) at eight C-DMRs that overlapped with protein-coding regions. In four of these genes, the gain or loss of DNA methylation was correlated with a large decrease or increase in mRNA abundance, respectively, and with the presence of 24-nt smRNAs at each silenced epiallele (Fig. 3, D to F, and fig. S8). These findings reveal that changes in epiallelic state can lead to major effects on transcriptional output (fig. S9).

We also observed that the methylation status of one C-DMR resulted in alternative promoter usage of *ACTIN RELATED PROTEIN 9* (At5g43500) (fig. S10C). The loss of DNA methylation within the 5' untranslated region (UTR) of the At5g43500.1 isoform led to an increase in



**Fig. 4.** Methylation status of all 72 epialleles in methylation and demethylation mutant backgrounds. Most of the epialleles become unmethylated in *met1-3*, whereas a smaller number become re-methylated in the DNA demethylase triple mutant *rdd*.

mRNA expression, whereas expression of isoform At5g43500.2, with a transcriptional start site located further downstream, was unaffected (fig. S10, D and E).

Although epialleles can have major impacts on phenotypic diversity, until now their identi-

fication was not trivial. Even more puzzling is the origin of “pure” alleles, which are defined by their formation in the absence of any genetic variation in cis or trans (8). One route to epiallele formation may be the failure to correctly maintain the proper methylation status through-

out epigenetic reprogramming that occurs post-fertilization (29, 30). It is noteworthy that 63 of the 72 C-DMRs overlap with regions previously shown to have altered methylation patterns in methylation enzyme mutants (Fig. 4) (3). Of the 14 C-DMRs that overlap with genes, 5 become reexpressed in *metl-3* and 1 transcript becomes silenced in *rdt* (3). These results suggest that a failure to faithfully maintain genome-wide methylation patterns by *MET1* and/or *RDD* is likely one source of spontaneous epiallele formation.

Regardless of their origin, the majority of epialleles identified in this study are meiotically stable and heritable across many generations in this population. Understanding the basis for such trans-generational instability and the mechanism(s) that trigger and/or release these epiallelic states will be of great importance for future studies.

#### References and Notes

1. J. A. Law, S. E. Jacobsen, *Nat. Rev. Genet.* **11**, 204 (2010).
2. S. J. Cokus *et al.*, *Nature* **452**, 215 (2008).
3. R. Lister *et al.*, *Cell* **133**, 523 (2008).
4. X. Zhang *et al.*, *Cell* **126**, 1189 (2006).
5. D. Zilberman, M. Gehring, R. K. Tran, T. Ballinger, S. Henikoff, *Nat. Genet.* **39**, 61 (2007).
6. S. W.-L. Chan *et al.*, *Science* **303**, 1336 (2004).
7. J. Paszkowski, U. Grossniklaus, *Curr. Opin. Plant Biol.* **14**, 195 (2011).

8. E. J. Richards, *Nat. Rev. Genet.* **7**, 395 (2006).
9. K. Shibuya, S. Fukushima, H. Takatsuji, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 1660 (2009).
10. P. Cubas, C. Vincent, E. Coen, *Nature* **401**, 157 (1999).
11. K. Manning *et al.*, *Nat. Genet.* **38**, 948 (2006).
12. A. J. Thompson *et al.*, *Plant Physiol.* **120**, 383 (1999).
13. M. W. Vaughn *et al.*, *PLoS Biol.* **5**, e174 (2007).
14. F. Johannes *et al.*, *PLoS Genet.* **5**, e1000530 (2009).
15. F. K. Teixeira *et al.*, *Science* **323**, 1600 (2009); 10.1126/science.1165313.
16. A. Vongs, T. Kakutani, R. A. Martienssen, E. J. Richards, *Science* **260**, 1926 (1993).
17. R. G. Shaw, D. L. Byers, E. Darms, *Genetics* **155**, 369 (2000).
18. S. Ossowski *et al.*, *Science* **327**, 92 (2010).
19. Additional experiments and descriptions of methods used to support our conclusions are presented as supporting material on Science Online.
20. C. M. Nievergelt *et al.*, *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **141B**, 234 (2006).
21. M. A. Zapala, N. J. Schork, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 19430 (2006).
22. R. K. Chodavarapu *et al.*, *Nature* **466**, 388 (2010).
23. J. Liu, Y. He, R. Amasino, X. Chen, *Genes Dev.* **18**, 2873 (2004).
24. J. Bender, G. R. Fink, *Cell* **83**, 725 (1995).
25. S. Melquist, B. Luff, J. Bender, *Genetics* **153**, 4017 (1999).
26. S. E. Jacobsen, E. M. Meyerowitz, *Science* **277**, 1100 (1997).
27. H. Saze, T. Kakutani, *EMBO J.* **26**, 3641 (2007).
28. W. J. Soppe *et al.*, *Mol. Cell* **6**, 791 (2000).
29. R. A. Mosher *et al.*, *Nature* **460**, 283 (2009).
30. R. K. Slotkin *et al.*, *Cell* **136**, 461 (2009).

**Acknowledgments:** We thank M. White, R. Lister, M. Galli, and R. Amasino for discussions; R. Shaw and E. Darms for seeds; J. Nery for sequencing operations; and M. Axtell for Southern blot protocol. R.J.S. was supported by an NIH National Research Service Award postdoctoral fellowship (F32-HG004830). M.D.S. was supported by a NSF Integrative Graduate Education and Research Traineeship grant (DGE-0504645). M.G.L. was supported by an European Union Framework Programme 7 Marie Curie International Outgoing Fellowship (project 252475). O.L. and N.J.S. are supported by NIH/National Center for Research Resources grant number UL1 RR025774. This work was supported by the Mary K. Chapman Foundation, the NSF (grants MCB-0929402 and MCB1122246), the Howard Hughes Medical Institute, and the Gordon and Betty Moore Foundation (GBMF) to J.R.E. J.R.E. is a HHMI-GBMF Investigator. Analyzed data sets can be viewed at [http://neomorph.salk.edu/30\\_generations/browser.html](http://neomorph.salk.edu/30_generations/browser.html). Sequence data can be downloaded from National Center for Biotechnology Information Sequence Read Archive (SRA035939). Correspondence and requests for materials should be addressed to J.R.E. (ecker@salk.edu).

#### Supporting Online Material

[www.sciencemag.org/cgi/content/full/science.1212959/DC1](http://www.sciencemag.org/cgi/content/full/science.1212959/DC1)  
Materials and Methods

SOM Text

Figs. S1 to S11

Tables S1 to S16

References

22 August 2011; accepted 7 September 2011

Published online 15 September 2011;

10.1126/science.1212959

# Computation-Guided Backbone Grafting of a Discontinuous Motif onto a Protein Scaffold

Mihai L. Azoitei,<sup>1\*</sup> Bruno E. Correia,<sup>1,2\*</sup> Yih-En Andrew Ban,<sup>1†</sup> Chris Carrico,<sup>1,3</sup> Oleksandr Kalyuzhnyi,<sup>1</sup> Lei Chen,<sup>4</sup> Alexandria Schroeter,<sup>1</sup> Po-Ssu Huang,<sup>1</sup> Jason S. McLellan,<sup>4</sup> Peter D. Kwong,<sup>4</sup> David Baker,<sup>1,5</sup> Roland K. Strong,<sup>3</sup> William R. Schief<sup>1,6,7‡</sup>

The manipulation of protein backbone structure to control interaction and function is a challenge for protein engineering. We integrated computational design with experimental selection for grafting the backbone and side chains of a two-segment HIV gp120 epitope, targeted by the cross-neutralizing antibody b12, onto an unrelated scaffold protein. The final scaffolds bound b12 with high specificity and with affinity similar to that of gp120, and crystallographic analysis of a scaffold bound to b12 revealed high structural mimicry of the gp120-b12 complex structure. The method can be generalized to design other functional proteins through backbone grafting.

Computational protein design tests our understanding of protein structure and folding and provides valuable reagents for biomedical and biochemical research; long-term goals include the design of field- or clinic-ready biosensors (1), enzymes (2), therapeutics (3), and vaccines (4, 5). A major limitation has been an inability to manipulate backbone structure; most computational protein design has involved sequence design on predetermined backbone structures or with minor backbone movement (1–5). Accurate backbone remodeling presents a substantial challenge for computational methods owing to limited conformational sampling and imperfect energy functions (6).

Novel recognition modules (7), inhibitors (8, 9), enzymes (2), and immunogens (4, 5, 10, 11) have been designed by grafting functional constellations of side chains onto protein scaffolds of predefined backbone structure. In all cases, the restriction to using predetermined scaffold backbone structures limited the complexity of the functional motifs that could be transplanted. For example, the de novo enzymes could accommodate grafting of only three or four catalytic groups, whereas many natural enzymes have six or more (12), and the immunogens were limited to continuous (single-segment) epitopes even though most antibody epitopes are discontinuous (involving two or more antigen segments) (13, 14).

To address the challenge of incorporating backbone flexibility modeling into grafting design, we developed a hybrid computational-experimental method for grafting the backbone and side chains of functional motifs onto scaffolds (Fig. 1). We tested this method by grafting a discontinuous HIV gp120 epitope, targeted by the broadly neutralizing monoclonal antibody b12 (15), onto an unrelated scaffold. b12 binds to a conserved epitope within the CD4-binding site (CD4bs) of gp120 (16), an area of great interest for vaccine design. We focused on transplantation of two segments from gp120: residues 365 to 372, known as the CD4b (CD4 binding) loop (17), and residues 472 to 476, known as the ODe (outer domain exit) loop (16). The b12-gp120 interaction involves six or seven backbone segments on gp120 (16), but 60% of the buried surface area on gp120 lies on the CD4b and ODe loops, and a Rosetta energy calculation (18) suggested that these two

<sup>1</sup>Department of Biochemistry, University of Washington, Seattle, WA 98195, USA. <sup>2</sup>Ph.D. Program in Computational Biology, Instituto Gulbenkian de Ciência, Oeiras, Portugal. <sup>3</sup>Division of Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA. <sup>4</sup>Vaccine Research Center, National Institute of Allergy and Infectious Diseases, Bethesda, MD 20892, USA. <sup>5</sup>Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA. <sup>6</sup>IAVI Neutralizing Antibody Center, The Scripps Research Institute, La Jolla, CA 92037, USA. <sup>7</sup>Department of Immunology and Microbial Science, The Scripps Research Institute, La Jolla, CA 92037, USA.

\*These authors contributed equally to this work.

†Present address: Arzeda Corporation, Seattle, WA 98102, USA.

‡To whom correspondence should be addressed. E-mail: schief@scripps.edu