UNIVERSITY OF CALIFORNIA

Los Angeles

5-methylcytosine: from deposition to detection of the 5th base in mammalian genomes

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Molecular, Cell and Developmental Biology

by

Marco Morselli

2017

ABSTRACT OF THE DISSERTATION

5-methylcytosine: from deposition to detection

of the 5th base in mammalian genomes

by

Marco Morselli

Doctor of Philosophy in Molecular, Cell and Developmental Biology

University of California, Los Angeles, 2017

Professor Matteo Pellegrini, Chair

In multicellular organisms, there are many different cell types possessing the same genetic information, each performing a particular role. Many players, such as transcription factors, nucleosome positioning, histone post-translational modifications and non-coding RNAs, contribute to regulate the expression of specific genes, defining the specific cell state. Once the expression pattern is established during development, it is faithfully maintained throughout the life of an organism. In many organisms, a key component of this epigenetic regulation is a covalent modification of cytosines (5meC) in the genome, known as DNA methylation. DNA methylation is associated with repression of transcription if present in gene promoters and it can suppress the transcription of aberrant intragenic transcripts. The focus of my doctoral research has been the development of methods to map the distribution of 5meC genome-wide and understand how DNA methyltransferases are recruited to their targets. Since mammalian genomes are large, current approaches to map the methylation of the entire genome are expensive. Several methods have been developed to assess the methylation status of part of the

genome. Some of them are based on enrichment probes, others on enzymatic digestion. Chapters 1 and 2 are based on methods that assess the methylation status of part of the genome. Reduced-Representation Bisulfite Sequencing (RRBS) captures the majority of CpG islands and promoters. Since only 1% of the genome is assessed with this technique, costs associated with sequencing are dramatically reduced. In chapter 1, this technique is used to discover methylation levels at specific CpG sites associated with complex disease traits. Most of the CpG sites in the genome are methylated and do not have variable methylation levels between different cell types, suggesting that some of the fragments isolated by RRBS do not provide useful information. In order to overcome this limitation, we improved an existing method, Methylation-sensitive Restriction Enzyme-seq (MRE-seq), which enriches for regions poorly methylated (approximately 20% of the genome). This method, called MRE-BS (MRE-Bisulfite Sequencing), is described in chapter 2. The costs are similar to RRBS, but the development of a multiple regression model has allowed us to estimate differential methylation between two samples across 60% of the genome.

Chapter 3 focuses on how *de novo* DNA methyltransferases are recruited to their target sites. This work has shown that the murine DNMT3b is guided by histone post-translational modifications, both in yeast and primordial germ-cells. In general, DNMT3b is absent from regions marked by H3K4me3 and it is recruited in gene-bodies by H3K36me3.

The last chapter focuses on the presence of 5meC in messenger RNA, the function of which is still unknown. We discovered several hundred putative methylation sites that are associated with predicted secondary structures in mRNAs. This finding might be explained either by the recruitment of RNA methyltransferases (RMTs) by structural motives or by the limitations of the method utilized. More experiments are needed to understand what signal is needed for the specific recruitment of RMTs to their target sites and what is the function of this mark.

The dissertation of Marco Morselli is approved.

Arnold J. Berk

Steven E. Jacobsen

Atsushi Nakano

Matteo Pellegrini, Committee Chair

University of California, Los Angeles

2017

To my family and my wife, whose support, love and constant encouragement made this work possible

TABLE OF CONTENTS **PAGE**

LIST OF FIGURES AND SUPPLEMENTARY FIGURES **PAGE**

Chapter 1

Chapter 2

Chapter 3

Chapter 4

LIST OF TABLES AND SUPPLEMENTARY TABLES **PAGE**

Chapter 1

Chapter 2

Chapter 3

Chapter 4

AKNOWLEDGEMENTS

This is probably the most important part of the thesis.

I would like to start, of course, from Professor Pellegrini. There are no words that can express my gratitude for this unique experience. And this is not only for being the best mentor and wonderful person I could possibly dream during my Graduate Studies, but also for having offered me a position as a lab tech before that. It might seem not a big deal for you, but it was and it still is for me. Everything that happened in the past 5 years really derives from that. I hope you didn't regret that decision. As I said, it's really difficult to express my gratitude with words, so just a sincere: Thank You!!

I would also like to thank Professor Steve Jacobsen, for having accepted to be my co-mentor during my graduate studies and for having considered me as part of his lab. Thank you, Steve, and thanks to all the Jacobsen Lab, especially to: Mahnaz, Suhua, Javi, Sylvain, Martin, Ash, Will, Jake, Linda, Qikun, Magda, Basu and Wanlu. I also want to thank the other two members of the Doctoral Committee. Professor Atsushi Nakano, who introduced me a little into the world of stem cells and their potential use as future therapeutic agents. Arnie Berk, co-author of the "Bible" I used during my undergraduate studies, and the first Professor I met on campus. And it was practically magic when you remembered who I was several months later. That episode made me really feel part of the UCLA community.

I would like to continue thanking other two cornerstones of this experience: Mila and Roberto. Roberto, thanks for everything, I think your help during the first weeks after my arrival here in Los Angeles really made a huge difference. I know that I might have been a little bit bothersome at times, if not most of the times, but I really saw you as a brother here. And if now I'm always ready to help other people, well it's also because of you. Thank you. Talking about newly acquired "family members", I would also like to thank Mila Rubbi, who was basically my American mom for the past five years. Apart from the professional perspective, and I've learned a great deal from you, I am most grateful to you for letting me be part of your family. I'm

honored of having shared the ups and downs of the past few years, sometimes with a tear, sometimes with loud laughs (con occhiataccia successiva..e ci siamo intesi).

Now, I'm going to thank my real family members, in Italian, so you don't have to deal with messy online translations. Grazie di tutto, di avermi sempre sostenuto, anche se col magone in gola. Ma il magone in gola ce l'ho avuto sempre anche io. Mi ricordo come se fosse ieri l'ultima notte in Italia prima della mia partenza e tutte le volte che la rivivo, non riesco a trattenermi dalle lacrime. Ciao nonna. Mamma, papà, non so se ve l'ho mai detto, ma penso di aver pianto per quattro ore di fila in aereo dopo che ho letto la vostra lettera. E a poco contano i sorrisi della mamma che, da buona Borrini, cerca di mascherare cosi la tristezza del distacco, anche se più di qualche lacrima alla fine sfugge sempre. Poco conta anche il silenzio del papà, che, da buon Morselli, vuole cosi nascondere la voce tremula causata dalla forte emozione. Grazie anche a te Massi, che seppur nella tua riservatezza, mi hai sempre sostenuto e difeso a spada tratta in ogni mia decisione. Da quando l'inossidabile Bric ci ha lasciato, hai trovato un nuovo fratello in Pepino…fallo arrabbiare come si deve anche da parte mia. È in momenti di riflessione cosi che ti chiedi il motivo per cui hai abbandonato tutto questo. Si, certo, una grande opportunità, una esperienza senza pari…ma sono convinto che nulla di tutto ciò sarebbe accaduto se non fosse per mia moglie. All'epoca ancora non lo era, ma fu lei che mi convinse ad intraprendere questo cammino negli States, per un'occasione che, secondo lei, non era possibile rifiutare. Il merito è tutto suo. Lei ha preso anche la decisione più rischiosa, abbandonare un posto fisso in Italia e tuffarsi in questa esperienza, quasi un salto nel vuoto. In più, ha anche avuto il coraggio di sposarmi. Spero tu non abbia rimpianti, non riuscirei mai a perdonarmelo. Spero che invece mi abbiano gia perdonato Angelo e Milena per avere "rubato" loro la figlia. Grazie del supporto per le nostre decisioni e per sdrammatizzare, a volte, momenti di difficoltà. Non so se e quando torneremo.

Intanto che ci sono, finisco direttamente questa scatola di Kleenex ringraziando un'altra persona speciale: il Prof. Simone Ottonello. Mi ricordo come se fosse ieri quando ci siamo salutati per

l'ultima volta e, ovviamente, mi e scappata "qualche" lacrima. Mi ricordo ancora la frase che mi ha detto: "Le lacrime durano da qui (indicandosi l'occhio) a qui (indicandosi la bocca, su cui era stampato un sorriso)". Aveva già previsto che sarebbe stata un'esperienza positiva. Se lo avessi saputo anche io, sarebbe stato tutto più facile…ma forse è giusto cosi, sicuramente è più bello cosi. Grazie Barbara, volevo dire Prof.ssa Montanini per avermi insegnato insieme ad Andrea ed Elisabetta le basi, e non solo, di fare ricerca in lab. E se le lacrime quel giorno mi sono scappate, beh è perché li a Biochimica (quando ancora così si chiamava, now I've lost track) si era creata una grande famiglia, che saluto: Andre, Mery, Gioietta, Marty, Cri, Bea e Nick, Skizzo (ed Ele), Bomber, Vince, Cavazza, Roby e tutti i Mozza.

Allright, going back to English here, and I hope no more tears from now on.

Thanks to all the members of the always changing family here in the Pellegrini Lab (random order): Giancarlo, Pao-Yang, Weilong, Larry, David, Dennis, Shawn, Eric, Davide, Roberto, Arturo, Giorgia, Alessandra, Kai, Kianoush and the current wet lab line-up with Anela, Giorgia (e Alberto), Simona and Davide (e Chiara). I would also thank my classmates from the ACCESS XIX (I heard it was the best class ever…). I really miss our Saturday nights during the first year! Thanks especially to Roy (and Michelle and best wishes for your little Roy/Michelle), Arthur, Anna, Nathan, Josh, Ao, Ruhi, Sophie, Brittany, Lynnea, Mehmet, Elliot(s), Michael, Katie(s), Angelyn, Matt, Brian, Spencer, Jeff and Preet. I'm also extremely grateful to all the ACCESS and MBI IDP Staff: Pamela, Jennifer, Jody and Frank.

I would like to thank my friends back in Vicobellignano and nearby towns, who are always waiting for my return. Fabio, Anna mi dispiace non essere stato fisicamente vicino a voi durante quel periodo durissimo della vostra vita. Vasty e Silvia, in bocca al lupo per tutto! Volevo ringraziare anche i miei amici qui a L.A., che mitigano la nostalgia degli amici lasciati in patria: Fede, Silvia, Umbe, Ale, Alice, Maria Letizia, e gli ultimi arrivati: Davide, Simo, Albe e Giorgia. Grazie anche alle mie due squadre di calcio del cuore: A.C. Martignana e F.C. United Nations (SBPSL Champion 2017). Finally, I want to thank my bike and the Traffic Volunteer lady.

VITA

PhD Candidate Molecular, Cell and Developmental Biology — 2012 – 2017
Advisor: Matteo Pellegrini. *Co-Advisor:* Steven E. Jacobsen
University of California, Los Angeles

M.Sc. Molecular Biotechnology — 2009 – 2011
University of Parma, Italy

B.Sc. Biotechnology — 2006 – 2009
University of Parma, Italy

Research Experience

Research Staff — Jan 2012 – Sept 2012
Laboratory of Matteo Pellegrini
University of California, Los Angeles

Original Research Thesis Project — Mar 2009 – Nov 2011
Laboratory of Simone Ottonello
University of Parma, Italy

Awards

Dissertation Year Fellowship — 2016-2017
Philip Whitcome Fellowship — 2015-2016
Philip Whitcome Fellowship — 2014-2015

Teaching Experience

187: Research Immersion Laboratory in Genomic Biology — Winter 2015
University of California, Los Angeles

104: Research Immersion Laboratory in Developmental Biology — Fall 2013
University of California, Los Angeles

Publications

Traller JC, Cokus SJ, Lopez DA, Gaidarenko O, Smith SR, McCrow JP, Gallaher SD, Podell S, Thompson M, Cook O, Morselli M, Jaroszewicz A, Allen EE, Allen AE, Merchant SS, Pellegrini M, Hildebrand M. Genome and methylome of the oleaginous diatom *Cyclotella cryptica* reveal genetic flexibility toward a high lipid phenotype. *Biotechnol Biofuels*. 2016 Nov 25;9:258.

Vega-Vaquero A*, Bonora G*, Morselli M*, Vaquero-Sedas MI, Rubbi L, Pellegrini M, Vega-Palas MA. Novel features of telomere biology revealed by the absence of telomeric DNA methylation. *Genome Res*. 2016 Aug;26(8):1047-56.

Meng Q, Ying Z, Noble E, Zhao Y, Agrawal R, Mikhail A, Zhuang Y, Tyagi E, Zhang Q, Lee JH, Morselli M, Orozco L, Guo W, Kilts TM, Zhu J, Zhang B, Pellegrini M, Xiao X, Young MF, Gomez-Pinilla F, Yang X. Systems Nutrigenomics Reveals Brain Gene Networks Linking Metabolic and Brain Disorders. *EBioMedicine*. 2016 May;7:157-66.

Lopez D, Hamaji T, Kropat J, De Hoff P, Morselli M, Rubbi L, Fitz-Gibbon S, Gallaher SD, Merchant SS, Umen J, Pellegrini M. Dynamic Changes in the Transcriptome and Methylome of *Chlamydomonas reinhardtii* throughout Its Life Cycle. *Plant Physiol*. 2015 Dec;169(4):2730-43.

Clark RI, Salazar A, Yamada R, Fitz-Gibbon S, Morselli M, Alcaraz J, Rana A, Rera M, Pellegrini M, Ja WW, Walker DW. Distinct Shifts in Microbiota Composition during *Drosophila* Aging Impair Intestinal Function and Drive Mortality. *Cell Rep*. 2015 Sep 8;12(10):1656-67.

Orozco LD, Morselli M, Rubbi L, Guo W, Go J, Shi H, Lopez D, Furlotte NA, Bennett BJ, Farber CR, Ghazalpour A, Zhang MQ, Bahous R, Rozen R, Lusis AJ, Pellegrini M. Epigenome-wide association of liver methylation patterns and complex metabolic traits in mice. *Cell Metab*. 2015 Jun 2;21(6):905-17.

Morselli M, Pastor WA, Montanini B, Nee K, Ferrari R, Fu K, Bonora G, Rubbi L, Clark AT, Ottonello S, Jacobsen SE, Pellegrini M. *In vivo* targeting of *de novo* DNA methylation by histone modifications in yeast and mouse. *Elife*. 2015 Apr 7;4:e06205.

Chen PY, Montanini B, Liao WW, Morselli M, Jaroszewicz A, Lopez D, Ottonello S, Pellegrini M. A comprehensive resource of genomic, epigenomic and transcriptomic sequencing data for the black truffle *Tuber melanosporum*. *Gigascience*. 2014 Oct 30;3:25.

Montanini B, Chen PY, Morselli M, Jaroszewicz A, Lopez D, Martin F, Ottonello S, Pellegrini M. Non-exhaustive DNA methylation-mediated transposon silencing in the black truffle genome, a complex fungal genome with massive repeat element content. *Genome Biol*. 2014 Jul 31;15(7):411.

Chen PY, Ganguly A, Rubbi L, Orozco LD, Morselli M, Ashraf D, Jaroszewicz A, Feng S, Jacobsen SE, Nakano A, Devaskar SU, Pellegrini M. Intrauterine calorie restriction affects placental DNA methylation and gene expression. *Physiol Genomics*. 2013 Jul 15;45(14):565-76.

OVERVIEW

Multicellular organisms possess cells with functional and phenotypic differences, despite sharing the same genetic information. Each cell's identity is established during development and faithfully maintained during mitotic cell divisions and it is determined by the specific set of genes expressed at a particular time. *Cis*-regulatory regions are sequences on the DNA that can be bound by sequence-specific transcription factors (TFs) influencing the activity of nearby (or distant, if a looping mechanism is involved) genes (1). In addition to a DNA-binding module, TFs possess one or more effector domains that can recruit other transcription factors for cooperative binding, the basal transcriptional machinery, general co-activators, or co-repressors (2). An important set of proteins recruited by TFs consists of enzymes that can modify the chromatin therefore influencing the binding to DNA and/or activity of other regulatory proteins. Example of these are chromatin remodelers, histone-modifying enzymes and DNA methyltransferases. In Eukaryotic cells, the fundamental building blocks of chromatin are called nucleosomes, which provide an efficient wrapping of DNA around the histone octamer. This can influence the binding of other DNA-binding proteins. However, since there are no covalent bonds between histones and DNA, cells use enzymes called chromatin remodelers and chaperones to dynamically regulate this interaction: nucleosomes can be moved, inserted or removed from the DNA molecule (3). Another way of regulating the interaction between nucleosomes and DNA is to post-translationally modify histone tails. Acetylation and phosphorylation of specific residues, for example, can change the highly basic nature of histones, reducing their affinity to the negatively charged DNA backbone (4). Similarly to nucleosome positioning, also the covalent modifications introduced by histone-modifying enzymes, "writers", are dynamic and can be reversed by other enzymes identified as "erasers" (e.g. histone acetylases and de-acetylases). Combinations of histone post-translational

modifications, the so called “histone code” postulated more than 15 years ago, are recognized by proteins called “readers” that can directly or indirectly mediate fundamental processes in gene activation/repression, DNA replication and repair (5-9).

In many, but not all, eukaryotic organisms there are enzymes called DNA methyltransferases able to attach a methyl group on cytosine residues on the DNA (10-12). This modified base is called 5-methylcytosine (5meC) and is regarded as the “fifth base”. The methyl group is projected in the major groove of the double helix and does not interfere with the pairing properties of cytosines. Similarly to histone post-translational modifications, this mark can be recognized by readers, referred to as methyl-DNA binding proteins (MBDs), and removed by erasers (7).

The term Epigenetics (the Greek prefix epi- means “above”) was coined to describe heritable changes in gene activity not accompanied by changes in DNA sequence (13). However, now the term is loosely used to describe mechanisms involved in gene regulation that do not involve alteration in the nucleotide sequence, regardless of its heritable nature (14). Recent improvements in sequencing technologies have played a pivotal role in the study of Epigenetics. The second generation of sequencing technologies, known as Next-Generation Sequencing (NGS), relies on highly multiplexed reactions, increasing the throughput and decreasing the cost of sequencing per base (15, 16). Several techniques have now been adapted to NGS approaches, making it possible to test not just single loci, but extend the analysis to the entire genome (17). An example of such a technology is ChIP-seq (Chromatin Immunoprecipitation coupled to sequencing), which is based on the sequencing of all the DNA fragments associated with a cross-linked protein of interest, provided that an antibody specifically recognizing the target exists (18). Similar affinity-based techniques can be applied to map 5meC: MeDIP-seq (Methyl DNA Immuno-Precipitation) or MBD-seq (Methyl DNA Binding Domain). These techniques are able to enrich for fragments containing 5meC, however they do not have single-nucleotide resolution. Although real-time sequencing can provide information on modified bases, the gold standard for

single-nucleotide resolution of 5meC remains the utilization of sodium bisulfite (BS) (19). Treatment of denatured genomic DNA with sodium bisulfite leads to deamination only of un-methylated cytosines to deoxy-uracil, which has identical base pairing properties of thymine. By contrast, 5meC residues are not efficiently converted by BS. The comparison of the BS-treated with an untreated DNA (also called reference) can reveal what cytosines are methylated in the original DNA sample. Several methods developed to test single or few loci rely on PCR: Methylation-Specific PCR (MSP) and Bisulfite Sequencing PCR (BSP). MSP is based on the design of primers that can only anneal to a specific status of a target cytosine (methylated or un-methylated) after bisulfite treatment, while BSP requires the design of primers that can amplify the region of interest independently of the methylation status of target cytosines. If the readout of the first method can be the presence of a discrete PCR band on a gel or real-time quantification, the second technique requires sequencing of PCR products or cloned PCR products. Although these two methods can produce single nucleotide information, they are not suitable to test the methylation status of all the cytosines in the entire genome. For this, a technique called whole-genome bisulfite sequencing (WGBS) is used instead. This method relies on the high-throughput capabilities of next generation sequencing methods, and is able to measure the methylation status of all the cytosines in a genome at single-nucleotide resolution. WGBS has been used to study the distribution of 5meC in several organisms: DNA methylation is present in mammals, plants, many fungi and animals, with the notable exception of important model organisms such as *Drosophila melanogaster*, *Caenorhabditis elegans* and *Saccharomyces cerevisiae*. 5meC is found at different contexts: symmetric (CpG and CpHpG, where H = A,T,C) or asymmetric (CpHpH) sequences. In plants cytosines in all three contexts are methylated, while the majority of mammalian 5meC is restricted to CpG dinucleotides (20). In mammals, CpG dinucleotides are depleted throughout the genome (less than 1%, compared to the expected 4.4%), due to spontaneous deamination. However, they are clustered in so-called CpG islands, regions dense in CpGs dinucleotides. CpG islands are present in most

promoters (60-70%) of mammalian genes, enriched for highly expressed housekeeping genes and they are generally depleted in methylcytosines. The remaining promoters show a dynamic pattern of 5meC and are usually tissue-specific. The rest of the genome, which has low CpG density, is usually methylated (21). DNA methylation is regarded as a repressive mark, present at promoters of silenced genes and on repetitive elements. However, 5meC is also enriched in gene bodies of highly transcribed genes.

Mapping the distribution of methylcytosines genome-wide remains expensive for large-scale studies. Numerous methods have been developed to reduce the cost per sample and maintain single-nucleotide resolution. One of them, called Reduced-Representation Bisulfite Sequencing (or RRBS), utilizes a restriction enzyme (*Msp*I) to fragment the genome. DNA fragments smaller than 400bp are enriched in promoters, CpG islands and nearby CpG island shores (22-24). The main advantage of RRBS is that it can capture an informative fraction of the genome for a reduced cost per sample. This approach is particularly helpful for large-scale genomic projects. An example of such a project is described in Chapter 1 (25). RRBS was used to measure 5meC levels of liver DNA of 90 inbred mouse strains and it was shown to be associated with complex molecular and metabolic traits (such as bone mineral density, insulin resistance and protein/metabolite levels) through Epigenome-Wide Association Studies (EWAS). The study suggests that the distribution of DNA methylation is controlled by genetic variants, but since it's more dynamic than the "static" genomic sequence, it can provide additional information about cell type composition of the tissue tested in response to external stimuli or disease states. The findings that many associations were found only with EWAS and not with genetic variants highlight the importance of 5meC profiling in quantitative trait modeling.

Although RRBS enriches for CpG-rich regions, it still interrogates a small fraction of them (6-12% of all the CpGs present in the human genome). WGBS can accurately define the methylation status of most of the cytosines in the genome, but is expensive for large genomes (e.g. mammalian genomes). Moreover, most of the CpG sites are methylated in the genome and

80% of them do not vary between different cell types. In order to overcome these limitations, other approaches have been developed. One such method, called MRE-seq (Methylation-sensitive Restriction Enzyme), relies on the sequencing of small fragments produced by the digestion of the genome by three methylation-sensitive restriction enzymes (*Hpa*II, *Hin*6I and *Aci*I) (26). In contrast to the methylation-insensitive restriction enzyme *Msp*I, these three enzymes cut the DNA only if the cytosines in their recognition site are unmethylated. The advantage of this technique is that it focuses sequencing on the smaller fraction of the genome that is unmethylated, reducing the costs. However, the resolution is only limited to the restriction sites, since the methylation status of cytosines within each fragment is unknown. One approach used MeDIP (which enriches for methylated fragments) to complement the data obtained from MRE-seq (27). Even if this combined approach improves the methylation estimates, it increases the cost per sample associated with library preparation and sequencing and it doesn't have single-nucleotide resolution. We, therefore, developed a method that has comparable costs to RRBS, has single-nucleotide resolution (as RRBS and WGBS), but provides information of differential methylation over a fraction of the genome comparable to WGBS. Chapter 2 is a manuscript in preparation describing the approach, called MRE-BS: Methylation-sensitive restriction Enzyme Bisulfite Sequencing. The addition of a bisulfite treatment step, not only allows us to use the coverage information as in MRE-seq, but also provides us single nucleotide methylation estimates for a small fraction of the sites. A multiple regression model was built to combine these two features allowing us to estimate differential methylation (only through read coverage) across 60% of the genome and improved estimates for a small fraction of the genome (1.5%) when the coverage was sufficiently high for accurate methylation calling (through bisulfite sequencing). This approach will be useful for projects comparing several samples with large genomes.

Methylation of cytosines is a covalent modification introduced by a class of enzymes called DNA methyltransferases (DNMTs). DNA methyltransferases are found in both

Prokaryotes and Eukaryotes, even if missing in some species (11, 28-30). DNA methylation systems in Prokaryotes have evolved to prevent the degradation of the host DNA by restriction enzymes, used as a defense mechanism from invading nucleic acids. However, in Eukaryotes, DNA methylation plays an important role in the control of several cellular processes, such as imprinting and X-chromosome inactivation (in mammals), transposon repression (in fungi, plants and animals) and gene regulation. Even if some functions of DNA methylation are conserved among fungi, plants and animals, their respective DNA methyltransferases are different. In fungi DNA methylation targets repeated sequences in order to silence its expression. The two main studied organisms are *Ascobolus immersus*, which methylates and silence repetitive DNA (MIP: Methylation-Induced Pre-meiotically), and *Neurospora crassa*, which targets repeats with a DNA methylation-linked process that induces mutation at the nucleotide level (RIP: Repeat-Induced Point mutation) (31, 32). Likewise, both plants and mammals methylate repetitive sequences to repress their activity, however, there are many differences in the DNMTs. Eukaryotic DNMTs can be divided in two groups: *de novo* DNA methyltransferases (which act on an unmethylated DNA substrate) and maintenance DNA methyltransferases (that prefer a hemi-methylated DNA substrate). *De novo* DNMTs establish methylation patterns by transferring a methyl group from the donor S-Adenosyl Methionine (SAM) to an unmethylated cytosine. Both MET1 in *Arabidopsis* (model organisms for DNA methylation in flowering plants) and DNMT1 in mammals function as DNA maintenance methyltransferases on symmetric CpG sites. They function by recognizing and methylating a hemi-methylated CpG dinucleotide and they perpetuate the pattern during DNA replication for the entire life of the organism, unless an external signal interferes with it. In contrast to mammals, plants have evolved additional mechanisms of DNA methylation maintenance for non-CpG sites (through CMT3 and CMT2), which are based on a reinforcing-loop involving methylation of lysine 9 of histone 3 (H3K9me) (30). *De novo* methylation in plants depends on DRM2 (Domains Rearranged Methyltransferase 2) and requires other components of RNA

interference, two RNA polymerases (RNA pol IV and RNA pol V), chromatin remodeling factors and several other proteins. This pathway is known as RNA-dependent DNA Methylation (RdDM) and is not present in animals. Mammals have two active *de novo* DNA methyltransferases (DNMT3a and DNMT3b) and one truncated inactive version (DNMT3L) (11). *De novo* DNMTs in mammals are expressed at higher levels in undifferentiated and germ cells precursors compared to somatic cells. This is in agreement with the fact that DNMT3s establish DNA methylation patterns during development after global DNA demethylation events and the signal is then faithfully maintained by DNMT1. DNMT1 is recruited to its target sites through interaction of its N-terminus to other proteins, such as PCNA (targeting to replication forks) and UHRF1 (which recognizes hemi-methylated DNA) (33). Similarly to DNMT1, active *de novo* DNMTs possess a catalytic domain at the C-terminus. A defective methyltransferase domain is present in the inactive DNMT3L protein. However, DNMT3L, along with the two active *de novo* DNMTs, possess an ATRX-DNMT3-DNMT3L (ADD) domain, that interacts specifically with unmethylated histone 3 lysine 4 residues (H3K4me0). Moreover, the binding to H3K4me0 has been shown to disrupt the auto-inhibitory activity of the N-terminus portion of DNMT3a. Di- or tri-methylation of H3K4 (H3K4me2 and H3K4me3) is sufficient to disrupt the interaction of the ADD domain with the H3 histone tail. A third domain is present on DNMT3a and DNMT3b and it consists of a proline-tryptophan motif (PWWP) (33). Previous evidence showed that the PWWP domain of DNMT3a interacts with tri-methylated histone 3 lysine 36 (H3K36me3) *in vitro*, but since additional interactions with DNA and other histone modifications have been reported, further characterization of the *de novo* DNMTs is needed (34). In order to address this question, we introduced one of the two mammalian *de novo* DNA methyltransferases (DNMT3b from mouse) in the budding yeast *Saccharomyces cerevisiae*, which doesn't have endogenous DNA methylation. This system has several advantages: it has histone sequences and many modifications are conserved with higher Eukaryotes, it can be easily manipulated, and its genome is small, reducing the costs associated with NGS approaches.

Moreover, the absence of 5meC in yeast mimics the status of the mammalian genome that undergoes global demethylation during development (35, 36). Chapter 3 is a published paper showing how DNMT3b is guided *in vivo* by chromatin cues. Briefly DNMT3b prefers linker to nucleosomal DNA, and methylation occurs at H3K4me3-negative and H3K36me3-positive regions (37). This pattern of unmethylated transcriptional start sites (TSSs, where H3K4me3 is present) and methylated gene bodies (rich in H3K36me3) is consistent with the distribution found in mammals. Moreover, the deletion of Set1, the enzyme responsible for the deposition of H3K4me3, and Set2, which tri-methylates H3K36, increases DNMT3b-dependent DNA methylation over TSSs and decreases 5meC levels in gene bodies, respectively. The distribution of H3K4me3 and H3K36me3 in embryonic germ cells predicts the regions of the genome that undergo *de novo* DNA methylation, suggesting that the same mechanism is targeting DNMT3b in mammals (38).

We conclude that the chromatin marks themselves are responsible for targeting DNA methylation. In agreement with that, another study showed that in mouse embryonic stem cells gene body DNA methylation is dependent on H3K36me3 and a functional PWWP domain of DNMT3b (39). Although DNA methylation in promoters has been associated for a long time with transcriptional repression, the function of gene-body methylation remained largely unknown. Only recently, it has been shown that DNMT3b-mediated gene-body methylation prevents intragenic cryptic transcription (40). These functions are mediated by 5meC-recognizing proteins (41).

5meC is not only present in DNA, but also in RNA molecules. RNA modifications have been discovered several years ago, initially in abundant non-coding RNAs, such as ribosomal RNA (rRNAs), transfer RNA (tRNAs) and small nuclear RNA (snRNAs). As of today, more than 150 modified residues have been identified in RNA molecules, such as N6-methyladenosine (m6A), N6,2'-O-dimethyladenosine (m6Am), 5-methylcytidine (5meC), 5-hydroxymethilcytidine (5hmeC), inosine (I), pseudouridine (Ψ) and N1-methyladenosine (m1A)

(42, 43). The first methods used to identify the modified residues were based on chemical or enzymatic digestions coupled with chromatography (44). However, this approach suffers from the inability to assign a specific modification to its sequence context. A breakthrough in this field has been the ability to map RNA modifications genome-wide using NGS approaches, a field called “Epitranscriptomics” (45, 46). The best-characterized modification in Eukaryotic coding transcripts is m6A, since it is the most abundant and several proteins involved in its pathway have been described (47, 48). Similarly to DNA, RNA modifications are introduced post-transcriptionally by RNA methyltransferase enzymes (RMT). Mutations in the RMTs have been linked to complex disorders, such as developmental defects, mental retardation and cancer (49). Several enzymes have been shown to methylate cytosine residues on the RNA, among them Nsun2, Nsun7 and DNMT2 (50). Since this modification is present on tRNAs and rRNAs, making it difficult to work with 5meC-RMTs mutants, little is known about the function of this modification in coding transcripts. Several approaches have been developed to map 5meC on RNA. RNA ImmunoPrecipitation (5meC-RIP) is based on antibody specific to the methylated cytosine, but it doesn’t have single nucleotide resolution. Three additional methods have single nucleotide resolution: bisulfite-RNAseq (BS-RNAseq, similar to the technique used for DNA), aza-IP (based on the formation of a covalent intermediate between RMTs and the cytosine-analog 5-azaC) and miCLIP (based on the expression of a mutated NSun2 RMT, followed by IP) (47). The only technique that doesn’t require expression of transgenes or media analog supplementation is the treatment of RNA with bisulfite, which causes deamination of unmodified cytosines to uracils. We, therefore, developed a method to detect 5meC in an enriched fraction of RNA, polyA-tailed or rRNA-depleted, based on the treatment with sodium bisulfite followed by reverse transcription and library preparation, compatible with the Illumina platform (Chapter 4).

Despite the ability to map methylation of cytosines in RNA at single-base resolution, there are disadvantages associated with BS-RNAseq that can lead to the production of false positives.

First, cytosines can be resistant to the bisulfite conversion if located in double stranded RNA. Second, since the treatment of RNA with bisulfite is carried out in milder conditions than the one on DNA, it can cause incomplete conversion. Third, 5meC is not the only modification preventing the BS-mediated conversion (i.e. hydroxyl-methylcytosine). Different studies have shown that 5meC identified sites cannot be reproduced or confirmed by different methods (51-53). In Chapter 4, a bioinformatics pipeline, adapted from the one used to detect 5meC in BS-treated DNA (54), identified about 500-1000 methylated sites for each of two different mouse genetic background (C57 and DBA). The C57 strain shows global levels of RNA methylation higher than DBA, in accordance with higher expression levels of two known 5meC-RMTs: Nsun7 and DNMT2. Interestingly, many methylated sites identified by our method are associated with stable secondary structures predicted with bioinformatics software (mfold) or experimentally measured with the Parallel Analysis of RNA Structure (PARS) in human cell lines (55, 56). Two different hypotheses can explain this finding. One possible explanation is that stretches of structured RNA are not converted efficiently, resulting in higher 5meC false positives. Alternatively RMTs don't have inherent sequence specificity, but they recognize secondary structures instead. This also would be in agreement with the hypothesis that many of the 5meC sites are the result of the aspecific activity of 5meC RMTs, which usually act on non-coding RNAs, a class rich in secondary and tertiary structures. Different questions still remain to be answered. For example what fraction of the identified 5meC sites on RNA are false positives and what is the biological significance of RNA methylation. In order to address these questions, improvements of the bisulfite-treatment conditions, validation with orthogonal approaches and more in depth study of the targeting and activity of RNA methyltransferases are needed.

REFERENCES

1. Spitz F & Furlong EE (2012) Transcription factors: from enhancer binding to developmental control. *Nature reviews. Genetics* 13(9):613-626.

2. Frietze S & Farnham PJ (2011) Transcription factor effector domains. *Sub-cellular biochemistry* 52:261-277.

3. Langst G & Manelyte L (2015) Chromatin Remodelers: From Function to Dysfunction. *Genes* 6(2):299-324.

4. Bowman GD & Poirier MG (2015) Post-translational modifications of histones that influence nucleosome dynamics. *Chem Rev* 115(6):2274-2295.

5. Jenuwein T & Allis CD (2001) Translating the histone code. *Science (New York, N.Y.)* 293(5532):1074-1080.

6. Patel DJ & Wang Z (2013) Readout of epigenetic modifications. *Annual review of biochemistry* 82:81-118.

7. Rothbart SB & Strahl BD (2014) Interpreting the language of histone and DNA modifications. *Biochimica et biophysica acta* 1839(8):627-643.

8. Strahl BD & Allis CD (2000) The language of covalent histone modifications. *Nature* 403(6765):41-45.

9. Yun M, Wu J, Workman JL, & Li B (2011) Readers of histone modifications. *Cell research* 21(4):564-578.

10. Feng S*, et al.* (2010) Conservation and divergence of methylation patterning in plants and animals. *Proceedings of the National Academy of Sciences of the United States of America* 107(19):8689-8694.

11. Jurkowska RZ & Jeltsch A (2016) Enzymology of Mammalian DNA Methyltransferases. *Advances in experimental medicine and biology* 945:87-122.

12. Zemach A, McDaniel IE, Silva P, & Zilberman D (2010) Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science (New York, N.Y.)* 328(5980):916-919.

13. Ptashne M (2007) On the use of the word 'epigenetic'. *Curr Biol* 17(7):R233-236.

14. Berger SL, Kouzarides T, Shiekhattar R, & Shilatifard A (2009) An operational definition of epigenetics. *Genes & development* 23(7):781-783.

15. Goodwin S, McPherson JD, & McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. *Nature reviews. Genetics* 17(6):333-351.

16. Morey M, *et al.* (2013) A glimpse into past, present, and future DNA sequencing. *Molecular genetics and metabolism* 110(1-2):3-24.

17. Reuter JA, Spacek DV, & Snyder MP (2015) High-throughput sequencing technologies. *Molecular cell* 58(4):586-597.

18. Pellegrini M & Ferrari R (2012) Epigenetic analysis: ChIP-chip and ChIP-seq. *Methods in molecular biology* 802:377-387.

19. Yong WS, Hsu FM, & Chen PY (2016) Profiling genome-wide DNA methylation. *Epigenetics & chromatin* 9:26.

20. Law JA & Jacobsen SE (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature reviews. Genetics* 11(3):204-220.

21. Li E & Zhang Y (2014) DNA methylation in mammals. *Cold Spring Harbor perspectives in biology* 6(5):a019133.

22. Boyle P, *et al.* (2012) Gel-free multiplexed reduced representation bisulfite sequencing for large-scale DNA methylation profiling. *Genome biology* 13(10):R92.

23. Gu H, *et al.* (2011) Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nature protocols* 6(4):468-481.

24. Meissner A, *et al.* (2005) Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic acids research* 33(18):5868-5877.

25. Orozco LD, *et al.* (2015) Epigenome-wide association of liver methylation patterns and complex metabolic traits in mice. *Cell metabolism* 21(6):905-917.

26. Maunakea AK, *et al.* (2010) Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* 466(7303):253-257.

27. Li D, Zhang B, Xing X, & Wang T (2015) Combining MeDIP-seq and MRE-seq to investigate genome-wide CpG methylation. *Methods* 72:29-40.

28. Casadesus J (2016) Bacterial DNA Methylation and Methylomes. *Advances in experimental medicine and biology* 945:35-61.

29. Cheng X (1995) Structure and function of DNA methyltransferases. *Annual review of biophysics and biomolecular structure* 24:293-318.

30. Du J (2016) Structure and Mechanism of Plant DNA Methyltransferases. *Advances in experimental medicine and biology* 945:173-192.

31. Barra JL, Holmes AM, Gregoire A, Rossignol JL, & Faugeron G (2005) Novel relationships among DNA methylation, histone modifications and gene expression in Ascobolus. *Molecular microbiology* 57(1):180-195.

32. Selker EU*, et al.* (2003) The methylated component of the Neurospora crassa genome. *Nature* 422(6934):893-897.

33. Tajima S, Suetake I, Takeshita K, Nakagawa A, & Kimura H (2016) Domain Structure of the Dnmt1, Dnmt3a, and Dnmt3b DNA Methyltransferases. *Advances in experimental medicine and biology* 945:63-86.

34. Rondelet G, Dal Maso T, Willems L, & Wouters J (2016) Structural basis for recognition of histone H3K36me3 nucleosome by human de novo DNA methyltransferases 3A and 3B. *Journal of structural biology* 194(3):357-367.

35. Ambrosi C, Manzo M, & Baubec T (2017) Dynamics and Context-Dependent Roles of DNA Methylation. *Journal of molecular biology*.

36. Stewart KR, Veselovska L, & Kelsey G (2016) Establishment and functions of DNA methylation in the germline. *Epigenomics* 8(10):1399-1413.

37. Morselli M*, et al.* (2015) In vivo targeting of de novo DNA methylation by histone modifications in yeast and mouse. *eLife* 4:e06205.

38. Du J, Johnson LM, Jacobsen SE, & Patel DJ (2015) DNA methylation pathways and their crosstalk with histone methylation. *Nature reviews. Molecular cell biology* 16(9):519-532.

39. Baubec T*, et al.* (2015) Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. *Nature* 520(7546):243-247.

40. Neri F*, et al.* (2017) Intragenic DNA methylation prevents spurious transcription initiation. *Nature* 543(7643):72-77.

41. Baubec T, Ivanek R, Lienert F, & Schubeler D (2013) Methylation-dependent and -independent genomic targeting principles of the MBD protein family. *Cell* 153(2):480-492.

42. Harcourt EM, Kietrys AM, & Kool ET (2017) Chemical and structural effects of base modifications in messenger RNA. *Nature* 541(7637):339-346.

43. Helm M & Motorin Y (2017) Detecting RNA modifications in the epitranscriptome: predict and validate. *Nature reviews. Genetics* 18(5):275-291.

44. Grosjean H, Keith G, & Droogmans L (2004) Detection and quantification of modified nucleotides in RNA using thin-layer chromatography. *Methods in molecular biology* 265:357-391.

45. Frye M, Jaffrey SR, Pan T, Rechavi G, & Suzuki T (2016) RNA modifications: what have we learned and where are we headed? *Nature reviews. Genetics* 17(6):365-372.

46. Saletore Y*, et al.* (2012) The birth of the Epitranscriptome: deciphering the function of RNA modifications. *Genome biology* 13(10):175.

47. Li X, Xiong X, & Yi C (2016) Epitranscriptome sequencing technologies: decoding RNA modifications. *Nature methods* 14(1):23-31.

48. Slobodin B*, et al.* (2017) Transcription Impacts the Efficiency of mRNA Translation via Co-transcriptional N6-adenosine Methylation. *Cell* 169(2):326-337 e312.

49. Khoddami V, Yerra A, & Cairns BR (2015) Experimental Approaches for Target Profiling of RNA Cytosine Methyltransferases. *Methods in enzymology* 560:273-296.

50. Schaefer M (2015) RNA 5-Methylcytosine Analysis by Bisulfite Sequencing. *Methods in enzymology* 560:297-329.

51. Amort T*, et al.* (2017) Distinct 5-methylcytosine profiles in poly(A) RNA from mouse embryonic stem cells and brain. *Genome biology* 18(1):1.

52. Hussain S, Aleksic J, Blanco S, Dietmann S, & Frye M (2013) Characterizing 5-methylcytosine in the mammalian epitranscriptome. *Genome biology* 14(11):215.

53. Jeltsch A*, et al.* (2016) Mechanism and biological role of Dnmt2 in Nucleic Acid Methylation. *RNA biology*:1-16.

54. Guo W*, et al.* (2013) BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. *BMC genomics* 14:774.

55. Kertesz M*, et al.* (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature* 467(7311):103-107.

56. Wan Y*, et al.* (2014) Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* 505(7485):706-709.

CHAPTER 1:

Epigenome-Wide Association of Liver Methylation Patterns and Complex Metabolic Traits in Mice

Orozco LD*, et al.* (2015). *Cell metabolism* 21(6):905-917.

Epigenome-Wide Association of Liver Methylation Patterns and Complex Metabolic Traits in Mice

Luz D. Orozco,[1] Marco Morselli,[1] Liudmilla Rubbi,[1] Weilong Guo,[2] James Go,[3] Huwenbo Shi,[4] David Lopez,[1] Nicholas A. Furlotte,[5] Brian J. Bennett,[6] Charles R. Farber,[7] Anatole Ghazalpour,[8] Michael Q. Zhang,[2] Renata Bahous,[9] Rima Rozen,[9] Aldons J. Lusis,[8] and Matteo Pellegrini[1,*]

[1]Department of Molecular, Cell and Developmental Biology, University of California, Los Angeles, Los Angeles, CA 90095, USA
[2]Center for Synthetic & Systems Biology, TNLIST, Tsinghua University, Beijing 100084, China
[3]Department of Biology, California State University, Northridge, Northridge, CA 91330, USA
[4]Department of Bioinformatics
[5]Department of Computer Science
University of California, Los Angeles, Los Angeles, CA 90095, USA
[6]Department of Genetics, University of North Carolina, Kannapolis, NC 28081, USA
[7]Department of Medicine, University of Virginia, Charlottesville, VA 22904, USA
[8]Department of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA
[9]Departments of Human Genetics and Pediatrics, McGill University, Montreal, QC 514, Canada

*Correspondence: matteop@mcdb.ucla.edu
http://dx.doi.org/10.1016/j.cmet.2015.04.025

SUMMARY

Heritable epigenetic factors can contribute to complex disease etiology. Here we examine the contribution of DNA methylation to complex traits that are precursors to heart disease, diabetes, and osteoporosis. We profiled DNA methylation in the liver using bisulfite sequencing in 90 mouse inbred strains, genome-wide expression levels, proteomics, metabolomics, and 68 clinical traits and performed epigenome-wide association studies (EWAS). We found associations with numerous clinical traits including bone density, insulin resistance, expression, and protein and metabolite levels. A large proportion of associations were unique to EWAS and were not identified using GWAS. Methylation levels were regulated by genetics largely in *cis*, but we also found evidence of *trans* regulation, and we demonstrate that genetic variation in the methionine synthase reductase gene *Mtrr* affects methylation of hundreds of CpGs throughout the genome. Our results indicate that natural variation in methylation levels contributes to the etiology of complex clinical traits.

INTRODUCTION

Methylation of DNA cytosine bases is evolutionarily conserved in multiple species from plants to humans. In mammalian species, DNA methylation plays an important role in imprinting, X chromosome inactivation, cell differentiation, gene silencing, and regulation of gene expression. Similar to genetic variation, epigenetic modifications are variable between individuals and regulated by genetics (Orozco et al., 2014). Methylation QTL (metQTL) studies in human adipose tissue found that 28% of CpGs were associated with nearby SNPs (Grundberg et al., 2013). DNA methylation is variable between inbred strains in *Arabidopsis*, rice, and mice and in human populations. Strain-specific methylation patterns were maintained across generations in mouse strains, and twin studies in humans have shown a higher concordance of methylation patterns in monozygotic twins relative to dizygotic twins (Gordon et al., 2012; McRae et al., 2014), suggesting that DNA methylation is under genetic control.

Both human populations and mouse strains show variation in multifactorial traits like heart disease and osteoporosis. In the past decade, genome-wide association studies (GWAS) have identified hundreds of genetic variants influencing clinical traits (Welter et al., 2014). DNA methylation has also been associated with gene expression (Bell et al., 2011; Grundberg et al., 2013) and complex traits including cancer (Shenker et al., 2013; Xu et al., 2013), aging (Heyn et al., 2013; Horvath, 2013), multiple sclerosis (Huynh et al., 2014), rheumatoid arthritis (Liu et al., 2013), and obesity in humans (Dick et al., 2014). However, although DNA methylation is influenced by genetics and could account for part of the heritability of clinical traits, epigenetic variation has typically not been considered in GWAS for complex traits.

In this study, we performed epigenome-wide association studies (EWAS) to determine the contribution of DNA methylation to complex clinical traits related to heart disease, diabetes, obesity, and osteoporosis. Our study integrates systems genetics data that includes SNP genotypes, DNA methylation bisulfite sequencing data, genome-wide gene expression, proteomics, metabolomics, and clinical phenotypes. Our results reveal a large number of associations between DNA methylation variants and clinical or molecular traits, many of which we could not identify using traditional GWAS. We explored the contribution of genetics to DNA methylation patterns and found that 52% of highly variable CpGs were under genetic control. The narrow sense heritability for CpG methylation levels was on average 27% and was 60% for highly variable CpGs. We also present evidence of common genetic variation affecting DNA methylation patterns in *trans* and experimentally validate the role of *Mtrr* in regulating methylation levels of CpGs across the genome.

RESULTS

Data

We constructed reduced representation bisulfite sequencing (RRBS) libraries using liver genomic DNA from 16-week-old male mice using a previously described protocol (Smith et al., 2009), corresponding to 90 mouse inbred strains from the Hybrid Mouse Diversity Panel (HMDP) (Bennett et al., 2010). We sequenced the libraries using the Illumina HiSeq platform and obtained an average of 90 ± 11 million reads per sample, then aligned the data to the mouse genome using BS-Seeker2 (Guo et al., 2013) for an average of 41 ± 7 million uniquely aligned reads per sample (Figure S1A). This corresponded to 46% mappability and 48× coverage per sample on average (Figure S1B). We filtered the cytosines based on 10× or more coverage for a total of 11,520,175 cytosines present in at least 90% of the samples, of which 2,047,165 were CG, 2,737,475 were CHG, and 6,735,535 were in CHH context. The mouse genome contains 21.3 million CpGs, and we observed approximately 2 million (9.6%) of all CpGs using RRBS.

Global methylation levels in the adult mouse livers were 44% ± 1% for CpG cytosines, 1.1% ± 0.4% for CHG, and 0.8% ± 0.4% for CHH cytosines, where H is any base other than G (Figure S1C). Since non-CG methylation was too low to be studied in these samples (Figures S1C and S1E), we focused our analyses on CG cytosines only. We defined a set of 360,324 *Variable* CpGs, which showed a 50% absolute change (delta) in methylation levels in at least one sample. We further identified a set of 22,227 *Hypervariable* CpGs, which showed 50% or higher methylation delta, relative to the median methylation level of the CpG in 5 or more samples. An example of a *Variable* and a *Hypervariable* CpG can be found in Figures S2A and S2B. We excluded 6,993 CpGs that were also SNPs in the mouse strains, since the changes in methylation observed correspond to the loss of a CpG in strains carrying the SNP.

The liver is one of the main tissues involved in energy metabolism. Because of its roles in carbohydrate and fat metabolism, the liver has a significant impact on clinical phenotypes such as plasma glucose, cholesterol and lipid levels, body weight, adiposity, and atherosclerosis. It would also be important to consider methylation levels in other metabolically relevant tissues such as adipose, muscle, pancreas, and intestine in future studies. For the same mouse strains, we measured 68 clinical traits including atherosclerosis, diabetes, obesity, osteoporosis, and blood-cell-related traits, as well as genome-wide expression levels in the liver (Bennett et al., 2010) using Affymetrix arrays. We obtained liver proteomics from 1,543 peptides measured by liquid chromatography-mass spectrometry (Ghazalpour et al., 2011). We also profiled 260 liver and plasma metabolites using mass spectrometry, comprising eight classes of molecules including lipids, carbohydrates, amino acids, peptides, xenobiotics, vitamins, cofactors, and nucleotides (Ghazalpour et al., 2014).

DNA Methylation Has Lower Correlation in *Cis* than SNPs

Correlations between pairs of alleles that are near each other on a chromosome, or linkage disequilibrium (LD), can result in large genomic blocks that contain multiple genes. A given association may have a few or dozens of candidate genes depending on the level of LD at that locus. We were interested in determining the correlation in pairwise CpG methylation levels, and hence the level of resolution we could achieve using CpGs in our association studies. We determined pairwise correlations of CpGs at different distances from each other. For example, we took CpGs separated by 100 kb or less and estimated the correlation between the CpG methylation levels in the HMDP strains. We then estimated the average correlation between all pairs of CpGs in the genome at that distance from each other and repeated this estimate at various distances. We compared this to the level of correlation in SNPs (LD) that we had previously calculated for the same HMDP strains (Bennett et al., 2010). Pairwise correlations in CpG methylation levels for a locus on chromosome 1 are shown in Figure 1A for *Variable* CpGs, *Hypervariable* CpGs (Figure 1B), and SNPs (Figure 1C). Methylation levels in mouse strains for a sample locus in chromosome 1 are shown in Figure 1D, where methylation levels vary between 0% and 100%. Correlations plots for whole chromosomes can be found on Figures S3D–S3I.

At the genome-wide level, we found that the distance-dependent correlation between CpGs was lower than that of SNPs. The average correlation across the genome between CpGs within 100 kb was $r^2 = 0.06$ for *Variable* CpGs and $r^2 = 0.43$ for *Hypervariable* CpGs, and at 2 Mb, the average correlation was $r^2 = 0.03$ for *Variable* CpGs and $r^2 = 0.17$ for *Hypervariable* CpGs. In contrast, the average correlation across the genome was $r^2 = 0.88$ for SNPs at 100 kb, dropping to $r^2 = 0.49$ at 2 Mb. The genome-wide average of pairwise correlations is shown on Figure 1E, for CpGs or SNPs at various distances from each other in increasing 100 kb bins. Since methylation levels in mammals are bimodal for a large proportion of CpGs (Figure S1D), we examined pairwise correlations between CpGs with low or high methylation levels. We found that *Hypervariable* CpGs with low methylation levels were generally more highly correlated with nearby CpGs than CpGs with high methylation levels. For example, the average correlation between *Hypervariable* CpGs at 100 kb was $r^2 = 0.68$ for CpGs with 0%–20% methylation levels, and $r^2 = 0.58$ for CpGs with 80%–100% methylation. The average correlation was $r^2 = 0.32$ for CpGs with 0%–20% methylation levels, and $r^2 = 0.18$ for CpGs with 80%–100% methylation, at a distance of 2 Mb. We observed no differences in pairwise correlation levels between low and high methylated *Variable* CpGs (Figure 1F).

Natural Variation in DNA Methylation Is Associated with Complex Traits

To determine the association of epigenetic variation with complex clinical and molecular traits, we performed EWAS between CpG methylation levels and (1) 68 clinical traits, including plasma cholesterol, fatty acids, glucose and insulin, body weight, adiposity, blood cell counts, and bone mineral density phenotypes; (2) 260 plasma and liver metabolites; (3) protein levels from 1,543 peptides corresponding to 480 genes; and (4) genome-wide microarray expression levels corresponding to 12,980 genes (Figure 2). Similar to GWAS using SNPs, we used a linear mixed model (Kang et al., 2008) to determine associations between traits and CpGs and to correct for population structure. Since we used CpG methylation as the predictors

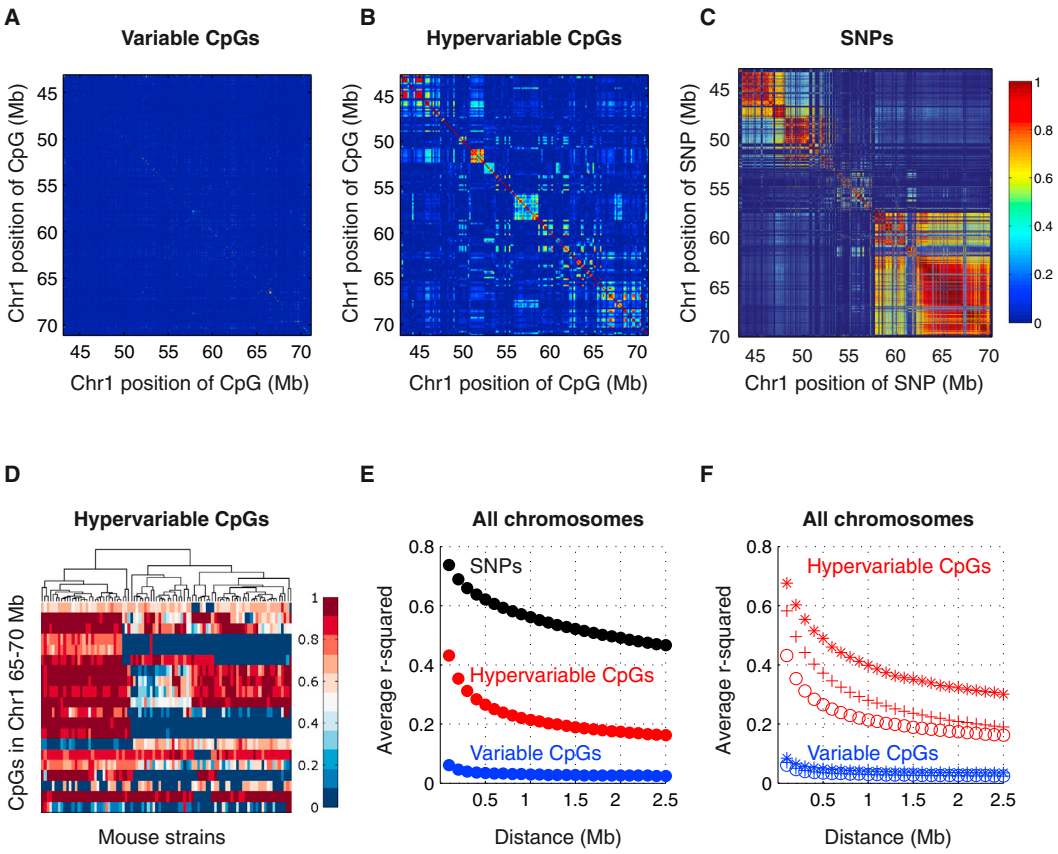


Figure 1. Methylation and SNP Correlations

(A–C) Correlations for a locus in chromosome 1 in (A) *Variable* CpG methylation, (B) *Hypervariable* CpG methylation, and (C) SNPs. The x and y axes denote the chromosome position, and the color represents the correlation (r^2) between CpGs, or SNPs.

(D) Methylation levels of *Hypervariable* CpGs for a representative locus, strains are on the x axes, CpGs are on the y axes, and the color represents percent methylation levels.

(E and F) Genome-wide average correlation between CpGs, or SNPs, at various distances. SNPs are shown in black; *Variable* CpGs are in blue; *Hypervariable* CpGs are in red. Each point is the average correlation at increasing 100 kb bins.

(E) Genome-wide average correlation between CpGs binned by their methylation level. Open circles (o) represent the average for all CpGs; asterisks (*) represent the average for CpGs with methylation levels between 0% and 20%, and plus symbols (+) represent the average for CpGs with methylation levels between 80% and 100%. See also Figure S3.

instead of SNPs, we also employed a methylation-based kinship matrix instead of a SNP kinship matrix in the model. We and others have shown that this approach corrects for false-positive associations due to population structure (Bennett et al., 2010) and potential tissue heterogeneity in the methylation data (Zou et al., 2014).

Each of the EWAS plots in Figure 2 summarizes associations between CpGs across the mouse genome and traits. Due to the large number of traits in the proteomics and gene expression data sets, only associations to *Hypervariable* CpGs are shown for these data sets. All associations shown are significant at the Bonferroni threshold (Table S1). In summary, we found that natural variation in CpG methylation was associated with numerous complex clinical and molecular traits. A table with the number of EWAS hits and the significance threshold used can be found Table S1, and tables with the individual associations we identified for all clinical and molecular traits can be downloaded from http://ewas.mcdb.ucla.edu/download.html. We found no evidence of inflation in our EWAS results (Figures S4A–S4D; Supplemental Information) and no evidence of macrophage contamination in our liver samples (Figures S4E and S4F).

EWAS Identifies Both Known and Novel Associations

We identified numerous associations between clinical traits and CpG methylation near genes known to influence those traits, including associations not identified using traditional GWAS. We previously performed GWAS for clinical traits and expression levels in the HMDP (Bennett et al., 2010). We also previously examined LD for SNPs in the HMDP and estimated an average resolution of 2 Mb (Bennett et al., 2010; Ghazalpour et al., 2011; Orozco et al., 2012), although LD blocks could be smaller or larger depending on the genomic region. We took the average 2 Mb resolution in the HMDP and called associations in 2 Mb bins, such that more than one association in a bin was

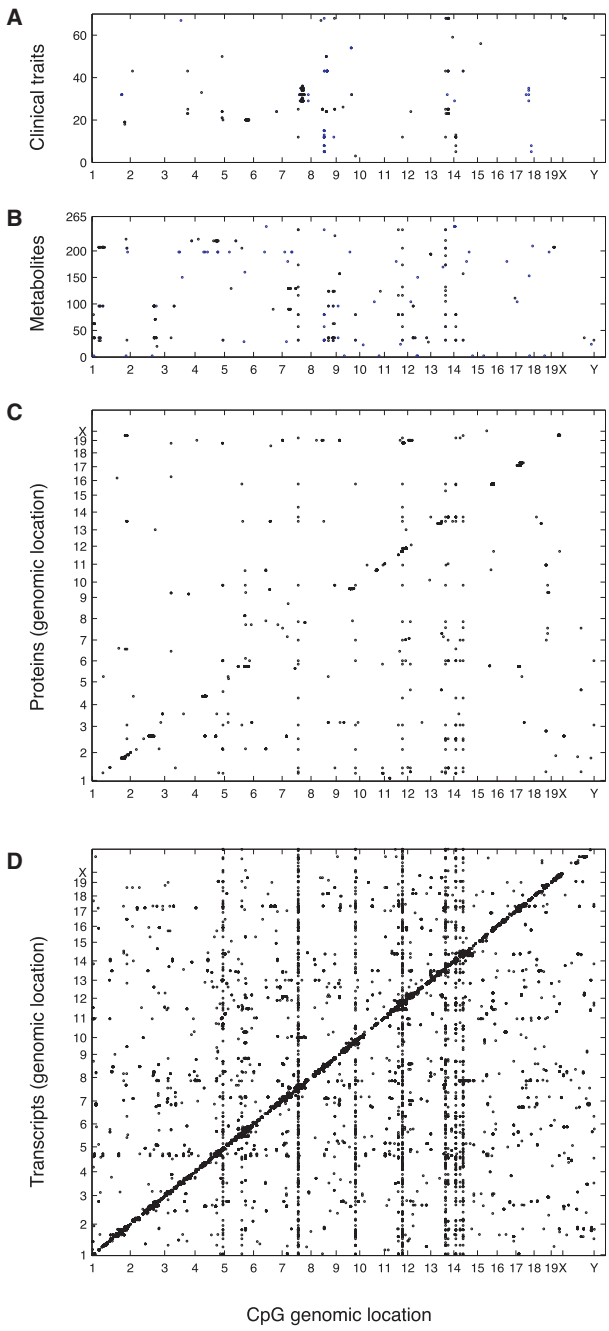


Figure 2. EWAS

(A–D) Association between CpG methylation and (A) clinical traits, (B) metabolites, (C) protein, and (D) gene expression levels. Each point is a significant EWAS at the corresponding Bonferroni thresholds across all CpGs and traits tested. The genomic position of CpGs is on the x axis, the y axis denote traits, and the position in the genome of the associated proteins and genes. Black points are EWAS hits for *Hypervariable* CpGs, and blue points are EWAS hits for *Variable* CpGs. For simplicity, only associations to *Hypervariable* CpGs are shown for the proteomics and gene expression data sets. See also Figure S4.

considered to be the same locus. To look for the overlap between GWAS and EWAS associations, we considered associations to be overlapping if they were within 2 Mb of each other. For clinical traits common to both studies, we identified 266 EWAS hits and 300 GWAS hits, and 41 were identified by both EWAS and GWAS where the associated loci were within 2 Mb of each other. We also observed that the overlap between EWAS and GWAS hits was much higher for expression *cis*-eQTL (77%) and protein (37%) *cis*-pQTL (Figure S5A). Clinical traits are typically more complex than molecular traits such as gene expression or protein levels. For increasingly complex traits, a smaller proportion of the variance in the phenotype can be explained by genetic or epigenetic variation. Therefore, for a given statistical power, we are likely to detect far more molecular QTL than clinical trait QTL. The low overlap between EWAS and GWAS clinical trait hits may be due to a lack of sufficient power to detect associations for complex traits, since we also observe a higher overlap between EWAS and GWAS associations for expression *cis*-eQTL (77%) than for clinical traits (15%).

As an example, we found an EWAS hit on chromosome 13 for adipose tissue insulin resistance (ATIRI), glucose-to-insulin ratio, a measure of insulin sensitivity, and percent of monocytes in the blood (Figures 3A–3C). We did not find an association using GWAS for measures of insulin resistance or monocyte levels at this locus. The gene *Bhmt* encoding betaine-homocysteine methyltransferase located at 94.3 Mb on chromosome 13 was a candidate gene in this locus, since we also identified a *cis* association for protein levels of this gene using EWAS, or *cis*-pQTL, but no expression associations (Figures 3D and 3E). Methylation levels at this locus were correlated with glucose-to-insulin ratio (Figure 3F) and inversely correlated with protein levels of the gene (Figure 3G). Protein levels of *Bhmt* were also correlated with the trait (Figure 3H). Previous work in *Bhmt* knockout mice demonstrated that *Bhmt* plays a role in energy metabolism, specifically in lipid synthesis, and insulin sensitivity (Teng et al., 2012). Additional examples for associations with plasma cholesterol and total bone mineral density are describe in the Supplemental Information and Figure S6.

Conditional Association Studies

We used conditional association studies to determine whether hits identified with both EWAS and GWAS were (1) caused by the same signal or (2) co-localizing but independent signals. To accomplish this, we performed EWAS for overlapping associations using the CpG as the predictor and the SNP genotype as a covariate. In this approach, if an EWAS hit remains significant when we use the SNP as a covariate, this would suggest that the overlapping EWAS and GWAS association was independent but co-localizing. In contrast, if the EWAS hit goes away when we use the SNP as a covariate, then we can conclude that the EWAS and GWAS overlapping association was arising from the same signal. Out of 41 overlapping clinical trait associations, only three (7%) remained significant when we used the SNP genotype as a covariate at the Bonferroni threshold ($p < 1.2 \times 10^{-3}$), and none were significant at $p < 1 \times 10^{-7}$. Similarly, 71 (4%) of overlapping *cis* expression associations, corresponding to 64 unique genes, remained significant after using the SNP genotype for the locus as a covariate at the Bonferroni threshold

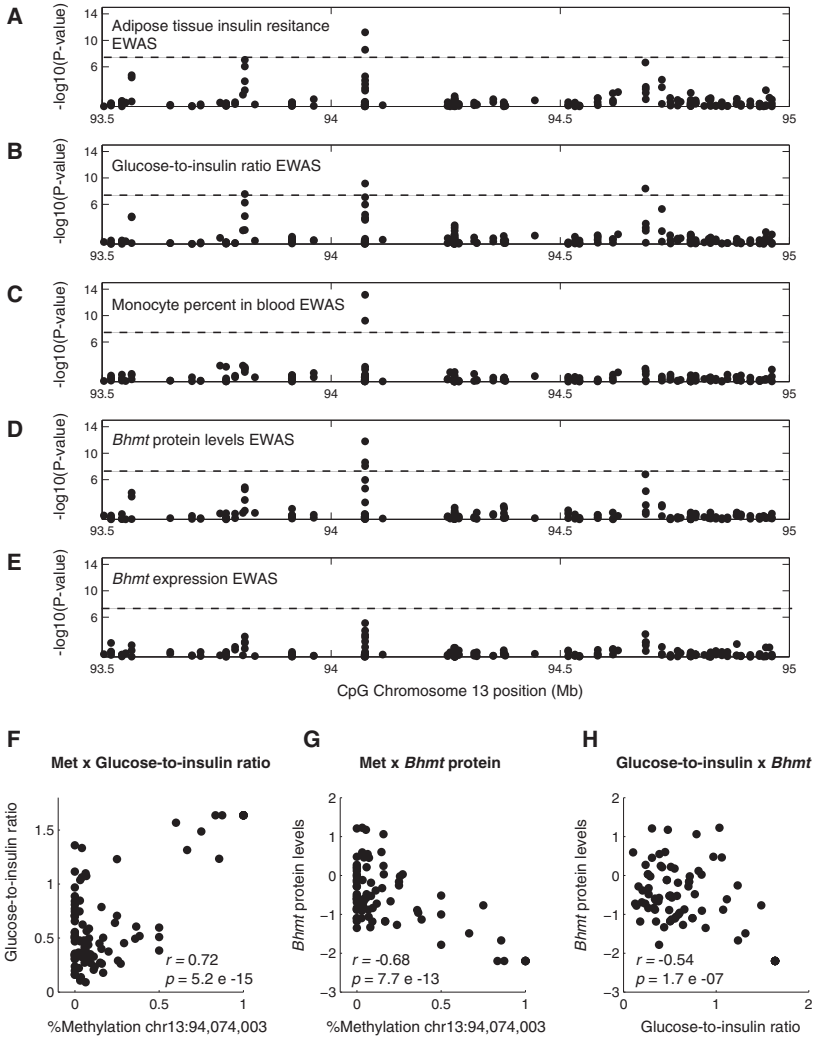


Figure 3. Insulin Resistance and *Bhmt* EWAS

(A–E) Manhattan plots showing association of methylation levels to (A) ATIRI, (B) plasma glucose-to-insulin ratio, (C) monocyte percent in the blood, (D) liver protein levels of *Bhmt*, and (E) liver expression levels of *Bhmt*. Chromosome location is on the x axis, the p value for the association is on the y axis, and each point represents a CpG. The dotted line is drawn at $p < 1 \times 10^{-7}$, the Bonferroni threshold for a single phenotype.

(F–H) Each point represents a mouse sample, showing correlation between (F) methylation levels for the peak associated CpG and glucose-to-insulin ratio levels, (G) methylation levels for the peak associated CpG and liver protein levels of *Bhmt*, and (H) glucose-to-insulin ratio and *Bhmt* protein levels. See also Figure S6.

(2.7×10^{-5}), and 30 of the associations (2%) were significant at $p < 1 \times 10^{-7}$. These results suggest that the majority of associations we found using both EWAS and GWAS were likely arising from the same signal at the associated locus.

Causal Inference Test

To determine whether associations identified by both EWAS and GWAS were mediated by differential methylation levels, we performed causal inference tests (CITs) using the R statistical package CIT (Millstein et al., 2009). The CIT performs a series of conditional probability tests to determine if the association between a genetic locus (L) and a trait (T) is mediated by DNA methylation (M), in this case, by testing for the following conditions: (1) the trait is associated with the locus, L→T; (2) the trait is associated with the methylation mediator given the locus M→T | L; (3) the methylation mediator is associated with locus given the trait, L→M | T; and (4) the locus is independently associated with both the mediator and the trait given the mediator, L→T | M. If the first three conditions are met, we can say that genetic variation at the locus is associated with the trait and that this association is mediated entirely by DNA methylation, L→M→T. Alternatively, if all four conditions are met, we can say that the association is still causal and mediated by DNA methylation, but the genotype at the locus also affects methylation and the trait independently. In summary, the two causal models tested by the CIT are the causal model, where the association is mediated entirely by DNA methylation, and the causal independent model, where the association is causal but the locus is also independently associated with DNA methylation and the trait.

We performed the CIT for clinical trait associations identified by both EWAS and GWAS and found that ten out of the 41 (24%) overlapping associations were causal and mediated by DNA methylation, and nine of the ten were causal independent associations (Table S2). However, it is possible that there are additional causal relationships between genetic variants, DNA methylation, and traits, but we lack sufficient power to detect these. Similarly, we performed the CIT for the gene expression associations (eQTL) we identified using both EWAS and GWAS, corresponding to 1,530 unique genes. We found 352 (22%) *cis*-eQTL mapping genes were mediated by variation in DNA methylation, where the locus genotype influenced methylation and DNA methylation in turn influenced the trait, L→M→T, and 321 of these were causal independent. Overall, results from the CIT indicate that a proportion (22%–24%) of overlapping EWAS and GWAS hits were causal associations mediated by DNA methylation. These results in conjunction with the conditional association studies suggest that that remaining EWAS hits are likely to be secondary to the genetic associations.

Principal-Component Analysis

The clinical traits in our study have a complex correlation structure (Figure S5B), since several traits such as insulin resistance,

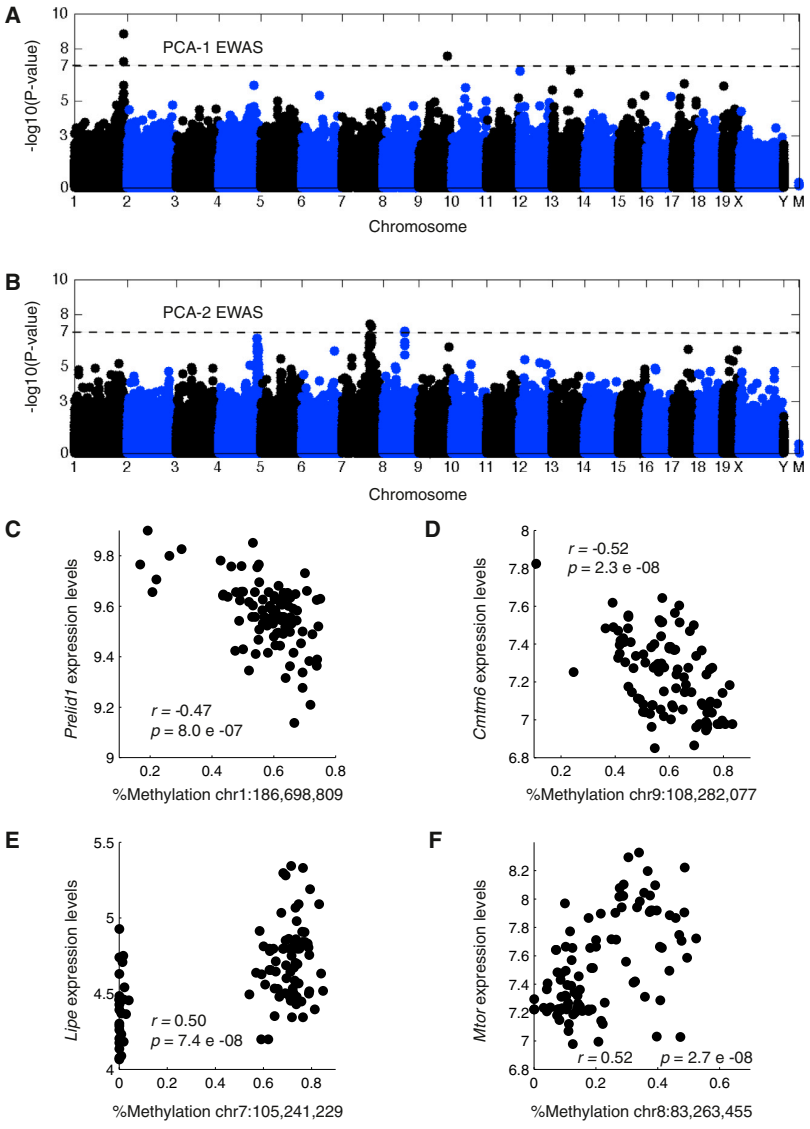


Figure 4. Principal Component EWAS

(A and B) Association between methylation and principal components (A) one and (B) two. Each dot represents a CpG, the genomic position of CpGs is on the x axis, and the $-\log_{10}$ of the p value for the association is on the y axis; chromosomes are shown in alternating colors.

(C–F) Each dot represents a mouse sample, showing correlation between (C) liver expression levels of *Prelid1* and methylation levels of a CpG associated with PCA1, (D) expression of *Cmtm6* and methylation of a CpG associated with PCA1, (E) expression of *Lipe* and methylation of a CpG associated with PCA2, and (F) expression of *Mtor* and methylation of a CpG associated with PCA2. See also Figure S5.

plasma cholesterol levels, and obesity are interrelated. To account for correlations between clinical traits, and to identify loci that drive multiple correlated traits, we performed principal-component analysis on the clinical traits (Figure S5C). The first component explained 24% of the variation in the traits and had the highest weights for glucose-to-insulin ratio and fat-related traits. The second component explained 12% of the variation in the traits and had the highest weights for blood- and fat-related traits. We performed EWAS between CpGs and the two first principal components as phenotypes and found two significant associations with the first component on chromosomes 1 and 9 at the Bonferroni threshold ($p < 6.9 \times 10^{-8}$, Figure 4A). We searched for individual methylation sites in this locus that were also associated with gene expression or protein levels. Genes associated with the chromosome 1 locus methylation levels include *Prelid1* (Figure 4C), a gene involved in lipid transport, and *1110057K04Rik*, a gene recently found to be involved in lipid storage (Goo et al., 2014). Methylation levels at the chromosome 9 locus were associated with *Cmtm6* (Figure 4D), a gene structurally related to chemokines, although its exact function is still unknown. The second component was strongly associated with a locus on chromosome 7 spanning approximately 10 Mb that also coincided with the Hemoglobin beta locus (Figure 4B). Consistent with the correlation between the second principal component and fat-related traits, we found that CpGs at this locus were associated with several genes including *Lipe* (Figure 4E), a lipase gene involved in free fatty acid oxidation (Reid et al., 2008). In addition, we identified CpGs in chromosome 8 associated with the second principal component and with expression levels of *Mtor* (Figure 4F). *Mtor* plays a role in metabolic regulation, response to nutrients, insulin, and diabetes (Zhu et al., 2013).

DNA Methylation Can Be Used to Infer Phenotypes

Genetics and genomics data is a highly valuable resource that can be used to model disease susceptibility and risk. An individual's genome is predominantly static, but the epigenome is variable in different tissues and is affected by transcription patterns. To determine if CpGs could be used to infer clinical traits in other individuals based on their methylation status, we built linear models with CpG sites using the generalized linear model package *glmnet* (Friedman et al., 2010), and tested their power to infer clinical traits in test individuals where the CpG status was known. For each trait, we randomly selected a test set of ten mice that were kept hidden from the training set and used the remaining mice as the training set, where the methylation status was known in both sets. We used *glmnet* to select CpGs from the 20,000 most variable CpGs and built a linear model for the trait based on these CpGs using the training set. We then used the resulting linear model to infer trait values on the test set of ten mice and determined the accuracy of the model using Pearson's correlation between the observed and inferred trait values. We found

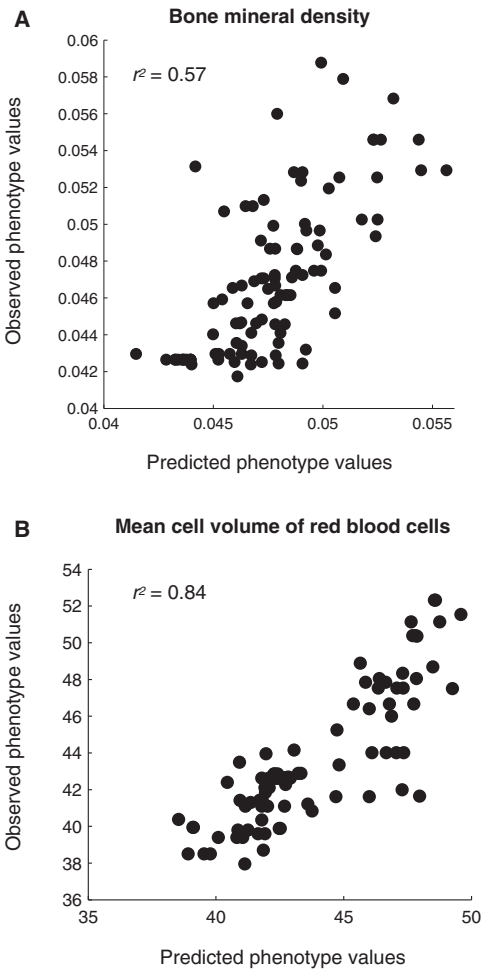


Figure 5. Phenotype Inference

(A and B) Phenotype predictions for (A) bone mineral density and (B) mean cell volume of red blood cells. The predicted phenotype value is on the x axis and the measured phenotype value is on the y axis. Each point is a mouse sample. See also Table S4.

several clinical traits that could be accurately inferred on test individuals from a set of CpG methylation sites, where the trait values predicted by the model were highly correlated with the clinical trait values measured in the mice. We found eight clinical traits with $r^2 > 0.5$, including plasma total cholesterol levels, plasma HDL cholesterol levels, total bone mineral density, plasma fatty acids, and red blood cell phenotypes. Examples of inferred and measured clinical trait values are shown for bone mineral density (Figure 5A) and for mean cell volume of red blood cells (Figure 5B). A list of inferred phenotypes, the correlations between predicted and observed phenotypes, and the top ten CpGs selected to model each phenotype can be found in Table S4.

Natural Genetic Variation Influences Genome-Wide DNA Methylation Levels

To determine the extent to which genetics affects natural variation in DNA methylation, we used a linear mixed model to perform GWAS of CpG methylation levels as traits to SNPs across the mouse genome. We and others have shown that this method reduces false-positive associations that are due to population structure (Bennett et al., 2010; Kang et al., 2008). For each CpG, we associated methylation levels to SNPs with MAF greater than 10%. We called significant associations at the Bonferroni threshold ($p < 1.4 \times 10^{-12}$) by considering each CpG and SNP pair as an independent test (Table S5). We chose this stringent threshold to call significant associations in order to minimize the possibility of examining false-positive associations. However, all association results at $p < 1 \times 10^{-6}$ can be obtained from our website at http://ewas.mcdb.ucla.edu/download.html. We identified 3,017,453 associations between methylation levels and genetics at the Bonferroni threshold (Figure 6A), corresponding to 26,563 unique CpGs, or 7% of all methylation sites tested, and 92,959 unique SNPs. Approximately 51% of all significant associations were for *Hypervariable* CpGs, and 52% of all *Hypervariable* CpGs (11,644) were under genetic regulation. We found that 12% of the associations involved SNPs that abolished a CpG in a fraction of the mouse strains (CG-SNPs), corresponding to 2,533 of the CG-SNPs present in the strains. We note that these CG-SNPs were not used in our EWAS, since differences in methylation between strains were the result of a cytosine change to a different DNA base.

We estimated the variance explained by genetics, or the narrow sense heritability of DNA methylation levels for individual CpGs using an additive model. When we examined all *Variable* CpGs, which display variation in methylation levels in at least one strain, the variance explained by genetics was on average 27% (Figure S7A). In contrast, on average 60% of the variance was explained by genetics for *Hypervariable* CpGs, which display higher variation in methylation among the strains. The variance explained by genetics was on average 75% for CG-SNPs. These CG-SNPs do not show 100% heritability, likely because methylation levels can still be controlled in *trans* for the strains with the C allele. A large proportion of the associations were *local* or *cis*. We previously estimated the GWAS mapping resolution to be 2 Mb on average (Bennett et al., 2010). Here, 54% of the associations were *local* or *cis*, where SNP and CpG pairs were within 2 Mb of each other. However, *trans* associations where SNP and CpG pairs were more than 2 Mb from each other were also found in the same chromosome for approximately 79% of *trans* associations, suggesting that many of these associations may be *cis* associations caused by long-range LD in SNPs. The distance between SNPs and *Hypervariable* CpGs was on average 1.4 times smaller than the overall distribution of SNP to CpG distances, and the distance between SNPs and CG-SNPs was on average 3.6 times smaller than the overall distribution. The distribution of the distances between SNP and CpGs pairs is shown on Figure S7B, and results of the DNA methylation GWAS are summarized in Table S5.

Mtrr Influences Methylation Levels of CpGs across the Genome

Methylation levels of hundreds of CpGs across the genome mapped to a QTL hotspot on chromosome 13, defined as a 2 Mb bin at 68–70 Mb, near the gene Methionine synthase reductase, *Mtrr*, located at 68.7 Mb (Figure 6C). This methylation hotspot can also be seen as a vertical band on the GWAS plot

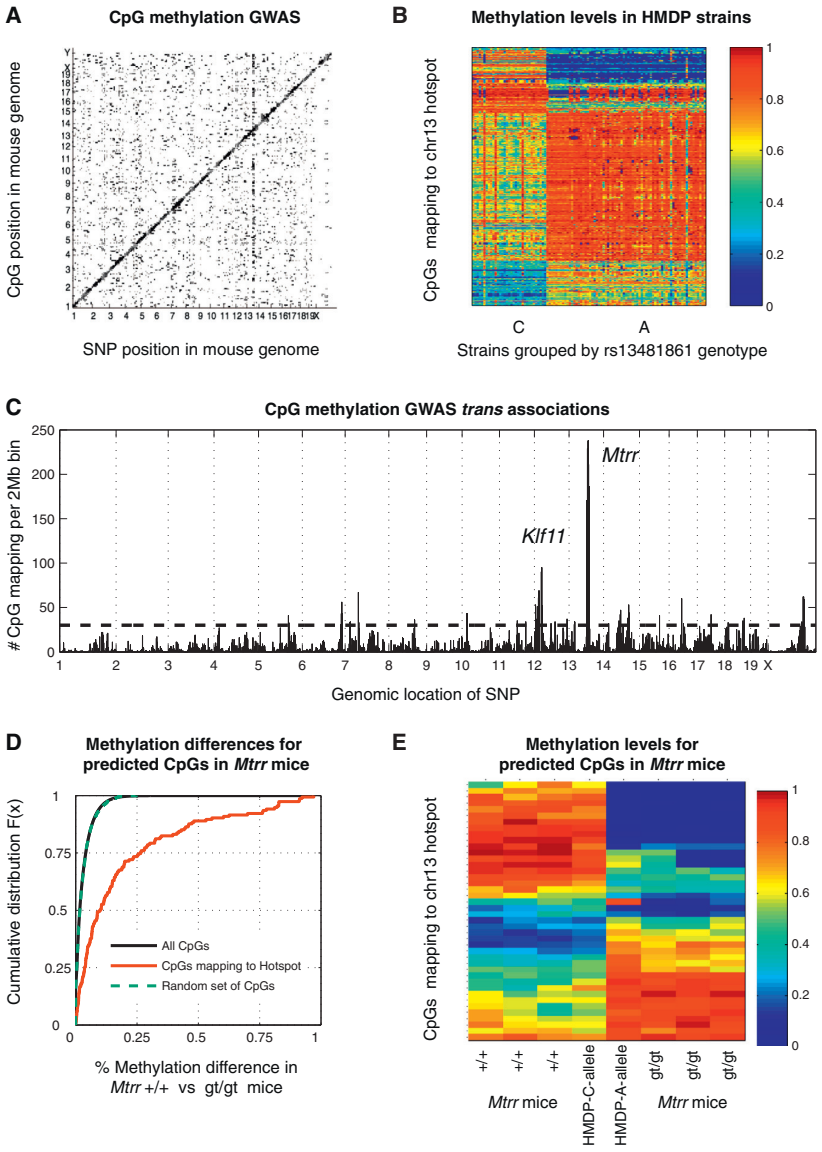


Figure 6. Natural Genetic Variation Influences Genome-Wide DNA Methylation

(A) GWAS using *Variable* CpG methylation levels as phenotypes and SNPs as predictors. Genomic position of SNPs is on the x axis, and the genomic position of CpGs is on the y axis. Each point is a significant association at the Bonferroni threshold $p < 1.4 \times 10^{-12}$.

(B) Methylation levels for CpGs mapping to the chromosome 13 GWAS hotspot. Strains are on the x axis grouped by their genotype of rs13481861 at the *Mtrr* locus, and CpGs are on the y axis. The color is the methylation level between 0% and 100%.

(C) CpG methylation GWAS hotspots. The number of CpGs mapping in *trans* to each 2 Mb bin is on the y axis. The genomic position of each bin is on the x axis. The horizontal dotted line is the Poisson significance threshold for each hotspot bin.

(D and E) Experimental validation of the hotspot at the *Mtrr* locus using RRBS in livers of wild-type mice (+/+) and mice homozygous for the *Mtrr* gene-trapped allele (gt/gt).

(D) Distribution of methylation differences. The difference in methylation between +/+ versus gt/gt mice is on the x axis, and the cumulative distribution function is on the y axis. The curves show the distribution of all CpGs (black), randomly sampled CpGs (green dotted), and CpGs predicted to be affected by the *Mtrr* genotype (red).

(E) Differentially methylated CpGs between $Mtrr^{+/+}$ and gt/gt mice at FDR < 5%. Mice are on the x axis, and CpGs are on the y axis. The color denotes methylation levels between 0% and 100%. See also Figure S7.

(Figure 6A). Expression levels of *Mtrr* were variable among the mouse strains and were regulated in *cis*, since we observed both a *cis*-eQTL for *Mtrr* using GWAS ($p = 1.82 \times 10^{-14}$) and an association between *Mtrr* expression and methylation levels 314 bp from *Mtrr* using EWAS ($p = 4.97 \times 10^{-14}$). Expression levels of *Mtrr* were highly correlated with methylation levels of CpGs mapping to the locus both in *cis* and in *trans*, with an average absolute Pearson's $r = 0.48$. The distribution of these correlations was significantly different (KS test $p = 4.98 \times 10^{-187}$) from the correlation between *Mtrr* expression and all CpGs, which had average absolute $r = 0.09$ (Figure S7C). *Mtrr* is necessary for the utilization of methyl groups from the folate cycle, which donates methyl groups to multiple cellular pathways including DNA methylation (Crider et al., 2012). *MTRR* was recently associated with methylation levels in autoimmune thyroid disease in humans (Arakawa et al., 2012), along with *DNMT1*, *DNMT3A*, *DNMT3B*, and *MTHFR*. All these suggested that *Mtrr* was an ideal candidate gene for the methylation hotspot.

Based on our GWAS results, we observed 471 CpGs from the HMDP mapping to the chromosome 13 hotspot (Figure 6B). These CpGs were physically located throughout the mouse genome, and we hypothesized that their methylation levels were influenced by *Mtrr*. To experimentally validate *Mtrr* as a causal gene for the chromosome 13 methylation hotspot, we measured DNA methylation levels in the livers of *Mtrr* wild-type (+/+) and homozygous gene-trapped mice (gt/gt) using RRBS. Mice homozygous for the gene-trapped allele display reduced expression, protein, and activity of *Mtrr* (Elmore et al., 2007). Of the 471 CpGs mapping to the hotspot, 154 CpGs were represented in the RRBS data set of *Mtrr* mice, and 42 of the 154 (27%) were differentially methylated between *Mtrr* wild-type and gt/gt mice at 5% false discovery rate (FDR). Methylation levels at these differentially methylated sites are shown on Figure 6E for wild-type and gt/gt mice, as well as the average methylation levels of mouse strains with the reference allele (C) or alternate allele (T) for SNP rs13481861 located in an exon of *Mtrr*. The list of 154 CpGs we tested using RRBS, including the methylation delta; the p value for differential

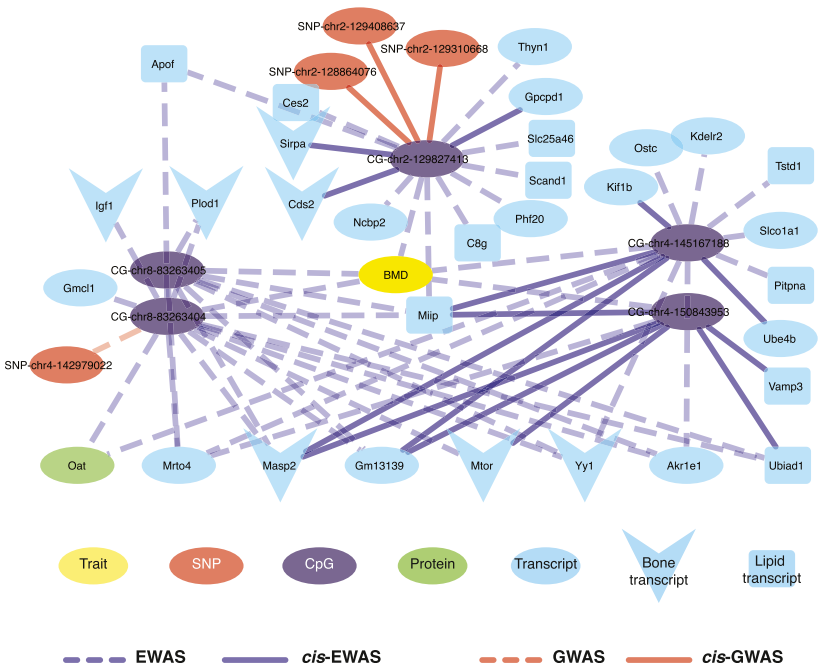


Figure 7. Bone Mineral Density Association Graph

Association graph of EWAS and GWAS hits. Edges are defined by EWAS, purple lines; GWAS, red lines; *Cis*-associations, solid lines; and *Trans*-associations, dotted lines. Node colors are trait, yellow; CpG, purple; SNP, red; gene expression, light blue; and protein levels, green. Genes implicated in bone mineral density and/or bone biology are shown with V shape, and genes implicated in fat or lipid metabolism are shown as squares.

methylation between *Mtrr* wild-type and gt/gt mice; and the average methylation of allele C and allele T HMDP strains can be found in Table S6. We were not able to observe all 471 CpGs because of the decrease in coverage, since we multiplexed six samples in one lane for the validation experiments and two samples per lane for all the HMDP samples.

The difference in methylation (delta) between wild-type and gt/gt mice was on average 6.7-fold higher in the 154 predicted CpGs, relative to all CpGs observed in the *Mtrr* mice (KS test $p = 1.8 \times 10^{-33}$). To test if the differences between *Mtrr* wild-type and gt/gt mice at these CpGs were due to chance, we randomly sampled sets of CpGs across the genome and calculated the delta between wild-type and gt/gt mice in 1,000 samplings, but we found no significant differences between randomly sampled CpGs and all observed CpGs. Random sets of CpGs had an average delta of 0.99% and an average KS test $p = 0.52$ for *Mtrr* +/+ versus gt/gt mice across all 1,000 samplings, with a minimum KS test $p = 0.001$ and maximum delta of 1.1%. In contrast, the 154 CpGs that we hypothesized were regulated by *Mtrr* showed an average delta of 20% between $Mtrr^{+/+}$ and gt/gt, with KS test $p = 1.8 \times 10^{-33}$. The distribution of methylation differences in *Mtrr* wild-type and gt/gt mice is shown on Figure 6D for the 154 predicted CpGs, a set of randomly sampled CpGs, and all CpGs observed in the *Mtrr* mouse RRBS data set.

Online Databases for the Identification of Candidate Genes

All of our EWAS and GWAS results are available through an online database to facilitate the identification of candidate genes for clinical traits using DNA methylation patterns, which can be accessed at http://ewas.mcdb.ucla.edu. We also created an different online tool that generates association graphs based on our results at http://pathways.mcdb.ucla.edu/network. A sample association graph for bone mineral density is shown in Figure 7. Further details on these online databases can also be found in the Supplemental Information.

DISCUSSION

In this study we leveraged a powerful mouse systems genetics platform to ask what the relationship is between DNA variation and methylation. What are the loci that control methylation levels? How does methylation relate to clinical traits that are precursors to heart disease and diabetes? Can methylation be incorporated into a network and causal models of complex disease? Our results demonstrate that using DNA methylation for GWAS (i.e., EWAS) complements traditional GWAS.

We identified thousands of associations between DNA methylation levels and clinical traits such as bone mineral density, adiposity, plasma cholesterol, glucose, insulin, triglyceride levels, and molecular traits such as metabolites, protein, and gene expression levels (Figure 2). Roughly 15% of EWAS hits for clinical traits could also be found using GWAS in the same panel of mouse strains, but the remaining associations were unique to the EWAS (Figure S5A). The low overlap between methylation and genetic associations for clinical traits may be due a lack of statistical power to detect associations with small effects. Since molecular traits are typically less complex relative to clinical traits, we were more likely to detect associations for gene expression and protein levels, and indeed, we observed a much larger overlap for *cis* expression (77%) and *cis* protein associations (37%) identified with EWAS and GWAS. In cases where associations are identified by both EWAS and GWAS, we asked whether methylation was mediating the effect on the trait. We used a CIT to address this question, and found evidence that DNA methylation was mediating the effect on the trait in 24% of the overlapping clinical trait associations, and 22% of the overlapping *cis* expression associations. However, we note that it is often difficult to establish causal relationships for either genetic or epigenetic associations for a number of reasons, including the complexity of biological pathways, LD, and statistical power. In addition, although the CIT examines the model where a locus affects methylation and methylation affects the trait (L→M→T), there are other possible scenarios not tested

by the CIT, such as where DNA methylation changes are reactive and are not mediators of the association.

DNA methylation levels are dynamic and can be modified in response to disease, age, and environmental perturbations. In contrast, the genome remains largely static throughout an individual's lifetime and is modified only in certain diseases such as cancer. We found that correlations in CpG methylation patterns were significantly smaller than correlations in SNP genotypes (LD), leading to a dramatic increase in our association mapping resolution (Figures 1 and S3). LD blocks are typically much larger in laboratory mice than in human populations, due the breeding history of existing mouse strains. Hence, we would not expect a similar increase in mapping resolution using EWAS in human populations. The extensive LD in mice has been a major difficulty for candidate gene identification in mouse genetics studies, and we found that this can be largely overcome using EWAS in mice. We hypothesize that the plasticity in the epigenome allows methylation patterns to be at least partly decoupled from local genetic patterns, since CpG methylation can be modified by chromatin binding proteins in *trans*. In contrast, SNP genotypes are static, and LD patterns remain fixed in the population.

In addition, DNA methylation patterns can vary across different tissues. One of the advantages of studies in mammalian model organisms such as the mouse is the ability to sample tissues that are not readily available in human studies. A previous study identified methylation associations using blood for rheumatoid arthritis in humans (Liu et al., 2013), and in the current study, we found associations for bone mineral density and methylation levels in the liver (Figures 7 and S6). These findings suggest that it is possible to uncover significant associations for methylation patterns that are conserved between the tissue that is sampled and the tissue relevant to the trait of interest, but associations to methylation levels that are not conserved are likely to be missed. We believe that there is potentially a large amount of information to be gained from studying relevant tissues whenever possible.

We examined the degree to which methylation is controlled by genetics by taking methylation patterns of individual CpGs as phenotypes and mapping them to the SNP genotypes using GWAS. We found that 7% of all CpGs and 52% of *Hypervariable* CpGs were under genetic control (Figure 6A). A large proportion (55%) of these associations were in *cis*, where the CpG and SNP were found within 2 Mb of each other. Although only 7% of all CpGs we examined were significantly associated with SNPs, it is possible that we did not have sufficient power to identify additional associations, particularly for genetic variants with subtle effects on DNA methylation. In addition, we note that CpGs with minimal or no variation in methylation levels would not be significantly associated even if they were stably maintained across generations. Previous human methylation studies, or mQTL studies, found that 20% of variable CpGs were associated with genetic variation in blood leukocytes (McRae et al., 2014) and 28% in adipose tissue (Grundberg et al., 2013). We observed an average heritability of 27% for all CpGs and 60% for *Hypervariable* CpGs in the liver of mouse inbred strains, excluding CG-SNPs. In comparison, heritability in previous human twin studies was on average 12% in cord blood mononuclear cells, 7% in human umbilical vein endothelial cells (Gordon et al., 2012), 19% in blood leukocytes (McRae et al., 2014), and median heritability of 34% in adipose tissue in the MuTHER cohort (Grundberg et al., 2013). These results support the notion that DNA methylation levels are indeed associated with genetic variation, although the extent of the heritability and genetic effects is variable across different tissues and population samples.

Although the majority of associations between CpG methylation levels and genetics were in *cis*, we found a *trans* association hotspot where methylation levels of CpGs across the genome map to a locus in chromosome 13 near *Mtrr* (Figures 6A and 6C), a gene necessary for the utilization of methyl groups from the folate cycle. This suggested that natural genetic variation in a population can influence genome-wide DNA methylation levels. The FAST kinase gene *Fast3kd* was also located on chromosome 13 near *Mtrr*, and is another candidate gene for the hotspot. Unfortunately, *Fastkd3* was not represented in either our gene expression or protein data sets, and we could not observe any *cis*-eQTL or *cis* expression EWAS hits for this gene. Although *Fastkd3* may also be a candidate for the hotspot, it contains a mitochondrial-targeting domain and functions primarily in the mitochondria, making it a less likely candidate than *Mtrr* for influencing CpG methylation levels in *trans*. We confirmed the role of *Mtrr* in 27% CpGs predicted to be affected by the chromosome 13 hotspot, since these CpGs were differentially methylated between wild-type and gt/gt mice (Figure 6E). The validation results of the chromosome 13 hotspot we present here are consistent with our previous work on a gene expression hotspot on mouse chromosome 8 in primary macrophages (Orozco et al., 2012), where we experimentally validated 12% of the genes predicted to map to the chromosome 8 hotspot.

One of the most desirable applications since the advent of the human genome project has been to be able to determine a person's phenotype from their genome sequence. However, understanding how genetic variation alters cellular behavior and organismal phenotypes, and accurately inferring phenotypes from raw genotypes has proved to be an extremely difficult endeavor. A recent study demonstrated that modeling of SNPs from whole-genome sequencing data could be used to infer starvation resistance and startle response in *Drosophila* (Ober et al., 2012). Here we show that DNA methylation patterns can be used to infer complex phenotypes in a mammalian organism, including bone mineral density, blood cell phenotypes, and plasma cholesterol levels (Figure 5; Table S4). We built linear models that incorporate the DNA methylation status at specific CpGs and used these models to infer clinical traits in other individuals in the same cohort whose methylation status was known. We note that the statistical inference approach we used to model, or explain, a phenotype is distinct from longitudinal prediction of phenotypes for a given individual at a future time.

Association studies over the past 10 years have found that the majority of genetic polymorphisms associated with traits are outside protein-coding regions. The associated genetic polymorphisms are thought not to alter genes themselves, but rather regulatory elements that control gene expression (Furey and Sethupathy, 2013). Our findings suggest that DNA variants can act by regulating DNA methylation, which in turn affects regulation of gene expression or protein levels of genes that function in biological mechanisms important for the phenotype expression. We hypothesize that it is the plasticity in DNA methylation that makes

it ideal for quantitative trait modeling. Since DNA methylation patterns are specific to developmental stages and cell types and can vary in response to the environment or disease, they can capture the cellular status and provide a more detailed picture of dynamic cellular behavior than our static genomes. Ultimately, our studies suggest that CpG methylation patterns are themselves under genetic control, but because they are more responsive to an organism's state, they can provide added information that cannot be obtained from the genetic sequence alone.

EXPERIMENTAL PROCEDURES

A more detailed version of the experimental procedures can be found in the Supplemental Information.

Mice

All animals were handled in strict accordance with good animal practice as defined by the relevant national and local animal welfare bodies, and all animal experiments and work were carried out with UCLA IACUC approval.

Data Access

All RRBS sequencing and SNP data can be obtained from GEO: GSE67507. The EWAS and GWAS results can be accessed in our online databases to search for candidate genes at http://ewas.mcdb.ucla.edu and to generate association graphs at http://pathways.mcdb.ucla.edu/network. Individual tables with all methylation associations can be downloaded from http://ewas.mcdb.ucla.edu/download.html. The GWAS results can also be accessed at http://systems.genetics.ucla.edu/data/hmdp.

RRBS Libraries

We prepared RRBS libraries as previously described (Smith et al., 2009), with minor modifications. We sequenced the libraries by multiplexing two libraries per lane in an Illumina HiSeq sequencer, with 100 bp reads.

Alignment

We aligned the reads with BS-Seeker2 (Guo et al., 2013) to the mm9 mouse reference genome. We used Bowtie as the base aligner, trimmed adapters, allowed for up to five mismatches, and selected uniquely aligned reads.

LD and CpG Correlation Studies

We computed the Pearson's r^2 between pairs of SNPs, or pairs of CpGs, excluding missing values.

EWAS

We used the linear mixed model package pyLMM (https://github.com/nickFurlotte/pylmm) to test for association and to account for population structure and relatedness among the mouse strains. This method was previously described as EMMA (Kang et al., 2008), and we implemented the model in Python to allow for continuous predictors, such as CpG methylation levels that vary between 0 and 1. We applied the model $y = \mu + x\beta + u + e$, where μ = mean, x = CpG, β = CpG effect, and u = random effects due to relatedness, with Var(u) = σ_g^2K and Var(e) = σ_e^2, where K = IBS (identity-by-state) matrix across all *Variable* CpGs. We computed a restricted maximum likelihood estimate for σ_g^2K and σ_e^2, and we performed association based on the estimated variance component with an F test to test that β does not equal 0. Each phenotype was log transformed for the association test.

Inflation

We calculated the inflation factor lambda by taking the chi-square inverse cumulative distribution function for the median of the association p values, with one degree of freedom, and divided this by the chi-square probability distribution function of 0.5 (the median expected p value by chance) with one degree of freedom.

Overlap of EWAS and GWAS

We defined an overlap between EWAS and GWAS if the associations were found within 2 Mb (Figure S5A). To decrease the chance of not finding an overlap based on our stringent Bonferroni EWAS thresholds, we used the per phenotype Bonferroni threshold of $p < 1 \times 10^{-7}$ for EWAS and $p < 4.1 \times 10^{-6}$ for the GWAS as previously described (Bennett et al., 2010).

Conditional EWAS

We performed EWAS for clinical traits or *cis* expression associations identified with both EWAS and GWAS. We used the pyLMM package as described with one modification: for each EWAS, we used the SNP genotype for the GWAS hit as a covariate.

CIT

We performed CITs using the R statistical package CIT developed by Millstein and colleagues (Millstein et al., 2009), according to the user's manual.

PCA

We performed a principal-component analysis on the clinical traits. The first and second principal components explained 24% and 12% of the variation in the traits, respectively. We mapped the first two principal components as traits to CpG methylation levels across the genome using EWAS as described above.

Methylation GWAS

We tested for association between methylation levels as phenotypes, and SNPs as predictors using EMMA as previously described (Bennett et al., 2010). The difference between the EWAS model described above, and the GWAS linear mixed model is that in GWAS x = SNP, β = SNP effect, and K = IBS (identity-by-state) matrix across all SNPs.

Methylation GWAS Hotspots

We divided the genome into 2 Mb bins and counted the number of all unique CpGs with a significant GWAS hit in that bin and called these "*cis* and *trans*" associations (Figure S7E). We also defined a set of *trans* association hotspots (Figure 6C), where we counted CpGs mapping to each bin in *trans*, such that the CpG was physically located at least 10 Mb away from the bin. We considered CpGs to be associated at the Bonferroni threshold with $p < 1.4 \times 10^{-12}$. We used the Poisson distribution to determine if individual bins had a higher than expected number of associations.

Validation of *Mtrr* Hotspot

We generated RRBS libraries from *Mtrr* gene-trapped mice (Elmore et al., 2007), using three wild-type and three homozygous gene trapped (gt/gt) male mice at 3 months of age. We sequenced the libraries by multiplexing all six libraries in one lane and aligned the data using BS-Seeker2 as described above. We compared CpG methylation levels in +/+ and gt/gt mice using a t test and estimated the FDR using the Storey method (Storey, 2002). We calculated the difference in methylation levels at each CpG by taking the absolute difference in methylation between the average methylation *Mtrr*+/+ and the average in −/− mice (i.e., delta methylation).

Phenotype Inference

We used the *glmnet* package in R for building linear models, which fits a generalized linear model via penalized maximum likelihood (Friedman et al., 2010).

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures, six tables, and Supplemental Experimental Procedures and can be found with this article online at http://dx.doi.org/10.1016/j.cmet.2015.04.025.

AUTHOR CONTRIBUTIONS

L.D.O., A.J.L., and M.P. conceived the studies. L.R. and M.M. prepared the libraries. W.G. aligned the libraries. J.G. performed PCA and phenotype predictions. H.S. and D.L. created the online databases. N.A.F. created pyLMM. R.B. and R.R. generated the *Mtrr* mice. B.J.B., C.R.F., A.G., and L.D.O. collected

the tissues, clinical traits, DNA, metabolite-, protein-, and gene expression data. L.D.O. performed analyses and wrote the manuscript. M.P. directed the study.

ACKNOWLEDGMENTS

We want acknowledge Joshua Millstein for his valuable advice on performing the CIT and Hong Xiu Qi, Pingzi Wen, and Sharda Charugundla for their valuable help in the animal experiments. L.D.O. was supported by the Ruth L. Kirschstein National Research Service Award T32AR059033. J.G. was supported by the grant CIRM TB1-01183. M.Z. and W.G. were supported by the grant NBRPC 2012CB316503 and the grant NSFC 91010016. W.G. was supported by the China Scholarship Council. This work was supported by the National Institutes of Health grant GM095656-01A1 and HL28481.

Received: October 7, 2014
Revised: January 12, 2015
Accepted: April 22, 2015
Published: June 2, 2015

REFERENCES

Arakawa, Y., Watanabe, M., Inoue, N., Sarumaru, M., Hidaka, Y., and Iwatani, Y. (2012). Association of polymorphisms in DNMT1, DNMT3A, DNMT3B, MTHFR and MTRR genes with global DNA methylation levels and prognosis of autoimmune thyroid disease. Clin. Exp. Immunol. *170*, 194–201.

Bell, J.T., Pai, A.A., Pickrell, J.K., Gaffney, D.J., Pique-Regi, R., Degner, J.F., Gilad, Y., and Pritchard, J.K. (2011). DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biol. *12*, R10.

Bennett, B.J., Farber, C.R., Orozco, L., Kang, H.M., Ghazalpour, A., Siemers, N., Neubauer, M., Neuhaus, I., Yordanova, R., Guan, B., et al. (2010). A high-resolution association mapping panel for the dissection of complex traits in mice. Genome Res. *20*, 281–290.

Crider, K.S., Yang, T.P., Berry, R.J., and Bailey, L.B. (2012). Folate and DNA methylation: a review of molecular mechanisms and the evidence for folate's role. Adv. Nutr. *3*, 21–38.

Dick, K.J., Nelson, C.P., Tsaprouni, L., Sandling, J.K., Aïssi, D., Wahl, S., Meduri, E., Morange, P.E., Gagnon, F., Grallert, H., et al. (2014). DNA methylation and body-mass index: a genome-wide analysis. Lancet *383*, 1990–1998.

Elmore, C.L., Wu, X., Leclerc, D., Watson, E.D., Bottiglieri, T., Krupenko, N.I., Krupenko, S.A., Cross, J.C., Rozen, R., Gravel, R.A., and Matthews, R.G. (2007). Metabolic derangement of methionine and folate metabolism in mice deficient in methionine synthase reductase. Mol. Genet. Metab. *91*, 85–97.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. *33*, 1–22.

Furey, T.S., and Sethupathy, P. (2013). Genetics. Genetics driving epigenetics. Science *342*, 705–706.

Ghazalpour, A., Bennett, B., Petyuk, V.A., Orozco, L., Hagopian, R., Mungrue, I.N., Farber, C.R., Sinsheimer, J., Kang, H.M., Furlotte, N., et al. (2011). Comparative analysis of proteome and transcriptome variation in mouse. PLoS Genet. *7*, e1001393.

Ghazalpour, A., Bennett, B.J., Shih, D., Che, N., Orozco, L., Pan, C., Hagopian, R., He, A., Kayne, P., Yang, W.P., et al. (2014). Genetic regulation of mouse liver metabolite levels. Mol. Syst. Biol. *10*, 730.

Goo, Y.H., Son, S.H., Kreienberg, P.B., and Paul, A. (2014). Novel lipid droplet-associated serine hydrolase regulates macrophage cholesterol mobilization. Arterioscler. Thromb. Vasc. Biol. *34*, 386–396.

Gordon, L., Joo, J.E., Powell, J.E., Ollikainen, M., Novakovic, B., Li, X., Andronikos, R., Cruickshank, M.N., Conneely, K.N., Smith, A.K., et al. (2012). Neonatal DNA methylation profile in human twins is specified by a complex interplay between intrauterine environmental and genetic factors, subject to tissue-specific influence. Genome Res. *22*, 1395–1406.

Grundberg, E., Meduri, E., Sandling, J.K., Hedman, A.K., Keildson, S., Buil, A., Busche, S., Yuan, W., Nisbet, J., Sekowska, M., et al.; Multiple Tissue Human Expression Resource Consortium (2013). Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. Am. J. Hum. Genet. *93*, 876–890.

Guo, W., Fiziev, P., Yan, W., Cokus, S., Sun, X., Zhang, M.Q., Chen, P.Y., and Pellegrini, M. (2013). BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. BMC Genomics *14*, 774.

Heyn, H., Moran, S., and Esteller, M. (2013). Aberrant DNA methylation profiles in the premature aging disorders Hutchinson-Gilford Progeria and Werner syndrome. Epigenetics *8*, 28–33.

Horvath, S. (2013). DNA methylation age of human tissues and cell types. Genome Biol. *14*, R115.

Huynh, J.L., Garg, P., Thin, T.H., Yoo, S., Dutta, R., Trapp, B.D., Haroutunian, V., Zhu, J., Donovan, M.J., Sharp, A.J., and Casaccia, P. (2014). Epigenome-wide differences in pathology-free regions of multiple sclerosis-affected brains. Nat. Neurosci. *17*, 121–130.

Kang, H.M., Zaitlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., Daly, M.J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. Genetics *178*, 1709–1723.

Liu, Y., Aryee, M.J., Padyukov, L., Fallin, M.D., Hesselberg, E., Runarsson, A., Reinius, L., Acevedo, N., Taub, M., Ronninger, M., et al. (2013). Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat. Biotechnol. *31*, 142–147.

McRae, A.F., Powell, J.E., Henders, A.K., Bowdler, L., Hemani, G., Shah, S., Painter, J.N., Martin, N.G., Visscher, P.M., and Montgomery, G.W. (2014). Contribution of genetic variation to transgenerational inheritance of DNA methylation. Genome Biol. *15*, R73.

Millstein, J., Zhang, B., Zhu, J., and Schadt, E.E. (2009). Disentangling molecular relationships with a causal inference test. BMC Genet. *10*, 23.

Ober, U., Ayroles, J.F., Stone, E.A., Richards, S., Zhu, D., Gibbs, R.A., Stricker, C., Gianola, D., Schlather, M., Mackay, T.F., and Simianer, H. (2012). Using whole-genome sequence data to predict quantitative trait phenotypes in Drosophila melanogaster. PLoS Genet. *8*, e1002685.

Orozco, L.D., Bennett, B.J., Farber, C.R., Ghazalpour, A., Pan, C., Che, N., Wen, P., Qi, H.X., Mutukulu, A., Siemers, N., et al. (2012). Unraveling inflammatory responses using systems genetics and gene-environment interactions in macrophages. Cell *151*, 658–670.

Orozco, L.D., Rubbi, L., Martin, L.J., Fang, F., Hormozdiari, F., Che, N., Smith, A.D., Lusis, A.J., and Pellegrini, M. (2014). Intergenerational genomic DNA methylation patterns in mouse hybrid strains. Genome Biol. *15*, R68.

Reid, B.N., Ables, G.P., Otlivanchik, O.A., Schoiswohl, G., Zechner, R., Blaner, W.S., Goldberg, I.J., Schwabe, R.F., Chua, S.C., Jr., and Huang, L.S. (2008). Hepatic overexpression of hormone-sensitive lipase and adipose triglyceride lipase promotes fatty acid oxidation, stimulates direct release of free fatty acids, and ameliorates steatosis. J. Biol. Chem. *283*, 13087–13099.

Shenker, N.S., Polidoro, S., van Veldhoven, K., Sacerdote, C., Ricceri, F., Birrell, M.A., Belvisi, M.G., Brown, R., Vineis, P., and Flanagan, J.M. (2013). Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. Hum. Mol. Genet. *22*, 843–851.

Smith, Z.D., Gu, H., Bock, C., Gnirke, A., and Meissner, A. (2009). High-throughput bisulfite sequencing in mammalian genomes. Methods *48*, 226–232.

Storey, J.D. (2002). A direct approach to false discovery rates. J. R. Stat. Soc. Series B Stat. Methodol. *64*, 479–498.

Teng, Y.W., Ellis, J.M., Coleman, R.A., and Zeisel, S.H. (2012). Mouse betaine-homocysteine S-methyltransferase deficiency reduces body fat via increasing energy expenditure and impairing lipid synthesis and enhancing glucose oxidation in white adipose tissue. J. Biol. Chem. *287*, 16187–16198.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., and Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. *42*, D1001–D1006.

Xu, Z., Bolick, S.C., DeRoo, L.A., Weinberg, C.R., Sandler, D.P., and Taylor, J.A. (2013). Epigenome-wide association study of breast cancer using prospectively collected sister study samples. J. Natl. Cancer Inst. *105*, 694–700.

Zhu, Y., Soto, J., Anderson, B., Riehle, C., Zhang, Y.C., Wende, A.R., Jones, D., McClain, D.A., and Abel, E.D. (2013). Regulation of fatty acid metabolism by mTOR in adult murine hearts occurs independently of changes in PGC-1α. Am. J. Physiol. Heart Circ. Physiol. *305*, H41–H51.

Zou, J., Lippert, C., Heckerman, D., Aryee, M., and Listgarten, J. (2014). Epigenome-wide association studies without the need for cell-type composition. Nat. Methods *11*, 309–311.

Cell Metabolism, Volume *21*

Supplemental Information

Epigenome-Wide Association of Liver Methylation

Patterns and Complex Metabolic Traits in Mice

Luz D. Orozco, Marco Morselli, Liudmilla Rubbi, Weilong Guo, James Go, Huwenbo Shi, David Lopez, Nicholas A. Furlotte, Brian J. Bennett, Charles R. Farber, Anatole Ghazalpour, Michael Q. Zhang, Renata Bahous, Rima Rozen, Aldons J. Lusis, and Matteo Pellegrini

Figure S1, related to Experimental Procedures

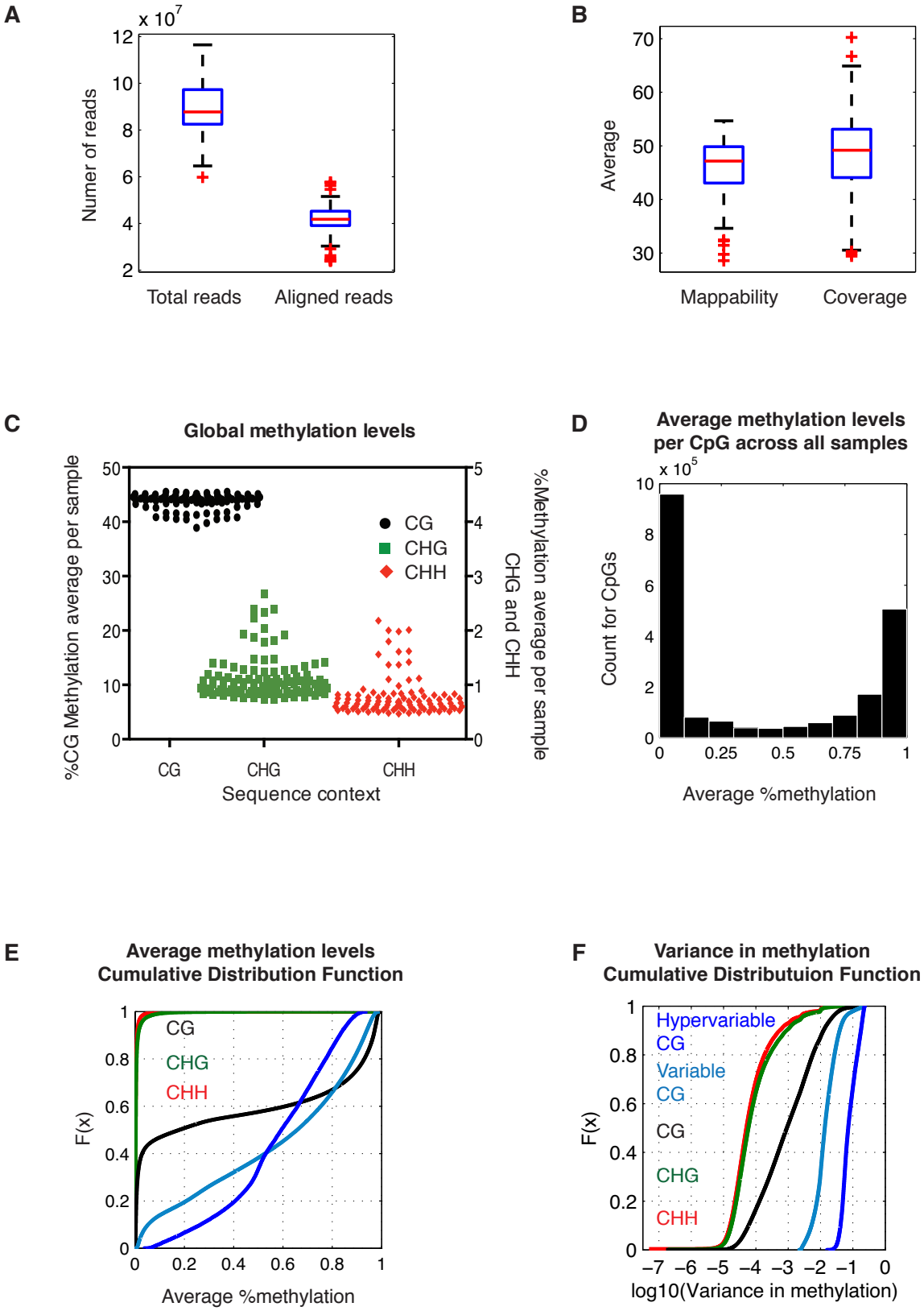


Figure S2, related to Experimental Procedures

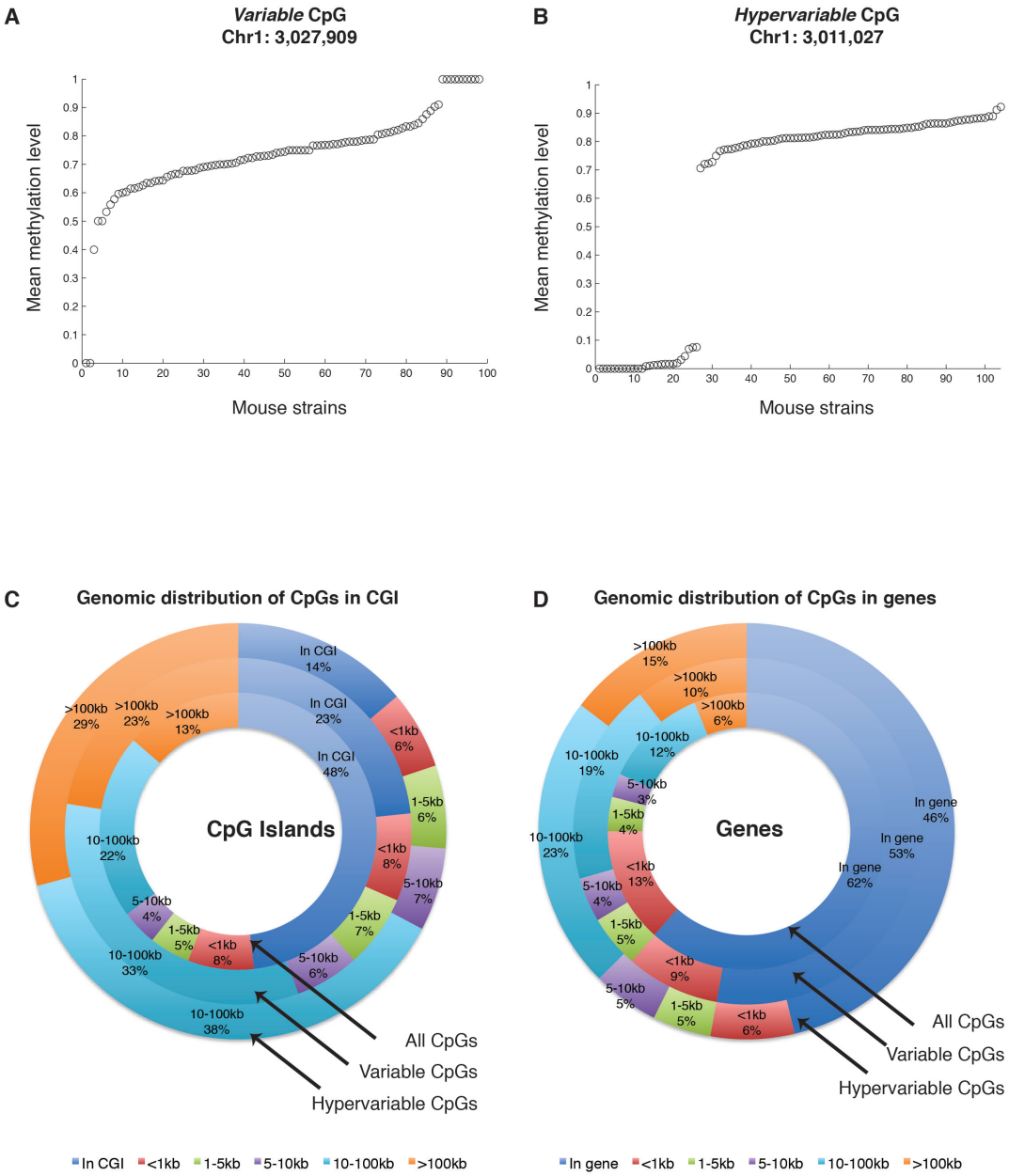


Figure S3, related to Figure 1

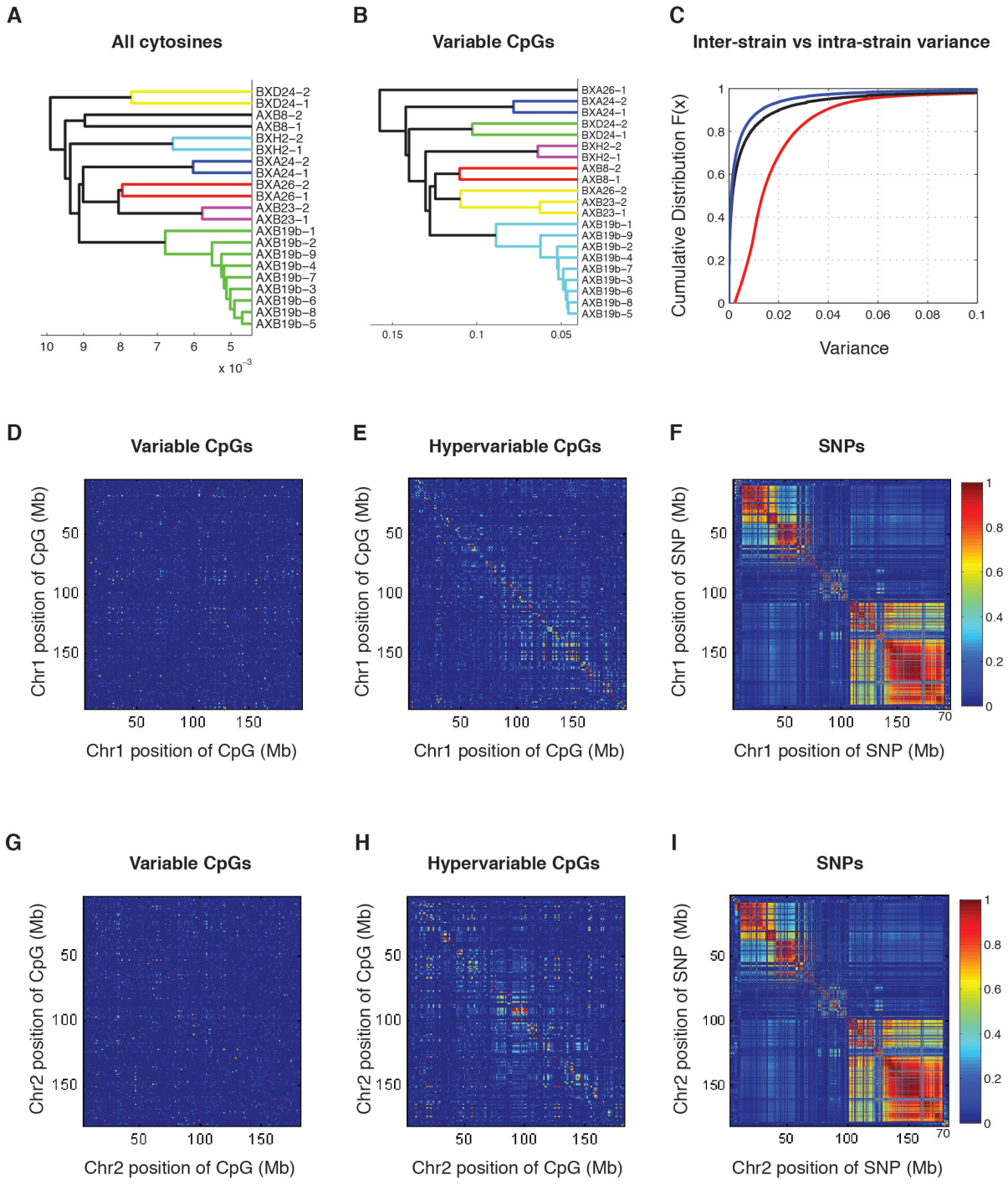
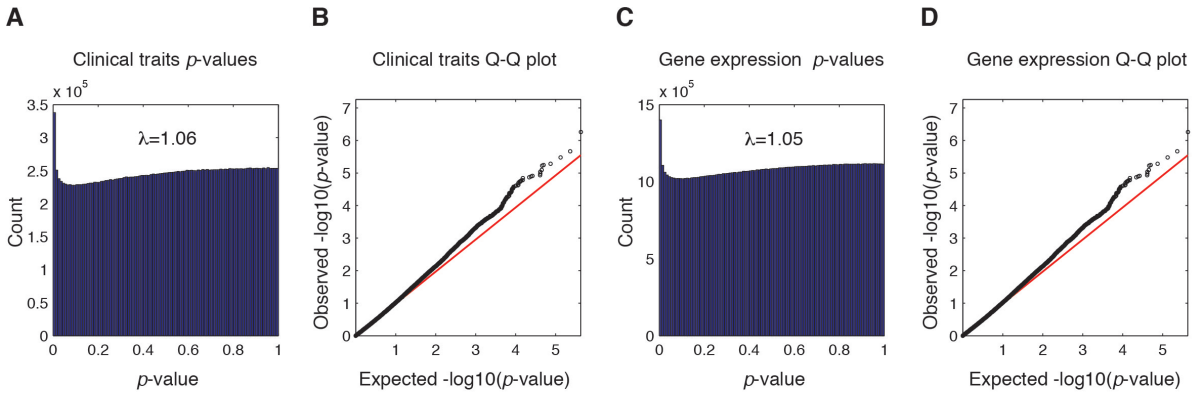
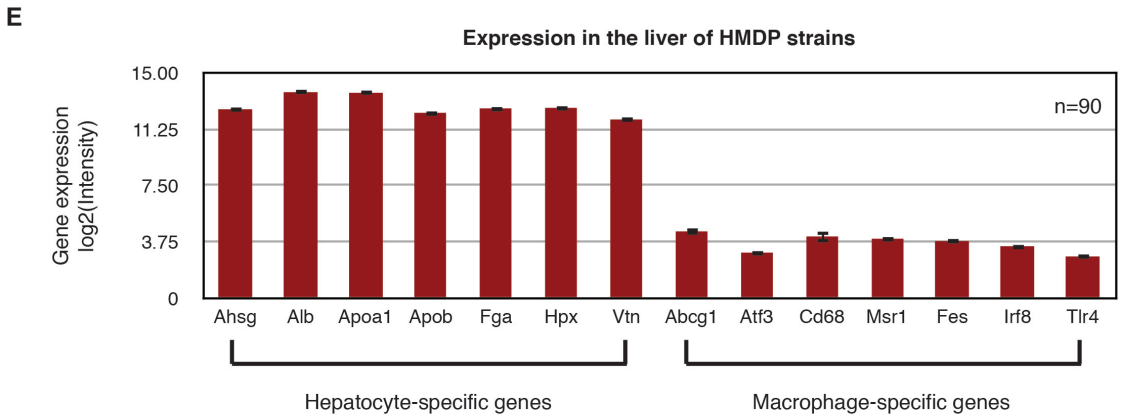


Figure S4, related to Figure 2





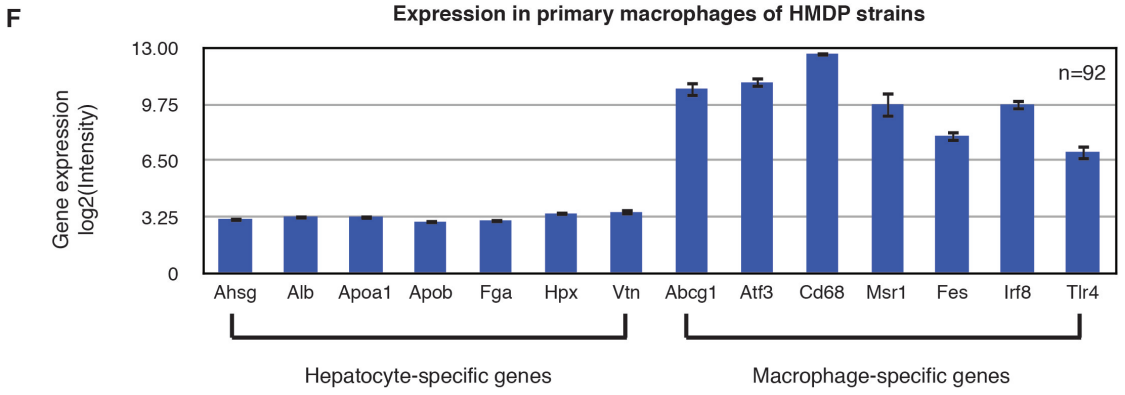
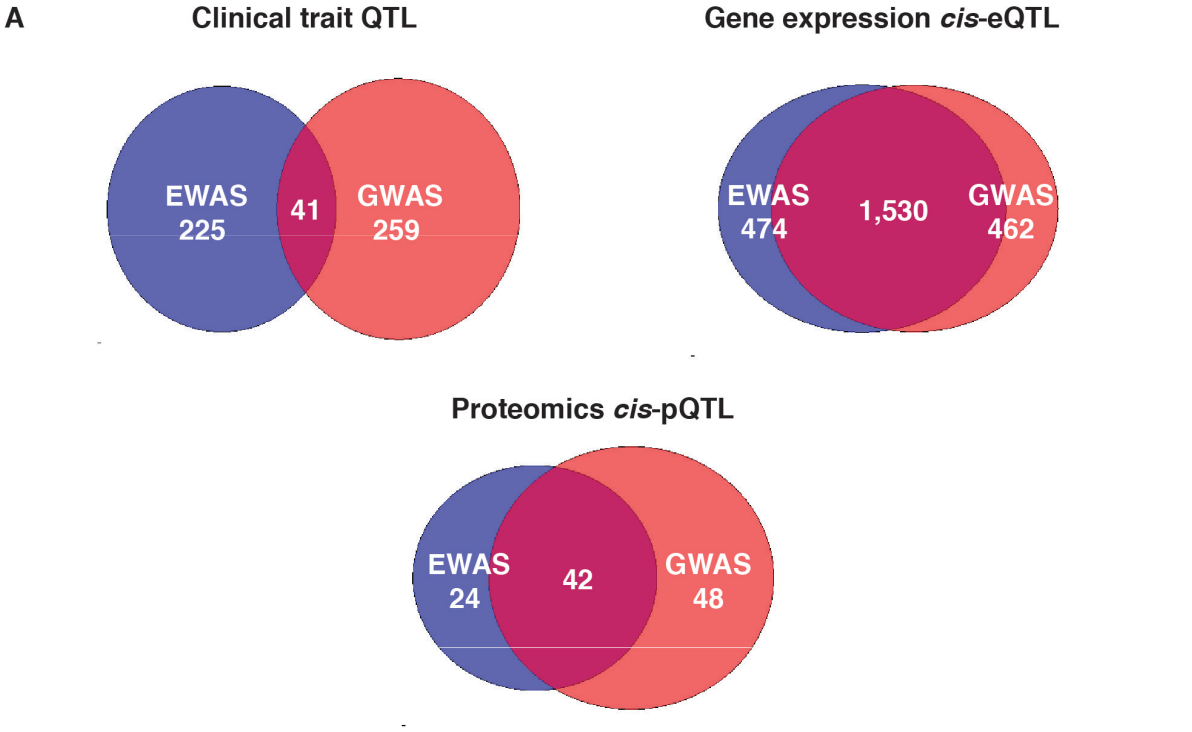


Figure S5, related to Figure 4



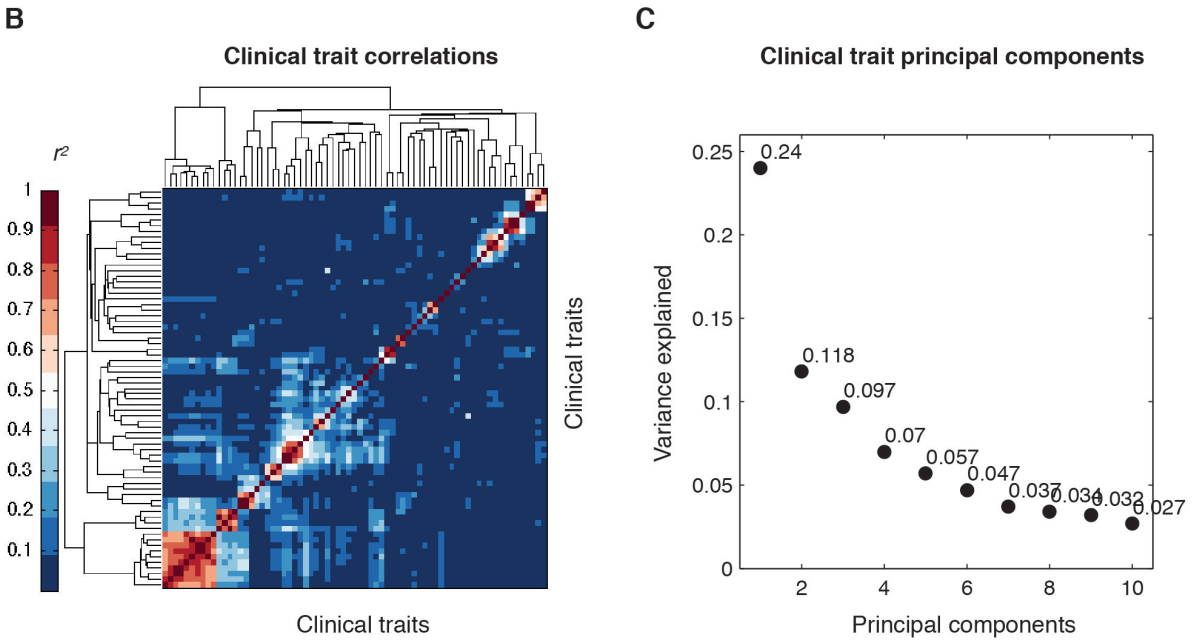


Figure S6, related to Figure 3

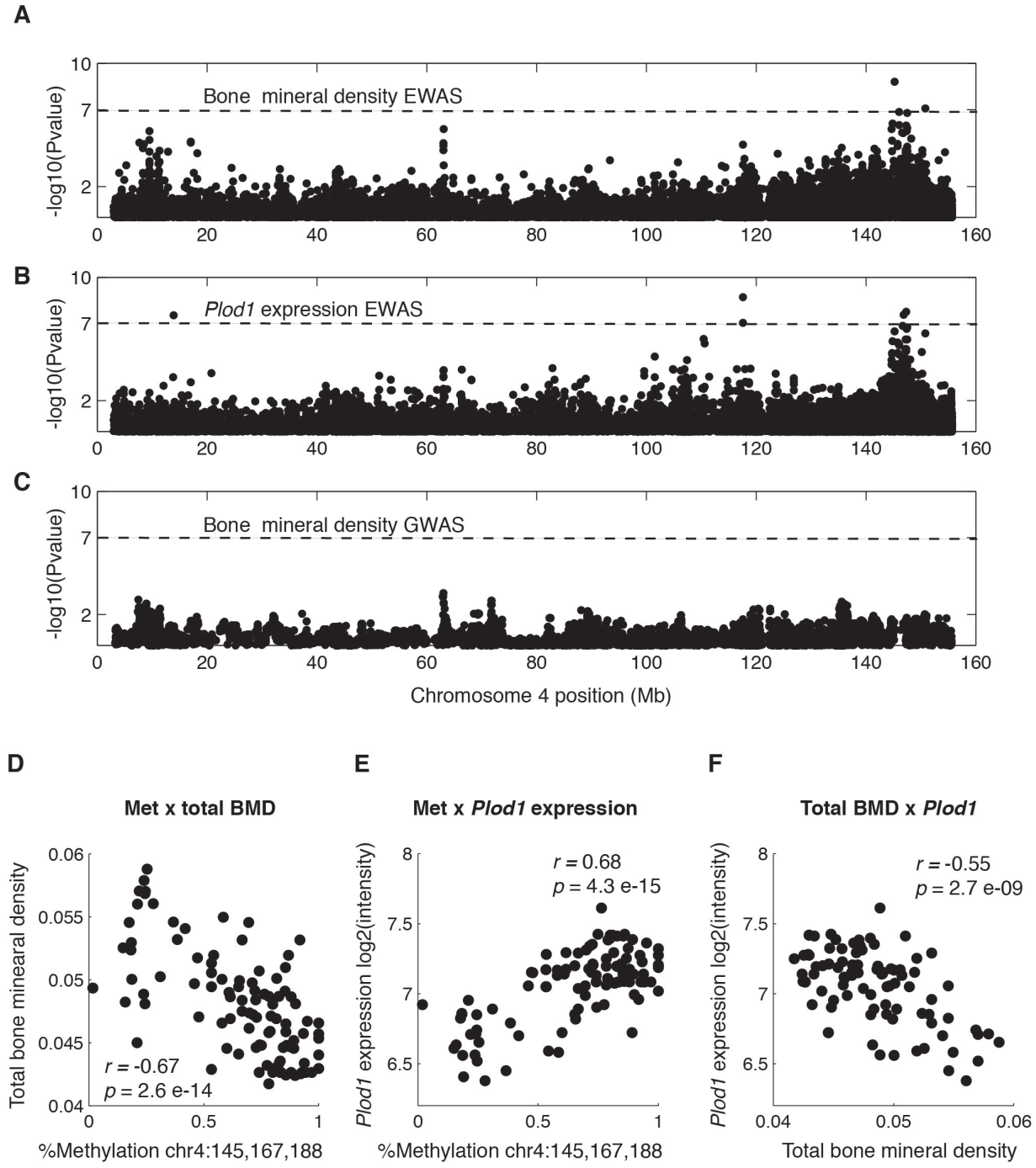
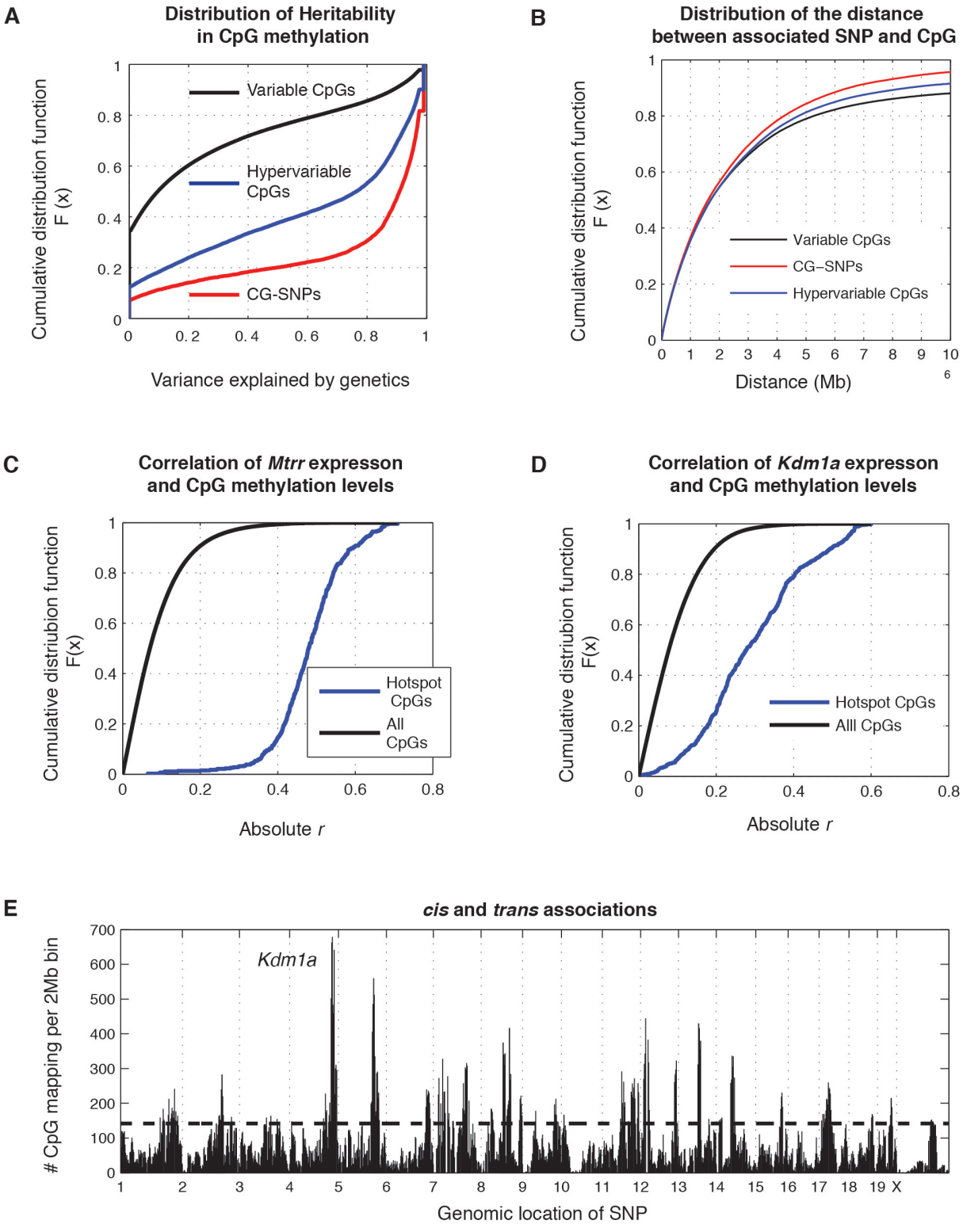


Figure S7, related to Figure 6



SUPPLEMENTAL FIGURE LEGENDS

Figure S1. Sample statistics. Related to Experimental Procedures. (A) Total number of reads and aligned reads across all RRBS mouse libraries. (B) Mappability and Coverage in all RRBS libraries. (C) Global percent methylation levels per sample in CG, CHG and CHH cytosine contexts. Each point represents an RRBS mouse library. H=any base other than G. (D) Histogram showing distribution of average methylation levels for each *Variable* CpG. (E) Distribution of the average methylation levels across samples for all cytosines. The x-axis is average methylation levels, and the y-axis is the cumulative distribution function F(x). (F) Distribution of the variance in methylation levels across samples for all cytosines. The x-axis is the $-\log_{10}$(variance), with increasing variance towards the right, and the y-axis is the cumulative distribution function F(x). Error bars on boxplots indicate the minimum and maximum values, excluding outliers.

Figure S2. Variable and Hypervariable CpGs. Related to Experimental Procedures. Methylation patterns in a (A) *Variable* CpG and (B) *Hypervariable* CpG. The x-axes are individual samples, where each dot is a sample. The y-axes are the methylation levels for the CpG in that sample. (C)-(D) Nested pie charts showing the percent fraction of CpGs found within or near (C) CpG islands, and (D) Genes. The inner plot shows all CpGs represented in our libraries, the second outer plot shows all *Variable* CpGs, the third outer plot shows *Hypervariable* CpGs.

Figure S3. Methylation reproducibility and correlations in full chromosomes. Related to Figure 1. (A) Clustering of replicate samples based on methylation levels of all cytosines covered in all samples, and (B) *Variable* CpGs only. (C) Distribution of the variance in methylation levels across all *Variable* CpGs, where the variance is shown on the x-axis, and the cumulative distribution function is plotted on the y-axis. The distribution of the variances among different mouse strains is shown in the red line (inter-strain variance), for biological replicates using different mice of the same strain is shown in the black line (intra-strain variance), and for technical replicates using different libraries of the same sample in the blue line. Correlation between CpG methylation levels or Linkage disequilibrium for SNPs in sample chromosomes. (D-F) Chromosome 1 and (G-I) Chromosome 2. The x and y axes denote the chromosome position, and the color represents the level of pairwise correlation (r^2) between CpGs, or SNPs.

Figure S4. EWAS *p*-value distributions and expression of tissue specific genes. Related to Figure 2. Histogram of the *p*-value distributions in the EWAS for (A) clinical traits, and (C) gene expression. Q-Q plots between the theoretical uniform distribution (red) and the EWAS *p*-value distribution (black) for a random phenotype for each phenotype category in (B) clinical traits, and (D) gene expression. Expression levels in HMDP strains for (E) liver tissue and (F) primary peritoneal macrophages. The y-axis shows the expression levels as the mean of all strains on a $\log_2$ scale, and the error bars show the standard deviation. The x-axis are genes known to be expressed in liver hepatocytes or macrophages.

Figure S5. Overlap of EWAS and GWAS associations and principal component analysis. Related to Figure 4. (A) Venn diagrams showing the number of associations identified using EWAS, GWAS or both for clinical traits, gene expression *cis*-eQTL, and proteomics *cis*-pQTL. Associations were considered to overlap if the associated CpG (EWAS) or SNP (GWAS) were within 2Mb of each other. (B) Pairwise correlations between clinical traits in our study. Traits are plotted on the x and y-axes. The color represents the Pearson's *r*-squared. (C) Variance in clinical phenotypes explained by the first 10 principal components.

Figure S6. Bone mineral density and *Plod1* EWAS. Related to Figure 3. Manhattan plots showing association of CpG methylation levels in mouse chromosome 4 for (A) bone mineral density, and (B) liver expression levels of *Plod1*. (C) Manhattan plot showing GWAS for bone mineral density. Chromosome location is shown on the x-axis, the *p*-value for the association is on the y-axis, and each dot represents a CpG. The dotted line is drawn at $p<1x10^{-7}$, the Bonferroni threshold for a single phenotype. (D)-(F) Each dot represents a mouse sample. (D) Correlation between methylation levels for the peak associated CpG (x-axis) and bone mineral density (y-axis). (E) Correlation between methylation levels for the peak associated CpG (x-axis) and liver expression levels of *Plod1* (y-axis). (F) Correlation between bone mineral density (x-axis) and *Plod1* expression levels (y-axis).

Figure S7. CpG methylation GWAS. Related to Figure 6. (A) Distribution of the narrow sense heritability in CpG methylation levels. The x-axis is the variance explained by genetics and the y-axis is the cumulative distribution function F(x) for the distribution of the variance. (B) Distribution of the distance between CpGs and SNPs for the CpG methylation GWAS

associations. The SNP-to-CpG distance is shown on the x-axis and the y-axis is the cumulative distribution function F(x) for the distance. (C) Distribution of the correlation between *Mtrr* gene expression and methylation levels of CpGs mapping to the *Mtrr* hotspot locus (blue). The distribution of the correlation between *Mtrr* expression and all CpGs is shown in black. The absolute Pearson's correlation coefficient *r* is shown on the x-axis, and the y-axis is the cumulative distribution function F(x) for the distribution of *r*. (D) Distribution of the correlation between *Kdm1a* gene expression and methylation levels of CpGs mapping to the *Kdm1a* hotspot locus (blue). The distribution for the correlation between *Kdm1a* expression and all CpGs is shown in black. The absolute Pearson's correlation coefficient *r* is shown on the x-axis, and the y-axis is the cumulative distribution function F(x) for the distribution of *r*. (E) CpG methylation GWAS hotspots. The y-axis denotes the total number of CpGs mapping, in *cis* and *trans,* to each 2Mb bin across the mouse genome, at the Bonferroni threshold $p<1.4\text{x}10^{-12}$. The x-axis denotes the genomic position of each bin. The horizontal dotted line is the Poisson significance threshold for each hotspot bin.

SUPPLEMENTAL TABLES

Table S1. EWAS counts for individual CpG-phenotype associations, Related to Experimental Procedures

Traits	Predictor	Traits x Predictors (# of tests)	Bonferroni alpha	#EWAS hits total	#EWAS hits *cis* (<=2Mb)	% EWAS hits *cis* (<=2Mb)
Clinical traits	Variable CpG excluding CG-SNP	68x360324	2.04E-09	214	NA	NA
Metabolites	Variable CpG excluding CG-SNP	260x360324	5.34E-10	164	NA	NA
Proteins	Variable CpG excluding CG-SNP	1543x360324	8.99E-11	13,334	1,499	11.2%
Gene expression	Variable CpG excluding CG-SNP	22416x360324	6.19E-12	124,522	22,741	18.3%
Clinical traits	Hypervariable CpG excluding CG-SNP	68x22227	3.31E-08	468	NA	NA
Metabolites	Hypervariable CpG excluding CG-SNP	260x22227	8.65E-09	490	NA	NA
Proteins	Hypervariable CpG excluding CG-SNP	1543x22227	1.46E-09	2,300	992	43.1%
Gene expression	Hypervariable CpG excluding CG-SNP	22416x22227	1.00E-10	34,198	15,742	46.0%

Table S2. Causal inference test for clinical trait associations, Related to Figure 2								
Trait	**EWAS *pval***	**EWAS CpG chr**	**EWAS CpG bp**	**GWAS *pval***	**GWAS SNP chr**	**GWAS SNP bp**	**GWAS SNP**	**CIT *pval***
High density lipoprotein cholesterol (HDL) in plasma	3.20E-09	1	173115750	1.12E-08	1	172213506	rs31465983	**4.44E-04**
Mean cell hemoglobin concentration (MCHC) in blood	3.40E-10	7	97591666	1.63E-08	7	96827691	rs6401951	0.78
Mean cell hemoglobin concentration (MCHC) in blood	9.50E-14	7	103359297	6.09E-11	7	103186910	rs31489892	0.38
Mean cell hemoglobin concentration (MCHC) in blood	7.74E-15	7	105287895	1.67E-13	7	105685575	rs31836285	0.07
Mean cell hemoglobin concentration (MCHC) in blood	3.39E-21	7	107641776	2.62E-17	7	107683538	rs3722049	0.26
Mean cell hemoglobin concentration (MCHC) in blood	2.31E-27	7	109650998	6.95E-23	7	108918190	rs3713052	0.62
Mean cell hemoglobin concentration (MCHC) in blood	3.81E-27	7	110584221	7.83E-24	7	110098213	rs31048502	0.34
Mean cell hemoglobin concentration (MCHC) in blood	1.54E-19	7	113527803	2.30E-17	7	112061872	rs31547013	0.14
Mean cell hemoglobin concentration (MCHC) in blood	4.16E-18	7	114353622	7.77E-17	7	114016970	rs13479450	0.96

Table S2. Causal inference test for clinical trait associations, Related to Figure 2								
Mean cell hemoglobin concentration (MCHC) in blood	2.32E-10	7	119433112	5.71E-09	7	119020177	rs31887032	0.85
Mean cell volume of red blood cells (MCV) in blood	3.63E-09	7	97591666	4.81E-07	7	96779987	rs31999165	0.58
Mean cell volume of red blood cells (MCV) in blood	3.62E-12	7	103359365	3.89E-10	7	103186910	rs31489892	0.95
Mean cell volume of red blood cells (MCV) in blood	5.09E-14	7	105287940	4.36E-12	7	105685575	rs31836285	0.25
Mean cell volume of red blood cells (MCV) in blood	3.23E-18	7	107641776	3.81E-15	7	107683538	rs3722049	0.28
Mean cell volume of red blood cells (MCV) in blood	4.57E-24	7	109650998	2.41E-21	7	109825010	rs31424200	0.12
Mean cell volume of red blood cells (MCV) in blood	9.50E-24	7	110584221	4.23E-23	7	110098213	rs31048502	0.76
Mean cell volume of red blood cells (MCV) in blood	6.18E-17	7	113527803	1.88E-15	7	112061872	rs31547013	0.31
Mean cell volume of red blood cells (MCV) in blood	1.71E-16	7	114353621	7.51E-15	7	114016970	rs13479450	0.05
Mean cell volume of red blood cells (MCV) in blood	2.97E-08	7	109650998	3.91E-07	7	109825010	rs31424200	**0.04**

Table S2. Causal inference test for clinical trait associations, Related to Figure 2								
Mean cell volume of red blood cells (MCV) in blood	7.26E-09	7	111693801	2.85E-07	7	110098213	rs31048502	**8.82E-03**
Mean cell volume of red blood cells (MCV) in blood	1.36E-08	7	114353621	2.19E-06	7	114016970	rs13479450	**9.83E-03**
Red blood cell distribution width as percent (RDW %) in blood	6.35E-09	7	107641776	1.55E-06	7	107884366	rs3665475	0.35
Red blood cell distribution width as percent (RDW %) in blood	2.36E-14	7	109379956	2.88E-13	7	109825010	rs31424200	**0.02**
Red blood cell distribution width as percent (RDW %) in blood	2.55E-14	7	111768244	5.75E-14	7	111015676	37-7-111015	0.15
Red blood cell distribution width as percent (RDW %) in blood	2.37E-11	7	113527803	2.71E-10	7	113954724	rs31378954	0.53
Red blood cell distribution width as percent (RDW %) in blood	1.08E-11	7	114353622	8.53E-11	7	114287225	rs31411962	0.32
Red blood cell distribution width in absolute number (RDWa) in blood	6.34E-11	7	103359366	4.82E-08	7	102789599	rs31678255	0.21
Red blood cell distribution width in absolute number (RDWa) in blood	3.14E-11	7	105287940	8.66E-10	7	105639548	rs32319939	0.56

Table S2. Causal inference test for clinical trait associations, Related to Figure 2								
Red blood cell distribution width in absolute number (RDWa) in blood	1.40E-13	7	107641776	6.13E-09	7	107683608	rs3722575	0.93
Red blood cell distribution width in absolute number (RDWa) in blood	3.25E-15	7	109650998	3.40E-13	7	109389815	rs6357312	0.77
Red blood cell distribution width in absolute number (RDWa) in blood	1.63E-16	7	110637216	1.12E-11	7	110098213	rs31048502	0.46
Red blood cell distribution width in absolute number (RDWa) in blood	1.41E-10	7	113527803	1.94E-08	7	112061872	rs31547013	0.71
Red blood cell distribution width in absolute number (RDWa) in blood	4.78E-11	7	114353621	3.46E-08	7	114287225	rs31411962	0.14
Total cholesterol in plasma	1.86E-08	1	173115750	1.48E-07	1	172213506	rs31465983	**5.53E-04**
Glucose-to-insulin ratio in plasma	6.37E-09	8	62624331	5.12E-07	8	62671482	rs33530751	0.17
Glucose-to-insulin ratio in plasma	3.06E-09	13	91473377	2.73E-07	13	90090833	rs29897087	**0.04**
Glucose-to-insulin ratio in plasma	6.61E-10	13	94074003	3.31E-07	13	95677756	rs29610795	**1.58E-03**
Glucose-to-insulin ratio in plasma	2.13E-09	13	97342942	1.40E-06	13	96030495	rs29738590	**7.11E-03**

Table S2. Causal inference test for clinical trait associations, Related to Figure 2								
Glucose-to-insulin ratio in plasma	5.73E-08	13	101382031	3.45E-06	13	101871731	rs29252234	0.36
Glucose-to-insulin ratio in plasma	4.74E-08	13	104403649	2.43E-07	13	105606102	rs6306099	0.12
Glucose-to-insulin ratio in plasma	3.23E-09	13	106755583	4.80E-07	13	106546115	rs29516615	**0.02**

Table S3. Candidates genes in EWAS hotspots, Related to Figure 2						
Hotspot traits	**Chr**	**Start of hotspot locus (Bp)**	**End of hotspot locus (Bp)**	**Candidate gene(s)**	**Function of candidate(s)**	***cis*-EWAS hit**
HDL, Total cholesterol	1	173,000,001	174,000,000	*Apoa2*	*Apoa2* is the second most abundant protein in HDL cholesterol	Protein *cis*-EWAS hit for *Apoa2*
Proteomics, Gene expression	4	150,000,001	151,000,000	*Mtor,Tnfrsf9, Camta1*	*Mtor* is a regulator of metabolism, growth and survival in response to hormones, growth factors, nutrients, energy and stress	Expression *cis*-EWAS hit for *Mtor*
Glucose-to-insulin ratio, Femoral fat pad %weight, Metabolites, Proteomics, Gene expression	7	88,000,001	89,000,000	***Cpeb1***, *Btbd1, Zfp592, Rps17*	*Cpeb1* is involved mRNA processing, regulated by DNA methylation, insulin signaling	
Hemoglobin concentration (MCHC), Percent ed blood cell in blood (HCT), Red blood cell volume (MCV), Red blood cell average and absolute width (%RDW, RDWa)	7	110,000,001	111,000,000	Hemoglobin beta cluster	Oxygen transport in red blood cells	
Free fluid, Metabolites, Proteomics, Gene expression	9	104000001	105000000	*Mrpl3*	Mitochondrial ribosomal protein	Expression *cis*-EWAS hit for *Mrpl3*
Femoral fat pad %weight, Metabolites, Proteomics, Gene expression	11	97,000,001	98,000,000	*Mrpl45*	Mitochondrial ribosomal protein	Expression *cis*-EWAS hit for *Mrpl45*
Glucose-to-insulin ratio, Femoral fat pad %weight, Monocyte percent in blood, Metabolites, Proteomics, Gene expression	13	81,000,001	82,000,000	*Gpr98,* ***Polr3g***, *Cetn3*	*Polr3g* DNA-directed RNA polymerase	

Table S3. Candidates genes in EWAS hotspots, Related to Figure 2						
Adipose tissue insulin resistance (ATIRI), Plasma insulin, Glucose-to-insulin-ratio, Monocyte percent in blood, Proteomics, Gene expression	13	94,000,001	95,000,000	*Bhmt*	Involved in lipid accumulation, fat metabolism, insulin sensitivity	Protein *cis*-EWAS hit for *Bhmt*
Femoral fat pad weight and %weight, Gonadal fat pad %weight, Body fat percent by NMR, Metabolites, Proteomics, Gene expression	14	14,000,001	15,000,000	*Thoc7*	RNA transport	Expression *cis*-EWAS hit for *Thoc7*
Monocyte percent in blood, Proteomics, Gene expression	14	53,000,001	54,000,000	*Tcra*	T-cell receptor alpha locus	

Table S4. Clinical trait inference, Related to Figure 5			
Clinical trait	**Observed and predicted trait correlation (r^2)**	***p*-value for r^2**	**Top CpGs chosen in model**
Mean cell volume of red blood cells (MCV) in blood	0.84	0.001	chr2:106433316, chr7:109650998, chr7:111693801, chr1:78198288, chr7:144808207, chr15:84673742, chr2:105699152, chr8:126506671, chr4:129427810, chr3:80307596
Mean cell hemoglobin concentration (MCHC) in blood	0.77	0.002	chr7:110637216, chr7:110584221, chr7:111627057, chr3:87411731, chr1:173211748, chr17:17823699, chr12:18352759, chr3:80426374, chr7:96907685, chr7:111693801
Red blood cell distribution width in absolute number (RDWa) in blood	0.71	0.004	chr2:106433316, chr7:109650998, chr16:85901018, chr7:104251059, chr7:111693801, chr6:93571975, chr7:38152759, chr7:110637216, chr3:80426374, chr5:142240807
Total cholesterol in plasma	0.61	0.009	chr17:93791057, chr1:173123267, chr15:19722260, chr5:24278274, chr1:158939766, chr6:3361731, chr4:9513779, chr1:22287020, chr7:107811213, chr14:121891692
High density lipoprotein cholesterol (HDL) in plasma	0.60	0.012	chr14:121689788, chr1:173123267, chr4:9513779, chr7:146362390, chr1:158939766, chr4:149571737, chr1:172018514, chr5:24190850, chr6:32784381, chr12:55828602
High density lipoprotein cholesterol log(HDL) in plasma	0.58	0.011	chr1:173123267, chr4:9513779, chr7:146362390, chr4:149571737, chr1:172018514, chr15:19722260, chr10:21993317, chr12:55828602, chr15:85863990, chr14:121689788
Total bone mineral density	0.57	0.013	chr7:20147544, chr6:107901150, chr11:28239859, chr16:51400929, chr3:123220640, chr1:159461569, chr9:41716166, chr8:86781301, chr4:147505750, chr7:19385586
Free fatty acids in plasma	0.50	0.023	chr7:78032906, chr4:59197781, chr16:68199356, chr3:121248386, chr19:6219728, chr14:16560588, chr14:31342640, chr16:32767085, chr17:44915174, chr10:89620236

Table S4. Clinical trait inference, Related to Figure 5			
Clinical trait	**Observed and predicted trait correlation (r^2)**	***p*-value for r^2**	**Top CpGs chosen in model**
Unesterified cholesterol in plasma	0.48	0.028	chr17:93791057, chr1:173123267, chr6:3361731, chr4:9513779, chr1:22287020, chr16:32767055, chr6:32784381, chr3:106968184, chr5:118079923, chr1:141566515
Glucose-to-insulin ratio in plasma	0.46	0.052	chr8:126680394, chr13:106755624, chr3:124695940, chr8:62156796, chr1:183802024, chr11:4566339, chr12:30828542, chr8:62624331, chr13:97342942, chr3:58566785
LDL plus VLDL cholesterol in plasma	0.44	0.054	chr15:19722260, chr17:93791057, chr5:118079923, chr1:106285591, chr15:6041729, chr15:8947724, chr1:141566515, chr1:173123267, chr5:24278274, chr5:114012265
Free fluid by NMR	0.42	0.042	chr17:35825528, chr9:88935896, chr8:90589021, chr10:124287267, chr5:138793552, chr5:140167600, chr1:188792013, chr1:48912457, chr4:107253570, chr4:148255721
Bone mineral density in lumbar spine L1-L6	0.42	0.045	chr12:58756641, chr6:107901150, chr4:140581221, chr12:14855022, chr7:19385586, chr2:168854983, chr4:17061832, chr14:53282704, chr4:133742504, chr6:137021617
Mean cell hemoglobin, i.e. average mass of hemoglobin per red blood cell (MCH) in blood	0.40	0.082	chr2:119486633, chr17:33229525, chr11:37795541, chr4:118063013, chr17:13607758, chr4:141022032, chr12:12687287, chr12:29483529, chr3:34865628, chr17:39981353
Glucose in plasma (colorimetric assay)	0.38	0.070	chr1:72640416, chr1:78826882, chr18:56877109, chr6:137638518, chr1:13462840, chr3:50506277, chr17:66672504, chr18:77085310, chr4:111608424, chr1:184087524
Red blood cell distribution width as percent (RDW%) in blood	0.37	0.094	chr11:83659976, chr7:111627057, chr7:111693801, chr3:129307299, chr15:85785313, chr2:136756323, chr4:136821943, chr17:43592013, chr4:129195413, chr4:129427810

Table S4. Clinical trait inference, Related to Figure 5			
Clinical trait	**Observed and predicted trait correlation (r^2)**	***p*-value for r^2**	**Top CpGs chosen in model**
Hematocrit, i.e. percentage of the volume in whole blood that consists of red blood cells (HCT)	0.35	0.123	chr7:110637216, chr12:12960282, chr7:110584221, chr1:166956547, chr5:92650222, chr7:107641776, chr3:27012064, chr17:33288773, chr3:80426374, chr10:89620236
Retroperitoneal fat pad percent weight	0.34	0.076	chr4:56230515, chr4:43435465, chr8:69045642, chr2:172168272, chr1:172974907, chr3:70482851, chr6:134591327, chr4:51031381, chr7:92130374, chr4:16362492
Retroperitoneal fat pad weight	0.34	0.077	chr3:31785637, chr11:36196077, chr5:147797209, chr2:172168272, chr1:172974873, chr19:6219728, chr4:16362492, chr9:47224582, chr12:50485430, chr1:158674522
Monocyte concentration in absolute number (MONO) in blood	0.33	0.155	chr13:56788932, chrX:90023204, chr17:45823427, chr12:29343801, chr12:104059623, chr3:136118906, chr7:39155188, chr18:69632285, chr1:34232571, chr15:41703296
Syndecan-1 ectodomains (percent relative to C57BL/6J) in plasma	0.32	0.182	chr4:124528418, chr7:32998073, chr5:55599332, chr12:120991166, chr13:75993475, chr12:113414824, chr3:113733631, chr2:170521944, chr10:125446784, chr4:135935109
Fat mass weight by NMR	0.31	0.117	chr4:16362492, chr11:36196077, chr12:50485430, chr3:31785637, chr9:100195211, chr6:138521149, chr1:191697765, chr19:6219728, chr14:14117106, chr5:147797209
Monocyte concentration as percent (MONO%) in blood	0.30	0.127	chr15:52858818, chrX:73649226, chr10:124769898, chr18:9680388, chr11:90238164, chr2:26009521, chr3:126199837, chr2:20821494, chr1:130537022, chr4:149302743
Femur bone mineral density	0.30	0.103	chr7:20147544, chr6:107901150, chr4:136928341, chr14:53282704, chr15:57123066, chr13:96698837, chr4:134955904, chr4:139137044, chr6:32784306, chr17:39981353

Table S5. GWAS for CpG methylation levels, Related to Figure 6	
Traits	Methylation of Variable CpG
Predictor	SNP, MAF>10%
TraitsxPredictors (# of tests)	367317x94498
Bonferroni alpha	1.44E-12
#Associations total	3,017,453
#Associations (<=1Mb)	1,088,729
%Associations <=1Mb)	36.1%
#Associations *cis* (<=2Mb)	1,645,835
%Associations *cis* (<=2Mb)	54.5%
#Associations (<=5Mb)	2,382,320
%Associations (<=5Mb)	79.0%
#*Trans* Associations (>2Mb)	1,371,618
#*Trans* Associations (>2Mb) in same chromosome	1,079,442
%*Trans* Associations (>2Mb) in same chromosome	78.7%
#SNP-CpG pairs	3,017,453
#Individual SNPs	92,959
#Individual CpGs	26,563
#CpGs that are CG-SNP	2,533
#CpGs that are Hypervariable	11,644

Table S6. *Mtrr* validation results, Related to Figure 6								
CpG chr	**CpG bp**	**FDR**	***p*-value**	**Methylation delta**	**Met mean in WT**	**Met mean in gt/gt**	**Met mean in HMDP allele C strains***	**Met mean in HMDP allele A strains***
1	14075228	84.17%	4.30E-01	0.01	0.91	0.92	0.79	0.91
1	14075294	45.59%	3.05E-01	0.05	0.86	0.91	0.65	0.87
1	14075350	76.79%	4.07E-01	0.01	0.91	0.92	0.67	0.91
1	18068341	25.55%	2.29E-01	0.15	0.91	0.76	0.66	0.87
1	18068657	55.16%	3.40E-01	0.08	0.80	0.72	0.50	0.75
1	18068658	97.92%	4.61E-01	0.00	0.75	0.75	0.49	0.75
1	22881270	3.92%	6.38E-02	0.17	0.68	0.85	0.68	0.88
1	22881330	35.20%	2.83E-01	0.05	0.80	0.84	0.72	0.91
1	85044352	56.60%	3.46E-01	0.04	0.92	0.96	0.80	0.90
1	152185173	29.97%	2.62E-01	0.11	0.55	0.67	0.61	0.81
1	152185232	15.04%	1.66E-01	0.15	0.64	0.79	0.66	0.91
1	152185244	2.73%	4.77E-02	0.16	0.66	0.82	0.58	0.89
1	168870254	97.70%	4.63E-01	0.00	0.44	0.44	0.53	0.87
1	183712801	53.05%	3.33E-01	0.03	0.97	0.94	0.95	0.89
2	82881478	91.59%	4.49E-01	0.01	0.85	0.84	0.77	0.91
2	96875718	0.02%	9.23E-04	0.42	0.46	0.88	0.50	0.88
2	96875732	0.02%	9.92E-04	0.28	0.61	0.89	0.63	0.93
2	96875779	0.39%	9.59E-03	0.29	0.63	0.92	0.58	0.87
2	96875812	1.06%	2.17E-02	0.33	0.55	0.88	0.56	0.91
3	77509441	6.50%	8.62E-02	0.14	0.94	0.80	0.81	0.90
3	96210033	18.11%	1.85E-01	0.11	0.53	0.42	0.61	0.84
3	96210101	69.77%	3.84E-01	0.05	0.43	0.48	0.46	0.76
3	96210202	54.22%	3.38E-01	0.04	0.77	0.74	0.77	0.94
3	96210204	28.93%	2.56E-01	0.10	0.73	0.63	0.66	0.91
3	96210213	45.17%	3.05E-01	0.06	0.70	0.64	0.72	0.90
3	96210216	34.39%	2.83E-01	0.09	0.53	0.45	0.49	0.78
3	159337823	90.45%	4.53E-01	0.02	0.79	0.77	0.60	0.81
3	159337868	70.63%	3.86E-01	0.06	0.66	0.72	0.56	0.81
3	159337906	39.37%	2.82E-01	0.08	0.76	0.84	0.56	0.83
4	68825234	67.47%	3.80E-01	0.05	0.93	0.88	0.76	0.86
5	33103043	0.00%	4.69E-05	0.76	0.97	0.20	0.78	0.12
6	14496933	91.25%	4.51E-01	0.01	0.56	0.55	0.53	0.71
6	26784829	7.30%	9.33E-02	0.15	0.56	0.71	0.58	0.83
6	60645050	0.05%	1.74E-03	0.53	0.94	0.41	0.93	0.49
6	68312345	35.37%	2.78E-01	0.07	0.25	0.19	0.32	0.47
6	69443856	8.53%	1.05E-01	0.16	0.61	0.77	0.58	0.78
6	69753111	19.31%	1.89E-01	0.14	0.85	0.98	0.79	0.93
6	69753160	69.28%	3.87E-01	0.03	0.93	0.90	0.65	0.89

Table S6. *Mtrr* validation results, Related to Figure 6								
6	69753161	83.97%	4.32E-01	0.01	0.94	0.94	0.66	0.90
6	69753165	15.04%	1.63E-01	0.11	0.71	0.83	0.43	0.74
6	69753166	78.18%	4.12E-01	0.01	0.81	0.79	0.43	0.73
6	69753189	18.87%	1.88E-01	0.20	0.83	0.64	0.33	0.63
6	69753209	94.27%	4.56E-01	0.01	0.75	0.76	0.55	0.78
6	69753210	78.69%	4.11E-01	0.02	0.83	0.85	0.60	0.79
6	69753226	35.35%	2.81E-01	0.04	0.97	0.93	0.67	0.91
6	69753246	37.68%	2.78E-01	0.07	0.95	0.88	0.58	0.90
6	145590674	47.59%	3.13E-01	0.03	0.86	0.89	0.78	0.93
7	13258547	25.17%	2.28E-01	0.10	0.74	0.85	0.66	0.82
7	55387401	65.07%	3.79E-01	0.01	0.01	0.02	0.01	0.24
7	55387420	59.93%	3.64E-01	0.01	0.03	0.04	0.04	0.49
7	55387445	64.76%	3.80E-01	0.01	0.03	0.04	0.05	0.42
8	4380072	74.59%	4.02E-01	0.02	0.91	0.93	0.77	0.94
8	19045211	33.88%	2.85E-01	0.13	0.85	0.72	0.60	0.85
8	36320023	13.73%	1.54E-01	0.12	0.73	0.85	0.71	0.89
8	112537572	67.03%	3.81E-01	0.03	0.90	0.87	0.86	0.54
9	17813589	64.30%	3.80E-01	0.03	0.84	0.88	0.73	0.91
9	30679901	6.47%	8.73E-02	0.12	0.82	0.94	0.62	0.87
9	89373252	1.01%	2.12E-02	0.14	0.78	0.92	0.72	0.89
9	109322097	5.52%	8.23E-02	0.05	0.82	0.87	0.83	0.59
9	109322126	21.87%	2.09E-01	0.14	0.50	0.64	0.58	0.43
10	27569592	49.89%	3.22E-01	0.05	0.75	0.80	0.78	0.31
10	29160856	21.89%	2.06E-01	0.07	0.80	0.86	0.64	0.84
10	69590376	0.04%	1.40E-03	0.46	0.25	0.71	0.23	0.74
10	69590377	0.01%	5.31E-04	0.43	0.15	0.58	0.21	0.70
10	69590413	0.00%	3.56E-04	0.48	0.14	0.62	0.19	0.65
10	69590440	5.08%	7.91E-02	0.20	0.05	0.24	0.11	0.34
10	99062510	0.77%	1.72E-02	0.32	0.43	0.11	0.43	0.87
10	99062571	5.14%	7.83E-02	0.28	0.43	0.15	0.40	0.85
10	112826327	38.63%	2.82E-01	0.08	0.76	0.84	0.64	0.85
11	3093388	0.06%	1.96E-03	0.45	0.63	0.18	0.55	0.28
11	3093439	0.02%	9.12E-04	0.46	0.83	0.37	0.72	0.32
11	3093448	0.03%	1.24E-03	0.41	0.87	0.46	0.79	0.38
12	25042018	0.02%	9.50E-04	0.47	0.43	0.90	0.59	0.95
12	25042037	0.04%	1.50E-03	0.35	0.32	0.67	0.41	0.82
13	14657134	12.40%	1.43E-01	0.16	0.56	0.72	0.48	0.70
13	14657167	40.81%	2.78E-01	0.04	0.87	0.83	0.70	0.88
13	14657168	22.56%	2.10E-01	0.07	0.82	0.89	0.68	0.89
13	14657188	34.86%	2.84E-01	0.07	0.86	0.93	0.74	0.92
13	19530472	40.14%	2.79E-01	0.07	0.84	0.90	0.53	0.86
13	19530485	69.75%	3.87E-01	0.06	0.47	0.41	0.15	0.44

Table S6. *Mtrr* validation results, Related to Figure 6								
13	19530597	45.64%	3.03E-01	0.10	0.61	0.72	0.28	0.63
13	19531062	66.62%	3.82E-01	0.04	0.91	0.87	0.67	0.87
13	19531155	7.22%	9.40E-02	0.16	0.86	0.70	0.61	0.83
13	50623749	92.13%	4.49E-01	0.01	0.58	0.57	0.52	0.67
13	59127984	13.28%	1.51E-01	0.02	0.98	1.00	0.92	0.68
13	60537835	0.65%	1.49E-02	0.76	0.92	0.16	0.87	0.58
13	60565529	0.29%	7.75E-03	0.83	0.97	0.15	0.88	0.60
13	60639030	0.87%	1.89E-02	0.81	0.99	0.18	0.91	0.52
13	60826611	39.33%	2.84E-01	0.03	0.96	0.99	0.91	0.58
13	62406458	36.52%	2.78E-01	0.01	0.92	0.91	0.93	0.84
13	62406495	18.00%	1.87E-01	0.17	1.00	0.83	0.96	0.85
13	62406504	16.77%	1.77E-01	0.04	0.96	0.91	0.89	0.77
13	62988985	60.82%	3.66E-01	0.01	0.09	0.08	0.66	0.42
13	64002236	0.01%	3.86E-04	0.83	0.83	0.00	0.74	0.02
13	64088377	0.03%	1.21E-03	0.60	0.13	0.73	0.19	0.76
13	64299068	33.11%	2.86E-01	0.03	0.95	0.97	0.96	0.69
13	64435227	36.65%	2.76E-01	0.07	0.74	0.81	0.71	0.85
13	64496672	5.81%	8.32E-02	0.29	0.36	0.07	0.41	0.17
13	65360332	0.20%	5.42E-03	0.25	0.25	0.00	0.46	0.02
13	65691239	3.65%	6.08E-02	0.22	0.19	0.40	0.23	0.46
13	66781925	8.21%	1.03E-01	0.12	0.48	0.60	0.49	0.66
13	66807216	23.41%	2.15E-01	0.19	0.15	0.34	0.46	0.61
13	67515142	6.33%	8.72E-02	0.18	0.74	0.57	0.79	0.71
13	67522055	2.79%	4.76E-02	0.20	0.32	0.12	0.31	0.21
13	67729395	0.00%	2.86E-06	0.97	0.97	0.00	0.91	0.04
13	68850423	0.00%	3.74E-04	0.91	0.91	0.00	0.89	0.03
13	69595580	0.02%	9.17E-04	0.78	0.78	0.00	0.83	0.02
13	69595591	0.19%	5.53E-03	0.63	0.63	0.00	0.62	0.02
13	69674062	61.97%	3.70E-01	0.00	0.01	0.01	0.70	0.09
13	69811585	0.00%	5.47E-06	0.82	0.84	0.02	0.87	0.02
13	69811641	0.00%	2.53E-07	0.92	0.94	0.02	0.96	0.05
13	69878173	NaN	NaN	0.00	0.00	0.00	0.48	0.09
13	69878797	NaN	NaN	0.00	0.00	0.00	0.02	0.24
13	69890792	0.00%	1.37E-04	0.70	0.70	0.00	0.70	0.04
13	69890887	0.00%	1.23E-05	0.91	0.92	0.01	0.93	0.07
13	69910712	0.00%	2.64E-04	0.83	0.84	0.01	0.76	0.02
13	69944654	20.55%	1.99E-01	0.09	0.66	0.75	0.65	0.08
13	78498140	1.69%	3.10E-02	0.19	0.48	0.67	0.53	0.83
13	82668080	96.92%	4.63E-01	0.00	0.94	0.93	0.86	0.79
13	96260119	5.66%	8.27E-02	0.56	0.76	0.21	0.85	0.38
14	121048631	39.44%	2.80E-01	0.09	0.68	0.76	0.63	0.81
15	49676164	51.04%	3.26E-01	0.07	0.82	0.90	0.69	0.86

Table S6. *Mtrr* validation results, Related to Figure 6								
15	79285383	15.68%	1.68E-01	0.13	0.08	0.20	0.10	0.20
15	79285446	33.96%	2.83E-01	0.12	0.28	0.40	0.35	0.46
16	64872950	1.13%	2.25E-02	0.27	0.23	0.49	0.25	0.62
16	64872952	2.25%	4.02E-02	0.26	0.19	0.45	0.18	0.55
16	68548140	72.17%	3.91E-01	0.04	0.91	0.87	0.52	0.77
16	68548141	96.52%	4.64E-01	0.00	0.74	0.74	0.49	0.71
16	68548142	74.65%	3.99E-01	0.04	0.75	0.79	0.43	0.70
16	68548163	49.76%	3.24E-01	0.04	0.94	0.90	0.59	0.80
16	68548164	39.62%	2.78E-01	0.09	0.87	0.78	0.58	0.80
16	68548167	84.84%	4.31E-01	0.01	0.85	0.84	0.52	0.75
16	68548168	66.34%	3.83E-01	0.05	0.86	0.81	0.54	0.76
16	68548225	91.24%	4.54E-01	0.01	0.93	0.93	0.77	0.90
16	83450561	10.16%	1.23E-01	0.06	0.91	0.85	0.82	0.95
16	83450563	79.70%	4.13E-01	0.00	0.88	0.88	0.72	0.92
17	16879208	51.54%	3.27E-01	0.04	0.47	0.52	0.48	0.64
17	32280497	18.25%	1.84E-01	0.10	0.03	0.13	0.27	0.43
17	32280583	36.08%	2.78E-01	0.09	0.11	0.19	0.27	0.40
18	4228525	0.60%	1.43E-02	0.32	0.71	0.38	0.66	0.50
18	41624633	1.48%	2.87E-02	0.23	0.68	0.91	0.61	0.93
18	41624728	1.68%	3.17E-02	0.34	0.57	0.92	0.53	0.82
18	41625156	0.09%	2.77E-03	0.40	0.56	0.96	0.55	0.88
19	39323373	0.31%	8.04E-03	0.44	0.47	0.92	0.74	0.91
19	39323464	6.33%	8.89E-02	0.24	0.60	0.84	0.76	0.92
19	39323480	0.06%	2.00E-03	0.30	0.35	0.65	0.53	0.82
X	36160452	86.84%	4.38E-01	0.02	0.57	0.55	0.51	0.67
X	37992988	33.31%	2.84E-01	0.20	0.37	0.57	0.36	0.65
X	47819925	5.03%	8.01E-02	0.27	0.49	0.75	0.43	0.71
X	47819937	10.83%	1.27E-01	0.15	0.59	0.74	0.63	0.79
X	82912686	37.35%	2.79E-01	0.07	0.82	0.89	0.75	0.93
X	108294397	35.57%	2.77E-01	0.04	0.91	0.96	0.91	0.79
X	111652978	10.36%	1.24E-01	0.03	0.87	0.90	0.80	0.94
X	152164563	40.51%	2.79E-01	0.07	0.82	0.75	0.85	0.74
*Genotype corresponds to rs13481861								

SUPPLEMENTAL INFORMATION

Data

The average CG methylation levels for all were 40 times higher than CHG methylation (KS-test $p<1x10^{-16}$), and 55 times higher than CHH methylation ($p<1x10^{-16}$, Figure S1E). The variance in methylation levels across the samples was on average 7 and 10 times higher in CG cytosines than CHG (KS-test $p<1x10^{-16}$) or CHH (KS-test $p<1x10^{-16}$), respectively (Figure S1F). To define *Variable* and *Hypervariable* CpGs, we selected a change in methylation (delta) of 50% or more since we observed that selecting smaller deltas lead to a high false positive discovery rate. Using simulation, we previously determined that CpGs falsely identified as differentially methylated increases as the delta in methylation decreases (Orozco et al., 2014). In addition, we (Figure S1D) and others have observed that the distribution of CpG methylation levels in mammals is largely bimodal (Chen et al., 2011; Meissner et al., 2008), where CpG methylation levels appear to be on or off for a large proportion of CpGs. Therefore, we wanted to focus on CpGs with a low false positive discovery rate and which were more likely to be biologically relevant. We observed that *Variable* and *Hypervariable* CpGs tended to be further away from genes relative to all CpGs. For example, 62% of all CpGs were intragenic while 53% of *Variable* and 46% of *Hypervariable* CpGs were intragenic. The location of CpGs relative to CpG islands and genes is shown in Figure S2C-D.

Mappability of bisulfite sequencing data

The average mapping efficiency of 46% we observed is reasonable for RRBS libraries. We and others have perviously observed that mapping efficiencies are lower for RRBS libraries relative to whole genome libraries (Chatterjee et al., 2012; Doherty and Couldrey, 2014; Guo et al., 2013). We also aligned a sample from the current mouse RRBS dataset using different aligners and observed that the mapping efficiency was comparable using BS-Seeker2 (47.29%), Bismark (46.96%) and BSMAP (45.64%).

RRBS methylation data is reproducible

We examined reproducibility in our dataset by comparing methylation levels in biological replicates for a subset of the mouse strains, using RRBS libraries from different mice of the same strain, as well as in technical replicates, using different RRBS libraries from the same DNA sample. Different mice of the same strain are genetically identical, like monozygotic (MZ)

twins, but unlike MZ twins they did not share a prenatal environment. The technical replicates allowed us to examine experimental variation not due to true biological differences among the samples. We clustered samples based on their methylation levels using data from all cytosines (Figure S3A) and *Variable* CpG cytosines (Figure S3B), and found that samples from the same strain cluster together in both. We compared the distribution of the variance in methylation levels across CpG cytosines, and found that the variance in methylation between different mouse strains, or inter-strain variance, was on average 2 times higher than the variance in biological replicates, or intra-strain variance (KS-test $p<1x10^{-16}$). Furthermore, the variance in CpGs among different strains was 3.3 times higher than the variance in technical replicates (KS-test $p<1x10^{-16}$, Figure S3C). Technical replicates measured for different RRBS libraries of the same DNA sample were highly correlated with $r^2=0.99$. We have previously validated RRBS data relative to traditional bisulfite sequencing, by cloning DNA fragments into bacterial colonies followed by Sanger sequencing, and found a high degree of concordance between RRBS and traditional bisulfite sequencing results (Chen et al., 2013; Orozco et al., 2014).

EWAS inflation and purity of liver tissue samples

To examine inflation in our EWAS results, we computed the inflation factor lambda, where lambda values over 1 indicate inflation, lambda values under 1 indicate deflation, and lambda of 1 indicates neither. We observed no evidence of inflation, with lambda values of 1.06 for clinical traits, 1.07 for metabolites, 1.06 for proteomics, and 1.05 for gene expression associations. The *p*-value distribution and qqplots for sample phenotypes are shown on Figure S4A-D. We confirmed that our liver samples were derived primarily from hepatic cells by examining expression of hepatocyte-specific and macrophage-specific genes. Liver samples of HMDP strains had high expression levels for hepatocyte-specific genes such as Alpha2-HS glycoprotein (*Ahsg*), albumin (*Alb*), apolipoproteins (*Apoa1* and *Apob*), fibrinogen (*Fga*), hemopexin (*Hpx*) and vitronectin (*Vtn*), and virtually undetectable expression levels of genes highly expressed in macrophages such as *Abcg1*, *Atf3*, *Cd68*, *Msr1*, *Fes*, *Irf8* and *Tlr4* (Figure S4E). As a control, we show that primary peritoneal macrophages from the HMDP strains show high expression levels of macrophage genes, but not hepatocyte specific genes (Figure S4F).

EWAS identifies both known and novel associations

We identified an association for plasma high-density lipoprotein cholesterol levels (HDL) in distal chromosome 1 at 173.1Mb ($p=3.2x10^{-09}$), where methylation levels at the locus were

correlated with HDL (r=-0.67, p=9.6x10^{-15}). This result was consistent with a GWAS hit for HDL at this locus. A candidate gene underlying this association is *Apoa2*, which is the second most abundant lipoprotein in HDL cholesterol particles. We found a *cis* association for protein levels of *Apoa2* using EWAS (p=9.5x10^{-08}), methylation was inversely correlated with Apoa2 protein levels (r=-0.51, p=2.3x10^{-07}), and *Apoa2* was correlated with plasma HDL (r=0.51, p=1.9x10^{-07}). We and others have previously identified a genetic association for HDL cholesterol at the same locus using GWAS in the HMDP strains (Bennett et al., 2010) and linkage in a mouse cross (Wang et al., 2007), and shown that altered protein levels of Apoa2 influence plasma HDL cholesterol levels (Warden et al., 1993).

We also found an association between methylation and total bone mineral density (BMD) on distal chromosome 4, even though there was no significant GWAS hit for BMD on this chromosome. We searched for candidate genes in the locus and found a *cis* association for expression levels of procollagen-lysine, 2-oxoglutarate 5-dioxygenase, *Plod1,* suggesting that expression levels of *Plod1* were variable in the population and regulated in *cis* (Figure S6A-C). Methylation levels at the locus were correlated with BMD (r=-0.67, p=2.6x10^{-14}) and *Plod1* expression levels (r=0.68, p=4.3x10^{-15}), and *Plod1* expression was correlated with the BMD trait (r=-0.55, p=2.7x10^{-9}, Figure S6D-F). These results suggest that *Plod1* is an ideal candidate gene for the association between methylation levels and BMD, and indeed *Plod1* has previously been shown to play a role in bone mineral density in humans (Tasker et al., 2006).

Although we did not measure DNA methylation levels in bone, we found that total bone mineral density was associated with liver methylation levels in chromosome 4. A candidate gene for this association was *Plod1*, since expression of *Plod1* was also associated in *cis* at this locus (Figure S6). *Plod1* catalyzes the hydroxylation of lysine residues in procollagen molecules, a critical step in collagen synthesis. Procollagen molecules are exported from the cell at a later stage during collagen synthesis, but there is no evidence that collagen molecules are transported to bone tissue. A possible explanation for this association is that cleavage of secreted collagen molecules that enter the circulation may serve as signaling peptides. An alternative and more likely explanation is that methylation levels at this locus are conserved between liver and bone tissue.

Causal inference test using CG-SNPs

We also examined associations using SNPs that abolish a CpG site (i.e. CG-SNPs). These CG-SNPs alter methylation levels by changing the cytosine base of a CpG to another

base, and hence can no longer be methylated. We can potentially identify associations between traits and the SNP genotype using GWAS and/or the methylation levels of these CG-SNPs using EWAS. Furthermore, we can test whether SNPs mediate their effect through altered DNA methylation levels using the causal inference test. We identified 79 associations between clinical traits and methylation levels of CG-SNPs using EWAS. However, none of these were associated with the SNP genotype using GWAS, and the causal inference test did not support association between SNP and trait mediated by DNA methylation. It is possible that we did not identify significant GWAS associations due to the very small minor allele frequency of these CG-SNPs, since the majority of the CG-SNPs were present in only one strain. Similarly, the 79 clinical trait associations to methylation levels may in fact be spurious associations and we would not pursue these going forward with our studies.

EWAS Hotspots

Previous genetics and genomics studies have identified and validated QTL hotspots, where a genetic polymorphism(s) at a locus affects many traits. Hotspots can help us identify genes that function as global regulators of gene expression and clinical traits. They can be seen as vertical bands on genome-wide association plots, and we observed several such bands in our EWAS results (Figure 2). To find hotspots, we divided the genome into 1Mb bins and counted the number of associations between methylation levels in that bin and traits. We identified association hotspots for clinical traits, metabolites, proteins, and gene expression traits, and observed that many of the hotspots were shared among the different types of traits. For example, a hotspot on chromosome 7 at 88Mb was associated with metabolites, proteomics, gene expression, glucose-to-insulin ratio and femoral fat pad weight (Table S3). A candidate gene underlying this associations is the cytoplasmic polyadenylation element binding protein 1 (*Cpeb1*), which was located in the 1Mb interval of this locus. *Cpeb1* is a gene involved in mRNA processing, insulin signaling and insulin resistance (Alexandrov et al., 2012), and is itself regulated by DNA methylation (Xiaoping et al., 2013). Another hotspot on chromosome 4 was associated with proteomics and gene expression. The Mechanistic target of rapamycin, *Mtor,* is a candidate gene for this hotspot since its expression levels map to methylation levels in *cis* (i.e. *cis*-eQTL), and several genes known to interact with *Mtor* also mapped to the locus, such as *Rictor*. The hotspot on chromosome 7 at 110Mb coincides with the Hemoglobin beta locus, and was associated with multiple blood cell phenotypes such as hemoglobin

concentration, percent red blood cells, red blood cell volume and size. A table listing the top EWAS hotspots, and candidate genes for the hotspots, can be found in Table S3.

DNA methylation GWAS hotspots

We found a QTL hotspot regulating methylation levels of ~100 CpGs across the genome on chromosome 12 at 26Mb (Figure 6C). A plausible candidate for this association is the gene *Klf11*, found 1.3Mb from the hotspot. *Klf11* is a transcription factor involved in tumor suppression and metabolic disease (Lomberk et al., 2012), and it is known to couple to histone acetyltransferase and histone methyltransferase chromatin remodeling pathways in transcription regulation (Seo et al., 2012).

When we examined QTL hotspots controlling CpGs in *cis* and *trans*, we found several loci that primarily influence methylation levels of nearby CpGs (Figure S7E). For example, we found such a "local" methylation hotspot in chromosome 4 at 136Mb, roughly 100kb from the lysine-specific histone demethylase 1A, *Kdm1a*. Expression levels of *Kdm1a* were correlated with methylation levels of CpGs mapping to the locus with an average absolute r=0.29. The distribution of these correlations was significantly different (KS-test p=1.2x10^{-205}) from the correlation between *Kdm1a* expression and all CpGs, with average absolute r=0.09 (Figure S7D). In summary, our results show that natural genetic variation can influence both local and distant CpG methylation levels across the genome.

Note on experimental validation of candidate gene *Mtrr*

We experimentally tested our hypothesis that *Mtrr* was influencing CpG methylation levels across the genome, using bisulfite sequencing data of *Mtrr* wild-type and homozygous gene trapped mice (gt/gt). We confirmed the role of *Mtrr* in 27% CpGs predicted to be affected by the chromosome 13 hotspot that were differentially methylated between wild-type and gt/gt mice (Figure 6E). It is possible that we were able to validate only 27% of all CpGs predicted to map to the chromosome 13 hotspot due to lack of power in our validation studies, since we examined three *Mtrr* wild-type and and three gt/gt mice. Furthermore, gt/gt mice had decreased expression and activity of *Mtrr,* but it was not completely absent (Elmore et al., 2007). It is also possible that a fraction of the CpGs predicted to map to the hotspot locus are false positives. Alternatively, we hypothesize that the chromosome 13 hotspot is complex, like previously described hotspot loci, such that there may be more that one gene that is causally related to CpG methylation levels.

Comparison of mouse and human DNA methylation

To determine how DNA methylation in mice compared to human methylation profiles, we obtained a public dataset from normal human liver from GEO (GSM916049). We selected this dataset since it was generated with the same sequencing technology as the mouse data presented here, and because it corresponded to an adult liver human sample similar to our adult mouse liver samples, although the human dataset was a whole genome bisulfite sequencing library preparation and our mouse data were RRBS. We used *liftover* to obtain the mouse chromosomal locations that corresponded to the human genetic loci. There were 51,435,834 cytosines represented in the human methylation sample on both plus and minus strands, and we identified 17,140,641 syntenic mouse positions using *liftover*. From these, 80,123 cytosines were also represented in our mouse data. We compared human methylation levels to the average mouse methylation levels across all our mouse samples and found they were correlated with Pearson's r=0.62 (p<1x10^{-16}). The mean methylation level across all sites was 31.2% in human and 37.6% in mouse, although the median methylation level in the human sample was 5.6%, compared to the median mouse methylation level of 30.4%. As we might expect, there are both similarities and differences between human and mouse methylation patterns. However, it is challenging to compare DNA methylation profiles from different specifies for biological and technical reasons. For example, although a large proportion of the mouse genome is syntenic to the human genome, we cannot always find a one to one concordance between human and mouse genetic loci. In addition, variables such as diet, environment, age, library preparation, bisulfite conversion protocols, coverage, whole genome versus RRBS libraries and batch effects can all contribute to variability in DNA methylation levels.

Online databases for the identification of candidate genes

Candidate gene identification. Our database incorporates: (i) EWAS and GWAS results for clinical traits, metabolites, gene expression and proteomics data in our study, (ii) gene annotations for candidate gens using PubMed publications, (iii) published GWAS associations from the online GWAS catalog (Welter et al., 2014), and (iv) the Citeline database to provide information on existing drugs targeting candidate genes. This database can be accessed at http://ewas.mcdb.ucla.edu. Our database allows the user to query associations for clinical traits, and to identify candidate genes based on proximity to the associated CpG. The user has the option search for genes associated with a trait by EWAS and/or GWAS. In addition,

one can narrow down candidate genes by selecting genes with *cis* associations for gene expression (i.e. *cis*-eQTL), and/or protein levels (*cis*-pQTL).

Our database also incorporates gene annotations based on all PubMed publications. This functionally allows us to narrow down candidate genes that have been previously implicated with the clinical trait, or other traits related to it. For example, suppose we are searching for candidate genes for a clinical trait measured in the HMDP, such as “Plasma insulin levels”. In the online database we can select from a list of curated terms that are related to insulin, such as *body mass index, glucose*, *diabetes*, *insulin*, *insulin resistance, islet cell, leptin,* etc. Then the search for candidate genes will include publications where a given gene and the curated term were found together in a publication, either in the abstract or the full text, including publication links, and the total number of publications (PMID count) where the gene and the curated term were found together. Alternatively, we can enter our own term or list of terms to be used for the search instead of choosing from the curated terms. The online database also provides information for any known GWAS association between the candidate genes and clinical traits, by incorporating all published hits from the online GWAS catalog (Welter et al., 2014). Finally, for each candidate gene we provide information on drugs known to target the gene, and whether they are currently in clinical trials based on the Citeline database.

Association graphs. Cellular and organismal phenotypes arise from the concerted action of thousands genes, transcription factors, genetic, epigenetic and environmental variation. To help us understand and visualize how different cellular markers such as gene expression, protein levels, metabolite levels, genetic and epigenetic associations work together to influence clinical phenotypes, we created a different online tool that generates association graphs based on our results. This website allows a user to select a clinical trait of interest and a *p*-value threshold and displays associations between the trait and CpGs at the given threshold. Each of the clinical trait associations is further extended to the associations for gene expression, protein and metabolites to these CpGs using EWAS. We provide a graph to visualize connections between clinical traits, gene expression, proteins, metabolites, and CpGs that allows us to more easily identify how these different cellular systems are interconnected, and how they interact with each other based on epigenetic associations. As an example, we generated a bone mineral density graph (Figure 7) which displays associations for the trait and individual CpGs. These CpGs were also associated with gene with expression and protein levels. Genes such as *Plod1*, *Igf1*, and *Mtor* are known to affect bone mineral density and/or bone biology, while other genes are involved in lipid or fat metabolism pathways, which are

often correlated with bone density. Additional genes in the association graph are not known to directly impact bone mineral density, such as *Cds2,* a gene involved in calcium metabolism, *Masp2* a calcium-binding gene involved in complement activation, *Yy1*, a transcription factor which inhibits bone morphogenic protein (Kurisaki et al., 2003), but their membership in this graph allows us to confidently hypothesize that these genes in fact can influence bone mineral density and/or bone metabolism. This tool can can accessed at http://pathways.mcdb.ucla.edu/network.

Novel SNPs in 90 mouse inbred strains

Eighteen mouse inbred strains and wild-derived strains have been sequenced to date (Keane et al., 2011). We used RRBS data with 48X average coverage to identify SNPs on 90 mouse classical inbred and recombinant inbred strains, 12 of which have been previously sequenced by the Mouse Genome Project. We note that we included the *Mus musculus castaneus* strain CAST/EiJ in the SNP analyses but not in the EWAS or GWAS, since it is genetically widely divergent from the other *M. m. domesticus* mouse strains and would confound our association studies. Overall, we identified 135,213 SNPs with 20X coverage or better on both strands, consisting of 42,031 known SNPs and 93,182 novel SNPs. Approximately 45% of SNPs (60,943) were present in a single mouse strain, and 26% of SNPs (35,731) were present in more than ten percent of the samples. We used *SnpEff* (Cingolani et al., 2012) to annotate the SNPs and found 10,327 missense, 56 nonsense and 14,264 silent SNPs.

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Data access. All RRBS sequencing and SNP data can be obtained from GEO: GSE67507. The EWAS and GWAS results can be accessed in our online databases to search for candidate genes at http://ewas.mcdb.ucla.edu and to generate association graphs at http://pathways.mcdb.ucla.edu/network. Individual tables with all methylation associations can be downloaded from http://ewas.mcdb.ucla.edu/download.html. The GWAS results can also be accessed at http://systems.genetics.ucla.edu/data/hmdp.

Mice and sample collection. Male mice were purchased from Jackson Labs (Bar Harbor, Maine, USA) between 6 and 10 weeks of age. To ensure adequate acclimatization to a common environment the mice were aged until 16 weeks of age. All mice were maintained on a chow diet (Ralston-Purina Co., St. Louis, MO, USA) until sacrificed at 16 weeks of age. Following an overnight 16-hour fast, mice were bled retro-orbitally under isoflurane anesthesia and euthanized by cervical dislocation in the morning between the hours of 8am and 12pm. Livers were dissected out and flash frozen in liquid nitrogen. All animals were handled in strict accordance with good animal practice as defined by the relevant national and local animal welfare bodies, and all animal experiments and work were carried out with UCLA IACUC approval.

HMDP mouse data. Clinical traits were measured in HMDP strains using 8-12 mice per strain. Expression array profiling was performed on liver tissue using three mice per strain (Bennett et al., 2010). We measured expression in primary peritoneal macrophages in four mice per strain using cells incubated overnight in 20% FBS DMEM media, followed by a 4 hour incubation in 1% FBS DMEM (Orozco et al., 2012), as previously described. We measured proteomics data using Liquid Chromatography–Mass Spectrometry in one mouse per strain (Ghazalpour et al., 2011). We profiled metabolite data using one or two mice per strain (Ghazalpour et al., 2014). Detailed protocols for all phenotype measurements can be found at http://systems.genetics.ucla.edu/protocols/hmdp and http://systems.genetics.ucla.edu/protocols/hmdp_secondset.

RRBS Libraries. We prepared RRBS libraries as previously described (Smith et al., 2009), with minor modifications. Briefly, we isolated genomic DNA from flash frozen livers using a phenol-chloroform extraction, digested 1µg of DNA with MspI restriction enzyme (NEB, Ipswich, MA, USA), carried out end-repair/adenylation (NEB) and ligation with TruSeq barcoded adapters (Illumina, San Diego, CA, USA). We selected DNA fragments of size range 200-300bp with AMPure magnetic beads (Beckman Coulter, Brea, CA, USA), followed by bisulfite treatment on

the DNA (Millipore, Billerica, MA, USA), and PCR amplification (Bioline, Taunton, MA, USA). We sequenced the libraries by multiplexing two libraries per lane in an Illumina HiSeq sequencer, with 100bp reads. We made one RRBS library per strain for the majority of the strains, 2 libraries per strain for AXB23/PgnJ, AXB8/PgnJ, BXA24/PgnJ, BXA26/PgnJ, BXH2/TyJ and BXD24/TyJ strains, and 9 libraries for strain AXB19b/PgnJ.

Alignment. We aligned the reads with BS-Seeker2 (Guo et al., 2013) to the mm9 mouse reference genome. We used Bowtie as the base aligner, trimmed adapters, allowed for up to 5 mismatches and selected uniquely aligned reads.

Reproducibility of RRBS data. We clustered the replicates using hierarchical clustering and a 'correlation' distance metric. To cluster the samples, we selected cytosines with no missing data, which corresponded to 5,366,593 sites for all cytosines (Figure S3A) and 150,144 sites for *Variable* CpGs (Figure S3B). To examine the distribution of the variance in methylation levels, we computed the variance in percent methylation levels for each cytosine across all samples, using either samples from all strains for the variance between-strains (inter-strain variance), or samples from different mice of the same strain for the within-strain variance (intra-strain variance), or from technical replicates using different libraries made from the same DNA sample. We plotted the empirical cumulative distribution of these variances and compared the distributions with the Kolmogorov-Smirnov test (KS-test), and took the mean of each variance distribution to compare the fold difference of the distributions.

Selection of CpGs. We observed a total of 47,063,780 cytosines with RRBS coverage in at least one strain. From this, we selected 11,520,175 cytosines present in at least 90% of the samples, with coverage of 10x or more, which corresponded to 2,047,165 CG, 2,737,475 CHG, and 6,735,535 CHH cytosines. We identified a set of 367,317 CpGs which show a change in methylation level (delta) of 50% or more, in at least one strain. We excluded 6,993 sites from the EWAS studies, since these coincided with SNPs that abolished the CpG site in mouse strains carrying the SNP, resulting in 360,324 *Variable* CpGs. We also identified a set of 22,227 *Hypervariable* CpGs which show a delta in methylation of at least 50%, between 5 or more samples and the median methylation level for all samples.

Linkage disequilibrium and CpG correlation studies. We computed the Pearson's *r*-squared between pairs of SNPs, or pairs of CpGs, excluding missing values. To determine the average *r*-squared, we calculated the distance in base-pairs between pairwise CpGs or SNPs, then selected all pairwise r^2 values between CpGs/SNPs that were found with 100kb of each other,

and computed the average r^2 at that distance. We then repeated this process for increasing pairwise distance bins, such as 200kb, 300kb, etc.

EWAS. We used the linear mixed model package pyLMM (https://github.com/nickFurlotte/pylmm) to test for association and to account for population structure and relatedness among the mouse strains. This method was previously described as EMMA (Kang et al., 2008), and we implemented the model in python to allow for continuous predictors, such as CpG methylation levels that vary between 0 and 1. We applied the model: **$y=\mu+x\beta+u+e$,** where μ=mean, x=CpG, β=CpG effect, and u=random effects due to relatedness, with $Var(u) = \sigma_g^2K$ and $Var(e) = \sigma_e^2$, where K=IBS (identity-by-state) matrix across all *Variable* CpGs. We computed a restricted maximum likelihood estimate for σ_g^2K and σ_e^2, and we performed association based on the estimated variance component with an F-test to test that β does not equal 0. Each phenotype was log transformed for the association test.

Inflation. We calculated the inflation factor lambda by taking the chi-squared inverse cumulative distribution function for the median of the association *p*-values, with one degree of freedom, and divided this by the chi-squared probability distribution function of 0.5 (the median expected *p*-value by chance) with one degree of freedom. Since it was not feasible to calculate this statistic using all *p*-values for the gene expression dataset, we calculated lambda using a sample of 108 million *p*-values, corresponding to *p*-values for 300 randomly selected probes. For the remaining datasets, we used the entire *p*-value distribution. We plotted qqplots for representative phenotypes using the *qqplot* function in Matlab, with a theoretical uniform distribution with parameters 0,1.

Overlap of EWAS and GWAS. We defined an overlap between EWAS and GWAS if the associations were found within 2Mb (Figure S5A). To decrease the chance of not finding an overlap based on our stringent Bonferroni EWAS thresholds, we used the per phenotype Bonferroni threshold of $p<1x10^{-7}$ for EWAS, and $p<4.1x10^{-6}$ for the GWAS as previously described (Bennett et al., 2010).

Published GWAS. We previously performed GWAS in the HMDP for clinical traits and microarray expression levels (Bennett et al., 2010), proteomics (Ghazalpour et al., 2011) and metabolomics (Ghazalpour et al., 2014). For all these associations, we employed the EMMA linear mixed model, using SNPs with at least 10% minor allele frequency and missing data in less than 10% of the samples, and selected significant associations where $p<4.1x10^{-6}$ as previously described (Bennett et al., 2010).

Conditional EWAS. We performed EWAS for clinical traits or *cis* expression associations identified with both EWAS and GWAS. We used the pyLMM package as described with one modification: for each EWAS we used the SNP genotype for the GWAS hit as a covariate.

Causal inference test. We performed causal inference tests using the R statistical package CIT developed by Millstein and colleagues (Millstein et al., 2009), according to the user's manual.

EWAS Hotspots. We divided the genome into 1Mb bins and counted the number of unique traits, or metabolites, or genes, with a significant association in that bin. We only considered associations from *Hypervariable* CpGs, at the corresponding Bonferroni significance threshold. We used the Poisson distribution to determine if individual bins had a higher than expected number of associations. A given bin was considered a significant hotspot if the number of unique associated traits in that bin was above 3 for clinical traits, 5 for metabolites, 6 for proteins, and 20 for the gene expression.

PCA. We performed a principal component analysis on the clinical traits. The first and second principal components explained 24% and 12% of the variation in the traits, respectively. We mapped the first two principal components as traits to CpG methylation levels across the genome using EWAS as described above.

Methylation GWAS. We tested for association between methylation levels as phenotypes, and SNPs as predictors using EMMA as previously described (Bennett et al., 2010). The difference between the EWAS model described above, and the GWAS linear mixed model is that in GWAS x=SNP, β=SNP effect, and K=IBS (identity-by-state) matrix across all SNPs. Inbred strains were previously genotyped by the Broad Institute (http://www.broadinstitute.org/mouse/hapmap), and they were combined with the genotypes from Wellcome Trust Center for Human Genetics (WTCHG). Genotypes of RI strains at the Broad SNPs were inferred from WTCHG genotypes by interpolating alleles at polymorphic SNPs among parental strains, calling ambiguous genotypes missing. Of the 140,000 SNPs available, 94,498 were informative with an allele frequency greater than 10% and missing values in less than 10% of the strains.

Heritability. We estimated the narrow sense heritability using a linear mixed-model approach (Yang et al., 2010). We assume each phenotype $\mathbf{y}$ follows the model $\mathbf{y} = \mathbf{1}_n\boldsymbol{\mu} + \mathbf{u} + \mathbf{e}$, where the random variable $\mathbf{u}$ follows a normal distribution centered at zero with variance $\sigma_g^2\mathbf{K}$, and $\mathbf{e}$ represents an independent noise component with variance σ_e^2. The matrix K is estimated using Identity by State (IBS) across all SNPs. For each trait we estimated σ_g^2 and σ_e^2 using REML and calculated the heritability as $h^2 = \sigma_g^2/(\sigma_g^2 + \sigma_e^2)$.

Methylation GWAS Hotspots. We divided the genome into 2Mb bins and counted the number of all unique CpGs with a significant GWAS hit in that bin and called these "*cis* and *trans*" associations (Figure S7E). We also defined a set of *trans* association hotspots (Figure 6C), where we counted CpGs mapping to each bin in *trans,* such that the CpG was physically located at least 10Mb away from the bin. We considered CpGs to be associated at the Bonferroni threshold with $p<1.4x10^{-12}$. We used the Poisson distribution to determine if individual bins had a higher than expected number of associations. A given bin was considered a significant hotspot if the number of unique CpGs mapping to that bin was above 142 for "*cis* and *trans*" GWAS, and 30 for *trans* GWAS. We used 2Mb bins instead of 1Mb bins because of the increased LD in the mouse SNPs used for GWAS.

Validation of *Mtrr* Hotspot. We generated RRBS libraries from *Mtrr* gene trapped mice (Elmore et al., 2007), using three wild-type and three homozygous gene trapped (gt/gt) male mice at three months of age. We sequenced the libraries by multiplexing all six libraries in one lane. We aligned the data using BS-Seeker2 as described above, and filtered the data by selecting only CpGs covered by 10 or more reads. Of the 471 CpGs predicted to be affected by *Mtrr*, 154 were represented in this dataset. We were not able to observe all 471 CpGs because of (1) the decrease in coverage, since we multiplexed six samples in one lane for the validation experiments and two samples per lane for all the HMDP samples, and (2) the inherent randomness of sequencing data in RRBS using the Illumina sequencing technology. We compared CpG methylation levels in +/+ and gt/gt mice using a *t*-test, and estimated the FDR using the Storey method (Storey, 2002). We calculated the difference in methylation levels at each CpG by taking the absolute difference in methylation between the average methylation *Mtrr* +/+ and the average in -/- mice, i.e. delta methylation. We examined the distribution of the methylation difference, or delta for (1) all CpGs, (2) the 154 CpGs predicted to be affected by *Mtrr*, and (3) random sets of CpGs. We compared one distribution to another using the Kolmogorov-Smirnov test. We selected random sets of CpGs from all CpGs observed in the RRBS dataset 1,000 times, and compared the distribution of the delta in *Mtrr* +/+ and gt/gt mice for each random set.

Phenotype inference. We used the *glmnet* package in R for building linear models, which fits a generalized linear model via penalized maximum likelihood (Friedman et al., 2010). For each trait, we randomly selected test sets consisting of 10 mice which were hidden from the training dataset, and used the remaining mice for model building. We selected the 20,000 most variable CpGs in the training set mice as features, and built a linear model based on these features. We

then used the linear model to infer traits on the test set, and measured the accuracy of the trait predictions relative to the measured clinical traits, by taking the Pearson's *r* between predicted and measured clinical trait values. We repeated this process ten times with ten different test sets. The list of inferred phenotypes, the correlations between predicted and observed phenotypes, and the top ten CpGs selected most frequently to model each phenotype can be found in Table S4. We used lasso regularization by setting the elastic-net penalty parameter *alpha* to 1, and selected the *lambda* value that minimized cross-validation error for each trait, where *lambda* is the tuning parameter that controls the overall strength of the penalty. We replaced missing data in with data from the closest mouse according to euclidean distance.

Association graphs. We defined edges in the BMD association graph based on EWAS for CpGs and clinical traits, metabolites, proteins and gene expression, for *Hypervariable* CpGs where $p<1x10^{-7}$. Associations were considered in *cis* if the distance between a gene and the CpG was arbitrarily within 5Mb. We also identified edges between CpGs and SNPs based on the methylation GWAS between CpG methylation traits and SNPs, at the Bonferroni significance threshold $2.4x10^{-11}$, using *Hypervariable* CpGs and 94,498 SNPs. Edges between clinical traits and SNPs were based on our published GWAS (Bennett et al., 2010; Ghazalpour et al., 2011), where $p<4.1x10^{-6}$. We constructed the figures using Cytoscape (www.cytoscape.org).

SNP calling. We developed a method to predict SNPs from the ATCGmap file generated by BS-Seeker2. First, we selected the sites covered by 20X reads on both strands, ensuring our prediction would not be affected by poor sampling. Second, as a T in the read can correspond to a T or unmethylated C in the genome in bisulfite sequencing, we re-calculated the counts supporting A, T, C or G calls, such that counts supporting C = (#Ts + #Cs). Third, we tested if counts support each nucleotide by chance using a Binomial test, assuming sequencing error rate = 0.2, and then we called nucleotides where $p<0.01$ for a given nucleotide at each position. Fourth, to avoid mapping bias between the two strands, we used the intersection of the predicted set of nucleotides from both strands to be the predicted polymorphism at that site. Fifth, we compared the polymorphisms with the reference genome to determine whether the site was a SNP, and whether it was a homogeneous or heterogeneous SNP. Finally, we annotated SNP categories and functional consequences using *SnpEff* (Cingolani et al., 2012).

SUPPLEMENTAL REFERENCES

Alexandrov, I.M., Ivshina, M., Jung, D.Y., Friedline, R., Ko, H.J., Xu, M., O'Sullivan-Murphy, B., Bortell, R., Huang, Y.T., Urano, F.*, et al.* (2012). Cytoplasmic polyadenylation element binding protein deficiency stimulates PTEN and Stat3 mRNA translation and induces hepatic insulin resistance. PLoS Genet *8*, e1002457.

Bennett, B.J., Farber, C.R., Orozco, L., Kang, H.M., Ghazalpour, A., Siemers, N., Neubauer, M., Neuhaus, I., Yordanova, R., Guan, B.*, et al.* (2010). A high-resolution association mapping panel for the dissection of complex traits in mice. Genome Res *20*, 281-290.

Chatterjee, A., Rodger, E.J., Stockwell, P.A., Weeks, R.J., and Morison, I.M. (2012). Technical considerations for reduced representation bisulfite sequencing with multiplexed libraries. J Biomed Biotechnol *2012*, 741542.

Chen, P.Y., Feng, S., Joo, J.W., Jacobsen, S.E., and Pellegrini, M. (2011). A comparative analysis of DNA methylation across human embryonic stem cell lines. Genome Biol *12*, R62.

Chen, P.Y., Ganguly, A., Rubbi, L., Orozco, L.D., Morselli, M., Ashraf, D., Jaroszewicz, A., Feng, S., Jacobsen, S.E., Nakano, A.*, et al.* (2013). Intrauterine calorie restriction affects placental DNA methylation and gene expression. Physiol Genomics *45*, 565-576.

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) *6*, 80-92.

Doherty, R., and Couldrey, C. (2014). Exploring genome wide bisulfite sequencing for DNA methylation analysis in livestock: a technical assessment. Front Genet *5*, 126.

Elmore, C.L., Wu, X., Leclerc, D., Watson, E.D., Bottiglieri, T., Krupenko, N.I., Krupenko, S.A., Cross, J.C., Rozen, R., Gravel, R.A.*, et al.* (2007). Metabolic derangement of methionine and folate metabolism in mice deficient in methionine synthase reductase. Mol Genet Metab *91*, 85-97.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw *33*, 1-22.

Ghazalpour, A., Bennett, B., Petyuk, V.A., Orozco, L., Hagopian, R., Mungrue, I.N., Farber, C.R., Sinsheimer, J., Kang, H.M., Furlotte, N.*, et al.* (2011). Comparative analysis of proteome and transcriptome variation in mouse. PLoS Genet *7*, e1001393.

Ghazalpour, A., Bennett, B.J., Shih, D., Che, N., Orozco, L., Pan, C., Hagopian, R., He, A., Kayne, P., Yang, W.P.*, et al.* (2014). Genetic regulation of mouse liver metabolite levels. Mol Syst Biol *10*, 730.

Guo, W., Fiziev, P., Yan, W., Cokus, S., Sun, X., Zhang, M.Q., Chen, P.Y., and Pellegrini, M. (2013). BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. BMC Genomics *14*, 774.

Kang, H.M., Zaitlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., Daly, M.J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. Genetics *178*, 1709-1723.

Keane, T.M., Goodstadt, L., Danecek, P., White, M.A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M.*, et al.* (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. Nature *477*, 289-294.

Kurisaki, K., Kurisaki, A., Valcourt, U., Terentiev, A.A., Pardali, K., Ten Dijke, P., Heldin, C.H., Ericsson, J., and Moustakas, A. (2003). Nuclear factor YY1 inhibits transforming growth

factor beta- and bone morphogenetic protein-induced cell differentiation. Mol Cell Biol *23*, 4494-4510.

Lomberk, G., Mathison, A.J., Grzenda, A., Seo, S., DeMars, C.J., Rizvi, S., Bonilla-Velez, J., Calvo, E., Fernandez-Zapico, M.E., Iovanna, J.*, et al.* (2012). Sequence-specific recruitment of heterochromatin protein 1 via interaction with Kruppel-like factor 11, a human transcription factor involved in tumor suppression and metabolic diseases. J Biol Chem *287*, 13026-13039.

Meissner, A., Mikkelsen, T.S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B.E., Nusbaum, C., Jaffe, D.B.*, et al.* (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature *454*, 766-770.

Millstein, J., Zhang, B., Zhu, J., and Schadt, E.E. (2009). Disentangling molecular relationships with a causal inference test. BMC Genet *10*, 23.

Orozco, L.D., Bennett, B.J., Farber, C.R., Ghazalpour, A., Pan, C., Che, N., Wen, P., Qi, H.X., Mutukulu, A., Siemers, N.*, et al.* (2012). Unraveling inflammatory responses using systems genetics and gene-environment interactions in macrophages. Cell *151*, 658-670.

Orozco, L.D., Rubbi, L., Martin, L.J., Fang, F., Hormozdiari, F., Che, N., Smith, A.D., Lusis, A.J., and Pellegrini, M. (2014). Intergenerational genomic DNA methylation patterns in mouse hybrid strains. Genome Biol *15*, R68.

Seo, S., Lomberk, G., Mathison, A., Buttar, N., Podratz, J., Calvo, E., Iovanna, J., Brimijoin, S., Windebank, A., and Urrutia, R. (2012). Kruppel-like factor 11 differentially couples to histone acetyltransferase and histone methyltransferase chromatin remodeling pathways to transcriptionally regulate dopamine D2 receptor in neuronal cells. J Biol Chem *287*, 12723-12735.

Smith, Z.D., Gu, H., Bock, C., Gnirke, A., and Meissner, A. (2009). High-throughput bisulfite sequencing in mammalian genomes. Methods *48*, 226-232.

Storey, J.D. (2002). A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology) *64*, 479-498.

Tasker, P.N., Macdonald, H., Fraser, W.D., Reid, D.M., Ralston, S.H., and Albagha, O.M. (2006). Association of PLOD1 polymorphisms with bone mineral density in a population-based study of women from the UK. Osteoporos Int *17*, 1078-1085.

Wang, S.S., Shi, W., Wang, X., Velky, L., Greenlee, S., Wang, M.T., Drake, T.A., and Lusis, A.J. (2007). Mapping, genetic isolation, and characterization of genetic loci that determine resistance to atherosclerosis in C3H mice. Arterioscler Thromb Vasc Biol *27*, 2671-2676.

Warden, C.H., Hedrick, C.C., Qiao, J.H., Castellani, L.W., and Lusis, A.J. (1993). Atherosclerosis in transgenic mice overexpressing apolipoprotein A-II. Science *261*, 469-472.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L.*, et al.* (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res *42*, D1001-1006.

Xiaoping, L., Zhibin, Y., Wenjuan, L., Zeyou, W., Gang, X., Zhaohui, L., Ying, Z., Minghua, W., and Guiyuan, L. (2013). CPEB1, a histone-modified hypomethylated gene, is regulated by miR-101 and involved in cell senescence in glioma. Cell Death Dis *4*, e675.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W.*, et al.* (2010). Common SNPs explain a large proportion of the heritability for human height. Nat Genet *42*, 565-569.

CHAPTER 2:

DNA methylation estimation using methylation-sensitive restriction enzyme bisulfite sequencing (MREBS)

DNA methylation estimation using methylation-sensitive restriction enzyme bisulfite sequencing (MREBS)

Giancarlo Bonora*, Liudmilla Rubbi*, Marco Morselli*, Constantinos Chronis, Kathrin Plath, and Matteo Pellegrini

*The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors

Abstract

Whole-genome bisulfite sequencing (WGBS) and reduced representation bisulfite sequencing (RRBS) are widely used for measuring DNA methylation levels on a genome-wide scale(1). Both methods have limitations: WGBS is expensive and prohibitive for most large-scale projects; RRBS only interrogates 6-12% of the CpGs in the human genome (16,19). Here, we introduce methylation-sensitive restriction enzyme bisulfite sequencing (MREBS) which has the reduced sequencing requirements of RRBS, but significantly expands the coverage of CpG sites in the genome. We built a multiple regression model that combines the two features of MREBS: the bisulfite conversion ratios of single cytosines (as in WGBS and RRBS) as well as the number of reads that cover each locus (as in MRE-seq (12)). This combined approach allowed us to estimate differential methylation across 60% of the genome using read count data alone, and

where counts were sufficiently high in both samples (about 1.5% of the genome), our estimates were significantly improved by the single CpG conversion information. We show that differential DNA methylation values based on MREBS data correlate well with those based on WGBS and RRBS. This newly developed technique combines the sequencing cost of RRBS and DNA methylation estimates on a portion of the genome similar to WGBS, making it ideal for large-scale projects of mammalian genomes.

Introduction

DNA methylation plays an important role in gene regulation and the maintenance of cell identity, although much remains to be uncovered regarding the specific mechanisms underlying the targeting and reading of methylcytosines(2-4). High-throughput DNA sequencing technologies have enabled the measurement of cytosine methylation on a genome-wide scale, leading to the application of these approaches in myriad studies(5-12). Many technologies have been developed over the past decade to measure DNA methylation(13). Some of these provide qualitative information about regions enriched for DNA methylation, such as approaches that select DNA fragments using proteins that selectively bind methylated cytosines (e.g. meDIP)(13,14), or methods to digest DNA with methylation- sensitive restriction enzymes (e.g. MRE)(15). Other approaches are able to probe the methylation state of single cytosines, by chemically converting unmethylated cytosines to uracils using sodium bisulfite(16). The fraction of unconverted cytosines provides an estimate of the DNA methylation level at a particular locus. This read out can be determined from microarrays (e.g. Illumina 450K)(17) or next-generation sequencing(13,18-20). In addition, some single molecule sequencing technologies

are able to directly detect methylcytosines by monitoring the incorporation rates of nucleotides (e.g. Pacific Biosciences)(21).

Among these approaches, two widely used next-generation sequencing methods for assessing DNA methylation levels at single cytosines on a genome-wide scale are whole-genome bisulfite sequencing (WGBS, also known as BS-seq, methyl-seq, or methylC-seq)(18,19,22) and reduced representation bisulfite sequencing (RRBS)(20,23). As implied by their names, both protocols are based on bisulfite treatment of DNA. However, to obtain high-confidence methylation estimates, one requires a minimum level of read coverage per site, typically at least 5–10X. WGBS is far more comprehensive and can in theory assess the methylation status of nearly every single cytosine in the genome, but requires very deep sequencing to arrive at modest levels of coverage, and, hence, can be very costly, especially if one is working with large genomes, such as those of mammals.

RRBS interrogates a smaller portion of the genome, significantly reducing the amount of sequencing required to obtain high-confidence methylation estimates at this subset of sites. The RRBS protocol introduces a step where genomic DNA is first digested, typically with the methylation-insensitive restriction endonuclease MspI, which cuts at the recognition sequence C˅CGG. Digested fragments are then size selected, typically in the range of 50 to 300 nucleotides. This fraction enriches for CpG-rich regions, including many regions involved in transcriptional regulation such as promoters and enhancers, but typically only covers 6–12% of CpGs genome-wide(23,24).

To address the respective limitations of WGBS and RRBS, we developed a new method, methylation-sensitive restriction enzyme bisulfite sequencing (MREBS), which adds a bisulfite step to an existing protocol: MRE-seq(15). Typically, MRE-seq utilizes three methylation-sensitive restriction endonucleases in parallel to digest DNA (HpaII (C˅CGG), HinP1I (G˅CGC), and AciI (C˅CGC)). Similarly to RRBS, a size selection step enriches for fragments between 50 bp and 300 bp. DNA methylation levels are inferred by the inverse relationship between MRE-

seq read coverage and CpG methylation at the restriction enzyme target sites. Although the three aforementioned restriction enzymes only cut DNA at unmethylated CpG dinucleotides, the DNA methylation state of other CpGs within the resulting fragments could still be either methylated or unmethylated. We reasoned that the addition of a bisulfite conversion step to the MRE-seq protocol would directly measure the methylation state of cytosines within MRE fragments. Typically 70 – 80 % of CpG dinucleotides in the genome are methylated(2), and the rationale behind this approach is that we direct our sequencing resources to the regions of the genome that are more likely to be unmethylated by using MRE to digest the DNA. Then, rather than simply relying on the inverse relationship between MRE-seq read coverage and DNA methylation levels around cut sites, we additionally directly measure the DNA methylation levels of their flanking regions. In principle, the advantage afforded by MREBS over WGBS and RRBS, is that we focus our sequencing effort on hypomethylated loci.

Since MREBS reads are expected to show an overall bias for lowly methylated regions due to the propensity of the restriction enzymes to digest demethylated regions, their methylation levels do not provide an unbiased measurement of absolute DNA methylation levels. However, the data can be readily used to determine differential methylation between two samples, which is often of greater interest. With this in mind, we developed a computational model that determines differential DNA methylation in two ways. First, based on read coverage alone, which is expected to anti-correlate with DNA methylation levels, we determined differential methylation within a region around each CpG dimer by looking at the difference in read counts between samples, as is done with traditional MRE-seq. Second, in those regions with sufficient read coverage for reliable estimates, we determined differential DNA methylation at single CpG resolution based on a model of bisulfite conversion ratios.

To test our approach, we first compared MREBS conversion-based methylation estimates and coverage to those based on WGBS and RRBS data, using two cell types that represent very different developmental stages, namely mouse embryonic stem cells (ESCs) and

an early somatic cell reprogramming intermediate obtained by inducing the expression of Oct4, Sox2, Klf4, and cMyc in mouse embryonic fibroblasts (MEFs) for 48 hours, where we had observed substantial differential methylation by WGBS and RRBS. We found that MREBS bisulfite conversion-based DNA methylation estimates correlated well with WGBS and RRBS-based values. The number of CpG dimers with sufficient read coverage to obtain MREBS conversion-based methylation estimates was comparable to that of RRBS. Importantly, in contrast to RRBS, we found that nearly 60% of all CpGs in the mouse genome had sufficient reads within the surrounding region in at least one of the two cell types to determine differential DNA methylation estimates based on MREBS read counts alone. Within lower coverage regions, we compute the counts in 1kb windows around CpGs to obtain approximate differential methylation as with traditional MRE-seq. In high coverage regions (~3% of CpGs), we use a multiple regression model that considers both cytosine methylation estimates from converted reads and read count data to predict differential DNA methylation values. The differential methylation estimates generated by this model compared favorably to measurements from RRBS data.

We found that MREBS provides a level of sequence coverage with nucleotide resolution similar to that obtained with RRBS. Additionally, with MREBS one can estimate DNA methylation levels for broader swathes of the genome based on differential MRE read counts around CpGs, thereby providing a level of coverage that begins to approach that obtained by WGBS, but at a fraction of the cost.

Results

Study design and data sets

We chose to test our approach on two cell lines where we expected to see significant differential methylation as they represented distinct developmental stages and differentially methylated regions (DMRs) were observed using WGBS and RRBS . The two cell types were: 1) MEFs that were induced to ectopically express the Yamanaka reprogramming factors OCT4 (O), SOX2 (S), KLF4 (K), and MYC (M; also known as 'cMYC')(25) for 48 hours, representing an early somatic cell reprogramming intermediate (EARLY), and 2) mouse embryonic stem cells (ESCs) representing the pluripotent stem cell state reached upon successful reprogramming (Figure 2.1A). Both states have been recently described in detail (26).

WGBS libraries for the two cell types were generated and reads were mapped to the mm9 genome using BS-Seeker2(27). The WGBS data sets for the two cell types showed comparable sequencing depth and CpG coverage (Figure 2.1B, Supplementary Table 2.2). RRBS and MREBS libraries for the same cell lines were also generated, with the MREBS libraries produced in duplicate to test reproducibility. RRBS reads were mapped to an in silico MspI-digested reduced reference mm9, and MREBS reads were mapped to the whole genome after being filtered for the expected 5' cut sites (Figure 2.1C/D, Supplementary Table 2.1). Although the sequencing depth for RRBS and MREBS was comparable, twice as many CpGs were covered by at least one read with MREBS (Figure 2.1C/D, Supplementary Table 2.2).

MREBS read counts provide high coverage of the genome

To determine DNA methylation estimates using bisulfite conversion rates, one typically requires at least 5X read coverage. As expected, the proportion of the 21.3 million CpGs in the mouse genome covered with a minimum of 5X coverage was substantially higher for the WGBS samples (~80%), than for either the RRBS (6%) or the MREBS samples (4–5%) (Supplementary Table

2.3). And, as was the case for the individual samples, the pairwise 5X coverage was substantially higher for the WGBS samples (75.5% of all CpG dimers between the two samples) than for either the RRBS (5.6%) or the MREBS samples (~3%) (Supplementary Table 2.4).

However, apart from using bisulfite conversion ratios to determine DNA methylation, we reasoned that for the MREBS samples we might be able to model DNA methylation based on differential read coverage alone as with traditional MRE since MREBS utilizes methylation sensitive digestion, read counts around each CpGs should anti-correlate with their methylation levels (Tables 2.1 and 2.2). In other words, MREBS read counts within windows around CpGs could be used to determine methylation, thereby providing broader coverage than one would obtain by relying only on high confidence DNA methylation calls at each CpG based on bisulfite conversion ratios. 42–48% of CpG dimers had two or more reads falling within a surrounding 1kb window (Supplementary Table 2.3), with nearly 60% of CpG dimers had at least two MREBS reads falling within the surrounding 1kb window in at least one of the two cell types (Supplementary Table 2.4). This suggested that MREBS could be utilized for determining differentially methylated regions (DMRs) between a pair of samples using both read counts and bisulfite conversion ratios.

MREBS conversion ratios correlate and MREBS read counts anti-correlate with WGBS and RRBS DNA methylation estimates, respectively

To investigate the relationship between WGBS, RRBS and MREBS-based DNA methylation estimates, we computed global correlations between them (Table 2.1). MREBS bisulfite conversion-based methylation estimates correlated more closely with the WGBS and RRBS-based estimates in the EARLY reprogramming intermediate, than they did with the ESC counterparts. Moreover, MREBS conversion-based methylation estimates for ESCs correlated more closely with the ESC RRBS and WGBS data than they did with the methylation estimates of EARLY reprogramming intermediate (Table 2.1). As expected, MREBS read counts anti-

correlated with the DNA methylation levels based on WGBS-, RRBS-, and MREBS data, though not in a particularly cell type-specific manner (Table 2.1).

Correlations based on differential data were substantially stronger than those based on absolute levels. MREBS differential data correlated with the differential WGBS and RRBS differential DNA methylation in the expected directions: differential read counts between the EARLY reprogramming intermediate and ESCs (EARLY - ESC) based on MREBS anti-correlated strongly with the differential DNA methylation values based on WGBS and RRBS, while MREBS differential bisulfite conversion ratio estimates correlated positively with those of WGBS and RRBS (Table 2.2). This suggested that MREBS data might be best utilized to estimate differential DNA methylation. Additionally, based on these observations, we hypothesized that by combining the MREBS differential conversion ratio estimates and MREBS differential read counts, we could make use of both domains of MREBS data to better estimate differential DNA methylation.

Distributions of MREBS bisulfite conversion-based DNA methylation estimates across different chromatin states are similar to those of WGBS and RRBS-based estimates

To ensure that DNA methylation estimates based on MREBS data corresponded to those based on WGBS and RRBS in all genomic contexts, we compared DNA methylation estimates across different chromatin states. To determine chromatin states, we took advantage of a hidden Markov model of chromatin states generated by using chromHMM(28). The model is described in detail in Chronis et al.(26). Briefly, the genomes of the two cell types were tiled into 200 bp windows and assigned to one of 18 chromatin states based on ChIP-seq signals for nine histone modifications a nd one histone variant (histone H3.3), including a native input library. Functional annotations were determined for each of the 18 states based on the prevalence and combination of the histone mark peaks and the enrichment of genomic features (Figure 2.2A).

Mean DNA methylation levels were estimated for all those 200 bp windows containing a minimum of one CpG with 5X read coverage. Distributions of DNA methylation levels genome-wide and within the 200 bp windows belonging to each chromatin state were then plotted for each cell type (EARLY intermediates (Figure 2.2B) and ESCs (Figure 2.2C)) and for each bisulfite sequencing method (WGBS (i), RRBS (ii), and MREBS (iii)). Different chromatin states showed characteristic DNA methylation distributions that were similar in both cell types (Figure 2.2B/C). For instance, the promoter-associated chromatin states (1 and 2) were comparatively hypomethylated, while several of the enhancer-related chromatin states (3, 4, 5, and 7) showed wide spread DNA methylation levels and an intermediate mean DNA methylation. Most of the other chromatin states were largely hypermethylated (Figure 2. 2B/C).

Apart from differences across chromatin states, there were also some differences between the cell types, as well as differences between approaches. For instance, the distributions of the MREBS-based DNA methylation estimates are systematically lower in most chromatin states as might be expected due to the use of methylation-sensitive endonucleases (Figure 2.2Biii/Ciii). Most notably the genome-wide DNA methylation levels based on MREBS estimates are low, (dark blue violin plots) close to that of the more demethylated chromatin states. Indeed, the MREBS samples are particularly enriched for two states of regulatory importance having the lowest DNA methylation levels, namely promoter and specific enhancer states (chromatin states 1–5, 7, and 15, Figure 2.2B/C). Although MREBS conversion-based DNA methylation estimates are systematically lower than those obtained by WGBS and RRBS data, their distributions within different chromatin states are very similar . Reassured that MREBS DNA methylation estimates mirrored patterns see by WGBS and RRBS across all chromatins states, we hypothesize that MREBS could be used to determine methylation levels in and between samples, if scaled appropriately, or incorporated into a model to predict differential DNA methylation.

Differential CpG-level methylation can be modeled using MREBS data

We hypothesized that WGBS-based differential DNA methylation values could be modeled using the MREBS data. To investigate this, we built four different linear regression models, as follows:

1. $y_i = \beta_0 + \beta r_i \ (i = 1..n_1)$
2. $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \ (i = 1..n_2)$
3. $y_i = \beta_0 + \beta_1 x_{i1} \ (i = 1..n_3)$
4. $y_i = \beta_0 + \beta_2 x_{i2} \ (i = 1..n_4)$

Here, y represents the WGBS conversion-based differential DNA methylation (EARLY - ESC), r the RRBS conversion-based differential DNA methylation, and x1 the MREBS conversion-based differential DNA methylation, in each case for those CpG dimers with 5X coverage in both cell types. x2 represents the differential MREBS read count within a 1kb window around each CpG dimer for windows with at least 2 reads in at least one of the two cell types. n1 = 666,214; n2 = 318,400 for MREBS replicate 1 and n2 = 322,431 for MREBS replicate 2; n3 = 319,304 for MREBS replicate 1 and n3 = 323,431 for MREBS replicate 2; n4 = 9,485,471 for MREBS replicate 1 and n4 = 9,670,440 for MREBS replicate 2.

In other words, in each case WGBS-based differential methylation serves as the response variable. In model 1, RRBS-based differential methylation is used as the explanatory variable. Model 2 is a multiple linear regression model that uses both MREBS conversion-based differential methylation and MREBS-based differential reads counts to predict WGBS-based differential methylation estimates, while model 3 and model 4 use each of these predictors independently. The lm() function from the R statistical software environment was used to implement these models(29). Coefficients for each model are provided in Supplementary Table 5. Model 1 was used for comparison purposes and shows that WGBS-based differential DNA

methylation values (EARLY - ESC) can be modeled using RRBS data for ~3% of the CpG dimers (666,214) in the mouse genome with sufficient coverage (5X) in both the WGBS and RRBS samples. The model had an R2 = 0.39, implying a correlation between the WGBS estimated differential DNA methylation values and the model-fitted ones of r = 0.63 (Table 2.3). The root-mean-square error (RMSE) between the observed and fitted values was 20.4% and the mean absolute error (MAE) was 15.2%. The metrics 'methyl15' and 'methyl25' give the percentage of CpG dimers where the difference between the WGBS differential DNA methylation estimate and that of the model was at most 15% and 25 %, respectively. Based on the methyl25 metric, the RRBS-based model-fitted DNA methylation values show 80% concordance with the WGBS-based estimates (Table 2.3).

Model 2 is a multiple regression model using both MREBS conversion-based differential DNA methylation (EARLY - ESC) and MREBS differential read count data to predict WGBS-based differential DNA methylation values. A model was built for each replicate pair. The fits for both replicates were similar (R2 = 0.35 and R2 = 0.36) and similar t o those obtained using the RRBS data (Table 2.3). Interestingly, both RMSE (~18% for both replicates) and MAE (~12.5% for both replicates) values were better than those seen for the RRBS-based model. The concordance metrics (methyl15 = ~73% and methyl25 = ~86%, for both replicates) were also superior to that obtained for the RRBS-based model (Table 3). However, the differential DNA methylation values of only 1.5% (n = ~320 thousand) of CpG dimers genome-wide could be predicted, half of that predicted by the RRBS data (n = 666,214 or ~3% of CpG dimers). However, these percentages could be increased with greater sequencing depth.

Models 3 and 4 are both simple linear models using each of the two independent variables from model 2, respectively. The fit for model 3, using only MREBS conversion-based differential DNA methylation (EARLY - ESC), was somewhat worse without the additional count data (R2 = 0.30 and R2 = 0.31 for the replicate pairs), but, interestingly, the RMSE (~18.5% for both replicates) and MAE (~12.5% for both replicates) is still superior to that of the RRBS based

model, as are the concordance metrics (methyl15 = ~73% and methyl25 = ~85%, for both replicates) (Table 2.3).

Model 4 is based only on MREBS differential read counts within 1kb window around each CpG (EARLY - ESC). Only those CpGs with at least two reads in the surrounding +/- 500bp in at least one sample were considered, amounting to 12.5 (58.8%) and 12.8 (59.7%) million CpG dimers for MREBS replicate 1 and 2, respectively (Table 2.3). This represents ~10X more CpG dimers than those that are available for use in the RRBS-based model 1 (1.2 million CpG dimers), and ~20X more CpG dimers than those that are available for use in models 2 and 3 based on MREBS sites with 5X coverage in both samples (~648–665 thousand CpG dimers). Although the fits for model 4 are worse than those based on the MREBS conversion-based DNA methylation estimates (R^2 = 0.11 for both replicate pairs), the RMSE (~24–25%) and MAE (~18.5%) are not that much worse than the RRBS based model, nor are the concordance metrics (methyl15 = ~53% and methyl25 = ~73%) (Table 2.3). However, the differential DNA methylation values for 44–45% of CpGs were predicted using model 4, representing the overlap of those CpG dimers with 5X coverage by WGBS and those CpGs with at least 2 MREBS reads within the surrounding 1kb window.

To sum the benefit of both the extended coverage of model 4 and the improved accuracy of model 2, we combined their results, updating the model 4 estimates with those of model 2 where available. This marginally improved all the applicable metrics discussed previously (Table 2.3). Figure 2.4 shows how these combined differential DNA methylation predictions (iii, green tracks, two replicates) compared to WGBS (i, dark blue tracks) and RBBS (ii, light blue tracks) at different length scales: 611kb (A), 19kb (B), and an extended locus partitioned in three 18kb panels (C). Below the modeled estimates are tracks showing the MREBS conversion-based differential DNA methylation (iv, orange, two replicates) and MREBS-based differential read counts (v, red, two replicates) – the data that was combined. While the MREBS conversion-based differential DNA methylation coverage is comparable to that of the RRBS data (cf. tracks

iv and ii, Table 2.3), the MREBS differential read count coverage approached that obtained using WGBS data (cf. tracks v and i, Table 2.3). In other words, the majority of the differential DNA estimates modeled on MREBS data are obtained by the read count data. These estimates track WGBS-based estimates, both for regions that are more methylated in the EARLY intermediates (Figures 2.4B/C and Supplementary Figure 2.1A), as well as regions that are more methylated in the ESCs (Figures 2.4A and Supplementary Figure 2.1B).

Discussion

WGBS(22) and RRBS(23) are two popular bisulfite sequencing based methods for assessing DNA methylation levels. WGBS can potentially determine the methylation status of every single cytosine, but the amount of sequencing required to obtain sufficient coverage to do so can be beyond the scope of most projects. Sequencing demands are significantly reduced by using RRBS, but one incurs an 80–90% loss in the number of cytosines that can be measured. In order to address these respective shortcomings, we introduce methylation-sensitive restriction enzyme bisulfite sequencing (MREBS), which adds a bisulfite conversion step to the existing MRE protocol, methylation-sensitive restriction enzyme digestion followed by high-throughput sequencing (MRE-seq)(15).

Due to MREs reliance on methylation sensitive endonucleases, the distributions of the MREBS conversion-based DNA methylation estimates were systematically lower than those obtained using WGBS or RRBS data. However, the MREBS conversion-based estimates followed similar trends across all chromatin states (Figure 2.2). Moreover, high-confidence MREBS conversion-based DNA methylation estimates (CpGs with 5X coverage) were particularly enriched in chromatin states with the lowest DNA methylation levels (Figure 2. 3). Since these chromatin states are known to be associated with gene regulation, their enrichment in MREBS

data is beneficial. For MREBS libraries, 4.3–4.6% of CpG dimers had at least 5X read coverage, which we set as a threshold to generate bisulfite conversion based DNA methylation level calls. This fraction was comparable to the coverage obtained from RRBS libraries at a similar level of sequencing. However, for MREBS, ~60% of CpG dimers had two or more reads falling within the surrounding 1kb window in at least one of the two cell types (Supplementary Table 2.4). These can be used for estimating differential DNA methylation of a high proportion of CpGs, since MREBS utilizes methylation sensitive digestion and therefore read counts around CpGs anti-correlate with their methylation levels (Tables 2.5 and 2.6).

To obtain estimates of differential DNA methylation based on MREBS data, we built a multiple regression model that incorporates both MREBS conversion fractions and read count data to predict differential DNA methylation values for ~3% of CpGs. The fits for both replicates were similar and correspondence metrics comparing the model-fitted values to WGBS estimates were superior to those obtained from models built using RRBS methylation data alone (Table 2.3). Differential DNA methylation estimates for a much greater proportion of CpGs (~60%) could be obtained using a model that used only MREBS differential read data within 1kb windows around CpG sites. The accuracy of MREBS read count-based models was lower than those based on conversion ratios, nonetheless, the dramatically higher coverage makes these data useful for low resolution differential methylation estimates (Table 2.3, Figures 2.4 and 2.5). In this study, we utilized 1kb windows around CpG dimers for this purpose, but one could use different windows, as one sees fit.

In summary, with respect to conversion-based DNA methylation estimates, MREBS provides a similar level of coverage to that obtained using RRBS. However, with MREBS one can additionally obtain DNA methylation estimates for a much larger proportion of the genome based on differential MREBS read counts around CpGs, providing a level of coverage that approaches that obtained by WGBS at a fraction of the cost.

Materials and methods

Methylation-sensitive restriction enzyme bisulfite sequencing (MREBS)

Three enzymatic digestions were performed on 1 μg of purified genomic DNA using 10 U of each one of the MRE restriction enzymes (HpaII, Hin6 and AciI - Fermentas) in a 50 μl final volume with TANGO buffer. 2.5 μl of RNase cocktail mix (Ambion) were added and the reaction was incubated overnight at 37°C. After the digestion, the three reactions were pooled and the DNA was purified using AMPure XP beads (Beckman Coulter). Subsequent reactions of DNA End Repair, A-tailing and Adapter Ligation were performed using Illumina TruSeq reagents, following manufacturer's instructions and the DNA was size selected between 200 and 500 bp using AMPure XP beads. Size selected DNA was then treated with bisulfite using the EpiTect kit (QIAGEN) according to the protocol suggested from the manufacturer, except that the conversion step was performed twice, for a total time of 10 h. For each bisulfite-converted sample, two parallel PCR reactions were set up in a final volume of 50 μl using MyTaq HS Mix (Bioline) and 2.5 μl of Illumina TruSeq PCR Cocktail Primers. The amplification cycles were as follows: 98°C – 2 min; 12 cycles of: 98°C – 15 sec, 60°C – 30 sec, 72°C – 30 sec; 72°C – 5 min. The final PCR products were purified using AMPure XP beads and the final concentration of the libraries was measured using Qubit DNA BR Assay (Life Technologies). Single-end sequencing for 100bp reads was performed on an Illumina Hiseq 2000.

Whole-genome bisulfite sequencing (WGBS)

Genomic DNA from induced MEFs (48h OSKM) and ESCs was isolated using the Blood and tissue DNeasy kit (Qiagen). Isolated DNA was treated with RNAseA for 30 min at 37oC and cleaned up using AMPure XP beads. 5 μg of treated DNA was fragmented to 100–500 bp using a Bioruptor Sonicator. 5 minutes in pulses of 30 sec on, 1 minute off. DNA fragments were visualized on 1% agarose gel, gel extracted and purified using a QIAGEN gel extraction minelute

kit. End-repair reactions (50 µl) contained 1x T4 DNA ligase buffer (NEB), ATP, 0.4 mM dNTPs, 15 units T4 DNA polymerase, 5 units Klenow DNA polymerase, 50 units T4 polynucleotide kinase (all NEB) and were incubated for 30 min at 20°C. DNA clean-up was performed using a 2x volume of AMPure XP beads and eluted in 32 µl of dH2O. Adenylation was performed for 30 minutes at 37°C in 50 µl volumes that contained 5 µl 1x Klenow buffer, 0.2 mM dATP and 15 units Klenow exo– (NEB). Adenylated DNA fragments and methylated adapters (Illumina) were ligated for 15 min at 20oC in a 50 µl reaction containing 5,000 units quick ligase (NEB) and 5 µl of adapters. Adaptor-ligated DNA of 200-600 bp, was size-selected on a 2% agarose gel. Bisulfite conversion was performed with an EpiTect Bisulfite Kit (QIAGEN) following the manufactures conditions. Bisulfite converted DNA was amplified for 15 cycles with PfuTurboCx Hotstart DNA polymerase (Agilent technologies). The final library DNA was quantified using a Qubit fluorometer and a Quant-iT dsDNA HS Kit (Invitrogen). Single-end sequencing for 100bp reads was performed on an Illumina Hiseq 2000.

Reduced Representation Bisulfite sequencing (RRBS)

5 µl of genomic DNA was digested with 50 units of MspI (NEB) in a 100 µl reaction for 6 hours at 37oC. Digested DNA was run on a 3% low-melt agarose gel (Lonza) and fragments of 25 to 300 bp were extracted and purified using a MinElute gel extraction kit (QIAGEN) according to the manufacturers instructions. DNA end-repair and adenylation was as described above with the exception of using a dNTP mix consistent of dATP, dGTP and 5medCTP. Ligation to methylated adapters and subsequent library construction was performed similarly to the WGBS protocol. Single-end sequencing for 100bp reads was performed on an Illumina Hiseq 2000.

Bisulfite sequencing data processing

DNA methylation calling was performed using BS-Seeker2(27) using Bowtie 0.12.9(30) for read alignment. WGBS and MREBS reads were mapped to the mm9 reference genome while

RRBS reads were mapped to a reduced reference that was in silico digested using the MspI recognition sequence and limited to fragments of 20–500bp in length. The 100bp reads were trimmed of adapter sequences and allowed 5 mismatches during mapping. MREBS reads were first filtered so that only the expected 5' trimers (CGG and CGC; Supplementary Table 1) were retained. For conversion based DNA methylation level calling, only CpG dimers covered by at least 5 reads on both were used in an effort to obtain reliable methylation levels.

ChIP-seq library preparation and chromatin states analysis

The protocol and the model is described in detail in a separate manuscript(26), but briefly 18 chromatin states in the MEFs, EARLY intermediates, LATE intermediates, and ESCs were identified at a resolution of 200 bp using chromHMM as described by Ernst and Kellis(31) using ChIP-seq data sets for nine histone modification, one histone variant (H3.3), and an input, as listed in Figure 2.

Differential DNA methylation modeling

Linear regression was used to model differential CpG dimer methylation estimates based on WGBS (the response vectors) using differential methylation estimates based on RRBS and MREBS, as well as differential read counts within 1kb windows based on MREBS data around corresponding CpG dimers with R's lm() function(29). The coefficients in in Supplementary Table 5 are outputs from R's summary.lm() function(29).

Author contributions

G.B. participated in project planning and data interpretation, performed bioinformatics analysis, and wrote the manuscript. M.M. and L.R. participated in project planning and data interpretation, and generated experimental data. C.C. produced experimental data. K.P. participated in data interpretation, provided supervision, and edited the manuscript. M.P.

conceived the study, supervised the project, interpreted the data, provided guidance for bioinformatics analysis, and edited the manuscript.

Acknowledgements

We thank Bernadett Papp for critical reading of this manuscript. G.B. was supported by a UCLA Philip Whitcome Pre-doctoral Training Fellowship, a UCLA Dissertation Year Fellowship and a UCLA Quantitative and Computational Biosciences Postdoctoral Fellowship; CC by a CIRM Training Grant and a Leukemia and Lymphoma Research Visiting Fellowship (10040); M.M. was supported by a UCLA Philip Whitcome Pre-doctoral Training Fellowship and a UCLA Dissertation Year Fellowship. KP by the UCLA Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, funds from the UCLA David Geffen School of Medicine, CIRM, and NIH P01 GM099134; and MP from NIH P01 GM099134. The authors have no conflict of interest to declare.

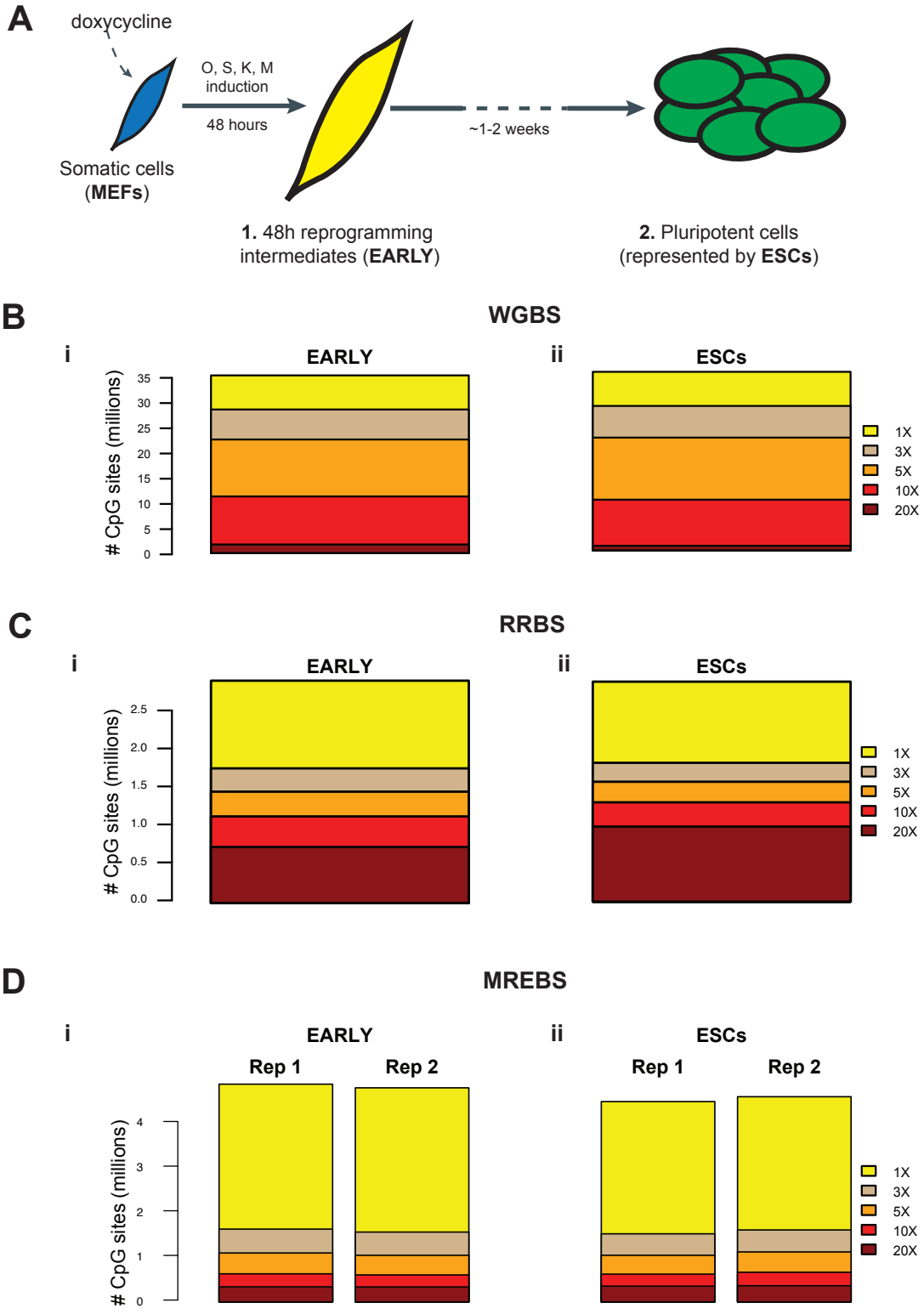


Figure 2.1. WGBS, RRBS, and MREBS for samples representing two stages of somatic cell reprogramming.

A. Schematic representation of the two cell types used in the study. Mouse embryonic fibroblasts (MEFs; blue), modified to harbor a 'stem cell cassette' allowing for the simultaneous induction of the four pluripotency factors (OCT4 (O), SOX2 (S), KLF4 (K), and MYC (M)) by the addition of doxycycline, were induced for 48 hours. These EARLY somatic cell reprogramming intermediates (yellow) were the first of the two cell types sampled, with embryonic stem cells (ESCs; green), representing the fully reprogrammed state, being the second.

B. Bar plots showing the number of CpGs obtained at five different coverage levels (1–20 X) in each of the two whole-genome bisulfite sequencing (WGBS) samples: i) EARLY intermediates (429 M mapped reads) and ii) ESCs (391 M mapped reads).

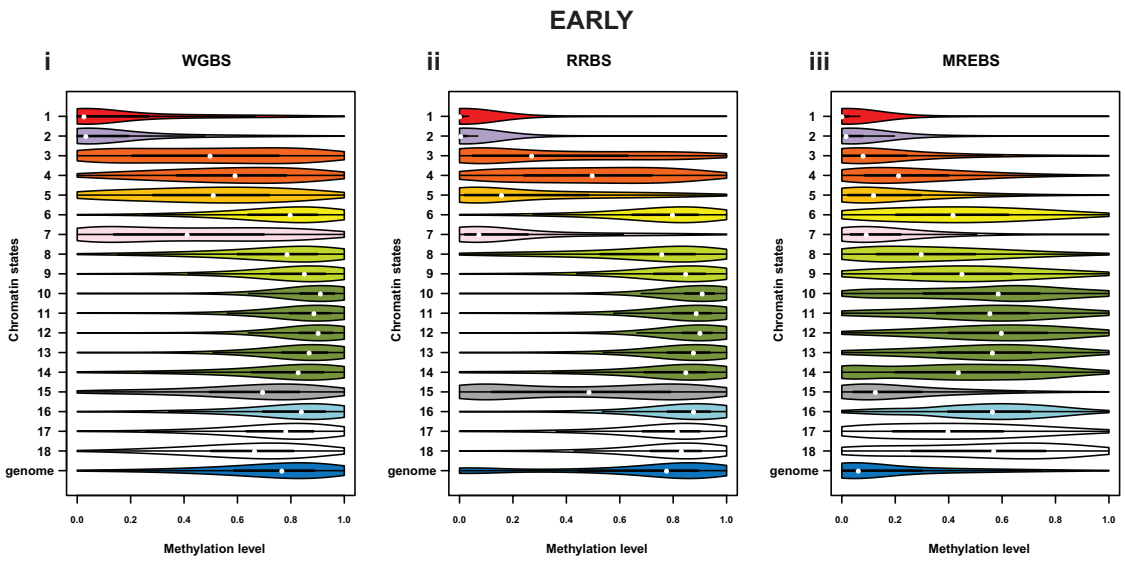
C. As in (B), but for reduced representation bisulfite sequencing (RRBS) samples (12.8 M mapped reads for i and 18.1 M mapped reads for ii).

D. As in (B), but for duplicate samples produced using methylation-sensitive restriction enzyme bisulfite sequencing (MREBS) (11.9 and 12.4 M mapped reads for i; 11.8 and12.2 M mapped reads for ii).

A

	State	Genome %	Input	K27ME3	K27AC	K4ME2	K4ME1	K9AC	K4ME3	K36ME3	K9ME3	K79ME2	H3.3
Promoter	1_PromA	0.5	3	2	87	93	32	84	95	6	4	74	46
	2_PromP	0.2	2	18	47	91	32	77	87	0	4	1	4
Enhancer	3_EnhA	0.9	2	0	84	91	87	27	8	6	4	2	74
	4_EnhA	1.4	3	1	56	96	96	7	1	1	1	1	3
	5_EnhM	1.3	2	1	3	74	45	20	1	0	2	1	2
	6_EnhW	2.6	3	1	20	3	54	1	0	1	1	0	1
	7_EnhP	1.1	22	69	7	39	69	2	1	5	3	1	0
Txscribed enhancer	8_TxEnhA	1.2	3	0	49	81	86	14	6	30	4	96	28
	9_TxEnhW	1.0	9	2	37	15	80	1	0	75	3	4	7
Transcription	10_Tx	2.0	3	1	8	1	19	1	0	86	10	89	14
	11_Tx5'	1.6	2	0	4	3	19	3	0	11	2	83	2
	12_Tx3'	6.7	4	1	5	0	1	0	0	84	1	2	2
	13_TxWk3'	4.5	4	0	1	0	2	4	0	13	0	1	0
	14_Tx3'	0.5	5	3	19	5	16	5	2	34	38	3	75
Polycomb	15_ReprPC	8.5	7	56	0	0	1	0	0	1	3	0	0
Repeats	16_Repeats	1.3	4	8	2	0	1	0	0	15	64	0	1
Low signal	17_Low	22.4	3	3	2	0	1	0	0	0	1	0	0
	18_LowL	42.2	0	0	0	0	0	0	0	0	1	0	0

B



C

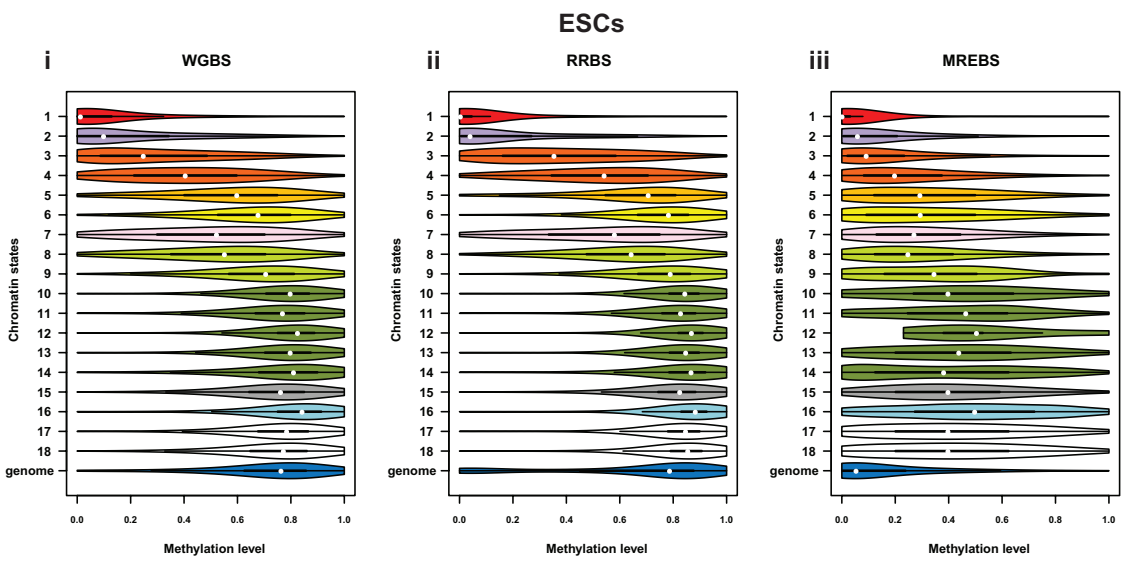


Figure 2.2. DNA methylation estimates based on WGBS, RRBS and MREBS data in different chromatin states.

A. Heat map and functional annotation for an 18-level chromatin state model at 200 bp resolution built using peak calls made using ChIP-seq data sets for 10 histone modifications, as well as an input library, for each of the three cell types described in Figure 1A: i) MEFs, ii) EARLY intermediates, and iii) ESCs, as well as a late reprogramming intermediate (LATE), partially induced pluripotent cells, or pre-iPSCs, not otherwise used in the study. Candidate functional annotations were assigned to each of the 18 chromatins states based on the prevalence and combination of histone mark peaks, which could in turn be classified into the seven categories indicated in the left-hand column. The probability of a window in each state to be contain a peak for a given histone modification is given as a percentage in each cell, and visually indicated by the intensity of color in the heat map. The proportion of the concatenated genome (MEFs + EARLY + LATE + ESCs) found in each of the 18 chromatin states is given in the third column.

B. Violin plots of the distributions of the DNA methylation estimates in each of the 18 chromatin states (described in A), as well as genome-wide (blue), for EARLY intermediates using i) whole-genome bisulfite sequencing (WGBS), reduced representation bisulfite sequencing (RRBS), and iii) methylation-sensitive restriction enzyme bisulfite sequencing (MREBS). The mean DNA methylation estimates within 200 bp windows corresponding to those used for the chromatin state model were used, only considering those windows containing at least one CpG with 5X coverage, in an effort to ensure high-confidence estimates. White circles represent median values.

C. As in (B), but for ESCs.

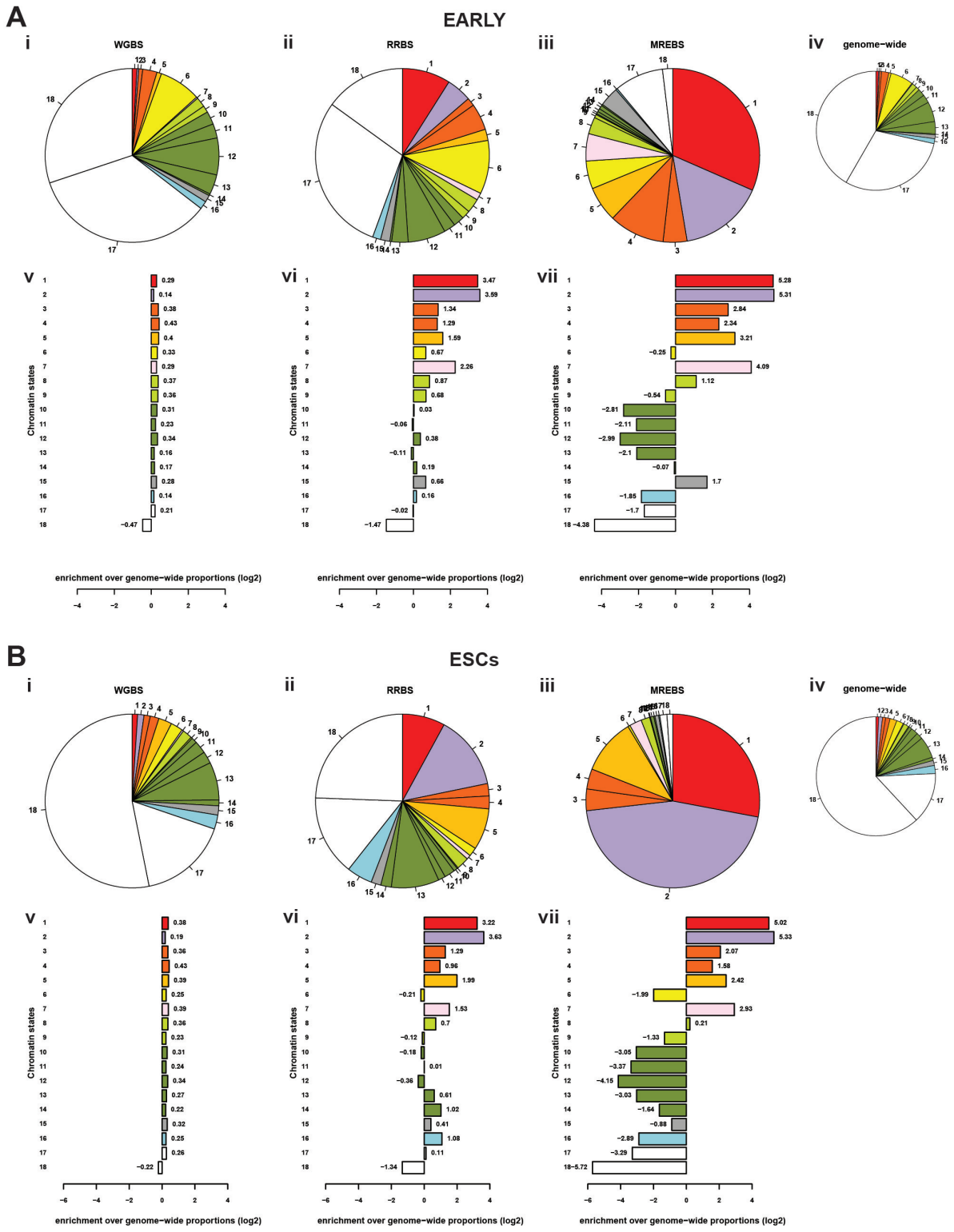


Figure 2.3. Chromatin state coverage by DNA methylation estimates by WGBS, RRBS, and MREBS.

A. Pie charts show the proportion of 200 bp windows with DNA methylation estimates found in each of the 18 chromatin states (as described in Figure 2.2A) using i) whole-genome bisulfite sequencing (WGBS), reduced representation bisulfite sequencing (RRBS), and iii) methylation-sensitive restriction enzyme bisulfite sequencing (MREBS) EARLY intermediate samples, as compared to the proportion of chromatin states in genome for all 13.3 million windows (iv). Bar plots show the log2 fold change (observed / expected) number of windows with estimates per method: i) WGBS), RRBS, and iii) MREBS. The mean DNA methylation estimates within 200 bp windows corresponding to those used for the chromatin state model were used, only considering those windows containing at least one CpG with 5X coverage, in an effort to ensure high-confidence estimates.

B. As in (A), but for ESCs.

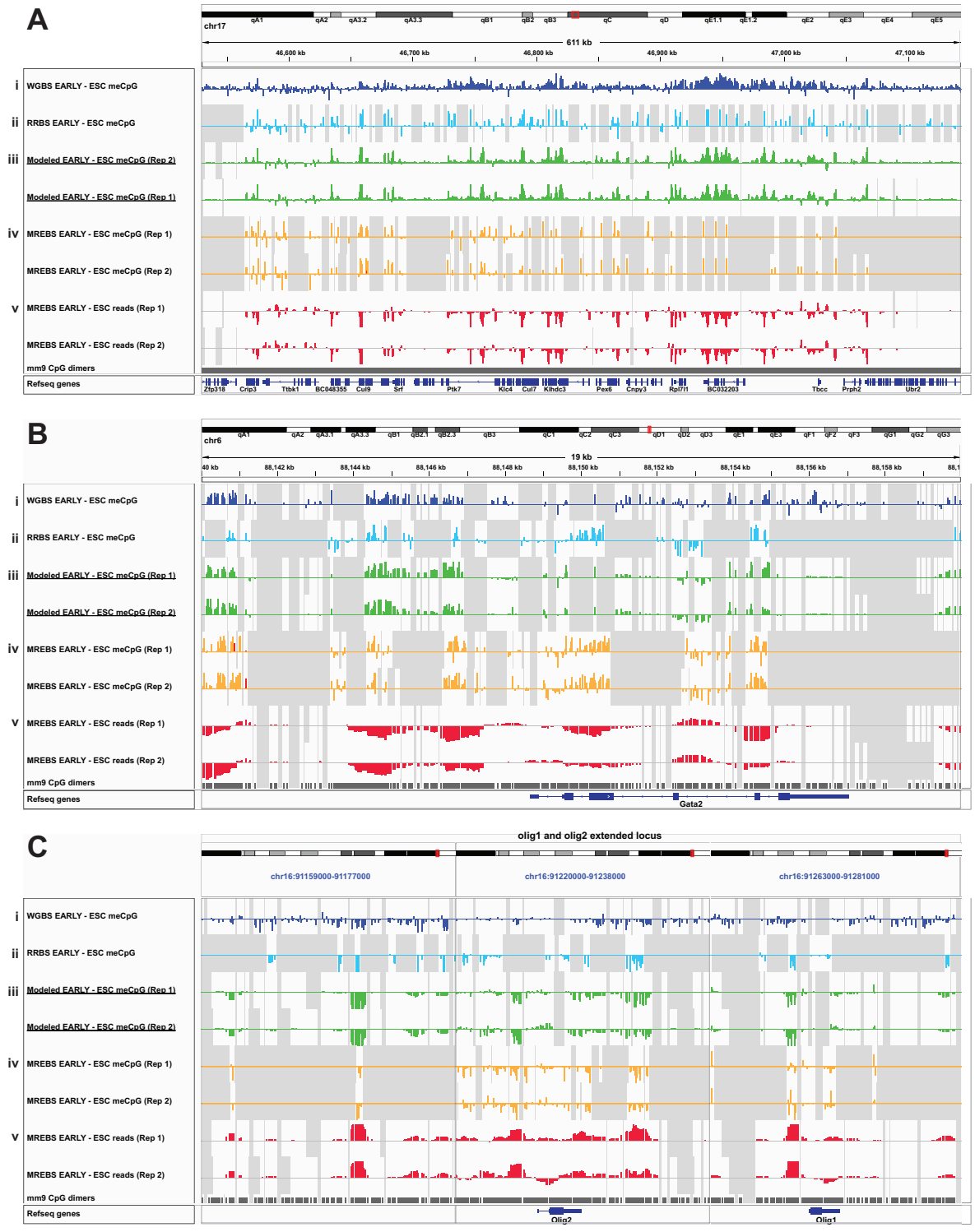
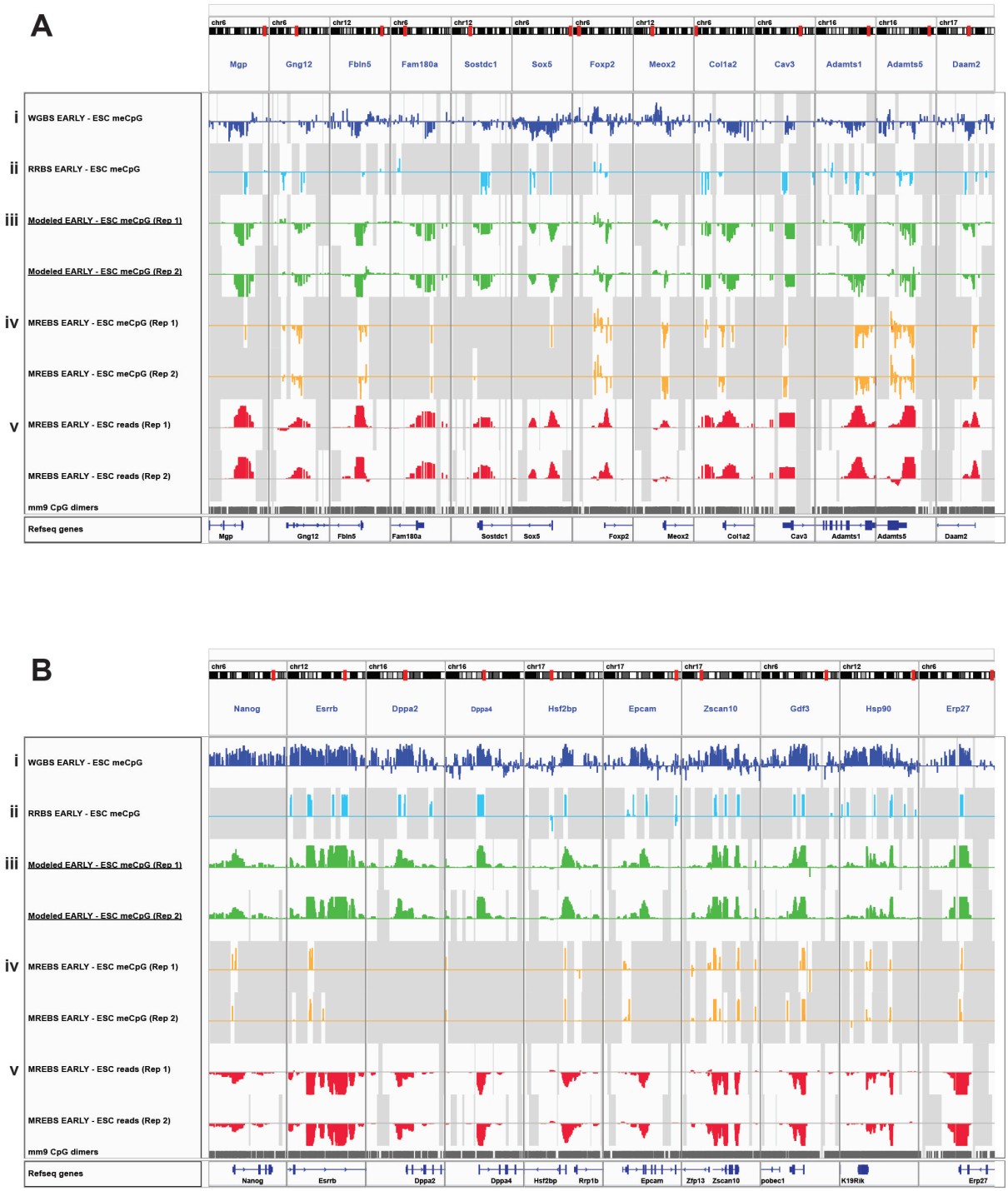


Figure 2.4. Differential DNA methylation levels modeled using MREBS data.

A. IGV tracks of differential DNA methylation estimates between EARLY intermediates and ESCs (EARLY - ESC) based on WGBS data (i, dark blue), RRBS data (ii, light blue), modeled data based on combined model 2 (iii, green, two replicates), MREBS conversion-based DNA methylation estimates (iv, orange, two replicates), and MREBS read count-based estimates (v, red, two replicates), within a 611 kb region of chr17. Bottom two tracks show CpG dimer and Refseq gene locations. Gray background reflects regions (CpG dimers) where data was not available.

B. As in (A), but for a 19kb region around the Gata2 gene.

C. As in (A), but for the extend Olig1/2 gene locus, divided into three 18 kb panels.



Supplementary Figure 2.1. Examples of modeled differential DNA methylation around gene loci.

A. As in Figure 4A, but for 13 genes up-regulated in EARLY intermediates relative to ESCs.

B. As in Figure 4A, but for 10 genes up-regulated in ESCs relative to EARLY intermediates.

			CpG dimer-level conversion-based methylation estimates *								MREBS read counts within 1kb windows around CpG dimers **			
			WGBS		RRBS		MREBS							
			EARLY	ESC	EARLY	ESC	EARLY Rep1	EARLY Rep2	ESC Rep1	ESC Rep2	EARLY Rep1	EARLY Rep2	ESC Rep1	ESC Rep2
CpG dimer-level conversion-based methylation estimates *	WGBS	EARLY	1.00	0.56	0.90	0.78	0.78	0.77	0.55	0.55	-0.14	-0.14	-0.18	-0.17
		ESC	0.56	1.00	0.77	0.90	0.53	0.53	0.70	0.71	-0.15	-0.15	-0.24	-0.22
	RRBS	EARLY	0.90	0.77	1.00	0.88	0.82	0.82	0.66	0.66	-0.21	-0.21	-0.27	-0.27
		ESC	0.78	0.90	0.88	1.00	0.64	0.64	0.80	0.80	-0.21	-0.21	-0.30	-0.29
	MREBS	EARLY Rep1	0.78	0.53	0.82	0.64	1.00	0.88	0.68	0.67	-0.07	-0.07	-0.09	-0.09
		EARLY Rep2	0.77	0.53	0.82	0.64	0.88	1.00	0.69	0.68	-0.07	-0.07	-0.10	-0.09
		ESC Rep1	0.55	0.70	0.66	0.80	0.68	0.69	1.00	0.83	-0.07	-0.07	-0.10	-0.10
		ESC Rep2	0.55	0.71	0.66	0.80	0.67	0.68	0.83	1.00	-0.07	-0.07	-0.10	-0.09
MREBS read counts within 1kb windows around CpG dimers **		EARLY Rep1	-0.14	-0.15	-0.21	-0.21	-0.07	-0.07	-0.07	-0.07	1.00	1.00	0.96	0.98
		EARLY Rep2	-0.14	-0.15	-0.21	-0.21	-0.07	-0.07	-0.07	-0.07	1.00	1.00	0.97	0.98
		ESC Rep1	-0.18	-0.24	-0.27	-0.30	-0.09	-0.10	-0.10	-0.10	0.96	0.97	1.00	0.99
		ESC Rep2	-0.17	-0.22	-0.27	-0.29	-0.09	-0.09	-0.10	-0.09	0.98	0.98	0.99	1.00

* For CpG dimers with 5X coverage in both cell lines

** For windows with at least two MREBS reads in at least one cell line

Table 2.1. CpG dimer-level correlations between bisulfite sequencing libraries. Pearson correlation values between WGBS, RRBS, and MREBS CpG dimer-level conversion-based methylation estimates, as well as binned read counts within 1kb windows around CpG dimers, for those with at least two MREBS reads. Red intensity signifies the strength of a positive correlation, while blue intensity signifies the strength of the anti-correlation.

			CpG dimer-level differential methylation using conversion-based estimates *				Differential MREBS read counts within 1kb windows around CpG dimers **	
			WGBS	RRBS	MREBS			
					Rep1	Rep2	Rep1	Rep2
CpG dimer-level differential methylation using conversion-based estimates *	WGBS		1.00	0.63	0.55	0.56	-0.34	-0.34
	RRBS		0.63	1.00	0.64	0.64	-0.51	-0.52
	MREBS	Rep1	0.55	0.64	1.00	0.64	-0.35	-0.35
		Rep2	0.56	0.64	0.64	1.00	-0.34	-0.34
Differential MREBS read counts within 1kb windows around CpG dimers **		Rep1	-0.34	-0.51	-0.35	-0.34	1.00	0.87
		Rep2	-0.34	-0.52	-0.35	-0.34	0.87	1.00

* For CpG dimers with 5X coverage in both cell lines

** For windows with at least two MREBS reads in at least one cell line

Table 2.2. CpG dimer-level correlations between differential values for all bisulfite sequencing library pairs. Pearson correlation values between WGBS, RRBS, and MREBS differential CpG dimer-level methylation estimates (EARLY - ESC), as well as differential read counts between all CpG dimers with at least two MREBS reads within a surrounding 1kb window. Red intensity signifies the strength of a positive correlation, while blue intensity signifies the strength of the anti-correlation.

			Model 1	Model 2		Model 3		Model 4		Model 4 estimates updated with Model 2 values where available	
	Total / optimal value	WGBS differential DNA methylation	RRBS differential DNA methylation	MREBS differential DNA me + counts		MREBS differential DNA me only		MREBS differential counts only			
				Rep1	Rep2	Rep1	Rep2	Rep1	Rep2	Rep1	Rep2
covered dimers*	21,342,492	16,113,172	1,204,249	648,635	664,446	649,614	665,431	12,542,720	12,746,080	12,542,720	12,746,080
%	100.00	75.5	5.6	3.0	3.1	3.0	3.1	58.8	59.7	58.8	59.7
fitted dimers	21,342,492	NA	666,214	318,400	322,431	319,304	323,336	9,485,471	9,670,440	9,485,471	9,670,440
%	100.00	NA	3.1	1.5	1.5	1.5	1.5	44.4	45.3	44.4	45.3
R-squared	1	NA	0.39	0.35	0.36	0.30	0.31	0.11	0.11	NA	NA
obs vs fitted correlation	1	NA	0.63	0.60	0.60	0.55	0.56	0.34	0.34	0.37	0.36
RMSE	0	NA	20.4	17.9	17.9	18.6	18.5	24.7	24.6	24.5	24.4
MAE	0	NA	15.2	12.5	12.4	12.7	12.5	18.5	18.4	18.3	18.2
methyl15	100	NA	61.16	73.13	73.51	73.27	73.40	53.11	53.27	53.71	53.85
methyl25	100	NA	80.09	86.11	86.12	85.16	85.14	73.30	73.46	73.69	73.82

*In the case of methylation levels, only CpG-dimers with 5X coverage in both DOX & ES were considered.

With respect to counts, only CpGs with 2+ reads in the surrounding 1Kb bin, in at least one sample, were considered.

Table 2.3. Differential DNA methylation model metrics. The table gives metrics (first column) for four different models, as well as a combined model, (top row and described in the text). The column labeled 'Total / optimal value' gives the maximum or best value achievable for each metric. The column labeled 'WGBS differential DNA methylation' provides coverage information for comparison purposes. *Note: In the case methylation levels, only CpG dimers with 5X coverage in both EARLY intermediates and ESCs were considered. With respect to counts, CpG dimes with 2+ reads in the surrounding 1Kb bin, in at least one sample, were considered. Red intensity signifies how close the metrics are to the optimal values.

	Total Mapped Reads	Mean CpG coverage depth	
WGBS EARLY	429,374,384	7.80	
WGBS ESC	391,724,853	7.23	
RRBS EARLY	12,826,209	12.49	*
RRBS ESC	18,131,716	18.88	
MREBS EARLY Rep1	11,963,716	5.72	**
MREBS EARLY Rep2	12,400,629	5.78	
MREBS ESC Rep1	11,835,343	6.35	
MREBS ESC Rep2	12,222,192	6.28	

* RRBS reads mapped to in silico MspI-digested reduced references genome.

** MREBS reads in silico filtered and mapped to whole genome.

Supplementary Table 2.1. Bisulfite sequencing library mapped reads and mean CpG coverage depth. WGBS and MREBS reads were mapped the whole genome (mm9). Mean CpG coverage determined for CpGs on either strand. Mean CpG coverage determined for CpGs on either strand. RRBS reads were mapped to an in silico MspI digested reduced reference genome. MREBS reads were filtered in silico to have the expected 5' start sites.

MRE restriction enzyme	4mer CpG context	Instances in mm9 genome (1 strand)
	ACGT	1,756,359
HpaII*	CCGG	1,594,148
	CCGT	1,454,486
	ACGG	1,449,336
	TCGG	1,404,375
	CCGA	1,401,476
	TCGT	1,392,167
	TCGA	1,391,828
	ACGA	1,391,206
	GCGT	1,258,753
	ACGC	1,255,903
AciI	CCGC	1,251,553
	GCGG	1,250,525
Hin6I	GCGC	1,102,589
	TCGC	995,280
	GCGA	992,773
	NCGA	7
	NCGC	4
	ACGN	3
	CCGN	3
	GCGN	2
	NCGG	2
	NCGT	1
	Total	21,342,779

* This is the same recognition sequence as for MspI,
the endonuclease typically used for RRBS-eq libraries,
albeit HpaII is methylation sensitive, as are AciI and Hin6I.

Supplementary Table 2.2. MRE endonuclease recognition sequence frequency within the mm9 genome. MRE endonuclease recognition sites are highlighted to show their position within the ranked frequencies for all the 4mer CpG, including chrM. *Note: HpaII has the same recognition sequence as MpsI (the endonuclease typically used for RRBS libraries), albeit HpaII is methylation sensitive, as are AciI and Hin6I.

	CpG dimers*	%	
WGBS EARLY DNAme	17,202,917	80.6%	CpG dimers with 5X coverage
WGBS ESC DNAme	17,022,903	79.8%	
RRBS EARLY DNAme	1,289,663	6.0%	
RRBS ESC DNAme	1,358,529	6.4%	
MREBS EARLY Rep1 DNAme	962,559	4.5%	
MREBS EARLY Rep2 DNAme	924,551	4.3%	
MREBS ESC Rep1 DNAme	918,508	4.3%	
MREBS ESC Rep2 DNAme	973,441	4.6%	
MREBS EARLY Rep1 Counts>0	14,695,688	68.9%	Read fallng within 1kb around CpG dimer
MREBS EARLY Rep2 Counts>0	14,718,855	69.0%	
MREBS ESC Rep1 Counts>0	13,602,796	63.7%	
MREBS ESC Rep2 Counts>0	13,947,144	65.3%	
MREBS EARLY Rep1 Counts>=2	10,250,065	48.0%	
MREBS EARLY Rep2 Counts>=2	10,315,026	48.3%	
MREBS ESC Rep1 Counts>=2	8,987,596	42.1%	
MREBS ESC Rep2 Counts>=2	9,227,587	43.2%	
MREBS EARLY Rep1 Counts>=5	5,317,311	24.9%	
MREBS EARLY Rep2 Counts>=5	5,387,683	25.2%	
MREBS ESC Rep1 Counts>=5	4,696,304	22.0%	
MREBS ESC Rep2 Counts>=5	4,659,571	21.8%	
MREBS EARLY Rep1 Counts>=10	3,557,311	16.7%	
MREBS EARLY Rep2 Counts>=10	3,610,365	16.9%	
MREBS ESC Rep1 Counts>=10	3,389,630	15.9%	
MREBS ESC Rep2 Counts>=10	3,352,308	15.7%	

* 21,342,493 CpG-dimers in mm9 (ex. chrM)

Supplementary Table 2.3. CpG dimer coverage per bisulfite sequencing library. The percentage of CpG dimers with at least 5X coverage for each bisulfite sequencing library, as well as the percentage of CpG dimers with the specified number of MREBS reads within a surrounding 1kb window. *Note: This is based on 21,342,493 CpG dimers in mm9, excluding chrM.

	CpG dimers with 5X coverage in both EARLY and ESC	%
WGBS	16,113,172	75.5%
RRBS Rep1	1,204,249	5.6%
MREBS Rep1	649,614	3.0%
MREBS Rep2	665,431	3.1%

	CpG dimers with at least 2 reads in the surrounding 1 kB region in at least one sample	%	
MREBS Rep1	12,542,720	58.8%	**
MREBS Rep2	12,746,080	59.7%	

* 21,342,493 CpG-dimers in mm9 (ex. chrM)

** Based on bins with 2+ reads in at least one sample

Supplementary Table 2.4. CpG dimer coverage for differential analysis per bisulfite sequencing library. The percentage of CpG dimers with at least 5X coverage in both the EARLY intermediate and ESC samples for each bisulfite sequencing library, as well as the percentage of CpG dimers with at least two MREBS reads within a surrounding 1kb window. *Note: This is based on 21,342,493 CpG dimers in mm9, excluding chrM.

	Model 1	Model 2		Model 3		Model 4	
	RRBS-based differential DNA methylation	MREBS conversion-based differential DNA methylation + differential read counts		MREBS conversion-based differential DNA methylation only		MREBS differential reads counts only	
		Rep1	Rep2	Rep1	Rep2	Rep1	Rep2
B0 (intecept)	5.29	5.09	4.92	5.45	5.01	2.61	2.73
B1	0.75	0.53	0.54	0.63	0.64	-0.38	-0.38
B2	N/A	-0.09	-0.09	N/A	N/A	N/A	N/A

Supplementary Table2.5. Differential DNA methylation model coefficients.

Coefficient values (first column) for four different models (top row and described in the text).

References

1. Yong, W.S., Hsu, F.M. and Chen, P.Y. (2016) Profiling genome-wide DNA methylation. Epigenetics Chromatin, 9, 26.

2. Jones, P.A. (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet, 13, 484-492.

3. Smith, Z.D. and Meissner, A. (2013) DNA methylation: roles in mammalian development. Nat Rev Genet, 14, 204-220.

4. Suzuki, M.M. and Bird, A. (2008) DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet, 9, 465-476.

5. Zemach, A., McDaniel, I.E., Silva, P. and Zilberman, D. (2010) Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science, 328, 916-919.

6. Feng, S., Cokus, S.J., Zhang, X., Chen, P.Y., Bostick, M., Goll, M.G., Hetzel, J., Jain, J., Strauss, S.H., Halpern, M.E. et al. (2010) Conservation and divergence of methylation patterning in plants and animals. Proceedings of the National Academy of Sciences of the United States of America, 107, 8689-8694.

7. Hon, G.C., Rajagopal, N., Shen, Y., McCleary, D.F., Yue, F., Dang, M.D. and Ren, B. (2013) Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. Nat Genet, 45, 1198-1206.

8. Shipony, Z., Mukamel, Z., Cohen, N.M., Landan, G., Chomsky, E., Zeliger, S.R., Fried, Y.C., Ainbinder, E., Friedman, N. and Tanay, A. (2014) Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. Nature, 513, 115-119.

9. Xie, W., Schultz, M.D., Lister, R., Hou, Z., Rajagopal, N., Ray, P., Whitaker, J.W., Tian, S., Hawkins, R.D., Leung, D. et al. (2013) Epigenomic analysis of multilineage differentiation of human embryonic stem cells. Cell, 153, 1134-1148.

10. Ziller, M.J., Gu, H., Muller, F., Donaghey, J., Tsai, L.T., Kohlbacher, O., De Jager, P.L., Rosen, E.D., Bennett, D.A., Bernstein, B.E. et al. (2013) Charting a dynamic DNA methylation landscape of the human genome. Nature, 500, 477-481.

11. Smith, Z.D., Chan, M.M., Humm, K.C., Karnik, R., Mekhoubad, S., Regev, A., Eggan, K. and Meissner, A. (2014) DNA methylation dynamics of the human preimplantation embryo. Nature, 511, 611-615.

12. Lee, D.S., Shin, J.Y., Tonge, P.D., Puri, M.C., Lee, S., Park, H., Lee, W.C., Hussein, S.M., Bleazard, T., Yun, J.Y. et al. (2014) An epigenomic roadmap to induced pluripotency reveals DNA methylation as a reprogramming modulator. Nat Commun, 5, 5619.

13. Laird, P.W. (2010) Principles and challenges of genomewide DNA methylation analysis. Nat Rev Genet, 11, 191-203.

14. Jacinto, F.V., Ballestar, E. and Esteller, M. (2008) Methyl-DNA immunoprecipitation (MeDIP): hunting down the DNA methylome. Biotechniques, 44, 35, 37, 39 passim.

15. Maunakea, A.K., Nagarajan, R.P., Bilenky, M., Ballinger, T.J., D'Souza, C., Fouse, S.D., Johnson, B.E., Hong, C., Nielsen, C., Zhao, Y. et al. (2010) Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature, 466, 253-257.

16. Frommer, M., McDonald, L.E., Millar, D.S., Collis, C.M., Watt, F., Grigg, G.W., Molloy, P.L. and Paul, C.L. (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proceedings of the National Academy of Sciences of the United States of America, 89, 1827-1831.

17. Sandoval, J., Heyn, H., Moran, S., Serra-Musach, J., Pujana, M.A., Bibikova, M. and Esteller, M. (2011) Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenetics, 6, 692-702.

18. Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M. and Jacobsen, S.E. (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature, 452, 215-219.

19. Lister, R. and Ecker, J.R. (2009) Finding the fifth base: genome-wide sequencing of cytosine methylation. Genome Res, 19, 959-966.

20. Meissner, A., Gnirke, A., Bell, G.W., Ramsahoye, B., Lander, E.S. and Jaenisch, R. (2005) Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res, 33, 5868-5877.

21. Yang, Y., Sebra, R., Pullman, B.S., Qiao, W., Peter, I., Desnick, R.J., Geyer, C.R., DeCoteau, J.F. and Scott, S.A. (2015) Quantitative and multiplexed DNA methylation analysis using long-read single-molecule real-time bisulfite sequencing (SMRT-BS). BMC Genomics, 16, 350.

22. Feng, S., Rubbi, L., Jacobsen, S.E. and Pellegrini, M. (2011) Determining DNA methylation profiles using sequencing. Methods Mol Biol, 733, 223-238.

23. Gu, H., Smith, Z.D., Bock, C., Boyle, P., Gnirke, A. and Meissner, A. (2011) Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. Nat Protoc, 6, 468-481.

24. Smith, Z.D., Gu, H., Bock, C., Gnirke, A. and Meissner, A. (2009) High-throughput bisulfite sequencing in mammalian genomes. Methods, 48, 226-232.

25. Takahashi, K. and Yamanaka, S. (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell, 126, 663-676.

26. Chronis, C., Fiziev, P., Papp, B., Butz, S., Bonora, G., Sabri, S., Ernst, J. and Plath, K. (2017) Cooperative Binding of Transcription Factors Orchestrates Reprogramming. Cell, 168, 442-459 e420.

27. Guo, W., Fiziev, P., Yan, W., Cokus, S., Sun, X., Zhang, M.Q., Chen, P.Y. and Pellegrini, M. (2013) BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. BMC Genomics, 14, 774.

28. Ernst, J. and Kellis, M. (2012) ChromHMM: automating chromatin-state discovery and characterization. Nat Methods, 9, 215-216.

29. Team, R.D.C. (2014). 3.1.2 ed. R Foundation for Statistical Computing, Vienna, Austria.

30. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol, 10, R25.

31. Ernst, J. and Kellis, M. (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol, 28, 817-825.

CHAPTER 3:

In vivo targeting of de novo DNA methylation by histone modifications in yeast and mouse

Morselli M*, et al.* (2015). *eLife* 4:e06205.

In vivo targeting of de novo DNA methylation by histone modifications in yeast and mouse

Marco Morselli[1]*, William A Pastor[1], Barbara Montanini[2], Kevin Nee[1], Roberto Ferrari[1], Kai Fu[1], Giancarlo Bonora[1,3], Liudmilla Rubbi[1], Amander T Clark[1], Simone Ottonello[2], Steven E Jacobsen[1,3,4]*, Matteo Pellegrini[1,3]*

[1]Department of Molecular, Cell and Developmental Biology, University of California, Los Angeles, Los Angeles, United States; [2]Biochemistry and Molecular Biology Unit, Department of Life Sciences, Laboratory of Functional Genomics and Protein Engineering, Parma, Italy; [3]Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, University of California, Los Angeles, Los Angeles, United States; [4]Howard Hughes Medical Institute, University of California, Los Angeles, Los Angeles, United States

Abstract Methylation of cytosines ($5^{me}C$) is a widespread heritable DNA modification. During mammalian development, two global demethylation events are followed by waves of de novo DNA methylation. In vivo mechanisms of DNA methylation establishment are largely uncharacterized. Here, we use *Saccharomyces cerevisiae* as a system lacking DNA methylation to define the chromatin features influencing the activity of the murine DNMT3B. Our data demonstrate that DNMT3B and H3K4 methylation are mutually exclusive and that DNMT3B is co-localized with H3K36 methylated regions. In support of this observation, DNA methylation analysis in yeast strains without Set1 and Set2 shows an increase of relative $5^{me}C$ levels at the transcription start site and a decrease in the gene-body, respectively. We extend our observation to the murine male germline, where H3K4me3 is strongly anti-correlated while H3K36me3 correlates with accelerated DNA methylation. These results show the importance of H3K36 methylation for gene-body DNA methylation in vivo.
DOI: 10.7554/eLife.06205.001

Introduction

In multicellular organisms, every cell type possesses the same genetic information, but manifests a different phenotype. Chromatin plays a fundamental role in both the establishment and maintenance of each cell's state. Many players contribute to chromatin states, including nucleosome organization, histone post-translational modifications, and non-coding RNAs (***Chen and Dent, 2014***; ***Maze et al., 2014***; ***Quinodoz and Guttman, 2014***). Another mechanism for maintaining the state of a cell through cell division is the methylation of cytosines at position 5 ($5^{me}C$), a widespread heritable DNA modification found in prokaryotes, plants, several fungi, and animals (***Iyer et al., 2011***). In mammals, DNA methylation plays a fundamental role in processes such as imprinting, X-chromosome inactivation, transposon inactivation, and gene expression regulation (***Smith and Meissner, 2013***). Dysregulation of DNA methylation is a common feature in cancer (***Eden et al., 2003***; ***You and Jones, 2012***) and a variety of human diseases are caused by defective imprinting (***Peters, 2014***).

Methylation is mainly found at symmetric CpG dinucleotides, where it is introduced by the de novo DNA methyltransferases (DNMT3a and DNMT3b) and can be copied faithfully during DNA replication by the activity of a 'maintenance' DNA methyltransferase, DNMT1 (***Law and Jacobsen, 2010***). However, DNA methylation is not static throughout mammalian development. In fact, $5^{me}C$ can either

eLife digest In animals and other multicellular organisms, there are many different types of cells that each perform particular roles in the body. This is possible because the genetic information—which is the same in all cells—is controlled so that only a subset of all the genes within an individual cell are 'switched on' at a particular time.

Genetic information is contained within molecules of DNA, which are wrapped around proteins called histones. The genes in regions of DNA where these histones are packed tightly together tend to be switched off, while genes in regions of DNA that are loosely packed tend to be switched on. The level of packaging is controlled by the addition of 'methyl' tags to the histone proteins.

These tags can also be added directly to the DNA in a process called DNA methylation. Enzymes called methyltransferases add the tags to the DNA, which tends to switch off the gene. The locations of the methyl tags can be copied when the DNA replicates before the cell divides so that the pattern of DNA methylation can be passed on to its daughter cells. However, it is not clear how the methyltransferases are able to target particular regions for methylation.

To address this question, Morselli et al. introduced a methyltransferase called DNMT3b into yeast, a single-celled organism that does not normally add methyl tags to its DNA. The experiments show that the activity of the enzyme is affected by the presence of methyl tags on certain histone proteins. For example, a methyl tag at one particular site on a histone, called H3K4, prevents the DNMT3b enzyme from adding methyl tags to DNA. However, a methyl tag at another site called H3K36 promotes DNA methylation.

Morselli et al. found that these two histone sites had similar effects on DNA methylation in mouse sperm cells. Morselli et al.'s findings may be useful in the future development of treatments for cancer and other diseases that are caused by defects in DNA methylation.

DOI: 10.7554/eLife.06205.002

be lost by a passive mechanism, such as the failure to maintain DNA methylation through cell division or by an active mechanism such as the removal of methylcytosine, typically via an oxidized intermediate (***Pastor et al., 2013***).

Demethylation and de novo methylation can occur in a locus-specific manner, typically in concert with the activation or silencing of promoters or enhancers. However, global demethylation and de novo methylation events can also occur during development (***Pastor et al., 2013***; ***Seisenberger et al., 2013***). For example, most DNA methylation is progressively lost between fertilization and the formation of the blastula and global de novo DNA methylation then occurs coincidently with implantation of the embryo. This de novo methylation event largely shapes the methylation pattern of the animal, with additional changes occurring in somatic tissues, which contribute to cellular identity. In the germline however, a second reprogramming event occurs. After specification of the germ cells, most DNA methylation is lost during early primordial germ cell (PGCs) development. Unlike in early embryogenesis, imprints are erased during this period. Genome-wide de novo methylation then occurs before birth in the male germline and upon oocyte maturation in females (***Smallwood et al., 2011***). This de novo methylation event establishes the imprints that are inherited in the next generation.

Considering the importance of local and global de novo methylation events in imprinting, gene regulation and cellular identity, it is important to understand how the de novo DNA methyltransferases are targeted to the correct genomic regions. DNMT3 proteins do not have strong sequence preferences beyond CpG dinucleotides (***Dodge et al., 2002***). We therefore sought to determine which factors are critical for the targeting of de novo DNA methyltransferases.

Active de novo DNA methyltransferases possess three different domains: the catalytic domain, found at the C-terminus of the protein, an ADD domain and a PWWP domain (***Figure 1A***) (***Law and Jacobsen, 2010***). In contrast, the inactive DNMT3L possesses only a functional ADD domain. The ADD domains of all three DNMTs have been shown to preferentially bind histone 3 tails that lack methylation at lysine 4 (H3K4me0) (***Ooi et al., 2007***; ***Zhang et al., 2010***), and this binding has been recently shown to relieve DNMT3a auto-inhibition (***Guo et al., 2015***). This is consistent with the observation that genomic regions bearing H3K4 methylation are generally depleted of $5^{me}C$ (***Singh et al., 2013***). The PWWP domain of several proteins has been shown to bind H3K36

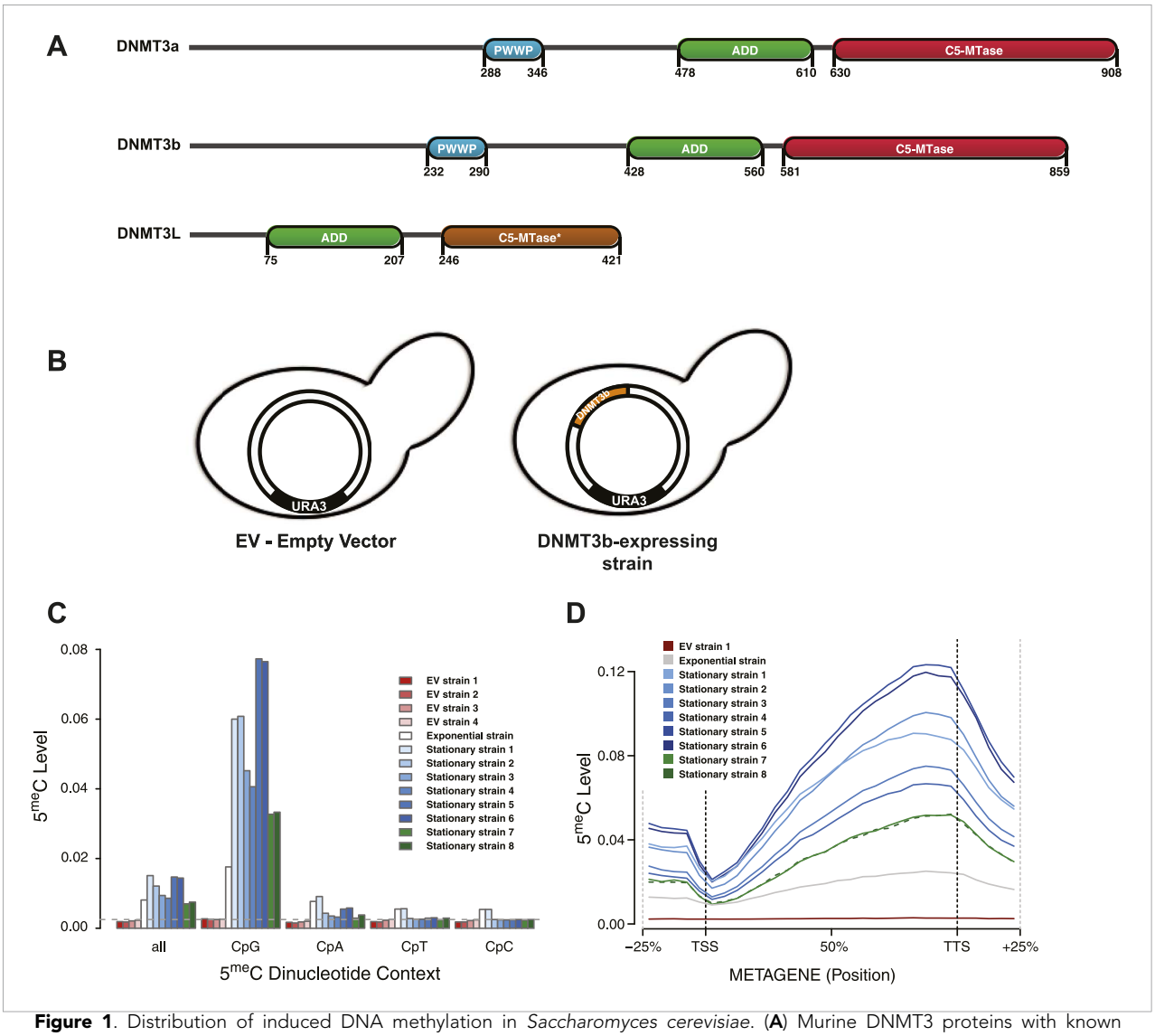


Figure 1. Distribution of induced DNA methylation in *Saccharomyces cerevisiae*. (**A**) Murine DNMT3 proteins with known domains: PWWP, ADD (ATRX–DNMT3–DNMT3L), and C-5 methyltransferase domain (not functional in DNMT3L). Accession numbers: DNMT3a = O88508; DNMT3b = O88509; DNMT3L = Q9CWR8. (**B**) Constructs used in this study. The empty vector (EV) is pYES2 (Life Technologies). DNMT3b expression is controlled by the GAL1 promoter. (**C**) Levels of 5meC in different dinucleotide contexts. The gray dotted line represents the unconversion rate. (**D**) Metagene plot of CpG methylation in cells expressing DNMT3b during logarithmic and stationary phase. EV (strain not expressing DNMT3b). Exponential and stationary strains 1–6 are derived from the W303 strain, while stationary strains 7 and 8 are in a BY4741 background.

DOI: 10.7554/eLife.06205.003

The following figure supplements are available for figure 1:

Figure supplement 1. Chromosome-wide view of DNA methylation and genomic features.

DOI: 10.7554/eLife.06205.004

Figure supplement 2. Distribution of 5meC around TSS and TTS.

DOI: 10.7554/eLife.06205.005

methylation (***Vermeulen et al., 2010***), and indeed the DNMT3a-PWWP domain has also been shown to interact with the tri-methylated lysine 36 of histone H3 (H3K36me3) in vitro (***Dhayalan et al., 2010***). The importance of these histone-binding domains in targeting DNA methyltransferase activity in vivo is still unclear. It is also possible that the PWWP domain's primary function is to bind DNA and not nucleosomes (***Dhayalan et al., 2010***). Recently, it has been reported that the PWWP domain is important in specifying the localization of DNMT3b in mouse embryonic stem cells (***Baubec et al., 2015***).

While there has been extensive characterization of DNMT3 in vitro, a comprehensive analysis of the mechanisms guiding the activity of a de novo DNMT in vivo is still incomplete. To address this question, we introduced DNMT3b into an organism that has no endogenous DNA methylation machinery, the budding yeast *Saccharomyces cerevisiae*, to study the chromatin components affecting the activity of a mammalian de novo DNA methyltransferase. This system has several advantages over the study of DNA methylation in mammalian cells. Yeast has conserved histone sequences and many residues are modified at the same sites as those found in higher eukaryotes. However, unlike mammalian cells, yeast cells can be easily manipulated and the small size of their genome reduces costs associated with next-generation sequencing-based approaches. Moreover, yeast has already been used to show the importance of the N-terminus of histone H3 in targeting the DNA methylation complex (***Hu et al., 2009***).

Our data show that the chromatin template guides the activity of DNMT3b. DNMT3b preferentially deposits methylation in linker DNA compared to nucleosomal DNA. Also, DNMT3b activity correlates positively with H3K36me3 and negatively with H3K4me3. In fact, mutation of the H3K36 methyltransferase Set2 decreases DNA methylation over regions that would normally contain H3K36me3. Thus the marks themselves, as opposed to genomic features that correlate with these marks, are responsible for targeting DNA methylation. We also demonstrate that the pattern of H3K4 and H3K36 methylation in embryonic male germ cells accurately predicts which regions undergo de novo methylation, indicating that the mechanism observed in yeast is conserved in mammals.

Results

Ectopically expressed DNMT3b methylates yeast genomic DNA

S. cerevisiae does not have any endogenous cytosine DNA methyltransferases, and its DNA is therefore unmethylated. To study the activity of a de novo methyltransferase in this organism, we introduced the murine DNMT3b under the control of the inducible GAL1 promoter (***Figure 1B***). We measured the levels of 5-methylcytosine ($5^{me}C$) in these strains using whole genome bisulfite sequencing (WGBS) (***Supplementary file 1A***). We observed significant levels of $5^{me}C$ of DNA extracted from the exponentially growing and stationary phases of the same strain culture (***Figure 1C*** and ***Supplementary file 2A***), with higher methylation levels observed in stationary phase. CpG dinucleotides were preferentially methylated, as expected from the previously characterized activity of mammalian DNMT3. The methylation levels of CpG dinucleotides range from 3.3 to 7.7%, depending on the yeast strain analyzed. These levels are about 10–20 times higher than the average of other dinucleotides levels (***Supplementary file 2A***), and well above the bisulfite non-conversion rate of 0.27%, as estimated from an unmethylated lambda DNA spike-in.

Despite some level of variability, we observe methylation across the entire yeast genome (***Figure 1—figure supplement 1A,B***). When mapping reads to the genome we only retain those that map to a single position. As a result we do not obtain methylation estimates for regions that contain repetitive sequences, such as the rRNA containing regions in chromosome XII.

We also observed a striking methylation distribution within genes (***Figure 1D***), with low levels at the transcription start site (TSS) and increasing methylation in the gene body, reaching a maximum close to the transcription termination site (TTS). The same pattern is found in mammals (***Lister et al., 2009***; ***Chodavarapu et al., 2010***), suggesting that equivalent mechanisms regulating DNMT3 activity in mammalian genes might also be present in yeast.

DNMT3b preferentially methylates linker DNA

In yeast, nucleosomes are well positioned at the beginning of a gene, with nucleosome-free regions (NFRs) immediately upstream of the TSS and downstream of the TTS (***Brogaard et al., 2012***). When average levels of $5^{me}C$ are calculated around the TSS, we observed a periodicity of about 170 bp (***Figure 1—figure supplement 2***). A similar periodicity is also observed at the TTS. This suggested that nucleosomes might influence the activity of de novo DNMTs.

To address this question, we measured nucleosome positioning genome-wide using micrococcal nuclease-digested chromatin and deep-sequencing (MNase-seq) (***Supplementary file 1B*** and ***Supplementary file 3A,B***). We profiled the distribution of methylated cytosines at the TSS (***Figure 2A***), TTS (***Figure 2B***), and around each nucleosome center (***Figure 2C***).

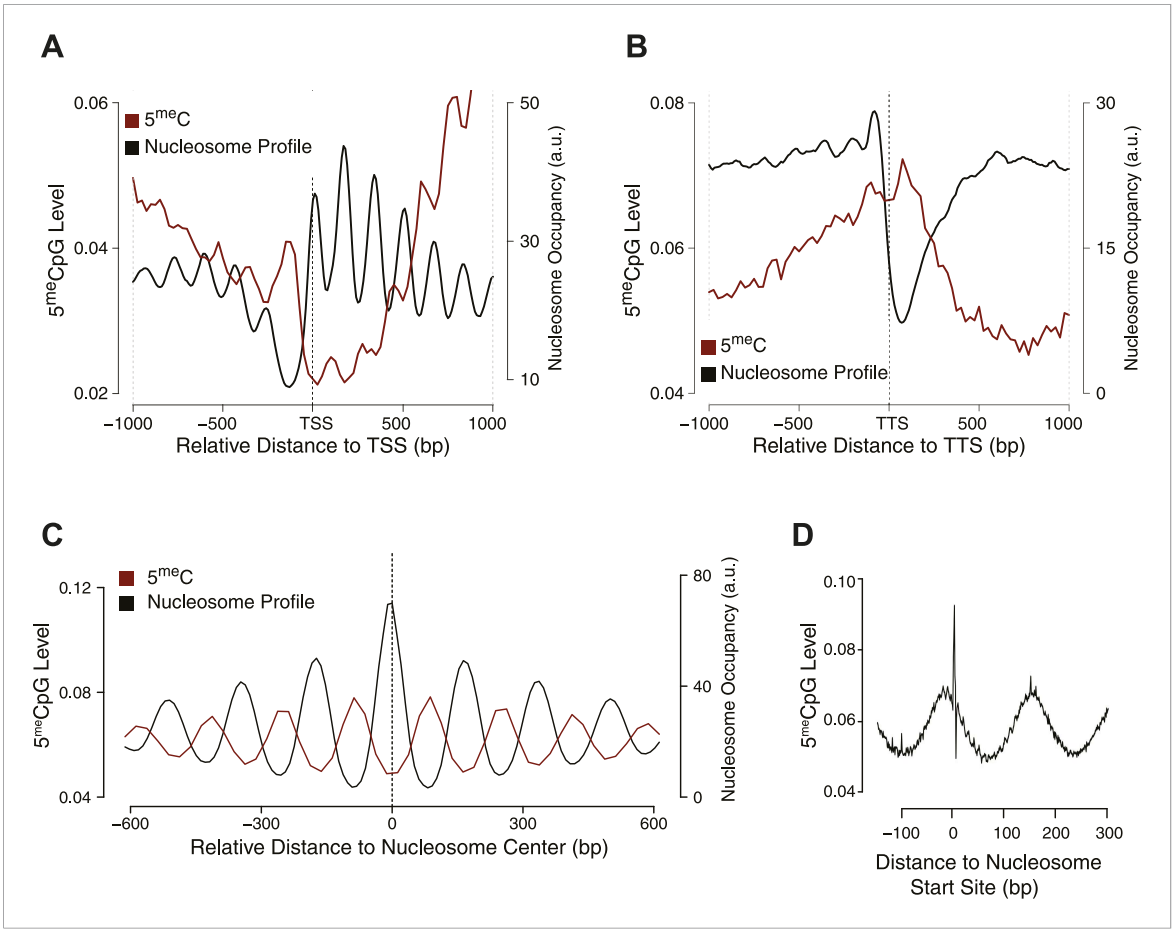


Figure 2. Influence of nucleosome positioning on DNA methylation. Average distribution of nucleosomes and DNA methylation (CpG context) around (**A**) Transcriptional Start Site (TSS), (**B**) Transcriptional Termination Site (TTS), and (**C**) nucleosome centers. (**D**) Meta-nucleosome plot of CpG methylation.

a.u. = Arbitrary units.

DOI: 10.7554/eLife.06205.006

The following figure supplement is available for figure 2:

Figure supplement 1. Differences in nucleosome occupancy between DNMT3b-expressing and non-expressing yeast strains.

DOI: 10.7554/eLife.06205.007

From these analyses, it is evident that DNMT3b preferentially methylates non-nucleosomal DNA. We observe a 50% increase in the methylation of linker DNA compared to nucleosome bound DNA (***Figure 2C***). We also observe a slight 10 bp periodicity of methylated CpG (***Figure 2D***), another feature shown in higher eukaryotes that reflects the periodicity of the DNA helix (***Klug and Lutter, 1981***).

Impact of DNA methylation on yeast nucleosome position and gene expression

We considered the possibility that introducing 5^{me}C would alter nucleosome distribution or gene expression in yeast. However, a comparison of DNMT3b-expressing and non-expressing strains showed no detectable change in nucleosome positioning by MNase treatment near the TSS, TTS (***Figure 2—figure supplement 1A,B*** and ***Supplementary file 3C***), or elsewhere in the genome.

RNA-seq analysis identified some differentially expressed genes (about 5% of the total number of genes, with an equal number of up- and down-regulated transcripts) between the strain expressing and non-expressing DNMT3b grown to stationary phase (***Figure 3*** and ***Supplementary file 1C*** and ***Supplementary file 4A***). The down-regulated genes showed enrichment for branched-chain aminoacid biosynthesis genes, while the up-regulated ones were enriched in ribosomal biogenesis

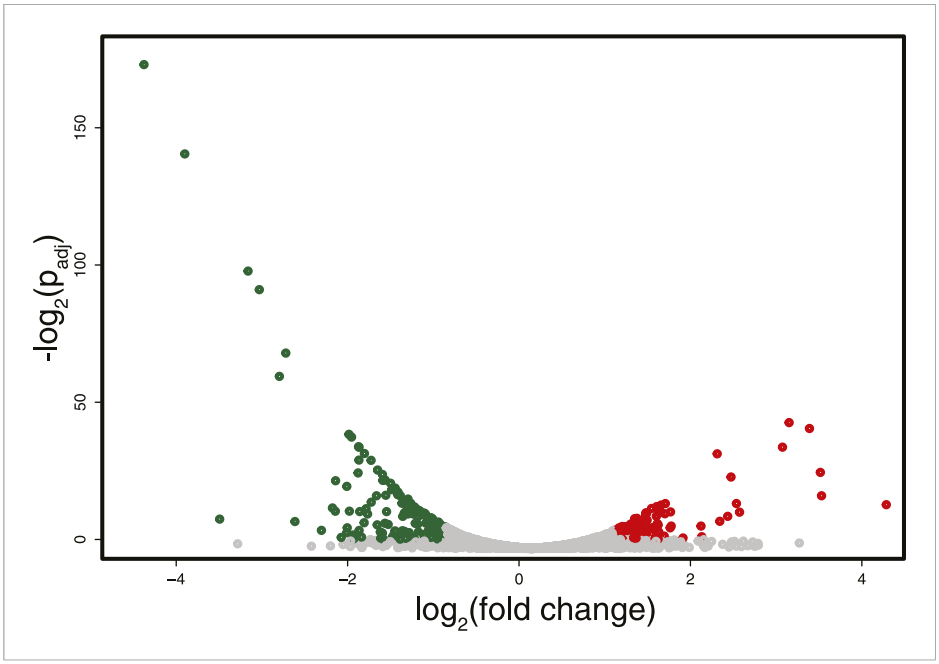


Figure 3. Differences in RNA expression between DNMT3b-expressing and non-expressing yeast strains. The expression difference in RNA expression between DNMT3b and EV strains is plotted on the x axis, and false discovery rate (FDR)-adjusted significance is plotted on the y-axis (–log2 scale). Upregulated and downregulated RNAs shown in red and green, respectively. Significantly expressed RNAs have a fold change bigger than two with a FDR smaller than 0.1.

DOI: 10.7554/eLife.06205.008

The following figure supplements are available for figure 3:

Figure supplement 1. DNA Methylation in up- and down-regulated genes.

DOI: 10.7554/eLife.06205.009

Figure supplement 2. DNA Methylation in ribosomal biogenesis genes.

DOI: 10.7554/eLife.06205.010

genes (***Supplementary file 4B–F***). However, these changes are likely due to stress response pathways that are triggered by the overexpression of MmDNMT3b, rather than by the changes in DNA methylation itself. In support of this view, when the levels of CpG, CpHpG, and CpHpH methylation in the up- and down-regulated genes were compared, no significant difference was evident (***Figure 3—figure supplement 1***). Moreover, the methylation levels of the differentially transcribed genes were not different from that of other members of the same Gene Ontology (GO) term (***Figure 3—figure supplement 2***). Since DNA methylation machinery is not native in yeast, it is likely that proteins able to recognize and mediate $5^{me}C$ effects are also absent.

DNMT3b activity is associated with specific histone tail modifications

We next sought to test whether the observed levels of $5^{me}C$ could be explained by the underlying distribution of specific histone tail modifications. To address this, we mapped the distribution of DNMT3b and of specific histone residue modifications via ChIP-seq in both the DNMT3b-expressing and wild type (wt) (non-expressing) strains (***Supplementary file 1D***).

We found that, as expected, DNMT3b co-localizes with methylated regions (***Figure 4A***). The distribution of DNMT3b is also consistent with the distribution of DNA methylation across the gene body (***Figure 4—figure supplement 1***). We also observed that DNMT3b and $5^{me}C$ are strongly anti-correlated with H3K4me3 and positively correlated with H3K36me3 (***Figure 4B*** and ***Figure 4—figure supplements 2–4***). By examining the distribution of histone marks across gene bodies, we found that H3K4me3 is concentrated at the promoter while H3K36me3 levels peak near the 3′ end of the gene (***Figure 4—figure supplement 1***). These observations suggest that the ADD and PWWP domains of DNMT3B play a role in targeting the activity of the enzyme. H3K4me1 shows a weak positive

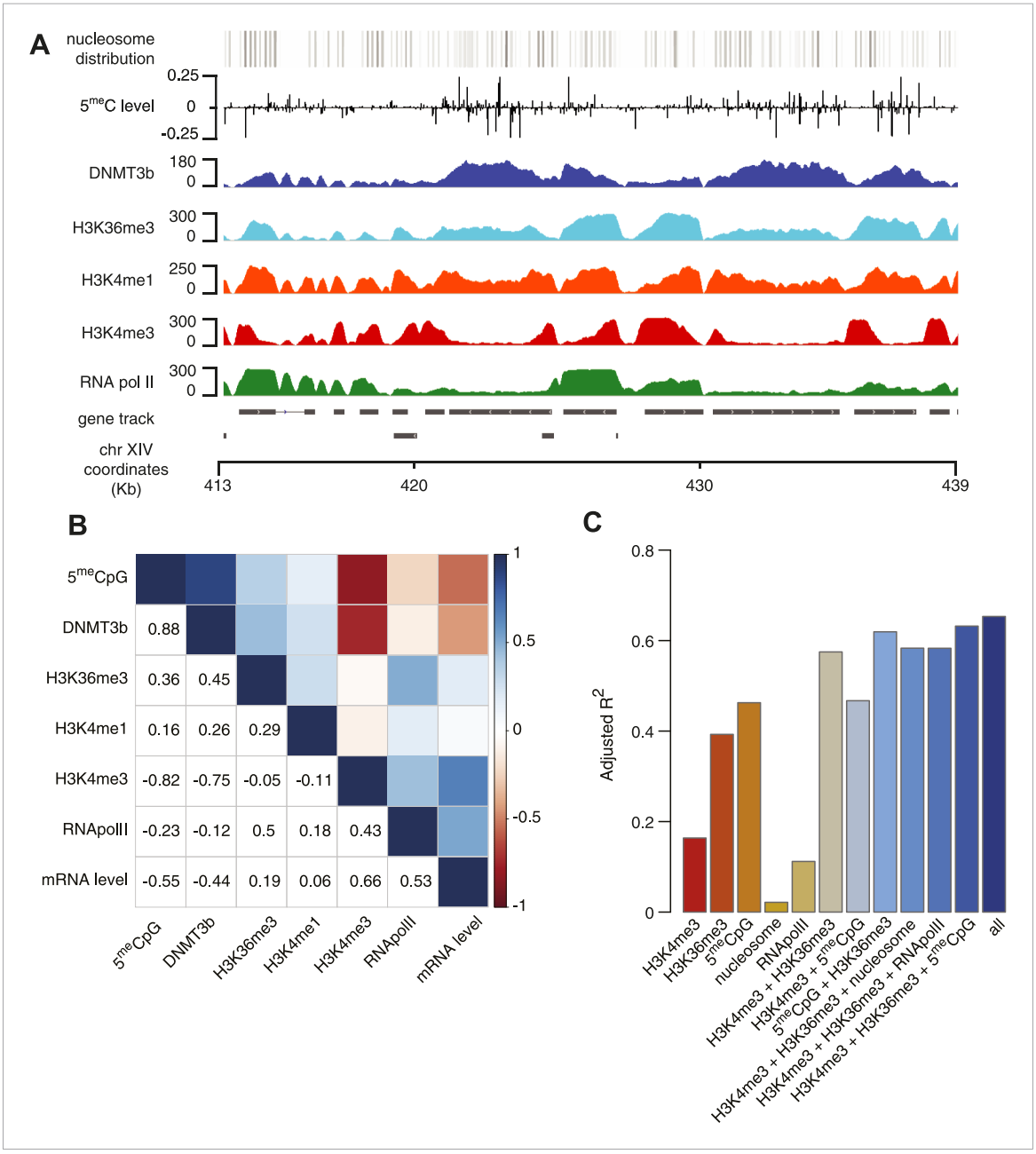


Figure 4. Correlation between histone marks and DNA methylation. (**A**) Genome-wide distribution of nucleosome, 5meC, DNMT3b, H3K36me3, H3K4me1, H3K4me3, and RNA polymerase II. (**B**) Spearman correlation coefficients between 5meC, histone marks, RNA polymerase II, DNMT3b and mRNA average levels for protein coding genes. (**C**) Prediction of DNMT3b levels using DNA methylation, H3K4 and H3K36 trimethylation, RNA polymerase II and nucleosome distribution as predictors. The y-axis shows the adjusted R^2 value between the predicted linear model and observed values. DOI: 10.7554/eLife.06205.011
The following figure supplements are available for figure 4:

Figure supplement 1. Metagene plot of ChIP sequencing in a DNMT3b-expressing strain. DOI: 10.7554/eLife.06205.012

Figure supplement 2. Relationship between transcription and 5meC or histone marks levels. DOI: 10.7554/eLife.06205.013

Figure supplement 3. Relationship between DNA methylation and histone marks levels. DOI: 10.7554/eLife.06205.014

Figure 4. continued on next page

Figure 4. Continued

Figure supplement 4. Relationship between H3K4me3 and 5meC or histone marks levels. DOI: 10.7554/eLife.06205.015

Figure supplement 5. 5meC levels prediction using chromatin marks.
DOI: 10.7554/eLife.06205.016

correlation with both $5^{me}C$ levels and DNMT3b. This might be due to the specific distribution of H3K4me1 within the gene body, partially overlapping to the H3K36me3 modification (***Figure 4A*** and ***Figure 4—figure supplement 1***).

$5^{me}C$ and DNMT3b distribution are also inversely correlated with gene transcription and Pol II abundance (***Figure 4B***, ***Figure 4—figure supplements 2E–F, 3A***). Both H3K4me3 and H3K36me3 correlate positively with transcription (***Figure 4B***, ***Figure 4—figure supplement 2C–D***). Since yeast genes are very small relative to mammalian genes, H3K4 methylation can spread well into the gene body (***Figure 4—figure supplement 2B–C***) and limit the deposition of $5^{me}C$ in highly transcribed genes. In support of this observation, we find that a higher level of H3K4me3 in the last third of the gene, is associated with a lower level of DNMT3b or $5^{me}C$ (***Figure 4—figure supplement 4***).

To determine whether the methylation of H3K4 and H3K36 is sufficient to explain the observed DNA methylation of our DNMT3b strains, we constructed a simple linear model of DNA methylation based on our ChIP-seq data. We used linear multivariate regression to model whether the distribution of one or a few histone marks, nucleosome positioning or RNA polymerase II occupancy could predict the levels of DNMT3b or $5^{me}C$ (***Figure 4C*** and ***Figure 4—figure supplement 5***). Strikingly, we found that H3K4me3 and H3K36me3 levels are sufficient to predict the distribution of both DNMT3b and $5^{me}C$ with very high accuracy. The prediction could only be slightly improved by using additional data, suggesting that H3K4me3 and H3K36me3 are the key factors in determining the targeting of DNA methylation (***Supplementary file 5***).

Deletion of histone lysine methyltransferases affect DNA methylation distribution

To determine whether H3K36me3 has a direct role in the recruitment/activity of DNMT3b in vivo, we measured the DNA methylation distribution in three mutant strains: *set1Δ*, *set2Δ*, and *dot1Δ* (***Supplementary file 1E***). In yeast, Set1 is responsible for the methylation of H3K4, Set2 is the methyltransferase for H3K36, and Dot1 catalyzes the methylation of H3K79. We included the *dot1Δ* strain as a control, since we do not expect its activity to influence the binding of DNMT3b. If the modification of H3K36 plays a role in DNMT3b activity we would expect a reduction in DNA methylation levels in gene bodies, which are the primary H3K36me3 positive regions.

Due to an impact of the set mutations on global transcription, the levels of the induced DNMT3b and the resulting DNA methylation were lower in deletion strains than the wt. Nonetheless, the resulting $5^{me}C$ levels were still significantly higher than background levels found in the wt strains (***Figure 5A*** and ***Supplementary file 2B***). To account for the variations in global methylation levels we adopted two types of normalization: the first normalized by the total amount of DNA methylation in the sample and the second was based on the expression of DNMT3b measured via RT-qPCR (***Figure 5B*** and ***Figure 5—figure supplement 1***). Both strategies gave similar results (data not shown).

As expected, we see no significant differences in $5^{me}C$ distribution in *dot1Δ* strains compared to wt (***Figure 5B***). In contrast, in the *set1Δ* strain, we found that regions close to the TSS, with high H3K4me3 and low DNA methylation in a wt strain, contain methylation levels that are not significantly different from other regions outside of the gene (***Figure 5B,C***). This suggests that H3K4 methylation plays an active role in suppressing DNA methylation in the wt, and that this effect disappears in the *set1Δ* strain (***Figure 5D***).

In a *set2Δ* strain, $5^{me}C$ levels are reduced over gene bodies compared to wt strains (***Figure 5C***). Moreover, in this strain maximum levels of DNA methylation peak outside of the gene, where H3K36me3 is not present (***Figure 5B***). Thus, in this mutant strain DNA methylation is redistributed from gene bodies (H3K36me3-rich regions) to intergenic regions compared to the wt, suggesting that H3K36me3 is responsible for recruitment of DNMT3B (***Figure 5C,E***).

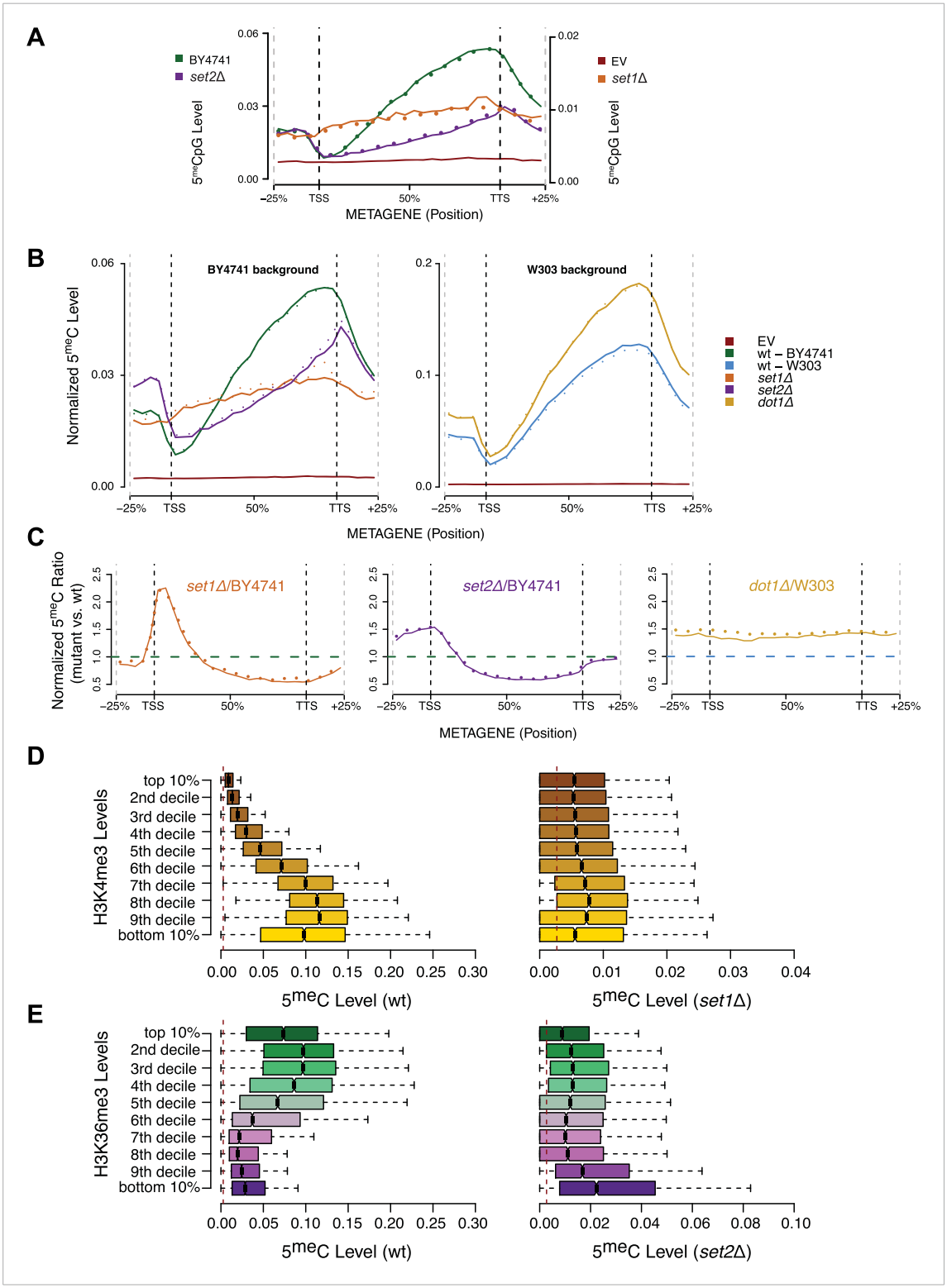


Figure 5. Effect of histone lysine methyltransferase deletions on the distribution of DNA methylation. (**A**) Metagene plot of CpG methylation in *set1Δ* and *set2Δ* cells expressing DNMT3b. Differently from ***Figure 5B***, 5meC levels are not normalized. Replicates of the same strain are represented as dotted lines. Data from BY4741-derived strains. BY4741 = Wild type (wt); EV = Empty vector. (**B**) Metagene plot of CpG methylation in *set1Δ*, *set2Δ*, and *dot1Δ* cells expressing DNMT3b. *set1Δ*, *set2Δ* are in a BY4741 background, while *dot1Δ* is in a W303 background. 5meC levels are normalized by DNMT3b *Figure 5. continued on next page*

Figure 5. Continued

expression measured by RT-qPCR. Two replicates for each strain are shown (solid and dotted line). (**C**) Metagene plots of CpG methylation ratio between the mutant and its wt counterpart. Two replicates for each mutant strain are shown (solid and dotted line). Wt ratios (=1) are represented by the horizontal dashed line (green or blue). (**D**) Boxplots showing levels of DNA methylation in the wt (left) and *set1Δ* strain (right) of 200-bp genome bins sorted into deciles by H3K4me3 level. (**E**) Boxplots showing levels of DNA methylation in the wt (left) and *set2Δ* strain (right) of 200-bp genome bins sorted into deciles by H3K36me3 level. The dashed red line represents background levels of DNA methylation due to incomplete bisulfite conversion (>99.7%). DOI: 10.7554/eLife.06205.017
The following figure supplement is available for figure 5:

Figure supplement 1. DNMT3b transcript levels in different yeast strains.
DOI: 10.7554/eLife.06205.018

Correspondence between H3K36me3 and early DNA methylation in mammalian cells

To extend our findings in yeast, we sought evidence to determine whether H3K36me3 also promotes de novo DNA methylation in mammals. The mouse germline is an excellent model for such studies. The mouse germline is specified from the epiblast at E7.25 and then progressively loses DNA methylation through subsequent rounds of cell division. By E13.5, almost all DNA methylation has been lost (***Popp et al., 2010***; ***Seisenberger et al., 2012***). In male germ cells, cell division halts, and the de novo methyltransferases and their co-factor DNMT3L are expressed between E13.5 and birth, when the genome undergoes global de novo DNA methylation (***Seisenberger et al., 2012***; ***Kobayashi et al., 2013***). Thus in this setting, DNA methyltransferases are introduced into hypomethylated cells, and are therefore an ideal model to study the targeting of de novo DNA methylation.

We mapped DNA methylation in the male germline at E16.5, P2.5, and P10.5 (***Supplementary file 1F***) (***Pastor et al., 2014***), and obtained E13.5 DNA methylation data from published sources (***Seisenberger et al., 2012***). Consistent with previous observations about the timing of de novo DNA methylation in the developing mouse germline, global CpG methylation rises from 7% at E13.5 to 55%at E16.5 and reaches at 75% by P2.5 (***Figure 6A***). Previous studies have shown that the entire male germline genome is methylated by default, except for regions of H3K4 methylation such as TSSs which antagonize de novo DNA methylation (***Singh et al., 2013***). However, charting the progression of DNA methylation over time, it is apparent that there exist 'early methylating' regions that reach their final methylation state by E16.5 and 'late methylating' regions that undergo substantial DNA methylation between E16.5 and P2.5. We observed that heavily transcribed regions of chromosomes showed much higher DNA methylation at E16.5 than less transcribed regions (***Figure 6B***). Furthermore, while the TSS of active genes was unmethylated, gene bodies of actively transcribed genes were typically early-methylators (***Figure 6B,D***). Thus, transcriptional initiation correlates negatively with de novo DNA methylation while transcriptional elongation correlates positively with de novo methylation.

In light of the data from yeast, we considered that the trends noted above could be caused by the underlying chromatin environment, with H3K4me3 antagonizing and H3K36me3 promoting de novo DNA methylation. Since transcriptional elongation causes H3K36me3 deposition, we asked whether the association of transcriptional read-through with DNA methylation could explain the observed phenomenon. To test this hypothesis, we analyzed published H3K4me3 ChIP-seq data (***Lesch et al., 2013***) and performed H3K36me3 ChIP-seq on sorted germ cells of pooled E13.5 testis (***Supplementary file 1G***). H3K4me3 at E13.5 correlates with low DNA methylation at all subsequent time points (***Figure 6D,E***). Genes with high H3K36me3 levels at E13.5 showed significantly elevated gene-body DNA methylation at E16.5, consistent with H3K36me3 accelerating DNA methylation (***Figure 6B,C,D,F***). This trend was still apparent at P2.5 (***Figure 6F***). Thus, H3K36me3 appears to direct DNA methylation in mammalian cells.

Discussion

Our study aimed to identify chromatin features that affect the activity of mammalian de novo DNMTs in the establishment of DNA methylation. The expression of the murine DNMT3b in a host with no detectable levels of $5^{me}C$ led to the methylation of CpG dinucleotides at different levels depending on the specific chromatin context. The presence of the H3K4me3 mark inhibits the activity of DNMT3b, while H3K36me3 promotes DNA methylation. This suggests that the activity of DNMT3B is guided by

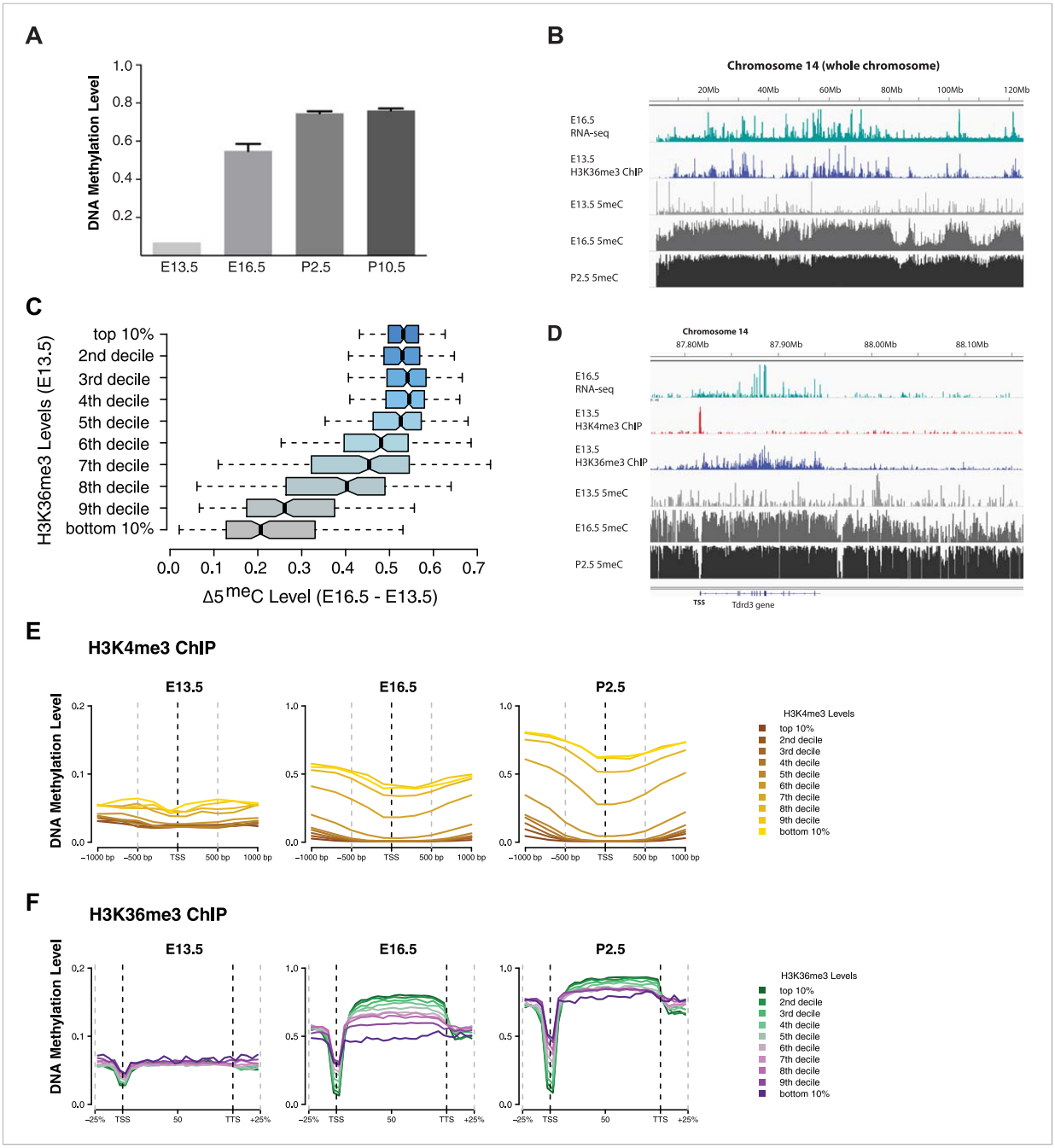


Figure 6. H3K4me3 and H3K36me3 distribution predicts de novo DNA methylation pattern in male germline. (**A**) Genome-wide CG methylation levels during murine development as measured by bisulfite sequencing. (**B**) RNA-seq, and ChIP read abundance and relative DNA methylation levels are plotted across chromosome 14. Note the correspondence between RNA-seq and H3K36me3 ChIP levels and rapid DNA methylation between E13.5 and E16.5.

(**C**) Boxplots showing the difference of DNA methylation levels between E13.5 and E16.5 of 1 Mb genome bins sorted into deciles by H3K36me3 level. (**D**) RNA-seq and ChIP read abundance and DNA methylation levels are plotted relative to transcriptionally active genes. The gene promoters contain high H3K4me3 and are not methylated, while the gene bodies contain high H3K36me3 and are methylated rapidly. (**E**) Metaplots showing DNA methylation level ±1000 bp relative to the TSS of genes sorted into deciles by H3K4me3 level. (**F**) Metagene plots showing DNA methylation across gene bodies sorted into deciles by H3K36me3 level.

DOI: 10.7554/eLife.06205.019

the interactions of the ADD and PWWP domains with histone tails. It has been recently shown (**Baubec et al., 2015**) that in embryonic stem cells the PWWP domain is responsible for the targeting of DNMT3b to regions enriched for the H3K36me3. Similarly to our finding in yeast, the reintroduction of DNMT3b into methylation deficient DNMT1/DNMT3A/DNMT3B triple KO (TKO) ES cells partially restores $5^{me}C$ levels. Methylation levels are higher at H3K36me3 sites, a trend eliminated by the ablation of the H3K36me3 methyltransferase Setd2 (**Baubec et al., 2015**). Our findings are in agreement with the Baubec et al. observations, both in a system where other factors guiding DNA methylation are absent (yeast), and during a period of biologically important de novo DNA methylation (germ cells).

In our yeast system, we detected an anti-correlation between transcript levels and DNA methylation, while we found a positive correlation in germ cells as was shown in ES cells (**Baubec et al., 2015**). According to our findings, the levels of DNA methylation are guided by the presence of two transcription-dependent marks: H3K4 and H3K36 methylation. The discrepancy between the findings in yeast and germ cells can be explained by the difference in the length of their genes. Yeast genes are relatively small compared to genes in higher eukaryotes so, H3K4 methylation can spread within the body of the gene, thus preventing the binding of the DNMT3-ADD domain to the N-terminus of histone H3 and reducing its activity. In contrast, in mammals, H3K4me is localized to the start of the gene, and does not spread significantly within the gene body. Hence, highly transcribed genes in mammals show a strong enrichment of H3K4me3 around the TSS and H3K36me3 into the gene-body, shaping their intragenic DNA methylation distribution.

The observation that transcriptional elongation is linked to DNA methylation has been noted in many contexts in addition to male germ cells. In mature oocytes, which have intermediate global levels of CpG methylation (~50%), similar to male E16.5 PGCs, actively transcribed gene bodies have far higher levels of DNA methylation than less transcribed genes and intergenic regions (**Smallwood et al., 2011**; **Kobayashi et al., 2012**). Also, in oocytes, intragenic CpG islands show far higher DNA methylation than other CpG islands (**Smallwood et al., 2011**). Transcriptional read-through is a common feature of maternally imprinted loci (**Weaver and Bartolomei, 2014**) and ablation of an upstream promoter prevents proper methylation of the imprinted *Gnas* locus (**Chotalia et al., 2009**). In mammalian soma, inactive X-chromosome shows higher promoter methylation, consistent with its silent state, but markedly lower intragenic methylation (**Hellman and Chess, 2007**). Transcriptional elongation is also correlated with DNA methylation in tumor cells (**Jin et al., 2012**). It has been suggested that transcriptional read-through could 'open' chromatin for DNMTs, or that heterochromatin is physically inaccessible to DNMTs. We suggest however that direct recruitment of DNMTs by H3K36me3 is the most likely mechanism for the correlation between transcriptional read-through and DNA methylation.

H3K36me3 functions both to suppress intragenic transcriptional initiation through recruitment of histone deacetylases, and to promote DNA methylation. These marks likely cooperate to induce lasting silencing of transcriptional initiation at target loci (**Figure 7**, **Figure 7—figure supplement 1**). Intragenic TSSs originating at transposons have the potential to generate truncated or transposon/gene hybrid transcripts that could be deleterious to cell survival. H3K36me3 and DNA methylation could cooperate to silence these transposons in the germline and other periods of de novo methylation, and to maintain silencing through development. Moreover, where multiple TSSs exist for a gene, as in many imprinted loci, H3K36me3-mediated DNA methylation may serve to ensure the dominance of one promoter in a given cell type.

A number of H3K36 methyltransferases exist in mammals but only one, SETD2, can catalyze the conversion of H3K36me2 to me3 (**Wagner and Carpenter, 2012**). *Setd2*$^{-/-}$ mice exhibit profound vascular defects and die at E10.5–E11.5 (**Hu et al., 2010**), while *Setd2*$^{-/-}$ are defective for differentiation toward endoderm (**Zhang et al., 2014**). *Setd2* is also a tumor suppressor mutated frequently in leukemia (**Zhu et al., 2014**). It will be important to determine how loss of *Setd2* affects the distribution of DNA methylation in the germline and soma, and whether loss of *Setd2* contributes to aberrant methylation in cancer.

More broadly, targeting of DNMT enzymes by association with H3K36me3 could explain methylation distribution across plants and animals. All catalytically active DNMT3-family methyltransferases in animals contain PWWP domains, and accordingly, gene body DNA methylation is observed in all animals that have retained DNMT3 enzymes. Preferential methylation of gene bodies over intergenic regions is observed for invertebrates such as honey bees (*Apis mellifera*), sea squirts (*Ciona intestinalis*), sea anemones (*Nematostella vectensis*) (**Zemach and Grafi, 2003**; **Feng et al., 2010**). While the relationship between relative gene expression and gene-body methylation varies

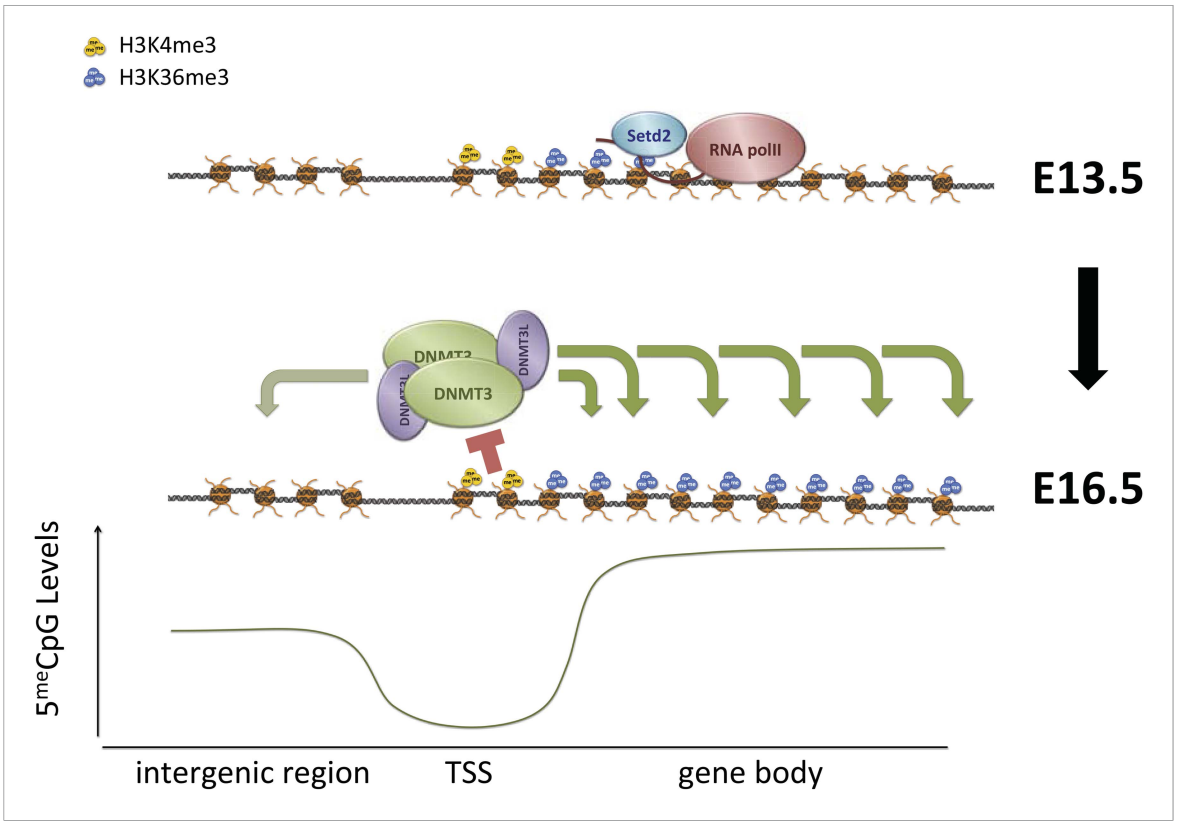


Figure 7. Proposed model for de novo DNA methylation establishment. Model proposed for the targeting of DNMT3 during events of de novo 5meC establishment after genome-wide erasure of DNA methylation. Our model suggests that the presence of transcription-dependent histone modifications, such as H3K4me3 and H3K36me3, determines the activity of DNMT3b in vivo.
DOI: 10.7554/eLife.06205.020
The following figure supplement is available for figure 7:

Figure supplement 1. Factors affecting DNA methylation deposition.
DOI: 10.7554/eLife.06205.021

across these species, there is a strong correlation between gene-body H3K36me3 in *Drosophila melanogaster* genes and DNA methylation of homologous gene bodies in other invertebrates (***Nanty et al., 2011***). DNA methylation is also associated with gene bodies in zebrafish (*Danio rerio*) (***Zemach and Grafi, 2003***) and in mammalian contexts as discussed above. Finally, some chlorophyte algae have a 'chlorophyte-type cytosine methylase', which evolved independently of DNMT3-family methyltransferases, which is fused to two PWWP domains (***Iyer et al., 2011***). Thus, H3K36me3 could be relevant to DNA methylation targeting throughout the plant and animal kingdoms.

Materials and methods

Experimental methods

Yeast strains, plasmids, media and culture Conditions

All the plasmids, primers, and strains used in this study are listed in ***Supplementary file 6A,B***. Murine DNMT3b isoform 1 was amplified from the plasmid pCR-Blunt II-TOPO and subcloned into pYES2 (Life Technologies, Carlsbad, CA) using *Hind*III and *Bam*HI. All the plasmids were introduced in yeast using the standard Lithium Acetate Procedure (***Gietz and Schiestl, 2007***). Mutant yeast strains *set1*Δ and *set2*Δ were kindly donated by the Kurdistani Lab (UCLA), while the *dot1*Δ strain (W303 background) was prepared via PCR-mediated gene disruption (***Wach, 1996***) using primers listed in ***Supplementary file 6C***.

All the yeast strains were grown at 30˚C in SC + Galactose 2% without uracil (Sunrise Science, San Diego, CA, cat 1652 and 1485-100) overnight. The next morning, cells were diluted to 0.3 OD_{600}/ml and grown to mid-log phase (0.8–1 OD_{600}/ml) or to stationary phase (5–6 OD_{600}/ml or for 24–30 hr).

Yeast WGBS libraries preparation

DNA was collected from yeast cells according to ***Hoffman (2001)*** with minor changes. Briefly, about 5 OD_{600} of yeast cells were disrupted by vortexing for 6 min in a Disruptor Genie (Scientific Industries, Inc., Bohemia, NY) in the presence of an equal volume of breaking buffer, acid-washed glass beads and phenol:chloroform (1:1). After the addition of TE buffer, aqueous phase is transferred into a new tube and precipitated with ethanol. The nucleic acid pellet is resuspended in TE buffer and treated for 1 hr at 37˚C with RNaseA, followed by incubation for 1 hr with 2 mg/ml proteinase K in the presence of 1% SDS at 60˚C. The resulting solution is treated twice with phenol:chloroform (1:1), once with chloroform and ethanol precipitated. The DNA pellet is resuspended in EB buffer (Qiagen, Valencia, CA). Between 500
and 1000 ng of extracted yeast DNA is added to 2 ng of λ unmethylated DNA (Promega, Madison, WI, D1521) and the mixture is sonicated with a Covaris S-2 to obtain fragments in the 200–300 bp range (Total time: 6 min; Duty cycle: 10%; Intensity: 5; Cycles/Burst: 200; Mode: Frequency sweeping). The reagents used in the library preparation are from the Illumina TruSeq DNA Sample Prep kit v2 (Illumina, San Diego, CA). End-Repair, purification and dA-tailing steps are performed according to
manufacturer's instructions. Ligation is performed according to the protocol except that 1 µl of Illumina TruSeq Adapters is used in the final reaction. The ligation reaction is purified using 1.2 vol of AMPure XP beads (Beckman Coulter Inc. Indianapolis, IN,) and DNA fragments with a 170–350 bp range are enriched using 0.7 and 0.3 vol of AMPure XP beads in the first and second size-selection step, respectively. Samples are treated with bisulfite (EpiTect kit, QIAGEN) according to manufacturer's protocol, except that two consecutive rounds of conversion are performed, for a total of 10 hr of incubation. Half of the converted DNA is amplified using MyTaq Mix (Bioline, Taunton, MA,) and Illumina TruSeq PCR Primer Cocktail according to the following protocol: initial denaturation at 98˚C for 30 s; 12 cycles of 98˚C for 15 s, 60˚C for 30 s, 72˚C for 30 s; final extension at 72˚C for 5 min. The final product is purified using AMPure XP beads before being submitted for sequencing. Libraries were sequenced with an Illumina HiSeq 2000 system using 50 bp or 100 bp single-end reads.

Yeast MNase-seq libraries preparation

Nucleosome mapping has been performed according to ***Rando (2010)*** with minor modifications. Stationary phase yeast culture (≈6 OD/ml) is cross-linked with 1% formaldehyde for 20 min with occasional rotation at room temperature. The reaction is quenched with glycine for 5 min at room temperature. About 60 OD of yeast cells are washed twice with PBS buffer and then resuspended in
2 ml of Z buffer (1 M sorbitol, 50 mM Tris-HCl pH 7.4 with freshly added 10 mM β-mercaptoethanol) containing 3.6 mg of Zymolyase-20T (from *Arthrobacter luteus*, AMS Biotechnology, Cambridge, MA,) and incubated at 37˚C in agitation. After 45 min the same amount of Zymolyase is added and each sample which is incubated for an additional 45 min at 37˚C in agitation. Spheroplasts are then pelleted by centrifugation for 5 min at 4˚C at 1500 g. The pellet is washed once with NP-buffer, then resuspended in 1.6 ml of NP-buffer and divided in three tubes. An increasing amount of MNase (Sigma, N3755, St. Louis, MO,) is added to each tube: 0.25 U, 0.5 U, and 1 U. After incubation for 20 min at 37˚C, each reaction is stopped by the addition of SDS and EDTA to a final concentration of 1%and 10 mM, respectively. The reaction is then treated with 0.2 mg/mg of proteinase K (NEB, Ipswich, MA) at 65˚C overnight. The sample is then purified with two rounds of phenol:chloroform (1:1) and the aqueous solution precipitated. The resuspended DNA pellet is treated for 1 hr with RNase A at 37˚C. For the naked DNA digestion, 200 ng of extracted DNA is incubated at 37˚C with 0.01 U of MNase. After 7 min the reaction is stopped as described before. Both naked and RNaseA-treated nucleosomal DNAs are then purified using 1.8 vol of AMPure XP beads and the libraries prepared using NEBNext DNA Library Prep Master Mix Set for Illumina (NEB, E6040S) with few modifications of the manufacturer's protocol. Only the digestion pattern obtained with 0.5 U of MNase was used for the library preparation. The DNA is end-repaired (in half of the suggested volume), dA-tailed, and 1 µl of Illumina TruSeq Adapters is added to a 40 µl ligation reaction. Purification after each step is performed using AMPure XP beads according to the protocol. The size selection step is carried out with 0.8× of AMPure beads in the first step and 0.2× of AMPure XP beads in the second step. Half of the DNA is amplified using Illumina PCR Master Mix and Illumina TruSeq PCR Primer Cocktail (TruSeq DNA Sample Prep kit v2) with the following protocol: initial denaturation at 98˚C for 30 s; 12 cycles of 98˚C for 15 s, 60˚C for 30 s, 72˚C for 30 s; final extension at 72˚C for 5 min. The final product is purified using AMPure XP beads before being submitted for sequencing. Libraries were sequenced with an Illumina HiSeq 2000 system using 50 bp single-end reads.

Yeast RNA-seq libraries preparation

RNA was collected from 5 OD of yeast cells (***Collart and Oliviero, 2001***). Approximately 500–1000 ng of extracted yeast RNA are used to prepare RNA-seq libraries using Illumina TruSeq mRNA Library Prep Kit v2 according to manufacturer's instructions. Libraries were sequenced with an Illumina HiSeq 2000 system using 50 bp single-end reads.

RT-qPCR

Quantitative RT-PCR was used to determine the relative expression of DNMT3b in wild-type and mutant yeast strains. Briefly, 1 µg of total RNA was subject to polyA enrichment using TruSeq oligo-dT magnetic beads (part # 15026778, Illumina) and reverse transcribed using SuperScript III (cat # 18080-044, Life Technologies) according to manufacturer's instruction. An equal amount of cDNA was used for each qPCR reaction, using primers listed in ***Supplementary file 6C***. Murine DNMT3b expression levels were normalized to TDH1 levels and the relative expression between the wild-type and each mutant was calculated using the ΔΔCt method (***Schmittgen and Livak, 2008***).

Yeast ChIP-seq libraries preparation

Chromatin immunoprecipitation experiments were conducted according to ***Kitada et al. (2012)***, with minor modifications. Briefly, 50 OD of yeast cells are crosslinked using 1% formaldehyde for 15 min at room temperature and quenched with glycine 125 mM for 5 min at room temperature. After two washes with ice-cold PBS, the cells are resuspended in yeast lysis buffer (with 140 mM NaCl for DNMT3b and RNApolII or 500 mM NaCl for histone post-translational modifications) and the same volume of acid-washed glass beads. We disrupted the cells by vortexing for 5 min in a Disruptor Genie at 4˚C and incubating in iced-water for 2 min. We repeated the cycle for an additional 5 times. We collected the lysate by centrifugation after creating a hole on the bottom of the tube with a 25-G needle. We transferred a fraction of the lysate into a microTube (AFA filter—Covaris, Woburn, MA) and proceeded with the sonication using the Covaris S2 system according to the following parameters: 14 cycles of 30 s ON, 30 s OFF; Duty cycle = 5%; Intensity = 5%; Cycles/Burst = 200. The sonicated lysate is clarified via centrifugation and 50 µl of the supernatant is incubated overnight at 4˚C with a specific antibody (***Supplementary file 6D***). 10 µl of the clarified lysate is used as input control. The next day, immunoprecipitations are incubated 2 hr at 4˚C with Protein A Dynabeads (Life Technologies). Each wash is performed twice in the following order: low-salt buffer (50 mM HEPES pH 7.5, SDS 0.1%, 1% Triton X-100, 0.1% Deoxycholate, 1 mM EDTA, 140 mM NaCl), high salt buffer (50 mM HEPES pH 7.5, SDS 0.1%, 1% Triton X-100, 0.1% Deoxycholate, 1 mM EDTA, 500 mM NaCl), LiCl buffer (10 mM Tris-HCl pH 8, 250 mM LiCl, 5 mM EDTA, 1% Triton-X, 0.5% NP-40), TE buffer (100 mM Tris-HCl pH 8, 10 mM EDTA). Elution is performed at 65˚C with TE/SDS buffer (100 mM Tris-HCl pH 8, 10 mM EDTA, 1% SDS). Tubes containing the eluted immunoprecipitations and input controls (additioned of TE/SDS buffer) are incubated overnight at 65˚C to reverse the cross-links. RNase treatment is performed at 37˚C for 1 hr, followed by a proteinase K treatment for 1 hr at 60˚C. Each reaction is then purified using 1.8 vol of AMPure XP beads according to manufacturer's instructions. Libraries were prepared with Ovation Ultralow DR kit (Nugen Technologies, San Carlos, CA) starting from 1 ng of purified DNA according to the protocol. Libraries were sequenced with an Illumina HiSeq 2000 system using 50 bp single-end reads.

Mice

Mice homozygous for a characterized Oct4-IRES-GFP allele (***Wernig et al., 2007***) were used for murine H3K36me3 ChIP. Embryonic male germ cells express the GFP marker and can be sorted efficiently (***Vincent et al., 2011***; ***Pastor et al., 2014***).

Bisulfite sequencing and RNA-seq data (mouse)

Whole genome bisulfite sequencing data from sorted E16.5, P2.5, and P10.5 germ cells was generated as part of a parallel project studying the transposon silencer *Morc1* (***Pastor et al., 2014***), with the data from the phenotypically normal *Morc1*$^{+/-}$ controls from that study serving as methylomes in this study. Briefly, germ cells from between three to five male mice at each time point were harvested and libraries generated, and reads from these libraries were pooled. E13.5 bisulfite sequencing data were taken from replicate two of (***Seisenberger et al., 2012***). Genome-wide bisulfite sequencing average coverage was 5.36 (E13.5), 14.57 (E16.5), and 8.52 (P2.5). RNA-seq data from two E16.5 *Morc1*$^{+/-}$ controls from (***Pastor et al., 2014***) were also used in this study.

Mouse germ cells purification for ChIP

Collection of embryonic testes was performed following institutional approval for appropriate care and use of laboratory animals, according to published protocols (***Pastor et al., 2014***). Pregnant females were euthanized using CO_2 and the embryos removed from the womb and stored on a 10 cm dish filled with chilled 1× PBS. Testicles were removed from the embryos, placed in an individual 15 ml
falcon tube with 3 ml of 0.25% Trypsin with 3 µl of DNAse I 1 U/1 µl (Life Technologies). Testes were incubated for 15 min at 37˚C. After incubation the cells were agitated into suspension gently by pipetting. The trypsin was then quenched using 5 ml DMEM/10% FBS (Life Technologies). The cells
were centrifuged at 278 g for 5 min and resuspended in 500 µl FACS buffer (1× PBS 1% BSA). 7-AAD was added at a 1:50 dilution (BD Biosciences, San Jose, CA) and the cells strained through BD FACS tubes (Corning, Union City, CA) before analysis. GFP positive cells were sorted for ChIP.

Mouse ChIP-seq

The ChIP-seq protocol was adapted from published sources (***Ng et al., 2013***; ***Pastor et al., 2014***). FACS sorted cells from four male, E13.5 embryos were diluted to 292 µl with room temperature 1× PBS. 8.11 µl 37% Formaldehyde (Sigma) was added and the sample was incubated 10 min at room temperature with rocking. 48.8 µl of 1 M glycine was then added to yield a final concentration of 0.14 M and the samples were quenched 30 min with rocking. Cells were then spun 425 g for 10 min at RT. The cell pellet was flash frozen.

After thawing, the cells were resuspended in 300 µl Lysis buffer (50 mM Tris-Cl pH 8.0, 20 mM EDTA pH 8.0, 0.1% SDS, 1× Complete Protease Inhibitor [Roche]) and incubated on ice 10 min. Samples were then sonicated by Covaris S2 (Intensity 5, cycles/burst = 200, duty cycle = 5%, 10 × 30 s on 30 s off sonication). Samples were spun 14000 g 10 min to remove insoluble material. The soluble
sample was diluted to 600 µl with dilution buffer (16.7 mM Tris pH 8, 0.01% SDS, 1.1% Triton X-100, 1.2 mM EDTA, 167 mM NaCl) and 10% of material was saved as input. Sample was precleared with
30 µl Protein A Dynabeads (Life Technologies) and preincubated 1 hr. The cleared material was incubated with 1 µl L anti-H3K36me3 antibody (Abcam Ab9050) overnight.

The samples were incubated with 30 µl Protein A Dynabeads and the precipitated material was recovered with a magnet. The beads were washed 2 × 4 min with Buffer A (50 mM HEPES pH 7.9, 1%Triton X-100, 0.1% Deoxycholate, 1 mM EDTA, 140 mM NaCl), 2 × 4 min with Buffer B (50 mM HEPES pH 7.9, 0.1% SDS, 1% Triton X-100, 0.1% Deoxycholate, 1 mM EDTA, 500 mM NaCl) and 2 × 4 min with 10 mM Tris/1 mM EDTA. Bound material was eluted with 100 µl Elution buffer (50 mM Tris pH 8.0, 1 mM EDTA, 1% SDS) at 65˚C for 10 min and then eluted a second time with 150 µl elution buffer.

The input samples were thawed and diluted with 200 µl buffer. Crosslinking of ChIP and input samples was reversed by incubating 16 hr at 65˚C. Samples were cooled and treated with 1.5 µl of 10 mg/ml RNaseA (PureLink RNAse A, Invitrogen #12091-021) for 30 min at 37˚C. 100 µg of Proteinase K was then added and the samples treated for 2 hr at 56˚C. The samples were then purified using a Qiagen MinElute kit.

Samples were amplified by a SeqPlex DNA Amplification kit (Sigma) and then converted to libraries using an Ovation Rapid Library kit.

Data processing and analysis

Bisulfite sequencing

Reads from bisulfite-treated yeast and mouse genomic DNA (***Seisenberger et al., 2012***; ***Pastor et al., 2014***) were aligned using BS-Seeker2 v2.0.3 (***Guo et al., 2013***) against the sacCer3 and mm9 genome assemblies, respectively. Up to four mismatches were allowed and bowtie (v0.12.8) was specified as the aligner. Methylation was called using default parameters of BS-Seeker2.

MNase-sequencing

Reads from both naked and nucleosomal DNA sequencing were aligned using bowtie v0.12.8 (***Langmead et al., 2009***) against the sacCer3 genome assembly, allowing up to two mismatches. Nucleosome calling was performed using DANPOS v2.1.3 (***Chen et al., 2013***) subtracting the naked DNA-derived reads from the nucleosomal reads and using the '-k1 -e1' parameters (***Supplementary files 1B, 4***).

RNA sequencing

RNAseq reads from mouse germ cells (***Pastor et al., 2014***) and yeast were aligned against the mm9 and sacCer3 genome assemblies using STAR v2.3.1 (***Dobin et al., 2013***) with the following parameters: --outFilterMismatchNoverLmax 0.04 --outFilterMultimapNmax 1.

Differential expression was performed using the DEseq package (***Anders and Huber, 2010***) in R-Bioconductor. Differentially expressed genes are defined as having more than twofold difference in the level of the corresponding RNA and a false discovery rate (p-adj) smaller than 0.1. GO term enrichment for upregulated and dowregulated genes in the DNMT3b-expressing compared to the EV was performed using the Gene Ontology Term Finder tool on the Sccharomyces Genome Database website (http://www.yeastgenome.org/cgi-bin/GO/goTermFinder.pl). RPKM values were calculated using rpkmforgenes.py (available at http://sandberg.cmb.ki.se/media/data/rnaseq/rpkmforgenes.py) specifying the following options: -fulltranscript -nocollapse -rmnameoverlap –allmapnorm (***Supplementary file 4***).

ChIP sequencing

Reads from yeast and mouse (this study and [***Lesch et al., 2013***]) were first mapped against the yeast (sacCer3) and mouse (mm9) genome, respectively, using bowtie v0.12.8 (***Langmead et al., 2009***), then aligned reads were processed according to ***Ferrari et al. (2012)***.

Linear model of methylation

The yeast genome was divided in 200-bp bins and log-transformed average levels of each feature calculated for each bin. The model was built using simple linear regression lm() function in R and the resulting prediction correlated (Pearson) with the observed values for both $5^{me}C$ levels and DNMT3b occupancy.

Data access

Data can be accessed at GEO (Gene Expression Omnibus) under the accession GSE6691.

Acknowledgements

We are grateful to Dr Maria Vogelauer for the helpful discussion and insights on the project. Yeast mutant strains were kindly provided by the Kurdistani Lab (UCLA). We also thank the Broad Stem Cell Research Center High-Throughput Sequencing and Flow Cytometry Cores.

Additional information

Funding

Funder	Grant reference	Author
National Institutes of Health (NIH)	R01 GM095656-01A1	Matteo Pellegrini
National Institutes of Health (NIH)	GM60398	Steven E Jacobsen
National Institute of Child Health and Human Development (NICHD)	R01HD058047	Amander T Clark
Howard Hughes Medical Institute (HHMI)		Steven E Jacobsen
Jane Coffin Childs Memorial Fund for Medical Research		William A Pastor
Whitcome Fellowship		Marco Morselli

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

MM, WAP, Conception and design, Acquisition of data, Analysis and interpretation of data, Drafting or revising the article; BM, Conception and design, Acquisition of data, Drafting or revising the article; KN, LR, Acquisition of data, Drafting or revising the article; RF, KF, GB, Analysis and interpretation of data, Drafting or revising the article; ATC, SEJ, MP, Conception and design, Drafting or revising the article; SO, Drafting or revising the article, Contributed unpublished essential data or reagents.

Ethics

Animal experimentation: All animal experimentation was conducted with the highest ethical standards in accordance with UCLA policy and procedures (DHHS OLAW A3196-01, AAALAC #000408 and protocol # 2008-070), and applicable provisions of the USDA Animal Welfare Act Regulations, the Public Health Service Policy on Humane Care and Use of Laboratory Animals, and the Guide for the Care and Use of Laboratory Animals.

Additional files

Supplementary files

- Supplementary file 1. (**A**) Yeast Whole Genome Bisulfite Sequencing Data. (**B**) Yeast MNase Sequencing Stats. (**C**) Yeast mRNA Sequencing Stats. (**D**) Yeast ChIP Sequencing Stats. (**E**) Yeast Whole Genome Bisulfite Sequencing Data for mutant strains. (**F**) Yeast Whole Genome Bisulfite Sequencing in mouse. (**G**) ChIP Sequencing Stats in mouse.
DOI: 10.7554/eLife.06205.022
- Supplementary file 2. (**A**) Yeast dinucleotide context methylation. (**B**) Yeast mutant strains dinucleotide context methylation. (**C**) Mouse germ cells dinucleotide context methylation.
DOI: 10.7554/eLife.06205.023
- Supplementary file 3. (**A**) Nucleosome called in a DNMT3b-expressing strain. (**B**) Nucleosome called in a non DNMT3b-expressing strain (EV). (**C**) Differential nucleosomes between DNMT3b-expressing and non-expressing strain.
DOI: 10.7554/eLife.06205.024
- Supplementary file 4. (**A**) Yeast mRNA differential expression using Deseq. (**B**) Upregulated genes in DNMT3b-expressing strain vs EV. (**C**) Downregulated genes in DNMT3b-expressing strain vs EV. (**D**) Gene Ontology (GO) term analysis for upregulated genes in DNMT3b-expressing strain vs EV. (**E**) GO term analysis for downregulated genes in DNMT3b-expressing strain vs EV. (**F**) RPKM values of yeast verified ORF in DNMT3b-expressing strain.
DOI: 10.7554/eLife.06205.025
- Supplementary file 5. Correletion coefficients of DNMT3b occupancy and $5^{me}C$ levels predictions.
DOI: 10.7554/eLife.06205.026
- Supplementary file 6. (**A**) Plasmids used in this study. (**B**) Yeast strains used in this study. (**C**) Oligonucleotides used in this study. (**D**) Antibodies used in this study.
DOI: 10.7554/eLife.06205.027

Major datasets

The following dataset was generated:

Author(s)	Year	Dataset title	Dataset ID and/or URL	Database, license, and accessibility information
Morselli M, *et al.*,	2015	In vivo targeting of de novo DNA methylation by histone modifications in yeast and mouse	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=anmpigiuppmnfeh&acc=GSE66911	Publicly available at the NCBI Gene Expression Omnibus (GSE6691).

The following previously published datasets were used:

Author(s)	Year	Dataset title	Dataset ID and/or URL	Database, license, and accessibility information
Seisenberger S, *et al.*,	2012	The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells	http://www.ebi.ac.uk/ena/data/view/ERP001953	Publicly available at the EBI European Nucleotide Archive (ERP001953).
Pastor W, *et al.*,	2014	MORC1 represses transposable elements in the mouse male germline	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63048	Publicly available at the NCBI Gene Expression Omnibus (GSE63048).

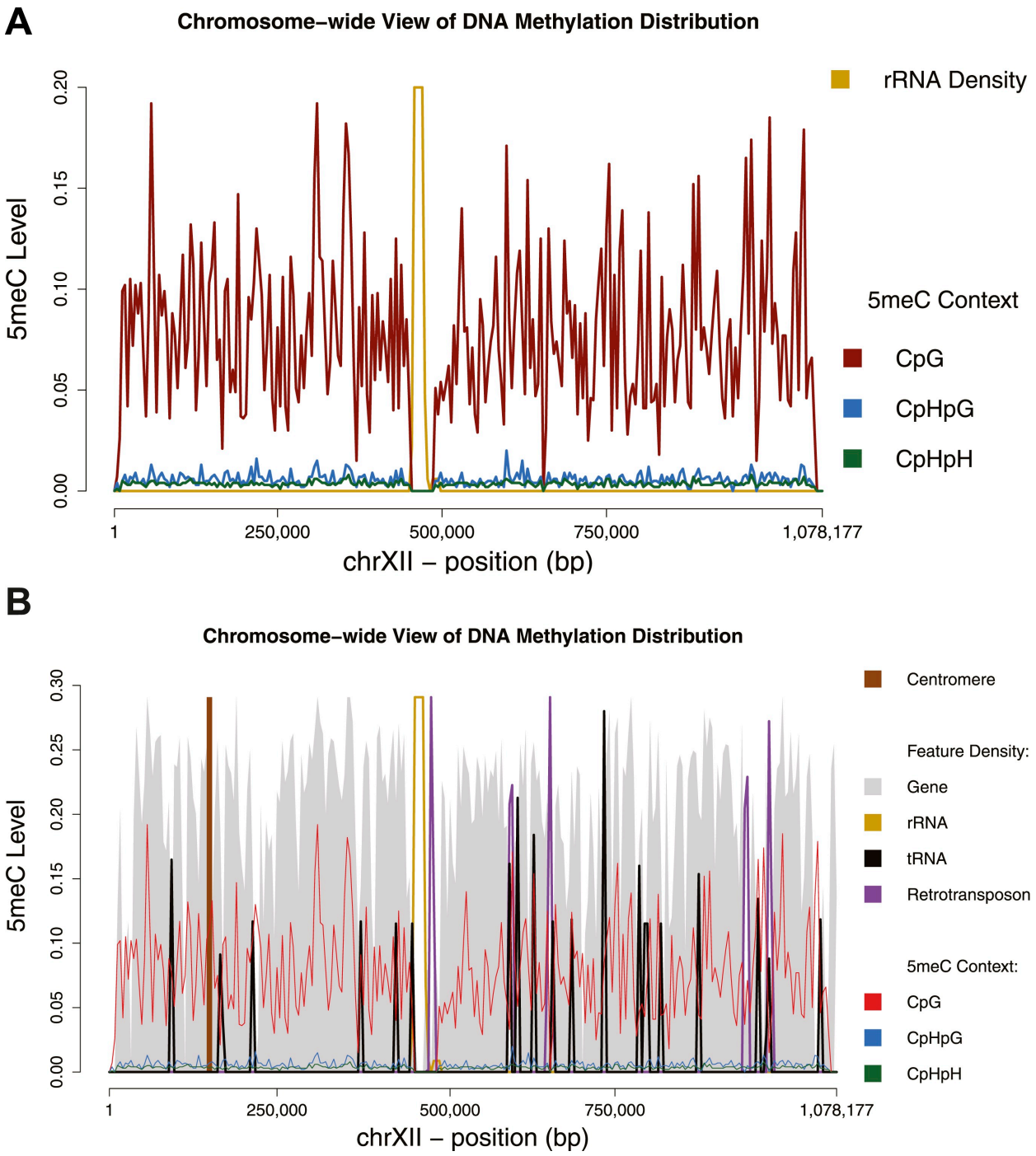


Figure 1—figure supplement 1. Chromosome-wide view of DNA methylation and genomic features. Distribution of DNA methylation on chromosome XII of S. cerevisiae (A and B). In (B) the density of other genomic features is shown (arbitrary units). Averages for DNA methylation and genomic features are calculated on 4 Kb bins. Areas of repetitive sequences (such as rRNA and transposable elements) show very little to no coverage. Gene-rich bins also correspond to peaks in DNA methylation levels.
DOI: http://dx.doi.org/10.7554/eLife.06205.004

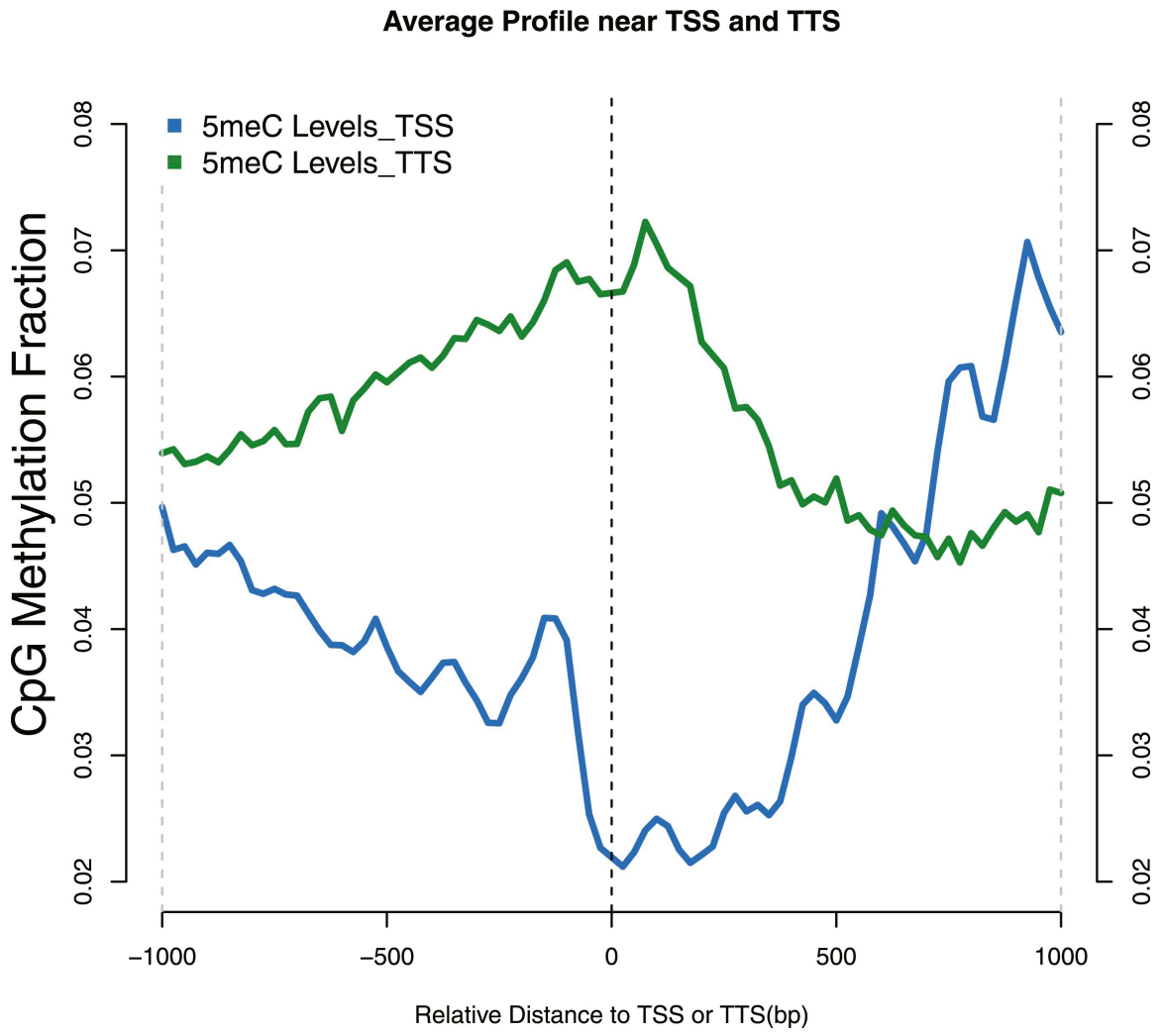


Figure 1—figure supplement 2. Distribution of 5meC around TSS and TTS.
CpG methylation levels around (TSSs—blue) and (TTSs—green) of yeast genes. Periodic peaks of DNA methylation are evident at the TSS, where nucleosomes form a well positioned array.
DOI: http://dx.doi.org/10.7554/eLife.06205.005

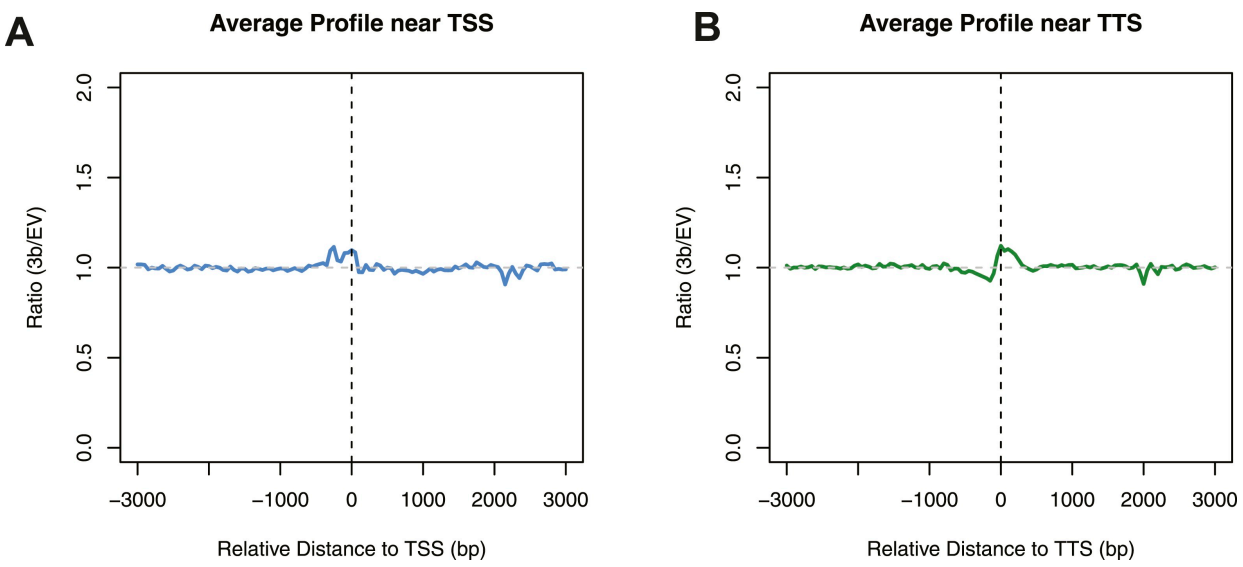


Figure 2—figure supplement 1. Differences in nucleosome occupancy between DNMT3b-expressing and non-expressing yeast strains.
Ratio of nucleosome occupancy between DNMT3b-expressing (3b) and non-expressing (EV) yeast strains at TSS (A) and TTS (B).
DOI: http://dx.doi.org/10.7554/eLife.06205.007

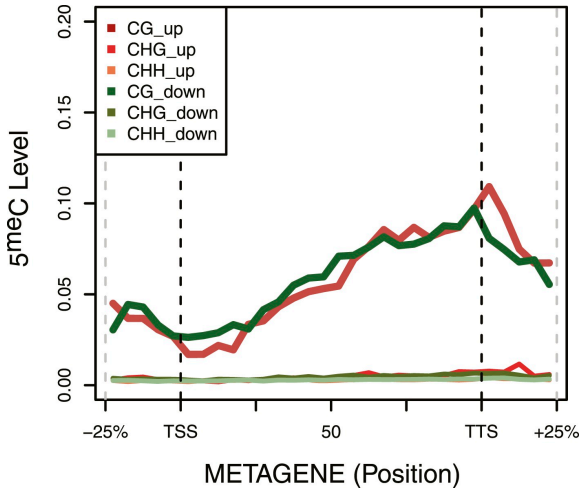


Figure 3—figure supplement 1. DNA Methylation in up- and down-regulated genes. Metagene plot of 5meC in different contexts (CpG, CpHpG, CpHpH) of upregulated (red) and downregulated (green) genes.

DOI: http://dx.doi.org/10.7554/eLife.06205.009

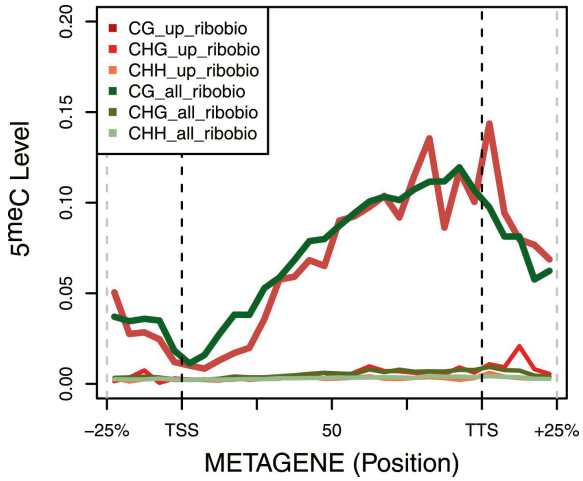


Figure 3—figure supplement 2. DNA Methylation in ribosomal biogenesis genes. Metagene plot of 5meC in different contexts (CpG, CpHpG, CpHpH) of upregulated ribosomal biogenesis genes (red) compared to all the genes of the same class (green).

DOI: http://dx.doi.org/10.7554/eLife.06205.010

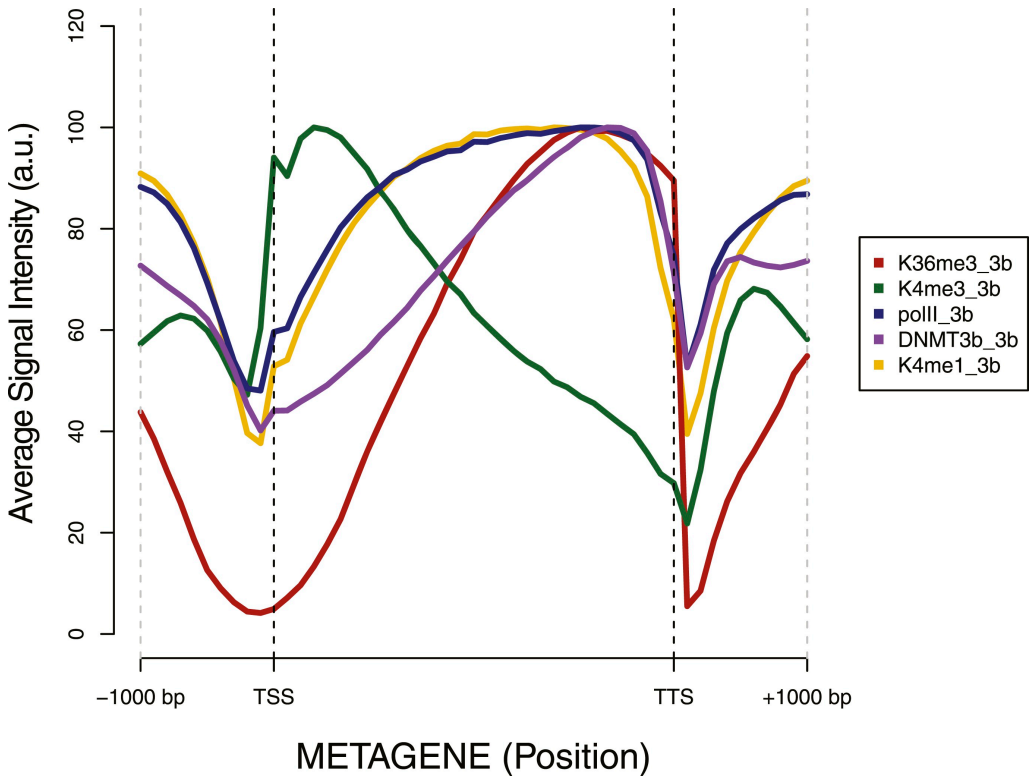


Figure 4—figure supplement 1. Metagene plot of ChIP sequencing in a DNMT3b-expressing strain.
ChIP-seq reads average intensity across yeast genes and 1 Kb upstream and downstream.
DOI: http://dx.doi.org/10.7554/eLife.06205.012

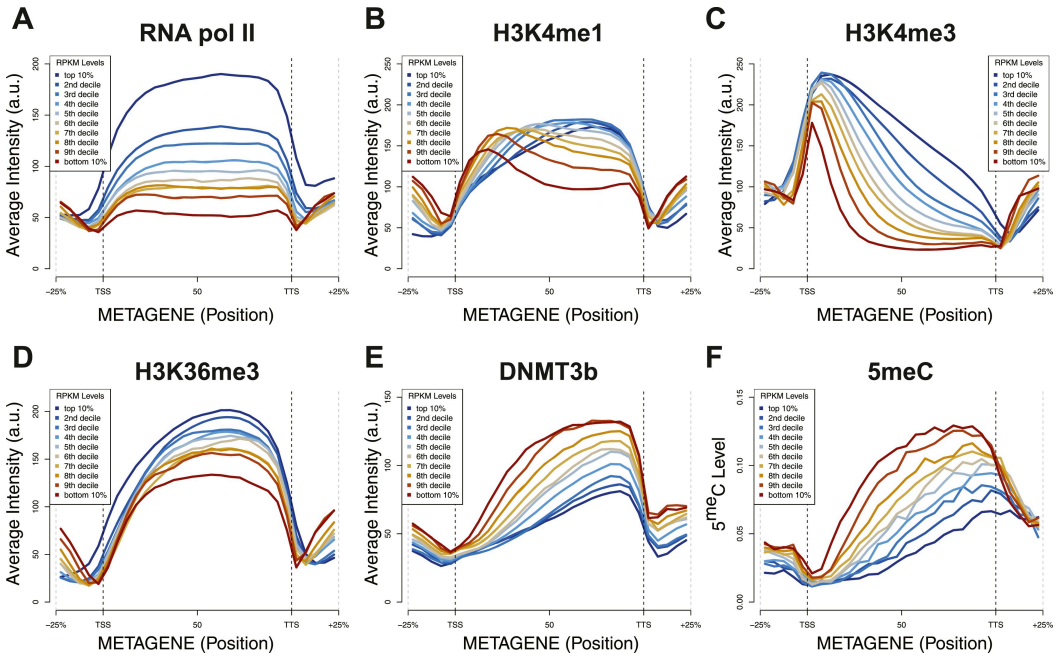


Figure 4—figure supplement 2. Relationship between transcription and 5meC or histone marks levels.

Average ChIP-seq intensity (A–E) or 5meC levels (F) across yeast genes divided in deciles based on RNA values (RPKM). a.u. = Arbitrary units.

DOI: http://dx.doi.org/10.7554/eLife.06205.013

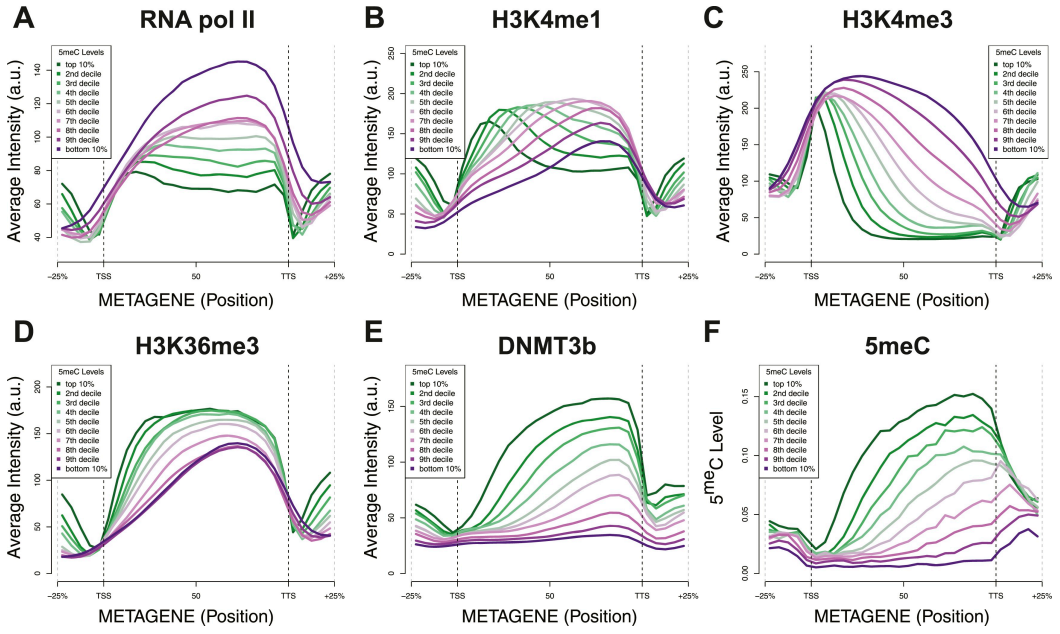


Figure 4—figure supplement 3. Relationship between DNA methylation and histone marks levels.

Average ChIP-seq (A–E) or DNA methylation (F) distribution across yeast genes divided in deciles based on average 5meCpG intragenic levels. a.u. = Arbitrary units.

DOI: http://dx.doi.org/10.7554/eLife.06205.014

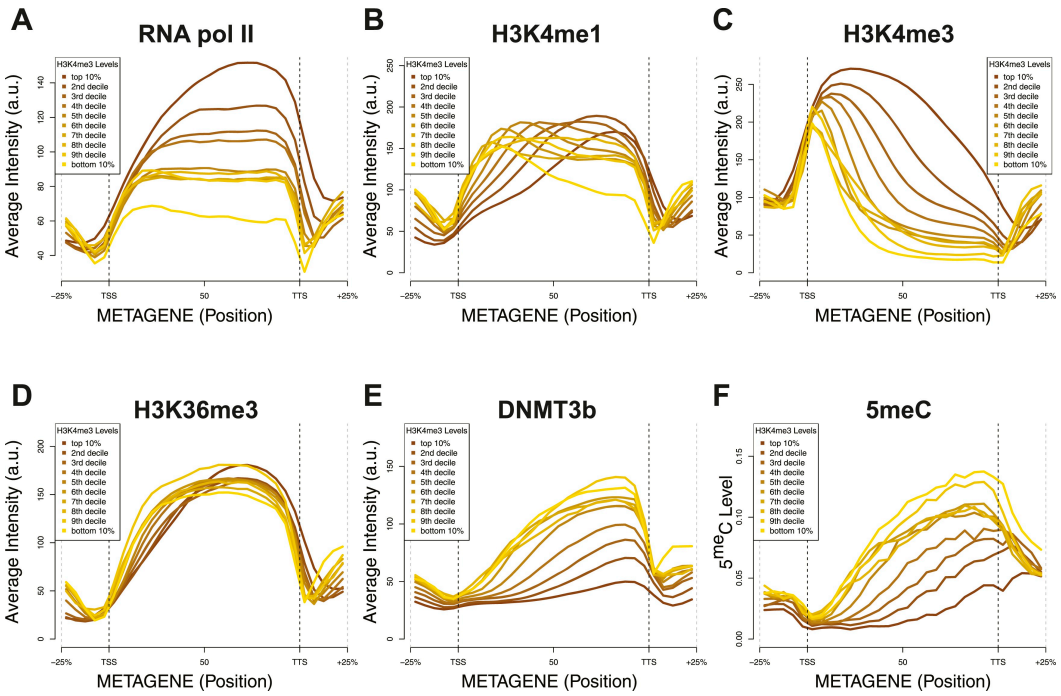


Figure 4—figure supplement 4. Relationship between H3K4me3 and 5meC or histone marks levels.
Average ChIP-seq intensity (A–E) or 5meC levels (F) across yeast genes divided in deciles based on H3K4me3 average in the last third of each gene. a.u. = Arbitrary units.
DOI: http://dx.doi.org/10.7554/eLife.06205.015

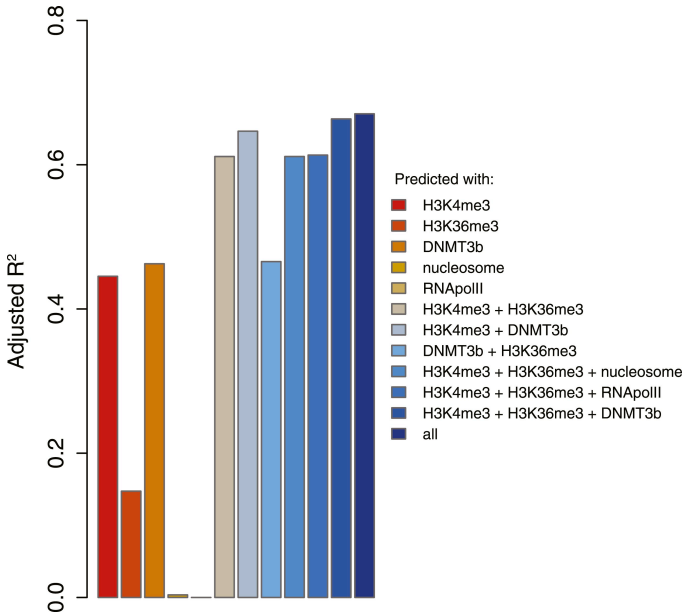


Figure 4—figure supplement 5. 5meC levels prediction using chromatin marks.
Prediction of 5meC levels across the genome divided in 200-bp bins with a linear multivariate regression method using several combinations of chromatin marks. On the y-axis the adjusted R^2 value is reported.
DOI: http://dx.doi.org/10.7554/eLife.06205.016

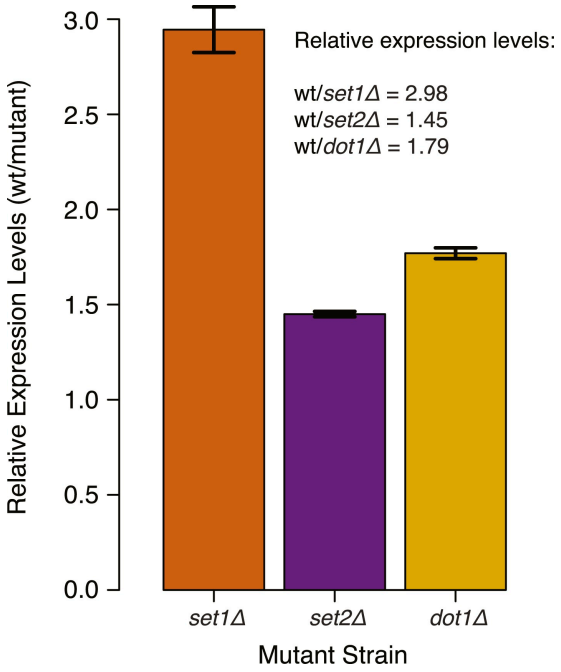


Figure 5—figure supplement 1. DNMT3b transcript levels in different yeast strains. Expression levels of DNMT3b of the wild-type strain compared to yeast mutants (set1Δ, set2Δ, and dot1Δ). The relative levels were calculated using the ΔΔCt method (Schmittgen and Livak, 2008) using TDH1 gene as reference. DNA methylation levels of each mutant used to produce Figure 5B,C were linearly scaled according to the reported averages of RT-qPCR replicates.
DOI: http://dx.doi.org/10.7554/eLife.06205.018

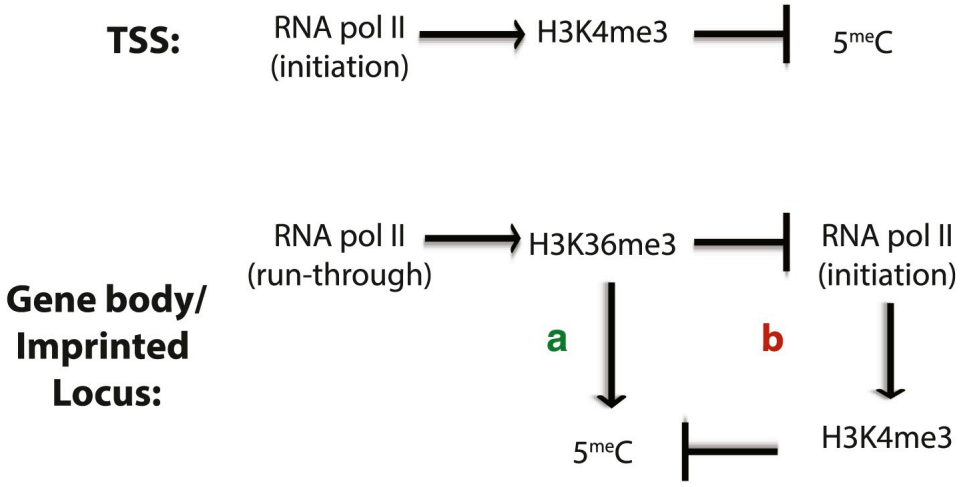


Figure 7—figure supplement 1. Factors affecting DNA methylation deposition. (Top) DNA methylation is negatively affected by H3K4me3 at the TSS. (Bottom) H3K36me3 promotes DNA methylation by (A) direct recruitment of DNMT3b and by (B) preventing the methylation of H3K4, which antagonizes 5meC deposition.
DOI: http://dx.doi.org/10.7554/eLife.06205.021

Supplementary File 1

A: Yeast Whole Genome Bisulfite Sequencing Data

NAME	STRAIN	GROWTH PHASE	READ LENGTH	# READS	MAPPED READS	MAPPABILITY (%)
EV strain 1	W303	stationary phase	50	5618700	4238470	75.44
EV_strain 2	BY4741	stationary phase	50	5693179	4184515	73.50
EV strain 3	W303	stationary phase	100	18656763	13587923	72.83
EV_strain 4	W303	stationary phase	100	18334345	13334549	72.73
3b_exp	W303	exponential growth	100	30969585	20311078	65.58
3b strain 1	W303	stationary phase	100	30168485	19944982	66.11
3b strain 2	W303	stationary phase	100	16815631	12460190	74.10
3b strain 3	W303	stationary phase	100	18976985	14054606	74.06
3b strain 4	W303	stationary phase	100	15158680	11099360	73.22
3b strain 5	W303	stationary phase	50	10538019	8543601	81.07
3b strain 6	W303	stationary phase	50	11206761	9269664	82.71
3b strain 7	BY4741	stationary phase	50	11155426	8762918	78.55
3b strain 8	BY4741	stationary phase	50	8821778	7240328	82.07

B: Yeast MNase Sequencing Stats

NAME	STRAIN	GROWTH PHASE	TYPE	READ LENGTH	# READS	MAPPED READS	MAPPABILITY (%)
EV strain 1	W303	stationary	naked-DNA	50	18421958	17828771	96.78
3b strain 1	W303	stationary	naked-DNA	50	19889834	19215569	96.61
EV strain 1	W303	stationary	MNase-digested chromatin	50	18746212	18061975	96.35
3b strain 1	W303	stationary	MNase-digested chromatin	50	18536100	17957774	96.88

C: Yeast mRNA Sequencing Stats

NAME	STRAIN	GROWTH PHASE	READ LENGTH	# READS	MAPPED READS	MAPPABILITY (%)
EV strain 3	W303	stationary phase	50	17028551	14576036	85.60
EV_strain 4	W303	stationary phase	50	16542825	13793873	83.38
3b strain 2	W303	stationary phase	50	16739897	13835362	82.65
3b strain 3	W303	stationary phase	50	16534790	13435040	81.25
3b strain 4	W303	stationary phase	50	15793246	13140669	83.20

D: Yeast ChIP Sequencing Stats

NAME	SALT (mM)	STRAIN	GROWTH PHASE	READ LENGTH	# READS	MAPPED READS	MAPPABILITY (%)
polII	140	W303	stationary	50	13558635	9428180	69.5
DNMT3b	140	W303	stationary	50	11345100	7548780	66.5
H3K4me1	500	W303	stationary	50	18396633	13745135	74.7
H3K4me3	500	W303	stationary	50	15892023	12360992	77.8
H3K36me3	500	W303	stationary	50	16039999	12993099	81.0
INPUT_1	140	W303	stationary	50	18552692	13700075	73.8
INPUT_2	500	W303	stationary	50	15269649	8392024	55.0

E: Yeast Whole Genome Bisulfite Sequencing Data for mutant strains

NAME	STRAIN	GROWTH PHASE	READ LENGTH	# READS	MAPPED READS	MAPPABILITY (%)
set1Δ replicate 1	BY4741	stationary phase	50	12825798	10723394	83.6
set1Δ replicate 2	BY4741	stationary phase	50	9443269	7989638	84.6
set2Δ replicate 1	BY4741	stationary phase	50	10521217	8621585	81.9
set2Δ replicate 2	BY4741	stationary phase	50	11537314	9252601	80.2
dot1Δ replicate 1	W303	stationary phase	50	11307035	9018367	79.8
dot1Δ replicate 2	W303	stationary phase	50	10711735	8624989	80.5

F: Whole Genome Bisulfite Sequencing in mouse

NAME	TIME	READ LENGTH	# READS	MAPPED READS	MAPPABILITY (%)
E13.5	E13.5	50	244054365	175922769	72.1
E16.5	E16.5	100	1080044130	750630672	69.5
P2.5	P2.5	100	621842708	416532028	67.0

G: ChIP Sequencing Stats in mouse

NAME	TIME	READ LENGTH	# READS	MAPPED READS	MAPPABILITY (%)
E16.5 INPUT	E16.5	50	103447430	37359661	36.1
E16.5 K36me3 IP	E16.5	50	72711515	28013262	38.5

Supplementary File 2

A: Yeast dinucleotide context methylation

NAME	STRAIN	GROWTH PHASE	5^{me}C CONTEXT (METHYLATION PERCENTAGE)				
			all	CpG	CpA	CpT	CpC
EV strain 1	W303	stationary phase	0.19	0.27	0.17	0.19	0.18
EV strain 2	BY4741	stationary phase	0.18	0.26	0.15	0.18	0.18
EV strain 3	W303	stationary phase	0.21	0.25	0.18	0.23	0.22
EV strain 4	W303	stationary phase	0.23	0.26	0.20	0.25	0.24
3b exp	W303	exponential growth	0.81	1.76	0.77	0.55	0.54
3b strain 1	W303	stationary phase	1.51	6.00	0.91	0.56	0.54
3b strain 2	W303	stationary phase	1.21	6.08	0.43	0.28	0.25
3b strain 3	W303	stationary phase	0.94	4.52	0.35	0.26	0.24
3b strain 4	W303	stationary phase	0.86	4.06	0.32	0.26	0.24
3b strain 5	W303	stationary phase	1.47	7.73	0.55	0.29	0.24
3b strain 6	W303	stationary phase	1.44	7.65	0.58	0.30	0.25
3b strain 7	BY4741	stationary phase	0.70	3.27	0.29	0.24	0.23
3b strain 8	BY4741	stationary phase	0.75	3.33	0.38	0.29	0.25

B: Yeast mutant strains dinucleotide context methylation

NAME	STRAIN	GROWTH PHASE	5^{me}C CONTEXT (METHYLATION PERCENTAGE)				
			all	CpG	CpA	CpT	CpC
set1Δ replicate 1	BY4741	stationary phase	0.34	0.84	0.27	0.25	0.23
set1Δ replicate 2	BY4741	stationary phase	0.30	0.78	0.20	0.22	0.21
set2Δ replicate 1	BY4741	stationary phase	0.48	1.90	0.24	0.24	0.23
set2Δ replicate 2	BY4741	stationary phase	0.52	1.97	0.31	0.27	0.25
dot1Δ replicate 1	W303	stationary phase	1.17	5.96	0.44	0.27	0.24
dot1Δ replicate 2	W303	stationary phase	1.20	6.16	0.50	0.29	0.25
EV strain 1	W303	stationary phase	0.19	0.27	0.17	0.19	0.18
EV strain 2	BY4741	stationary phase	0.18	0.26	0.15	0.18	0.18
EV strain 3	W303	stationary phase	0.21	0.25	0.18	0.23	0.22
EV strain 4	W303	stationary phase	0.23	0.26	0.20	0.25	0.24

C: Mouse Germ Cells dinucleotide context methylation

NAME	TIME	5^{me}C CONTEXT (METHYLATION PERCENTAGE)				
		all	CpG	CpA	CpT	CpC
E13.5	E13.5	3.3	6.8	3.3	2.9	3.1
E16.5	E16.5	5.1	56.2	5.8	1.8	0.7
P2.5	P2.5	7.99	77.1	9.9	2.9	0.9

Supplementary File 3: available at
https://elifesciences.org/content/4/e06205/article-data#fig-data-datasets

Supplementary File 4: available at
https://elifesciences.org/content/4/e06205/article-data#fig-data-datasets

Supplementary File 5

Correletion coefficients of DNMT3b occupancy and $5^{me}C$ levels predictions

PREDICTOR(S)			PREDICTED: 5^{me}CPG LEVELS		PREDICTED: DNMT3B OCCUPANCY	
			CORRELATION	ADJ R^2	CORRELATION	ADJ R^2
H3K4me3			0.675381	0.4453	0.4309058	0.1637
H3K36me3			0.3837221	0.1474	0.6271073	0.3927
DNMT3b/$5^{me}C$			0.7000648	0.4627	0.7000648	0.4627
nucleosome			0.0715297	0.003606	0.147471	0.02157
RNApolII			0.004768579	-3.43E-05	0.3148327	0.1119
H3K4me3	H3K36me3		0.7868484	0.6114	0.7713618	0.5749
DNMT3b/$5^{me}C$	H3K36me3		0.7034986	0.4657	0.8002787	0.6194
H3K4me3	H3K36me3	nucleosome	0.786882	0.6114	0.7783655	0.5832
H3K4me3	H3K36me3	RNApolII	0.7873445	0.6134	0.7760999	0.583
H3K4me3	H3K36me3	DNMT3b/$5^{me}C$	0.8215763	0.6635	0.8090604	0.632
all			0.8247778	0.6706	0.8221772	0.6533

Supplementary File 6

A: Plasmids used in this study

PLASMIDS	NAME IN THE PAPER	EXPRESSION OF
pYES2	**EV**	**N/A**
pYES2-DNMT3b	**DNMT3b**	**MmDNMT3b**

B: Yeast strains used in this study

PLASMIDS	GENETIC BACKGROUND	GENOTYPE
W303	W303	MATa, leu2-3,112 trp1-1 can1-100 ura3-1 ade2-1 his3-11,15
BY4741	BY4741	MATa, his3Δ1, leu2Δ0, met15Δ0, ura3Δ0
***set1*Δ (KLY170)**	BY4741	MATa, his3Δ1, leu2Δ0, met15Δ0, ura3Δ0, set1::KAN
***set2*Δ (KLY156)**	BY4741	MATa, his3Δ1, leu2Δ0, met15Δ0, ura3Δ0, set2::HIS3
***dot1*Δ**	W303	MATa, leu2-3,112 trp1-1 can1-100 ura3-1 ade2-1 his3-11,15, dot1::KAN

C: Oligonucleotides used in this study

NAME	TARGET	SEQUENCE	NOTES
3b_Fw	DNMT3b	TAAATATAAA AAGCTT C GGT CCG GCC TCA CGA CAG GAA ACA AT	Used to amplify pCR-BLUNT II-TOPO DNMT3b
3b_Rev	DNMT3b	AATTATTTTA GGATC CGG ACC GTCCCCAGTCCTGGGTAGAAC	
Dot1_UP45	KanMX	CACCAGTAATTGTGCGCTTTGGTTACATTTTGTTGTACAGTAATGATAACTTCGTATAATGTATGC	Used to amplify kanMX cassette with loxP flanking sequences for PCR-based gene disruption
Dot1_DOWN45	KanMX	CTTAGTTATTCATACTCATCGTTAAAAGCCGTTCAAAGTGCCTCATGATAACTTCGTATAGCATAC	
yTDH1_qFw	TDH1	TGCTGCTAAGGCTGTCGGTA	qPCR primers
yTDH1_qRev	TDH1	CAACGGCATCTTCGGTGTAA	
mDNMT3b_qFw	DNMT3b	CTGTGGAGTTTCCGGCTACC	qPCR primers
mDNMT3b_qRev	DNMT3b	TGCTCTCTGCATCCACCTGT	

D: Antibodies used in this study

TARGET	SUPPLIER	CATALOG #	LOT #	USED FOR ChIP (µl)
RNA pol II	Covance	MMS-126R-200	D13HF02305	2.5
H3K4me1	Abcam	ab8895	GR149140-1	2.5
H3K4me3	Active Motif	39159	12613005	3
H3K36me3	Abcam	ab9050	GR114293-1	4
DNMT3b	Abcam	ab2851	GR101720-2	3

References

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biology* **11**:R106. doi: 10.1186/gb-2010-11-10-r106.

Baubec T, Colombo DF, Wirbelauer C, Schmidt J, Burger L, Krebs AR, Akalin A, Schu¨ beler D. 2015. Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. *Nature* **520**:243–247.doi: 10.1038/nature14176.

Brogaard K, Xi L, Wang JP, Widom J. 2012. A map of nucleosome positions in yeast at base-pair resolution. *Nature* **486**:496–501. doi: 10.1038/nature11142.

Chen K, Xi Y, Pan X, Li Z, Kaestner K, Tyler J, Dent S, He X, Li W. 2013. DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Research* **23**:341–351. doi: 10.1101/gr.142067.112.

Chen T, Dent SY. 2014. Chromatin modifiers and remodellers: regulators of cellular differentiation. *Nature Reviews Genetics* **15**:93–106. doi: 10.1038/nrg3607.

Chodavarapu RK, Feng S, Bernatavichute YV, Chen PY, Stroud H, Yu Y, Hetzel JA, Kuo F, Kim J, Cokus SJ, Casero D,Bernal M, Huijser P, Clark AT, Kramer U, Merchant SS, Zhang X, Jacobsen SE, Pellegrini M. 2010. Relationship between nucleosome positioning and DNA methylation. *Nature* **466**:388–392. doi: 10.1038/nature09147.

Chotalia M, Smallwood SA, Ruf N, Dawson C, Lucifero D, Frontera M, James K, Dean W, Kelsey G. 2009.Transcription is required for establishment of germline methylation marks at imprinted genes. *Genes &Development* **23**:105–117. doi: 10.1101/gad.495809.

Collart MA, Oliviero S. 2001. Preparation of yeast RNA. *Current Protocols in Molecular Biology*. Chapter 13,Unit13.12. doi: 10.1002/0471142727.mb1312s23.

Dhayalan A, Rajavelu A, Rathert P, Tamas R, Jurkowska RZ, Ragozin S, Jeltsch A. 2010. The Dnmt3a PWWP domain reads histone 3 lysine 36 trimethylation and guides DNA methylation. *The Journal of Biological Chemistry* **285**:26114–26120. doi: 10.1074/jbc.M109.089433.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**:15–21. doi: 10.1093/bioinformatics/bts635.

Dodge JE, Ramsahoye BH, Wo ZG, Okano M, Li E. 2002. *De novo* methylation of MMLV provirus in embryonic stem cells: CpG versus non-CpG methylation. *Gene* **289**:41–48. doi: 10.1016/S0378-1119(02)00469-9.

Eden A, Gaudet F, Waghmare A, Jaenisch R. 2003. Chromosomal instability and tumors promoted by DNA hypomethylation. *Science* **300**:455. doi: 10.1126/science.1083557.

Feng S, Cokus SJ, Zhang X, Chen PY, Bostick M, Goll MG, Hetzel J, Jain J, Strauss SH, Halpern ME, Ukomadu C,Sadler KC, Pradhan S, Pellegrini M, Jacobsen SE. 2010. Conservation and divergence of methylation patterning in plants and animals. *Proceedings of the National Academy of Sciences of USA* **107**:8689–8694. doi: 10.1073/pnas.1002720107.

Ferrari R, Su T, Li B, Bonora G, Oberai A, Chan Y, Sasidharan R, Berk AJ, Pellegrini M, Kurdistani SK. 2012. Reorganization of the host epigenome by a viral oncogene. *Genome Research* **22**:1212–1221. doi: 10.1101/gr.132308.111.

Gietz RD, Schiestl RH. 2007. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nature Protocols* **2**:31–34. doi: 10.1038/nprot.2007.13.

Guo W, Fiziev P, Yan W, Cokus S, Sun X, Zhang MQ, Chen PY, Pellegrini M. 2013. BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. *BMC Genomics* **14**:774. doi: 10.1186/1471-2164-14-774.

Guo X, Wang L, Li J, Ding Z, Xiao J, Yin X, He S, Shi P, Dong L, Li G, Tian C, Wang J, Cong Y, Xu Y. 2015. Structural insight into auto-inhibition and histone H3-induced activation of DNMT3A. *Nature* **517**:640–644. doi: 10.1038/nature13899.

Hellman A, Chess A. 2007. Gene body-specific methylation on the active X chromosome. *Science* **315**:1141–1143.doi: 10.1126/science.1136352.

Hoffman CS. 2001. Preparation of yeast DNA. *Current Protocols in Molecular Biology*. Chapter 13, Unit13.11.doi: 10.1002/0471142727.mb1311s39.

Hu JL, Zhou BO, Zhang RR, Zhang KL, Zhou JQ, Xu GL. 2009. The N-terminus of histone H3 is required for *de novo* DNA methylation in chromatin. *Proceedings of the National Academy of Sciences of USA* **106**:22187–22192.doi: 10.1073/pnas.0905767106.

Hu M, Sun XJ, Zhang YL, Kuang Y, Hu CQ, Wu WL, Shen SH, Du TT, Li H, He F, Xiao HS, Wang ZG, Liu TX, Lu H,Huang QH, Chen SJ, Chen Z. 2010. Histone H3 lysine 36 methyltransferase Hypb/Setd2 is required for embryonic vascular remodeling. *Proceedings of the National Academy of Sciences of USA* **107**:2956–2961. doi: 10.1073/pnas.0915033107.

Iyer LM, Abhiman S, Aravind L. 2011. Natural history of eukaryotic DNA methylation systems. *Progress in Molecular Biology and Translational Science* 101:25–104. doi: 10.1016/B978-0-12-387685-0.00002-0.

Jin B, Ernst J, Tiedemann RL, Xu H, Sureshchandra S, Kellis M, Dalton S, Liu C, Choi JH, Robertson KD. 2012. Linking DNA methyltransferases to epigenetic marks and nucleosome structure genome-wide in human tumor cells. *Cell Reports* **2**:1411–1424. doi: 10.1016/j.celrep.2012.10.017.

Kitada T, Kuryan BG, Tran NN, Song C, Xue Y, Carey M, Grunstein M. 2012. Mechanism for epigenetic variegation of gene expression at yeast telomeric heterochromatin. Genes & Development **26**:2443–2455. doi: 10.1101/gad.201095.112.

Klug A, Lutter LC. 1981. The helical periodicity of DNA on the nucleosome. *Nucleic Acids Research* **9**:4267–4283.doi: 10.1093/nar/9.17.4267.

Kobayashi H, Sakurai T, Imai M, Takahashi N, Fukuda A, Yayoi O, Sato S, Nakabayashi K, Hata K, SotomaruY, Suzuki Y, Kono T. 2012. Contribution of intragenic DNA methylation in mouse gametic DNA methylomes to establish oocyte-specific heritable marks. *PLOS Genetics* **8**:e1002440. doi: 10.1371/journal.pgen.1002440.

Kobayashi H, Sakurai T, Miura F, Imai M, Mochiduki K, Yanagisawa E, Sakashita A, Wakai T, Suzuki Y, Ito T,Matsui Y, Kono T. 2013. High-resolution DNA methylome analysis of primordial germ cells identifies gender specific reprogramming in mice. *Genome Research* **23**:616–627. doi: 10.1101/gr.148023.112.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**:R25. doi: 10.1186/gb-2009-10-3-r25.

Law JA, Jacobsen SE. 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Reviews Genetics* **11**:204–220. doi: 10.1038/nrg2719.

Lesch BJ, Dokshin GA, Young RA, Mccarrey JR, Page DC. 2013. A set of genes critical to development is epigenetically poised in mouse germ cells from fetal stages through completion of meiosis. *Proceedings of the National Academy of Sciences of USA* **110**:16061–16066. doi: 10.1073/pnas.1315204110.

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, Ecker JR. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**:315–322. doi: 10.1038/nature08514.

Maze I, Noh KM, Soshnev AA, Allis CD. 2014. Every amino acid matters: essential contributions of histone variants to mammalian development and disease. *Nature Reviews Genetics* **15**:259–271. doi: 10.1038/nrg3673.

Nanty L, Carbajosa G, Heap GA, Ratnieks F, van Heel DA, Down TA, Rakyan VK. 2011. Comparative methylomics reveals gene-body H3K36me3 in Drosophila predicts DNA methylation and CpG landscapes in other invertebrates. *Genome Research* **21**:1841–1850. doi: 10.1101/gr.121640.111.

Ng JH, Kumar V, Muratani M, Kraus P, Yeo JC, Yaw LP, Xue K, Lufkin T, Prabhakar S, Ng HH. 2013. In vivo epigenomic profiling of germ cells reveals germ cell molecular signatures. *Developmental Cell* **24**:324–333.doi: 10.1016/j.devcel.2012.12.011.

Ooi SK, Qiu C, Bernstein E, Li K, Jia D, Yang Z, Erdjument-Bromage H, Tempst P, Lin SP, Allis CD, Cheng X, BestorTH. 2007. DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature* **448**:714–717. doi: 10.1038/nature05987.

Pastor WA, Aravind L, Rao A. 2013. TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. *Nature Reviews Molecular Cell Biology* **14**:341–356. doi: 10.1038/nrm3589.

Pastor WA, Stroud H, Nee K, Liu W, Pezic D, Manakov S, Lee SA, Moissiard G, Zamudio N, Bourc'his D, Aravin AA,Clark AT, Jacobsen SE. 2014. MORC1 represses transposable elements in the mouse male germline. *Nature Communications* **5**:5795. doi: 10.1038/ncomms6795.

Peters J. 2014. The role of genomic imprinting in biology and disease: an expanding view. *Nature Reviews Genetics* **15**:517–530. doi: 10.1038/nrg3766.

Popp C, Dean W, Feng S, Cokus SJ, Andrews S, Pellegrini M, Jacobsen SE, Reik W. 2010. Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. *Nature* **463**:1101–1105. doi: 10.1038/nature08829.

Quinodoz S, Guttman M. 2014. Long noncoding RNAs: an emerging link between gene regulation and nuclear organization. *Trends in Cell Biology* 24:651–663. doi: 10.1016/j.tcb.2014.08.009.

Rando OJ. 2010. Genome-wide mapping of nucleosomes in yeast. *Methods in Enzymology* **470**:105–118. doi: 10.1016/S0076-6879(10)70005-7.

Schmittgen TD, Livak KJ. 2008. Analyzing real-time PCR data by the comparative C(T) method. *Nature Protocols* **3**:1101–1108. doi: 10.1038/nprot.2008.73.

Seisenberger S, Andrews S, Krueger F, Arand J, Walter J, Santos F, Popp C, Thienpont B, Dean W, Reik W. 2012.The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells. *Molecular Cell* **48**:849–862. doi: 10.1016/j.molcel.2012.11.001.

Seisenberger S, Peat JR, Hore TA, Santos F, Dean W, Reik W. 2013. Reprogramming DNA methylation in the mammalian life cycle: building and breaking epigenetic barriers. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences* **368**:20110330. doi: 10.1098/rstb.2011.0330.

Singh P, Li AX, Tran DA, Oates N, Kang ER, Wu X, Szabo´ PE. 2013. De novo DNA methylation in the male germ line occurs by default but is excluded at sites of H3K4 methylation. *Cell Reports* **4**:205–219. doi: 10.1016/j.celrep.2013.06.004.

Smallwood SA, Tomizawa S, Krueger F, Ruf N, Carli N, Segonds-Pichon A, Sato S, Hata K, Andrews SR, Kelsey G.2011. Dynamic CpG island methylation landscape in oocytes and pre-implantation embryos. *Nature Genetics* **43**:811–814. doi: 10.1038/ng.864.

Smith ZD, Meissner A. 2013. DNA methylation: roles in mammalian development. *Nature Reviews Genetics* **14**:204–220. doi: 10.1038/nrg3354.

Vermeulen M, Eberl HC, Matarese F, Marks H, Denissov S, Butter F, Lee KK, Olsen JV, Hyman AA, Stunnenberg HG,Mann M. 2010. Quantitative interaction proteomics and genome-wide profiling of epigenetic histone marks and their readers. *Cell* **142**:967–980. doi: 10.1016/j.cell.2010.08.020.

Vincent JJ, Li Z, Lee SA, Liu X, Etter MO, Diaz-Perez SV, Taylor SK, Gkountela S, Lindgren AG, Clark AT. 2011.Single cell analysis facilitates staging of Blimp1-dependent primordial germ cells derived from mouse embryonic stem cells. *PLOS ONE* **6**:e28960. doi: 10.1371/journal.pone.0028960.

Wach A. 1996. PCR-synthesis of marker cassettes with long flanking homology regions for gene disruptions in *S. cerevisiae*. *Yeast* **12**:259–265. doi: 10.1002/(SICI)1097-0061(19960315)12:3<259::AID-YEA901>3.0.CO;2-C.

Wagner EJ, Carpenter PB. 2012. Understanding the language of Lys36 methylation at histone H3. *Nature Reviews Molecular Cell Biology* **13**:115–126. doi: 10.1038/nrm3274.

Weaver JR, Bartolomei MS. 2014. Chromatin regulators of genomic imprinting. *Biochimica Et Biophysica Acta* **1839**:169–177. doi: 10.1016/j.bbagrm.2013.12.002.

Wernig M, Meissner A, Foreman R, Brambrink T, Ku M, Hochedlinger K, Bernstein BE, Jaenisch R. 2007. In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. *Nature* **448**:318–324. doi: 10.1038/nature05944.

You JS, Jones PA. 2012. Cancer genetics and epigenetics: two sides of the same coin? *Cancer Cell* **22**:9–20.doi: 10.1016/j.ccr.2012.06.008.

Zemach A, Grafi G. 2003. Characterization of Arabidopsis thaliana methyl-CpG-binding domain (MBD) proteins. *The Plant journal : for cell and molecular biology* **34**:565–572. doi: 10.1046/j.1365-313X.2003.01756.x.

Zhang Y, Jurkowska R, Soeroes S, Rajavelu A, Dhayalan A, Bock I, Rathert P, Brandt O, Reinhardt R, Fischle W, Jeltsch A. 2010. Chromatin methylation activity of Dnmt3a and Dnmt3a/3L is guided by interaction of the ADD domain with the histone H3 tail. *Nucleic Acids Research* **38**:4246–4253. doi: 10.1093/nar/gkq147.

Zhang Y, Xie S, Zhou Y, Xie Y, Liu P, Sun M, Xiao H, Jin Y, Sun X, Chen Z, Huang Q, Chen S. 2014. H3K36 histone methyltransferase Setd2 is required for murine embryonic stem cell differentiation toward endoderm. *Cell Reports* **8**:1989–2002. doi: 10.1016/j.celrep.2014.08.031.

Zhu X, He F, Zeng H, Ling S, Chen A, Wang Y, Yan X,Wei W, Pang Y, Cheng H, Hua C, Zhang Y, Yang X, Lu X,Cao L, Hao L, Dong L, Zou W, Wu J, Li X, Zheng S, Yan J, Zhou J, Zhang L, Mi S, Wang X, Zhang L, Zou Y,Chen Y, Geng Z, Wang J, Zhou J, Liu X, Wang J, Yuan W, Huang G, Cheng T, Wang QF. 2014. Identification of functional cooperative mutations of SETD2 in human acute leukemia. *Nature Genetics* **46**:287–293.doi: 10.1038/ng.2894.

CHAPTER 4:

Bisulfite RNA-seq: Detection and analysis of 5-methyl cytosine in polyA-RNA with next generation sequencing

Bisulfite RNA-seq: Detection and analysis of 5-methyl cytosine in polyA-RNA with next generation sequencing

Kianoush Sadre-Bazzaz, Liudmilla Rubbi, Marco Morselli, Larry Lam, Arshad H. Khan, Desmond J. Smith, Matteo Pellegrini

ABSTRACT

Methylation at the 5' carbon of cytosine in RNA is an epigenetic mark prevalent in all of life. Though well studied in non-coding RNA, methylation of messenger RNA (mRNA) is less understood. We report a method to quantify 5-methylcytosine in polyA(+)-RNA extracts from mouse hypothalamus and human stem cells using next-generation sequencing. Following bisulfite treatment, and library preparation, we sequenced hypothalamus samples from two mouse strains (C57 and DBA) and human stem cells on the Illumina platform. A pipeline was developed to analyze data using BS-seeker2 software. Unlike DNA, where methylation is enriched at CpG, we found that methylcytosines in polyA-RNA are equally represented in CG and CH (H=ATC) contexts. We report approximately 6000 uniquely methylated sites in our samples, with a correlation of 0.89 (p-value: 2.2e10^-16) for sites shared in samples from the same mouse strains, of which 200 were methylated across all samples. Among these conserved sites, a small set showed strain-specific methylation that was correlated with the expression of RNA methyltransferase enzymes (Nsun7 and Dnmt2) in

the hypothalamus. Finally, we performed an analysis of the RNA secondary structure of fragments proximal to methylcytosines and found that these are enriched for low free energy regions with secondary structure.

INTRODUCTION

Methylation at the 5' carbon of cytosine (5mC) is common in RNA and is found across a diverse range of organisms (1). In contrast to N6-methyladenosine (m6A), less is understood about the biological significance of 5mC in RNA (2). Numerous RNA-methyltransferases (RMTs) have been identified, and deletions or mutations of these enzymes have been implicated in developmental defects, mental retardation, and cancer (3). However, the RMT(s) responsible for 5mC in RNA are incompletely characterized.

Bisulfite RNA sequencing (bsRNAseq) is the primary approach used to detect 5mC in RNA. The method has been successfully applied to HeLa cell RNA extracts which were sequenced on the SOLiDTM platform, revealing over 10,000 mRNA and non-coding RNA methylation sites (4). Methylation of tRNA has also been analyzed with bsRNAseq, and in mice it was shown that 5mC is associated with tRNA stability (5).

Recently, a number of studies have begun to elucidate the role of RNA cytosine methylation in biological processes. For example, enhancer-RNA methylation by the RMT NSun7 can increase transcription of PGC-1α-regulated genes with metabolic consequences for the cell (6). PGC-1α transcription is increased in certain melanoma cell lines and correlates with MITF, a transcription factor that regulates melanin production in response to ultraviolet (UV) light (7). Another RMT, NSun2, is a target of the Myc transcription factor and is up-regulated in breast cancer (8). Methylation of mRNAs by Nsun2 may prolong transcript half-life (9). This finding is

consistent with a recent study showing that a lack of vault-RNA methylation by Nsun2 resulted in elevated processing of substrates into microRNAs (10).

To further characterize the targeting of cytosine methylation in mRNA we have developed a profiling technique that uses the Illumina sequencing platform. This approach relies on the bisulfite treatment of polyA RNA to yield transcriptome-wide views of the polyA RNA methylome. We have developed an informatics pipeline for processing the data that builds on previous methods for analyzing 5mC in DNA, in particular the BSSeeker2 program. Using these experimental and informatic strategies we have identified hundreds of sites consistently methylated in mRNA across independent samples from mouse brain tissues and human stem cells.

RESULTS

Generation of bisulfite mRNA libraries

To leverage the efficiency of the latest generation of DNA sequencers, we have developed a protocol to measure transcriptome wide levels of methylcytosine in mRNA using the Illumina platform. Bisulfite treatment of mRNA converts cytosine residues to uracil, leaving 5-methylcytosine residues unaffected. Thus, after PCR amplification, unmethylated cytosines are read as thymine ("T"), while methylated cytosines are protected from bisulfite and read as cytosine ("C"). This bisulfite sequencing method provides single-nucleotide resolution information on the 5mC status of RNA.

The procedure for library preparation is described in detail in the Methods section and shown schematically in Figure 4.1. In short, RNA is enriched for polyA-containing transcripts and then treated with sodium bisulfite. Following RNA fragmentation and adaptor ligation, DNA is synthesized from these templates by PCR and sequenced on the Illumina HiSeq2500.

The RNA used in our study was extracted from the hypothalamus of two mouse strains, C57BL/6J (C57) and DBA/2J (DBA). A total of six samples: two from C57 and four from DBA, were prepared for sequencing and analysis. We also generated a library from a human embryonic stem cell line.

Methylation levels observed in the hypothalamus

To measure DNA methylation profiles reads were aligned end-to-end and only those with an alignment length > 30, and the sense strand were used for calculation of methylation. Between 12-25 million reads were uniquely aligned to the transcriptome with a mappability of ~50% for each dataset (Table 4.1). This level of mappability is similar to what we usually observe for DNA methylation libraries. The reason for restricting our analysis to reads that align to the forward direction of transcription is that our protocol is stranded, and predominantly generates reads for the forward strand, with little signal on the reverse strand. Moreover, reverse reads are potentially due to DNA contamination or antisense transcription and do not necessarily reflect the methylation status of the mRNA.

To test the accuracy of our pipeline we analyzed the methylation status of a positive and negative control. As a negative control, a sample of in vitro transcribed RNA from a kanMX containing plasmid was spiked in one of the DBA samples. Since the plasmid was transcribed in vitro with only unmethylated ribonucleotides, we expect the methylation level to reflect the background rates of methylation in our assay. Reassuringly, we found that kanMX cytosines are mostly (~99.8%) converted to uracil by the bisulfite treatment. Moreover, we were able to estimate these conversion rates with high accuracy as the average coverage of the transcript was > 4000 per base (Figure 4.2a).

As a positive control we looked at the methylation of ribosomal RNA. Although our library used poly(dT) enrichment to select for mRNA and poly(A)+ non-coding RNA (ncRNA), we also obtained some reads that mapped to rRNA, very likely due to its high abundance. It is well

established that rRNA is significantly methylated (15). We therefore examined the methylation of rRNA in our libraries, each aligned against a reference containing a single copy of mouse ribosomal and transfer RNA genes. For each sample we observed multiple highly methylated sites in rRNA, supporting the notion that our approach is able to identify methylated cytosines in RNA (Figure 4.2b).

Having identified both positive and negative controls, we next turned our attention to the methylation of cytosines in poly(A)+ RNA. We observe numerous methylated cytosines across poly(A)+ RNA, with two representatives in exonic and UTR regions shown in Figures 4.2c and 4.2d. When analyzing DNA methylation we typically separate cytosines into two groups, depending on whether they are followed by guanine (CpG) or not followed by a guanine (CpH). This is because mammalian DNA methyltransferases preferentially act on CpG dinucleotides. We found that unlike mammalian DNA, which is heavily methylated at CpG sites, but mostly unmethylated at CpH sites, RNA methylation patterns of CG and CH are quite similar (Figure 4.3).

To identify significantly methylated cytosines, we counted the number of altered (T) and unaltered (C) cytosines aligned to each cytosine within our transcripts, and computed the methylation ratio as the fraction of C/(C + T) at that base. We used a test based on the binomial distribution to identify significantly methylated sites, and set the background rate to the average methylation of the kanMX unmethylated control (~ 0.2%). In addition, we used the Benjamini-Hochberg approach to correct for multiple testing, and used a false discovery rate of 1% as our threshold. Finally, we also required that each methylation call was supported by at least 5 cytosines. This approach yielded between 500-1,000 sites that were significantly methylated in each sample, and a total of ~6,000 unique sites across all six samples. In a pair-wise comparison, about 30-40% of these sites were observed in common between DBA mice, or DBA and C57 mice (Figure 4.4).

The methylation levels of common sites were significantly correlated between

individuals from the same (Figure 4.5a), or different strains of mice (DBA2/3 correlation is 0.89, p-value: < 2.2e-16, and DBA3/C57-1 correlation is 0.71, p-value < 2.2e-16, Figure 4.5b). To observe the patterns of methylated sites across the different mice, we identified the significant sites in common from all six datasets and plotted their methylation levels in a clustered heat map (Figure 4.6). The plot contained about 200 of the 6,000 unique sites (Supplemental Figure 4.1). We find ~10 sites heavily methylated (> 60%) across all of our mice in both strains (Figure 4.6, arrow). In addition, there are sites that are differentially methylated between C57 and DBA strains (Figure 4.6, brackets). The distribution of methylation levels showed strain specific effects across sites that are significantly methylated in all samples. Specifically, C57 mice tend to have more highly methylated sites (above 50% methylation, Figure 4.7) and a bimodal distribution of methylation intensity, which is not observed in DBA mice. Using the DAVID functional annotation tool (16) we found that the transcripts that contain significantly methylated sites across all mice are enriched in neuronal maintenance and cell sorting pathways (Table 4.2). We asked if RMT expression in the hypothalamus of C57 and DBA mice differs, and might be associated with the methylation of these transcripts. Using a hypothalamic transcriptome (11) reported for 99 mouse strains, the RMT genes Nsun2 and Wbscr22 were found to have the highest expression across both mice strains. Moreover, Nsun7 (p-value 0.0002) and Dnmt2 (p-value 0.05) found to be differentially expressed between these strains (Supplementary Table 4.1) suggesting a possible mechanism for the methylation profile observed for the strain specific methylated mRNAs. Nsun7, in particular has higher average FPKM in C57 compared to DBA and correlates with higher methylation levels observed in that strain (Figure 4. 7).

Properties of Methylated Cytosines

To determine if methylation of polyA-RNA might have positional specificity, we calculated the fractional position of all methylated sites from each dataset. We find that the methylation is

enriched toward the 3' ends of the transcripts (Figure 4.8). However, we did not observe a consensus sequence in methylated cytosines (Supplementary Figure 4.2). To test if methylated sites are part of regions with secondary structure, we calculated the free energy of folding of regions surrounding methylated cytosines. The sequence of 50 bases flanking each conserved methylated site (~200) was extracted, and its free energy was calculated using the mfold server (17). As a negative control, 200 randomly generated sequences of the same length from the mouse transcriptome were also computed. The distribution of these free energy values suggests that the two groups originate from different distributions, with methylated RNAs forming more stable structures (Figure 9a, Kolmogorov-Smirnov p-value: 5.5e-05). Some of these regions are heavily methylated (above 70%), at numerous positions (up to ~70 sites), which occur at both single and double stranded regions (Figure 4.10). We also tested if sites of methylation in poly-A RNA might overlap with microRNA (miRNA) binding sites. For this we queried the uniquely methylated sites against the mouse databank for miRNA binding sites (18). Compared to a set of randomly selected transcriptome coordinates, methylated sites have a higher proportion of at least one miRNA binding site (p-value: 2.2e-16, Supplementary Figure 4.3). It is possible that our transcripts are overly represented in the miRNA database, nevertheless, it is tempting to speculate that methylation might stabilize miRNA binding.

Human polyA-RNA methylation and structure

We prepared a human polyA-RNA library from embryonic stem cells followed by bisulfite RNA sequencing. Using the human hg19 transcriptome as reference, our pipeline identified about 500 significantly methylated sites. To determine if methylation is influenced by unique RNA secondary structure, we analyzed our data against Parallel Analysis of RNA Structure (PARS) measurements (19). The PARS value for each base is an estimate of whether that base resides in single or double stranded RNA.

The RNA is cleaved with V1 nuclease that cuts 3' end of double stranded RNA. Following library preparation of the digested molecules and high throughput sequencing, the base at the 5' end of each aligned read represents the (nth + 1) site of RNA cleavage. Together with S1 nuclease that cuts the 3' end of single-stranded RNA, a transcriptome picture of RNA secondary structure is generated.

When we compared the PARS values for our methylated sites to a random set from the hg19 transcriptome, we observed that the two sets were differentially distributed (Figure 9b, Kolmogorov-Smirnov p-value: 2.2e-06). Interestingly, about 54% of methylated sites were associated with PARS > 0, compared to ~30% of sites with PARS < 0, suggesting that that this modification is targeted to double stranded RNA.

DISCUSSION

We developed an informatics pipeline to characterize polyA-RNA methylation using bisulfite sequencing. Our method takes advantage of the Illumina DNA sequencing platform. Our pipeline builds on the BS-Seeker2 software, a commonly used tool for the analysis of DNA methylation. Using bisulfite-treated RNA from mouse hypothalamus, we were able to detect methyation of ribosomal RNA molecules to confirm the efficacy of our method. Subsequently, ~6000 uniquely methylated sites were detected in non-ribosomal transcripts. The methylation measurements were reproducible across individual mice, with a correlation of ~0.9 between mice of the same strain. We find that methylated transcripts are enriched in pathways involved in neuronal maintenance and cell sorting (Table 4.2). Notably we observe strain-specific differences in the methylation percentage and distributions of RNA sites (Figures 4.6 and 4.7), which correspond to expression differences of Nsun7 and Dnmt2 RMTs (Supplementary Table 4.1 and (11)).

Most RMTs in higher eukaryotes are localized in the nucleus (1). One of the differentially expressed enzymes from our analysis, Dnmt2, is similar to DNA methyltransferases with a single conserved cysteine in motif IV in contrast to two cysteines found in motifs IV and VI of other RMTs. Moreover, according to a prior study, Dnmt2 is the only RMT associated with the cytoplasmic compartment (20).

Our analysis of methylated sites reveals that they preferentially occur in hairpin loops or double stranded RNA (Figure 4.10). A study using molecular dynamics simulation of a bacterial RNA methyltransferase (Fmu) suggested that 16S rRNA binds in a folded state with target cytosines in close proximity to the enzyme's active site (21). We observed that the sites with significant methylation are energetically more stable, and structured (Figure 4.9), when compared to a population of randomly selected sequences from the transcriptome. Despite these similarities, it is possible that mRNA substrates will have different structured conformations than rRNA. Moreover, different RMTs may have diverse substrate requirements (2). For example the bacterial RMT gene (YebU) encodes a C-terminal RNA binding domain (22) for substrate recruitment, while a mitochondrial orthologue (Nsun4) requires a cofactor protein for this purpose (23).

Additional cell biological studies, including knock-down or over-expression combined with methylation profiling will help reveal enzymatic requirements for mRNA methylation in the future. Our method can be robustly applied to various cell types and also in genetic mapping experiments to determine epigenetic mechanisms affecting RNA.

MATERIALS AND METHODS

Bisulfite RNA library preparation

Samples were extracted from brains of males from two mouse strains C57BL/6J (C57) and DBA/2J (DBA) fed a high fat/high sugar diet, (11), and from human embryonic stem cells. Total RNA (2-3 μg) was treated with DNaseI and enriched for polyA-containing transcripts using oligo-dT beads (Illumina TruSeq mRNA kit). The RNA was then treated with bisulfite (EZ RNA Methylation kit, Zymo Research) according to the manufacturer's instructions. The converted RNA was fragmented to an average of 150 nt using an RNA-fragmentation buffer (NEB) for 3 minutes at 94°C. Fragmented RNA was then purified using RNA clean and concentrator-5 (Zymo Research). The RNA was treated with T4 PNK (NEB) for 30 minutes at 37°C, and purified again using RNA clean and concentrator-5. The 3'-Adapter (Illumina TruSeq Small RNA kit) was ligated to the RNA using T4 RNA ligase 2 Truncated KQ (NEB cat# M0373S) according to the manufacturer's instruction. The 5'-Adapter ligation and reverse transcription were performed according to the Illumina TruSeq Small RNA kit instructions. First strand products were amplified through 20 cycles of PCR, following TruSeq small RNA kit conditions. The final libraries were purified using AMPure XP beads with a 1:1.2 ratio (DNA:AMPure XP beads).

Alignment of bisulfite mRNA reads

To analyze the 5mC RNA data, we modified an existing pipeline for processing DNA bisulfite sequencing data. Reads were mapped using the BS-seeker2 software specifically designed for bisulfite sequencing (12). Following demultiplexing, the six mouse datasets were aligned against a mouse transcriptome as reference. This reference was generated from the Ensembl database by selecting cDNA sequences for the mouse genes (version GRCm38.p4). The FASTA-formatted file was filtered so only one transcript represents each gene. For this purpose we selected the

longest fragment that encompassed all the exons. The human reference FASTA file was assembled from hg19 transcriptome as prepared by Wan, Y. et al. (19).

Mapping and analysis

The coordinates for significantly methylated RNA were obtained by applying the binomial distribution test (discussed below) on the entire map obtained from BS-seeker2. The methylated positions were used to navigate the bam-formatted alignments for visualization with the IG-Viewer software (13). Venn and Logos plots were prepared using the bioinformatics servers Venn Diagram from Ghent University and WebLogo (14). Statistical analysis and figures were prepared using the programming language R and packages: vioplot, gplot, and ggpot2.

ACKNOWLEDGMENTS

We thank Dr. Atsushi Nakano for providing RNA samples from human embryonic stem cells. The human FASTA-formatted reference file was kindly provided by Dr. Howard Chang from Stanford University.

FUNDING

Supported by UCLA Department of Energy, Institute for Genomics and Proteomics and grants: NIH/NIGMS R01 GM098273 and R21 HG007405.

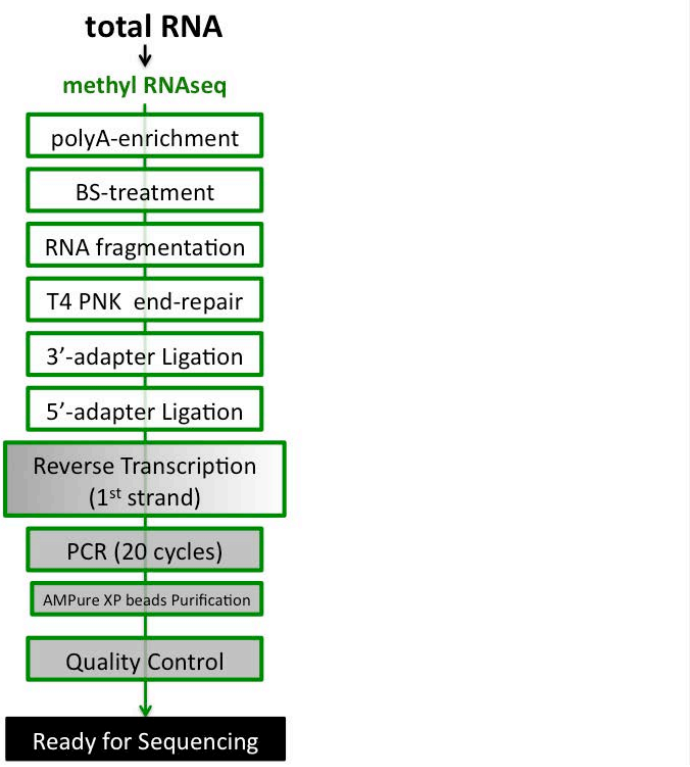


Figure 4.1: Schematic of polyA-enriched RNA-BS seq workflow. The reagents used are from the Illumina TruSeq mRNA v2 kit, Illumina small RNA TruSeq kit, NEB and Zymo Research.

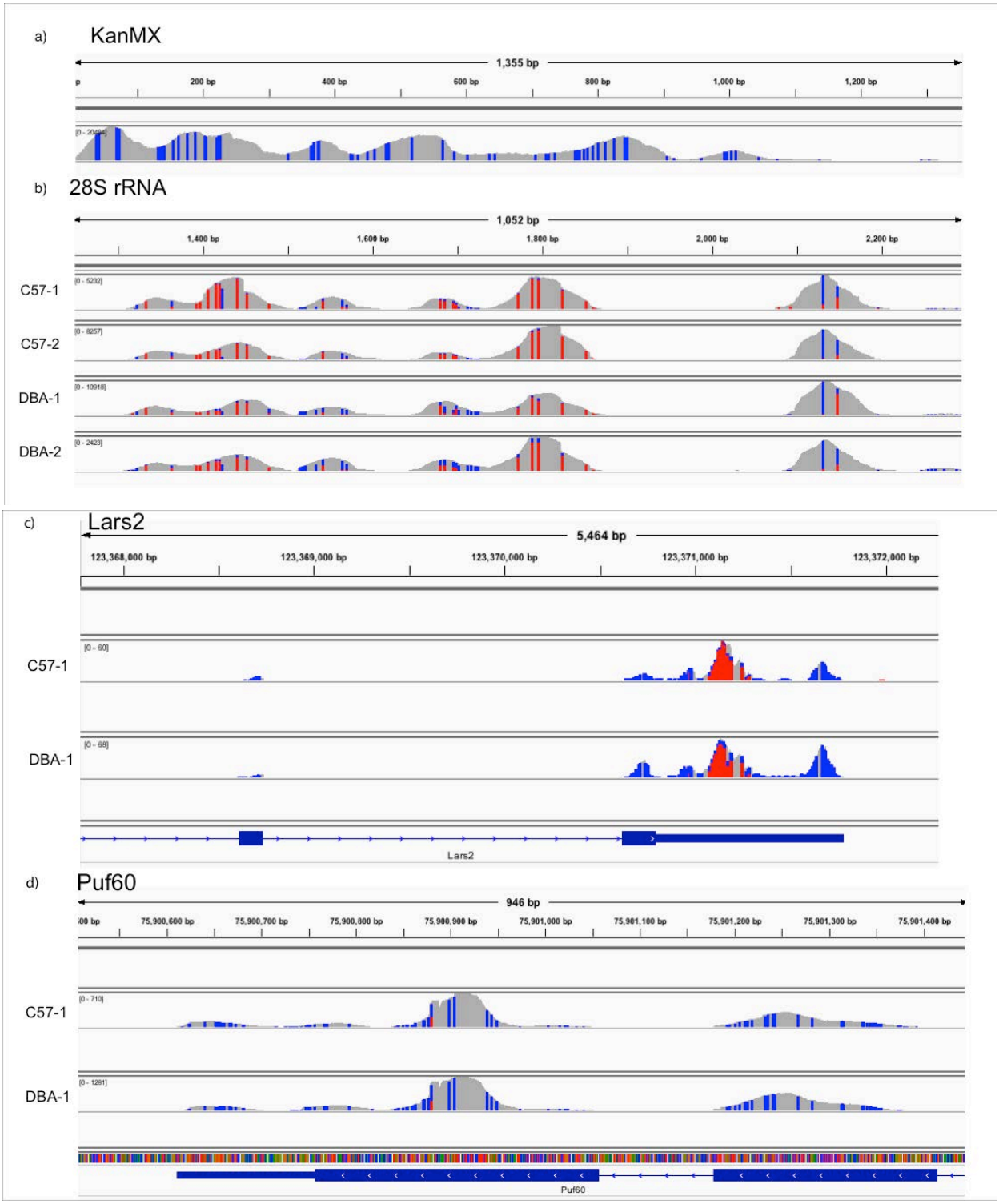


Figure 4.2: Coverage and methylation profile for (a) an un-methylated kanMX control, and (b) a section of the 28S ribosomal RNA across four independent samples. Methylated-CpG sites are shown in red and un-methylated in blue. Methylation profiles of Lars2 3'-UTR (c) and a section of Puf60 (d)**.**

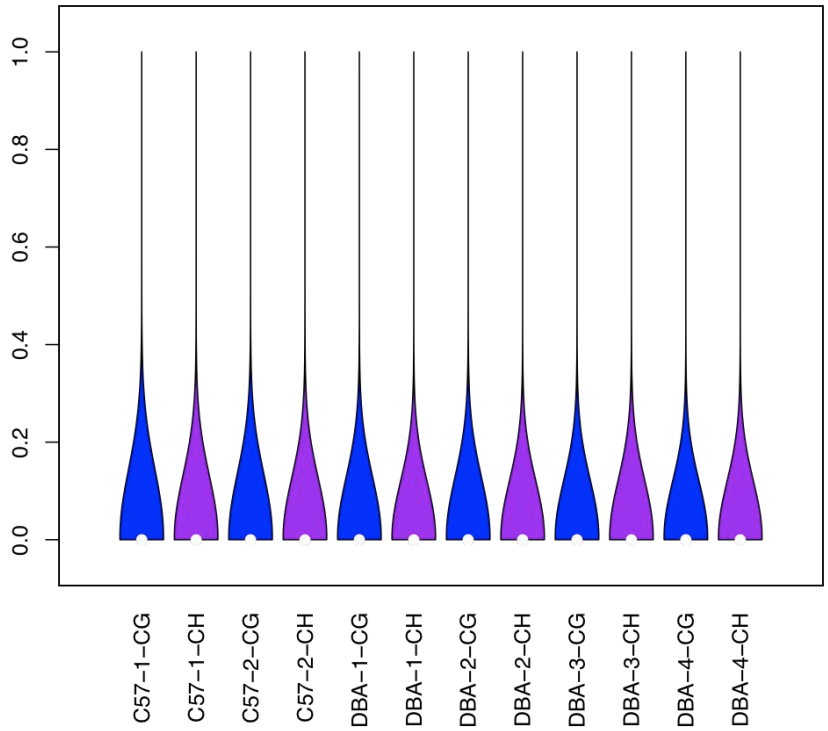


Figure 4.3: Global RNA cytosine methylation levels of six mouse samples in CG and CH contexts**.**

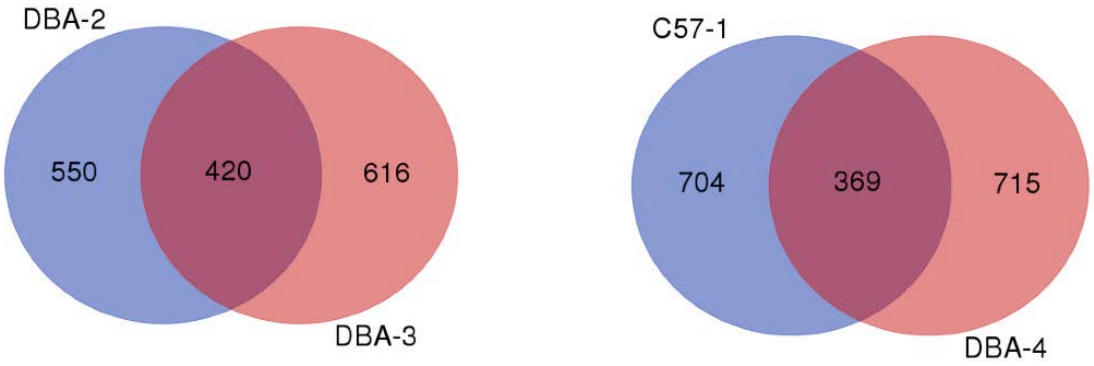


Figure 4.4: Common methylation sites between two DBA, or C57 and DBA mice. Each colored circle represents a different individual mouse**.**

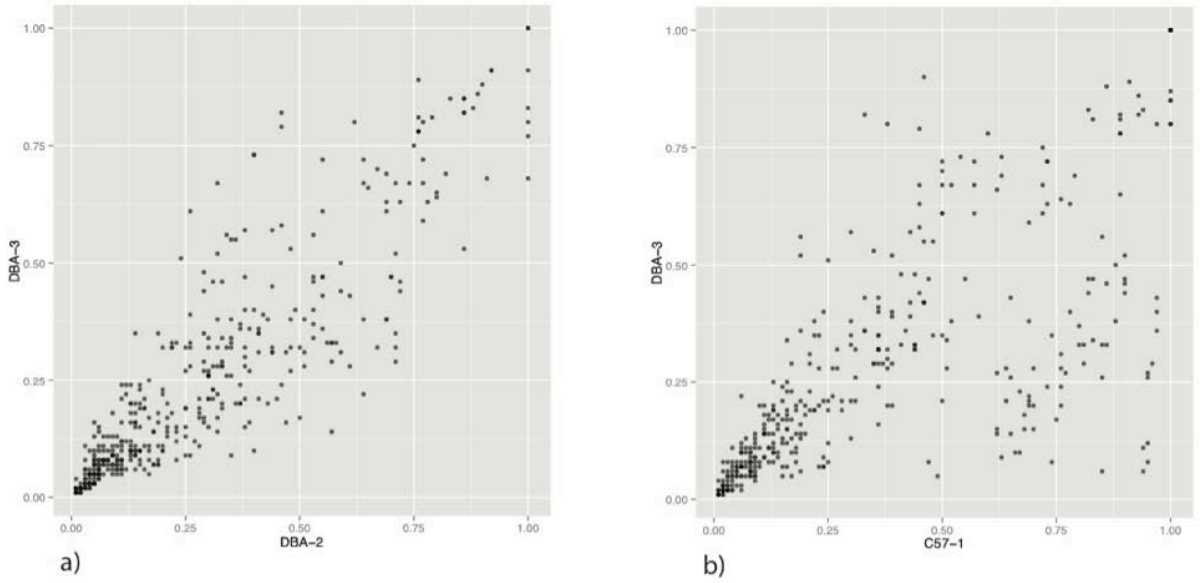


Figure 4.5: Significantly methylated common mRNA sites observed between (a) the same strain, DBA, or (b) different strains, C57 vs. DBA mice**.**

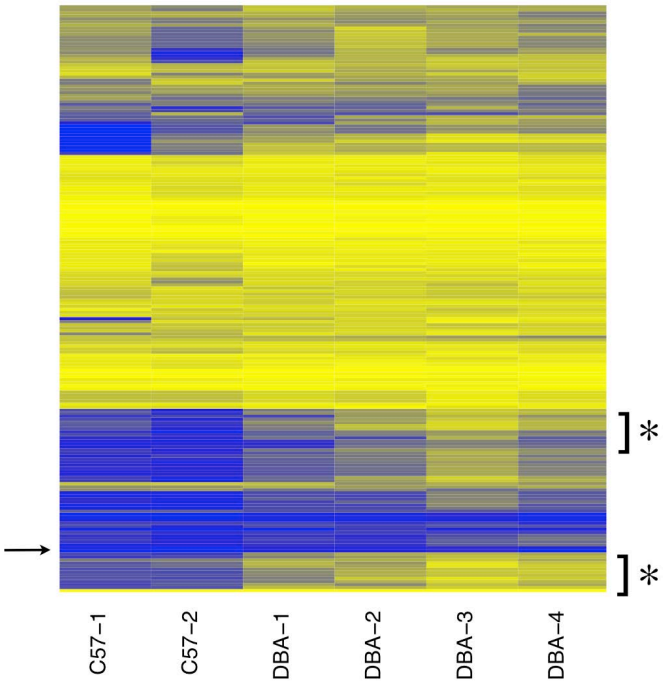


Figure 4.6: RNA methylation of common sites across all samples. Each row is a different site across the genome. Highly methylated sites (arrow) and differentially methylated sites (*) are shown. Highly methylated sites are blue, while lowly methylated sites are colored in yellow.

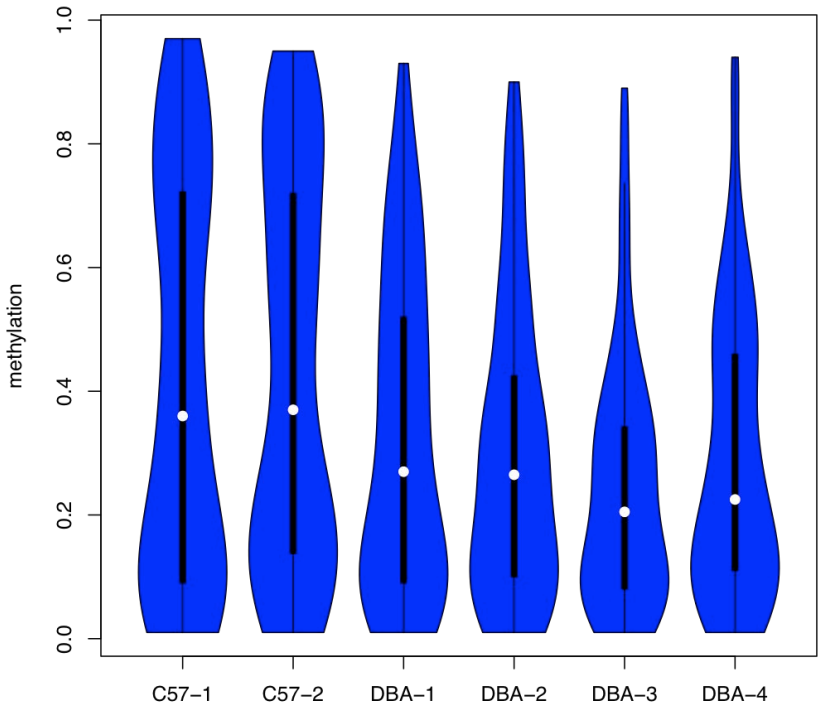


Figure 4.7: Distribution of methylation levels for all significant sites determined by Benjamini-Hochberg test at 1% FDR.

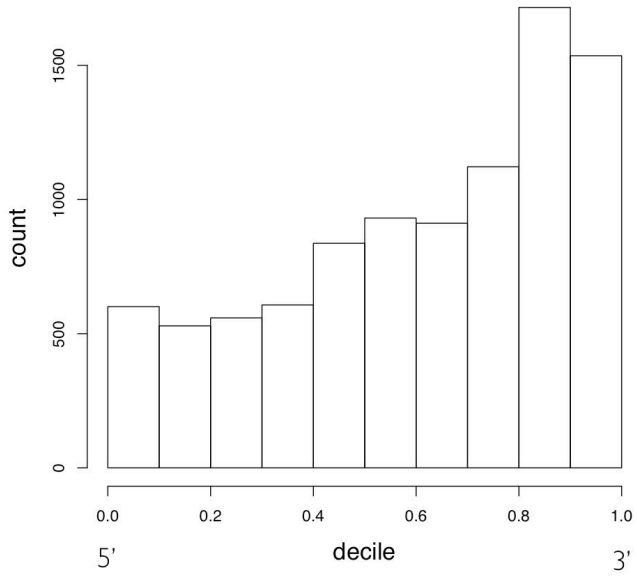


Figure 4.8: Metagene plot showing the distribution of all significant methylated sites determined by Benjamini-Hochberg test at 1% FDR.

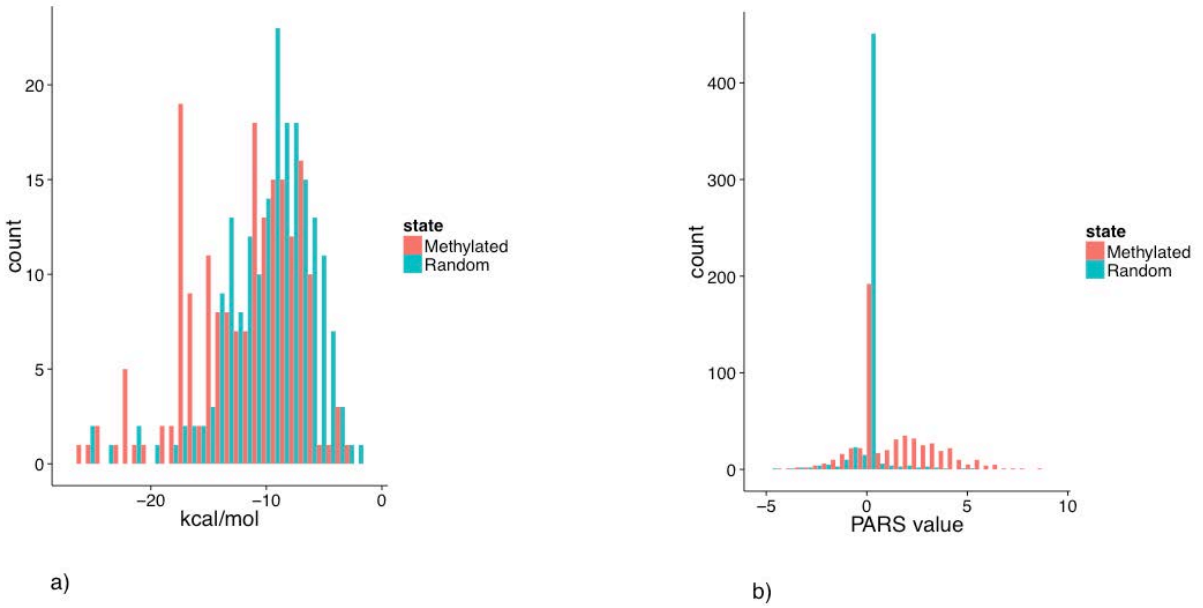


Figure 4.9: a) Free energy of RNA folding of the ~200 methylated sites flanked by 50 bases (red), compared to a sample of 200 random RNA sequences (blue). b) PARS score distribution of human methylated sites.

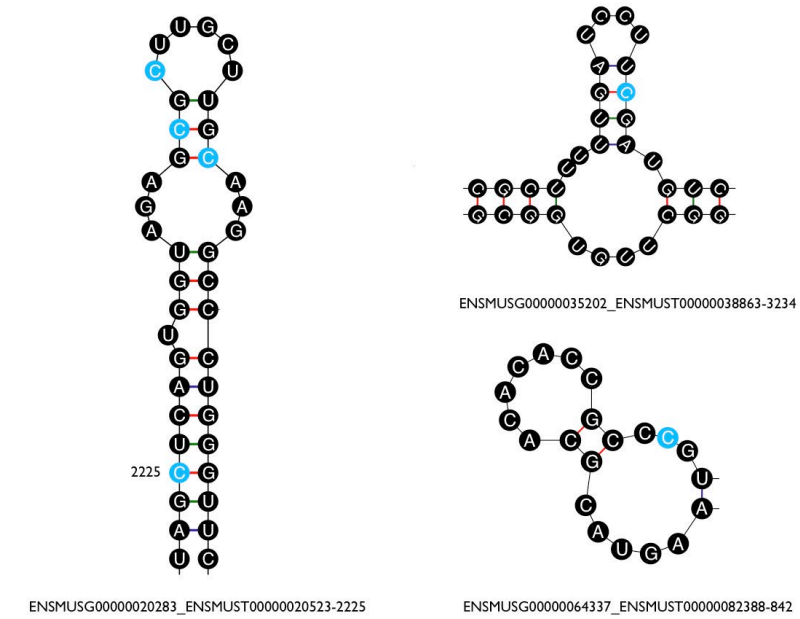
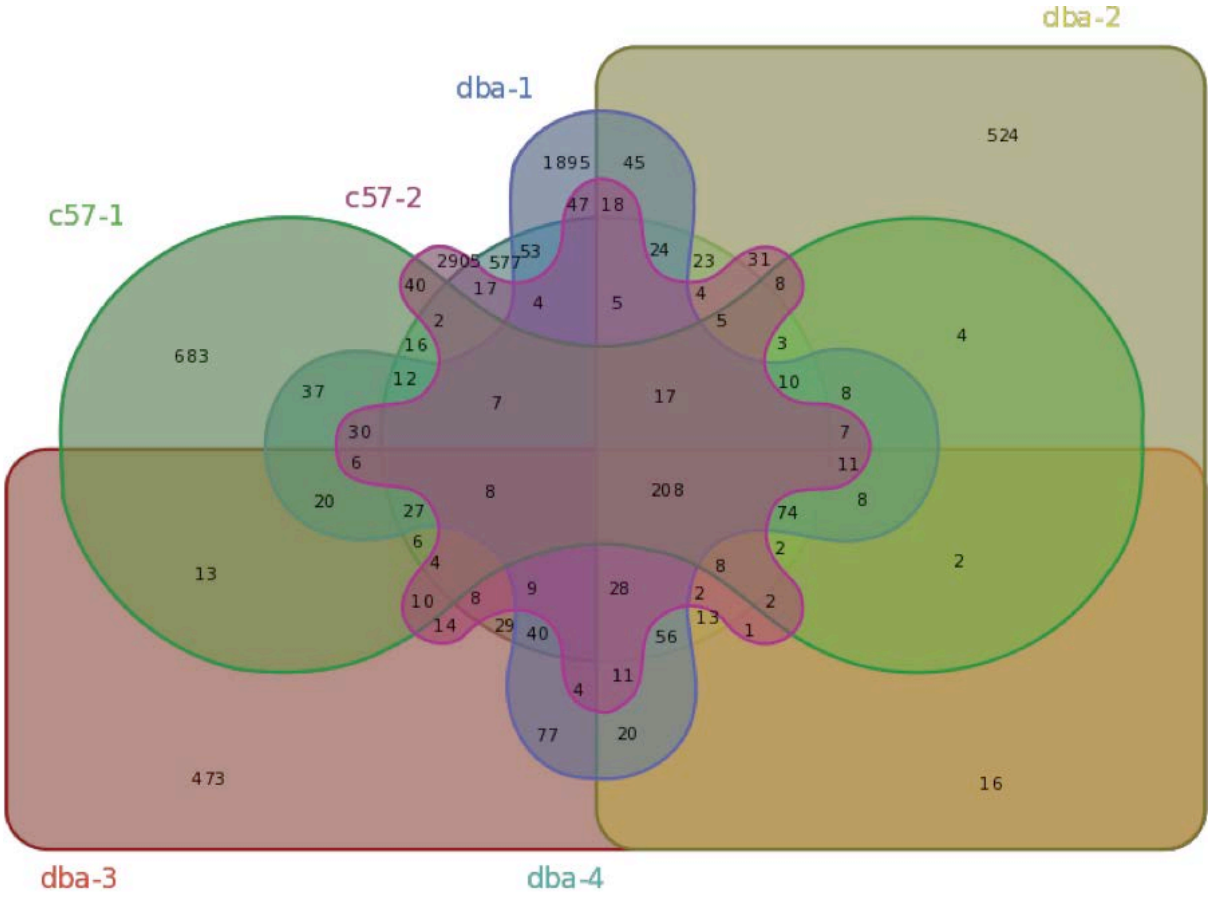
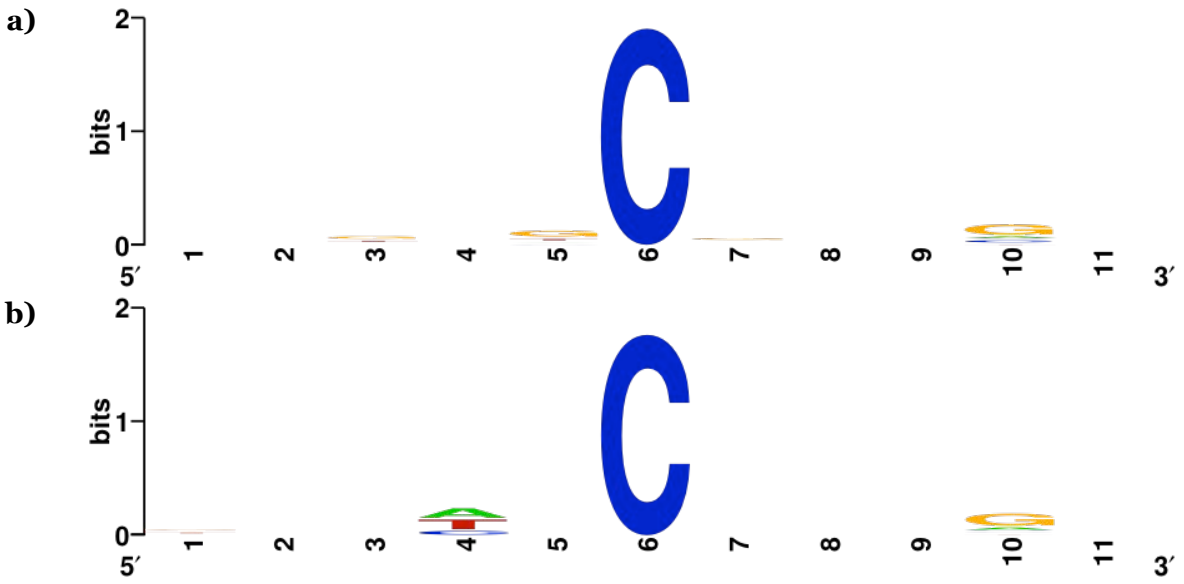


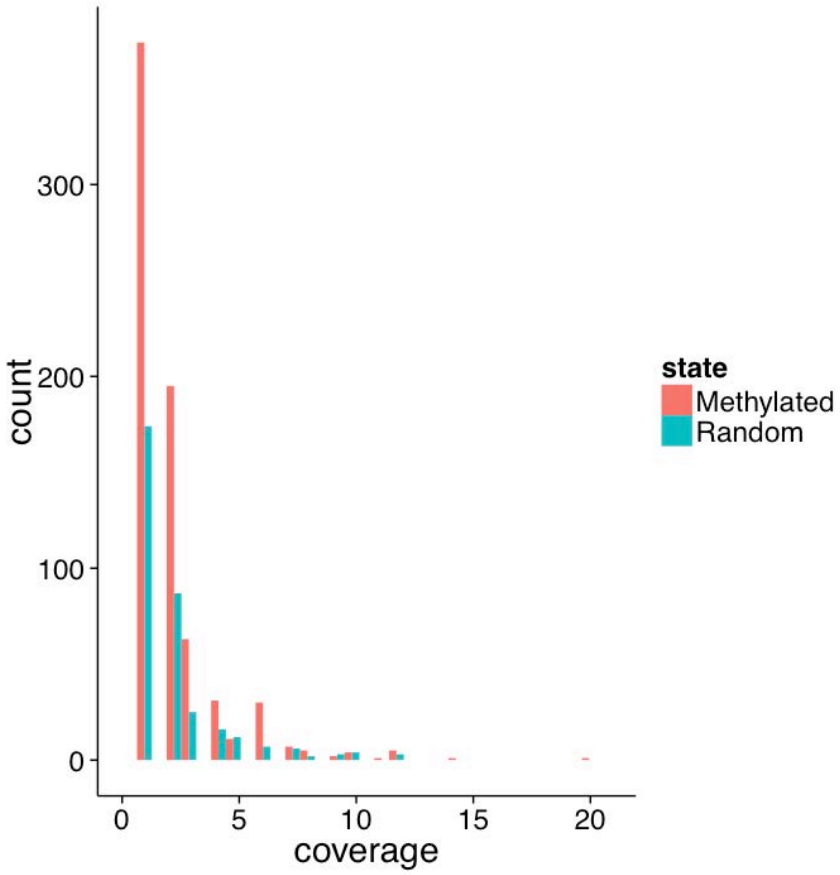
Figure 4.10: Methylated sites (blue cytosines) in selected transcript RNAs. Label format: ensemble gene ID-ensemble transcript ID-site number.



Supplementary Figure 4.1: Venn diagram showing overlap of significantly methylated sites across six (C57 and DBA) mice samples.



Supplementary Figure 4.2: Consensus sequence of five bases flanking highly methylated cytosines. a) greater than 40%, b) greater than 60% methylation.



Supplementary Figure 4.3: Distribution of poly-A RNA methylated sites with coverage greater than or equal to one (> 1) for miRNA binding sites.

SAMPLE	TOTAL READS	UNIQUE READS	% MAPPABILITY	% mCG	% mCH
C57-1	29,656,975	14,815,991	50	0.1	0.08
C57-2	27,732,482	12,286,177	44.4	0.1	0.09
DBA-1	44,854,774	24,702,130	55.1	0.1	0.08
DBA-2	27,702,064	15,035,502	54.3	0.1	0.08
DBA-3	28,497,048	16,535,606	58.1	0.1	0.08
DBA-4	28,035,715	14,993,569	53.6	0.1	0.08
HESC	10,422,699	3,940,175	38	0.3	0.2

Table 4.1: Alignment statistics for bisulfite RNA libraries.

Term	Count	p-value
GO:0043209~myelin sheath	12	4.7e-09
GO:0043005~neuron projection	12	8.3e-06
GO:0019901~protein kinase binding	11	4.2e-05
GO:0070062~extracellular exosome	26	7.6e-05
GO:0008021~synaptic vesicle	6	0.00050
GO:0019904~protein domain specific binding	7	0.0027
GO:0009611~response to wounding	4	0.0032
GO:0050998~nitric-oxide synthase binding	3	0.0049
GO:0005829~cytosol	18	0.0056
GO:0032403~protein complex binding	7	0.0085

Table 4.2: Gene ontology analysis of transcripts significantly methylated in C57 and DBA mice. DAVID annotation background genes definition: union of genes with > 10X coverage across mouse samples.

Genetic background									gene	T-test
C57-4	C57-5	C57-6	C57-7	C57-8	C57-9	DBA-1	DBA-2	DBA-3		
23.9657	18.8946	23.1716	21.4372	21.2099	20.6741	20.4012	18.6511	20.7781	**Nsun2**	0.0825
3.86925	5.09761	4.83456	3.00928	3.90069	4.55794	6.05667	4.20869	3.82026	**Nsun3**	0.3835
6.2411	5.56129	5.69054	7.13309	6.62158	6.29095	5.26426	6.70116	7.19113	**Nsun4**	0.3364
4.38068	4.94361	4.99657	6.29304	4.84751	7.13965	6.18942	6.68835	5.675	**Nsun5**	0.1793
22.6027	16.3546	23.2186	20.8711	18.7248	21.4556	19.4947	23.9264	21.7452	**Wbscr22**	0.2241
5.8011	5.66596	6.15771	6.52772	5.9457	7.24576	3.50999	2.7152	3.34148	**Nsun7**	0.0002
15.6128	14.0128	14.7951	14.4944	16.0073	16.3015	13.3061	14.6871	13.8562	**Dnmt1**	0.0569
3.50478	4.91572	4.15067	2.83591	2.73274	3.13031	3.33926	3.02353	2.09883	**Trdmt1-Dnmt2**	0.0476
0.56395	0.53414	1.07848	0.84008	0.71239	0.51309	4.42361	6.11518	6.31669	**Tuba1c**	0.0056
543.26	640.739	622.909	555.09	497.804	450.216	517.396	549.068	525.169	**Plp1**	0.0631
3.26664	4.24157	2.69313	3.33013	2.21881	3.02748	2.71447	2.7551	2.98336	**Nsun6**	0.0951

Supplementary Table 4.1: Expression of the various RNA methyltransferases from mouse hypothalamus (shown are RPKM values).

REFERENCES

1. Motorin, Y., Lyko F Fau - Helm, M. and Helm, M. (2010) 5-methylcytosine in RNA: detection, enzymatic formation and biological functions. Nucleic Acids Res, 38, 1415-1430.

2. Schaefer, M. (2015) RNA 5-Methylcytosine Analysis by Bisulfite Sequencing. Methods in Enzymology, 560, 297-329.

3. Khoddami, V., Yerra, A. and Cairns, B.R. (2015) Experimental Approaches for Target Profiling of RNA Cytosine Methyltransferases. Methods in Enzymology, 560, 273-296.

4. Squires, J.E., Patel Hr Fau - Nousch, M., Nousch M Fau - Sibbritt, T., Sibbritt T Fau - Humphreys, D.T., Humphreys Dt Fau - Parker, B.J., Parker Bj Fau - Suter, C.M., Suter Cm Fau - Preiss, T. and Preiss, T. (2012) Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. Nucleic Acids Res, 40, 5023-5033.

5. Tuorto, F., Liebers R Fau - Musch, T., Musch T Fau - Schaefer, M., Schaefer M Fau - Hofmann, S., Hofmann S Fau - Kellner, S., Kellner S Fau - Frye, M., Frye M Fau - Helm, M., Helm M Fau - Stoecklin, G., Stoecklin G Fau - Lyko, F. and Lyko, F. (2012) RNA cytosine methylation by Dnmt2 and NSun2 promotes tRNA stability and protein synthesis. Nature Structural Molecular Biology, 19(9), 900-906.

6. Aguilo, F., Li, S., Balasubramaniyan, N., Sancho, A., Benko, S., Zhang, F., Vashisht, A., Rengasamy, M., Andino, B., Chen, C.-h. et al. (2016) Deposition of 5-Methylcytosine on Enhancer RNAs Enables the Coactivator Function of PGC-1α. Cell Reports, 14, 479-492.

7. Shoag, J., Haq, R., Zhang, M., Liu, L., Rowe, G.C., Jiang, A., Koulisis, N., Farrel, C., Amos, C.I., Wei, Q. et al. (2013) PGC-1 Coactivators Regulate MITF and the Tanning Response. Molecular Cell, 49, 145-157.

8. Frye, M., Dragoni I Fau - Chin, S.-F., Chin Sf Fau - Spiteri, I., Spiteri I Fau - Kurowski, A., Kurowski A Fau - Provenzano, E., Provenzano E Fau - Green, A., Green A Fau - Ellis, I.O., Ellis Io Fau - Grimmer, D., Grimmer D Fau - Teschendorff, A., Teschendorff A Fau - Zouboulis, C.C. et al. (2010) Genomic gain of 5p15 leads to over-expression of Misu (NSUN2) in breast cancer. Cancer, Lett, 289, 71-80.

9. Zhang, X., Liu, Z., Yi, J., Tang, H., Xing, J., Yu, M., Tong, T., Shang, Y., Gorospe, M. and Wang, W. (2012) The tRNA methyltransferase NSun2 stabilizes p16INK4 mRNA by methylating the 3'-untranslated region of p16. Nat Commun, 3, 712.

10. Hussain, S., Sajini, A.A., Blanco, S., Dietmann, S., Lombard, P., Sugimoto, Y., Paramor, M., Gleeson, J.G., Odom, D.T., Ule, J. et al. (2013) NSun2-Mediated Cytosine-5 Methylation of Vault Noncoding RNA Determines Its Processing into Regulatory Small RNAs. Cell Reports, 4, 255-261.

11. Hasin-Brumshtein, Y.A.-O.h.o.o.X., Khan, A.H., Hormozdiari, F., Pan, C., Parks, B.W., Petyuk, V.A., Piehowski, P.D., Brummer, A., Pellegrini, M., Xiao, X. et al. (2016) Hypothalamic transcriptomes of 99 mouse strains reveal trans eQTL hotspots, splicing QTLs and novel non-coding genes. . Elife, e15614.

12. Guo W Fau - Fiziev, P., Fiziev P Fau - Yan, W., Yan W Fau - Cokus, S., Cokus S Fau - Sun, X., Sun X Fau - Zhang, M.Q., Zhang Mq Fau - Chen, P.-Y., Chen, P.Y. and Pellegrini, M. BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. BMC Genomics, 14, 774.

13. Robinson Jt Fau - Thorvaldsdottir, H., Thorvaldsdottir H Fau - Winckler, W., Winckler W Fau - Guttman, M., Guttman M Fau - Lander, E.S., Lander Es Fau - Getz, G., Getz G Fau - Mesirov, J.P. and Mesirov, J.P. (2011) Integrative genomics viewer. Nature Biotechnology, 1, 24-26.

14. Crooks, G.E., Hon G Fau - Chandonia, J.-M., Chandonia Jm Fau - Brenner, S.E. and Brenner, S.E. (2004) WebLogo: a sequence logo generator. Genome Research, 6, 1188-1190.

15. Popis, M.C., Blanco, S. and Frye, M. (2016) Posttranscriptional methylation of transfer and ribosomal RNA in stress response pathways, cell differentiation, and cancer. Curr Opin Oncol, 28, 65-71.

16. Huang da, W., Sherman Bt Fau - Lempicki, R.A. and Lempicki, R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols, 4, 44-57.

17. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res, 31, 3406-3415.

18. Betel, D., Wilson M Fau - Gabow, A., Gabow A Fau - Marks, D.S., Marks Ds Fau - Sander, C. and Sander, C. (2008) The microRNA.org resource: targets and expression. Nucleic Acids Res, 36, D149-153.

19. Wan, Y., Qu, K., Zhang, Q.C., Flynn, R.A., Manor, O., Ouyang, Z., Zhang, J., Spitale, R.C., Snyder, M.P., Segal, E. et al. (2014) Landscape and variation of RNA secondary structure across the human transcriptome. Nature, 706-712.

20. Schaefer, M., Steringer Jp Fau - Lyko, F. and Lyko, F. (2008) The Drosophila cytosine-5 methyltransferase Dnmt2 is associated with the nuclear matrix and can access DNA during mitosis. PLoS One, 3, e1414.

21. Foster, P.G., Nunes Cr Fau - Greene, P., Greene P Fau - Moustakas, D., Moustakas D Fau - Stroud, R.M. and Stroud, R.M. (2003) The first structure of an RNA m5C methyltransferase, Fmu, provides insight into catalytic mechanism and specific binding of RNA substrate. Structure, 11, 1609-1620.

22. Hallberg, B.M., Ericsson Ub Fau - Johnson, K.A., Johnson Ka Fau - Andersen, N.M., Andersen Nm Fau - Douthwaite, S., Douthwaite S Fau - Nordlund, P., Nordlund P

Fau - Beuscher, A.E.t., Beuscher Ae 4th Fau - Erlandsen, H. and Erlandsen, H. (2006) The structure of the RNA m5C methyltransferase YebU from Escherichia coli reveals a C-terminal RNA-recruiting PUA domain. J Mol Biol, 21, 774-787.

23. Spahr, H., Habermann B Fau - Gustafsson, C.M., Gustafsson Cm Fau - Larsson, N.-G., Larsson Ng Fau - Hallberg, B.M. and Hallberg, B.M. (2012) Structure of the human MTERF4-NSUN4 protein complex that regulates mitochondrial ribosome biogenesis. Proc Natl Acad Sci U S A, 109, 15253-15258.