Inheritance and evolution of epigenetic reprogramming in Mammalian germ cells

Antoine Molaro

Ecole doctorale Complexité du Vivant - ED515

Thèse de Doctorat de l'Université Pierre et Marie Curie, Paris VI

Présentée par

Antoine Molaro

Pour obtenir le grade de
DOCTEUR de l'université Paris VI

Intitulé :

Inheritance and Evolution of Epigenetic Reprogramming in Mammalian Germ Cells

Soutenue le 9 MAI 2012 devant le jury composé de :

Dr. Frédéric Devaux	Président
Dr. Vincent Colot	Examinateur
Dr. Jonathan Weitzman	Rapporteur
Dr. Hervé Seitz	Rapporteur
Dr. Philip Avner	Examinateur
Dr. Gregory Hannon	Directeur

UPMC - Ecole Doctorale
Complexité du Vivant, ED 515
7 Quai Saint-Bernard, case 32,
75252 Paris cedex 05, France

Howard Hughes Medical Institute
Cold Spring Harbor Laboratory
One Bungtown Road
Cold Spring Harbor, NY, USA 11724

Résumé:

Héritabilité et Evolution de la Reprogrammation Epigénétique des Cellules Germinales chez les Mammifères

Chez les mammifères, les cellules germinales sont induites à partir des tissus somatiques de l'embryon post-implantatoire. Les cellules germinales primordiales nouvellement induites voient l'ensemble de leurs marques de méthylation de l'ADN intégralement effacées puis rétablies *de novo*. Cette reprogrammation épigénétique rétablit leur pluripotence et leur permet d'acquérir les marques d'empreintes parentales. Chez les mâles, la méthylation *de novo* nécessite une voie d'ARN interférence impliquant les protéines PIWI et leurs petits ARNs associés (piRNAs). Les souris mutantes pour les protéines PIWIs sont stériles et présentent une méthylation incomplète des transposons.

Nous avons généré des souris transgéniques permettant d'étudier les signaux nécessaires à la production des piRNAs. Nous montrons que des loci reprogrammés sont capable de produire des piRNAs exogènes. Nous avons ensuite étudié l'impact de la perte des piRNAs sur les profils de méthylation des spermatocytes : alors que la majorité du génome reste correctement méthylé, seul un nombre réduit de transposons, transitoirement réactivés dans les cellules germinales primordiales, semble être affecté. Troisièmement, nous avons identifié chez l'Homme des différences structurelles entre les profils de méthylation *de novo* des cellules ES et du sperme. Enfin, la comparaison des profils de méthylation du sperme d'Homme et de Chimpanzé a révélé que le génome et l'épigénome évoluent de manière distincte ou concertée selon les régions. Dans leur ensemble, nos résultats illustrent l'étonnante plasticité des interactions existantes entre le génome et l'épigenome au cours du développement et de l'évolution.

Mots-clefs : épigénétique – piRNA - cellule germinale - méthylation – transposon - évolution

Summary:

Inheritance and Evolution of Epigenetic Reprogramming in Mammalian Germ Cells

During mammalian post-implantation development, germ cells are induced from the somatic tissues of the embryo. Following their induction, primordial germ cells undergo a genome-wide erasure and *de novo* re-establishment of DNA methylation marks. This epigenetic reprogramming re-instates pluripotency and allows parental imprints to be deposited. In the male germ line, a unique RNAi pathway involving PIWI proteins and their associated small RNAs (piRNAs) is

necessary for proper *de novo* methylation. PIWI mutant mice are infertile and display methylation defects over transposon sequences.

Using a transgenic approach, we investigated the signals necessary for piRNA production. We show that artificial piRNAs can be produced from reprogrammed loci outside of their native context. We then studied the genome-wide impact of piRNA loss on germ cell methylation. Whereas most of the genome is properly methylated, only a small group of transposons transiently reactivated in primordial germ cells is affected. Also we identified important structural differences in *de novo* methylation profiles between human sperm and ES cells. Finally, we compared sperm methylation profiles between human and chimpanzee and showed that the genome and the epigenome can evolve independently. Taken together, our results highlight the surprising plasticity of genome and epigenome interactions during development and evolution.

Keywords: epigenetic – piRNA – germ cell – methylation – transposon – evolution

Table of Contents

List of figures

List of Abbreviations

CpG: Cytosine-phophate-Guanine

DNA: Deoxyribonucleic Acid

DMR: Differentially Methylated Region

(X)dpc/E: (X) Days Post Coitum, can be replaced by E(X)

(**X)dpp/P**: (X) Days Post Partum, can be replaced by P(X)

ESC: Embryonic Stem Cells

ExE: Extraembryonic Ectoderm

HMR: Hypomethylated Region

ICM: Inner Cell Mass

IP: Immunoprecipitation

LINE: Long Nuclear Interspersed Elements

LTR: Long Terminal Repeats

MASC: Multipotent Adult Stem Cells

mRNA: Messenger RNA

PGC: Primordial Germ Cells

RNA: Ribonucleic Acid

RNAi: RNA-intereference

RNAse: RNA nuclease

SINE: Short Interspersed Nuclear Element

SSC: Spermatogonial Stem Cells

TE: Trophoectoderm

VE: Visceral Endoderm

WT: Wild Type

(X)n(Y)C: (X) copies of the genome (Y) number of chromatids

Preface/Acknowledgments

Chapter 1: Introduction

In sexually reproducing organisms, germ cells embody the concept of heredity. Upon fertilization, male and female germ cells (or gametes) have the ability to re-create a fully developed and fertile offspring. On the one hand, germ cells constitute the “continuous” link between generations, ensuring the transmission of all the traits accumulated over a species’ evolutionary history. On the other, they re-ignite embryogenesis at each life cycle, allowing the production of a unique individual, different from its parent but still somewhat similar. Germ cells are directed toward the “future”- similar to a new throw of dice in the never-ending game of evolution - however, because they constitute one of the most potent state of a cell, they rewind development and differentiation. Consequently, studying the biology of germ cells implies taking a deep dive into the biology and regulation of genomes as well as into some of the most fundamental aspects of developmental biology.

In 1889, August Weismann was one of the first biologists to popularize the idea that the physical separation between the immortal “germen” (what we would now call germ cells) and the perishable “soma” (constituting the rest of the body) distinguished ontogeny from phylogeny. According to his germ plasm theory, the inheritance of new traits occured exclusively via alterations of the germen, excluding any somatic “soft inheritance”. In light of contemporary molecular, cellular and developmental biology, this vision has become substantially more

refined. Development and evolution cannot be considered as two independent phenomena and our current knowledge on the formation of germ cells points towards a more plastic concept of heredity. However, his initial observation retains something to reflect upon as a key question remains: what makes germ cells so different from somatic cells with their relationship to heredity?

In mammals germ cells undergo an atypical mode of development during which a pool of cells primed toward a somatic fate is induced to become germ cell precursors. These cells are somehow “de-differentiating” from their previous fate or, in other words, “re-acquiring” a totipotent state. This phenomenon of cell fate reprogramming is by essence an epigenetic process and requires the activation of characteristic pathways involving chromatin remodeling and RNA interference machineries. Embedded within this phenomenon lies the core of this Thesis work, where I studied the epigenetic pathways necessary for the formation of a viable male gamete and compared the outcome of these pathways with other cell types as well as between closely related organisms. In the following introductory points I will outline the specification and development of germ cells, followed by an overview the role of DNA methylation during this process, and finish by introducing a small RNA pathway linked with the establishment of methylation marks in male germ cells.

1.1: Induction and Development of Male Germ Cells in Mammals

1.1.1 Key Aspects of Early Embryonic Development

Mammalian embryonic development begins with the fertilization of a fully-grown oocyte by a mature sperm cell. During this event, the transcriptionally silent male and female pronuclei fuse to form a diploid 1-cell zygote. In mouse, this zygote will remain transcriptionally silent for approximately 20 hours and rely on maternally deposited mRNAs for protein synthesis (Figure 1.1, Evsikov et al., 2004; Wang et al., 2004). The maternal-zygotic transition occurs at the mid-2-cell stage. Maternal mRNAs are degraded and zygotic transcription begins (Flach et al., 1982; Latham et al., 1991; Bouniol et al., 1995). The zygote follows a series of rapid cell divisions most of which are asynchronous past the 8-cell stage. As cells divide, they become specified into 2 major lineages: the trophoblast, contributing to extra-embryonic tissues, and the inner cell mass (ICM), contributing to all 3 future embryonic germ layers (endoderm, ectoderm and mesoderm). By the blastocyst stage (around 4.5 dpc in mouse and day 6-7 in human), the trophoblast lineage surrounds the ICM and its adjacent cavity – the blastocoel. By the time of implantation, the blastocyst is composed of 3 cell lineages: the trophectoderm, derived form the trophoblast, the epiblast and the primitive endoderm, derived from the ICM (Figure 1.2).

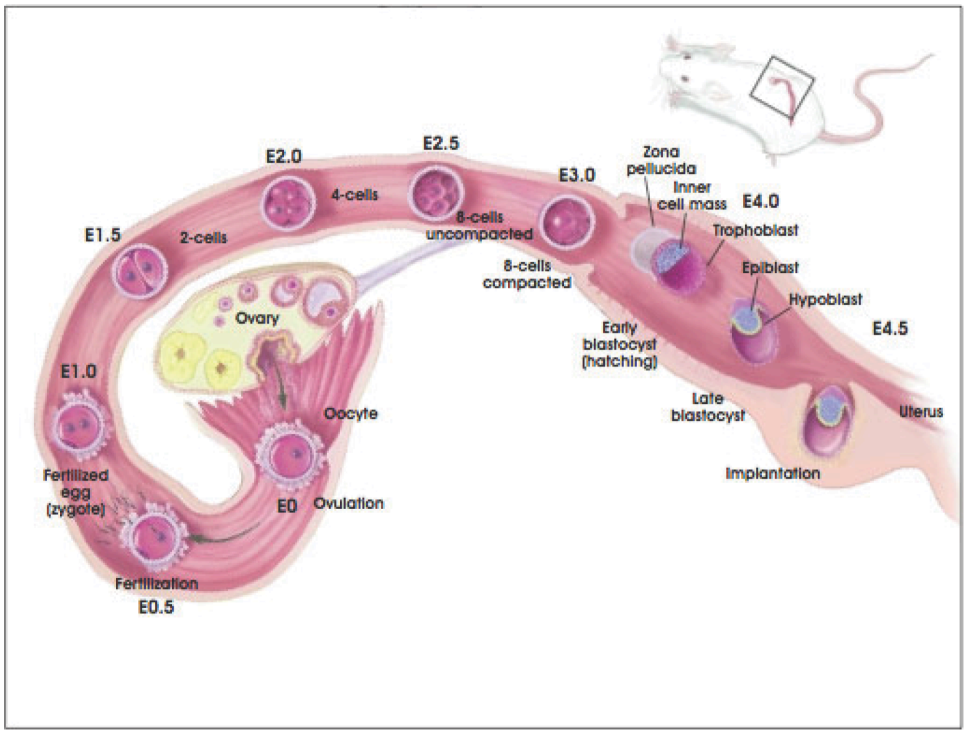


Figure 1.1: Mouse pre-implantation development
This image depicts a cross section of the female adult gonads and uterine horn. It shows the migration and early steps of pre-implantation development of the mouse zygote. Post fertilization, the zygote undergoes cleavage and compaction at E3.0. Blastocyst specification occurs between E4.0 and E4.5; the ICM is shown in blue and the trophoblast in red. Implantation occurs past E4.5. *Image credit: stemcells.nih.gov.*

During these key steps of pre- and post-implantation development, signaling pathways and transcriptional programs act coordinately to maintain the developmental potency of the embryo, as exemplified by the well-characterized stem/pluripotent-cell transcriptional network involving Oct4, Nanog and Sox2 (reviewed by Chambers and Smith, 2004; Surani et al., 2007; Silva and Smith, 2008). This network, in addition to key epigenetic modifiers, such as the Ezh2 complex (mammalian polycomb like complex), have been shown to be essential for proper ICM development in mouse embryos as well as for maintaining the

stemness of ICM derived embryonic stem cells (ESCs) *ex-vivo* (Chambers et al., 2003 and 2007; Mitsui et al., 2003; Nichols et al., 1998; Masui et al., 2007). In the epiblast and in cultured ES cells, Oct4, Nanog and Sox2 dominate the transcriptional network, directing their target genes to continuously repress lineage specification and maintain self-renewal. From the zygote to the establishment of the epiblast, developmental potency decreases (from totipotent to pluripotent), as epiblast cells cannot give rise to trophoblastic tissues following implantation.

1.1.2 Induction of Primordial Germ Cells from Somatic Tissues at E6.5

The induction of primordial germ cells occurs soon after blastocyst implantation. PGCs arise from a competent region located at the junction of the extraembryonic ectoderm (ExE), visceral endoderm (VE) and the adjacent anterior-proximal epiblast (Figure 1.2, Ginsburg et al., 1990; Lawson and Hage, 1994). At ~E6.0 the most proximal epiblast is still pluripotent but is being restricted towards a somatic fate (Saitou et al., 2002; Yabuta et al., 2006; Kurimoto et al., 2008). Indeed, single cell expression profiles revealed that this tissue expresses the key genes priming the development of the embryonic and extra-embryonic mesoderm (*T-brachyury*, *Hoxa1*, *Hoxb1*) and that early PGC precursors initially display these expression features. Interestingly, via heterotopic and orthotopic transplant experiments in mouse, it was shown that the proximal and distal regions of the epiblast maintain their fate in a non cell-

autonomous fashion, and that PGCs can arise from distal epiblast cells transplanted in the proximal region.

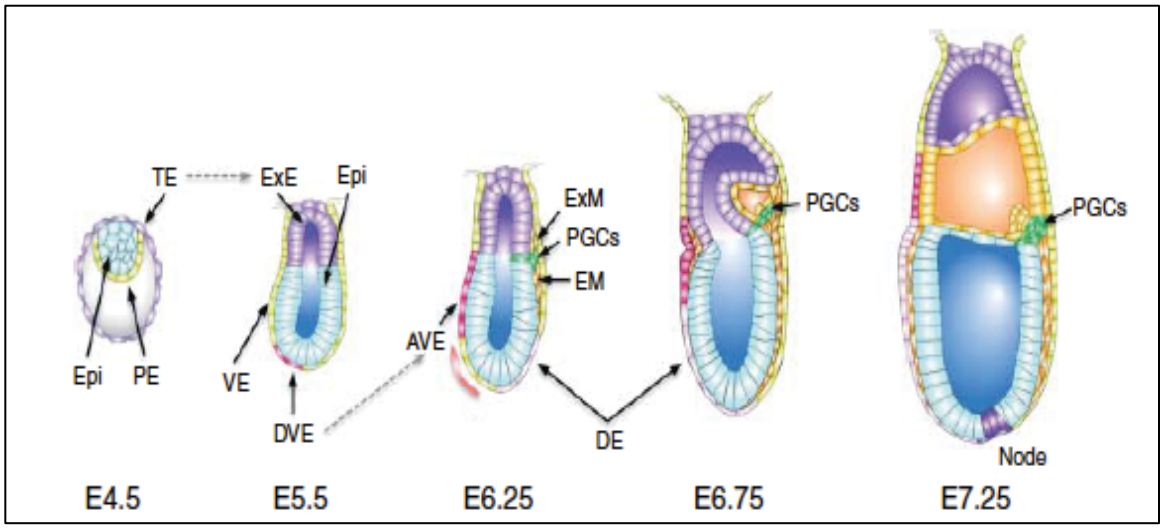


Figure 1.2: Primordial germ cell induction in the mouse embryo.
At E4.5 the blastocyst is composed of three cell types, trophoectoderm (TE, purple), primitive endoderm (PE, yellow), and epiblast (Epi, blue). The TE cells in direct contact with the epiblast proliferate and from the extraembryonic ectoderm (ExE) at E5.5. The initial embryonic patterning including anterior–posterior polarity formation, gastrulation, and germ cell specification is mediated by signalings from the ExE- and PE-derived visceral endoderm (VE) that cover the epiblast. Primordial germ cell (PGCs, green) induction occurs between E6.75 and E7.5.
DVE, distal visceral endoderm; AVE, anterior visceral endoderm; ExM, extraembryonic mesoderm; EM, embryonic mesoderm; DE, definitive endoderm. *Image adapted form Mitinori Saitou and Masashi Yamaji, 2010.*

It is believed that antagonistic signals originating from the proximal VE and BMP4/8 signaling from the ExE restrict a pool of ~40 cells to become PGCs (Lawson et al., 1999; Chang et al., 2001; Tremblay et al., 2001; Ohinata et al., 2009). Between 6.5-7.5 dpc, these PGCs acquire a unique fate characterized by the sequential expression of *Blimp1* (B-Lymphocyte induced maturation protein

1, also known as *Prmd1*), *Prmd14* (PR domain containing transcription factor 14), and *Stella/Dppa3* (Developmental pluripotency-associated gene 3) (Ohinata et al., 2005; Vincent et al., 2005; Yamaji et al., 2008; Payer et al., 2003). They also stain positively for alkaline phosphatase (Tam et al., 1996). Whereas Blimp1 has been shown to be associated with the suppression of the prior mesodermal-somatic fate (for example down regulating Hoxa1-2-3 and Tbx6), Prmd14 is necessary for the re-acquisition of pluripotency, in particular via the up-regulation of *Sox2* (Yamaji et al., 2008).

Following a brief period of proliferation and concomitant with gastrulation, newly induced PGCs migrate along the dorsal mesentrery and colonize the future gonads in the dorsal part of the emryo by ~E10-11.5. In the current model for PGC migration, homotypic cell adhesion and heterotypic repulsion serves to direct cellular migration and directionality (Tanaka et al., 2005). An interferon-inducible transmembrane protein (Ifitm3, also known as Fragilis) is expressed after induction by BMP signaling during PGC specification (Saitou et al., 2002). A similar protein, Ifitm1/Fragilis2, is later expressed in the developing PGC and the nascent mesodermal cells (Tanaka et al. 2001; Lange et al. 2003). A subsequent down-regulation of *Ifitm1,* but the persistence of Ifitm3 on the cell surface, results in a heterotypic repulsion between the PGC and its surrounding mesoderm. This repulsion restricts the PGC to the endoderm, thus facilitating their passive migration along the hindgut as it elongates (Lawson and Hage, 1994). The subsequent exit of the PGC from the endoderm (towards the developing embryonic gonads) might also be mediated by the activation of *Ifitm1* or

deactivation of *Ifitm3*, eliminating the heterotypic repulsion (Tanaka et al., 2005). However, the model remains disputed, as the knockout of *Ifitm3* or the *Ifitm* cluster does not appear to affect PGC development (Lange et al., 2008).

1.1.3 Colonization of Embryonic Gonads by Mitotic and Post Mitotic PGCs between E11.5 and P2

Between 11.5 and 13.5dpc PGCs proliferate, undergo a first wave of major epigenetic changes (discussed in detail in section 1.2) and acquire their sex-specific fates. By 13.5dpc they are referred to as gonocytes and their total number doesn't exceed ~10000 cells per embryo or ~5000 cells per gonads. Hereafter, PGCs undergo sex-specific developmental paths, which will lead to the formation of highly dimorphic post-meiotic male and female gametes later in the life of the animal (spermatozoa and oocytes respectively). Briefly, whereas male PGCs enter a long G1 phase at 13.5dpc from which they exit only after birth (~2dpp or P2), female PGCs begin meiosis during embryonic development and arrest at the end of prophase 1 until post-pubertal development.

As PGCs move from their site of induction to colonize the gonad gene expression is extremely dynamic. In addition to the genes essential for lineage restriction and fate specification, once in the gonads, PGC activate the expression of *Gcna-1* (germ cell nuclear antigen 1, Enders et al., 1994) and the widely conserved RNA-helicase VASA (*Mvh*, Mouse vasa homolog, Toyooka et al., 2000) both of which are exclusively found in this lineage. Another important marker of post migratory male PGCs is the sustained expression of *Oct4*, which

in conjunction with Nanog, has been shown to be essential for PGC survival past 11.5dpc (Chambers et al., 2007; Kehler et al., 2004). These genes and others constitute useful markers for cell staining and/or isolation of PGCs from dissected embryos.

1.1.4 From Spermatogonial Stem Cells to Mature Sperm

Male germ cell development is ultimately achieved soon after birth when PGCs occupy their final niche at the basement membrane of seminiferous tubules and become a dedicated stem-cell population termed spermatogonial stem cells (SSCs, Figure 1.3). SSCs isolated and cultured from mouse or human adult testes have been shown to differentiate into various cell types *in-vitro* and maintain their ability to form teratomas in immune-compromised mouse models even after multiple passages (Seandel et al., 2007, Guan et al., 2006). More importantly, adult SSCs and their cultured derivative, known as multipotent adult stem cells (MASCs), display an ESC like potential as they can contribute to all 3 germ-layers (endoderm, mesoderm and ectoderm) when injected into early blastocysts (Guan et al., 2006).

Mouse and human adult testis share a common structural organization. Seminiferous tubules are packaged into each gonad, and spermatogenesis occurs in a staged fashion from the basal side to the luminal side of each tubule. The SSCs and their daughters (type A and B spermatogonia) lie basally, they divide throughout the life of the individual. Moving towards the lumen, cells undergo meiosis and spermiogenesis, which include the final differentiation steps

of mature spermatozoa. Pre-puberty, the first wave of meiosis is believed to occur in a somewhat synchronous fashion (Bellve et al., 1977, see Evaluation of the Testis, Cache River Press 1990) albeit with some discrepancies between tubules. By taking advantage of this synchronized first wave of meiosis, staged isolation of meiotic cells can be achieved and their proprieties characterized. As SSCs divide, their daughters - type-A spermatogonia - move away from the basal membrane, enlarge in size and become type-B spermatogonia. Type-B spermatogonia maintain their mitotic potential and increase in number prior to their entry into meiosis as primary spermatocytes - characterized by fully replicated and condensed chromatin (with a chromosome count of 2n4C, Figure 1.3). During the first meiotic division, homologous chromosomes pair, crossing-overs occur and each homolog moves into distinct secondary spermatocytes (chromosome count of 1n2C). These cells enter the second division of meiosis, where sister chromatid segregate into distinct daughter cells to produce haploid spermatids (chromosome count of 1n1C). During meiosis, germ cell developmental potential decreases and *Oct4* expression is progressively lost (Pesce 1998). In the most final stage of sperm differentiation (spermiogenesis), canonical nucleosomes are replaced by protamines, compacting the genome into a highly dense, transcriptionally inactive, structure (Coffigny et al. 1999; Cho et al. 2001).

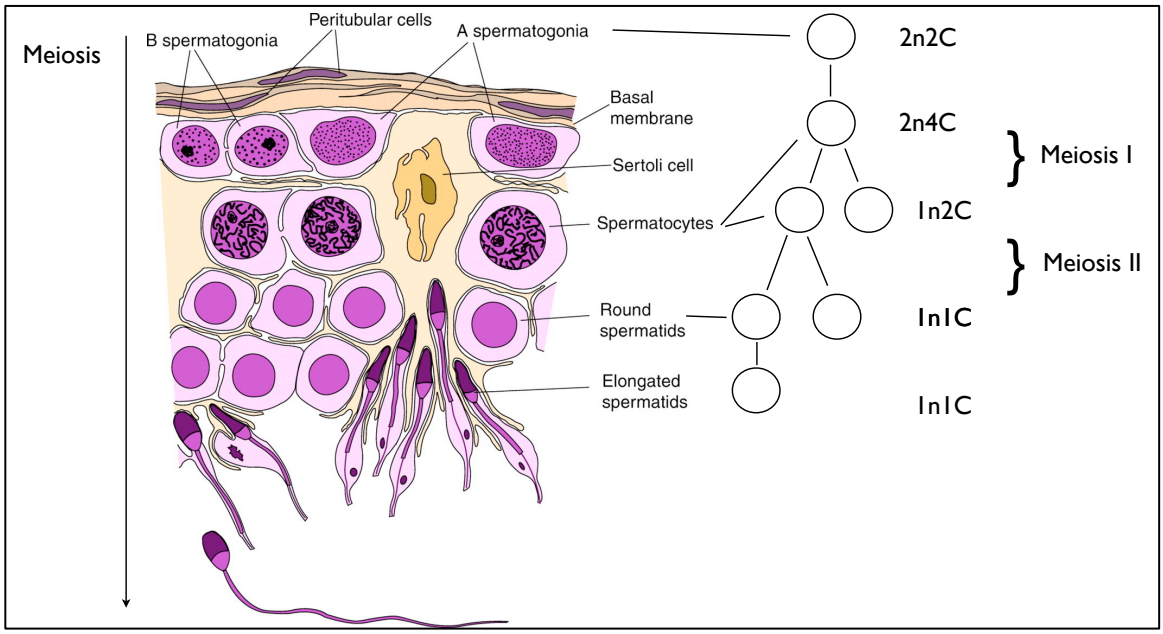


Figure 1.3: Structure of the mouse seminiferous tubule and mouse meiosis
Schematic representation of a seminiferous tubule cross-section. A schematic view of meiosis is shown on the right together with the corresponding DNA contents. Meiosis occurs in a polarized fashion from the basal membrane to the lumen of the tubule. Spermatogonial stem cell derived Type-A and -B spermatogonia are found at the most basal side of the tubule. Their DNA content is 2n2C - 2 copies of the genome, diploid, with a total of 2 chromatids, one per homolog. Fully replicated spermatogonia enter Meiosis I as primary spermatocytes and exit as secondary spermatocytes – DNA content going from 2n4C (replicated diploid genome) to 1n2C (replicated haploid genome). Secondary spermatocytes undergo meiosis II and exit as round spermatids with DNA content of 1n1C. *Adapted form de Rooij, 2003*.

1.2: DNA Methylation Dynamics During Mammalian Development

Organogenesis and cell-fate specification is a complex process whereby a pool of cells with equivalent potential progressively acquires a dedicated function in the organism – a process known as differentiation. At the molecular level, differentiation is characterized by cell-type specific transcript expression (of both coding and non-coding RNAs), reflecting the progressive restriction of the genome to a specific state. Chromatin based epigenetic regulation drives the establishment and maintenance of these states through mitosis and in some cases meiosis. Epigenetic modifications rely on direct chemical modification of the DNA molecule (e.g. DNA methylation) as well as post-translational modification of the N-terminal tail of histones, which compact the genome into the chromatin fiber. Histone modifications and DNA methylation unequally mark the genome in a time and context dependent fashion (e.g. at promoters, enhancer, repeats…) and contribute to the regulation of its transcriptional activity.

The previous section overviewed the key developmental steps leading to the formation of mature male germ cells. Prior to their induction, germ cells belong to a population of cells primed to differentiate into somatic tissues. The specification of the germ cell lineage is associated with the re-acquisition of pluripotency, a feature previously seen in the ICM of growing blastocysts. During these reprogramming events the epigenome is particularly dynamic, with genome-wide erasure and re-establishment of histone and DNA methylation marks. Focusing on DNA methylation, the following section will review the key

aspects of the deposition/removal of methylation marks and finish by a detailed description of DNA methylation re-programming in developing PGCs.

1.2.1 Key Aspects of DNA methylation

In mammals, DNA methylation is the deposition of a methyl (-CH3) group on the fifth carbon of cytosines. Methyl-Cs are extremely prevalent across the genome, with 60-70% of all cytosine being methylated in somatic tissues (Bird et al., 1985; Ehrlich et al., 1982). In most of the human and mouse cell types investigated thus far, DNA methylation is primarily detected in the context of a CpG di-nucleotide. In this context, methylation is predominantly found to be symmetrical on both strands, allowing the maintenance of methylation patterns during replication (Holliday et al., 1975; Riggs AD 1975; Wigler et al., 1981a/b). In 1975, Holliday and Pugh postulated that chemical modifications of the DNA molecule at/or around regulatory elements would accompany the modulation of gene transcription involved in controlling cell proliferation and differentiation during development. In addition, they proposed a model explaining how these marks would be deposited and maintained during cell division in a semi-conservative fashion. Their predictions turned out to be valid, with the discovery of maintenance and *de novo* methyl-transferase enzymes that catalyse the active deposition of methyl marks throughout the genome. Four genes with putative DNA methyl-transferase activity can be found in most mammalian genomes (*dnmt1*, *2*, *3a* and *3b*). Whereas DNMT1, 3a and 3b have been shown to be catalytically active *in-vivo*, the function of DNMT2 is still largely mysterious.

1.2.1a Establishment of DNA methylation in Mammals

DNMT1 was the first isolated DNA methyl-transferase in mammals (Bestor et al., 1988). Several lines of evidence suggest that DNMT1 is preferentially involved in the maintenance of methylation marks. During S-phase, DNMT1 is targeted to sites of DNA replication via several N-terminal domains interacting with component of the replication fork such as PCNA and UHRF1 (Leonhardt et al., 1992; Liu et al., 1998; Sharif et al., 2007; Arita et al., 2008). In addition, inhibition of DNMT1, using chemical inhibitors or antisense blocking oligonucleotides, affects the formation of replication forks (Knox et al., 2000). Finally, *in vitro* experiments on purified DNMT1 have shown that DNMT1 preferentially catalyzes the addition of a methyl group in the context of hemimethylated dsDNA (Zucker et al., 1985; Flynn et al., 1996; Pradhan et al., 1999). Thus the recruitment of DNMT1 during DNA replication is likely responsible for the re-establishment of symmetrical methylation at hemimethylated sites generated upon daughter strand synthesis (Figure 1.4). The recent solving of the crystal structures of DNMT1:DNA complexes provided some evidence for DNMT1 strong preference for CpG sites and showed that, when not engaged in catalysis, DNMT1 folding inhibits its interaction with fully unmethylated sites (Song et al., 2011).

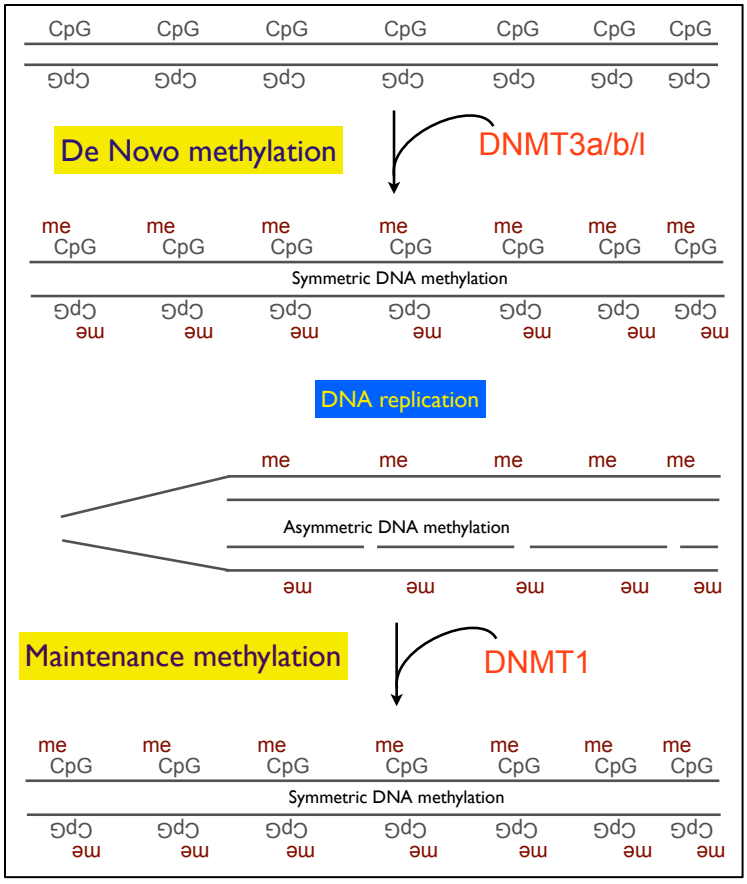


Figure 1.4: Establishment of DNA methylation
In the context of CpG dinucleotides, methyl-groups (red) are symmetrically deposited on both strands of the DNA sequence by *de novo* methyl-transferases (top). Upon DNA replication, the newly synthesized strands are hypomethylated, creating asymmetric sites (middle). The recruitment of DNMT1 at replication forks re-establishes symmetrical methylation (bottom).

Because of the inherent dynamic nature of the genome during development and differentiation, methylation marks are often deposited *de novo*, outside of a maintenance context. Whereas DNMT1 has been shown to have *de novo* methyl-transferase activity, albeit at low frequency, DNMT3a and b enzymes primarily achieve this function. The *de novo* methyl-transferase activity of these two enzymes on unmethylated DNA has been shown both *in-vivo* and *in-vitro* (Okano et al., 1998 and 1999; Gowher et al., 2001; Hsieh CL, 1999).

Despite their close sequence similarity, DNMT3a and b act on different genomic targets and catalyze DNA methylation in slightly different ways. DNMT3b has been shown to be essential for *de novo* methylation of peri-centromeric regions and act as a processive enzyme (Okano et al., 1999; Gowher et al., 2002). In contrast, DNMT3a was proposed to act preferentially on intergenic and genic single copy loci and requires *de novo* targeting following each methylation event (Lin et al., 2002; Hata et al., 2002).

Interestingly, de novo methyltransferases require the presence of a non-catalytic co-factor named DNMT3L to be functional in-vivo. DNMT3L is strongly up-regulated in zygotes and developing gonads and its presence in DNMT3 complexes enhances the catalytic activity of the enzymes (Hata et al., 2002; Chedin et al., 2002; Suetake et al., 2004). The structural analysis of DNMT3a:DNMT3L complexes revealed that through its C-terminal interaction with the catalytic site of DNMT3a, DNMT3L promotes the formation of a tetramer including 2 copies of the DNMT3a:3L complex and enhances targeting to the chromatin fiber (Jia et al., 2007; Jurkowska et al., 2008).

1.2.1b Erasure of DNA Methylation Marks: Current Models

DNA methylation is a reversible modification. Two modes of demethylation have been proposed in the literature: active and passive. Passive demethylation occurs when methyl marks are not re-established during cell division and get diluted out after several rounds of division. Recently, a flurry of studies suggested

that, similar to plants, hydroxylation of methyl-cytosines and base excision repair mechanisms could be involved in active DNA demethylation in mammals (reviewed by Bhutani et al., 2011). In this model, hydroxylation of methylcytosines by TET proteins (ten-eleven translocation) is followed by an AID/APOBEC mediated deamination into hydroxymethyluridine. This base is then recognized by the base pair excision pathway, which mediates its replacement with unmethylated cytosines. For example, these studies showed that knocking-down TET or AID (Activation Induced Deaminase) proteins impaired ES cell differentiation and induced pluripotent stem cell reprogramming, demonstrating that removal of DNA methylation memory was an essential step in the transition between distinct cellular states (Ficz et al., 2011; Bhutani et al., 2010). Moreover, animal mutants for TDG (a glycosyase involved in the base excision repair pathway) display demethylation defects during early embryogenesis (Cortellino et al., 2011).

There is still much debate as to which mode of demethylation is more prevalent during development. Because of its unspecific nature, passive demethylation seems like an attractive model to explain a fast and global erasure of methylation profiles. Active demethylation, on the other hand, seems more adapted to target a small subset of regulatory regions that fluctuate in methylation state during stem cell maintenance and differentiation. The reality might be more complex as both processes could co-occur in the same cellular context (see for example Rougier et al., 1998; Mayer et al., 2000; Inoue et al., 2011).

1.2.1c Interaction between DNA Methylation and other Chromatin Modifications

DNA methylation is functionally linked to the modification of histone tails, which can be methylated, acetylated, ubiquitinated or phosphorylated (reviewed by Cedar and Bergman, 2009). Briefly, whereas di- and tri-methylation of histone H3 lysine 9 (H3K9me) is associated with constitutive heterochromatin and transcriptional silencing, H3K4 and/or H3K36 methylation and H3K9 acetlyation mark actively transcribed regions in euchromatic domains. Other marks, such as H3K27 methylation, are found across regions displaying a more plastic pattern of expression named "bivalent domains" – switching rapidly from permissive to repressive (Bernstein et al., 2006). DNMTs are found in complex with histone modifiers such as de-acetylases (HDACs, Fuks et al., 2001; Rountree et al., 2000; Jones et al., 1998; Nan et al., 1998), methyltransferases (e.g. SUV39, G9a) and the heterochromatin-associated protein HP1 (Tachibana et al., 2002; Fuks et al., 2003; Esteve et al., 2006; Smallwood et al., 2007), showing how the interplay between higher order chromatin and DNA methylation dynamically structure the genome.

Several studies have now shown that un-methylated DNA is generally associated with H3K9acetylated/H3K4methylated enriched domains (see for example Edwards et al., 2010; Hashimoto et al., 2010). On the other hand, methylated regions are often found overlapping HP1, H3K9me2 and H3K9me3 heterochromatin. It is still unclear which chromatin signal (DNA methylation or

histone modifications) triggers the establishment of the other. Both epigenetic signals could act in a self-enforcing loop as seen in plants and fungi (Tariq et al., 2003; Tamaru et al., 2001; Jackson et al., 2002; Henckel et al., 2009). For example, inhibition of de novo methylation or the ectopic introduction of hypomethylated DNA reduces H3K9me deposition, enhances histone de-acetylation and recruits H3K4 methyl-transferases (Hashimshony et al., 2003; Nguyen et al., 2002; Thomson et al., 2010). In turn H3K4 methylation inhibits DNMT3L recruitment at promoters and prevents de novo methylation (Ooi et al., 2007).

1.2.2 DNA Methylation: Functional Relevance and Evolutionary Consequences

1.2.2a Function of DNA Methylation during Mammalian Development

For the past 30 years, DNA methylation has been studied in a wide variety of developmental contexts, including human and mouse differentiated cells, stem cells and whole tissues. These studies were focused on establishing the profiles, and studying the regulatory effects, of DNA methylation over a limited set of genomic loci or over ectopically introduced sequences – such as integrated viruses and transfected plasmids. These studies revealed that the maintenance of gene expression profiles during cell fate restriction and differentiation largely relies on methylation gain and losses at promoters or other regulatory elements (see Bird A, 2002; Goll and Bestor, 2005). In addition, mono-allelic and imprinted gene expression are mediated by the methylation of control regions on the silenced allele during embryonic and germ cell development (Li et al. 1993; reviewed in Surani MA, 1998); the most extreme example of this is X-chromosome inactivation, where, in females, an entire chromosome is mono-allelically expressed. The establishment and maintenance of X-chromosome inactivation depends on epigenetic modifications including a chromosome wide enrichment for methyl-Cs (for example see Kaslow et al., 1987; Csankovszki et al., 2001; Panning et al., 1996). Finally, and probably most importantly, the silencing of genomic repeats rely almost exclusively on DNA methylation. Preventing the establishment of DNA methylation over transposons leads to their

transcriptional up-regulation and induces genomic damage by non-homologous recombination (Okano et al., 1999; Walsh et al., 1998; Bourc'his and Bestor, 2004; Yoder and Bestor, 1997).

The essential role of DNA methylation is further highlighted by the deleterious effect of *dnmt* knockouts. Targeted deletion of DNMT1, 3a and 3b all lead to embryonic lethality (Li et al., 1992; Okano et al., 1999), albeit with different phenotypes. Interestingly, while none of these enzymes have been shown to be required for the maintenance of ES cell self-renewal, but instead affect differentiation in culture. As expected, *dnmt1* targeting leads to a 3-5 fold genome-wide reduction of cytosine methylation. More importantly, homozygous mouse embryos arrest their development between 9dpc and 11dpc and die *in-utero*. Similarly, *dnmt3a/b* targeted deletions lead to severe embryonic phenotypes. Interestingly, whereas dnmt3b null embryos can't survive past 9.5dpc, dnmt3a nulls develop to term but die soon after birth at around day 4. This strongly supports the notion that these two enzymes have essential and non-redundant roles. A deeper look at methylation defects harbored in these mutants revealed that peri-centromeric and transposon methylation is more affected in dnmt3b and double nulls than in dnmt3a mutants alone. The methylation of imprinted loci in whole-embryos is not affected in these mutants; however, proper imprinting fails to be established during ES cell differentiation in culture.

Because of the early embryonic lethal phenotypes displayed by these mutants, it has been hard to study their direct effect on germ cell methylation,

which occurs past 13.5dpc in mouse. The answer came from the study of dnmt3l null animals and revealed that this cofactor was essential for male germ line development and for the establishment of maternal imprints in developing oocytes (Bourc'his et al., 2001 and 2004). *Dnmt3l* nulls fail to establish *de novo* methylation during PGC development and display a strong up-regulation of retro-transposon transcripts. These germ cells enter meiosis but undergo apoptosis around the pachytene stage of meiosis 1. On the other hand, heterozygous embryos born from homozygous females die a mid-gestation due to biallelic expression of imprinted loci.

1.2.2b Evolutionary Impact of CpG Methylation

One interesting long-term effect of DNA methylation is its genome-wide impact on C to T transitions over evolutionary time scales (Duncan and Miller, 1980; Bird, A 1980; Cooper et al., 1989; Ehrlich et al., 1990; Schorderet et al., 1992). Methylated Cs can undergo spontaneous deamination into uracil and are subsequently replaced by thymines. Consequently, CpGs are strikingly under represented in mammalian genomes (Human and Mouse genome sequencing consortium). However, mammalian genomes retain areas of relatively high CpG density, called "CpG islands" (CGIs) (Gardiner-Garden and Frommer, 1987). These regions are conserved across vertebrate genomes and have somehow avoided CpG depletion over evolutionary time. In mouse and humans, CGIs typically overlap other important genomic elements such as transcriptional start sites (TSS) (Takai and Jones 2002, Gardiner-Garden and Frommer, 1987).

There is still much mystery as to what evolutionary force shapes CGIs over short and long evolutionary periods. One likely explanation is that hypomethylated regions display such enrichment over time because of lower rates of deamination, especially if hypomethylated in germ cells (Bird A, 1985). Alternatively, higher CpG densities could regulate transcription factor binding affinity or be retained due to selective pressure in the course of speciation, independently of methylation. Recently, Cohen, Kenigsberg and Tanay (Cohen et al., 2011) explored all possible scenarios, and suggested an interesting model in which the combination of CGI genomic context (promoters, intergenic…), methylation status (in germ cells or somatic tissues) and positive selection at individual CpG sites over regulatory or coding regions could explain the current CGI content of primate genomes. Of course only a detailed profiling of methylation across species and tissues would put these models to the test.

1.2.3 Germ Cells and ES Cells: Outcome of Epigenetic Reprogramming

As highlighted in section 1.1, early PGCs and the ICM of pre-implantation embryos share some interesting molecular and developmental features - the most striking of them being the shared expression of the stem-pluripotent transcriptional network mediated by Oct4, Nanog and Sox2. In addition to these overlapping transcriptional programs, both lineages undergo a wave of genome-wide erasure and re-establishment of methylation marks prior to their

specification (Monk et al., 1987; reviewed by Reik et al., 2001; Surani et al., 2007). The function of this epigenetic reprogramming still remains elusive, but it is associated with the establishment and maintenance of pluripotency and is needed for the acquisition of the unique chromatin signature of the ICM and the establishment sex specific epigenetic states in PGCs (Figure 1.5).

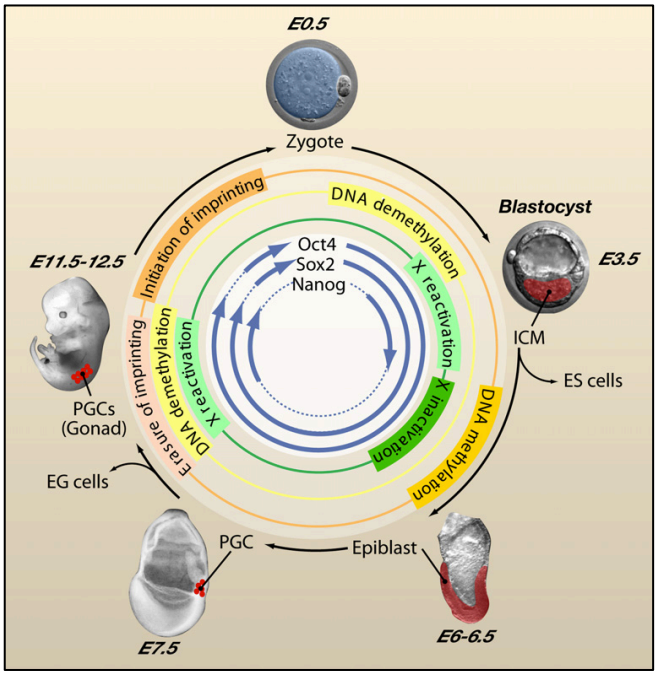


Figure 1.5: Reprogramming during mammalian development. The figure depicts the main epigenetic changes occurring during critical stages of development. The totipotent zygote contains maternally inherited epigenetic modifiers and transcription factors, including Oct4, Sox2, and Nanog. These, together with the embryonic transcripts, regulate development to the blastocyst stage, where the pluripotent ICM is established. PGCs exhibit epigenetic and transcriptional states that are associated with pluripotency, and the ensuing epigenetic reprogramming re-generates totipotency. *Adapted from Surani 2007*.

1.2.3a DNA Methylation Reprogramming in Pre-implantation Embryos

In the mouse 1-cell zygote, several reports have shown that the paternal genome undergoes a rapid wave of DNA demethylation prior to the onset of the first cleavage (Mayer et al., 2000; Oswald et al., 2000; Santos et al., 2002; Wossidlo et al., 2010). Using 5mC antibodies as well as targeted bisulfite sequencing, these studies have shown that, as protamine-to-histone exchange

occurs in the paternal pronucleus, overall levels of DNA methylation decrease at most genic and intergenic sequences with the exception of imprinted loci. In contrast, the maternal genome gradually loses methylation as cleavage occurs in what looks like a replication dependent process (Howlett et al., 1991; Kafri et al., 1992; Rougier et al., 1998; Inoue et al., 2011). By the 8-cell stage most zygotic methylation marks have been erased, and only imprinted methylation can be detected.

Re-acquisition of DNA methylation is observed in the developing ICM; however, the trophoblast lineage seems to be re-methylated to a lesser extent (Monk M, 1990; Santos et al., 2002). DNMT expression is very dynamic during pre and post-implantation development. DNMT3b is the first *de novo* methyl-transferase to be expressed in the ICM between 4.5 and 7.5dpc (Watanabe et al., 2002; Hirasawa et al., 2008). Past 9.5dpc DNMT3b is replaced by DNMT3a, which is then detected throughout the embryo during the rest of development. There is still much to be learned about the precise methylation profiles the ICM harbors after *de novo* methylation. An attempt to answer this question come from the study of methylation patterns in cultured ES cells, which are derived from the ICM and are thought to preserve its developmental potency. ES cells express all 3 DNMTs and display dynamic change in methylation, resembling those seen during organogenesis, upon *in-vitro* differentiation. The recent study of human H1 and H9 ES cell methylomes revealed that ES cells have methylation levels approaching, but not reaching, those seen in somatic tissues, indicating a global *de novo* methylation (Lister et al., 2009, Laurent et al., 2010). They also

confirmed earlier observations suggesting that, in ES cells, promoters are generally hypomethylated at developmentally regulated genes (e.g. HOX clusters). By comparing ES cells to various differentiated tissues, these studies showed that reduced methylation at promoters and high methylation in gene bodies was observed at transcriptionally active genes. Interestingly, they also found regions displaying non-CpG methylation, a feature thought to be restricted to plant and fungi. Although still under investigation, this form of cytosine methylation could affect about 20% of cytosines found outside a CpG context in undifferentiated ES cells.

1.2.3b Reprogramming in Germ Cells

During PGC induction and development, an important remodeling of the chromatin fiber precedes the erasure of DNA methylation genome-wide, which is completed by 13.5dpc in mouse (Hajkova et al., 2008; Seki et al., 2005 and 2007; Popp et al., 2010). Following their induction and during G2 arrest (between ~7.5 and ~9.5dpc), the nuclei of PGCs progressively enlarge and immunofluorescence stainings for H3K9me2/3, H3K9ac and H3K27me3, are greatly reduced. By 10.5dpc, PGC chromatin is believed to be extremely loose but pervasive transcription is maintained at a low level via RNA polymerase II Serine 5 and Serine 2 dephosphorylation (Seki et al., 2007). Histone variants are also transiently incorporated (e.g. H3.3 and H2A.Z), and the canonical histone linker H1 is lost, indicating that deeper changes in nuclear chromatin might also prevent aberrant gene expression (Hajkova et al., 2008 and 2010). Again using

immunofluorescence, it was shown that by 12.5dpc the genome wide distribution of repressive histone modifications (e.g. H3K27me3 and H3K9me2/3 but not H3K9ac), histone variants, linker H1 and active RNA polII, revert to the levels seen in surrounding somatic tissues. A detailed picture of the distribution of these chromatin changes is still lacking due to the low abundance of these cells and the lack of culture models recapitulating these events.

In 7.5dpc PGCs, focused studies revealed that imprinted loci, transposons and single copy genes still bear substantial methylation indicative of their somatic origin – one X chromosome is inactivated in females (Sugimoto et al., 2007; Tam et al., 1994) and at least some copies of LINE-1s and IAP retro-transposons retain over 70% of methylated Cs at their regulatory sequences (Hajkova et al., 2002).

As PGCs enter G2 arrest, DNMT3a and b are transcriptionally down regulated. In contrast, Dnmt1 is still present albeit at low levels (Seki et al., 2005; Kurimoto et al., 2008). During this time frame, the global levels of 5mC staining begin to reduce (Seki et al., 2005) and reach their lowest when PGC resume proliferation past ~9.5dpc suggesting that a portion of the genome is loosing methylation. Detailed analysis of the methylation status of PGCs between 10.5 and 13.5dpc using methylation sensitive restriction assays and focused bisulfite sequencing showed that the dynamics of DNA demethylation of imprinted loci, single copy genes and transposons is more heterogeneous than previously thought (Walsh et al., 1998; Lees-murdock et al., 2003; Lane et al., 2003; Hajkova et al., 2002; Lee et al., 2002). Some copies of LINE-1 and IAP

retrotransposons retain substantial methylation until 11.5, and loose most of this signal between 12.5 and 13.5dpc. Notably, more copies of IAP seem to retain methylation even after 13.5dpc and demethylation is less prominent in female PGCs (Lees-murdock et al., 2003; Lane et al., 2003). In addition, whereas some single copy genes and imprinted loci (maternally imprinted Nnat, and paternally imprinted Peg3/5 and H19) show partial demethylation as early as 10.5dpc, most achieve full demethylation rapidly between 12.5 and 13.5dpc (Xist promoter, Peg10…, Lee et al., 2002; Hajkova et al., 2002). These data suggest that both passive and active demethylation could be involved in the drastic erasure of methylation reaching its lowest at 13.5dpc (Figure 1.6). To gain further insight into PGCs methylation reprogramming, Popp and colleagues recently reported a low coverage survey methylation of PGCs at E13.5 using bisulfite sequencing, and compared them to sperm and ES cells (Popp et al., 2010). Despite the low coverage of this study, they were able to show that by 13.5dpc overall methylation levels are below 10% in PGCs compared to over 70% in all other tissues analyzed. They also provide evidence for the involvement of AID in active demethylation of early PGCs as AID deficient mice show a small but significant increase in methylation at 13.5pdc compared to WT.

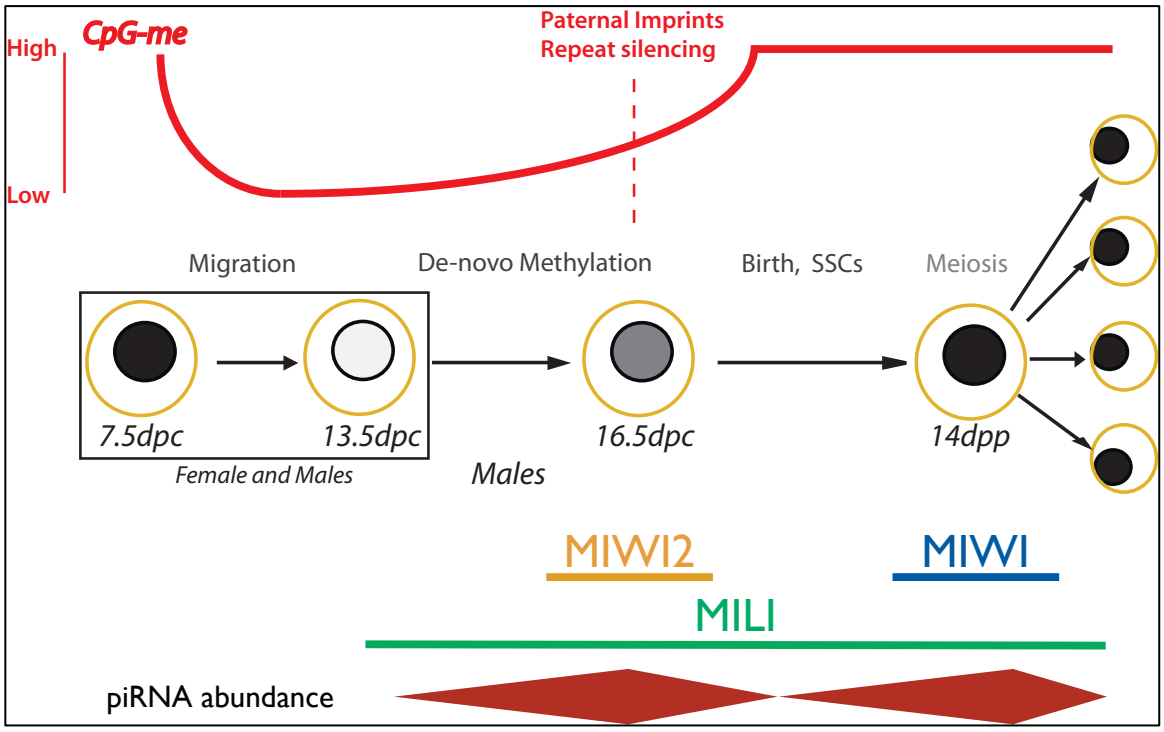


Figure 1.6: Timing of *de novo* methylation during male germ cell development. During PGC migration to the gonads, DNA methylation levels (red) are erased in both male and female PGCs (black box). During male gametogenesis, *de novo* methylation occurs between 13.5dpc and SSC establishment post-birth. Methylation levels remain high throughout meiosis. The relative timing of PIWI protein expression and piRNA abundance is also depicted (see section 1.3).

Male PGCs start de novo methylation soon after their entry into embryonic gonads. In contrast, the genome of female PGCs remains hypomethylated until arrested oocytes resume growth and meiosis after birth. In male PGCs, DNMT3L and DNMT3a start to accumulate at ~13.5, peak by ~15.5 and revert to their somatic level at ~18.5dpc for 3L and 6 days post birth for 3a. In contrast,

DNMT3B and DNMT1 are found at much lower levels throughout PGC development (La Salle et al., 2004). Consequently, transposon sequences and imprinted loci initiate a rapid wave of de novo methylation between 14.5 and 18.5dpc. However, *de novo* methyl-marks are continually deposited until day 2 post-birth when PGCs colonize their niche as SSCs (Walsh et al., 1998; Ueda et al., 2000; Kato et al., 2007; Lees-Murdock et al., 2003; Kuramochi-Miyagawa et al., 2008). Methylation profiles are believed to undergo little if any changes during male meiosis. DNMT3B and DNMT1 levels are elevated in SSCs and reduce both during meiosis and spermiogenesis as sperm nuclei become transcriptionally silent. Nucleosomes a exchanged for protamines late in spermiogenesis. Hammoud and colleagues recently showed in human sperm that about 4% of the genome in still packaged in nucleosome retaining domains (Hammoud et al., 2009). These nucleosomes consist of either canonical or histone variants, such as the testes-specific histone H2B (TH2B). Interestingly, regions protected from protamine-exchange were also shown to be under methylated, further connecting the methylation status of the genome to higher order chromatin structures.

1.3: Germ Cell Associated Small RNA Pathways

In addition to protein coding messenger RNAs, eukaryotic cells express a wide range of non-coding transcripts ranging from 19-30bp, for small regulatory RNAs, to several hundred base pairs (e.g. long intergenic non-coding RNAs, lincRNAs). They often engage in transcriptional and post-transcriptional gene regulation, relying on RNA:DNA or RNA:RNA interactions (reviewed by Mercer et al., 2010 and Ghildiyal et al., 2010). During metazoan development, germ cells are characterized by the expression of a distinct RNAi pathway involving PIWI proteins and their 24-30bp associated piRNAs (PIWI-interacting RNAs). Male mice deficient for piRNAs are infertile, as germ cells fail to undergo productive meiosis (Deng et al., 2002; Kuramochi-Miyagawa et al., 2001 and 2004; Carmel et al., 2007). It is believed that the interplay between piRNAs and *de novo* methylation drives the re-methylation of retro-transposons during PGC development (Aravin and Bourc'his 2008; Aravin and Hannon 2008). Considering the extreme sequence diversity of piRNAs, and the high abundance of retro-elements in mammalian genomes, this model provides an interesting framework to study the connection between transposon dynamics and germ cell development.

The previous section highlighted the dynamic character of DNA methylation patterns during germ cell development at both single copy loci and repeated sequences. However, beyond 12.5dpc, PGCs are transcriptionally active, raising the question of what the transcriptional statuses of repeats,

piRNAs and other coding and non-coding transcripts are prior to *de novo* methylation. The following points cover the known biogenesis pathways leading to the production of small interfering RNAs in various cellular contexts, and focus on the newly characterized piRNA pathway in germ cells, mode of action of which still remains uncharacterized.

1.3.1 Overview of RNAi

From the characterization of silencing phenotypes induced by the introduction of multi-copy transgenes in plants, or the cloning of the first micro-RNA (miRNA) in C.elegans (lin-4, Lee et al., 1993), RNA interference has been shown to be a highly conserved biological pathway used to dampen gene expression via the targeting of cellular mRNA and, sometimes, via transcriptional inhibition. At the core of this pathway lies the interaction between a small RNA and an Argonaute protein. Whereas small RNAs guide this RNA Induced Silencing Complex (RISC) to the cognate targets using base pairing, Argonaute proteins mediate silencing via mRNA cleavage, translational inhibition or chromatin remodeling (on the latter see Volpe and Martienssen 2011). Different classes of small RNA have now been characterized in unicellular and multicellular eukaryotes as well as in prokaryotes and archea (Hannon 2002; Marraffini and Sontheimer 2010). Over the past 20 years, small RNAs have been shown to regulate development, cell cycle, epigenetic inheritance and hundred small RNAs have been show to be mis-expressed in various diseases including

cancer and are currently being used for diagnosis and therapy (Silva et al., 2004).

In metazoans, small interfering RNAs (siRNA) and miRNAs are typically 19 to 23nt in size and constitute the most ubiquitously expressed classes of small RNAs. A third class, termed piRNAs (for PIWI interacting RNAs, also known as repeat-associated-small-interfering RNA, Aravin et al., 2003), is known to be exclusively expressed in germ cells and range form 24 to 31nt. miRNAs and siRNA share common steps in their biogenesis as both of them require the endonuclease III activity of Dicer to produce fully functional single stranded small RNAs from double stranded (ds) RNA structures (Bernstein et al., 2001; reviewed in Carmell and Hannon, 2004). They differ mostly in the nature of their precursors: siRNA are produced from long ds-RNA whereas miRNA are processed from a single stranded primary transcript that folds into a stem loop structure cleaved by the nuclear RNAse III enzyme Drosha (Lee et al., 2004). In contrast, piRNAs are thought to be initially processed from long primary single stranded transcripts by an unknown, Dicer independent, mechanism (Vagin et al., 2006; Houwing et al., 2007).

Following these processing steps, single stranded small RNAs are loaded into an Argonaute protein forming a functional RISC. Argonautes constitute a large conserved RNA binding protein family. They possess two related domains: the PAZ domain interacting with the 3' end of a small RNA and the PIWI domain, interacting with the 5'end, which forms the RNAseH catalytic domain (Song et al., 2004, Liu et al., 2004). Based on sequence analysis, Argonaute proteins can be

separated into two closely related clades: the Ago and the Piwi subfamily (Carmell et al., 2002). Argonautes in the Ago clade are most similar to ARGONAUTE1 found in *Arabidopsis thaliana*, and primarily interact with miRNAs and siRNA. The Piwi clade however, shares most similarities with the drosophila PIWI protein.

In mammals, the association of Argonaute proteins with a small RNA is primarily occurring in the cytoplasm where targeting is also thought to happen. RISC can mediate target cleavage 10nt away from the 5' end of the small RNA when perfectly matched. However, in most cases, mismatched interactions leads to translational inhibition (Olsen and Ambros, 1999). When engaged in silencing, RISC is found in processing and stress bodies throughout the cytoplasm (Liu et al., 2005a/b, reviewed by Leung and Sharp 2006). These bodies contain numerous structural and catalytic proteins promoting the silencing function of RISCs associated with miRNAs, siRNAs and even piRNAs (on the latter see Siomi and Aravin 2011).

1.3.2 PIWI Proteins and Germ Cell Specific RNAi

In contrast to the ubiquitously expressed Agos, the Piwi clade is restricted to the germ line. Preliminary studies in *drosophila* revealed that the PIWI proteins, *piwi* and *aubergine*, are essential for gametogenesis and germline stem cell renewal (Cox et al., 1998 and 2000; Schmidt et al., 1999). Disruption of the *piwi* gene causes improper stem cell divisons in the male and female germline,

and disrupts gametogenesis progression preventing cyst formation and meiosis. The mouse and human genomes encode three conserved PIWI proteins: MIWI/HIWI (PIWIL1), MILI/HILI (PIWIL2) and MIWI2/HIWI2 (PIWIL4). During mouse gametogenesis PIWI proteins are detected in PGCs as early as 12.5-13.5 dpc. In the male germ-line, MILI is expressed throughout spermatogenesis until the round spermatid stage following meiosis, first occurring around 20dpp (Figure 1.6). Similarly, MILI is detected in the female germ-line soon after PGC migration and is continuously expressed in both arrested and growing oocytes (Aravin et al., 2008). MIWI2 expression starts at ~15.5dpc and stops prior to SSC establishment after birth. Finally, MIWI expression starts at the pachytene stage of meiosis, at around 14 dpp, and stops at the round spermatid stage. Cellular localization of these proteins addressed by immuno-fluorescence or using GFP-fused transgenic animals showed that while MILI and MIWI are found exclusively in the cytoplasm, MIWI2 can shuttle to the nucleus when loaded with piRNAs (Aravin et al., 2008).

Disrupting either *mili* or *miwi* genes both causes male sterility. However, while *mili*-KO leads to in meiotic arrest at the pachytene stage, *miwi*-KO animals undergo meiosis properly, but germ cells fail to develop beyond the haploid round spermatid stage (Deng et al., 2002; Kuramochi-Miyagawa et al., 2001 and 2004). These data suggest that MILI and MIWI have non-redundant functions in spermatogenesis. Recently, two *miwi2* deficient mice were independently generated and revealing that MIWI2 is essential for SSC maintenance, with MIWI2 loss resulting in a progressive depletion of undifferentiated germ cells in

adult testes (Carmell et al., 2007, Kuramochi-Miyagawa et al., 2008). *miwi2*-KO animals also fail to undergo full meiosis and arrest cell division before the pachytene stage. Considering MIWI2 expression during embryogenesis, this effect is certainly due to defects occurring before meiosis.

In addition to germ line defects, it has been shown that *Drosphila piwi* and *aub* and mouse *miwi2* and *mili* mutant animals derrepress transposons in the germ line (Sarot et al., 2004; Brennecke et al., 2007; Carmell et al., 2007; Aravin et al., 2007). It was therefore suggested that PIWI proteins might play a critical role in transposon silencing in the germ line, and that defects observed during gametogenesis are mainly caused by the reactivation of transposable elements.

1.3.3 PiRNA Biogenesis during Male Germ Cell Development

In mammals, two distinct waves of piRNA expression have been described so far (Girard et al., 2006; Aravin et al., 2007 and 2008). The first wave of piRNA expression peaks at around the time when MIWI2 and MILI are co-expressed in PGCs (between 15.5dpc and birth) and those piRNAs are referred to as embryonic piRNAs. The second wave occurs during MIWI and MILI co-expression when cells enter the first division of meiosis; those are referred to as meiotic piRNAs (Figure 1.6). The purification of PIWI protein complexes and subsequent cloning and sequencing of their associated piRNAs revealed that these two waves differ mostly in the type of transcripts they can target.

Embryonic piRNAs are strongly, but not exclusively, enriched for transposable elements sequences, including all three classes of retro-

transposons (LINEs, SINEs and LTRs, Figure 1.7). Consequently, a large fraction of these piRNAs maps to multiple locations in the genome, offering the potential for robust and redundant silencing of transposons during epigenetic reprogramming of PGCs. Unlike the strong strand preference seen for the different PIWI complexes in *drosophila* (Brenecke et al., 2007), MILI and MIWI2 only show a slight bias toward sense and anti-sense piRNAs respectively (Aravin et al., 2008). In addition, MILI and MIWI2 have distinct piRNA size preferences: 26-27nt for MILI and 28-29nt for MIWI2. Interestingly, in a MILI mutant background MIWI2 fails to load piRNAs and to localize in the nucleus, suggesting that MILI functions upstream of MIWI2 (Aravin et al., 2008, Kuramochi-Miyagawa et al., 2008). Meiotic piRNAs are enriched for sequences mapping to large un-annotated portions of the genome both in rat, mouse and human (Lau et al., 2006, Girard et al., 2006, Aravin et al., 2006). MILI and MIWI bind virtually the same piRNA sequences though each complex is associated with its characteristic piRNA size - ~26nt for MILI and ~30nt for MIWI. The lack of information about the cellular function of the transcripts regulated by meiotic piRNAs makes it challenging to study their function, especially in an attempt to explain the MIWI sterility phenotype (Deng et al., 2002). Finally, between these two waves, MILI is continuously expressed in late PGCs, SSCs and pre-meiotic spermatogonia. Despite the absence of any partner, MILI is loaded with a “transition” population switching from transposons rich piRNAs to a much more gene enriched population in SSCs and ultimately meiotic piRNAs as cells enter differentiation (Aravin et al., 2007).

There is still much mystery regarding how piRNAs are processed from primary transcripts. Key signatures of piRNA sequences include a 5'Uracil and a 2'-O-methyl group at the 3' end, both of which have been shown to be important for their functional association with PIWI proteins (Horwich et al., 2007; Saito et al., 2007; Kirino et al., 2007a/b). The analysis of piRNA sequences uniquely mapping to the genome revealed the existence of large piRNA loci, 10kb to a 100kb in size, producing both sense and antisense reads (Brenecke et al., 2007; Girard et al., 2006; Aravin et al., 2006, 2007 and 2008). These large piRNA clusters produce about 10-20% of all embryonic piRNAs (the remaining piRNAs being produced from individual transposons or genes) and more than 90% of meiotic piRNAs (see Figure 1.7A). Whereas embryonic piRNA clusters can generate primary transcripts on both strands, meiotic piRNA clusters are transcribed from one strand but often show a typical bidirectional structure with a central promoter firing in opposite directions. Several lines of evidence from *drosophila* and mouse suggest that primary piRNA 5'end production requires the activity of a phospholipase D, MITOPLD in mouse and Zucchini in *drosophila* (Watanabe et al., 2011; Haase et al., 2010). Mutant animals for these proteins accumulate cluster transcripts and display sterility phenotypes. Recently, an *in-vitro* study using insect cells lysates, provided evidence that 3' ends of piRNAs are generated by the trimming of long transcript loaded into PIWIs by an unknown 3' to 5' exonucelase (Kawaoka et al., 2011). These observations provide an attractive model to explain the size and sequence preference of piRNAs bound to PIWI proteins.

Interestingly, part of primary transcript cleavage is dependent on the slicer activity of PIWI proteins. This slicer-mediated cleavage is believed to participate in an amplification loop, named the ping-pong loop, whereby a primary piRNA directs the cleavage of a complementary transcript. The 5'end of the cleaved product constitutes the 5'end of a new piRNA (or secondary piRNA). Secondary piRNAs can, in turn, mediate the cleavage of a complementary transcript generating the 5'end of a new piRNA closing the loop on itself (Figure 1.7C) (Brenecke et al., 2007; Aravin et al., 2008). Recently, catalytically inactive mutants of all 3 PIWI proteins were generated in mouse (De Fazio et al., 2011; Reuter et al., 2011). These studies revealed that while MILI and MIWI catalytic activity is critical for piRNA amplification and male fertility (recapitulating null phenotypes), MIWI2 catalytic mutants are fertile and transposon silencing is established normally, decoupling its silencing function from slicer cleavage. However, the catalytic domain of MILI has been shown to be critical for MIWI2 loading and silencing function, confirming previous observation made in null animals. Interestingly, MILI slicing activity is shown to be essential for LINE elements silencing, while exhibiting little effect on LTRs, suggesting that these two classes of retro-transposons engage in slightly different piRNA "pathways".

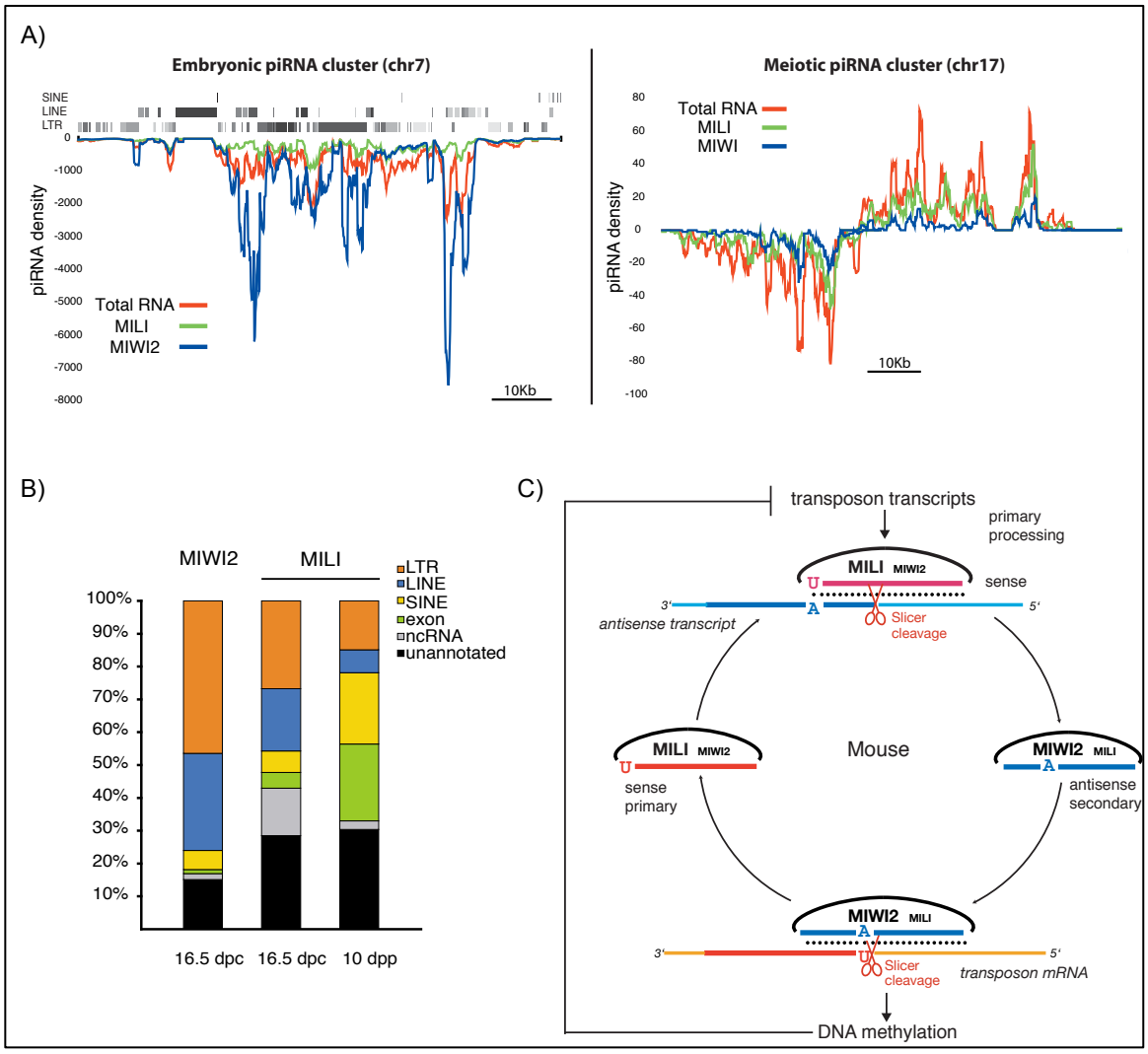


Figure 1.7: Mouse piRNA biogenesis and silencing function.
A) Two representative examples of an embryonic (on chr7, left panel) and a meiotic piRNA cluster (on chr17, right panel). Whereas embryonic piRNA clusters are enriched for transposon sequences, meiotic clusters are not associated with any annotation. The densities of piRNAs mapping across the clusters are shown both for PIWI proteins and total RNA cloned at these stages. B) Annotation of MIWI2 and MILI bound piRNAs in 16.5dpc and 10dpp male gonads (displayed as percent of total read mapped). C) Proposed ping-pong model occurring during mouse PGC maturation. *B) and C) are adapted from Aravin et al., 2008.*

1.3.4 Link between piRNAs and *De Novo* Methylation in Male PGCs

As mentioned above MILI and MIWI2 mutants display a strong reactivation of transposable elements, with LINE-1 and IAP transcripts accumulating to high levels in mutant meiotic germ cells (Carmel et al., 2007; Aravin et al., 2007). In addition, these mutants display a reduction in transposon *de novo* methylation in PGCs as early as 16.5dpc (Kuramochi-Miyagawa et al., 2008). They also phenocopy *dnmt3L* mutants (Bourc'his et al., 2001 and 2004) suggesting that the piRNA and the *de novo* methylation machinery might help each other in the establishment of transposon methylation during PGC maturation. More evidence for an existing epistatic relationship between de novo methylation and piRNAs came from the study of DNMT3l loss on piRNA biogenesis, where, in this context, transposon associated sense primary piRNA levels were dramatically increased (Aravin et al., 2008). A recent study also showed that piRNAs help to establish *de novo* methylation over the paternal allele of a known repeat associated differentially methylated region (DMR) controlling the imprinted expression of *Rasgrf1* (Watanabe et al., 2011). Taken together, these data suggest an elegant model whereby MILI and MIWI2 complexes act upstream of DNMT3L, which itself acts upstream of DNMT3a to establish transposon *de novo* methylation in germ cells. However, a detailed analysis of the true mechanistic connections between piRNAs and *de novo* methylation is still lacking.

1.4 Epigenetic Inheritance and Evolution some Open Questions.

The study of epigenetic pathways has not only refined our theories on phenotypis determination and robustness, but it has also profoundly challenged our concept of development. Despite the last 20 years of astonishing molecular characterization of chromatin based epigenetic mechanisms, we are witnessing a small revolution in the field with the use of next-generation sequencing technologies. We can now ask questions to unprecedented scales and within unexplored portions of genomes. The thesis work presented here was largely aimed at using these next-generation sequencing tools to tackle relevant open questions in the field.

1) Looking at germ cell small RNAs and chromatin dynamics: which components are innate and which are adaptive relative to genomic sequence?

2) Comparing the epigenetic reprogramming of germ cell and ES cell: what can we learn about epigenetic inheritance and determination during development?

3) Comparing epigenetic states in closely related species: does the epigenome affect genome evolution and vice-versa?

Chapter 2: Results

2.1 Study of a Transgenic piRNA Cluster in Meiotic Mouse Germ Cells.

2.1.1 Résumé en Français.

Chez les métazoaires, une classe de petits ARNs associée aux protéines de la famille PIWI est spécifiquement exprimée au cours du développement de la lignée germinale. Ces petit ARNs ou PIWI associated RNAs (piRNAs), répriment leurs cibles au niveau transcriptionnel et post-transcriptionnel par interférence ARN. Les animaux mutants pour les protéines PIWI sont stériles et présentent notamment une perte progressive des cellules souches nécessaires au maintient des tissus reproductifs. Au cours de la gamétogénèse embryonnaire et post-embryonnaire, ces petits ARNs participent à la répression de retro-transposons, de gènes codants et non-codants, ainsi que de vastes régions non-annotées du génome. De nombreuses questions subsistent quand aux mécanismes permettant la biogénèse des piRNAs. Les piRNAs semblent être produits à partir de transcrits primaires simple brin, sens ou antisens à leurs cibles. Ces transcrits peuvent être codés par des régions pouvant atteindre plus de 100Kb appelées « piRNA clusters ». Cependant, les signaux distinguant les transcrits entrant dans la voie des piRNAs d'autres transcrits cellulaires restent inconnus.

Dans l'étude qui suit, nous avons analysé chez la souris et la drosophile le comportement de différents piRNA clusters hors de leurs contextes naturels. La génération de piRNAs à partir de ces clusters transgéniques se produit de manière similaire aux régions endogènes, indiquant que la localisation génomique et le nombre de piRNA clusters par génome n'ont qu'un effet réduit sur leur production primaire. De plus, nous démontrons que l'introduction de séquences exogènes au sein de ces clusters transgéniques n'interrompt pas la production de piRNAs en amont ou en aval et que ces séquences sont elles même clivées en piRNAs présentant les caractéristiques typiques des piRNAs endogènes. Toutefois, la position ainsi que le type de cluster modifié influencent localement l'abondance et la distribution de ces piRNAs.

2.1.2 Specific contribution to the publication

This manuscript combines the analysis of 3 drosophila and 2 mouse constructs. I produced and analyzed the mouse part of this study. My specific contributions were: generation of BAC constructs GFP-NEO tagged, BAC preparation, establishment and maintenance of transgenic lines, small RNA cloning on both total and immuno-precipitated RNAs, sequencing annotation and analysis of the small RNA libraries presented here. I also contributed to the writing of the manuscript.

Dr. Sang Young Kim, from the Cold Spring Harbor Laboratory mouse facility, was in charge of BAC DNA injections into mouse zygotes.

Felix Muerdter, Ivan Olovnikov and Nikolay V. Rozhkov carried out the drosophila section of the manuscript.

2.1.3 Publication reference

Felix Muerdter*, Ivan Olovnikov*, Antoine Molaro*, Nikolay V. Rozhkov*, Benjamin Czech, Assaf Gordon, Gregory J. Hannon, and Alexei A. Aravin. Production of artificial piRNAs in flies and mice. ***RNA*** *January 2012 18: 42-52; Published in Advance November 17, 2011, doi:10.1261/rna.029769.111*

* Equal contribution

Production of artificial piRNAs in flies and mice

FELIX MUERDTER,[1,2,5] IVAN OLOVNIKOV,[3,4,5] ANTOINE MOLARO,[1,5] NIKOLAY V. ROZHKOV,[1,5] BENJAMIN CZECH,[1,2] ASSAF GORDON,[1] GREGORY J. HANNON,[1,6] and ALEXEI A. ARAVIN[3,6]

[1]Howard Hughes Medical Institute, Watson School of Biological Sciences, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA
[2]Zentrum für Molekularbiologie der Pflanzen, Entwicklungsgenetik, University of Tübingen, 72076 Tübingen, Germany
[3]California Institute of Technology, Division of Biology, Pasadena, California 91125, USA
[4]Institute of Molecular Genetics, Russian Academy of Sciences, 123182 Moscow, Russia

ABSTRACT

In animals a discrete class of small RNAs, the piwi-interacting RNAs (piRNAs), guard germ cell genomes against the activity of mobile genetic elements. piRNAs are generated, via an unknown mechanism, from apparently single-stranded precursors that arise from discrete genomic loci, termed piRNA clusters. Presently, little is known about the signals that distinguish a locus as a source of piRNAs. It is also unknown how individual piRNAs are selected from long precursor transcripts. To address these questions, we inserted new artificial sequence information into piRNA clusters and introduced these marked clusters as transgenes into heterologous genomic positions in mice and flies. Profiling of piRNA from transgenic animals demonstrated that artificial sequences were incorporated into the piRNA repertoire. Transgenic piRNA clusters are functional in non-native genomic contexts in both mice and flies, indicating that the signals that define piRNA generative loci must lie within the clusters themselves rather than being implicit in their genomic position. Comparison of transgenic animals that carry insertions of the same artificial sequence into different ectopic piRNA-generating loci showed that both local and long-range sequence environments inform the generation of individual piRNAs from precursor transcripts.

Keywords: piwi; noncoding RNA; piRNA

INTRODUCTION

In several animals, including *Drosophila* and mammals, piRNAs have been shown to form the core of a small RNA-based innate immune system that recognizes and represses mobile elements (Saito et al. 2006; Vagin et al. 2006; Aravin et al. 2007a; Brennecke et al. 2007; Gunawardane et al. 2007; Malone and Hannon 2009; Siomi et al. 2011). This function is essential for proper germ-line development, and mutations in the piRNA pathway lead to male and/or female sterility (Cox et al. 2000; Harris and Macdonald 2001; Li et al. 2009; Malone and Hannon 2009). In essence, piRNAs play a major role in defining genomic content as being transposon related; piRNAs comprise a catalog of transposon sequences that an organism has defined as targets for repression (Brennecke et al. 2007). Omission from that catalog can mean that an element escapes repression. In the case of flies, the lack of an effective piRNA-based definition for the *I-* or *P-element* in some strains means that introduction of even this single transposon can lead to highly penetrant sterility (Pelisson 1981; Rubin et al. 1982; Brennecke et al. 2008).

Sequencing of piRNA populations has revealed their extreme diversity; literally, millions of distinct piRNA sequences can be identified in a single individual (Aravin et al. 2006, 2007b; Girard et al. 2006; Brennecke et al. 2007; Houwing et al. 2007; Lau et al. 2009). Genomic mapping indicates that piRNAs arise from three different types of loci. First, the dominant source of piRNAs can be found in so-called piRNA clusters (Aravin et al. 2006, 2007b; Brennecke et al. 2007). These loci range from a few kilobases to >200 kb in size. They are often strongly enriched in transposon sequences, in accord with a function of the piRNA pathway in transposon control (Vagin et al. 2006; Brennecke et al. 2007; Gunawardane et al. 2007). In the majority of cases, clusters generate a mixture of small RNAs, with some sense and some antisense to each targeted transposon. Second, piRNAs can be derived from protein-coding genes, with these almost invariably being sense species from 3′ UTRs (Aravin et al. 2008; Robine et al.

[5]These authors contributed equally to this work.
[6]**Corresponding authors.**
E-mail aaa@caltech.edu.
E-mail hannon@cshl.edu.
Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.029769.111.

2009; Saito et al. 2009). It is as yet unclear whether a single transcript isoform can be either translated into protein or processed into small RNAs or whether a specific transcript variant serves as a piRNA precursor. Only a few genes give rise to piRNAs, and these do not show uniformly high expression, suggesting that some specific determinant or motif, rather than a high-transcript abundance marks specific genes for processing. Third, piRNAs can arise from dispersed, euchromatic transposon copies (Brennecke et al. 2007, 2008; Aravin et al. 2008). These are often full length and close to consensus, representing the potentially active representatives of each transposon family.

The three types of piRNA generative loci produce small RNAs through two different mechanisms. piRNA clusters and genic loci generate "primary" piRNAs, which appear to be sampled from long, single-stranded transcripts through the action of an unknown nucleolytic machinery (Aravin et al. 2006, 2007b; Brennecke et al. 2007; Malone et al. 2009). Abundant primary piRNAs share no apparent sequence or structural motifs except for the presence of a 5′ terminal U residue (1U), which may reflect a binding preference of some Piwi family proteins. Secondary piRNAs are produced through a slicer-dependent mechanism, termed the ping-pong cycle and have a characteristic bias for an A at position 10 (paired with the 1U in the primary piRNA) (Brennecke et al. 2007; Gunawardane et al. 2007).

Combined analysis of piRNA sequences and animals bearing mutations in piRNA pathway components has led to a model for the role of these small RNAs (Malone and Hannon 2009; Saito and Siomi 2010; Senti and Brennecke 2010; Siomi et al. 2011). piRNA clusters produce a multitude of individual piRNAs, and the sequence content of piRNA cluster defines sequences of mature piRNAs generated from it. With the notable exception of pachytene piRNAs that are expressed during male meiosis in mouse, piRNA clusters in both flies and mice are highly enriched in transposable element sequences. The sequence content of the piRNA clusters determines the capacity of the system to respond to a given element, in essence comprising an organisms' evolving molecular definition of transposons. Inherent in this scenario is the ability of the system to adapt to colonization by new elements by incorporating their sequence into a piRNA cluster. A clear example can be found in the *P-element*, which swept through global *Drosophila melanogaster* populations after the sequestration of common laboratory strains (Rubin et al. 1982). Laboratory strains have no ability to repress the *P-element*. In retrospect, studies of strains with natural or acquired *P-element* resistance suggested that integration of the element into a piRNA cluster was key to its control (Ronsseray et al. 1991, 1996, 2003).

Here, we sought to test whether the ability to translate new genomic content into small RNAs was a general characteristic of piRNA loci in flies and mice. We find that clusters can be programmed to produce artificial piRNAs (apiRNAs). Furthermore, we were able to separate functional piRNA clusters from their native genomic locations, indicating that the clusters themselves contain sufficient information to funnel their RNA products into the piRNA biogenesis pathway. We made use of marked transgenic clusters that carry insertions of the same sequence into different contexts to evaluate the features that lead to the production of individual piRNA species. We find that critical determinants lie both in the local and long-range sequence environments of the piRNA cluster.

RESULTS AND DISCUSSION

The current model for acquiring piRNA-dependent resistance against new transposon invasion implies that insertion of active transposons into an existing piRNA cluster leads to the generation of new piRNA species and enables element repression. This model suggests that any sequence, if inserted into a piRNA cluster, will lead to the generation of new piRNAs. Though attractive, this model has not been rigorously tested. Acquisition of natural resistance against transposable elements by transposition into piRNA clusters is difficult to study in an experimental setting. However, this scenario can be modeled using transgenes carrying new sequence information within a piRNA-generating locus.

Over the years, large collections of *Drosophila* stocks have been produced that carry transgenes integrated randomly throughout the genome. We took advantage of these tools by searching for integration events in native piRNA clusters. The line P{lArB}A171.1F1 (also known as P-1152) has a 18.3-kb construct P{lArB} integrated into a telomeric piRNA cluster on the X-chromosome (chromosomal location 1A) (Wilson et al. 1989; Roche and Rio 1998). The P{lArB} transgene contains sequences derived from the *hsp70*, *Adh*, and *rosy* genes of *D. melanogaster* and a bacterial *lacZ* gene. Unlike P{lArB} insertions in other genomic sites, P-1152 has unusual properties. It is able to suppress the expression of other *lacZ* transgenes in germ cells, a phenomenon termed *trans*-silencing (Fig. 1A; Supplemental Fig. S1; Ronsseray et al. 1991). The P{lArB} insertion in P-1152 is mapped to the Telomere Associated Sequence (TAS) repeats that produce abundant piRNAs from both genomic strands. These piRNAs are loaded into Piwi, Aub, and Ago3 in the germ cells of *D. melanogaster* ovaries (Brennecke et al. 2007). Aub and Ago3-loaded piRNAs derived from TAS repeats display the characteristic features of the ping-pong amplification cycle, including a prevalent 10-nt 5′ overlap of sense and antisense species and an enrichment for an A at position 10 of secondary piRNAs. The *trans*-silencing properties of P-1152 transgene and the association of these properties with its localization in the piRNA cluster suggested that insertion of *lacZ* into an existing piRNA cluster led to the generation of new anti-*lacZ* piRNAs that are able to suppress cognate transcripts in germ cells. Indeed, the presence of small

RNAs complementary to *lacZ* was recently demonstrated using RNAse-protection assay in ovaries of the P-1152 line (Todeschini et al. 2010).

To analyze more deeply any artificial piRNAs derived from the P{lArB} transgene, we sequenced small RNAs from ovaries of the P-1152 line, examining a size range from 18 to 29 nt. This includes piRNAs, siRNAs, and miRNAs. P{lArB} generated abundant small RNA species that mapped to both genomic strands (Fig. 1B). Their size profile indicated that the majority were likely piRNAs, ranging from 23 to 27 nt, while a minor fraction corresponded to 21-nt endo-siRNAs that are also a product of bidirectionally transcribed piRNA loci (Fig. 1C; Czech et al. 2008; Lau et al. 2009). Further analysis confirmed that the 23- to 27-nt RNAs were genuine piRNAs that could be separated into primary (1U-biased) and secondary (10A-biased) populations (Fig. 1D; data not shown). Transgene piRNAs mapping to opposite genomic strands tended to have a 10-nt overlap between their 5′ ends that is a characteristic feature of the ping-pong cycle (Fig. 1E). Notably, P{lArB} contains the only *lacZ* sequence information in the P-1152 strain. Since signatures of the ping-pong cycle were evident for *lacZ*-derived piRNAs, this demonstrates unequivocally that cluster transcripts derived from the plus and minus genomic strands can participate in the piRNA amplification loop. Native *Adh* and *rosy* transcripts are not processed into piRNAs in wild-type flies (data not shown). Therefore, it is unlikely that any specific signals that trigger piRNA processing might be present in these genes. Moreover, bacterial sequences are unlikely to have evolved as a trigger for piRNA production. Thus, our results indicate that, when present in the context of a piRNA cluster, virtually any sequence can serve as a substrate for piRNA biogenesis. We confirmed previous observations that the P{lArB} transgene inserted in TAS is able to silence *lacZ* expression from separate, euchromatic locations (Supplemental Fig. S1), demonstrating that artificial anti-*lacZ* piRNAs are functional and able to silence transcripts that share sequence content in *trans*.

piRNAs are processed from the entire P{lArB} transgene independently of the origin of the inserted fragments; both *D. melanogaster* and bacterial sequences generate piRNAs with similar efficiency (Fig. 1B). Throughout the construct there are approximately twofold more piRNAs derived from the plus than from the minus genomic strand independently of the orientation of the genes within the construct, just as is observed for native components of the cluster. For example, *Adh* and *rosy* have different orientations, but for both fragments the majority of piRNAs are mapped to the plus genomic strand. RT–PCR shows that *rosy* transcripts are present in ovaries of P-1152 females, but absent in wild-type flies or flies that have a P{lArB} insertion outside of the piRNA cluster (Supplemental Fig. S2), indicating that *rosy* expression is dependent on insertion of P{lArB} into TAS. Overall, both the distribution of piRNAs along P{lArB} transgene and RT–PCR results suggest that transcript of both plus- and minus-strand RNAs, which are processed to piRNAs, initiates outside of the transgenic construct, likely within adjacent TAS sequences.

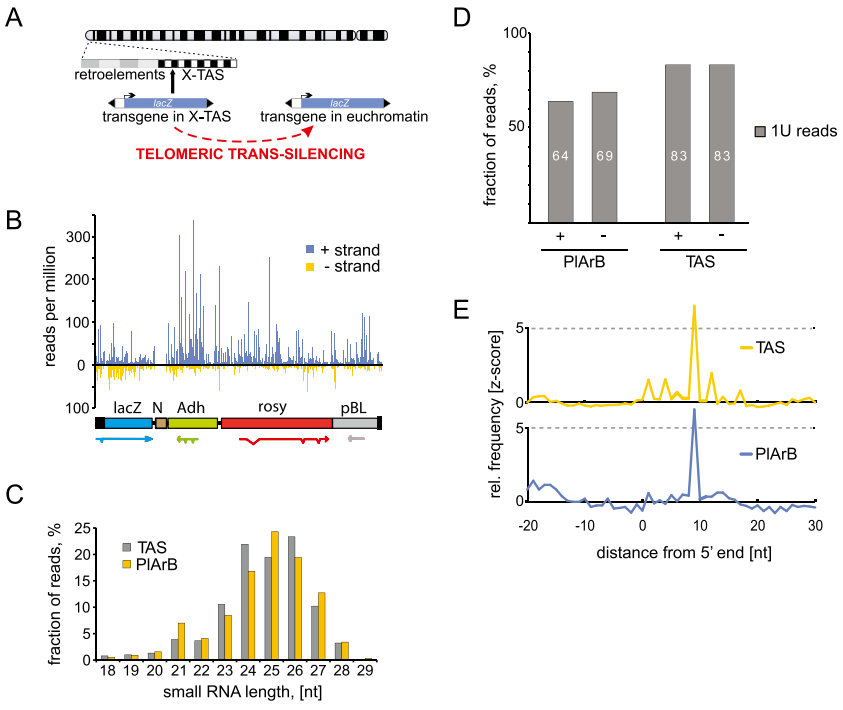


FIGURE 1. Production of artificial piRNAs (apiRNAs) from the *Drosophila* X-TAS cluster. (*A*) The P{lArB} insertion into the X-TAS cluster is shown schematically along with an illustration of *trans*-silencing. (*B*) *Below* is a schematic of the P{lArB} insert with the inferred structures of the transcripts it can produce (see text). N is an area where the sequence is unknown. *Above* is a plot of piRNA read frequencies along the plus and minus strands (indicated) of the element. (*C*) Small RNA lengths are plotted as a fraction of reads for TAS and for the inserted element. (*D*) Fractions of reads beginning with a 5′ U are plotted for the P{lArB} and TAS plus and minus strands. (*E*) The degree of 5′ overlap for piRNAs from the plus and minus strands for P{lArB} and TAS were quantified and plotted as relative frequencies (Z-scores). The spike at position 9 is a signature of the ping-pong amplification cycle.

Mapping of piRNAs to P{lArB} revealed that intronic sequences present within *Adh* and *rosy* gave rise to piRNA from both genomic strands. Even when present in the sense orientation, where the intron could have been removed by the splicing apparatus, piRNA levels remained comparable in adjacent intronic and exonic regions. The generation of piRNA from intronic sequence is unexpected, as primary piRNA bio-

genesis is thought to occur in the cytoplasm and has been linked to specific cytoplasmic bodies, e.g., nuage and Yb bodies, which concentrate components such as zucchini and armitage, which are implicated in piRNA processing (Tomari et al. 2004; Lim and Kai 2007; Pane et al. 2007; Malone et al. 2009; Haase et al. 2010; Olivieri et al. 2010; Saito et al. 2010; Qi et al. 2011). Furthermore, genic piRNAs that are processed from mRNA of protein-coding genes in *Drosophila* and mice are mapped almost exclusively to exonic sequences (Aravin et al. 2008; Robine et al. 2009; Gan et al. 2011). To reconcile these disparities, we searched explicitly for piRNAs that crossed predicted exon–exon junctions, since these must arise from spliced mRNAs. We did detect a few such small RNAs for *rosy* and *Adh*, coming only from the genomic strand with the intron in the appropriate orientation for splicing to occur. Considered together, these data suggest a model in which piRNA biogenesis normally occurs following intron removal, but that recognition of some RNA processing signals might be suppressed when they are present within a piRNA cluster. In this regard, strand-specific RT–PCR indicated that more than half of sense-oriented *rosy* transcripts are not spliced in P-1152 ovaries (Supplemental Fig. S2). Suppression of conventional RNA processing signals within piRNA clusters would make sense in many ways, since the insertion of a new element would often bring at least a polyadenylation signal, which under normal circumstances could negate the production of piRNAs downstream from that site by terminating transcription or preventing the export of piRNA precursors.

Generation of artificial piRNAs by insertion of a new sequence into a piRNA cluster provides a molecular tag that allows the monitoring of cluster function even if the native, nontagged cluster is present in the same genome. We exploited this fact to test whether the presence of piRNA clusters at precise genomic positions was important to their function.

In flies, piRNA clusters occur mainly at the boundaries between heterochromatin and euchromatin, particularly in pericentromeric regions (Brennecke et al. 2007). In mammals, piRNA clusters that are expressed in meiotic cells occur in strictly syntenic positions, even though the sequence content of these loci is not conserved (Aravin et al. 2006; Girard et al. 2006; Lau et al. 2006). These observations have strongly suggested that the genomic context of piRNA clusters might be key to their function. Precedent can be drawn from plants and fission yeast, where small RNAs are generated from loci whose function relies upon the presence of normally repressive chromatin marks (Huisinga and Elgin 2009; Lahmy et al. 2010). In turn, the repressive chromatin marks themselves are maintained by small RNA-directed complexes, closing the cycle. To determine whether specific chromatin environments, which are a property of the genomic context of piRNA clusters, are essential for piRNA production, we created ectopic insertions of tagged piRNA clusters in non-native sites.

As one test of the aforementioned hypothesis, we examined the position dependence of the *flamenco* cluster in *Drosophila* (Fig. 2A). *Flamenco* is present at the boundary between euchromatin and pericentromeric heterochromatin on the *Drosophila* X chromosome, and its position proximal to the *DIP1* gene is conserved through at least 12 M years of *Drosophila* evolution (Sarot et al. 2004; Brennecke et al. 2007; Malone et al. 2009). It produces piRNAs from only one genomic strand and is exclusively expressed in the somatic follicle cells of the ovary. We selected a P[acman] BAC clone that extended from a position ~30 kb upstream of the first annotated piRNA ~86 kb toward the X chromosome centromere (Venken et al. 2009). This encompassed ~30% of the *flamenco* cluster. To distinguish any ectopic copies of *flamenco* from the native locus, we marked the BAC by recombineering, inserting a cassette comprising a nonfunctional GFP sequence and a bacterial neomycin resistance gene (Copeland et al. 2001; Venken et al. 2006; Sharan et al. 2009). Marker sequences were inserted ~4 kb downstream from the first annotated piRNA in a site, which we had previously shown to produce abundant small RNAs.

For mice, we chose to modify a piRNA cluster on mouse chromosome 17 that is a major contributor to piRNA populations in developing male germ cells from the pachytene stage through the end of meiosis (Fig. 2B; Aravin et al. 2006; Girard et al. 2006). This cluster occurs in syntenic locations in rat and in human, indicating conservation through at least 80 M years of evolution. Like *flamenco*, each region of the ch17 cluster produces piRNAs from only one genomic strand. A mouse BAC clone comprising ~187 kb of chromosome 17 carried the complete ch17 cluster and extended 60 kb upstream of and 30 kb downstream from the locus. It was similarly marked by recombineering to insert a modified GFP/neo cassette.

In flies, we took advantage of a phiC-31 attachment site in the P[acman]-BAC to insert the modified *flamenco* cluster into a known genomic locus (Venken et al. 2006, 2009). Given that *flamenco* is normally present in a location annotated as heterochromatic (X chromosome, band 20A), we chose a gene-rich, euchromatic site to insert the transgene. Specifically, we created lines with one additional copy of *flamenco* on chromosome 3L at band 62E1 (landing pad 31) (Venken et al. 2006). For mice, we used standard pronuclear injection to create two independent founder lines (R13 and R37) with ch17 transgene insertions in presumably distinct random locations.

Small RNA cloning and Illumina sequencing revealed that abundant piRNAs derived from GFP were produced from ectopic clusters in both flies and mice (Fig. 2C,D). Like the native loci, these produced small RNAs from only one genomic strand. Unlike X-TAS, neither *flamenco* nor the ch17 cluster normally participate in the ping-pong

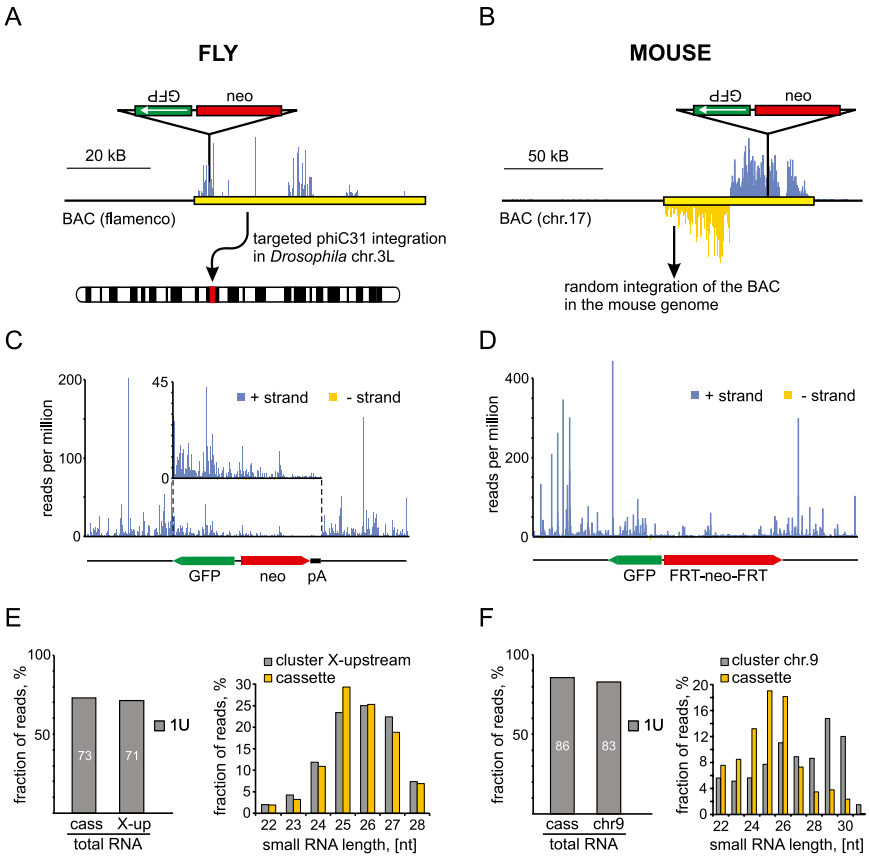


FIGURE 2. Generation of apiRNAs from ectopic clusters in flies and mice. (*A*) A schematic representation of the GFP/Neo cassette is shown along a diagram of the *flamenco* locus (in yellow, piRNA densities in blue) in the BAC used for transgenesis. *Below* is a schematic indicating that the transgene is inserted into chromosome 3L. (*B*) The GFP/neo insertion into the mouse chromosome 17 cluster is diagrammed as in *A*. (*C*) The structure of the *flamenco* GFP/Neo insertion is diagrammed *below* a plot of piRNA frequencies along the insert on the plus and minus strands (indicated). For reference, piRNAs are also mapped to flanking regions, though these represent a mixture of RNAs derived from the two native and one ectopic *flamenco* clusters. (*D*) A scheme of the GFP/Neo insert into the mouse chromosome 17 cluster is shown *below* piRNAs mapping to the insert and its context as in *C*. Again, piRNAs that flank the insert can be derived from the two native or inserted ectopic loci. (*E*) The 1U bias (*left*) and size distributions (*right*) of apiRNAs from the ectopic *flamenco* cluster are compared with another piRNA cluster (X-upstream) that also produces piRNAs from one genomic strand in follicle cells. (*F*) As in *E*, apiRNAs from the ectopic ch17 cluster in mice are compared with a similarly structured cluster on chromosome 9.

amplification loop, and the ectopic insertions also lacked signatures of the cycle, namely, small RNAs with a 10A bias and sense/antisense pairs that overlap by 10 nt. Small RNAs from the ectopic clusters did show the strong 1U bias that is a signature of primary piRNA populations (Fig. 2E,F; Supplemental Fig. S3A).

It seemed likely that the transgenic clusters would generate piRNAs both from the inserted marker gene and from sequences that represent their native content; however, it is impossible to distinguish the latter from piRNAs derived from endogenous loci. The ectopic cluster is present as a single copy in the genome, as compared with two endogenous copies. We might therefore expect piRNA levels coming from shared regions to increase by 1.5-fold if all copies were equally active. Indeed, we noted a 1.3-fold increase in piRNAs, which are derived from the portion of the *flamenco* cluster present in transgene. Similarly, the levels of MILI and MIWI piRNAs derived from the chr17 cluster in mouse increased by between 1.2- and 1.5-fold relative to a nonmodified cluster on ch9 in two independent transgenic lines. The profiles of piRNA mapped to the *flamenco* and ch17 clusters are very similar in wild-type and transgenic flies and mice (Supplemental Fig. S4). Therefore, the heterologous insertion of a marker gene does not appear to exert a strong influence on the processing of piRNAs from transgenic loci. Overall, our data indicate that transgenic piRNA clusters have similar activity to their endogenous counterparts, despite being present at non-native genomic positions.

Flamenco-derived piRNAs associate exclusively with Piwi, the only family member that is expressed in follicle cells (Sarot et al. 2004; Brennecke et al. 2007). Thus, they have a characteristic size profile, peaking at around 25 nt. piRNAs from the ectopic *flamenco* insertion shared this size distribution (Fig. 2E). piRNAs from the ch17 cluster (and other murine clusters expressed during meiosis) normally associate with both MILI and MIWI (Supplemental Fig. S3B). These complexes have distinct small RNA size profiles, with MILI to associate with a ~26-nt and MIWI harboring a ~30-nt species (Fig. 2F; Aravin et al. 2006; Girard et al. 2006). Overall, MIWI-bound species are substantially more abundant than MILI bound species (Aravin et al. 2006; Girard et al. 2006). While the ectopic ch17 cluster produced small RNAs with sizes characteristic of MILI and MIWI complexes, their ratio was very different than expected based upon the behavior of the native cluster (Fig. 2F; Supplemental Fig. S3B). RNAs with the size of MILI partners greatly outnumbered those with the size of MIWI-bound species. Thus, the ectopic cluster appeared to have a strong preference for one of its two potential Piwi-family partners (Fig. 2F; Supplemental Fig. S5).

Overall, our data indicate that piRNA clusters can function even when divorced from their normal genomic locale. With *flamenco*, the ectopic insertion behaved indistinguishably from the native locus, even though it had been substantially truncated on the centromere-proximal

side. For the ch17 cluster, piRNAs were still produced in abundance from the ectopic insertions, but the behavior of the small RNAs shifted toward preferential MILI association. This could indicate that some element of chromosomal context was important for signaling an ultimate association with MIWI or perhaps that critical signals that mark the cluster as a source of MIWI piRNAs were missing from our BAC clone, despite its extending well beyond the two ends of the cluster. Our results by no means rule out chromatin structure as a contributory element in defining piRNA clusters. However, if specific chromatin structures are important, the signals for their formation must be tightly linked to the piRNA loci themselves.

The inclusion of the same artificial sequence in piRNA clusters in multiple locations and in distinct organisms afforded the opportunity to probe the determinants of piRNA selection. In contrast to miRNAs and siRNAs, whose processing from longer precursors is informed by their specific secondary structure and is well understood, no rules that explain the selection of individual piRNAs have been defined. The only bioinformatic study that addressed this question came to the conclusion that the processing of individual piRNA from precursors is quasi-random, with only weak influences of local sequence (positions −1 to +4 relative to the 5′ end of the piRNA) (Betel et al. 2007). However, sequencing efforts from our and other groups showed that individual piRNAs are not produced uniformly along clusters. Instead, certain small RNAs appear substantially more abundant (Aravin et al. 2006, 2008; Girard et al. 2006; Brennecke et al. 2007). Characteristics underlying these inequalities could be intrinsic to the local sequence environment of each individual piRNA or could be conferred by long-distance interactions and formation of secondary structures within the precursor molecule. Alternatively, patterns could be essentially random, with the abundance of each species being determined stochastically.

As with native piRNAs, read distributions along the marker cassettes in the ectopic clusters were very uneven (Fig. 3A; Supplemental Fig. S5). Focusing on the GFP coding sequence, 1% of nucleotide residues contribute 19% of all 5′ ends of GFP-mapping piRNA reads in flies, while 10% of positions account for 70% of reads (Fig. 3B). In mouse, the distribution was even more skewed with 1% of GFP residues contributing 42% of piRNA reads (Fig. 3B). To probe the causes leading to these skewed distributions, we compared GFP-derived piRNAs in the two independent mouse transgenic lines. The correlation in the abundance of individual small RNAs was remarkable (R^2 = 0.99) (Fig. 3C), ruling out the notion that the patterns that we observe are random within each sample. Procedures for preparing small RNA libraries include steps with well-established sequence-based biases, namely, RNA adapter ligations and PCRs (Linsen et al. 2009). We therefore considered the possibility that those biases dominated apparent sequence preferences in apiRNA generation. However, very little correlation was seen between GFP piRNAs in flies and mice (R^2 = 0.01) (Fig. 3D), contrary to what one would expect if the patterns that we observed were strongly influenced by the biases of library preparation methods.

Considered as a whole, our data strongly support the existence of signals that determine the efficiency of production of individual piRNAs and raise several possibilities as to the nature of those signals. First, the biased distribution of piRNAs could be an exclusive consequence of their context within the cluster. This would imply that large-scale features, such as the structure of the transcript or preferential entry sites for the primary processing machinery determine differential piRNA production, akin to the generation of phased siRNAs from long dsRNAs in plants and animals (Zamore et al. 2000; Howell et al. 2007). Alternatively, determinants of efficient piRNA biogenesis could still be defined by the local sequence environment of each individual piRNA, with sequence determinants being interpreted differently in our two experimental models. To begin to discriminate between these possibilities, it was necessary to insert the same sequence (GFP) into different piRNA precursors that are expressed and processed in the same cell type.

The *traffic jam* (*tj*) gene encodes a basic leucine zipper transcription factor and is expressed in the follicle cells of the *Drosophila* ovary, just as is *flamenco* (Li et al. 2003; Saito et al. 2009). Importantly, *tj* generates piRNAs from a discrete segment of its 3′-UTR region (Saito et al. 2009). We created a marked, ectopic copy of *tj* by inserting a GFP coding sequence in the antisense orientation into its piRNA-producing domain and integrated this into a euchromatic site on chromosome 3L (Fig. 4A).

Sequencing of small RNAs (Fig. 4B) yielded abundant piRNAs from the inserted GFP sequence. These had the same characteristics as native *tj*-derived piRNAs, including being produced from the sense strand of the locus, having a size distribution characteristic of Piwi-associated species, and a strong bias for a 5′ terminal U residue (Fig. 4C,D). Position-dependent differences in piRNA abundance were also apparent, with the most abundant 10% of possible GFP piRNAs contributing 81% of all GFP-mapping reads (Fig. 3B).

To discriminate local- from long-distance sequence effects, we compared the abundance of individual piRNAs from the *tj* and *flamenco* transgenes. As compared with the patterns derived from independent insertions of the same transgenes in mice (R^2 = 0.99) (Fig. 3C), patterns of GFP piRNAs from *tj* and *flamenco* appeared quite different (R^2 = 0.24) (Fig. 4D). However, they were much more similar than patterns produced in mouse versus fly (R^2 = 0.01) (Fig. 3D). At the extremes, uridine positions in GFP that generate abundant piRNAs from the *flamenco* transgene tended also to generate abundant piRNAs from *tj* (Fig. 4E). Conversely, those that did not generate piRNAs from *flamenco* did not generate piRNAs from *tj*.

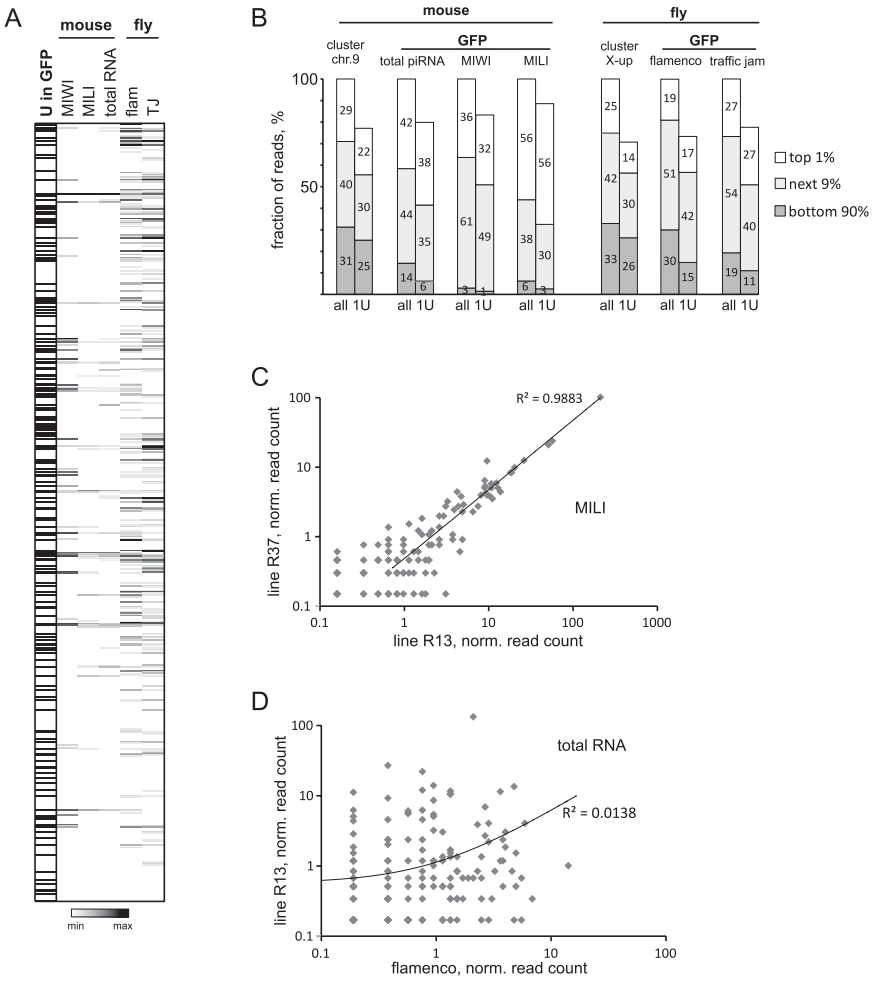


FIGURE 3. apiRNA production is not uniform along inserted sequences. (*A*) A heatmap of piRNA abundance is displayed for all positions in the GFP insert carried in ectopic piRNA clusters as indicated. Sequence measurements were from total RNAs except in mouse, where MIWI and MILI immunoprecipitates (indicated) were also analyzed. The first column simply indicates U positions relative to the heatmaps. (*B*) All possible positions for piRNA production from GFP sequences inserted into ectopic clusters (all sites or only U positions, indicated) were ranked by their contribution to actual piRNA populations. The fraction of piRNAs contributed by the top 1%, the next 9%, or the remaining 90% were measured and indicated. Native clusters (indicated) were similarly analyzed for reference. (*C*) MILI-bound piRNAs were quantified by sequencing from two independent lines carrying the ectopic ch17 cluster. Correlations between read counts for GFP-derived piRNAs are shown. Libraries were normalized as described in the Materials and Methods. (*D*) A similar analysis was performed for GFP-derived piRNAs in total reads, comparing the R13 mouse line carrying the ectopic ch17 cluster and the fly strain carrying the ectopic *flamenco* cluster.

Considered together, these results indicate effects of both local and long-range sequence environment on piRNA biogenesis. Small RNAs generated from GFP embedded in different piRNA precursor transcripts in the same species were more similar than expected by chance. Influences of sequence, however, seem species or cell-type specific, since these same biases did not extend from fly to mouse. Strong effects also appear to be exerted by the context within the cluster, given the near identity in GFP piRNA populations in independent mouse lines and their dissimilarity in comparisons of marked *flamenco* and *tj* transcripts. The precise nature of such context-dependent effects is unclear, but could depend upon the overall secondary or tertiary structures of piRNA precursors.

Our data are consistent with the model in which new insertions of transposable elements become incorporated into the piRNA repertoire as a mechanism of acquiring resistance. Indeed, our data indicate that any sequence will probably produce piRNAs immediately upon its incorporation into a functional piRNA cluster. Furthermore, our data demonstrate that the position of the cluster in the genome is not important, and, therefore, transgenic piRNA clusters can be created in heterologous genomic locations.

Previous bioinformatic analyses described the generation of individual piRNAs from long precursor molecules as a pseudo-random process with a weak influence of the local sequence environment of individual piRNA species (Betel et al. 2007). However, the distribution of individual piRNAs within the precursor is far from being random; different Us have drastically different propensities to generate piRNAs, and some non-U positions produce substantially more piRNAs than nonprocessed U positions. Here, we showed that patterns of individual piRNAs within the precursors are highly reproducible if the sequence is present within the same context. Patterns become less reproducible if the local sequence is embedded in a different context, indicating that both local and long-range sequence environments impact processing efficiency. This result explains a failure in the identification of simple rules that would explain the production of abundant piRNAs from a given precursor molecule.

The general approach we describe here, using marked ectopic piRNA clusters to produce apiRNA species, provides a path toward further dissection of elements that discriminate piRNA clusters and marks corresponding transcripts for piRNA processing. The ability to program the piRNA pathway to produce artificial piRNAs has implications for harnessing this system for controlling gene expression. In particular, in mammals this approach may present advantages over harnessing the miRNA pathway, since piRNAs can induce epigenetic silencing of loci through the recruitment, directly or indirectly, of the de

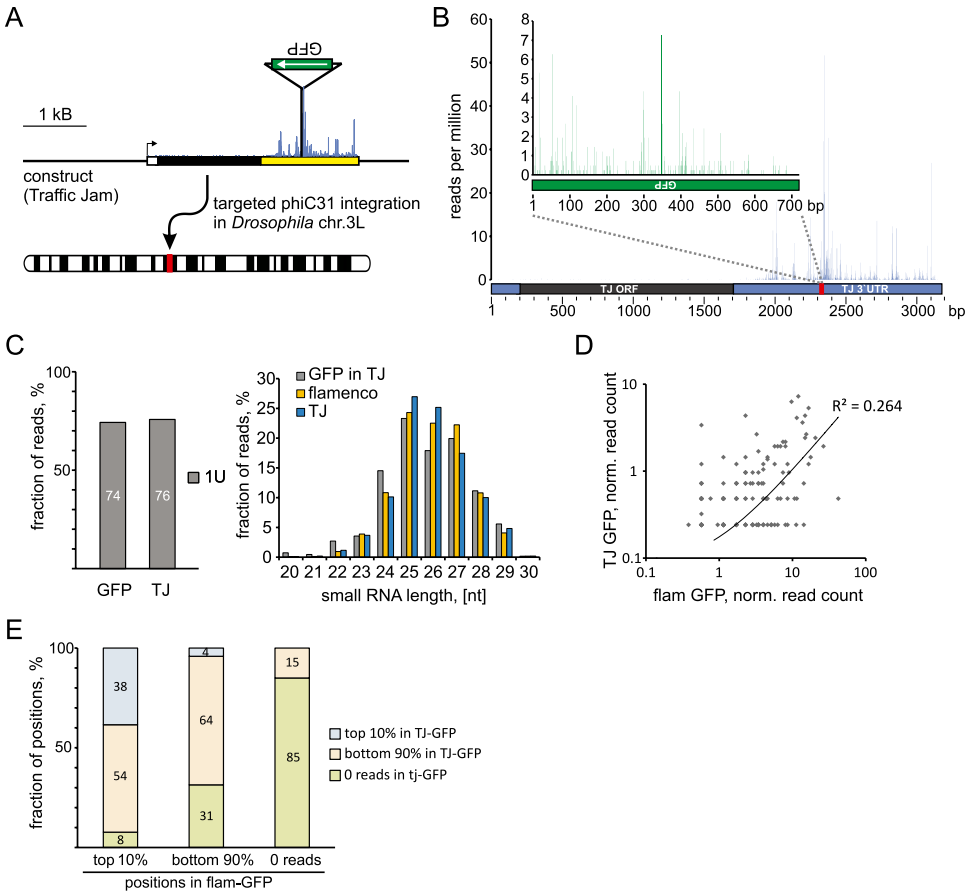


FIGURE 4. apiRNA production from the 3′ UTR of *traffic jam*. (*A*) A schematic of the GFP insertion into the 3′ UTR of the *traffic jam* gene indicates the transcriptional start site (arrow), the coding sequence (black box), and the 3′ UTR (yellow box). *Below*, a diagram indicates site-specific insertion into 3L. (*B*) piRNA read counts are plotted along the inserted GFP sequence (green *inset*) and the surrounding areas of the *tj* 3′ UTR. Note that sequences mapping outside of GFP could be produced from the ectopic insert or the two endogenous copies of *tj*. (*C*) The 1U bias (*left*) and the size distribution of piRNAs mapping to the GFP insert are shown with reference to piRNAs from the *flamenco* cluster. (*D*) Normalized piRNA read counts (see Materials and Methods) were compared for the GFP insertions into the ectopic *flamenco* or *tj* piRNA clusters. (*E*) Read counts are calculated for all possible piRNAs that start with uridine derived from the GFP insertion into *flamenco*. These were divided into the top 10%, the next 90%, and the subset that contributed no reads. For each subset, the number that were present in the top 10%, the next 90%, or the noncontributors for the GFP insertion into *tj* were plotted.

novo DNA methylation machinery (Carmell et al. 2007; Aravin et al. 2008; Kuramochi-Miyagawa et al. 2008; Siomi et al. 2011).

MATERIALS AND METHODS

D. melanogaster strains and crosses

The line, P-1152, which carries an insertion of the P{lArB} construct in telomeric sequences of X chromosome (site 1A) is described in Roche and Rio (1998). To test *trans*-silencing with P-1152, females of this line were crossed with males that have *lacZ* expressed from a euchromatic location on chromosome 2L (line BC69, site 35B10–35C1) (Lemaitre et al. 1993).

Cloning and recombineering—*D. melanogaster*

The *flamenco* transgene was created using P[acman] clone CH321-35A24, which contains an interval from chromosome X that includes ~20 kb of upstream sequence and the 5′ portion of the *flamenco* piRNA cluster (Venken et al. 2009). An antisense EGFP sequence was introduced into the BAC by recombineering as described in Sharan et al. (2009). The GFP-Neo insertion cassette was built by overlapping PCR based on a FRT-PGK-gb2-neo-FRT cassette (Gene Bridges). The position of the insertion within the *flamenco* cluster was selected based on uniqueness and high frequency of piRNA production from the surrounding region. The cassette was introduced into the the BAC using a pSim6 plasmid described in Datta et al. (2006). To promote recombination, *Escherichia coli* containing pSim6 were transferred to a 2-mL Eppendorf tube and induced at 42°C in an Eppendorf tabletop shaker. The linear DNA substrate was introduced by electroporation using the Gene Pulser XCell. Using exponential decay as a pulse-type, the cells were electroporated at 3000 V, 25 μF, and 200 Ω for 5 msec. After outgrowth and selection of cells, recombinant clones were screened for by PCR, sequencing and restriction digestion, followed by pulse-field gel electrophoresis.

The *D. melanogaster traffic jam* gene with 2 kb upstream and 0.5 kb downstream genomic regions was amplified from the CH322-145O22 P[acman] clone and inserted between the BspHI

and ClaI sites of the pIZ-V5-His vector (Invitrogen). A sequence ATTATTCTGATTGCGACAATAAATTCCGAT in the *TJ* 3′ UTR was substituted with the sequence CTTAAGCTGATTGCGACAA TAAATACCGGT by overlap PCR to introduce unique AflII and AgeI sites, which were used to insert the inverted EGFP sequence. The modified *traffic-jam* sequence was transferred into the pCasper5-attB vector (a modified *P-element* pCaSpeR5 vector [Le et al. 2007] with a phiC31 attB site to allow site-specific integration).

Cloning and recombineering—mouse

The transgene containing the modified chr17 piRNA cluster (Chr17: 27427600–27488899) was created using BAC clone RP23-131B16, which contains ~180 kb of genomic sequence that includes the whole chr17 piRNA cluster. We used the FRT-PGK-gb2-neo-FRT cassette (Gene Bridges) and a purified vector containing the EGFP sequence, to construct the GFP-Neo insert for recombineering. After three steps of overlapping PCR (KOD hot start DNA polymerase, Novagen), the recombineering inserts were cloned in a 2.1-TOPO vector (Invitrogen, Version U) according to the manufacturer's protocol. Homology arms for recombineering were added by PCR of purified plasmid.

Recombineering was carried using the Red/ET plasmid-expressing recombination proteins under an arabinose-inducible promoter (Counter-Selection BAC Modification Kit, Gene Bridges, 2007). We followed the manufacturer's protocol, except that recombined clones were selected on Kanamycin and the counter-selection step was skipped. The integrity of modified BAC DNAs were verified by restriction digests and sequencing.

Transgenic animal production—*D. melanogaster*

Tagged BAC DNA was purified with a Plasmid Maxi Kit (QIAGEN). The DNA was used for PhiC31 integrase-mediated transgenesis, which was carried out by BestGene (http://www.thebestgene.com/). *Flamenco* and *tj* transgenes were integrated into attP docking sites on chromosome 3 (VK00031—site 62E1, and VK00033—site 65B2, respectively).

Transgenic animal production—mouse

BAC DNAs were prepared from overnight *E. coli* cultures using Nucleobond BAC 100 columns (Clontech). DNA was eluted in Injection Buffer (10 mM TRIS, 0.1 mM EDTA, 100 mM NaCl, 1X polyamines) and linearized with PI-SceI enzyme for 4 h. Following linearization, BAC DNA was dialyzed overnight on a 25-mm, 0.025-μm filter (Millipore) by floating on Injection Buffer. Transgenic animals were obtained by pro-nuclear injection into B6xSJL F1 hybrids oocytes. Founder animals were crossed to C57BL/6J mice. R37 and R13 transgenic lines were initiated from two independent founder mice.

Immunoprecipitation of PIWI proteins

Immunoprecipitations from *D. melanogaster* ovaries were carried out according to previously described procedures (Brennecke et al. 2007). For mice, MILI and MIWI were immunoprecipitated from adult testis using antibodies and procedures previously described (Aravin et al. 2007b; Vagin et al. 2009). Briefly, testis were dounced in lysis buffer (10 mM Hepes at pH 7.0, 100 mM KCl, 5 mM MgCl2, 0.5% NP-40, 1% triton X-100, 10% Glycerol, 1 mM DTT, proteinase and RNAase inhibitors). Antibodies (MILI-N2 and MIWI-N2) were then added to the cleared lysates and binding reactions were allowed to proceed overnight at 4°C. Protein A beads are then added to the solution and incubated 3–4 h at 4°C with rotation. After three to four washes in NT-2 buffer (5 mM Tris at pH 7.4, 150 mM NaCl2, 1 mM MgCL2, 0.05% NP-40, RNAase inhibitors, 1 mM DTT), antibody complexes were proteinase K treated and RNAs ethanol precipitated following phenol/chlorophorm extraction. A fraction of the precipitated RNAs was radiolabeled and size profiles verified on 15% urea polyacrylamide gels.

Small RNA cloning

Small RNAs from IPs and total RNA extracts were cloned as previously described in Brennecke et al. (2007) and Aravin et al. (2008). Briefly, small RNAs within a 19–33-nt window for mouse samples or a 19–28-nt window for *D. melanogaster* were isolated from 12% polyacrylamide gels. 3′ and 5′ linkers were ligated, and products were reverse transcribed using Superscript III (Invitrogen). Following PCR amplification, libraries were submitted for sequencing using the Illumina GA2x platform.

Detection of β-galactosidase activity in *D. melanogaster* ovaries

Dissected ovaries from 3–5-d-old flies were fixed in freshly prepared 2% glutaraldehyde in PBS for 20 min, washed twice in PBS, and stained for several hours at 37°C in Fe/NaP buffer (3.1 mM $K_3Fe(CN)_6$; 3.1 mM $K_4Fe(CN)_6$; 10 mM $NaH_2PO_4xH_2O$; 0.15 M NaCl; 1 mM $MgCl_2$) with 0.25% X-Gal. Stained ovaries were mounted in 70% glycerol/PBS.

Bioinformatic analysis of small RNA libraries

After FASTQ to FASTA conversion, the Illumina dapter (CTGTAGGCACCATCAATTC) was clipped from the 3′ end of the read and sequences shorter than 16 nt were discarded from further analysis. The remaining sequences were collapsed into a nonredundant list and mapped to the *D. melanogaster* genome (*D. melanogaster* Apr. 2006 [BDGP R5/dm3]) or the mouse genome (mm9) using the short read aligner bowtie (Langmead et al. 2009). Up to two mismatches were allowed. Sequences that failed to map to the genome were mapped against the artificially introduced sequences. The multiplicity count of mapped sequences was normalized to the total number of reads that mapped to the genome. All further bioinformatic analysis on mapping sequences was done using Unix-based text utilities. Details of those scripts can be obtained upon request. Small RNA sequencing data are deposited at GEO, accession number GSE32435.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

We thank members of the Hannon and Aravin labs for helpful discussion and comments on the manuscript. We thank members

of the McCombie lab (CSHL) and Igor Antoshechkin (Caltech) for help with RNA sequencing. We thank Andres Canela (CSHL) for technical assistance and Simon Knott (CSHL) and Alex Zahn (Caltech) for help with statistical analysis. Sang Yong Kim (CSHL) created the transgenic mice used in this study. F.M. was supported by the Volkswagen Foundation and B.C. by the Boehringer Ingelheim Fonds. This work was supported by grants from the National Institutes of Health (DP2 OD007371A and R00 HD057233 to A.A.A.; 5R01GM062534 to G.J.H.), by the Ellison Medical Foundation (A.A.A.), and by a kind gift from Kathryn W. Davis (G.J.H.).

Received August 8, 2011; accepted September 26, 2011.

REFERENCES

Aravin A, Gaidatzis D, Pfeffer S, Lagos-Quintana M, Landgraf P, Iovino N, Morris P, Brownstein MJ, Kuramochi-Miyagawa S, Nakano T, et al. 2006. A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* **442:** 203–207.

Aravin AA, Hannon GJ, Brennecke J. 2007a. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science* **318:** 761–764.

Aravin AA, Sachidanandam R, Girard A, Fejes-Toth K, Hannon GJ. 2007b. Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science* **316:** 744–747.

Aravin AA, Sachidanandam R, Bourc'his D, Schaefer C, Pezic D, Toth KF, Bestor T, Hannon GJ. 2008. A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. *Mol Cell* **31:** 785–799.

Betel D, Sheridan R, Marks DS, Sander C. 2007. Computational analysis of mouse piRNA sequence and biogenesis. *PLoS Comput Biol* **3:** e222. doi: 10.1371/journal.0030222.

Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ. 2007. Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* **128:** 1089–1103.

Brennecke J, Malone CD, Aravin AA, Sachidanandam R, Stark A, Hannon GJ. 2008. An epigenetic role for maternally inherited piRNAs in transposon silencing. *Science* **322:** 1387–1392.

Carmell MA, Girard A, van de Kant HJ, Bourc'his D, Bestor TH, de Rooij DG, Hannon GJ. 2007. MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. *Dev Cell* **12:** 503–514.

Copeland NG, Jenkins NA, Court DL. 2001. Recombineering: a powerful new tool for mouse functional genomics. *Nat Rev Genet* **2:** 769–779.

Cox DN, Chao A, Lin H. 2000. piwi encodes a nucleoplasmic factor whose activity modulates the number and division rate of germline stem cells. *Development* **127:** 503–514.

Czech B, Malone CD, Zhou R, Stark A, Schlingeheyde C, Dus M, Perrimon N, Kellis M, Wohlschlegel JA, Sachidanandam R, et al. 2008. An endogenous small interfering RNA pathway in *Drosophila*. *Nature* **453:** 798–802.

Datta S, Costantino N, Court DL. 2006. A set of recombineering plasmids for gram-negative bacteria. *Gene* **379:** 109–115.

Gan H, Lin X, Zhang Z, Zhang W, Liao S, Wang L, Han C. 2011. piRNA profiling during specific stages of mouse spermatogenesis. *RNA* **17:** 1191–1203.

Girard A, Sachidanandam R, Hannon GJ, Carmell MA. 2006. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* **442:** 199–202.

Gunawardane LS, Saito K, Nishida KM, Miyoshi K, Kawamura Y, Nagami T, Siomi H, Siomi MC. 2007. A slicer-mediated mechanism for repeat-associated siRNA 5′ end formation in *Drosophila*. *Science* **315:** 1587–1590.

Haase AD, Fenoglio S, Muerdter F, Guzzardo PM, Czech B, Pappin DJ, Chen C, Gordon A, Hannon GJ. 2010. Probing the initiation and effector phases of the somatic piRNA pathway in *Drosophila*. *Genes Dev* **24:** 2499–2504.

Harris AN, Macdonald PM. 2001. Aubergine encodes a *Drosophila* polar granule component required for pole cell formation and related to eIF2C. *Development* **128:** 2823–2832.

Houwing S, Kamminga LM, Berezikov E, Cronembold D, Girard A, van den Elst H, Filippov DV, Blaser H, Raz E, Moens CB, et al. 2007. A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. *Cell* **129:** 69–82.

Howell MD, Fahlgren N, Chapman EJ, Cumbie JS, Sullivan CM, Givan SA, Kasschau KD, Carrington JC. 2007. Genome-wide analysis of the RNA-DEPENDENT RNA POLYMERASE6/DICER-LIKE4 pathway in *Arabidopsis* reveals dependency on miRNA- and tasiRNA-directed targeting. *Plant Cell* **19:** 926–942.

Huisinga KL, Elgin SC. 2009. Small RNA-directed heterochromatin formation in the context of development: what flies might learn from fission yeast. *Biochim Biophys Acta* **1789:** 3–16.

Kuramochi-Miyagawa S, Watanabe T, Gotoh K, Totoki Y, Toyoda A, Ikawa M, Asada N, Kojima K, Yamaguchi Y, Ijiri TW, et al. 2008. DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. *Genes Dev* **22:** 908–917.

Lahmy S, Bies-Etheve N, Lagrange T. 2010. Plant-specific multisubunit RNA polymerase in gene silencing. *Epigenetics* **5:** 4–8.

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

Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, Nakano T, Bartel DP, Kingston RE. 2006. Characterization of the piRNA complex from rat testes. *Science* **313:** 363–367.

Lau NC, Robine N, Martin R, Chung WJ, Niki Y, Berezikov E, Lai EC. 2009. Abundant primary piRNAs, endo-siRNAs, and microRNAs in a *Drosophila* ovary cell line. *Genome Res* **19:** 1776–1785.

Le T, Yu M, Williams B, Goel S, Paul SM, Beitel GJ. 2007. CaSpeR5, a family of *Drosophila* transgenesis and shuttle vectors with improved multiple cloning sites. *Biotechniques* **42:** 164–166.

Lemaitre B, Ronsseray S, Coen D. 1993. Maternal repression of the P element promoter in the germline of *Drosophila melanogaster*: a model for the P cytotype. *Genetics* **135:** 149–160.

Li MA, Alls JD, Avancini RM, Koo K, Godt D. 2003. The large Maf factor Traffic Jam controls gonad morphogenesis in *Drosophila*. *Nat Cell Biol* **5:** 994–1000.

Li C, Vagin VV, Lee S, Xu J, Ma S, Xi H, Seitz H, Horwich MD, Syrzycka M, Honda BM, et al. 2009. Collapse of germline piRNAs in the absence of Argonaute3 reveals somatic piRNAs in flies. *Cell* **137:** 509–521.

Lim AK, Kai T. 2007. Unique germ-line organelle, nuage, functions to repress selfish genetic elements in *Drosophila melanogaster*. *Proc Natl Acad Sci* **104:** 6714–6719.

Linsen SE, de Wit E, Janssens G, Heater S, Chapman L, Parkin RK, Fritz B, Wyman SK, de Bruijn E, Voest EE, et al. 2009. Limitations and possibilities of small RNA digital gene expression profiling. *Nat Methods* **6:** 474–476.

Malone CD, Hannon GJ. 2009. Small RNAs as guardians of the genome. *Cell* **136:** 656–668.

Malone CD, Brennecke J, Dus M, Stark A, McCombie WR, Sachidanandam R, Hannon GJ. 2009. Specialized piRNA pathways act in germline and somatic tissues of the *Drosophila* ovary. *Cell* **137:** 522–535.

Olivieri D, Sykora MM, Sachidanandam R, Mechtler K, Brennecke J. 2010. An in vivo RNAi assay identifies major genetic and cellular requirements for primary piRNA biogenesis in *Drosophila*. *EMBO J* **29:** 3301–3317.

Pane A, Wehr K, Schupbach T. 2007. *zucchini* and *squash* encode two putative nucleases required for rasiRNA production in the *Drosophila* germline. *Dev Cell* **12:** 851–862.

Pelisson A. 1981. The I–R system of hybrid dysgenesis in *Drosophila melanogaster*: are I factor insertions responsible for the mutator effect of the I–R interaction? *Mol Gen Genet* **183:** 123–129.

Qi H, Watanabe T, Ku HY, Liu N, Zhong M, Lin H. 2011. The Yb body, a major site for Piwi-associated RNA biogenesis and a gateway for Piwi expression and transport to the nucleus in somatic cells. *J Biol Chem* **286:** 3789–3797.

Robine N, Lau NC, Balla S, Jin Z, Okamura K, Kuramochi-Miyagawa S, Blower MD, Lai EC. 2009. A broadly conserved pathway generates 3′ UTR-directed primary piRNAs. *Curr Biol* **19:** 2066–2076.

Roche SE, Rio DC. 1998. *Trans*-silencing by *P* elements inserted in subtelomeric heterochromatin involves the *Drosophila* Polycomb group gene, *Enhancer of zeste*. *Genetics* **149:** 1839–1855.

Ronsseray S, Lehmann M, Anxolabéhère D. 1991. The maternally inherited regulation of *P* elements in *Drosophila melanogaster* can be elicited by two *P* copies at cytological site 1A on the *X* chromosome. *Genetics* **129:** 501–512.

Ronsseray S, Lehmann M, Nouaud D, Anxolabéhère D. 1996. The regulatory properties of autonomous subtelomeric *P* elements are sensitive to a *Suppressor of variegation* in *Drosophila melanogaster*. *Genetics* **143:** 1663–1674.

Ronsseray S, Josse T, Boivin A, Anxolabéhère D. 2003. Telomeric transgenes and *trans*-silencing in *Drosophila*. *Genetica* **117:** 327–335.

Rubin GM, Kidwell MG, Bingham PM. 1982. The molecular basis of P-M hybrid dysgenesis: the nature of induced mutations. *Cell* **29:** 987–994.

Saito K, Siomi MC. 2010. Small RNA-mediated quiescence of transposable elements in animals. *Dev Cell* **19:** 687–697.

Saito K, Nishida KM, Mori T, Kawamura Y, Miyoshi K, Nagami T, Siomi H, Siomi MC. 2006. Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the *Drosophila* genome. *Genes Dev* **20:** 2214–2222.

Saito K, Inagaki S, Mituyama T, Kawamura Y, Ono Y, Sakota E, Kotani H, Asai K, Siomi H, Siomi MC. 2009. A regulatory circuit for piwi by the large Maf gene traffic jam in *Drosophila*. *Nature* **461:** 1296–1299.

Saito K, Ishizu H, Komai M, Kotani H, Kawamura Y, Nishida KM, Siomi H, Siomi MC. 2010. Roles for the Yb body components Armitage and Yb in primary piRNA biogenesis in *Drosophila*. *Genes Dev* **24:** 2493–2498.

Sarot E, Payen-Groschene G, Bucheton A, Pelisson A. 2004. Evidence for a piwi-dependent RNA silencing of the gypsy endogenous retrovirus by the *Drosophila melanogaster* flamenco gene. *Genetics* **166:** 1313–1321.

Senti KA, Brennecke J. 2010. The piRNA pathway: a fly's perspective on the guardian of the genome. *Trends Genet* **26:** 499–509.

Sharan SK, Thomason LC, Kuznetsov SG, Court DL. 2009. Recombineering: a homologous recombination-based method of genetic engineering. *Nat Protoc* **4:** 206–223.

Siomi MC, Sato K, Pezic D, Aravin AA. 2011. PIWI-interacting small RNAs: the vanguard of genome defence. *Nat Rev Mol Cell Biol* **12:** 246–258.

Todeschini AL, Teysset L, Delmarre V, Ronsseray S. 2010. The epigenetic *trans*-silencing effect in *Drosophila* involves maternally-transmitted small RNAs whose production depends on the piRNA pathway and HP1. *PLoS ONE* **5:** e11032. doi: 10.1371/journal.pone.0011032.

Tomari Y, Du T, Haley B, Schwarz DS, Bennett R, Cook HA, Koppetsch BS, Theurkauf WE, Zamore PD. 2004. RISC assembly defects in the *Drosophila* RNAi mutant armitage. *Cell* **116:** 831–841.

Vagin VV, Sigova A, Li C, Seitz H, Gvozdev V, Zamore PD. 2006. A distinct small RNA pathway silences selfish genetic elements in the germline. *Science* **313:** 320–324.

Vagin VV, Wohlschlegel J, Qu J, Jonsson Z, Huang X, Chuma S, Girard A, Sachidanandam R, Hannon GJ, Aravin AA. 2009. Proteomic analysis of murine Piwi proteins reveals a role for arginine methylation in specifying interaction with Tudor family members. *Genes Dev* **23:** 1749–1762.

Venken KJ, He Y, Hoskins RA, Bellen HJ. 2006. P[acman]: a BAC transgenic platform for targeted insertion of large DNA fragments in *D. melanogaster*. *Science* **314:** 1747–1751.

Venken KJ, Carlson JW, Schulze KL, Pan H, He Y, Spokony R, Wan KH, Koriabine M, de Jong PJ, White KP, et al. 2009. Versatile P[acman] BAC libraries for transgenesis studies in *Drosophila melanogaster*. *Nat Methods* **6:** 431–434.

Wilson C, Pearson RK, Bellen HJ, O'Kane CJ, Grossniklaus U, Gehring WJ. 1989. P-element-mediated enhancer detection: An efficient method for isolating and characterizing developmentally regulated genes in *Drosophila*. *Genes Dev* **3:** 1301–1313.

Zamore PD, Tuschl T, Sharp PA, Bartel DP. 2000. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* **101:** 25–33.

Muerdter et al. Supplementary Figure 1

A

strain P1152;
P{lArB} in TAS

B

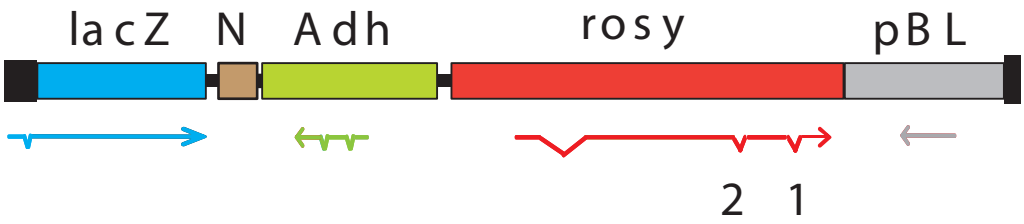
strain BC69;
lacZ transgene
in euchromatin

C

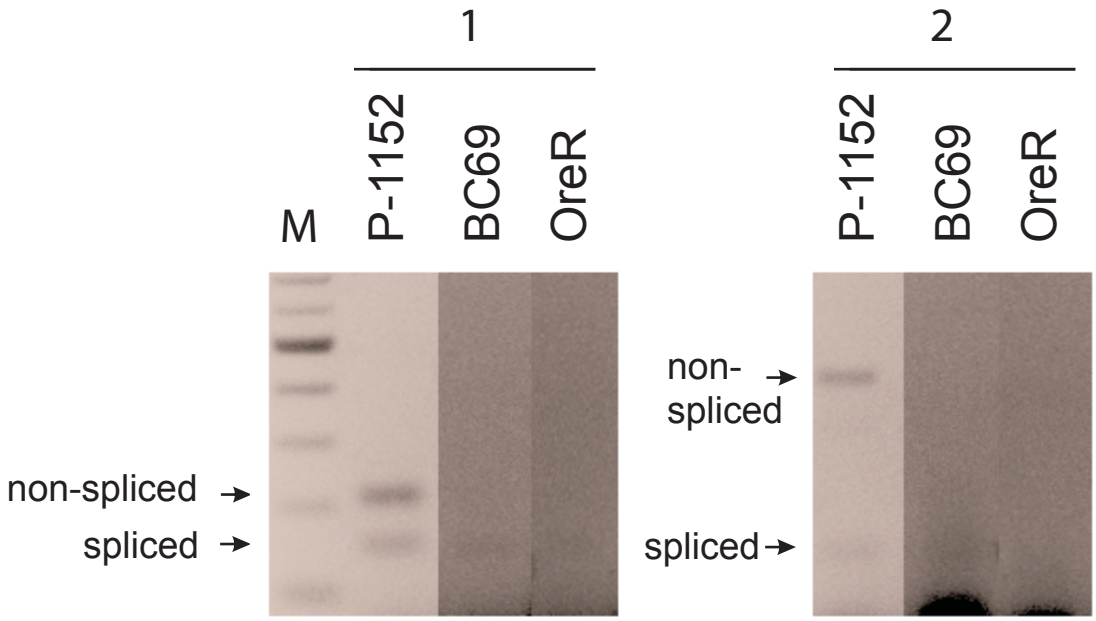
P1152 (female)
x BC69 (male)
(F1)

Muerdter et al. Supplementary Figure 2

A

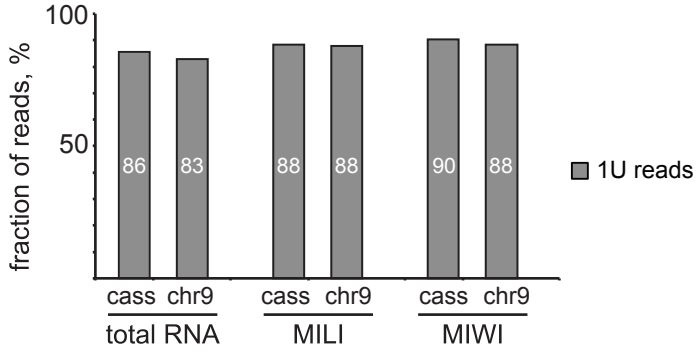


B

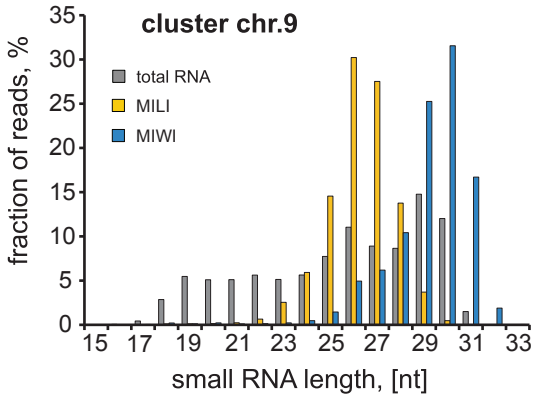


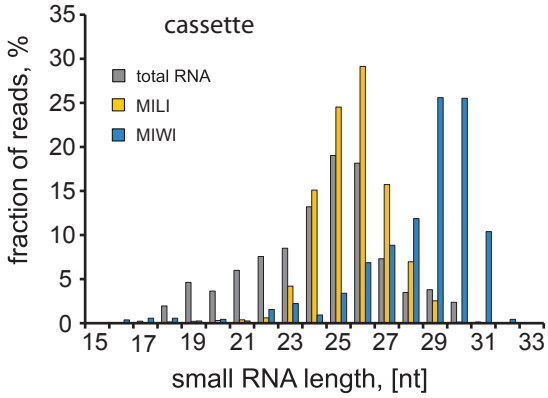
Muerdter et al. Supplementary Figure 3

A

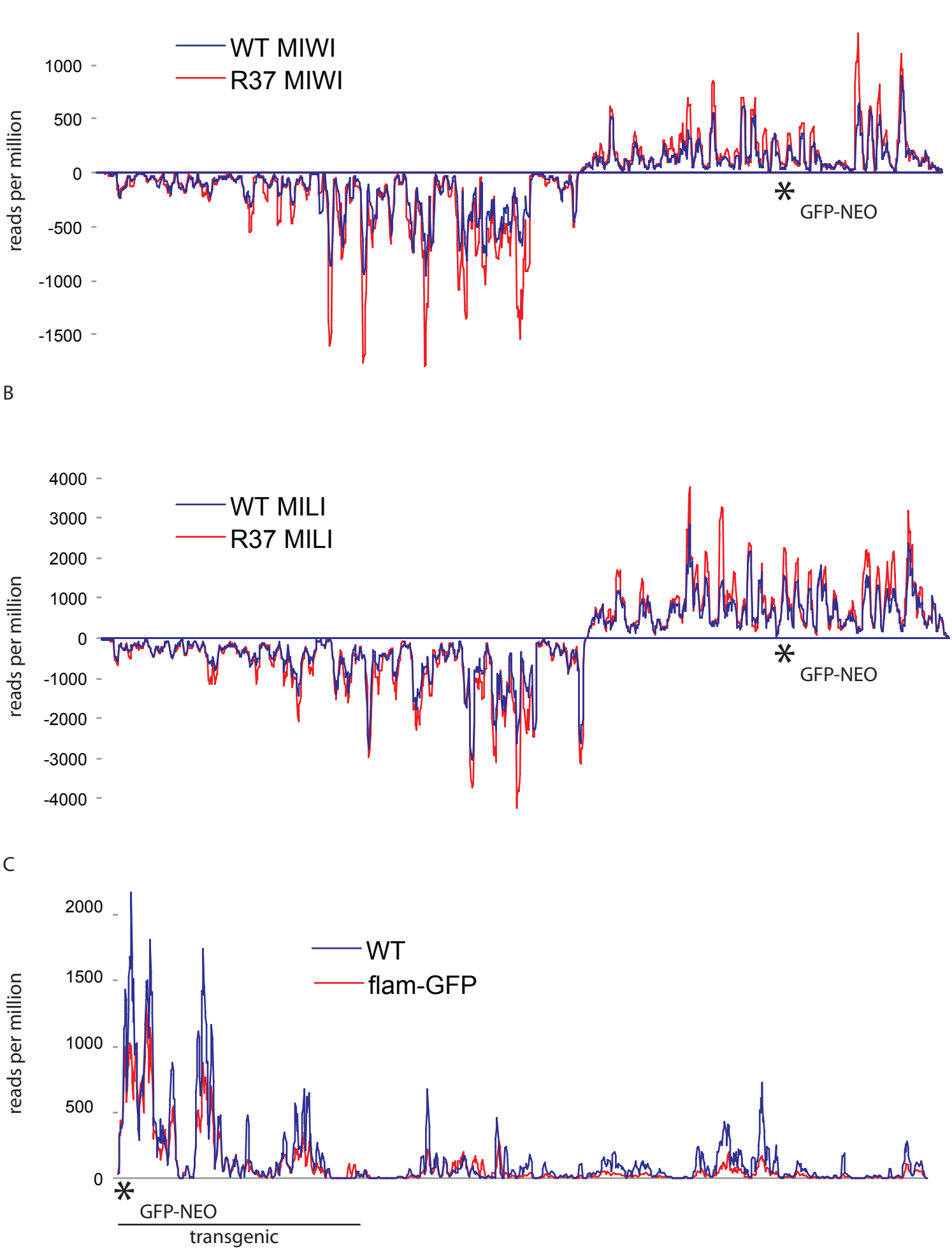


B



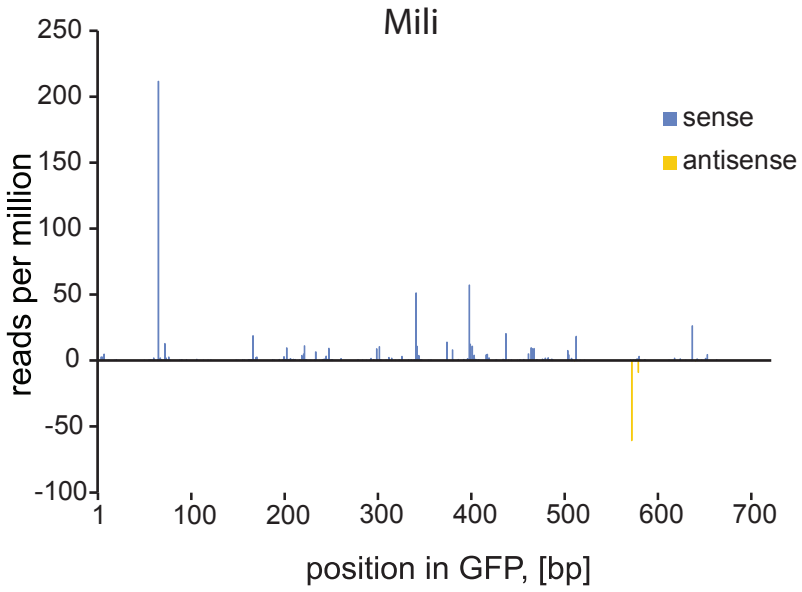


Muerdter et al. Supplementary Figure 4

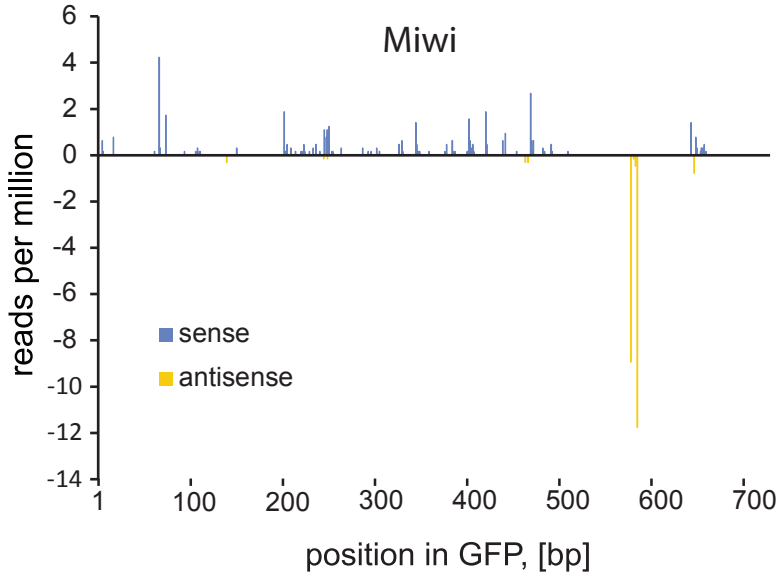


Muerdter et al. Supplementary Figure 5

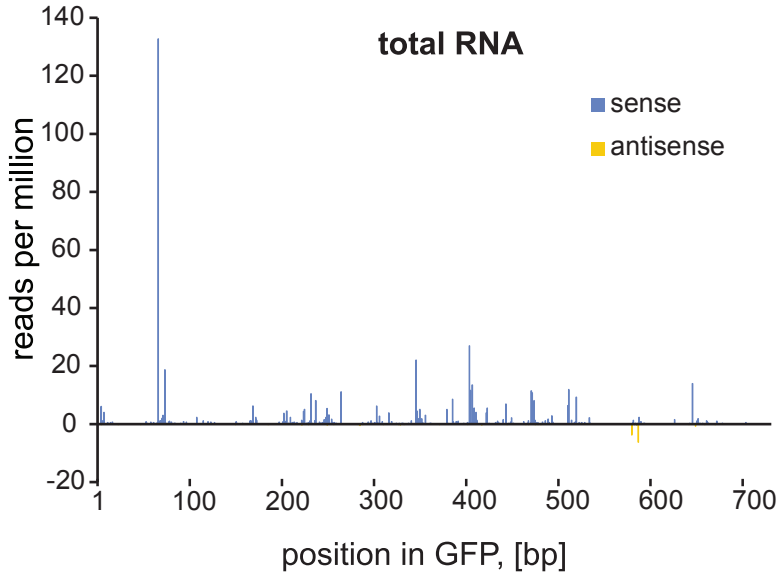
A



B



C



Supplementary Figure Legends

Figure S1. Trans-silencing of lacZ by P{lArB} derived apiRNAs. (A) Ovaries of strain P1152, which carries the P{lArB} insertion in TAS. (B) Ovaries of strain BC69 show lacZ expression from a euchromatic transgene. (C) Trans-silencing of lacZ expression in F1 ovaries of a cross between P1152 females and BC69 males. Note the slightly different levels of repression within different cells of the same ovary.

Figure S2. RT-PCR analysis of *rosy* transcripts. Reverse transcription with primer specific to sense strand of *rosy* transcripts was performed on total RNA from ovaries of strain P-1152 (TAS-inserted transgene), BC69 (same transgene inserted into euchromatin) or Oregon-R (wild type). (A) Position of PCR primers flanking 3rd and 2nd introns of *rosy* transcript. (B) Presence of longer PCR product indicates accumulation of non-spliced *rosy* transcripts in ovaries of P-1152, but not BC69 and Oregon flies.

Figure S3. Features of apiRNAs in mouse. (A) The 1U bias of apiRNAs mapping to the insertion cassette (cass) is compared to native piRNAs from another cluster on chr9. Sequences are derived from total RNA, MILI and MIWI immunoprecipitations (indicated). (B) Size distributions of native piRNAs mapping to a cluster on chr9 (upper panel) compared to apiRNAs mapping to the insertion cassette (lower panel). Sequences are derived from total RNA, MILI and MIWI immunoprecipitations (indicated).

Figure S4. piRNA profiles over wild-type and transgenic piRNA clusters in flies and mice (A) Read densities of piRNAs bound to MIWI are plotted along the cluster on chr 17 on the plus and minus strand (indicated). The site of the GFP cassette insertion is indicated with an asterix. (B) Read densities of piRNAs bound to MILI are plotted along the cluster on chr 17 on the plus and minus strand (indicated). (C) Read densities of piRNAs from total RNA are plotted along

flamenco on the plus strand. The portion of the cluster contained in the BAC is indicated as ‘transgenic’.

Figure S5. apiRNAs in mouse are preferentially bound by MILI. (A) Read counts of apiRNAs bound to MILI are plotted along the inserted GFP sequence on the plus and minus strand (indicated). (B) Read counts of apiRNAs bound to MIWI are plotted along the inserted GFP sequence on the plus and minus strand (indicated). (C) Read counts of apiRNAs from total RNA are plotted along the inserted GFP sequence on the plus and minus strand (indicated).

2.2 Establishment of De novo Methylation Profiles in Mouse PGCs: interplay between transcription, small RNAs and De novo Methylation.

2.2.1 Résumé en Français.

Au cours de l'induction des cellules germinale primordiales (PGCs), les marques de methylation de l'ADN sont intégralement effacées et rétablies *de novo*. Chez les mâles, il a été proposé qu'une classe de petits ARNs associés aux protéines de la famille PIWI (piRNAs) cible la machinerie de methylation *de novo* au niveau des éléments répétés du génome. Trois éléments étayent cette hypothèse : i) les souris mutantes pour la voie des piRNA sont stériles ; ii) elles présentent une altération des profils de methylation des transposons ; iii) les protéines PIWI, MILI (ou PIWIL2) et MIWI2 (ou PIWIL4), exprimées durant la maturation des PGCs sont associées à des piRNAs pouvant potentiellement cibler l'ensemble des transposons du génome. Cependant l'impact réel des piRNAs sur les profils de methylation des transposons n'a jamais été étudié à l'échelle génomique. Dans cette étude, nous présentons les profils de méthylation de spermatocytes sauvages ou mutants pour les piRNAs, ainsi que les profils de transcriptions des PGCs au cours de leur maturation. Nous montrons que, suite à la de-méthylation massive de leur génome, les PGCs réactivent transitoirement la transcription des retro-transposons. Par opposition, les copies de transposons résistantes à la dé-méthylation ne sont jamais induites. Les profils de méthylation des spermatocytes déficients pour la voie des

piRNAs se sont révélés étonnamment proches de ceux d'animaux sauvages, suggérant que, suite à une première vague de methylation *de novo* par défaut, les piRNAs s'engagent dans une seconde vague ciblant un nombre réduit d'éléments. Ces éléments diffèrent d'autres copies de la même famille par leur séquence et, nous établissons que leur methylation *de novo* dépend d'une population de piRNAs amplifiée par ping-pong et préférentiellement associée à MIWI2.

2.2.2 Specific contribution to the work

The following results constitute an unpublished work in the process of being submitted. I was involved in the production and analysis of all data reported here. Current author list:

Antoine Molaro, Emily Hodges, Ilaria Falciatori, Krista Marran, Tyler Garvin, Shahin Raffi, W. Richard McCombie, Alexei A. Aravin, Andrew D. Smith & Gregory J. Hannon

2.2.3 Manuscript and figures

Summary

During mammalian embryonic germ cell development, DNA methylation is thought to be entirely erased and *de novo* re-established genome-wide (Monk M, 1987; Reik et al., 2001; Surani et al., 2007). In male germ cells, the prevailing model suggests that repeat *de novo* methylation rely on targeting by PIWI interacting RNAs, or piRNAs (Aravin and Bourc'his, 2008). piRNA mutant mice are infertile and display methylation defects over, at least, some transposon loci

(Deng et al., 2002; Kuramochi-Miyagawa et al., 2001, 2004 and 2008; Carmell et al., 2007; Aravin et al., 2007). Mouse PIWI proteins expressed during primordial germ cell (PGCs) development, MILI (or PIWIL2) and MIWI2 (PIWIL4), associate with piRNAs spanning the full spectrum of transposons (Aravin et al., 2008, Kuramochi-Miyagawa et al., 2008). However, we previously uncovered thousands of naturally hypomethylated transposons in primate sperm methylomes, suggesting that some copies evade piRNA driven *de novo* methylation (Molaro et al., 2011). Thus, the rules leading to proper epigenetic reprogramming of transposons still remain to be studied. Here, we present the reference methylomes of WT and MILI mutant mouse spermatocytes as well as the small- and long-RNA profiles across PGC development. We show that de-methylated 13.5dpc PGCs transiently re-activate retro-transposon transcription, a feature never seen in somatic tissues. By contrast, elements resisting demethylation at E13.5 are never induced. Surprisingly, mutants and WT spermatocytes have very similar transposon methylomes, indicating that a widespread primary wave of *de novo* methylation is initiated by default and that piRNAs engage in a secondary wave, targeting only a small subset of repeats. These repeats display divergent regulatory sequences when compared to other copies of the same sub-family and depend on a transient ping-pong amplifying population of piRNAs enriched in MIWI2 complexes for *de novo* methylation.

Results

Reference methylomes

To gain global understanding of *de novo* DNA methylation occurring during mammalian germ cell epigenetic reprogramming, we generated single CpG resolution genome-wide methylation maps of purified ~13.5dpc PGCs, WT and *mili*-/- spermatocytes (see experimental procedures). Both Males and female 13.5dpc PGCs were sequenced to an average coverage of ~1.6-2X encompassing, respectively, 76% and 65% of all CpG sites in the mouse genome. Spermatocyte libraries were sequenced to an average coverage of ~8X and more than 90% of all sites were covered at least once in both WT and mutant methylomes (SupFig2.1). We also produced a lower coverage methylome of age-matched *mili*+/- animals, however, the following sections are restricted to a comparison of wild-type to mutant mice.

Consistent with previous reports, PGCs and spermatocytes represent the two extremes of the methylation spectrum (Popp et al., 2010). PGCs had virtually no methylated CpG sites (average methylation of 4%, SupFig2.1), strongly contrasting with the highly methylated spermatocyte genome (78%). This nearly complete erasure of CpG methylation at 13.5dpc was found across all genomic annotations, including all classes of retro-transposons (data not shown and Fig2.1a). However, both male and females PGCs displayed a substantial fraction of elements retaining methylation, preferentially within the LTR class, a feature already observed by other approaches (Popp et al., 2010, Lees-Murdock et al., 2003, Lane et al., 2003). Despite the low resolution of these methylomes, correlating the methylation status across neighboring CpGs allowed us to

confidently call about 8000 hypermethylated domains in male PGCs and 9000 in female PGCs. Enrichments for LTR subfamilies overlapping these domains were very similar in both sexes and strongly enriched for *musculus* specific IAPs (SupFig2.1).

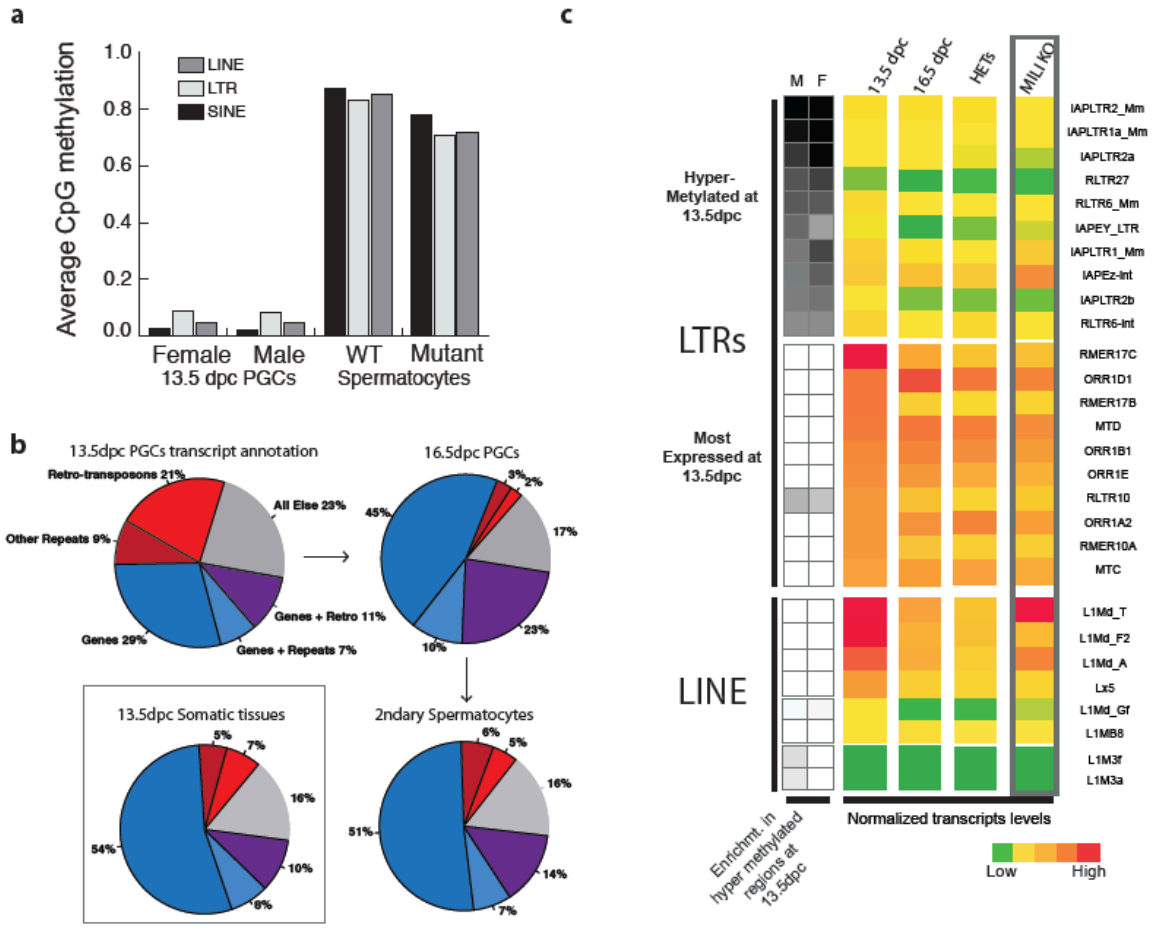


Figure 2.1: Retro-transposon reactivation upon epigenetic reprogramming.
a Average CpG methylation across all three retro-transposon classes in male and female 13.5dpc PGCs and in WT and mili-/- (Mutant) spermatocytes. **b** Relative abundance of annotated reads in 13.5dpc, 16.5dpc and meiotic (2ndary Spermatocytes) germ cell transcriptome. Transcript annotations of somatic tissues collected at 13.5dpc are also shown (inset). **c** Heat map of normalized transcript levels for key subfamilies of LTR and LINE retro-transposons. Subfamily enrichment (low in white, high in black) in 13.5dpc PGC hyper methylated domains in male and female is also show (left blocks).

Epigenetic reprogramming is associated with a transient re-activation of retro-transposon

Cytosine methylation is one of the major means by which retro-transposon transcription is repressed in mammalian genomes (Walsh et al., 1998). The vast erasure of methyl-marks measured across these repeats raised the question of whether their transcriptional status is affected. Therefore, we profiled transcription before, during and after *de novo* methylation (Fig2.1b). When we measured the relative abundance of transcripts mapping to genes, repeats or other locations, 13.5dpc PGCs showed the highest fraction of retro-transposon reads compared to somatic cells co-sorted from E13.5 gonads or to any later time point in germ cell maturation (Fig2.1b). Interestingly, the relative abundance of retro-transposons drastically decreases as PGCs undergo *de novo* methylation, dropping from 21% of all mapped reads in 13.5dpc PGCs to 2% at 16.5dpc (Fig1b).

Surprisingly, young and active sub-families of LINE-1 elements (L1s) ranked among the top most expressed elements (Fig2.1c). In fact, L1Md_T/A/F contributed about 4% of all annotated reads in 13.5dpc PGC transcriptomes (SupFig2.1). These *musculus* specific elements are strongly up-regulated in piRNA deficient animals, consistent with the idea that piRNAs contribute to the silencing of potentially threatening transposons (Fig2.1c and SupFig2.1). By contrast, LTR sub-families displayed a more complex expression pattern. First, the top sub-families enriched in hypermethylated domains at 13.5dpc only

displayed a weak induction suggesting that transposon transcription is a direct consequence of their hypomethylated state in post migratory PGCs. Second, LTR transcript levels in WT versus mutant spermatocytes show little change by comparison to that seen for L1s, and instead they are maintained at relatively constant levels during development. It is important to note that due to the potential equivocal read alignment to copies of the same sub-family, we cannot rule out that distinct copies contribute to this signal at different stages of development.

Retro-transposon de novo methylation: Default vs. piRNA-dependent

Next we assessed the contribution of the piRNA pathway to the genome-wide methylation profiles of meiotic germ cells. Comparing WT and mili-/- spermatocyte methylomes revealed very similar overall methylation levels (78% and 70% respectively). Unexpectedly, this was also true when focusing solely on retro-transposon sequences (Fig2.1a), suggesting that most of the repeat *de novo* methylation is piRNA independent. Identifying contiguous domains of low methylation, termed hypomethylated regions or HMRs (Molaro et al., 2011), also confirmed high overlap between these two samples with most variation seen across repeat associated HMRs (SupFig2.1). Next, we identified differentially methylated regions (DMRs) and isolated a total of ~6000 DMRs as being exclusively hypomethylated in mili-/- spermatocytes. As expected, these DMRs primarily overlap intergenic repeats (~69%, Fig2.2a). Of the fraction overlapping genic space (defined as +/- 10kb, ~27%), most DMRs are associated with

repeats located either in introns or surrounding regions (Fig2.2a and data not shown). The annotation of repeat copies associated with these DMRs identified ~17000 elements (Fig2.2b), which we refer to as DMR transposons or piRNA-targeted transposons.

We previously reported thousands of transposon copies naturally evading *de novo* methylation in chimp and human sperm methylomes (Molaro et al., 2011). Similarly, we report here ~94000 repeat copies evading methylation in both WT and mutant spermatocytes and refer to these as constitutively hypomethylated (Fig2.2b). When compared to the hypermethylated fraction of repeats (4.8Milion copies), DMR and constitutively hypomethylated repeats display a similar overall enrichment for known active mouse transposons families (Fig2.2b). Similar to what is observed for constitutively hypomethylated copies, DMR L1 retro-transposons were hypomethylated toward their 5'end, indicating that piRNAs responsible for their methylation in a WT context specifically target regulatory sequences involved in transcriptional regulation (Fig2.2c).

Comparison of repeat copies within identical sub-families that either evade *de novo* methylation or are targeted by piRNAs revealed signs of sequence divergence, both with respect to the consensus sequence or in terms of conservation, when mapped back to the closely related rat genome (SupFig2.2). In each case, piRNA targeted retro-transposons scored as being more recent insertions. This suggests that the piRNA pathway can discriminate between closely related repeat copies to specifically target the most recently integrated, and potentially threatening, ones. Looking for sequence motifs that could explain

this specificity, we aligned all known L1Md elements and looked for motifs enriched in differentially methylated copies between WT and mili-/-. As expected, most differences were found in their 5‘ regulatory region and strikingly, we detected four motifs found almost exclusively in DMR copies and never found in constitutively hypomethylated ones (Fig2.2d). This finding suggests that these motifs may drive differential transcription factor binding and transcriptional activation assisting the piRNA pathway during target recognition.

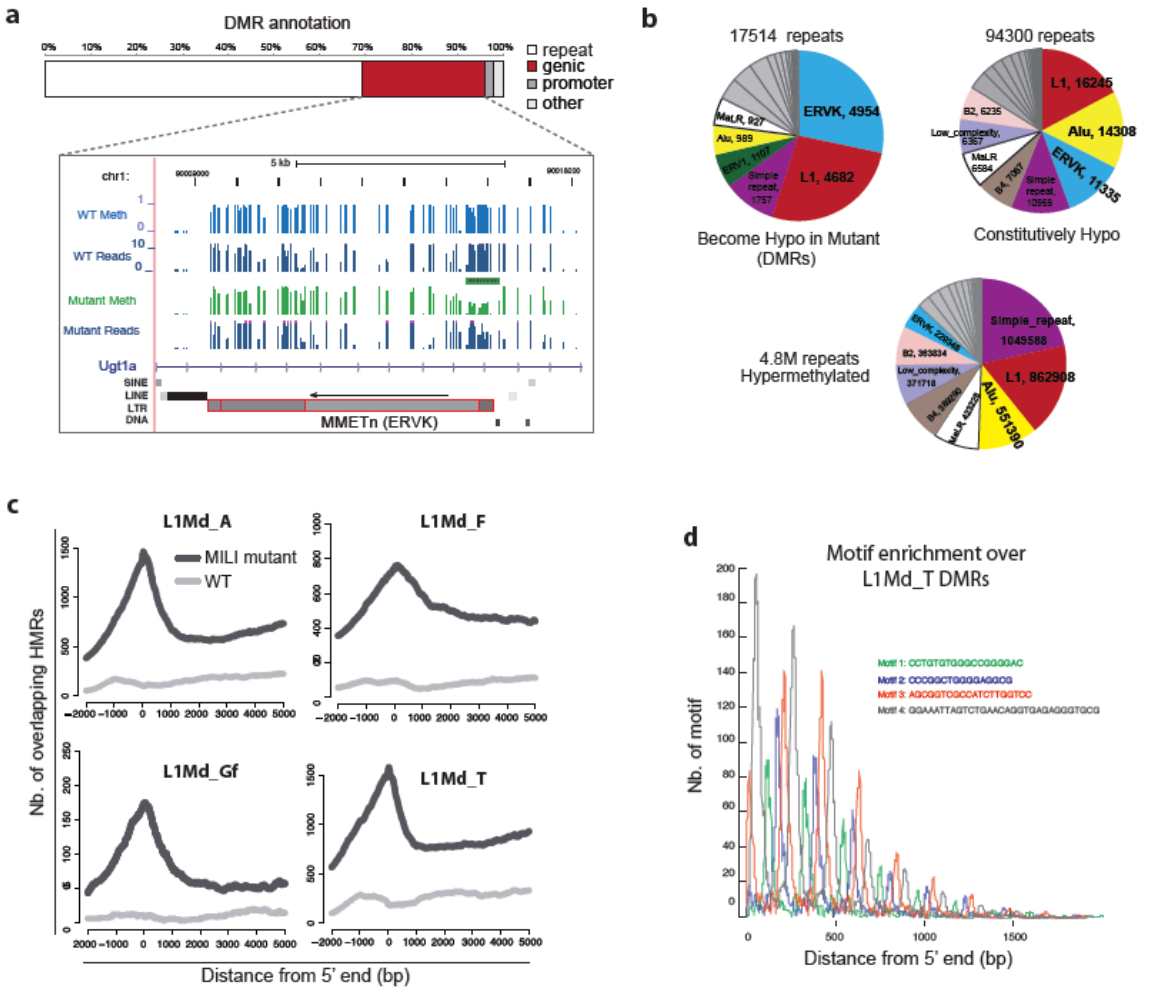


Figure 2.2: piRNA-mediated *de novo* methylation is restricted to distinct transposon copies.

a Genomic annotation of differentially methylated regions (DMRs) between WT and mili-/- spermatocytes. The inset depicts the UCSC browser view of an intronic repeat-associated DMR in first intron of Ugt1a. Read coverage (Reads), single CpG methylation levels (Meth) and hypomethyalted regions (solid bar) are shown for both WT and mutant. **b** Absolute counts of repeat copies (grouped by families) classified as overlapping DMRs, constitutively hypomethylated or hypermethylated in spermatocytes. **c** Metagene analysis of hypomethylated regions (HMRs) distribution over the copies, of four L1Md sub-families, differentially methylated in mili -/- . **d** Distribution of four motifs specifically enriched in differentially methylated L1Md_T.

An adaptive piRNA population mediates a second wave of repeat de novo methylation in PGCs.

Finally, we investigated how piRNAs could mediate the specific targeting of the aforementioned DMR retro-transposons,one hypothesis being that these copies would be strongly up-regulated at 13.5dpc and become preferentially integrated into PIWI proteins as a consequence of their abundance. A second option was that piRNA mapping to the differentially methylated domains of these copies would display discriminative signatures not shared by piRNAs that never engage in *de novo* methylation. To discriminate between these alternatives, which are not mutually exclusive, we cloned and sequenced 24 to 33nt small RNAs from male genital ridges at 13.5dpc and compared these to libraries generated from total, MILI and MIWI2 immuno-precipitated RNA from 16.5dpc genital ridges (Aravin et al., 2008). 13.5 genital ridges displayed an abundant fraction of reads resembling piRNAs, consistent with the activation of MILI expression early in PGCs development (Aravin et al., 2008). These piRNAs displayed a strong 5'U bias (80% of all reads), and a size range typical of this small RNA class (SupFig2.3). These piRNAs are likely to represent the most primary population produced by PGCs.

Whereas 16.5dpc piRNAs are strongly enriched for transposon reads (~50%), 13.5dpc piRNAs displayed an equal abundance of reads mapping to repeats, genes or other annotations (~30% each, Fig2.3a). This suggests that the relative abundance of each annotation class follows the underlying transcript abundance reported at 13.5dpc (Fig2.1b). However, this finding indicates that

transposon piRNAs undergo a strong secondary enrichment at 16.5dpc, as transposon transcript levels were found to occupy less that 10% of all mapped transcripts at the same stage. This is consistent with a secondary ping-pong amplification and stabilization of piRNAs in the presence of both MILI and MIWI2, after 15.5pdc (Aravin et al., 2008, Kuramochi-Miyagawa et al., 2008). Interestingly, support for the ping-pong model was also observed by the relative abundance of piRNAs and transcripts of various transposon subclasses at these stages, including LINEs and SINEs (SupFig2.3 and data not shown), suggesting that abundantly transcribed sub-families of L1s contribute the most to piRNA production (SupFig2.3). By contrast, LTR sub-families displayed a rather uniform abundance of piRNA reads, with the exception of the internal sequence and terminal repeats known to be associated with active IAP elements (SupFig2.3). This suggests that different piRNA based mechanisms lead to L1 and IAP silencing during PGC development. Their differential abundance in DMRs between WT and mili-/- reported here, together with a recent report showing that the catalytic domain of MILI is essential for L1 silencing (De Fazio et al., 2011), both support this idea.

We then attempted to characterize the differences between piRNAs underlying differentially methylated regions from other piRNAs (e.g. exclusively involved in post-transcriptional silencing). When the genomic sequence of a differentially methylated L1Md_T was extracted and used to focally re-map transcript and piRNAs, obvious differences were visible between the differentially methylated portion (5' repeats) and the rest of the element (Fig2.3b). piRNAs

mapping to the 5' repeats strongly amplified at 16.5dpc when compared to reads overlapping the open reading frame (ORF). In addition, a clear bias towards sense reads was found for 5'repeat-associated piRNAs in 13.5 and 16.5dpc total RNA libraries (Fig2.3b and SupFig2.3). In contrast, MIWI2-bound piRNAs were enriched in both sense and anti-sense reads throughout the element. 5' repeat and ORF associated reads also displayed different enrichments for ping pong pairs (SupFig2.3). These initial observations prompted us to interrogate all L1 and ERVK DMR-associated piRNAs in an unbiased fashion. Reads were aligned to the differentially methylated portion of these elements and compared to unaligned reads. L1 DMR-associated piRNAs displayed a bias towards sense reads in addition to a strong enrichment for perfect ping-pong pairs (Fig2.3c and d). Surprisingly, ERVKs derived piRNAs didn't show as strong enrichment, suggesting that LTR and LINE engage differently in the piRNA pathway. However, for both classes of elements, DMR-associated reads were found to be the most abundant population bound to MIWI2 (Fig2.3e).

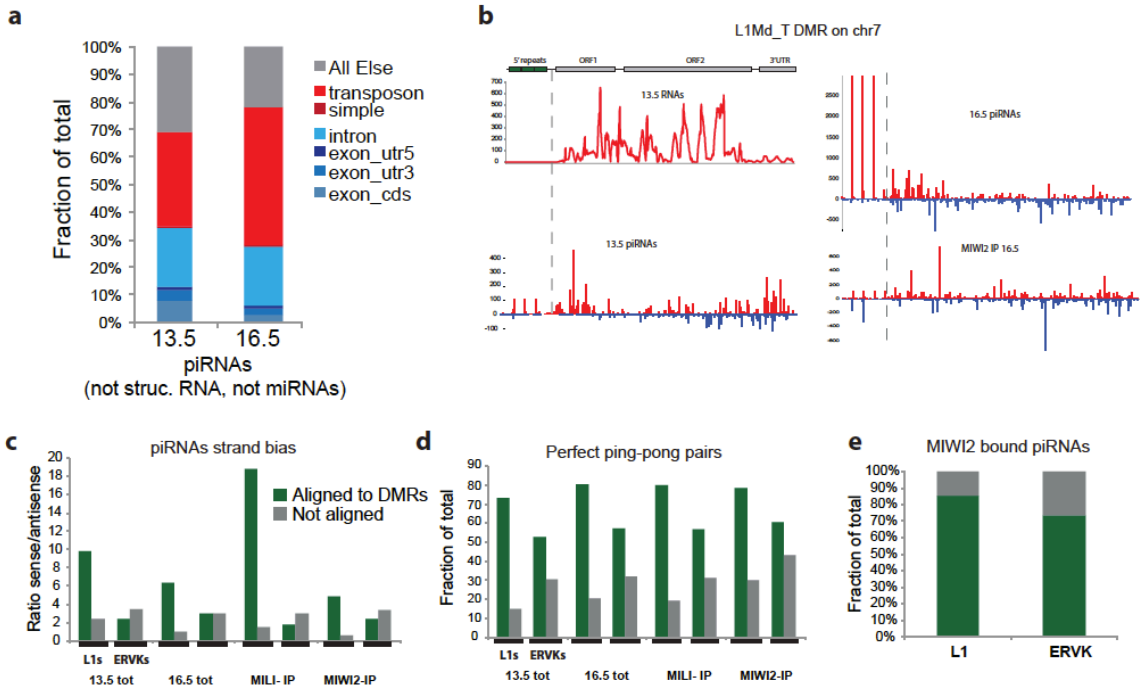


Figure2.3: Features and biogenesis of piRNAs mediating transposon *de novo* methylation

a Genomic annotation as indicated for 13.5dpc cloned piRNAs compared to 16.5dpc form Aravin *et al.,* 2008. **b** Mapping of 13.5dpc transcripts, 13.5, 16.5 and MIWI2-bound piRNAs to a representative L1Md_T (pos. chr7:62853712-62859551) differentially methylated in mili-/-. **c** Ratio of sense to antisense reads for piRNA sequences aligned (green) or not aligned (grey) to all L1 or ERVK-associated DMRs. 13.5, 16.5, MILI-IP and MIWI2-IP libraries are shown. **d** Fraction of total mapped reads with perfect ping-pong pairs for sequences as described in **c**. Fraction of total MIWI2-bound piRNAs aligned (green) or not aligned (grey) to DMRs for L1s and ERVKs.

Concluding remarks

Taken together the data presented here suggests that upon PGC reprogramming, hypomethylated repeats are transcriptionally up-regulated and converted into a primary pool of piRNAs. As a consequence of an initial genome-wide wave of default *de novo* methylation in PGCs after 13.5dpc (Walsh et al.,

1998; Ueda et al., 2000; Kato et al., 2007; Lees-Murdock et al., 2003; Kuramochi-Miyagawa et al., 2008), we suggest that retro-transposons become silent again. However, a small fraction of retro-transposon copies fail to be subjected to this default re-methylation, remain transcriptionally active, and engage in a ping-pong dependent secondary piRNA production. We also show that the group of transposon copies evading *de novo* methylation, despite a functional piRNA pathway, distinguishes divergent copies from threatening ones and might indicate their recent functionalization by the genome. Reminiscent of what has been observed in plant (Slotkin et al., 2009), the interplay described here between transcription, piRNA production and default *de novo* DNA methylation, provide a elegant model explaining how, at each generation, a unique and adaptive chromatin signature is established in germ cells.

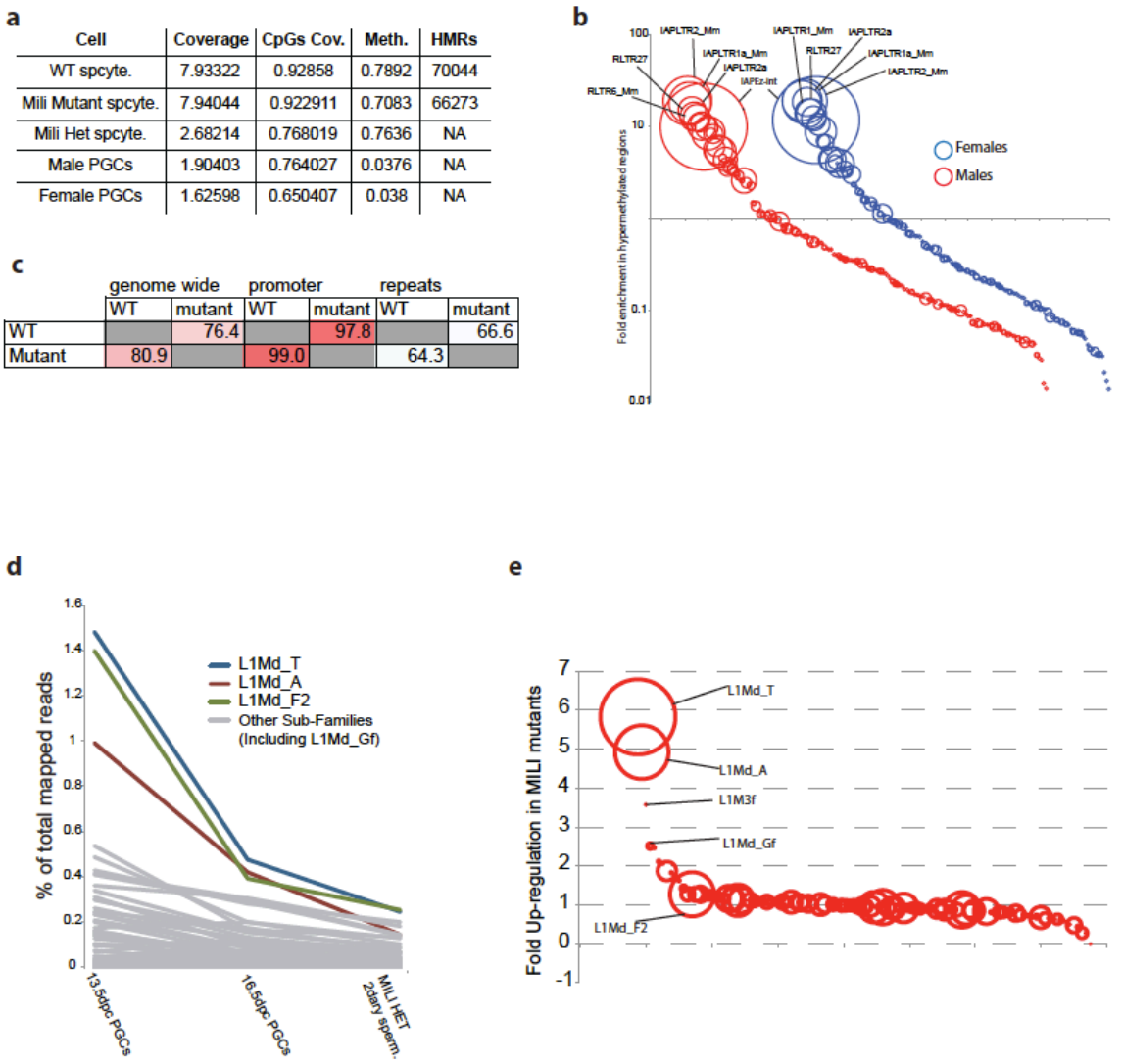
Supplementary figures

a

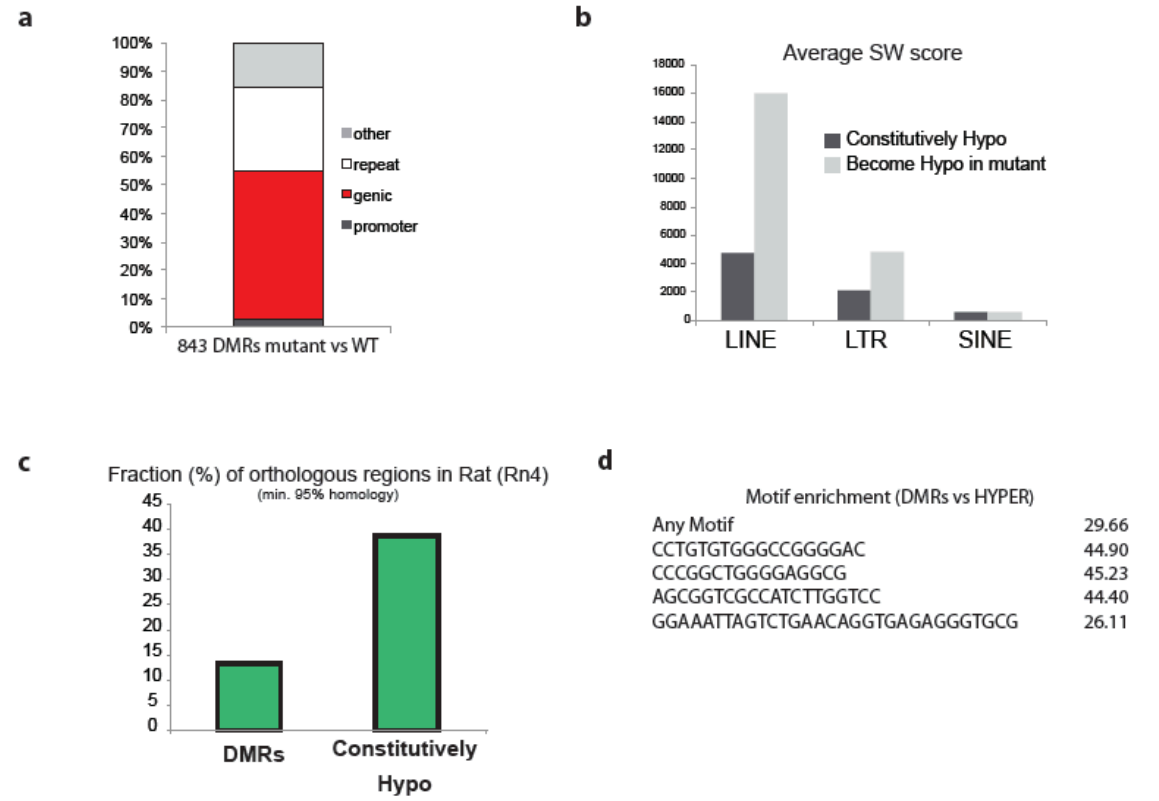
Cell	Coverage	CpGs Cov.	Meth.	HMRs
WT spcyte.	7.93322	0.92858	0.7892	70044
Mili Mutant spcyte.	7.94044	0.922911	0.7083	66273
Mili Het spcyte.	2.68214	0.768019	0.7636	NA
Male PGCs	1.90403	0.764027	0.0376	NA
Female PGCs	1.62598	0.650407	0.038	NA

c

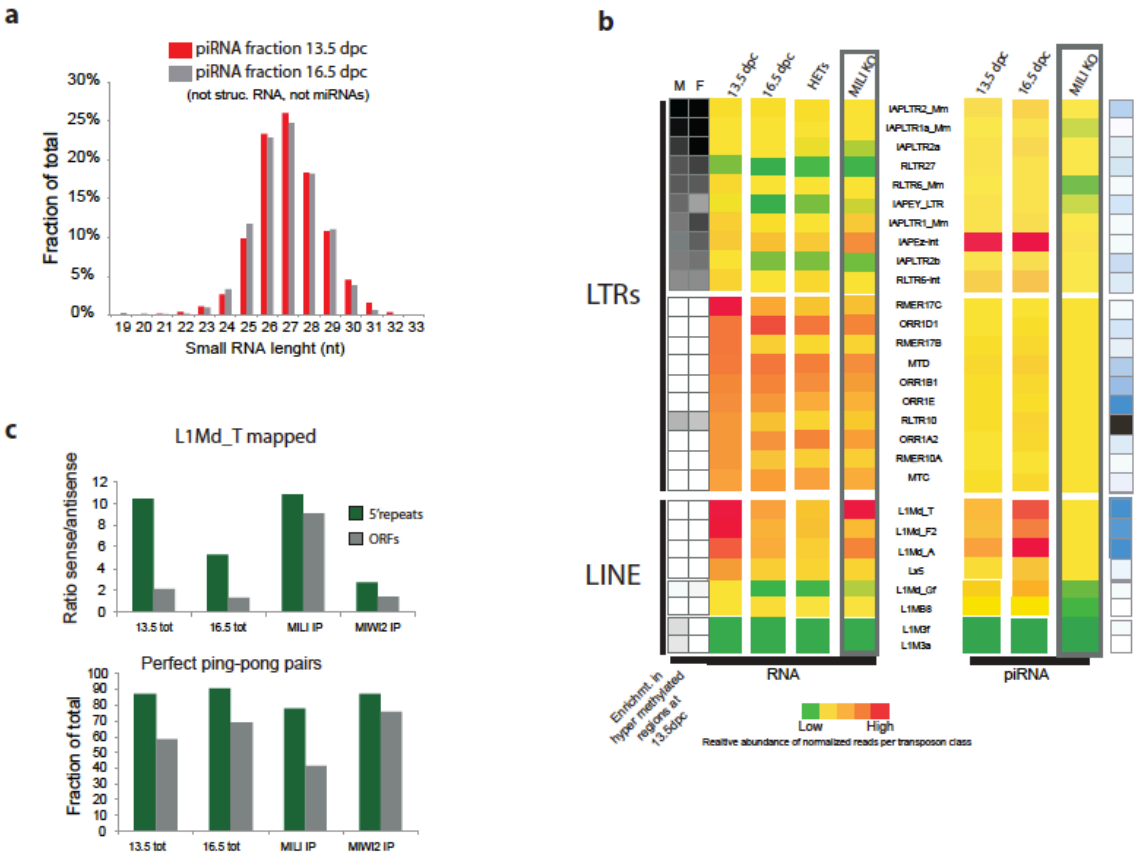
	genome wide		promoter		repeats	
	WT	mutant	WT	mutant	WT	mutant
WT		76.4		97.8		66.6
Mutant	80.9		99.0		64.3	



SupFig2.1:
a Mapping statistics of the methylomes analyzed in this study. **b** Bubble plot of LTR sub-families fold enrichement in hypermethylated domains at 13.5dpc in both males (red bubbles) and females (blue bubbles). Bubble size reflects the total number of elements called as hypermethylated in each subfamily. **c** Fraction of overlapping HMRs between WT and mili-/- spermatocytes as shown for the whole genome, promoters and repeats. Fraction as of number of HMRs from the row sample overlapping the column sample. **d** Relative expression of all L1 subfamilies as the fraction of total mapped reads in each of the shown transcriptome libraries. **e** Bubble plot showing the fold up-regulation measured for L1 subfamilies between WT and mili-/- spermatocytes. The size of the bubble reflects the sum of normalized transcript abundance in WT and mutant libraries.



SupFig2.2:
a Genomic annotation of differentially methylated regions identified between mili-/- and WT spermatocyte where WT is hypomethylated. **b** Average SW scores measured for constitutively hypomethyalted repeat copies (dark grey) and differentially methylated copies in mili-/- (light grey). **c** Fraction of repeat elements with and 95% homology in the rat genome as indicated for each group of repeats. **d** Enrichment score in all L1 for all or individual motifs characterized in differentially methylated L1s. Score shown between DMR and hypermethylated copies.



SupFig2.3:
a Size distribution of 13.5dpc piRNAs (red) compared to 16.5 piRNAs from Aravin *et al.*, 2008. **b** Heat map of normalized transcript (left map) and piRNA (right map) levels for key subfamilies of LTR and LINE retro-transposons. Heat maps were built separately for LTR and LINE. Enrichment for DMR in shown on the far right (low in white, high in dark blue). Enrichment in hypermethylated domains shown as in **Figure2.1c**. **c** Ratio of sense to antisense reads (top plot) and perfect ping-pong pairs (bottom plot) found in sequences mapping over a representative L1Md_T element (pos. chr7:62853712-62859551). Reads aligned to the 5'repeats (green) or the ORFs (grey) are shown for each analyzed libraries.

Experimental procedures:

Mouse strains: All strains used in this work were maintained on a C56Bl/6 background. For wild type secondary spermatocyte methylomes, mice were purchased form Charles River Laboratories. The *mili* knockout strain was obtained from Haifan Lin (Yale University) and is described in Kuramochi-Miyagawa et al., 2004. For PGC isolation, Oct4-EGFP mice, described in Lengner et al., 2007, were purchased form the Jackson Laboratory (Bar Harbor, Maine).

Cell sorting: Secondary and primary spermatocytes were FACS sorted (Aria II, BD Bioscience) form WT and *mili* mutant animals based on DNA content using Hoechst staining, as described in Bastos et al., 2005. Because *mili* mutant animals only produce a small 1n2C population (secondary spermatocytes), testis cells were also stained with Ep-Cam antibody (CD326, clone G8.8 from Biolegend), conjugated with Alexa 647, to enrich the 2n4C population for germ cells. PGCs at 13.5 and 16.5dpc were sorted using EGFP. GFP negative cells were also collected and referred to as "somatic" cells.

Shotgun bisulfite libraries preparation, sequencing and mapping: Shotgun bisulfite sequencing was performed as described in (Molaro et al., 2011). Briefly, purified genomic DNA was sonicated to an average size of 200-300bp, end-repaired and A-tailed using T4PNK (NEB), T4Polymerase (NEB) and Taq polymerase (Roche). *Illumina* paired-end adaptors were ligated at 25°C for 30mins using the Rapid DNA Ligase form Roche. The ligated products were bisulfite converted (EZ-methylation Gold Kit, Zymosearch) and amplified using the Illumina paired-end primers (Illumina) and the Expand High Fidelity plus PCR system (Roche). Amplicons were quantified by qPCR, and paired-end sequenced on the Illumina GAII platform (76PE and 100PE).
Sequenced Reads were mapped using RMAPs (Smith et al., 2009). HMR and DMRs calling was performed as described in Molaro et al., 2011 and Hodges et al., 2011.
Retro-transposon sub-faimily enrichment in hypermethylated domains at 13.5dpc was calculated as the ratio of Observed/Expected number of copies overlapping the domains.

Small RNA cloning: Small RNA cloning form total RNA was performed as described in Aravin et al., 2008.
Briefly, total RNAs form 13.5dpc whole gonads were extracted using Trizol (Invitrogen). Small RNAs within a 24 to 33-nt window were isolated from 12% polyacrylamide gels. 3' and 5' linkers were ligated, and products were reverse transcribed using Superscript III (Invitrogen). Following PCR amplification, libraries were submitted for sequencing using the Illumina GAII platform.

RNA-seq: RNA form sorted PGCs and adult spermatocytes were extracted using Trizol (Invotrogen). Following DNAse treatment, each sample was subjected to reverse transcription and linear amplification using the Ovation RNA-seq system according to manufacturer's protocol (Nugen). Both oligo-dT and random priming are used during this procedure. Finally, double stranded cDNAs were subjected to a standard Illumina paired-end genomic library preparation (Illumina), and sequenced to an average size of 76bp on the Illumina GAII platform.

Read Mapping of small and long RNA: After FASTQ to FASTA conversion, the Illumina adapter –CTGTAGGCACCATCAATTC for small RNA and

GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG for long RNAs - was clipped from the 3' end of the read and sequences shorter than 16 nt were discarded from further analysis. The remaining sequences were collapsed into a non-redundant list and mapped the mouse genome (mm9) using the short read aligner bowtie (Langmead et al., 2009). All further bioinformatic analysis on mapped sequences was done using Unix-based utilities (Galaxy, Goecks et al., 2010).

Transcript normalization was calculated for each transposon sub-families as the fraction of total mapped reads annotated as a given sub-family. The heat maps were colored independently for each repeat class.

To extract piRNAs mapping to DMRs, sequences in each small-RNA libraries were aligned to all L1 and ERVK associated DMR using the short read aligner bowtie.

Perfect ping-pong pairs were quantified within by counting the number of reads displaying a perfect 10nt offset with at least one other read.

2.3 DNA Methylation Profiles of Chimp and Human Sperm: A Look into Epigenetic Evolution

2.3.1 Résumé en Français

Chez les mammifères, l'embryon préimplantatoire ainsi que les cellules précurseurs de la lignée germinale voient l'ensemble de leurs marques de methylation intégralement effacées puis ré-établies *de novo*. Bien que les acteurs moléculaires impliqués dans ce processus aient déjà été caractérisés, l'influence de la séquence sous-jacente sur l'établissement des profils de methylation *de novo* n'a jamais été étudiée comparativement entre ces deux stades de développement ou dans uns contexte évolutif. La présente étude se propose de référencer et d'étudier, dans le sperme d'Homme et de Chimpanzé, les niveaux de methylation de l'ensemble des CpG du génome et de les comparer à ceux des cellules ES humaines. Dans un premiers temps, nous montrons que l'ensemble des régions hypométhylées s'étend bien au-delà des îlots CpGs et inversement, que la présence d'un îlot CpG n'est pas synonyme d'hypoméhtylation dans les cellules germinales. De plus, bien que la vaste majorité des promoteurs soit constitutivement hypométhylés aussi bien dans le sperme que dans les cellules ES, la structure des domaines hypomethylés diffère significativement entre ces deux types cellulaires. Notre étude révèle aussi la présence de milliers de copies d'éléments répétés résistants à la methylation *de novo* dans ces deux type cellulaires; certaines familles étant préférentiellement hypomethylées dans la lignée germinale. Enfin, la

comparaison des profils de methylation entre le Chimpanzé et l'Homme nous a permis de quantifier l'interdépendance existante entre l'évolution du methylome et du génome. Nous avons trouvé que des états de methylation divergents se produisent aussi bien dans des régions hautement orthologues qu'au niveau de régions polymorphiques. Dans leur ensemble, nos résultats suggèrent que des variations du génome et de l'épigénome pourraient indépendamment influencer le processus de spéciation.

2.3.2 Specific contribution to the publication

I was involved in the production and analysis of all data reported here (except for the ES cell methylome previously published by Laurent et al., 2010). The co-first author of this paper, Dr. Emily Hodges, helped with protocols, analyses and writing.

Dr. Andrew Smith developed the bisulfite sequencing mapping algorithm and performed the statistical and modeling analyses used for the study of HMRs.

2.3.3 Publication reference

Molaro A.*, Hodges E.*, Fang F., Song Q., McCombie W. R., Hannon G. J., Smith A. D. (2011). Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. **Cell** 146, 1029–1041. doi: 10.1016/j.cell.2011.08.016

* Equal contribution

Resource

Cell

Sperm Methylation Profiles Reveal Features of Epigenetic Inheritance and Evolution in Primates

Antoine Molaro,[1,3] Emily Hodges,[1,3] Fang Fang,[2] Qiang Song,[2] W. Richard McCombie,[1] Gregory J. Hannon,[1,*] and Andrew D. Smith[2,*]

[1]Howard Hughes Medical Institute, Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA
[2]Molecular and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA
[3]These authors contributed equally to this work
*Correspondence: hannon@cshl.edu (G.J.H.), andrewds@usc.edu (A.D.S.)
DOI 10.1016/j.cell.2011.08.016

SUMMARY

During germ cell and preimplantation development, mammalian cells undergo nearly complete reprogramming of DNA methylation patterns. We profiled the methylomes of human and chimp sperm as a basis for comparison to methylation patterns of ESCs. Although the majority of promoters escape methylation in both ESCs and sperm, the corresponding hypomethylated regions show substantial structural differences. Repeat elements are heavily methylated in both germ and somatic cells; however, retrotransposons from several subfamilies evade methylation more effectively during male germ cell development, whereas other subfamilies show the opposite trend. Comparing methylomes of human and chimp sperm revealed a subset of differentially methylated promoters and strikingly divergent methylation in retrotransposon subfamilies, with an evolutionary impact that is apparent in the underlying genomic sequence. Thus, the features that determine DNA methylation patterns differ between male germ cells and somatic cells, and elements of these features have diverged between humans and chimpanzees.

INTRODUCTION

In mammals, proper DNA methylation is essential for both fertility and viability of offspring (Bestor, 1998; Bourc'his and Bestor, 2004; Li et al., 1992; Okano et al., 1999; Walsh et al., 1998). DNA methylation in germ cells is required for successful meiosis (Bourc'his and Bestor, 2004), and blastocysts derived from embryonic stem cells (ESCs) lacking DNA methyltransferases (DNMTs) cannot survive past approximately 10 days of development (Li et al., 1992).

Mammalian germ cells are derived from somatic cells, rather than being set-aside during the first zygotic cleavages. During germ cell development, the genome undergoes a wave of nearly complete demethylation and remethylation (Popp et al., 2010; Walsh et al., 1998). This reprogramming event correlates with re-establishment of totipotency and with the creation of sex-specific methylation patterns at imprinted loci (reviewed by Sasaki and Matsui, 2008). Germ cell methylation patterns are erased and reset during a second wave of epigenetic reprogramming that occurs during preimplantation development. Post-fertilization, DNA methylation levels reach a nadir around the eight-cell stage, after which methylation is rewritten, attaining its somatic level by the blastocyst stage (Mayer et al., 2000). Because this is completed prior to the establishment of the inner cell mass from which cultured ESCs are derived, one can view ESCs and mature germ cells as the terminal products of the two landmark epigenetic reprogramming events in mammals.

Mobile genetic elements constitute roughly half of most mammalian genomes (Lander et al., 2001). Repression of transposons relies critically on DNA methylation and is essential for the maintenance of genomic stability in the long term and of germ cell function in the near term (Bestor, 1998; Bourc'his and Bestor, 2004; Okano et al., 1999; Walsh et al., 1998). At least in part, silencing of repeated DNA depends upon an abundant class of PIWI-associated small RNAs, called piRNAs (reviewed in Aravin and Hannon, 2008). In the absence of this pathway, methylation is lost on at least some element copies, transposons are derepressed, and germ cell development is arrested in meiosis.

CpG dinucleotides are underrepresented in mammalian genomes, most likely because a higher rate of spontaneous deamination of methylated cytosines exerts evolutionary pressure for CpG depletion by frequent CpG-to-TpG transitions (Duncan and Miller, 1980; Ehrlich et al., 1990). Mammalian genomes contain areas of relatively high CpG density, called "CpG islands" (CGIs) (Gardiner-Garden and Frommer, 1987), which have avoided CpG depletion over evolutionary time. CGIs are frequently observed at promoters and in some cases have been shown to exert regulatory effects. Thus, selection against CpG depletion may reflect the importance of specific CpG dinucleotides as sequence-based binding sites or simply the requirement for a certain regional density of CpGs. As an alternative, the existence of CGIs may simply be an artifact of longstanding hypomethylation of these regions, and consequent

Table 1. Shotgun Bisulfite Sequencing of Human and Chimp Sperm Methylomes

Species	Sample	Mapped	Distinct	Mismatches	BS Conversion	Methylation	CpG Coverage	CpGs Covered
Human	sperm (1)	609,127,589	388,835,058	1.58	0.992	0.724	8.8	0.96
	sperm (2)	588,920,777	316,860,245	1.84	0.983	0.674	7.3	0.94
	sperm (both)	1,198,048,366	705,695,303	1.70	0.988	0.701	16.1	0.96
	ESCs	940,731,922	366,844,212	0.64	0.988	0.663	14.1	0.93
Chimp	sperm (1)	459,258,834	255,193,493	1.87	0.985	0.665	6.2	0.95
	sperm (2)	520,905,232	327,796,614	1.70	0.984	0.672	7.4	0.94
	sperm (both)	980,164,066	582,990,107	1.78	0.985	0.669	13.6	0.96

Mapped: reads mapping optimally to a single location in the reference genome. Distinct: number of genomic locations to which a read maps; when multiple reads map to the same position, one with the best mapping score was selected at random, and all others discarded. Mismatches: average number of mismatches for the reads indicated in the distinct fragments column. Bisulfite (BS) conversion rate was calculated at non-CpG cytosines. Methylation: proportion of Cs in reads mapping over CpG dinucleotides.

relief from CpG erosion, in mammalian germ cells. Under this hypo-deamination model, selective pressure is independent of CpG density, per se, and CGIs may instead be a secondary consequence of protection from methylation at specific sites combined with prevalent methylation elsewhere in the genome (Cooper and Krawczak, 1989; Duncan and Miller, 1980; Ehrlich et al., 1990).

Studies encompassing evolutionarily distant species have shown that broad features of the epigenome, such as the high methylation levels of gene bodies and repeats, are deeply conserved (Zemach et al., 2010). In closely related species, however, fine-scale analysis of DNA methylation state reveals variation. The chimpanzee and human genomes share more than 95% sequence homology but display regions of differential methylation (Enard et al., 2004). Through focused studies, we have gained glimpses into the characteristics of the methylome and the evolutionary pressures that shape it. We wished to enable genome-wide comparisons of DNA methylation states in closely related species and to examine possible differences between the two major waves of epigenetic remodeling that occur during the mammalian life cycle. We therefore produced full-genome, single-CpG resolution DNA methylation profiles in human and chimp sperm and compared these with methylation maps from human ESCs (Laurent et al., 2010).

RESULTS

Methylomes of Mature Male Germ Cells in Human and Chimp

We conducted genome-wide shotgun bisulfite sequencing of sperm DNA samples isolated from two human and chimp donors (see Extended Experimental Procedures for details). Basic data analysis was conducted using a custom pipeline. We were able to determine methylation status for 96% of genomic CpGs in the human and chimp samples from a total of 28 million and 27 million CpGs, respectively (Table 1). Read coverage for CpGs on autosomes averaged 16× in human with an overall methylation level of ~70% for all CpG sites. For chimp we sequenced to an average coverage of nearly 14× and observed an average methylation level of ~67%. We did not observe significant methylation at non-CpG sites in either dataset. For comparison, we applied our analysis pipeline to a whole-genome bisulfite dataset from human ESCs (Laurent et al., 2010). This dataset was comparable to our own, with 93% of CpG dinucleotides covered and an average depth of 14× on CpGs genome-wide.

We identified contiguous domains of low methylation, termed hypomethylated regions or HMRs, in a manner independent of genomic annotations such as CGIs and promoters. Because methylation levels in sperm were generally high, HMRs appeared obvious on browser plots as valleys in which methylation dropped to very low levels. To call HMRs in a statistically principled manner, we designed a novel computational approach, based on a two-state hidden Markov model with Beta-Binomial emission distributions (see Extended Experimental Procedures). This algorithm identified ~79k HMRs in human sperm and ~70k HMRs in chimp sperm. Only ~44.5k HMRs were identified using the human ESC dataset, despite similar sequence coverage and overall methylation level (Laurent et al., 2010; see Table 1 and Table S1A available online). The sizes of HMRs also differed between germ and ESCs. In both chimp and human sperm, the mean size of HMRs was ~1.8 kb, and the median was ~1.3 kb. In ESCs, HMRs showed a mean size of ~1.2 kb with a median of 833 bp. HMRs overlapped all classes of genomic annotation (see Table S1B).

Global Comparisons among Primate Sperm Methylomes and with Human ESCs

Average methylation levels differed by a small amount among the human donors (donor 1: 72%; donor 2: 67%) but were more similar among chimp donors (donors 1 and 2: 67%). The methylation status of individual CpGs of HMRs correlated very highly between individuals, with divergence being higher in repeats as compared to promoters (Figures 1A and 1B). High interindividual correlations at the CpG and the HMR levels imply that our datasets permit accurate calling of CpG methylation genome-wide.

We also compared methylation between species at an individual nucleotide level (see Extended Experimental Procedures for details). As expected, the correlations between human and chimp sperm methylation are high, but the correlation remains generally highest within species.

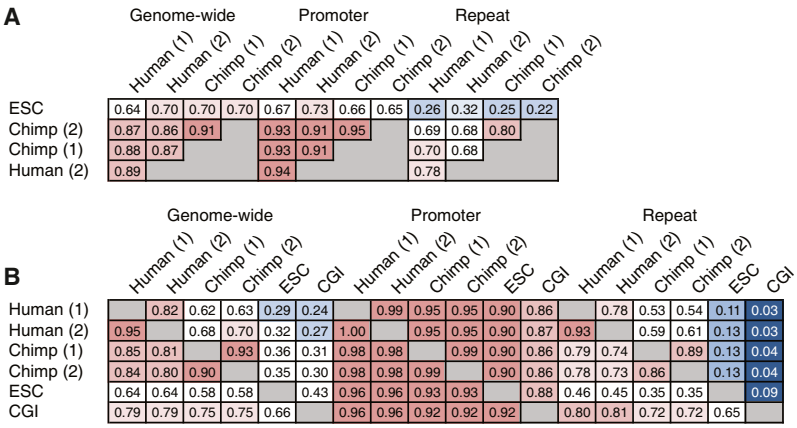


Figure 1. A Global View of Sperm and ESC Methylomes

(A) Correlations between methylomes with methylation levels measured at individual CpG sites. Correlations are displayed for CpGs genome-wide, within promoters, and within repeats, and correlation coefficients are colored blue to red to indicate low to high, respectively.

(B) Overlap between sets of HMRs from human sperm, chimp sperm, and ESC methylomes, along with annotated CGIs. Each cell gives the fraction of HMRs corresponding to the row that overlaps HMRs corresponding to the column. Colors are overlaid as in (A).

See also Table S1.

We also directly compared the methylomes from each of the human and chimp donors with the human ESC methylome. The nucleotide-level correlations between sperm methylation of each of the four primate individuals were higher than their correlations with ESC methylation patterns (Figure 1A). However, the human ESC methylome did show substantially higher correlation with the human germ cell methylomes than with those of chimp donors. Considered together these results indicate that, although waves of reprogramming in developing germ cells and embryos culminate in high genome-wide methylation, these two methylomes bear substantial differences overall.

Comparison of Hypomethylated Promoters between Sperm and ESC Methylomes

The majority of promoters are associated with HMRs in both sperm and ESCs, indicating widespread bookmarking of promoters during both waves of epigenetic reprogramming. A number of promoters did show differential methylation, with 1336 showing sperm-specific HMRs but only 201 showing ESC-specific HMRs (Figure 2A). Promoters hypomethylated in germ cells were strongly enriched for putative binding sites of transcription factors known to function in testis, including NRF1, NF-Y, YY1, and CREB (see Figure S1). A similar analysis of ESC-specific HMRs failed to yield significant results.

Only the genes with sperm-specific promoter hypomethylation revealed a strong enrichment for functional Gene Ontology (GO) categories. These were associated with germ cell functions (Figure 2B; Table S2) at distinct stages of gametogenesis (e.g., embryonic germ cell development and spermiogenesis). Thus, genes acting at developmental stages, potentially separated by decades, appear to maintain a permissive epigenetic state. Of the eight genes analyzed from the piRNA metabolic process category, seven showed promoter hypomethylation in sperm but not in ESCs, and one was hypomethylated in both (Figure 2B).

Retention of histones in human sperm was reported to be extensive (Hammoud et al., 2009). Our analysis of this data revealed a strong correlation between retained histones marked by H3K4me3 and HMRs at promoters. Among the 25.8k promoters marked by H3K4me3 in sperm, 91% overlapped an identified HMR. In general, these results support prior observations that the presence of H3K4me3 at promoters is often accompanied by hypomethylation (Hammoud et al., 2009; Ooi et al., 2007).

It was previously posited that genes involved in early embryonic development had a distinct chromatin status in sperm, being hypomethylated, histone-retained, enriched in H3K4me3 marks, and thus poised for expression (Hammoud et al., 2009). At least with respect to DNA methylation, we do not detect a preferential link between HMRs in sperm and developmental regulators but instead widespread HMRs. One potential explanation for this perceived discrepancy is that our comparisons involve sperm and ESCs, whereas prior studies used a differentiated cell type to contrast with sperm.

The genes with promoters that lack HMRs in both sperm and ESCs (n = 5,380; Figure 2A) show strong enrichment for G protein-coupled receptors and genes involved in neurological functions (Tables S2C and S2D). The reason why many of these genes, associated with highly specialized cell types, seem to lack promoter HMRs in sperm and ESCs remains obscure.

Shared HMRs Show Distinct Characteristics in Sperm and ESCs

Differences in average size and CpG densities suggest that the HMRs emerging after germ cell reprogramming differ qualitatively from those emerging after zygotic reprogramming (Figure 3A; Table S1A). The majority of HMRs have CpG density between 1% and 10%, and promoter HMRs fall almost exclusively in this range for the sperm methylomes. Those HMRs falling below 1% CpG density lie almost exclusively in repeats. These are overrepresented in human sperm relative to chimp sperm and human ESCs. Promoter-associated HMRs have sizes concentrated between 1 kb and 10 kb in human and chimp sperm, with an overall trend to be broader than promoter-associated HMRs in ESCs (Figure 3A). A notable increase in CpG density accompanies narrowing of HMRs and results in a significant portion of ESC HMRs with a CpG density above 10%.

To probe structural differences among HMRs in ESCs and sperm, we plotted the average methylation around HMR-associated transcriptional start sites (TSSs), genome-wide (Figure 3B, upper). This revealed a general principle, that a core HMR in ESCs, referred to as a nested HMR (Figure 3B, lower), often lies within an extended HMR in sperm. The median size of nested ESC HMRs is 1,498, less than half the median size of 3,109 for

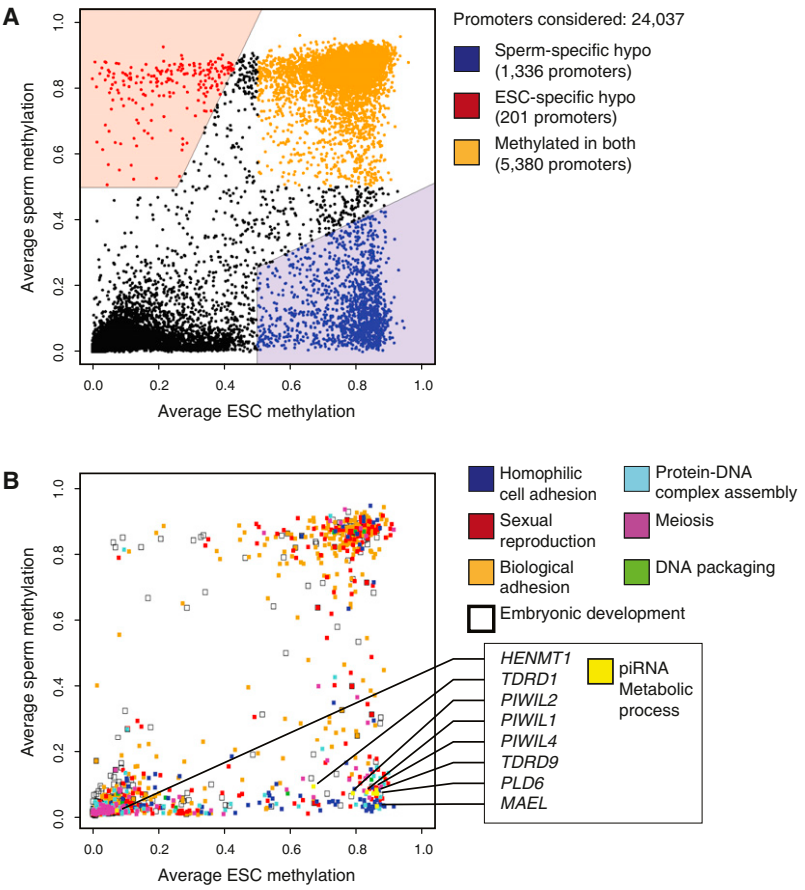


Figure 2. Differentially Reprogrammed Genes and Their Functions

(A) Average methylation through promoters (−1 kbp to +1 kbp) in human sperm and ESCs based on RefSeq gene annotations. Promoters that were hypomethylated only in sperm are shown in blue, those hypomethylated only in ESCs in red, and promoters methylated in both are shaded orange.

(B) Average methylation of promoters associated with GO terms found enriched in the sperm-specific hypomethylated fraction (see A), with the addition of genes from the "embryonic development" term. Individual genes involved in the "piRNA metabolic process" are indicated as an example.

See also Figure S1 and Table S2.

the sperm HMRs in which they reside. This phenomenon was also observed independently in a comparison of somatic and sperm HMRs, where variations in boundaries were additionally correlated with tissue-specific expression (Hodges et al., 2011). Extended HMRs are reminiscent of the concept of CpG shores (Doi et al., 2009), though in comparisons of sperm and ESCs, we made no attempt to correlate gene expression with the widespread phenomenon of nesting that we report herein.

The observation of nested HMRs could arise either from a true expansion of the hypomethylated domain in sperm or as an artifact of sperm having less precise HMR boundaries than ESCs. Examining degrees of change in methylation states across boundary CpGs in both cell types supports the former conclusion (Figure 3C). Thus, nesting appears to represent a general phenomenon and likely reflects differences in the underlying mechanisms by which the boundaries of hypomethylated regions are determined during the waves of de novo methylation that lead to sperm and ESCs.

As a step toward addressing such mechanisms, we asked whether any features are associated with HMR boundaries in either cell type. Two interesting characteristics emerged. Approaching the boundaries of either the extended sperm HMRs or the nested ESC HMRs, CpG densities dropped just prior to the start of the HMR and rose dramatically again thereafter, though overall densities were higher in the nested portions (Figure 3D). This reflects an increase in the average inter-CpG distance at the boundaries of HMRs (Figure 3E). Because our method of identifying HMRs is agnostic to inter-CpG distance, this is not simply an artifact of our approach. One could imagine increases in inter-CpG distance interrupting a processive activity, preventing the spread of de novo methylation either directly or indirectly.

Though we had no a priori expectation that sequence features would reside at sperm or ESC HMR boundaries, we searched for motifs that might occur at or near boundary CpGs, independent of CpG density. We noted a trend toward enrichment for an ACGT motif at ESC boundary CpGs with a corresponding depletion immediately outside ESC HMRs (Figure S2). This pattern was not significantly enriched at the boundaries of extended sperm HMRs. Building upon this observation, we also searched for larger motifs, focusing on those containing a central CpG core. Patterns with strong differences across HMR boundaries tended to have the ACGT core (Table S3). The most enriched pattern for sperm was AACGTT. For ESCs, we saw a well-known E box pattern, CACGTG. Plotting observed-to-expected (o/e) frequencies centered on CpGs around boundaries of extended and nested HMRs (Figure 3F), there was a clear depletion just outside each boundary followed by a sharp enrichment at the boundary CpG for each pattern in the appropriate cell type (Figure S2B). These results raise the possibility that one or more DNA-binding proteins might localize to HMR boundaries during waves of de novo methylation and help to define transitions in methylation states.

Differential Repeat Methylation in Sperm and ESCs

Consistent with prior observations and with the known role of DNA methylation in transposon silencing, most repeat elements were highly methylated in both sperm and ESCs. However, a substantial fraction of HMRs overlapped transposons in chimp and human sperm, with all repeat classes represented (Figure 4A; Table S1B). Fewer repeat-associated HMRs appeared in ESCs. In sperm, HMRs collectively contained 4%–5% of all bases assigned to repeats, compared to 1.3% in ESCs (see Table S1B). Overall, this suggests that different mechanisms,

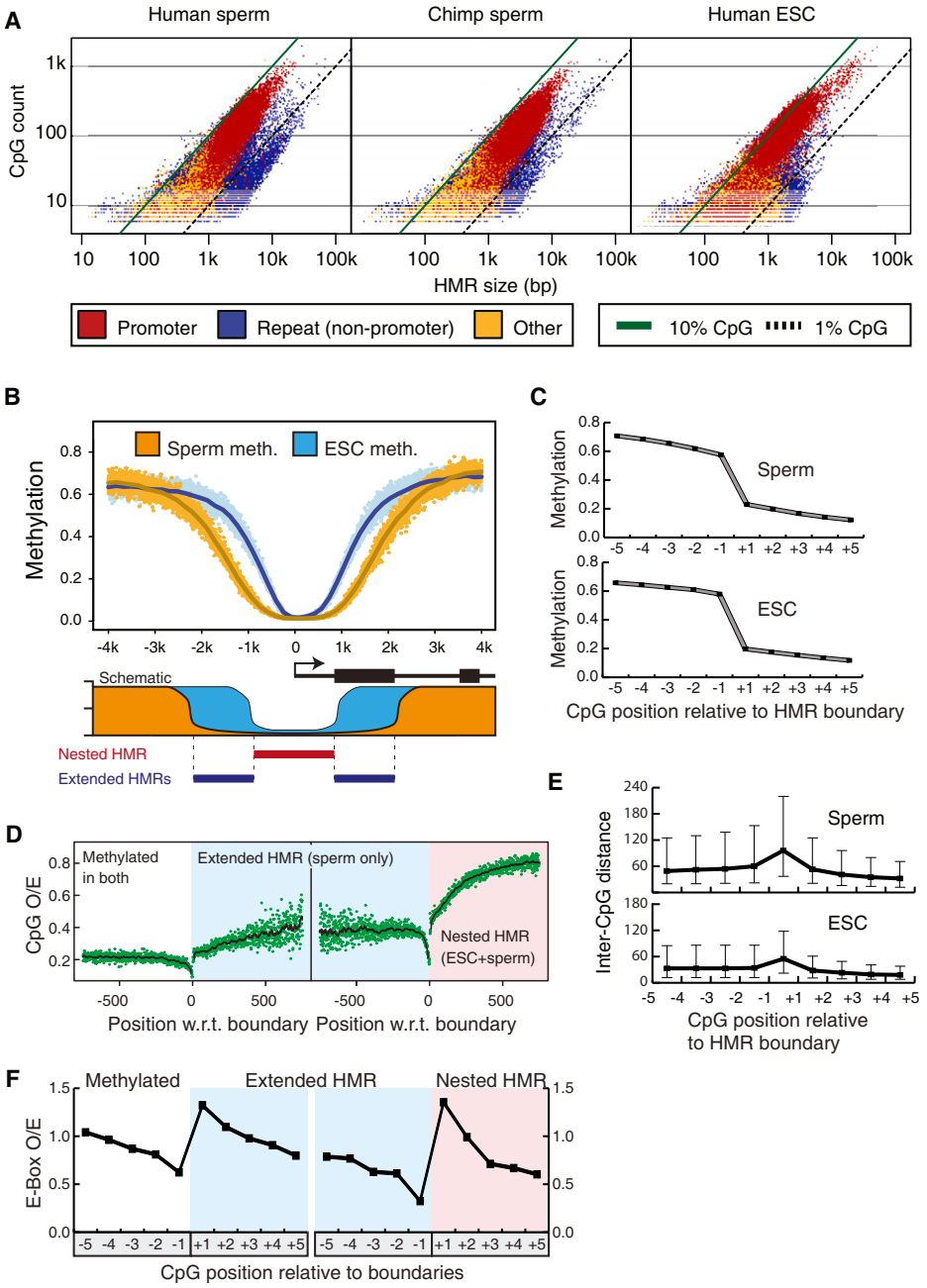


Figure 3. Characteristics of HMRs Emerging from Germline and Somatic Reprogramming

(A) Log-scale plot depicting the sizes (in bases) and numbers of CpGs for all identified HMRs in human sperm (left), chimp sperm (middle), and human ESCs (right). Diagonal lines indicate 10% CpG density (in green) and 1% CpG density (dashed line). HMRs are colored according to promoter overlap (red), overlap with repeats but not promoters (blue), or overlap with neither (orange).

(B) Average methylation around all TSS overlapping HMRs in both sperm (orange) and ESCs (blue); solid lines represent data smoothed using a 20 base sliding window. A schematic depicts the concepts of extended and nested HMRs at promoters.

(C) Average methylation at the −5 to +5 CpGs around boundaries of extended sperm HMRs and nested ESC HMRs (with the +1 CpG defined as the first inside an HMR on either side).

(D) Ratios of observed-to-expected (o/e) CpG density for each nucleotide position relative to boundaries of extended sperm HMRs (left) and nested ESC HMRs (right). Solid lines indicate values smoothed using a 20 base sliding window.

(E) Average inter-CpG distance for −5 to +5 CpGs around HMR boundaries of extended sperm and nested ESC HMRs. Upper and lower quartiles are reported for each position.

(F) Ratio of o/e frequencies of the CACGTG pattern at −5 to +5 CpGs for extended sperm and nested ESC HMRs.

See also Figure S2 and Table S3.

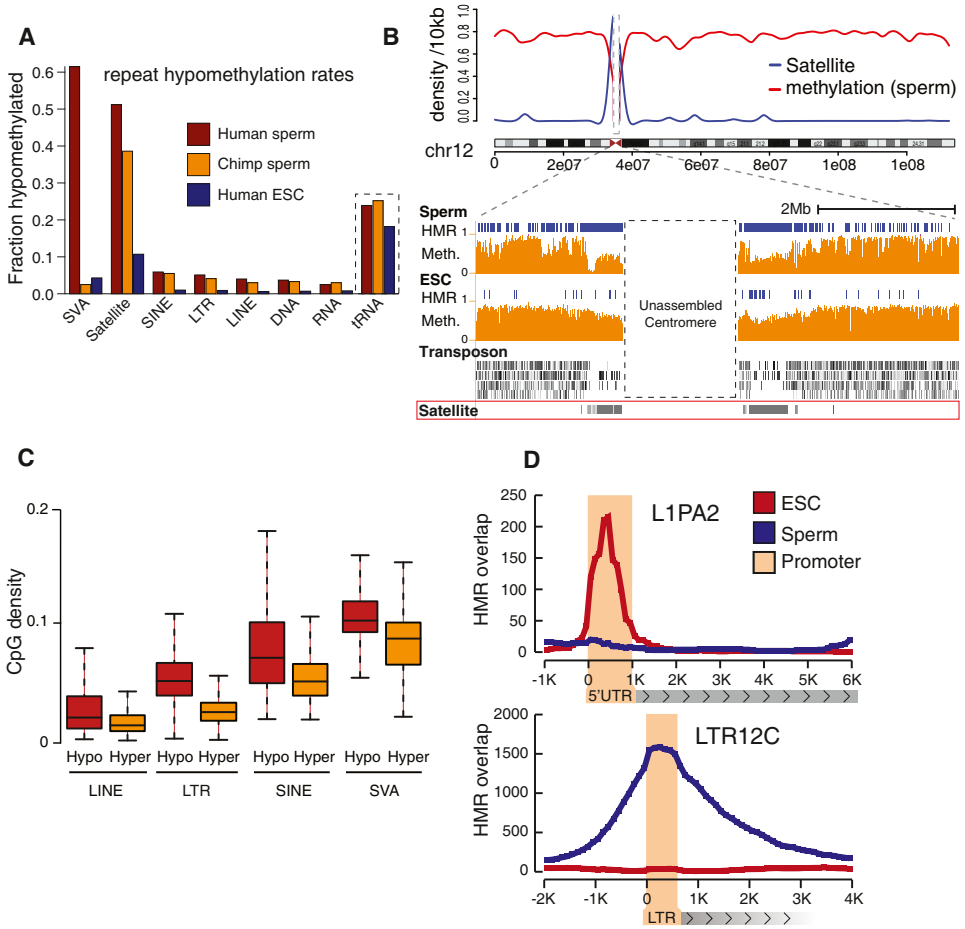


Figure 4. Differential Repeat Methylation during Male Germ Cell and Somatic Reprogramming

(A) For each repeat class, the proportion of elements that overlap HMRs is shown for human sperm (red), chimp sperm (orange), and ESCs (blue).

(B) Upper: Average methylation level (red) and satellite density (blue) in 10 kb sliding windows across chromosome 12. Lower: Chromosome 12 centromeric region with HMRs (blue) and methylation level (orange) for human sperm and ESCs.

(C) CpG densities of hypomethylated repeat copies (red) and methylated repeat copies (yellow) for LINEs, LTRs, SINEs, and SVAs.

(D) HMR overlap distribution around full-length L1PA2 and LTR12 ERV9 elements for human sperm (blue) and ESCs (red).

See also Figure S3 and Table S4.

with different stringencies, direct repeat methylation during germ cell and preimplantation development.

Sperm-Specific Satellite Hypomethylation Is Concentrated at Centromeres

We noted a strong decrease in methylation of sperm DNA within pericentromeric regions, extending several megabases outward from the unassembled core centromeres (Figure 4B). This was not seen in ESCs or in terminally differentiated cells (Hodges et al., 2011). This striking pattern was attributable to sperm-specific hypomethylation of ~75%–80% of the satellite repeats concentrated in pericentromeric regions (Figure 4A). In ESCs, only 16% of pericentromeric satellites were hypomethylated, a figure in accord with the overall hypomethylation rates of nonpericentromeric satellites in ESCs and sperm (Table S4A). Prior studies of mouse germ cells using methylation-sensitive restriction enzymes had noted selectively low methylation at pericentromeric satellites, suggesting that this is a conserved property (Yamagata et al., 2007).

Retroelement Methylation Patterns Are Determined at the Subfamily Level

Proper methylation of retrotransposons is required for transcriptional silencing of full-length and potentially active copies (Bourc'his and Bestor, 2004; Goodier and Kazazian, 2008; Walsh et al., 1998). However, specific retroelements can be active or unmethylated in male germ cells (e.g., AluY and AluYa5) (Schmid, 1991). Given our read lengths, we were able to address the methylation state of virtually all repeat families and most individual copies (see Table S4B).

Overall, retrotransposon copies that were full length or close to consensus showed a slight bias toward hypomethylation (Figures S3A and S3B). However, neither of these attributes could explain the variation observed in retrotransposon

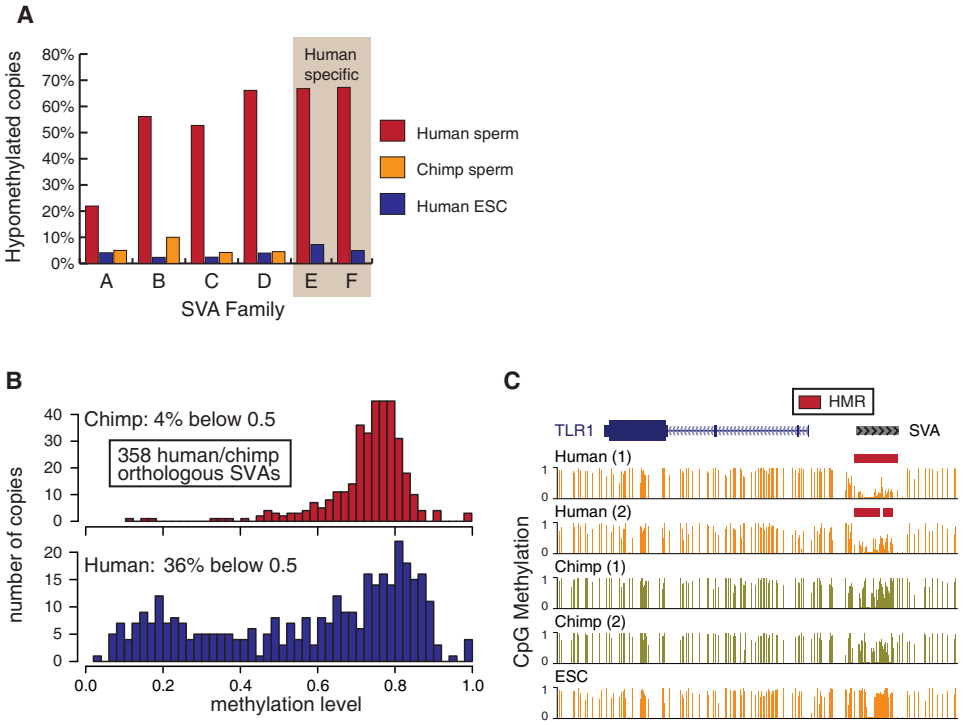


Figure 5. Divergent Methylation of SVA Elements between Human and Chimp

(A) Proportion of hypomethylated SVA copies hypomethylated according to subfamily (A to F) for human sperm (red), chimp sperm (orange), and ESCs (blue).

(B) The distribution of average methylation levels is shown for 358 human (lower) and chimp (upper) SVAs forming high-confidence orthologous pairs.

(C) An SVA insertion shared by human and chimp but with differential methylation between species.

methylation. Hypomethylated repeat copies did tend to have greater CpG density, especially within the LTR and SVA (SINE-R, VNTR, and Alu) classes (Figure 4C). For long interspersed nuclear elements (LINEs), LTR elements, and terminal repeats, HMRs concentrated within regulatory regions, which often show higher CpG density than their coding regions (Figures 4D; Figures S3C and S3D; Tables S4D and S4G). Short interspersed nuclear elements (SINEs) displayed a more uniform hypomethylation (Figure S4E). Thus, similar mechanisms appear to define HMRs in both repeat and nonrepeat portions of the genome, as for most repeats, there is a strong association of sperm HMRs with regulatory regions.

Among the LINEs, subfamilies of L1 were often hypomethylated in both sperm and ESCs, and these trended strongly toward the active groups (Tables S4E and S4H). L1PA subfamilies are considered the most active in the human genome (Khan et al., 2006), and the youngest of these (L1HS and L1PA2) were among the very few subfamilies enriched for hypomethylation in ESCs relative to sperm. Specifically in sperm, we noted hypomethylation of several other L1 families (e.g., L1PA4-16 and L1M3).

Among LTR subfamilies, sperm HMRs were enriched for ERV elements (Table S4C). Hypomethylated copies exist either as part of full-length provirus-like elements or as solo LTRs, with the greatest enrichment for LTRs belonging to "class I" elements (e.g., LTR12; see Tables S4D and S4G). The few LTR subfamilies with more hypomethylated copies in ESCs than sperm are all recently derived, human-specific ERVs (e.g., LTR5 and 13 and HERVH LTR7).

Sperm hypomethylation has been previously reported for primate Alu elements (Kochanek et al., 1993; Liu et al., 1994), and our data revealed several Alu subfamilies with differential methylation in sperm and ESCs, e,g., the AluY subfamily (Tables S4F and S4I). The more precisely defined AluYa5 (human) and AluYd4 (chimp) showed extreme enrichment for hypomethylation in sperm.

Species-Specific Methylation of the SVA Element

SVA elements showed strong, species-specific differences in methylation in human and chimp sperm (Figure 4A). SVAs are composite elements consisting of hexameric repeats, an Alu-like region, a VNTR (variable number of tandem repeats) region, and a SINE-R (Shen et al., 1994). SVA elements were active in the most recent common ancestor of chimp and human (Mills et al., 2006), and multiple examples of neoinsertions suggest that they still cause genomic rearrangements and disease in human (Ostertag et al., 2003).

Among the SVAs, the youngest subfamilies, D–F (Wang et al., 2005), showed the greatest frequency of hypomethylation in human sperm (Figure 5A). Notably, these have a higher CpG density than do older subfamilies. Three hundred and fifty-eight SVA insertions can be assigned as high-confidence orthologs between human and chimp, which remain highly similar in sequence (see Extended Experimental Procedures). Methylation

through these element copies was distributed through the full range from very low to very high average methylation, with two modes near 20% and 80% methylation (Figure 5B). In human sperm, 35% of orthologous SVAs had a methylation level below 50%. In sharp contrast, only 6% of copies fell below 50% methylation in chimp. We also annotated 921 SVA elements that appear to represent new insertions occurring after the human-chimp divergence (Mills et al., 2006). 852 (93%) of these were hypomethylated in sperm compared with only 62 (7%) in ESCs (Figure 5A). Considered together, our data indicate that SVA elements have come under different degrees of epigenetic control in the human and chimp lineages.

Many SVA insertions occur at or around promoters (Lander et al., 2001; Chimpanzee Sequencing and Analysis Consortium, 2005), and these elements often have a CpG content high enough to fit the traditional definition of a CpG island. Given their properties, SVA elements have the potential to introduce differential species- and cell type-specific methylation near genes that may be relevant for their regulation. Figure 5C exemplifies such a situation where, in the case of *TLR1*, no HMR exists near the promoter in chimp sperm or human ESCs, but one is contributed in human sperm by a nearby SVA element. Although sperm are largely transcriptionally silent, similar HMRs are expected to exist in transcriptionally active developing germ cells (data not shown).

Signatures of Selection Accompany Differential Methylation between Primates

CGIs are the most well known evolutionary signature of vertebrate DNA methylation. Their original definition required a CpG o/e ratio of at least 0.6. Although the full set of HMRs in human sperm and ESCs did not reach this empirical cut off, they did pass the 0.4 benchmark used by Weber and colleagues (Figure 6A) (Weber et al., 2007). In general, promoter-associated HMRs did surpass the 0.6 o/e cut off in both sperm and ESCs.

The differences in CpG density in nested and extended HMRs (Figure 3B) imply distinct CpG depletion pressure in these regions. Average CpG composition genome-wide is ~0.2 o/e but reaches ~0.35 in extended HMRs and 0.68 in nested HMRs. We analyzed sperm-specific and ESC-specific HMRs in an attempt to decompose the CpG depletion pressure exerted by the two methylomes. The ESC-specific HMRs reached only 0.35 o/e CpG composition, whereas the sperm-specific HMRs reached a CpG composition of 0.5.

The life cycle of a germ cell can be separated into two components. The first is the time from fertilization to the time that somatically derived primordial germ cells (PGCs) reach the genital ridge. Second is the time during which the PGC develops into a mature germ cell, which contributes to the zygote. The latter period generally spans from birth to the end of the reproductive life of the animal. Our data suggest a model in which methylation patterns present during both of these intervals shape genomic CpG distributions but indicate a greater influence of methylation profiles during germ cell maturation (Figure 6A).

We sought to measure the degree to which differential methylation could lead to CpG decay over the ~6 million years of divergent evolution separating human and chimp. We focused on regions that qualified as HMRs in either chimp or human, as these regions could have either lost methylation along one lineage or gained methylation along the other. For a given regional methylation level, we measured CpG decay as the proportion of regions having lost more than 5% of inferred ancestral CpGs (using gorilla as outgroup) and plotted the relationship between average methylation and decay rate (Figure 6B). The correlation between regional methylation level and CpG decay was extremely strong for both human and chimp. These results indicate that CpG decay is appreciable as a function of methylation even over relatively brief evolutionary periods.

This observation predicted that we might see signatures of selective pressure preventing erosion of some CpGs that are maintained despite germline methylation. To address this question, we analyzed segregating sites at CpG dinucleotides using data from the HapMap 3 project (CEU population; Altshuler et al., 2010). CpGs were treated symmetrically, so each derived allele at these sites can be classified as A, G, or T. As expected, segregating sites with T as the derived allele represent the vast majority.

We generated frequency spectra for each derived allele nucleotide with sites classified according to their methylation level in sperm (Figure S4). As methylation levels increased, derived allele frequencies shifted toward the low ends of the spectra (Figure 6C and Figure S4). This shift was observed not only for derived TpG alleles, which could be explained by an extreme bias in mutation rate, but also for ApG and GpG derived alleles. One interpretation of these findings is that selection is on average weaker at individual CpG sites with lower sperm methylation. Such an interpretation is consistent with recent findings of Cohen et al. (2011), who used sophisticated evolutionary models to posit that selection for high CpG content is not a significant factor contributing to maintenance of CGIs in the genome.

The strong connection between HMRs and gene promoters suggests that the evolutionary gain or loss of HMRs may be associated with changes in selective pressure on functional regulatory regions. To investigate this possibility, we analyzed sequence divergence in HMRs, focusing on those that are human or chimp specific. Because these differentially methylated regions will have different rates of C-to-T transitions, we counted changes from the inferred ancestor only at non-CpG sites. Genomic intervals differing by more than 1% relative to the inferred ancestor were counted as having divergent sequences.

Only 10% of HMRs shared between human and chimp showed divergence from the ancestral sequence at non-CpG sites (Figure 6D). At chimp-specific HMRs, 15% of human sequences and 19% of chimp sequences diverged from the inferred ancestor. At human-specific HMRs, 22% of human sequences diverged and 18% of chimp sequences diverged. These results indicated that changes in methylation state between human and chimp are associated with accelerated non-CpG sequence divergence. Interestingly, in both cases the species with the lower methylation state had a greater rate of divergence, which is consistent with adaptation at novel regulatory regions as a driver for these changes.

We only identified 104 promoters that are hypomethylated in human but not in chimp sperm and only 52 genes with differential promoter methylation in the opposite orientation. Neither set showed significant enrichment for any ontology category.

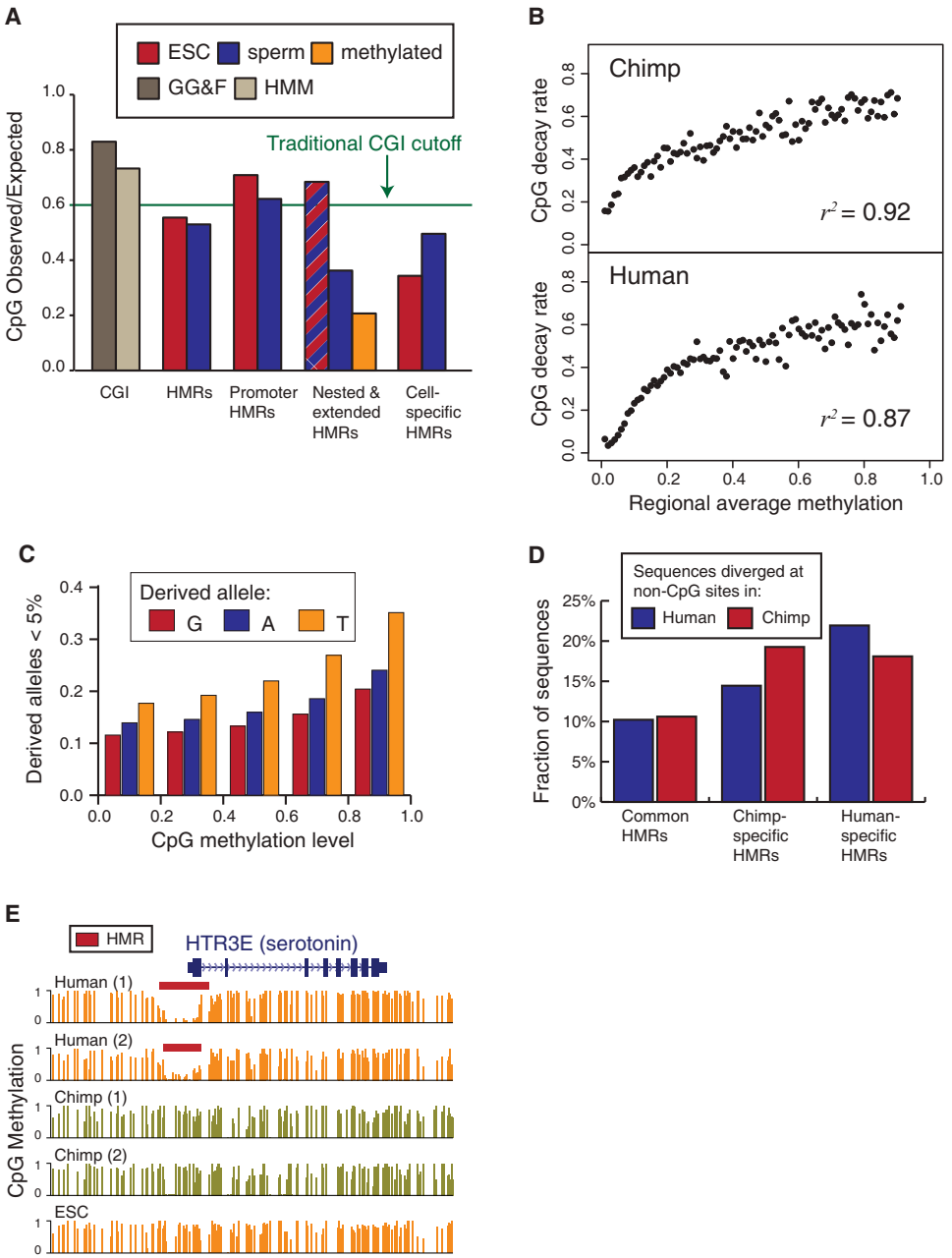


Figure 6. Sequence Features Associated with Methylome Divergence

(A) Ratio of o/e CpG density across all HMRs, those overlapping promoters, those sperm or ESC specific, and the extended/nested HMRs. Data for sperm are indicated in blue and for ESCs are indicated in red; orange indicates ratio immediately outside extended HMRs.

(B) Frequency of regions under CpG decay as a function of methylation for both human and chimp at locations of HMRs in the other species. Decay is presented for chimp in the upper panel and for human in the lower panel.

(C) Frequencies of rare derived alleles at CpG dinucleotides for each derived nucleotide, grouped according to methylation level in human sperm.

(D) Proportion of sequences displaying over 1% nucleotide divergence relative to the inferred ancestor using gorilla as an out-group and counting only non-CpG sites.

(E) The promoter of the human *HTR3E* (serotonin receptor) gene contains an HMR in both human donors but in neither chimp donor.

See also Figure S4 and Table S5.

However, analysis of genes with promoters within 10 kb of an identified human-specific sperm HMR revealed a strong enrichment for neuronal functions (see Table S5). The *HTR3E* gene, a serotonin receptor subunit, is an example of such a gene, whose promoter is selectively hypomethylated in human sperm (Figure 6E).

DISCUSSION

Sperm Methylation Patterns Are Conserved

Overall, sperm methylation patterns were highly similar in all our samples. However, there were differences, even among individuals. There has been much discussion regarding the role of germline transmission of epigenetic marks in interindividual variation (Curley et al., 2011). Changes in epigenetic state could allow flexibility in phenotype that could be reverted over short time spans if a trait became disadvantageous. Erosion of CpG content provides a mechanism to allow fixation of a positive trait in the long run. Thus, changes in DNA methylation patterns preceding changes in DNA sequence presents an attractive model for at least one mode of adaptation. Although evaluating such hypotheses will require many more datasets, the work presented here builds a firm foundation for such studies.

Most Promoters Have HMRs in Sperm

Global resetting of DNA methylation patterns happens twice during mammalian development: once during germ cell development and once early in embryogenesis. Our data permit a genome-scale analysis of these two events. Although high genome-wide levels of methylation are re-established during both waves of epigenetic remodeling, some regions are protected and establish HMR boundaries that appear relevant even in fully differentiated somatic cells (Hodges et al., 2011). A few promoters showed selective hypomethylation in sperm, and these are strongly enriched for annotations related to germ cell processes. Far fewer were selectively hypomethylated in ESCs, and these were not enriched in any particular annotation category. Promoters of genes retaining nucleosomes have recently been shown to be hypomethylated in human sperm (Hammoud et al., 2009), and both of these features have been proposed to aid rapid activation during development. We find that gene-associated hypomethylation in sperm can be extended to more than 70% of all annotated genes in both human and chimp. Among these we failed to find any enrichment for regulators of early development. Instead, it seems that promoter regions are generally identified and bookmarked in sperm (see Zaidi et al., 2010).

Distinct Processes of HMR Formation Shape Germ Cell and ESC Methylomes

Genome-wide, CpG sites seem to adopt a methylated state by default (Edwards et al., 2010). This raises the problem of precisely how regions that become HMRs are identified as such. Regions of hypomethylation at promoters have been correlated with regulatory DNA in various developmental contexts (Illingworth et al., 2008; Laurent et al., 2010; Rollins et al., 2006; Straussman et al., 2009). Based upon analysis of histone marks and on the proposed binding properties of DNMT3s (Dhayalan et al., 2010; Zhang et al., 2010), active transcription and accompanying methylation of K4 on histone H3 are thought to locally inhibit the methylation machinery. This could enable large-scale recognition of promoter regions if widespread transcription occurs during fetal germ cell development as genomic methylation patters are erased and reset. It is also plausible that specific protein/DNA complexes act locally even in the absence of active transcription, to prevent access by de novo methyltransferases. Proteins observed to function as boundary elements, such as CTCF and Sp1 (reviewed in Gaszner and Felsenfeld, 2006), provide candidates for such functions.

Despite overall similarity in the sets of promoters they mark, the HMRs observed at promoters in mature male germ cells usually extend beyond the boundaries of HMRs in ESCs when the two overlap. These wider HMRs do not seem to reflect less precision in HMR boundaries, as methylation differences across HMR boundaries are similar between sperm and ESCs. Because this "nested" HMR phenomenon is observed at so many promoters, it does not seem to be associated with the regulation of any specific genes during germ cell development. We have observed a clear increase in CpG content through the extended portion of these HMRs relative to the genome-wide average, suggesting that they have to some degree avoided pressure to decay and hence are more than a transient state. The phenomenon that we observe is similar to the concept of CpG shores (Doi et al., 2009). Perhaps the extended HMRs in germ cells presage the extent of "shores" that correlate with changes in gene expression.

Our data suggest that HMRs emerge from de novo methylation in male germ cells with sizes that differ from those that emerge from somatic reprogramming. Thus, despite involvement of similar methyltransferases and targeting of similar sets of sequences, the determinants of HMR sizes likely differ between the two reprogramming events. We have begun to see hints to the mechanisms determining such differences by comparing boundary-associated motifs in sperm and ESCs.

Transposon Hypomethylation in Sperm

It is thought that germ cell genomes must be closely guarded from the activity of mobile genetic elements. Although repeats were generally heavily methylated, we did find HMRs that overlapped repeats, and these were substantially more prevalent in sperm. We and others have characterized a conserved, small RNA-based silencing pathway, termed the piRNA pathway, that is important for recognizing and silencing mobile elements in germ cells (Aravin and Hannon, 2008). Our data indicate that both individual element copies and broader element subfamilies can evade piRNA-based silencing. Yet, both these element copies and element families are often efficiently silenced during preimplantation development. This suggests fundamental differences in the mechanisms that recognize repeats and mark them for repression during the two major waves of epigenetic reprogramming in mammals.

Examining patterns of repeat-associated HMRs is potentially enlightening. HMRs are more prevalent in younger transposon subfamilies, and the hypomethylated regions themselves tend to overlap with promoters or regulatory regions, just as they do in genes. Thus, it may be that active elements evade default methylation by being initially recognized as gene-like as a consequence of their binding transcription factors and possibly even being transcribed. In these cases, we imagine that silencing of most elements would be enforced by the piRNA pathway but that some sites, such as those we observe herein, might still escape. A number of examples can be cited in support of this hypothesis. The 5′ untranslated regions (UTRs) of the L1PA

subfamilies are known to carry conserved YY1-binding sites, whereas other recent subfamilies acquired RUNX3- and SRY-binding motifs, all of which could promote transcription in developing germ cells (Khan et al., 2006; Lee et al., 2010). Similarly, the sperm-enriched hypomethylated EVR9 LTR12 elements have been shown to bind NF-Y, MZF1, and GATA-2 in erythroid K562 cells (Yu et al., 2005). In each of these cases, HMRs within these elements tend to encompass such potential transcription factor-binding sites.

Similarly, Alu RNAs have been detected in human sperm (Kochanek et al., 1993). This suggests a potential link between Alu HMRs and the transcriptional activity of individual repeats, though previous studies also reported that the binding of SABP across Alu elements in sperm prevents their methylation (Chesnokov and Schmid, 1995). Interestingly, Alu hypomethylation is not seen in female germ cells (Liu et al., 1994) and has been proposed as one mediator of sex-specific imprints.

Centromeric Satellite Methylation

Satellites resist methylation in sperm when localized in clusters at centromeres but are generally methylated when located elsewhere even if they are clustered. This is consistent with previous observations made in mouse through the use of methylation-sensitive enzymes (Yamagata et al., 2007). Recent reports have shown that the transient transcriptional activation of paternal pericentromeric satellites was essential for centromeric heterochromatin formation in two-cell zygotes (Probst et al., 2010). This could indicate that hypomethylation of satellite repeats in male germ cell marks paternal centromeres, in a manner similar to imprinting, allowing their rapid transcriptional activation upon fertilization.

In addition to a characteristic location within chromocenters in sperm, centromeres display a distinct chromatin structure differentiating them regionally during meiosis from other chromosomal regions (reviewed by Dalal, 2009). This has prompted suggestions that centromeric chromatin states might be critical for proper meiosis, a hypothesis strongly supported by our observation of selective hypomethylation of megabase domains of centromeric satellite clusters. Prior studies have demonstrated that derepression of satellite repeats in mitotic cells creates segregation defects due to the formation of anaphase bridges (Frescas et al., 2008). Low methylation levels have also been correlated with the ability to bind cohesin complexes (Parelho et al., 2008). Considered as a whole, these observations suggest a model in which selective hypomethylation of centromeric satellites might be critical for accurate chromosome segregation during meiosis.

Differential Repeat Methylation between Species

The most striking example of species-specific methylation to emerge from our analysis involved the SVA elements. These primate-specific composite elements contain a high density of CpGs, remain active in human and chimp, and include many copies that are clear orthologs between human and chimp (Bantysh and Buzdin, 2009; Mills et al., 2006). Transduction of SVAs has been implicated in human diseases and gene formation (Damert et al., 2009; Ostertag et al., 2003). Our results indicate that for a subset of SVA elements, the ability to methylate these elements has either been acquired along the chimp lineage or lost in the human lineage during the past 6 million years, despite very little sequence change in these elements.

Mutual Canalization of the Genome and the Epigenome

It has been thought that CGIs arose as the result of protection from methylation-associated deamination over long evolutionary periods. This is consistent with the observed correlation between the location of CGIs and regions that lack methylation in both germline and somatic cells. However, recent results have pointed to functions for CGIs that may be associated with their high CpG density (Thomson et al., 2010), with the plausible interpretation that selection may be acting to preserve CpG density in CGIs. We find that although most CGIs fall within HMRs of sperm, most HMRs extend well beyond the annotated CGIs, even using weaker CGI definitions. Thus, hypomethylated regions in male germ cells do not appear to require a critical CpG density to avoid methylation. Instead, our results are consistent with CGIs arising as a consequence of different mutational pressures rather than selection for CpG density.

In our datasets, signatures of deamination-induced CpG depletion are clear. Yet we also observe CpG depletion from many sperm and ESC HMRs. Several scenarios could resolve this conundrum. For example, such regions may have been methylated for substantial periods prior to assuming their unmethylated status. Thus, they may have decayed at some time in the past but are now stabilized by their hypomethylated status. Such sites could also actually be methylated during a period of germ cell development to which our current datasets are blind (e.g., in fetal gonocytes or female germ cells). In accord with this explanation, we have observed distinct CpG densities associated with sperm-specific and ESC-specific HMRs. Moreover, at HMRs where the only central, nested portion is hypomethylated in ESCs, we observe greater CpG retention through regions hypomethylated in both ESCs and sperm. Overall, we cannot exclude a model in which selection acts to preserve critical functions requiring specific local CpG densities. However, our results lend additional support to recent conclusions of Cohen et al. (2011), whose sophisticated evolutionary modeling showed that CGIs can be explained without invoking selection on CpG sites. Our results suggest a refinement of the hypo-deamination model in which CpG retention is a function of the time spent hypomethylated during each generation in germ cells and their somatic precursors.

The detailed comparative analysis performed here has revealed that, over the ~6 million years since the divergence of human and chimp, most patterns of DNA methylation remain conserved in male germ cells. We have directly related evolutionary changes in CpG methylation with loss of CpG dinucleotides and have shown that even small differences in methylation can lead to substantial loss of CpGs over relatively short evolutionary periods. At the same time, there are many genomic regions that are highly conserved in sequence yet show quite different patterns of methylation. This could indicate an ability of the genome and the epigenome to evolve independently. However, we do find that the most drastic changes in methylation between human and chimp, where an HMR in one species shows high levels of methylation in the other, are accompanied

by an increased sequence divergence even at non-CpG dinucleotides. One interpretation is that most species-specific HMRs have arisen newly along one lineage with these novel functional elements showing signs of recent adaptation. On the other hand, if this accelerated sequence change were more a reflection of relaxed selective pressure, we would expect species-specific HMRs to more frequently result from loss of functional elements along the opposite lineage. Resolution of these questions can only come from a broadening to many more species of the studies reported herein.

EXPERIMENTAL PROCEDURES

Detailed methods can be found in the Extended Experimental Procedures.

Sperm Collection

Two anonymous human donors were used and data pooled after sequencing. Two chimp donors were used. Semen was collected at the New Iberia Research Center (New Liberia, LA) or the Southwest National Primate Research Center (San Antonio, TX, USA). Coagulated semen was separated from the liquid phase manually. Both human and chimp samples were diluted (1:1) in HBS buffer (0.01M HEPES, ph 7.4; 150 mM NaCl) and passed though a silica-based gradient, SpermFilter (Cryobiosystems), by centrifugation (according to manufacturer's instructions).

Library Preparation

DNA from ~100 million cells was extracted and sheared to a size of ~150–200 nt by sonication. Double-stranded DNA fragments were end repaired, A-tailed, and ligated to methylated Illumina adaptors. Ligated fragments were bisulfite converted using the EZ-DNA Methylation-Gold Kit (Zymo research). Following PCR enrichment, fragments of 340 to 360 bp were size selected and sequenced.

Computational Methods

Reads were mapped with RMAPBS (Smith et al., 2009). The accuracy of our mapping method is discussed in the Extended Experimental Procedures. Mapped reads were used to infer the methylation frequency at each CpG dinucleotide. These frequencies, along with the number of reads contributing to each frequency estimate, were supplied to a segmentation algorithm used to identify HMRs. Ortholog mapping between human and chimp was done with the liftOver tool available through the UCSC Genome Browser. Sequence conservation between human, chimp, and was measured based on MULTIZ 44-way vertebrate alignments, also available through the UCSC Genome Browser. Complete details of all computational methods are provided in the Extended Experimental Procedures.

ACCESSION NUMBERS

Data analyzed herein have been deposited in GEO with accession GSE30340.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, four figures, and five tables and can be found with this article online at doi:10.1016/j.cell.2011.08.016.

ACKNOWLEDGMENTS

We thank Michelle Rooks, Pramod Thekkat, and Colin Malone for help with experimental procedures and Assaf Gordon, Luigi Manna, and the CSHL and USC High Performance Computing Centers for computational support. We thank Babette Fontenot (New Iberia Research Center) and Jerilyn Pecotte (Southwest National Primate Center) for help with chimp sperm collection. We thank Sergey Nuzhdin, Ed Green, Peter Calabrese, Maren Friesen, Magnus Norborg, and Marie-Stanislas Remigereau for helpful discussions. This work was supported in part by grants from the NIH (R01HG005238 and 1RC2HD064459) and by a kind gift from Kathryn W. Davis.

Received: December 16, 2010
Revised: May 9, 2011
Accepted: August 10, 2011
Published: September 15, 2011

REFERENCES

Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., et al; International HapMap 3 Consortium. (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.

Aravin, A.A., and Hannon, G.J. (2008). Small RNA silencing pathways in germ and stem cells. Cold Spring Harb. Symp. Quant. Biol. *73*, 283–290.

Bantysh, O.B., and Buzdin, A.A. (2009). Novel family of human transposable elements formed due to fusion of the first exon of gene MAST2 with retrotransposon SVA. Biochemistry (Mosc.) *74*, 1393–1399.

Bestor, T.H. (1998). Cytosine methylation and the unequal developmental potentials of the oocyte and sperm genomes. Am. J. Hum. Genet. *62*, 1269–1273.

Bourc'his, D., and Bestor, T.H. (2004). Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. Nature *431*, 96–99.

Chesnokov, I.N., and Schmid, C.W. (1995). Specific Alu binding protein from human sperm chromatin prevents DNA methylation. J. Biol. Chem. *270*, 18539–18542.

Chimpanzee Sequencing and Analysis Consortium. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. Nature *437*, 69–87.

Cohen, N.M., Kenigsberg, E., and Tanay, A. (2011). Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. Cell *145*, 773–786.

Cooper, D.N., and Krawczak, M. (1989). Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. Hum. Genet. *83*, 181–188.

Curley, J.P., Mashoodh, R., and Champagne, F.A. (2011). Epigenetics and the origins of paternal effects. Horm. Behav. *59*, 306–314.

Dalal, Y. (2009). Epigenetic specification of centromeres. Biochem. Cell Biol. *87*, 273–282.

Damert, A., Raiz, J., Horn, A.V., Löwer, J., Wang, H., Xing, J., Batzer, M.A., Löwer, R., and Schumann, G.G. (2009). 5′-Transducing SVA retrotransposon groups spread efficiently throughout the human genome. Genome Res. *19*, 1992–2008.

Dhayalan, A., Rajavelu, A., Rathert, P., Tamas, R., Jurkowska, R.Z., Ragozin, S., and Jeltsch, A. (2010). The Dnmt3a PWWP domain reads histone 3 lysine 36 trimethylation and guides DNA methylation. J. Biol. Chem. *285*, 26114–26120.

Doi, A., Park, I.H., Wen, B., Murakami, P., Aryee, M.J., Irizarry, R., Herb, B., Ladd-Acosta, C., Rho, J., Loewer, S., et al. (2009). Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. Nat. Genet. *41*, 1350–1353.

Duncan, B.K., and Miller, J.H. (1980). Mutagenic deamination of cytosine residues in DNA. Nature *287*, 560–561.

Edwards, J.R., O'Donnell, A.H., Rollins, R.A., Peckham, H.E., Lee, C., Milekic, M.H., Chanrion, B., Fu, Y., Su, T., Hibshoosh, H., et al. (2010). Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. Genome Res. *20*, 972–980.

Ehrlich, M., Zhang, X.Y., and Inamdar, N.M. (1990). Spontaneous deamination of cytosine and 5-methylcytosine residues in DNA and replacement of 5-methylcytosine residues with cytosine residues. Mutat. Res. *238*, 277–286.

Enard, W., Fassbender, A., Model, F., Adorján, P., Pääbo, S., and Olek, A. (2004). Differences in DNA methylation patterns between humans and chimpanzees. Curr. Biol. *14*, R148–R149.

Frescas, D., Guardavaccaro, D., Kuchay, S.M., Kato, H., Poleshko, A., Basrur, V., Elenitoba-Johnson, K.S., Katz, R.A., and Pagano, M. (2008). KDM2A represses transcription of centromeric satellite repeats and maintains the heterochromatic state. Cell Cycle *7*, 3539–3547.

Gardiner-Garden, M., and Frommer, M. (1987). CpG islands in vertebrate genomes. J. Mol. Biol. *196*, 261–282.

Gaszner, M., and Felsenfeld, G. (2006). Insulators: exploiting transcriptional and epigenetic mechanisms. Nat. Rev. Genet. *7*, 703–713.

Goodier, J.L., and Kazazian, H.H., Jr. (2008). Retrotransposons revisited: the restraint and rehabilitation of parasites. Cell *135*, 23–35.

Hammoud, S.S., Nix, D.A., Zhang, H., Purwar, J., Carrell, D.T., and Cairns, B.R. (2009). Distinctive chromatin in human sperm packages genes for embryo development. Nature *460*, 473–478.

Hodges, E., Molaro, A., Dos Santos, C.O., Thekkat, P., Song, Q., Uren, P., Park, J., Butler, J., Rafii, S., McCombie, W.R., Smith, A.D., and Hannon, G.J. (2011). Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. Mol. Cell. Published online September 15 2011. 10.1016/j.cell.2008.06.028.

Illingworth, R., Kerr, A., Desousa, D., Jørgensen, H., Ellis, P., Stalker, J., Jackson, D., Clee, C., Plumb, R., Rogers, J., et al. (2008). A novel CpG island set identifies tissue-specific methylation at developmental gene loci. PLoS Biol. *6*, e22.

Khan, H., Smit, A., and Boissinot, S. (2006). Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Res. *16*, 78–87.

Kochanek, S., Renz, D., and Doerfler, W. (1993). DNA methylation in the Alu sequences of diploid and haploid primary human cells. EMBO J. *12*, 1141–1151.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al; International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C.T., Low, H.M., Kin Sung, K.W., Rigoutsos, I., Loring, J., et al. (2010). Dynamic changes in the human methylome during differentiation. Genome Res. *20*, 320–331.

Lee, S.H., Cho, S.Y., Shannon, M.F., Fan, J., and Rangasamy, D. (2010). The impact of CpG island on defining transcriptional activation of the mouse L1 retrotransposable elements. PLoS ONE *5*, e11353.

Li, E., Bestor, T.H., and Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell *69*, 915–926.

Liu, W.M., Maraia, R.J., Rubin, C.M., and Schmid, C.W. (1994). Alu transcripts: cytoplasmic localisation and regulation by DNA methylation. Nucleic Acids Res. *22*, 1087–1095.

Mayer, W., Niveleau, A., Walter, J., Fundele, R., and Haaf, T. (2000). Demethylation of the zygotic paternal genome. Nature *403*, 501–502.

Mills, R.E., Bennett, E.A., Iskow, R.C., Luttig, C.T., Tsui, C., Pittard, W.S., and Devine, S.E. (2006). Recently mobilized transposons in the human and chimpanzee genomes. Am. J. Hum. Genet. *78*, 671–679.

Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell *99*, 247–257.

Ooi, S.K., Qiu, C., Bernstein, E., Li, K., Jia, D., Yang, Z., Erdjument-Bromage, H., Tempst, P., Lin, S.P., Allis, C.D., et al. (2007). DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. Nature *448*, 714–717.

Ostertag, E.M., Goodier, J.L., Zhang, Y., and Kazazian, H.H., Jr. (2003). SVA elements are nonautonomous retrotransposons that cause disease in humans. Am. J. Hum. Genet. *73*, 1444–1451.

Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H.C., Jarmuz, A., Canzonetta, C., Webster, Z., Nesterova, T., et al. (2008). Cohesins functionally associate with CTCF on mammalian chromosome arms. Cell *132*, 422–433.

Popp, C., Dean, W., Feng, S., Cokus, S.J., Andrews, S., Pellegrini, M., Jacobsen, S.E., and Reik, W. (2010). Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. Nature *463*, 1101–1105.

Probst, A.V., Okamoto, I., Casanova, M., El Marjou, F., Le Baccon, P., and Almouzni, G. (2010). A strand-specific burst in transcription of pericentric satellites is required for chromocenter formation and early mouse development. Dev. Cell *19*, 625–638.

Rollins, R.A., Haghighi, F., Edwards, J.R., Das, R., Zhang, M.Q., Ju, J., and Bestor, T.H. (2006). Large-scale structure of genomic methylation patterns. Genome Res. *16*, 157–163.

Sasaki, H., and Matsui, Y. (2008). Epigenetic events in mammalian germ-cell development: reprogramming and beyond. Nat. Rev. Genet. *9*, 129–140.

Schmid, C.W. (1991). Human Alu subfamilies and their methylation revealed by blot hybridization. Nucleic Acids Res. *19*, 5613–5617.

Shen, L., Wu, L.C., Sanlioglu, S., Chen, R., Mendoza, A.R., Dangel, A.W., Carroll, M.C., Zipf, W.B., and Yu, C.Y. (1994). Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and the C4B genes in the HLA class III region. Molecular cloning, exon-intron structure, composite retroposon, and breakpoint of gene duplication. J. Biol. Chem. *269*, 8466–8476.

Smith, A.D., Chung, W.Y., Hodges, E., Kendall, J., Hannon, G., Hicks, J., Xuan, Z., and Zhang, M.Q. (2009). Updates to the RMAP short-read mapping software. Bioinformatics *25*, 2841–2842.

Straussman, R., Nejman, D., Roberts, D., Steinfeld, I., Blum, B., Benvenisty, N., Simon, I., Yakhini, Z., and Cedar, H. (2009). Developmental programming of CpG island methylation profiles in the human genome. Nat. Struct. Mol. Biol. *16*, 564–571.

Thomson, J.P., Skene, P.J., Selfridge, J., Clouaire, T., Guy, J., Webb, S., Kerr, A.R., Deaton, A., Andrews, R., James, K.D., et al. (2010). CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature *464*, 1082–1086.

Walsh, C.P., Chaillet, J.R., and Bestor, T.H. (1998). Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. Nat. Genet. *20*, 116–117.

Wang, H., Xing, J., Grover, D., Hedges, D.J., Han, K., Walker, J.A., and Batzer, M.A. (2005). SVA elements: a hominid-specific retroposon family. J. Mol. Biol. *354*, 994–1007.

Weber, M., Hellmann, I., Stadler, M.B., Ramos, L., Pääbo, S., Rebhan, M., and Schübeler, D. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat. Genet. *39*, 457–466.

Yamagata, K., Yamazaki, T., Miki, H., Ogonuki, N., Inoue, K., Ogura, A., and Baba, T. (2007). Centromeric DNA hypomethylation as an epigenetic signature discriminates between germ and somatic cell lineages. Dev. Biol. *312*, 419–426.

Yu, X., Zhu, X., Pi, W., Ling, J., Ko, L., Takeda, Y., and Tuan, D. (2005). The long terminal repeat (LTR) of ERV-9 human endogenous retrovirus binds to NF-Y in the assembly of an active LTR enhancer complex NF-Y/MZF1/GATA-2. J. Biol. Chem. *280*, 35184–35194.

Zaidi, S.K., Young, D.W., Montecino, M.A., Lian, J.B., van Wijnen, A.J., Stein, J.L., and Stein, G.S. (2010). Mitotic bookmarking of genes: a novel dimension to epigenetic control. Nat. Rev. Genet. *11*, 583–589.

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

Zhang, Y., Jurkowska, R., Soeroes, S., Rajavelu, A., Dhayalan, A., Bock, I., Rathert, P., Brandt, O., Reinhardt, R., Fischle, W., and 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 Res. *38*, 4246–4253.

Cell

EXTENDED EXPERIMENTAL PROCEDURES

Mapping Reads

Reads were mapped using the RMAPBS program. Our pipeline first removed adaptor sequence from any reads, discarding any reads with fewer than 40 high-quality bases after the adaptor was removed (reads were required to have at least 10 bases of overlap with the adaptor for any part to be trimmed). Ends of paired-end reads were mapped separately, and because adaptors were ligated to fragments prior to bisulfite treatment, the first end of each paired-end read was mapped using T→C wild-cards, and the second end of each read was mapped allowing A→G wild-cards (for details, see Smith et al., 2009). We allowed up to 10 mismatches when mapping reads, though the average was substantially lower, and low-quality positions in reads were never counted as a mismatch (recall that at least 40 high-quality positions were required). For each read, the mapping location was determined to be the location with the fewest mismatches. Reads for which two locations had the minimum number of mismatches were considered to map ambiguously and discarded.

In sequencing from the same library preparation, when multiple reads mapped to the exact same location, which we refer to as duplicate reads, we assumed these represent the same original molecule (e.g., PCR products of the same fragment). We discarded all but one read in the case of duplicates and retained the one with the fewest mismatches. This step of removing duplicates was only done prior to combining data from different library preparations. For paired-end reads, after mapping ends separately, any pairs found to overlap (indicating the original fragment had length less than 202 bases) were collapsed to prevent counting the same information twice in later analysis.

The reference genomes used were the hg18 (human) and panTro2 (chimp) genomes downloaded from the UCSC Genome Browser, and we excluded alternate haplotype sequences and "random" sequences for human. For chimp we excluded "random" sequences and the "unassembled" chromosome.

Accuracy of the Mapping Method

We conducted a simulation experiment to determine the portion of reads expected to be mapped to incorrect locations using the mapping method described above. The simulation used parameters for the following values:

- *Number of reads*. We set this value to 1 M.
- *Read length*. We used a read length of 101 nt (corresponding to the majority of our sequencing runs).
- *Methylation level*. Each CpG in sampled reads was considered methylated with probability 0.7. Although this does not simulate a specific methylation level for any given genomic CpG, the effect on mapping accuracy is the same.
- *Bisulfite conversion*. We set the simulation bisulfite conversion rate to 0.98, meaning that 98% of Cs that were not simulated as methylated were converted to Ts.
- *Sequencing errors*. We set the maximum number of sequencing errors per reads to 10. Each simulated read had 10 positions for errors sampled at random (though not uniformly; see below) with replacement. Errors were introduced after simulated bisulfite conversion.
- *Error distribution*. We used the error probabilities produced by the sequencing instrument in a 101 nt sequencing run to calibrate the probabilities for simulated errors occurring at any given position in the read. This results in a greater proportion of errors at the 3′ ends of simulated reads.

The simulation was done with human genome assembly hg18 (from UCSC Genome Browser) excluding unassembled centromeric regions. Simulated reads were mapped back to the genome using the procedure described above. Of the 1 million reads, 939,605 mapped back uniquely (94%). The portion mapping back to their location of origin was 935582 (99.6%). Because of sampling error positions with replacement, along with the nonuniform distribution for error locations, the average number of mismatches was 4.6 per mapped read, substantially greater than the average number of mismatches in our data. From this we conclude that any error introduced into downstream analysis by reads mapped to incorrect locations is sufficiently small to be negligible.

Association between Sets of Genomic Regions and Annotations

We stratified measures about CpG content and methylation in genomic regions according to their association with certain genomic annotations as follows. First we defined these associations so that they partition the set of regions in question. In other words, our definitions ensured that no HMR would be associated with both a promoter and a repeat element, even though a repeat could clearly exist inside the promoter of a gene. Our definitions were as follows:

- Promoter: Any region that overlaps the interval within 1 Kb of the transcription start site (TSS).
- Gene-proximal: Any nonpromoter region that overlaps the interval starting 10 Kb upstream of a TSS or 10 Kb downstream of a transcription termination site.
- Intergenic repeat: Any nonpromoter, non-gene-proximal region that overlaps a repeat.
- Intergenic nonrepeat: Any nonpromoter, non-gene-proximal region that does not overlap a repeat.

Repeat Definitions

We analyzed the following classes of repeats: LINE, SINE, LTR, Satellite, DNA, RNA, SVA, tRNA, low-complexity, and simple repeats. This list includes most of the repeats annotated in the RepeatMasker track from the UCSC Genome Browser.

SVA Elements with Identifiable Orthologs

We used SVA annotations from UCSC Genome Browser, which are based on RepBase. These annotations are constructed by matching repeat consensus sequences to the reference genome (hg18 and panTro2). SVA elements were retained in human if:

(1) The interval covered by the human copy lifts over to chimp
(2) The lift over target (in chimp) lifts back to human
(3) The target when lifting back from chimp to human is the same as the original interval

The same criteria were applied to chimp. This set of SVA elements was used in Figure 4A. This highly conservative criteria allowed us to compare methylation levels through copies of SVAs that existed in both species. The total number of these SVA copies included 358 pairs of high-confidence orthologs. The trends observed for this small, high-confidence set of elements is also reflected in the full sets of elements for human and chimp.

Calculation of Basic Statistics

Discarding Low-Quality Reads

Reads were first checked for the presence of adaptor sequence, indicating that the sequenced fragment was too short and sequencing proceeded into the adaptor at the other end of the fragment. We required at least a 10 base match starting from the beginning of the adaptor, excluding Ns in reads and allowing up to 2 mismatches. When such an adaptor sequence was found in a read, the read was trimmed after the beginning position of the match by replacing all subsequent bases (in the 3′ direction) with an N, which would not induce a mismatch during alignment. Any reads for which the final non-N base was at position 40 or less was discarded. Finally, any read with fewer than 28 non-N bases through its entire length of the read was discarded.

Estimating CpG Methylation Levels

For CpG i, define m_i as the number of reads showing methylation over position i, counting both strands. Define u_i as the number of reads showing lack of methylation over CpG i. The methylation level is estimated as $m_i/(m_i + u_i)$, which is an estimate of the probability that CpG i is methylated in a molecule sampled randomly from the cell population. Because CpG methylation is symmetric, m_i and u_i include observations associated with the cytosines on both strands for the *i-th* CpG.

Depth of Coverage and Bisulfite Conversion

All our measures of coverage are in terms of CpGs. Depth of coverage (fold coverage) is also measured only at CpGs and counts only T or C nucleotides (A or G for the second end of each read). Both these numbers are reflective of numbers calculated using all assembled bases. Bisulfite conversion is measured as the sum of the number of non-CpG cytosines that are converted to Ts (as indicated by Ts in reads mapping over non-CpG cytosines in the genome), divided by the total number of non-CpG cytosines in uniquely mapped reads.

Identifying Hypomethylated Regions

We identified hypomethylated regions (HMRs) using a stochastic segmentation to partition the methylome into alternating regions of hypermethylation and hypomethylation, the latter appearing as valleys in visual depictions of methylation profiles. More specifically, our method is based on a Hidden Markov Model (HMM; Durbin et al., 1999).

Our HMM consists of two states (for high and low methylation). To model the observations made at each individual CpG we use the following distributions. For a sequence of n CpGs in a contiguous chromosomal region, let p_i denote the true probability that CpG i is methylated in a molecule chosen at random from the sequenced sample. We assume that $p_i \sim \text{Beta}(\alpha, \beta)$. The BS-seq data provides the numbers m_i and u_i of methylated and unmethylated reads, respectively, from which we estimate $\hat{p}_i = m_i/(m_i + u_i)$. In calculating likelihoods of observations from a particular state (i.e., the emission distribution), we use a Beta-Binomial distribution. That is, we assume $m_i \sim \text{BetaBinom}(\alpha, \beta, m_i + u_i)$, and

$$\Pr(m_i|\alpha,\ \beta,\ m_i+u_i) = \binom{m_i+u_i}{m_i} B(m_i+\alpha,\ u_i+\beta)/B(\alpha,\ \beta),$$

where B denotes the beta function. Critically, using this distribution allows us to model methylation probabilities accounting for the amount of data at each CpG while keeping the variance independent of the mean.

To fit distribution parameters for numerical convenience we work directly with the estimates $\hat{p}_i$. This is because of the time required for maximum-likelihood computations directly with the Beta-Binomial. Instead, we estimate the maximum-likelihood parameters as though they were for a Beta distribution, and therefore satisfy

$$\psi(\hat{\alpha}) - \psi(\hat{\alpha}+\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \log(\hat{p}_i)$$

and

$$\psi(\widehat{\beta}) - \psi(\widehat{\alpha} + \widehat{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \log(1 - \widehat{p}_i)$$

with

$$\psi(x) = \frac{d}{dx} \log \Gamma(x).$$

To compute $\widehat{\alpha}$ and $\widehat{\beta}$, we use an iterative procedure. The initial parameter values are calculated as

$$\widehat{\alpha}^{(0)} = \psi^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \log(\widehat{p}_i)\right)$$

and

$$\widehat{\beta}^{(0)} = \psi^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \log(1 - \widehat{p}_i)\right).$$

This initialization corresponds roughly to the assumption of $\alpha + \beta = 1$, as $\psi(1) = 0$. At each iteration, these estimates are updated using the formulas

$$\widehat{\alpha}^{(k)} = \psi^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \log(\widehat{p}_i) + \psi\left(\widehat{\alpha}^{(k-1)} + \widehat{\beta}^{(k-1)}\right)\right)$$

and

$$\widehat{\beta}^{(k)} = \psi^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \log(1 - \widehat{p}_i) + \psi\left(\widehat{\alpha}^{(k-1)} + \widehat{\beta}^{(k-1)}\right)\right).$$

The inverse of the digamma (ψ) function can be calculated very easily by noting that $\psi^{-1}(x) = e^x + \epsilon$, for $0 \leq \epsilon \leq 1$ for any relevant values of x. We use a bisection search around e^x to evaluate ψ^{-1} and apply the iterative procedure until convergence criteria are satisfied.

After training the HMM parameters, HMRs were identified by posterior decoding, and then each was scored according to the sum of all $(1 - \widehat{p}_i)$ for each CpG i in the HMR. Because a single CpG with an very high number of reads and a very low methylation level can theoretically be identified as a single-CpG HMR under our model, we included a procedure to identify only significant HMRs based on their score. The CpGs were randomly permuted, and then the random permutation was decoded to obtain an empirical distribution of random HMR scores. We obtained p values from this random distribution, and then applied the method of Benjamini and Hochberg (1995) to identify a cutoff for a false discovery rate (FDR) of 0.05. Finally, we retained as HMRs only those regions having a score more extreme than the identified 0.05 FDR cutoff.

Measuring Sequence Divergence and CpG Decay

We measured nucleotide-level conservation between human (hg18), chimp (panTro2), and gorilla (gorGor1) by using the MULTIZ 44-way alignment available through the UCSC Genome Browser (Blanchette et al., 2004). This alignment is referenced on human. Alignments for genomic intervals were extracted by identifying the blocks containing the start and end points of the region in human. If one of the two end-points was not found in the alignment, the region was determined not to be alignable. Positions in the alignments that correspond to gaps were not counted. A sequence was called "under decay" if it lost more than 5% of its CpGs; we required the inferred ancestral sequence to have at least 20 CpGs in order to make this determination.

Analysis of Nucleosome Retention Data

Nucleosome retention data was taken from Hammoud et al. (2009). Data from different donors for histone ChIP-seq experiments were pooled and mapped to the hg18 assembly using RMAP. Domains of retained nucleosomes and the H3K4me3 and H3K27me3 modifications were inferred using the RSEG algorithm (Song and Smith, 2011). This method identified 118318, 105150, and 193158 enriched domains for H3K4me3, H3K27me3, and retained histones, respectively.

Gene Ontology Analysis

To measure Gene Ontology category enrichment we used the web interface to the DAVID tool (Huang et al., 2008). For sperm and ESC-specific hypomethylated promoters we required that the promoter (−1 kb to +1 kb) overlap an HMR in one cell type, have

a methylation level at least 0.5 in the other cell type, and have a difference of at least 2-fold between the lower and higher. We used RefSeq promoters downloaded from the UCSC Table Browser. To eliminate redundancy in the sets of Gene Ontology categories identified as enriched we used the REVIGO software through the web interface (Ŝkunca et al., 2009).

Motif Enrichment Analysis

We used programs for the CREAD package to analyze the HMR sequences for identifying enriched TFBS motifs. We used both libraries of known motifs from both TRANSFAC (Matys et al., 2006) and JASPAR (Sandelin et al., 2004). We measured enrichment relative to a randomly selected set of 5000 promoters from among those that had low methylation levels in both sperm and ESCs. To eliminate bias due to different CpG content, CpG dinucleotides were inserted (or deleted) randomly in the background sequence set to bring the level of CpG up to that in the foreground. When randomly removing CpGs, they were mutated to TpG or CpA. The enrichment was measured using the Binomial p value option in the motifclass program of CREAD.

Enrichment of Sequence Patterns at HMR Boundaries

To measure enrichment of sequence patterns at boundaries of nested and extended HMRs, we used only those HMRs where a sperm HMR fully contained exactly one ESC HMR. We only considered hexameric patterns that had a CpG dinucleotide at the center and no other CpG dinucleotides in order to avoid bias introduced by the fact that CpG content will differ on either side of an HMR boundary (which we already know). We determined the expected number of occurrences of a sequence pattern by counting the number of genomic CpGs centered on that pattern, and dividing by the number of genomic CpGs.

Use of Individual Variation Data from HapMap

Individual variation data from HapMap 3 (including phases II and III) were downloaded from http://hapmap.ncbi.nlm.nih.gov. We used the CEU population, as this most closely matched the sperm donors, and the amount of data was almost as high as any of the other 10 populations. In identifying sites to use, we took only sites where the HapMap annotated ancestral allele was at the C of a CpG site (on either strand), and we also required that at least 5 reads mapped over that CpG in our bisulfite sequencing data. We used Chi-squared goodness-of-fit tests to determine that the frequency spectra differed between low and high methylation levels for each type of derived nucleotide (A, G, or T).

SUPPLEMENTAL REFERENCES

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc., B *57*, 289–300.

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., et al. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. *14*, 708–715.

Durbin, R., Eddy, S.R., Krogh, A., and Mitchison, G. (1999). Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids (Cambridge, UK: Cambridge University Press).

Huang, D., Sherman, B., and Lempicki, R. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.

Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., et al. (2006). TRANSFAC(R) and its module TRANSCompel(R): transcriptional gene regulation in eukaryotes. Nucl. Acids Res. *34* (*suppl_1*), D108–D110.

Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W., and Lenhard, B. (2004). JASPAR: an open access database for eukaryotic transcription factor binding profiles. Nucl. Acids Res. *32*, D91–D94.

Ŝkunca, N., Ŝmuc, T., and Supek, F. (2009). REViGO: Redundancy Elimination and Visualization of Gene Ontology Term Lists. In The 3rd Adriatic Meeting on Computational Solutions in the Life Sciences.

Song, Q., and Smith, A.D. (2011). Identifying dispersed epigenomic domains from ChIP-Seq data. Bioinformatics *27*, 870–871.

	Factor	Logo	p-value
1.	NRF1		5.34e-09
2.	NFY/CP1/CBF/HAP2		6.39e-09
3.	NFY/CP1/CBF/HAP2		8.45e-09
4.	NFY/CP1/CBF/HAP2		6.83e-08
5.	NFY/CP1/CBF/HAP2		5.16e-07
6.	NFY/CP1/CBF/HAP2		8.18e-07
7.	YY1/NF-μE1		1.59e-06
8.	YY1/NF-μE1		2.46e-05
9.	ETS		3.41e-05
10.	CREB/ATF		1.13e-04
11.	NFY/CP1/CBF/HAP2		1.83e-04
12.	ETS		1.92e-04
13.	ETS		2.16e-04
14.	ETS		2.16e-04
15.	NF-κB		3.29e-04
16.	EBOX2		3.30e-04
17.	CREB/ATF		3.55e-04
18.	NFY/CP1/CBF/HAP2		3.59e-04
19.	FOX		3.67e-04
20.	CREB/ATF		4.04e-04

Figure S1. Related to Figure 2

Transcription factor-binding site motif enrichment associated HMRs overlapping promoters in human sperm but not in ESCs. p values of enriched motifs were calculated using a random subset of HMRs overlapping promoters in both cell types as a background.

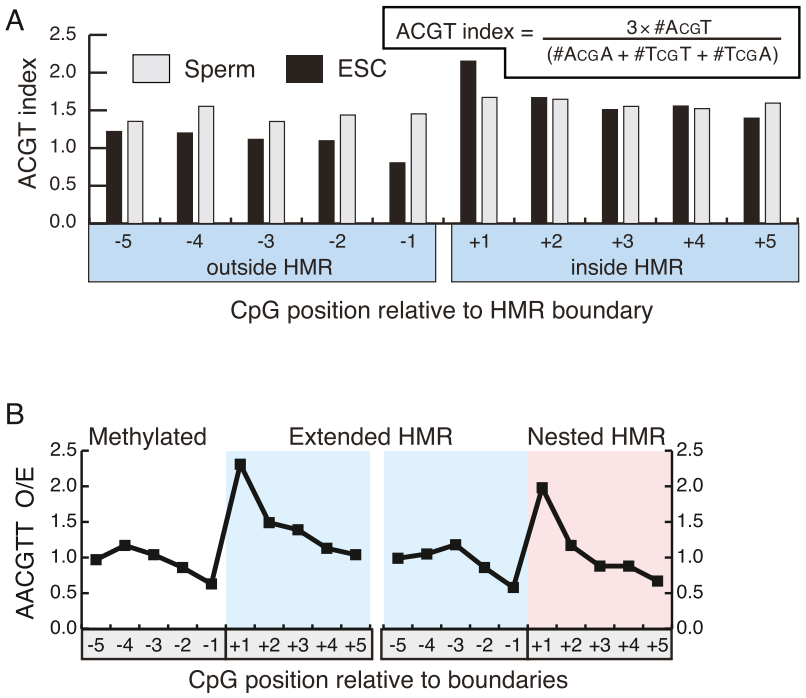


Figure S2. Related to Figure 3

(A) The AcgT index measured at CpG sites surrounding HMR boundaries in sperm (gray bars) and ESCs (black bars). Each data point corresponds to a CpG at positions −5 to +5 relative to HMRs boundaries.

(B) Observed-to-expected ratio for occurrences of the AACGTT pattern at each of the CpG positions from −5 to +5 relative to the boundaries of nested ESC and extended sperm HMRs.

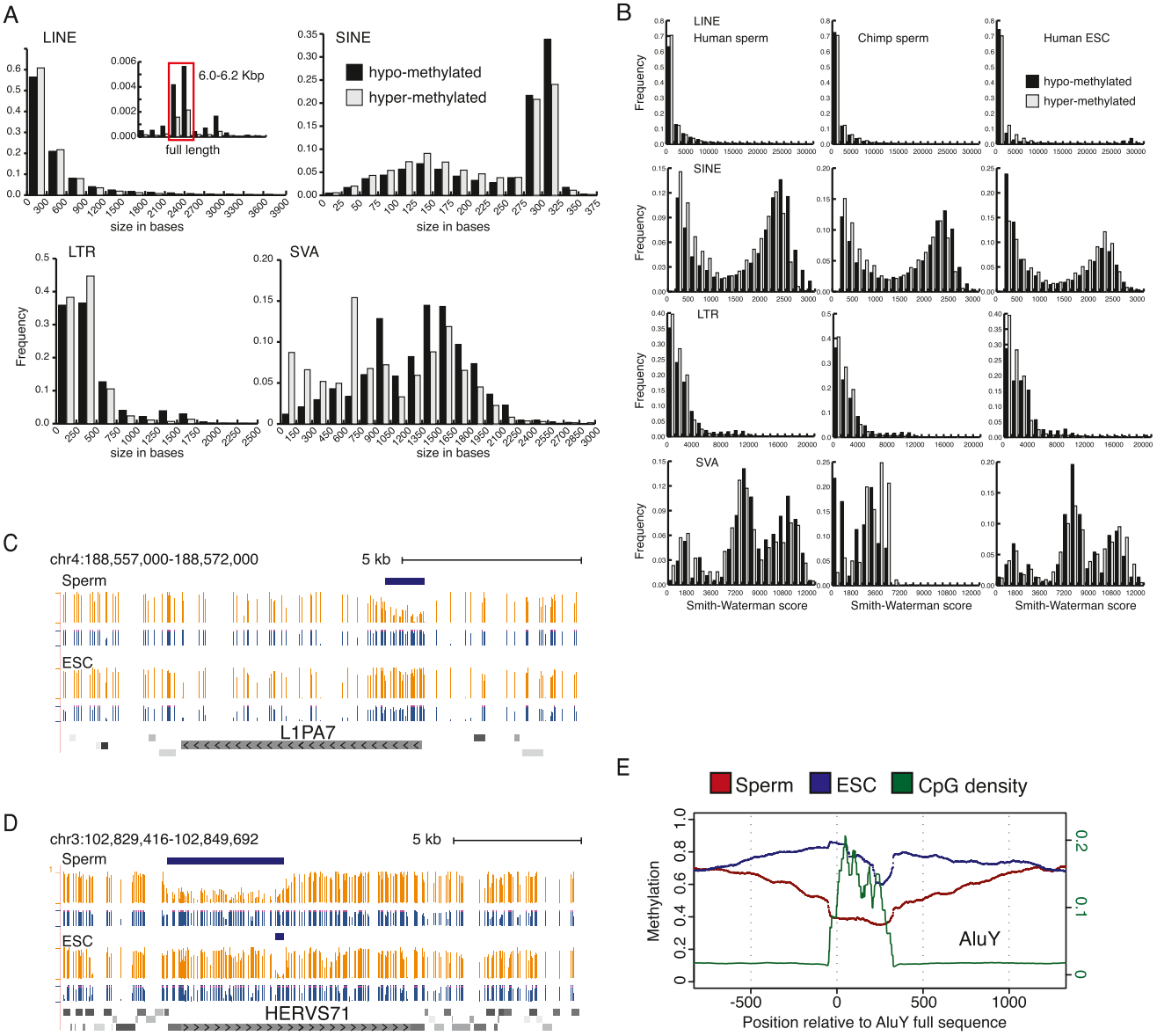


Figure S3. Related to Figure 4

(A) Size distribution of retrotransposons that are hypomethylated (black) and methylated (white) in human sperm. For each bin, the frequency of element copies is plotted.

(B) Histograms of Smith-Waterman scores of retro elements relative to their consensus sequences for hypomethylated and methylated copies. Separate histograms are given for LINE, SINE, LTR, and SVA elements, and for methylation status in human sperm, chimp sperm, and human ESCs.

(C) Browser tracks showing methylation (orange), read coverage (blue), and HMRs (blue bars) over a full-length LINE-1 element (L1PA7) hypomethylated in human sperm (upper tracks) but not in ESCs (bottom tracks).

(D) Browser track (as displayed in A) showing sperm-specific hypomethylation of the ERV HERVS71 in human sperm.

(E) Average methylation levels across all AluY SINE elements in human sperm (red) and ESCs (blue). CpG density is also shown in green. Methylation levels and CpG densities are also shown across flanking regions.

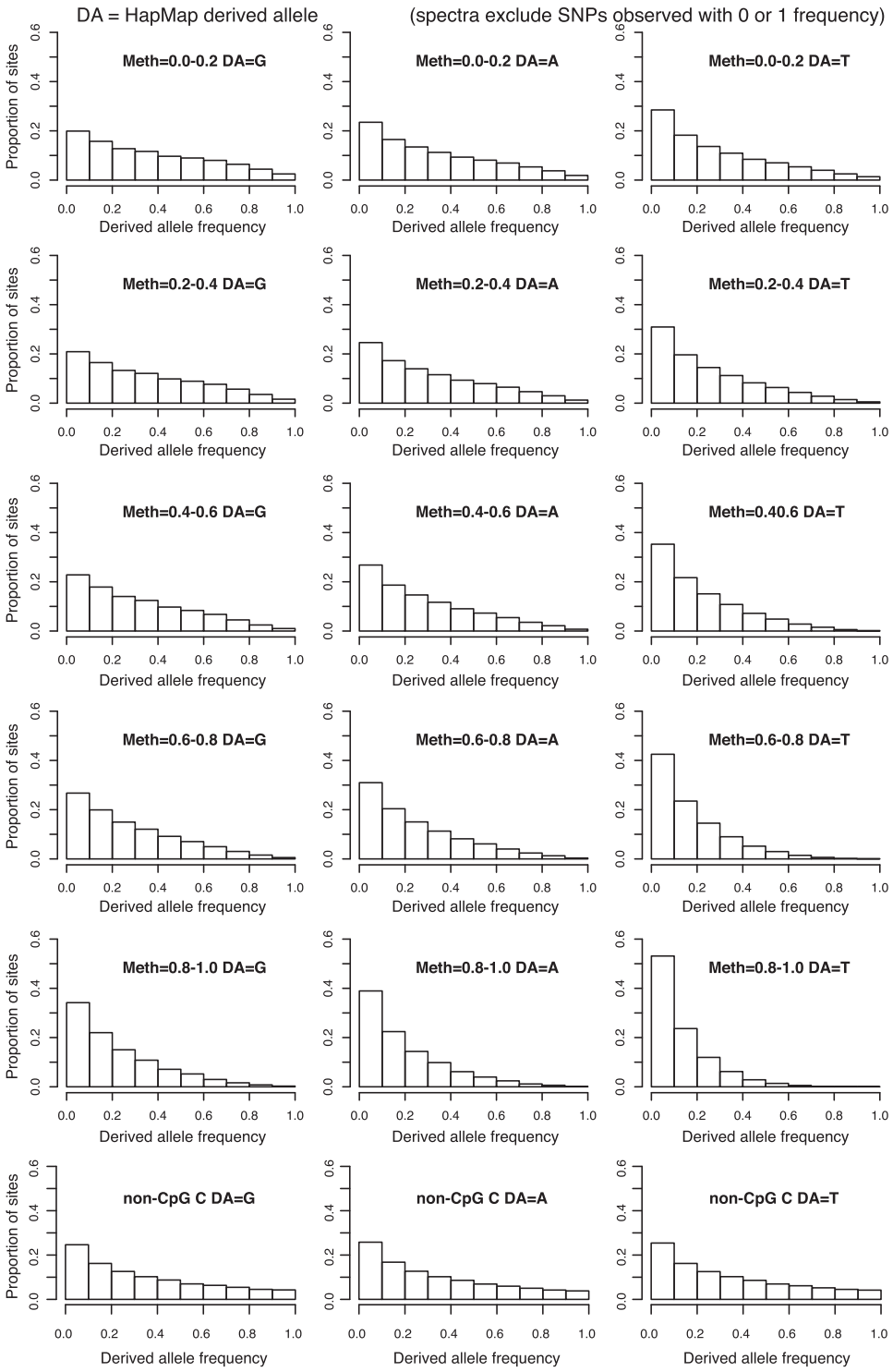


Figure S4. Related to Figure 6

Allele frequency spectra for each possible derived allele nucleotide at CpG sites treated symmetrically with cytosine as derived allele. For each derived allele, segregating sites were partitioned according to methylation levels in the intervals {[0.0, 0.2), [0.2, 0.4), … [0.8,1.0]}.

Chapter 3: Discussion and Perspectives.

Understanding the link existing between small RNAs and chromatin remodeling during mammalian germ cell development is not an easy task as it involves both innate and adaptive components. The piRNA pathway integrates such adaptive signals, such as those from previously unseen transposon insertions, and innate information using sequences stored within piRNA clusters. Similar to antigen presentation and selection which discriminates self from non-self during immune system development, epigenetic reprogramming transiently exposes the content of the genome to maturing PGCs allowing piRNAs and *de novo* methylation to bookmark each genome uniquely without the need for *a priori* information. This parallel with the immune system is far from being a new idea but, interestingly, it was proposed at different times in both fields independently (for example, on methylation see Yoder and Bestor 1997; about piRNAs see Aravin, Hannon and Brenecke 2007). The work presented here attempts to highlight how these pathways converge to establish a unique chromatin signature in germ cells at each generation.

Germ line inheritance of epigenetic states remains a central question. During male germ cell *de novo* methylation, regulatory elements with no apparent function in the germ line are isolated and transmitted in a hypomethylated state to the next generation. In some cases, these regulatory regions will only be used much later during post-fertilization development. One might view the transmission of these hypomethylated regions as a way the epigenome is set to

stabilize the future interactions between regulatory factors and their target genomic sequence. This germ cell pre-patterning offers the possibility for phenotypic variation to be inherited over short timescales. In terms of evolutionary time scales, this work highlights: 1) the potential for independent genome and epigenome evolution, providing evidence for the positive selection of epigenetic variants during speciation, 2) DNA sequence changes that sometimes accompany changes in methylation provide an insight into how multiple layers of selection converge to fix an (epi)allele within a lineage. These genome/epigenome interactions can be paralleled with the concepts of both canalization and genetic assimilation (Conrad H. Waddington, 1959). In the former, the effects of genetic variation are dampened by the epigenetic landscape. In the latter, changes in the epigenetic landscape become permanently fixed in the genome.

The following sections will extend the discussion presented in the preceding manuscripts and also address points that could not be mentioned in the result section. In addition, several “follow up” experiments are discussed in the context of the most recent literature.

3.1 Towards an understanding of piRNA cluster biology

Using various tagged piRNA clusters as transgenes in both mouse and *drosophila* genomes, we were able to show that piRNA clusters can be programmed to produce new piRNAs upon ectopic sequence insertion, outside of their native genomic context. These results suggest that the cues tagging a transcript for piRNA processing lay within cluster sequences themselves.

Surprisingly, meiotic piRNA clusters in mouse, rat and human are positioned within syntenic portions of their genomes despite cluster sequences being extremely divergent (Aravin et al., 2006; Girard et al., 2006). This indicates that different selective pressures drive the evolution of piRNA cluster position and cluster sequences. It has been suggested that piRNA clusters targeting transposons act as graveyards keeping a trace of previous waves of transposon activity and trapping new ones throughout the animal's lifespan. Our data, together with recent work by Kawaoka et al., (2011), strongly support this model.

The ability to ectopically program piRNA clusters could be of great use to control, for instance, gene expression during meiosis or to direct site specific chromatin changes if a cluster expressed during embryogenesis is targeted. In addition, our transgenic lines offer the opportunity to study the yet unknown, silencing capability of meiotic piRNAs. Some of these experiments are currently under investigation and are discussed in the following points. To address the silencing function of meiotic piRNAs, both GFP-tagged-cluster transgenic lines were crossed to a reporter mouse expressing a GFP tagged protein with appropriate patterns of expression (in this case a MILI-GFP tagged transgenic line was used from Aravin et al., 2008). Preliminary results from both GFP immuno-staining in testis cross-sections and sequencing of potential cleavage products by 5'RACE, revealed no significant differences between mice carrying the tagged cluster and the reporter, and mice bearing one or the other construct alone. These data could indicate that the co-expression of a meiotic piRNA and its cognate target sequence is not sufficient to trigger silencing or that meiotic

piRNAs don't act on cellular mRNAs. Maybe the reduction in MIWI loading seen for GFP piRNAs, compared to endogenous piRNAs, cause these small RNAs to never fully engage in silencing. To address this question one could perform MIIWI-IPs in animals expressing the target GFP-reporter and investigate if the presence of a target changes MIWI loading rate with GFP piRNAs. Finally, one could rule out the implication of local chromatin environment in both silencing potential and MIWI loading by generating a knock-in tagged piRNA clusters in mouse ES cells.

Finally, tagging a transposon rich embryonic piRNA cluster was obviously one of our original goals and constructs were generated in parallel. Unfortunately, stable maintenance of transgenic animals revealed itself to be difficult past F1. Due to the nature of these regions (transposon enriched, large size, etc.), it is possible that genetic background and lineage history impact the stability and heritability of this transgene. One could re-inject these BACs and try to cross F1 males and females to different background (e.g. mixed 129/BL6).

3.2 Transposon *de novo* methylation in the male germ line: insight into the ecology of our genomes

Through the detailed analysis of both meiotic and mature germ cell methylomes, it became clear that systematic methylation of repeated sequences was far from being the rule. In fact, centromeric repeats, DNA transposons and retro-transposons can be transmitted in a hypomethylated state with some of these elements even found hypomethylated later in development (e.g. in ES

cells). However, studying transposon methylation status in piRNA deficient animals showed that, at each generation, recent transposition events can be specifically recognized and transcriptionally silenced in an adaptive fashion. This highlights the non-deleterious status of constitutively hypomethylated copies and suggests that those might have been functionalized by the genome in recent evolutionary history. Using the dynamic nature of transposons to rapidly evolve new adaptive function lay at the core of their original discovery in maize by Barbara McClintock in the 40's and 50's. Since then, the impact of transposons on gene regulation has been studied in various contexts, including embryonic development and imprinting (e.g. see Peaston et al., 2004; Chow et al., 2010; also reviewed by Goodier and Kazazian, 2008; Levin and Moran, 2011). Understanding the regulatory function of these hypomethylated transposons during development could uncover some of the fundamental mechanism by which the genome and the epigenome interact and evolve.

The means by which transposons naturally escape methylation still remains a topic of investigation and cannot be decoupled from understanding how any other hypomethylated domain is established and maintained during *de novo* methylation (discussed in section 3.3). However, our study of piRNA-targeted transposons in PGCs suggests a model where a first wave of *de novo* methylation is established by default. To further test this model, we are currently investigating the methylation status of PGCs during *de novo* methylation at ~16.5dpc in a WT and MILI mutant context. This should help map the time course of methylation marks deposition as well as the genomic origins of this

default methylation. Profiling the transcriptome of MILI mutant PGCs at these stages could test whether piRNA-targeted transposons are indeed still transcriptionally active when MIWI2 localizes to the nucleus. In addition, performing strand specific transcriptome analysis on developing PGCs might uncover abundant antisense transcripts originating from these elements, explaining their preferential entry in the ping-pong amplification loop.

Finally, we proposed that the transient up-regulation of retro-transposon transcription, occurring in 13.5dpc PGCs, is necessary for adaptive piRNA targeting and proper epigenetic reprogramming. One could test this hypothesis, by introducing as a transgene a composite element containing the regulatory portion of a piRNA-targeted transposon fused to a reporter sequence. In this scenario, one could follow the kinetics of reporter transcriptional activation and silencing. Taking advantage of this unique genomic structure, one could also study the dynamics of polymerase recruitment and other chromatin marks preceeding and following piRNA targeting. A long-term project could also involve transiently culturing PGCs *ex-vivo* and trying to recapitulate *de novo* methylation in this context. For example, treating these cells with transcriptional inhibitors, such as Actinomycin D, should impact transposon silencing. However, previous attempts in deriving long-lived PGCs in culture have been challenging and we are still far from being able to recapitulate epigenetic reprogramming *in-vitro*.

3.3 HMR establishment and evolution

Mature sperm and to some extent ES cells represent the outcome of the extensive wave of *de novo* methylation occurring during germ cell and pre-implantation development, respectively. In light of many recent high profiling studies of DNA methylation, it is now becoming obvious that the default state of a cytosine is to be methylated. Hypomethylated regions (HMRs) seem to result from local protection from *de novo* methyl-transferases rather than a differential targeting signal. This is cocordant with HMRs occurring at both distal and proximal regulatory elements in sperm and ES cells. The involvement of DNA occupancy by DNA binding factors in preventing DNA methylation has been implicated in the establishment of parental imprints (involving CTCF, Pant et al., 2003; Schoenherr et al., 2003) and hypomethylation of a subset of CGIs (via SP1 binding, Brandeis et al., 1994; Macleod et al., 1994; or VEZF1 binding, Dickson et al., 2010). More recently, Lienert and colleagues showed that, when inserted elsewhere in the genome, HMRs were still protected from *de novo* methylation and that this protection was dependent on transcription factor, or insulator, binding (Lienert et al., 2011). Consequently, both common and cell-type specific transcription factor networks could account for the differences seen in HMR distribution and structure between ES cells and germ cells. Finally, this mode of HMR establishment offers an attractive model to try to characterize what factors are driving the evolution of HMRs between chimp and human, and allows this hypothesis to be tested both *in-vivo* and *in-silco*.

Decoupling the pleiotropic effects that the ablation of a given transcription factor could have on HMR protection and cell survival, or embryogenesis, is likely to prevent such model to be tested directly using mouse knock-outs (not to mention the redundant binding of other transcription factors). However, looking at transcription factor expression during *de novo* methylation in PGCs should help predict the HMR landscape. In addition, expanding the studies mentioned above, one could monitor the appearance of new germ cell HMRs forcing the ectopic expression of a transcription factor in PGCs (using a tissue specific or a species specific transcription factor). These ideas are currently being tested, for example, by sorting and sequencing germ cells from a previously published mouse model carrying a substantial portion of a human chromosome (e.g. TC1 mouse model for Down Syndrome, carrying an autonomously segregating fraction of human chromosome 21, O'Doherty et al., 2005). It is therefore expected to express human transcription factors as well as respond to endogenous mouse specific factors.

Interestingly, many HMRs that are established in sperm correspond to regulatory regions (including many promoters) with no obvious function during germ cell development. These have been interpreted, by us and others (Hammoud et al., 2009), as a form of pre-pattering necessary for early zygote development. The scenario mentioned above would have to account for the binding of these factors during PGC *de novo* methylation without them exerting their full effect on genome organization or transcription. These observations are calling for a deeper investigation of higher order chromatin structures that could

function to prevent these binding events from being deleterious to germ cells (histone modifications being one of them).

The comparative analysis performed on human and chimp sperm methylomes revealed that DNA sequence alone (including CpG density) was not sufficient to explain HMR evolution. However, it also showed that DNA sequence and methylation couldn't be fully separated; as signatures of DNA sequence divergence accompany HMR gain and losses. One can speculate about the relative selective pressure that the gain or loss of a binding site or a chromatin interacting factor might have on the HMR landscape of a lineage. But a question would remain: are DNA sequences changes helping to stabilize a new methylation state (e.g. becoming better binding sites) or are they a pre-requisite for *de novo* acquisition of an HMR within a lineage? The answer is probably the combination of both, with the genome and the epigenome "canalizing" each other towards the most stable state. Only the study of an out-group would deconvolute the relative contribution of both processes. Towards this goal, we are currently sequencing Gorilla and Bonobo sperm methylomes. The comparison between Bonobo and Chimpanzee will position the Human lineage as an out-group, whereas Gorilla would the out-group of all three species. In this context, we will characterize lineage specific DMRs and quantify the genomic changes preceding or following both ancestral and derived HMRs.

Literature cited

Aravin, A., Gaidatzis, D., Pfeffer, S., Lagos-Quintana, M., Landgraf, P., Iovino, N., Morris, P., Brownstein, M.J., Kuramochi-Miyagawa, S., Nakano, T., et al. (2006). A novel class of small RNAs bind to MILI protein in mouse testes. Nature 442, 203-207.

Aravin, A.A., and Bourc'his, D. (2008). Small RNA guides for de novo DNA methylation in mammalian germ cells. Genes Dev 22, 970-975.

Aravin, A.A., and Hannon, G.J. (2008). Small RNA silencing pathways in germ and stem cells. Cold Spring Harb Symp Quant Biol 73, 283-290.

Aravin, A.A., Hannon, G.J., and Brennecke, J. (2007). The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. Science 318, 761-764.

Aravin, A.A., Hannon, G.J., and Brennecke, J. (2007). The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. Science 318, 761-764.

Aravin, A.A., Lagos-Quintana, M., Yalcin, A., Zavolan, M., Marks, D., Snyder, B., Gaasterland, T., Meyer, J., and Tuschl, T. (2003). The small RNA profile during Drosophila melanogaster development. Dev Cell 5, 337-350.

Aravin, A.A., Sachidanandam, R., Bourc'his, D., Schaefer, C., Pezic, D., Toth, K.F., Bestor, T., and Hannon, G.J. (2008). A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. Mol Cell 31, 785-799.

Aravin, A.A., Sachidanandam, R., Girard, A., Fejes-Toth, K., and Hannon, G.J. (2007). Developmentally regulated piRNA clusters implicate MILI in transposon control. Science 316, 744-747.

Arita, K., Ariyoshi, M., Tochio, H., Nakamura, Y., and Shirakawa, M. (2008). Recognition of hemi-methylated DNA by the SRA protein UHRF1 by a base-flipping mechanism. Nature 455, 818-821.

Bastos, H., Lassalle, B., Chicheportiche, A., Riou, L., Testart, J., Allemand, I., and Fouchet, P. (2005). Flow cytometric characterization of viable meiotic and postmeiotic cells by Hoechst 33342 in mouse spermatogenesis. Cytometry A 65, 40-49.

Bellve, A.R., Cavicchia, J.C., Millette, C.F., O'Brien, D.A., Bhatnagar, Y.M., and Dym, M. (1977). Spermatogenic cells of the prepuberal mouse. Isolation and morphological characterization. J Cell Biol 74, 68-85.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125, 315-326.

Bernstein, E., Caudy, A.A., Hammond, S.M., and Hannon, G.J. (2001). Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature 409, 363-366.

Bestor, T., Laudano, A., Mattaliano, R., and Ingram, V. (1988). Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. J Mol Biol 203, 971-983.

Bestor, T.H., and Bourc'his, D. (2004). Transposon silencing and imprint establishment in mammalian germ cells. Cold Spring Harb Symp Quant Biol 69, 381-387.

Bhutani, N., Brady, J.J., Damian, M., Sacco, A., Corbel, S.Y., and Blau, H.M. Reprogramming towards pluripotency requires AID-dependent DNA demethylation. Nature 463, 1042-1047.

Bhutani, N., Burns, D.M., and Blau, H.M. DNA demethylation dynamics. Cell 146, 866-872.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes Dev 16, 6-21.

Bird, A., Taggart, M., Frommer, M., Miller, O.J., and Macleod, D. (1985). A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. Cell 40, 91-99.

Bird, A.P. (1980). DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res 8, 1499-1504.

Bouniol, C., Nguyen, E., and Debey, P. (1995). Endogenous transcription occurs at the 1-cell stage in the mouse embryo. Exp Cell Res 218, 57-62.

Bourc'his, D., and Bestor, T.H. (2004). Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. Nature 431, 96-99.

Bourc'his, D., Xu, G.L., Lin, C.S., Bollman, B., and Bestor, T.H. (2001). Dnmt3L and the establishment of maternal genomic imprints. Science 294, 2536-2539.

Brandeis, M., Frank, D., Keshet, I., Siegfried, Z., Mendelsohn, M., Nemes, A., Temper, V., Razin, A., and Cedar, H. (1994). Sp1 elements protect a CpG island from de novo methylation. Nature 371, 435-438.

Brennecke, J., Aravin, A.A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G.J. (2007). Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell 128, 1089-1103.

Carmell, M.A., Girard, A., van de Kant, H.J., Bourc'his, D., Bestor, T.H., de Rooij, D.G., and Hannon, G.J. (2007). MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. Dev Cell 12, 503-514.

Carmell, M.A., and Hannon, G.J. (2004). RNase III enzymes and the initiation of gene silencing. Nat Struct Mol Biol 11, 214-218.

Carmell, M.A., Xuan, Z., Zhang, M.Q., and Hannon, G.J. (2002). The Argonaute family: tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis. Genes Dev 16, 2733-2742.

Cedar, H., and Bergman, Y. (2009). Linking DNA methylation and histone modification: patterns and paradigms. Nat Rev Genet 10, 295-304.

Chambers, I., Colby, D., Robertson, M., Nichols, J., Lee, S., Tweedie, S., and Smith, A. (2003). Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. Cell 113, 643-655.

Chambers, I., Silva, J., Colby, D., Nichols, J., Nijmeijer, B., Robertson, M., Vrana, J., Jones, K., Grotewold, L., and Smith, A. (2007). Nanog safeguards pluripotency and mediates germline development. Nature 450, 1230-1234.

Chambers, I., and Smith, A. (2004). Self-renewal of teratocarcinoma and embryonic stem cells. Oncogene 23, 7150-7160.

Chang, H., and Matzuk, M.M. (2001). Smad5 is required for mouse primordial germ cell development. Mech Dev 104, 61-67.

Chedin, F., Lieber, M.R., and Hsieh, C.L. (2002). The DNA methyltransferase-like protein DNMT3L stimulates de novo methylation by Dnmt3a. Proc Natl Acad Sci U S A 99, 16916-16921.

Cho, C., Willis, W.D., Goulding, E.H., Jung-Ha, H., Choi, Y.C., Hecht, N.B., and Eddy, E.M. (2001). Haploinsufficiency of protamine-1 or -2 causes infertility in mice. Nat Genet 28, 82-86.

Chow, J.C., Ciaudo, C., Fazzari, M.J., Mise, N., Servant, N., Glass, J.L., Attreed, M., Avner, P., Wutz, A., Barillot, E., et al. (2010). LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. Cell 141, 956-969.

Coffigny, H., Bourgeois, C., Ricoul, M., Bernardino, J., Vilain, A., Niveleau, A., Malfoy, B., and Dutrillaux, B. (1999). Alterations of DNA methylation patterns in germ cells and Sertoli cells from developing mouse testis. Cytogenet Cell Genet 87, 175-181.

Cohen, N.M., Kenigsberg, E., and Tanay, A. Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. Cell 145, 773-786.

Cooper, D.N., and Krawczak, M. (1989). Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. Hum Genet 83, 181-188.

Cortellino, S., Xu, J., Sannai, M., Moore, R., Caretti, E., Cigliano, A., Le Coz, M., Devarajan, K., Wessels, A., Soprano, D., et al. Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair. Cell 146, 67-79.

Cox, D.N., Chao, A., Baker, J., Chang, L., Qiao, D., and Lin, H. (1998). A novel class of evolutionarily conserved genes defined by piwi are essential for stem cell self-renewal. Genes Dev 12, 3715-3727.

Cox, D.N., Chao, A., and Lin, H. (2000). piwi encodes a nucleoplasmic factor whose activity modulates the number and division rate of germline stem cells. Development 127, 503-514.

Csankovszki, G., Nagy, A., and Jaenisch, R. (2001). Synergism of Xist RNA, DNA methylation, and histone hypoacetylation in maintaining X chromosome inactivation. J Cell Biol 153, 773-784.

De Fazio, S., Bartonicek, N., Di Giacomo, M., Abreu-Goodger, C., Sankar, A., Funaya, C., Antony, C., Moreira, P.N., Enright, A.J., and O'Carroll, D. (2011). The endonuclease activity of Mili fuels piRNA amplification that silences LINE1 elements. Nature 480, 259-263.

Deng, W., and Lin, H. (2002). miwi, a murine homolog of piwi, encodes a cytoplasmic protein essential for spermatogenesis. Dev Cell 2, 819-830.

Dickson, J., Gowher, H., Strogantsev, R., Gaszner, M., Hair, A., Felsenfeld, G., and West, A.G. (2010). VEZF1 elements mediate protection from DNA methylation. PLoS Genet 6, e1000804.

Duncan, B.K., and Miller, J.H. (1980). Mutagenic deamination of cytosine residues in DNA. Nature 287, 560-561.

Edwards, J.R., O'Donnell, A.H., Rollins, R.A., Peckham, H.E., Lee, C., Milekic, M.H., Chanrion, B., Fu, Y., Su, T., Hibshoosh, H., et al. (2010). Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. Genome Res 20, 972-980.

Ehrlich, M., Gama-Sosa, M.A., Huang, L.H., Midgett, R.M., Kuo, K.C., McCune, R.A., and Gehrke, C. (1982). Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. Nucleic Acids Res 10, 2709-2721.

Ehrlich, M., Zhang, X.Y., and Inamdar, N.M. (1990). Spontaneous deamination of cytosine and 5-methylcytosine residues in DNA and replacement of 5-methylcytosine residues with cytosine residues. Mutat Res 238, 277-286.

Enders, G.C., and May, J.J., 2nd (1994). Developmentally regulated expression of a mouse germ cell nuclear antigen examined from embryonic day 11 to adult in male and female mice. Dev Biol 163, 331-340.

Esteve, P.O., Chin, H.G., Smallwood, A., Feehery, G.R., Gangisetty, O., Karpf, A.R., Carey, M.F., and Pradhan, S. (2006). Direct interaction between DNMT1 and G9a coordinates DNA and histone methylation during replication. Genes Dev 20, 3089-3103.

Evsikov, A.V., de Vries, W.N., Peaston, A.E., Radford, E.E., Fancher, K.S., Chen, F.H., Blake, J.A., Bult, C.J., Latham, K.E., Solter, D., et al. (2004). Systems biology of the 2-cell mouse embryo. Cytogenet Genome Res 105, 240-250.

Ficz, G., Branco, M.R., Seisenberger, S., Santos, F., Krueger, F., Hore, T.A., Marques, C.J., Andrews, S., and Reik, W. Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. Nature 473, 398-402.

Flach, G., Johnson, M.H., Braude, P.R., Taylor, R.A., and Bolton, V.N. (1982). The transition from maternal to embryonic control in the 2-cell mouse embryo. EMBO J 1, 681-686.

Flynn, J., Glickman, J.F., and Reich, N.O. (1996). Murine DNA cytosine-C5 methyltransferase: pre-steady- and steady-state kinetic analysis with regulatory DNA sequences. Biochemistry 35, 7308-7315.

Fuks, F., Burgers, W.A., Godin, N., Kasai, M., and Kouzarides, T. (2001). Dnmt3a binds deacetylases and is recruited by a sequence-specific repressor to silence transcription. EMBO J 20, 2536-2544.

Fuks, F., Hurd, P.J., Wolf, D., Nan, X., Bird, A.P., and Kouzarides, T. (2003). The methyl-CpG-binding protein MeCP2 links DNA methylation to histone methylation. J Biol Chem 278, 4035-4040.

Gardiner-Garden, M., and Frommer, M. (1987). CpG islands in vertebrate genomes. J Mol Biol 196, 261-282.

Ghildiyal, M., and Zamore, P.D. (2010). Small silencing RNAs: an expanding universe. Nat Rev Genet 10, 94-108.

Ginsburg, M., Snow, M.H., and McLaren, A. (1990). Primordial germ cells in the mouse embryo during gastrulation. Development 110, 521-528.

Girard, A., Sachidanandam, R., Hannon, G.J., and Carmell, M.A. (2006). A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature 442, 199-202.

Goecks, J., Nekrutenko, A., and Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11, R86.

Goll, M.G., and Bestor, T.H. (2005). Eukaryotic cytosine methyltransferases. Annu Rev Biochem 74, 481-514.

Goodier, J.L., and Kazazian, H.H., Jr. (2008). Retrotransposons revisited: the restraint and rehabilitation of parasites. Cell 135, 23-35.

Gowher, H., and Jeltsch, A. (2001). Enzymatic properties of recombinant Dnmt3a DNA methyltransferase from mouse: the enzyme modifies DNA in a non-processive manner and also methylates non-CpG [correction of non-CpA] sites. J Mol Biol 309, 1201-1208.

Gowher, H., and Jeltsch, A. (2002). Molecular enzymology of the catalytic domains of the Dnmt3a and Dnmt3b DNA methyltransferases. J Biol Chem 277, 20409-20414.

Guan, K., Nayernia, K., Maier, L.S., Wagner, S., Dressel, R., Lee, J.H., Nolte, J., Wolf, F., Li, M., Engel, W., et al. (2006). Pluripotency of spermatogonial stem cells from adult mouse testis. Nature 440, 1199-1203.

Haase, A.D., Fenoglio, S., Muerdter, F., Guzzardo, P.M., Czech, B., Pappin, D.J., Chen, C., Gordon, A., and Hannon, G.J. (2010). Probing the initiation and effector phases of the somatic piRNA pathway in Drosophila. Genes Dev 24, 2499-2504.

Hajkova, P., Ancelin, K., Waldmann, T., Lacoste, N., Lange, U.C., Cesari, F., Lee, C., Almouzni, G., Schneider, R., and Surani, M.A. (2008). Chromatin dynamics during epigenetic reprogramming in the mouse germ line. Nature 452, 877-881.

Hajkova, P., Erhardt, S., Lane, N., Haaf, T., El-Maarri, O., Reik, W., Walter, J., and Surani, M.A. (2002). Epigenetic reprogramming in mouse primordial germ cells. Mech Dev 117, 15-23.

Hajkova, P., Jeffries, S.J., Lee, C., Miller, N., Jackson, S.P., and Surani, M.A. (2010). Genome-wide reprogramming in the mouse germ line entails the base excision repair pathway. Science 329, 78-82.

Hammoud, S.S., Nix, D.A., Zhang, H., Purwar, J., Carrell, D.T., and Cairns, B.R. (2009). Distinctive chromatin in human sperm packages genes for embryo development. Nature 460, 473-478.

Hannon, G.J. (2002). RNA interference. Nature 418, 244-251.

Hashimoto, H., Vertino, P.M., and Cheng, X. Molecular coupling of DNA methylation and histone methylation. Epigenomics 2, 657-669.

Hashimshony, T., Zhang, J., Keshet, I., Bustin, M., and Cedar, H. (2003). The role of DNA methylation in setting up chromatin structure during development. Nat Genet 34, 187-192.

Hata, K., Okano, M., Lei, H., and Li, E. (2002). Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. Development 129, 1983-1993.

Henckel, A., Nakabayashi, K., Sanz, L.A., Feil, R., Hata, K., and Arnaud, P. (2009). Histone methylation is mechanistically linked to DNA methylation at imprinting control regions in mammals. Hum Mol Genet 18, 3375-3383.

Hirasawa, R., Chiba, H., Kaneda, M., Tajima, S., Li, E., Jaenisch, R., and Sasaki, H. (2008). Maternal and zygotic Dnmt1 are necessary and sufficient for the maintenance of DNA methylation imprints during preimplantation development. Genes Dev 22, 1607-1616.

Hodges, E., Molaro, A., Dos Santos, C.O., Thekkat, P., Song, Q., Uren, P.J., Park, J., Butler, J., Rafii, S., McCombie, W.R., et al. (2011). Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. Mol Cell 44, 17-28.

Holliday, R., and Pugh, J.E. (1975). DNA modification mechanisms and gene activity during development. Science 187, 226-232.

Horwich, M.D., Li, C., Matranga, C., Vagin, V., Farley, G., Wang, P., and Zamore, P.D. (2007). The Drosophila RNA methyltransferase, DmHen1, modifies germline piRNAs and single-stranded siRNAs in RISC. Curr Biol 17, 1265-1272.

Houwing, S., Kamminga, L.M., Berezikov, E., Cronembold, D., Girard, A., van den Elst, H., Filippov, D.V., Blaser, H., Raz, E., Moens, C.B., et al. (2007). A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. Cell 129, 69-82.

Howlett, S.K., and Reik, W. (1991). Methylation levels of maternal and paternal genomes during preimplantation development. Development 113, 119-127.

Hsieh, C.L. (1999). Evidence that protein binding specifies sites of DNA demethylation. Mol Cell Biol 19, 46-56.

Inoue, A., and Zhang, Y. Replication-dependent loss of 5-hydroxymethylcytosine in mouse preimplantation embryos. Science 334, 194.

Jackson, J.P., Lindroth, A.M., Cao, X., and Jacobsen, S.E. (2002). Control of CpNpG DNA methylation by the KRYPTONITE histone H3 methyltransferase. Nature 416, 556-560.

Jia, D., Jurkowska, R.Z., Zhang, X., Jeltsch, A., and Cheng, X. (2007). Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. Nature 449, 248-251.

Jones, P.L., Veenstra, G.J., Wade, P.A., Vermaak, D., Kass, S.U., Landsberger, N., Strouboulis, J., and Wolffe, A.P. (1998). Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. Nat Genet 19, 187-191.

Jurkowska, R.Z., Anspach, N., Urbanke, C., Jia, D., Reinhardt, R., Nellen, W., Cheng, X., and Jeltsch, A. (2008). Formation of nucleoprotein filaments by mammalian DNA methyltransferase Dnmt3a in complex with regulator Dnmt3L. Nucleic Acids Res 36, 6656-6663.

Kafri, T., Ariel, M., Brandeis, M., Shemer, R., Urven, L., McCarrey, J., Cedar, H., and Razin, A. (1992). Developmental pattern of gene-specific DNA methylation in the mouse embryo and germ line. Genes Dev 6, 705-714.

Kaslow, D.C., and Migeon, B.R. (1987). DNA methylation stabilizes X chromosome inactivation in eutherians but not in marsupials: evidence for multistep maintenance of mammalian X dosage compensation. Proc Natl Acad Sci U S A 84, 6210-6214.

Kato, Y., Kaneda, M., Hata, K., Kumaki, K., Hisano, M., Kohara, Y., Okano, M., Li, E., Nozaki, M., and Sasaki, H. (2007). Role of the Dnmt3 family in de novo methylation of imprinted and repetitive sequences during male germ cell development in the mouse. Hum Mol Genet 16, 2272-2280.

Kawaoka, S., Izumi, N., Katsuma, S., and Tomari, Y. (2011). 3' end formation of PIWI-interacting RNAs in vitro. Mol Cell 43, 1015-1022.

Kehler, J., Tolkunova, E., Koschorz, B., Pesce, M., Gentile, L., Boiani, M., Lomeli, H., Nagy, A., McLaughlin, K.J., Scholer, H.R., et al. (2004). Oct4 is required for primordial germ cell survival. EMBO Rep 5, 1078-1083.

Kirino, Y., and Mourelatos, Z. (2007). The mouse homolog of HEN1 is a potential methylase for Piwi-interacting RNAs. RNA 13, 1397-1401.

Kirino, Y., and Mourelatos, Z. (2007). Mouse Piwi-interacting RNAs are 2'-O-methylated at their 3' termini. Nat Struct Mol Biol 14, 347-348.

Knox, J.D., Araujo, F.D., Bigey, P., Slack, A.D., Price, G.B., Zannis-Hadjopoulos, M., and Szyf, M. (2000). Inhibition of DNA methyltransferase inhibits DNA replication. J Biol Chem 275, 17986-17990.

Kuramochi-Miyagawa, S., Kimura, T., Ijiri, T.W., Isobe, T., Asada, N., Fujita, Y., Ikawa, M., Iwai, N., Okabe, M., Deng, W., et al. (2004). Mili, a mammalian member of piwi family gene, is essential for spermatogenesis. Development 131, 839-849.

Kuramochi-Miyagawa, S., Kimura, T., Yomogida, K., Kuroiwa, A., Tadokoro, Y., Fujita, Y., Sato, M., Matsuda, Y., and Nakano, T. (2001). Two mouse piwi-related genes: miwi and mili. Mech Dev 108, 121-133.

Kuramochi-Miyagawa, S., Watanabe, T., Gotoh, K., Totoki, Y., Toyoda, A., Ikawa, M., Asada, N., Kojima, K., Yamaguchi, Y., Ijiri, T.W., et al. (2008). DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. Genes Dev 22, 908-917.

Kurimoto, K., Yabuta, Y., Ohinata, Y., Shigeta, M., Yamanaka, K., and Saitou, M. (2008). Complex genome-wide transcription dynamics orchestrated by Blimp1 for the specification of the germ cell lineage in mice. Genes Dev 22, 1617-1635.

La Salle, S., Mertineit, C., Taketo, T., Moens, P.B., Bestor, T.H., and Trasler, J.M. (2004). Windows for sex-specific methylation marked by DNA methyltransferase expression profiles in mouse germ cells. Dev Biol 268, 403-415.

Lane, N., Dean, W., Erhardt, S., Hajkova, P., Surani, A., Walter, J., and Reik, W. (2003). Resistance of IAPs to methylation reprogramming may provide a mechanism for epigenetic inheritance in the mouse. Genesis 35, 88-93.

Lange, U.C., Adams, D.J., Lee, C., Barton, S., Schneider, R., Bradley, A., and Surani, M.A. (2008). Normal germ line establishment in mice carrying a deletion of the Ifitm/Fragilis gene family cluster. Mol Cell Biol 28, 4688-4696.

Lange, U.C., Saitou, M., Western, P.S., Barton, S.C., and Surani, M.A. (2003). The fragilis interferon-inducible gene family of transmembrane proteins is associated with germ cell specification in mice. BMC Dev Biol 3, 1.

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.

Latham, K.E., Solter, D., and Schultz, R.M. (1991). Activation of a two-cell stage-specific gene following transfer of heterologous nuclei into enucleated mouse embryos. Mol Reprod Dev 30, 182-186.

Lau, N.C., Seto, A.G., Kim, J., Kuramochi-Miyagawa, S., Nakano, T., Bartel, D.P., and Kingston, R.E. (2006). Characterization of the piRNA complex from rat testes. Science 313, 363-367.

Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C.T., Low, H.M., Kin Sung, K.W., Rigoutsos, I., Loring, J., et al. (2010). Dynamic changes in the human methylome during differentiation. Genome Res 20, 320-331.

Lawson, K.A., Dunn, N.R., Roelen, B.A., Zeinstra, L.M., Davis, A.M., Wright, C.V., Korving, J.P., and Hogan, B.L. (1999). Bmp4 is required for the generation of primordial germ cells in the mouse embryo. Genes Dev 13, 424-436.

Lawson, K.A., and Hage, W.J. (1994). Clonal analysis of the origin of primordial germ cells in the mouse. Ciba Found Symp 182, 68-84; discussion 84-91.

Lee, J., Inoue, K., Ono, R., Ogonuki, N., Kohda, T., Kaneko-Ishino, T., Ogura, A., and Ishino, F. (2002). Erasing genomic imprinting memory in mouse clone embryos produced from day 11.5 primordial germ cells. Development 129, 1807-1817.

Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75, 843-854.

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Radmark, O., Kim, S., et al. (2003). The nuclear RNase III Drosha initiates microRNA processing. Nature 425, 415-419.

Lees-Murdock, D.J., De Felici, M., and Walsh, C.P. (2003). Methylation dynamics of repetitive DNA elements in the mouse germ cell lineage. Genomics 82, 230-237.

Lengner, C.J., Camargo, F.D., Hochedlinger, K., Welstead, G.G., Zaidi, S., Gokhale, S., Scholer, H.R., Tomilin, A., and Jaenisch, R. (2007). Oct4 expression is not required for mouse somatic stem cell self-renewal. Cell Stem Cell 1, 403-415.

Leonhardt, H., Page, A.W., Weier, H.U., and Bestor, T.H. (1992). A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei. Cell 71, 865-873.

Leung, A.K., and Sharp, P.A. (2006). Function and localization of microRNAs in mammalian cells. Cold Spring Harb Symp Quant Biol 71, 29-38.

Levin, H.L., and Moran, J.V. (2011). Dynamic interactions between transposable elements and their hosts. Nat Rev Genet 12, 615-627.

Li, E., Beard, C., and Jaenisch, R. (1993). Role for DNA methylation in genomic imprinting. Nature 366, 362-365.

Li, E., Bestor, T.H., and Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell 69, 915-926.

Lienert, F., Wirbelauer, C., Som, I., Dean, A., Mohn, F., and Schubeler, D. (2011). Identification of genetic elements that autonomously determine DNA methylation states. Nat Genet 43, 1091-1097.

Lin, I.G., Han, L., Taghva, A., O'Brien, L.E., and Hsieh, C.L. (2002). Murine de novo methyltransferase Dnmt3a demonstrates strand asymmetry and site preference in the methylation of DNA in vitro. Mol Cell Biol 22, 704-723.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462, 315-322.

Liu, J., Carmell, M.A., Rivas, F.V., Marsden, C.G., Thomson, J.M., Song, J.J., Hammond, S.M., Joshua-Tor, L., and Hannon, G.J. (2004). Argonaute2 is the catalytic engine of mammalian RNAi. Science 305, 1437-1441.

Liu, J., Rivas, F.V., Wohlschlegel, J., Yates, J.R., 3rd, Parker, R., and Hannon, G.J. (2005). A role for the P-body component GW182 in microRNA function. Nat Cell Biol 7, 1261-1266.

Liu, J., Valencia-Sanchez, M.A., Hannon, G.J., and Parker, R. (2005). MicroRNA-dependent localization of targeted mRNAs to mammalian P-bodies. Nat Cell Biol 7, 719-723.

Liu, Y., Oakeley, E.J., Sun, L., and Jost, J.P. (1998). Multiple domains are involved in the targeting of the mouse DNA methyltransferase to the DNA replication foci. Nucleic Acids Res 26, 1038-1045.

Macleod, D., Charlton, J., Mullins, J., and Bird, A.P. (1994). Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. Genes Dev 8, 2282-2292.

Marraffini, L.A., and Sontheimer, E.J. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nat Rev Genet 11, 181-190.

Masui, S., Nakatake, Y., Toyooka, Y., Shimosato, D., Yagi, R., Takahashi, K., Okochi, H., Okuda, A., Matoba, R., Sharov, A.A., et al. (2007). Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells. Nat Cell Biol 9, 625-635.

Mayer, W., Niveleau, A., Walter, J., Fundele, R., and Haaf, T. (2000). Demethylation of the zygotic paternal genome. Nature 403, 501-502.

Mercer, T.R., Dinger, M.E., and Mattick, J.S. (2010). Long non-coding RNAs: insights into functions. Nat Rev Genet 10, 155-159.

Mitsui, K., Tokuzawa, Y., Itoh, H., Segawa, K., Murakami, M., Takahashi, K., Maruyama, M., Maeda, M., and Yamanaka, S. (2003). The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. Cell 113, 631-642.

Molaro, A., Hodges, E., Fang, F., Song, Q., McCombie, W.R., Hannon, G.J., and Smith, A.D. (2011). Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. Cell 146, 1029-1041.

Monk, M. (1990). Changes in DNA methylation during mouse embryonic development in relation to X-chromosome activity and imprinting. Philos Trans R Soc Lond B Biol Sci 326, 299-312.

Monk, M., Boubelik, M., and Lehnert, S. (1987). Temporal and regional changes in DNA methylation in the embryonic, extraembryonic and germ cell lineages during mouse embryo development. Development 99, 371-382.

Muerdter, F., Olovnikov, I., Molaro, A., Rozhkov, N.V., Czech, B., Gordon, A., Hannon, G.J., and Aravin, A.A. (2012). Production of artificial piRNAs in flies and mice. RNA 18, 42-52.

Nan, X., Ng, H.H., Johnson, C.A., Laherty, C.D., Turner, B.M., Eisenman, R.N., and Bird, A. (1998). Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature 393, 386-389.

Nguyen, C.T., Weisenberger, D.J., Velicescu, M., Gonzales, F.A., Lin, J.C., Liang, G., and Jones, P.A. (2002). Histone H3-lysine 9 methylation is associated with aberrant gene silencing in cancer cells and is rapidly reversed by 5-aza-2'-deoxycytidine. Cancer Res 62, 6456-6461.

Nichols, J., Zevnik, B., Anastassiadis, K., Niwa, H., Klewe-Nebenius, D., Chambers, I., Scholer, H., and Smith, A. (1998). Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. Cell 95, 379-391.

O'Doherty, A., Ruf, S., Mulligan, C., Hildreth, V., Errington, M.L., Cooke, S., Sesay, A., Modino, S., Vanes, L., Hernandez, D., et al. (2005). An aneuploid mouse strain carrying human chromosome 21 with Down syndrome phenotypes. Science 309, 2033-2037.

Ohinata, Y., Ohta, H., Shigeta, M., Yamanaka, K., Wakayama, T., and Saitou, M. (2009). A signaling principle for the specification of the germ cell lineage in mice. Cell 137, 571-584.

Ohinata, Y., Payer, B., O'Carroll, D., Ancelin, K., Ono, Y., Sano, M., Barton, S.C., Obukhanych, T., Nussenzweig, M., Tarakhovsky, A., et al. (2005). Blimp1 is a critical determinant of the germ cell lineage in mice. Nature 436, 207-213.

Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell 99, 247-257.

Okano, M., Xie, S., and Li, E. (1998). Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. Nat Genet 19, 219-220.

Olsen, P.H., and Ambros, V. (1999). The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation. Dev Biol 216, 671-680.

Ooi, S.K., Qiu, C., Bernstein, E., Li, K., Jia, D., Yang, Z., Erdjument-Bromage, H., Tempst, P., Lin, S.P., Allis, C.D., et al. (2007). DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. Nature 448, 714-717.

Oswald, J., Engemann, S., Lane, N., Mayer, W., Olek, A., Fundele, R., Dean, W., Reik, W., and Walter, J. (2000). Active demethylation of the paternal genome in the mouse zygote. Curr Biol 10, 475-478.

Panning, B., and Jaenisch, R. (1996). DNA hypomethylation can activate Xist expression and silence X-linked genes. Genes Dev 10, 1991-2002.

Pant, V., Mariano, P., Kanduri, C., Mattsson, A., Lobanenkov, V., Heuchel, R., and Ohlsson, R. (2003). The nucleotides responsible for the direct physical contact between the chromatin insulator protein CTCF and the H19 imprinting control region manifest parent of origin-specific long-distance insulation and methylation-free domains. Genes Dev 17, 586-590.

Payer, B., Saitou, M., Barton, S.C., Thresher, R., Dixon, J.P., Zahn, D., Colledge, W.H., Carlton, M.B., Nakano, T., and Surani, M.A. (2003). Stella is a maternal effect gene required for normal early development in mice. Curr Biol 13, 2110-2117.

Peaston, A.E., Evsikov, A.V., Graber, J.H., de Vries, W.N., Holbrook, A.E., Solter, D., and Knowles, B.B. (2004). Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. Dev Cell 7, 597-606.

Popp, C., Dean, W., Feng, S., Cokus, S.J., Andrews, S., Pellegrini, M., Jacobsen, S.E., and Reik, W. (2010). Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. Nature 463, 1101-1105.

Pradhan, S., Bacolla, A., Wells, R.D., and Roberts, R.J. (1999). Recombinant human DNA (cytosine-5) methyltransferase. I. Expression, purification, and comparison of de novo and maintenance methylation. J Biol Chem 274, 33002-33010.

Reik, W., Dean, W., and Walter, J. (2001). Epigenetic reprogramming in mammalian development. Science 293, 1089-1093.

Reuter, M., Berninger, P., Chuma, S., Shah, H., Hosokawa, M., Funaya, C., Antony, C., Sachidanandam, R., and Pillai, R.S. (2011). Miwi catalysis is required for piRNA amplification-independent LINE1 transposon silencing. Nature 480, 264-267.

Riggs, A.D. (1975). X inactivation, differentiation, and DNA methylation. Cytogenet Cell Genet 14, 9-25.

Rougier, N., Bourc'his, D., Gomes, D.M., Niveleau, A., Plachot, M., Paldi, A., and Viegas-Pequignot, E. (1998). Chromosome methylation patterns during mammalian preimplantation development. Genes Dev 12, 2108-2113.

Rountree, M.R., Bachman, K.E., and Baylin, S.B. (2000). DNMT1 binds HDAC2 and a new co-repressor, DMAP1, to form a complex at replication foci. Nat Genet 25, 269-277.

Saito, K., Sakaguchi, Y., Suzuki, T., Siomi, H., and Siomi, M.C. (2007). Pimet, the Drosophila homolog of HEN1, mediates 2'-O-methylation of Piwi- interacting RNAs at their 3' ends. Genes Dev 21, 1603-1608.

Saitou, M., Barton, S.C., and Surani, M.A. (2002). A molecular programme for the specification of germ cell fate in mice. Nature 418, 293-300.

Saitou, M., Payer, B., Lange, U.C., Erhardt, S., Barton, S.C., and Surani, M.A. (2003). Specification of germ cell fate in mice. Philos Trans R Soc Lond B Biol Sci 358, 1363-1370.

Saitou, M., and Yamaji, M. Germ cell specification in mice: signaling, transcription regulation, and epigenetic consequences. Reproduction 139, 931-942.

Santos, F., Hendrich, B., Reik, W., and Dean, W. (2002). Dynamic reprogramming of DNA methylation in the early mouse embryo. Dev Biol 241, 172-182.

Sarot, E., Payen-Groschene, G., Bucheton, A., and Pelisson, A. (2004). Evidence for a piwi-dependent RNA silencing of the gypsy endogenous retrovirus by the Drosophila melanogaster flamenco gene. Genetics 166, 1313-1321.

Schmidt, A., Palumbo, G., Bozzetti, M.P., Tritto, P., Pimpinelli, S., and Schafer, U. (1999). Genetic and molecular characterization of sting, a gene involved in crystal formation and meiotic drive in the male germ line of Drosophila melanogaster. Genetics 151, 749-760.

Schoenherr, C.J., Levorse, J.M., and Tilghman, S.M. (2003). CTCF maintains differential methylation at the Igf2/H19 locus. Nat Genet 33, 66-69.

Schorderet, D.F., and Gartler, S.M. (1992). Analysis of CpG suppression in methylated and nonmethylated species. Proc Natl Acad Sci U S A 89, 957-961.

Seandel, M., James, D., Shmelkov, S.V., Falciatori, I., Kim, J., Chavala, S., Scherr, D.S., Zhang, F., Torres, R., Gale, N.W., et al. (2007). Generation of functional multipotent adult stem cells from GPR125+ germline progenitors. Nature 449, 346-350.

Seki, Y., Hayashi, K., Itoh, K., Mizugaki, M., Saitou, M., and Matsui, Y. (2005). Extensive and orderly reprogramming of genome-wide chromatin modifications associated with specification and early development of germ cells in mice. Dev Biol 278, 440-458.

Seki, Y., Yamaji, M., Yabuta, Y., Sano, M., Shigeta, M., Matsui, Y., Saga, Y., Tachibana, M., Shinkai, Y., and Saitou, M. (2007). Cellular dynamics associated with the genome-wide epigenetic reprogramming in migrating primordial germ cells in mice. Development 134, 2627-2638.

Sharif, J., Muto, M., Takebayashi, S., Suetake, I., Iwamatsu, A., Endo, T.A., Shinga, J., Mizutani-Koseki, Y., Toyoda, T., Okamura, K., et al. (2007). The SRA protein Np95

mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. Nature 450, 908-912.

Silva, J., Chang, K., Hannon, G.J., and Rivas, F.V. (2004). RNA-interference-based functional genomics in mammalian cells: reverse genetics coming of age. Oncogene 23, 8401-8409.

Silva, J., and Smith, A. (2008). Capturing pluripotency. Cell 132, 532-536.

Siomi, M.C., Sato, K., Pezic, D., and Aravin, A.A. PIWI-interacting small RNAs: the vanguard of genome defence. Nat Rev Mol Cell Biol 12, 246-258.

Slotkin, R.K., Vaughn, M., Borges, F., Tanurdzic, M., Becker, J.D., Feijo, J.A., and Martienssen, R.A. (2009). Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. Cell 136, 461-472.

Smallwood, A., Esteve, P.O., Pradhan, S., and Carey, M. (2007). Functional cooperation between HP1 and DNMT1 mediates gene silencing. Genes Dev 21, 1169-1178.

Smith, A.D., Chung, W.Y., Hodges, E., Kendall, J., Hannon, G., Hicks, J., Xuan, Z., and Zhang, M.Q. (2009). Updates to the RMAP short-read mapping software. Bioinformatics 25, 2841-2842.

Song, J., Rechkoblit, O., Bestor, T.H., and Patel, D.J. Structure of DNMT1-DNA complex reveals a role for autoinhibition in maintenance DNA methylation. Science 331, 1036-1040.

Song, J.J., Smith, S.K., Hannon, G.J., and Joshua-Tor, L. (2004). Crystal structure of Argonaute and its implications for RISC slicer activity. Science 305, 1434-1437.

Suetake, I., Shinozaki, F., Miyagawa, J., Takeshima, H., and Tajima, S. (2004). DNMT3L stimulates the DNA methylation activity of Dnmt3a and Dnmt3b through a direct interaction. J Biol Chem 279, 27816-27823.

Sugimoto, M., and Abe, K. (2007). X chromosome reactivation initiates in nascent primordial germ cells in mice. PLoS Genet 3, e116.

Surani, M.A. (1998). Imprinting and the initiation of gene silencing in the germ line. Cell 93, 309-312.

Surani, M.A., Hayashi, K., and Hajkova, P. (2007). Genetic and epigenetic regulators of pluripotency. Cell 128, 747-762.

Tachibana, M., Sugimoto, K., Nozaki, M., Ueda, J., Ohta, T., Ohki, M., Fukuda, M., Takeda, N., Niida, H., Kato, H., et al. (2002). G9a histone methyltransferase plays a dominant role in euchromatic histone H3 lysine 9 methylation and is essential for early embryogenesis. Genes Dev 16, 1779-1791.

Taft, R.J., Pang, K.C., Mercer, T.R., Dinger, M., and Mattick, J.S. Non-coding RNAs: regulators of disease. J Pathol 220, 126-139.

Takai, D., and Jones, P.A. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci U S A 99, 3740-3745.

Tam, P.P., and Zhou, S.X. (1996). The allocation of epiblast cells to ectodermal and germ-line lineages is influenced by the position of the cells in the gastrulating mouse embryo. Dev Biol 178, 124-132.

Tam, P.P., Zhou, S.X., and Tan, S.S. (1994). X-chromosome activity of the mouse primordial germ cells revealed by the expression of an X-linked lacZ transgene. Development 120, 2925-2932.

Tamaru, H., and Selker, E.U. (2001). A histone H3 methyltransferase controls DNA methylation in Neurospora crassa. Nature 414, 277-283.

Tanaka, S.S., and Matsui, Y. (2002). Developmentally regulated expression of mil-1 and mil-2, mouse interferon-induced transmembrane protein like genes, during formation and differentiation of primordial germ cells. Mech Dev 119 Suppl 1, S261-267.

Tanaka, S.S., Yamaguchi, Y.L., Tsoi, B., Lickert, H., and Tam, P.P. (2005). IFITM/Mil/fragilis family proteins IFITM1 and IFITM3 play distinct roles in mouse primordial germ cell homing and repulsion. Dev Cell 9, 745-756.

Tariq, M., Saze, H., Probst, A.V., Lichota, J., Habu, Y., and Paszkowski, J. (2003). Erasure of CpG methylation in Arabidopsis alters patterns of histone H3 methylation in heterochromatin. Proc Natl Acad Sci U S A 100, 8823-8827.

Thomson, J.P., Skene, P.J., Selfridge, J., Clouaire, T., Guy, J., Webb, S., Kerr, A.R., Deaton, A., Andrews, R., James, K.D., et al. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature 464, 1082-1086.

Toyooka, Y., Tsunekawa, N., Takahashi, Y., Matsui, Y., Satoh, M., and Noce, T. (2000). Expression and intracellular localization of mouse Vasa-homologue protein during germ cell development. Mech Dev 93, 139-149.

Tremblay, K.D., Dunn, N.R., and Robertson, E.J. (2001). Mouse embryos lacking Smad1 signals display defects in extra-embryonic tissues and germ cell formation. Development 128, 3609-3621.

Ueda, T., Abe, K., Miura, A., Yuzuriha, M., Zubair, M., Noguchi, M., Niwa, K., Kawase, Y., Kono, T., Matsuda, Y., et al. (2000). The paternal methylation imprint of the mouse H19 locus is acquired in the gonocyte stage during foetal testis development. Genes Cells 5, 649-659.

Vagin, V.V., Sigova, A., Li, C., Seitz, H., Gvozdev, V., and Zamore, P.D. (2006). A distinct small RNA pathway silences selfish genetic elements in the germline. Science 313, 320-324.

Vincent, S.D., Dunn, N.R., Sciammas, R., Shapiro-Shalef, M., Davis, M.M., Calame, K., Bikoff, E.K., and Robertson, E.J. (2005). The zinc finger transcriptional repressor Blimp1/Prdm1 is dispensable for early axis formation but is required for specification of primordial germ cells in the mouse. Development 132, 1315-1325.

Volpe, T., and Martienssen, R.A. RNA interference and heterochromatin assembly. Cold Spring Harb Perspect Biol 3, a003731.

Waddington, C.H. (1959). Canalization of development and genetic assimilation of acquired characters. Nature 183, 1654-1655.

Walsh, C.P., Chaillet, J.R., and Bestor, T.H. (1998). Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. Nat Genet 20, 116-117.

Wang, Q.T., Piotrowska, K., Ciemerych, M.A., Milenkovic, L., Scott, M.P., Davis, R.W., and Zernicka-Goetz, M. (2004). A genome-wide study of gene activity reveals developmental signaling pathways in the preimplantation mouse embryo. Dev Cell 6, 133-144.

Watanabe, D., Suetake, I., Tada, T., and Tajima, S. (2002). Stage- and cell-specific expression of Dnmt3a and Dnmt3b during embryogenesis. Mech Dev 118, 187-190.

Watanabe, T., Chuma, S., Yamamoto, Y., Kuramochi-Miyagawa, S., Totoki, Y., Toyoda, A., Hoki, Y., Fujiyama, A., Shibata, T., Sado, T., et al. (2011). MITOPLD is a mitochondrial protein essential for nuage formation and piRNA biogenesis in the mouse germline. Dev Cell 20, 364-375.

Watanabe, T., Tomizawa, S., Mitsuya, K., Totoki, Y., Yamamoto, Y., Kuramochi-Miyagawa, S., Iida, N., Hoki, Y., Murphy, P.J., Toyoda, A., et al. (2011). Role for piRNAs and noncoding RNA in de novo DNA methylation of the imprinted mouse Rasgrf1 locus. Science 332, 848-852.

Wigler, M., Levy, D., and Perucho, M. (1981). The somatic replication of DNA methylation. Cell 24, 33-40.

Wigler, M.H. (1981). The inheritance of methylation patterns in vertebrates. Cell 24, 285-286.

Wossidlo, M., Arand, J., Sebastiano, V., Lepikhov, K., Boiani, M., Reinhardt, R., Scholer, H., and Walter, J. Dynamic link of DNA demethylation, DNA strand breaks and repair in mouse zygotes. EMBO J 29, 1877-1888.

Yabuta, Y., Kurimoto, K., Ohinata, Y., Seki, Y., and Saitou, M. (2006). Gene expression dynamics during germline specification in mice identified by quantitative single-cell gene expression profiling. Biol Reprod 75, 705-716.

Yamaji, M., Seki, Y., Kurimoto, K., Yabuta, Y., Yuasa, M., Shigeta, M., Yamanaka, K., Ohinata, Y., and Saitou, M. (2008). Critical function of Prdm14 for the establishment of the germ cell lineage in mice. Nat Genet 40, 1016-1022.

Yoder, J.A., Walsh, C.P., and Bestor, T.H. (1997). Cytosine methylation and the ecology of intragenomic parasites. Trends Genet 13, 335-340.

Zucker, K.E., Riggs, A.D., and Smith, S.S. (1985). Purification of human DNA (cytosine-5-)-methyltransferase. J Cell Biochem 29, 337-349.

Appendix

Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment

Molecular Cell
Article

Directional DNA Methylation Changes and Complex Intermediate States Accompany Lineage Specificity in the Adult Hematopoietic Compartment

Emily Hodges,[1,2] Antoine Molaro,[1,2] Camila O. Dos Santos,[1,2] Pramod Thekkat,[1,2] Qiang Song,[3] Philip J. Uren,[3] Jin Park,[3] Jason Butler,[2,4] Shahin Rafii,[2,4] W. Richard McCombie,[1] Andrew D. Smith,[3,*] and Gregory J. Hannon[1,2,*]
[1]Watson School of Biological Sciences, Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA
[2]Howard Hughes Medical Institute
[3]Molecular and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA
[4]Department of Genetic Medicine and Ansary Stem Cell Institute, Weill Cornell Medical College, New York, NY 10065, USA
*Correspondence: andrewds@usc.edu (A.D.S.), hannon@cshl.edu (G.J.H.)
DOI 10.1016/j.molcel.2011.08.026

SUMMARY

DNA methylation has been implicated as an epigenetic component of mechanisms that stabilize cell-fate decisions. Here, we have characterized the methylomes of human female hematopoietic stem/progenitor cells (HSPCs) and mature cells from the myeloid and lymphoid lineages. Hypomethylated regions (HMRs) associated with lineage-specific genes were often methylated in the opposing lineage. In HSPCs, these sites tended to show intermediate, complex patterns that resolve to uniformity upon differentiation, by increased or decreased methylation. Promoter HMRs shared across diverse cell types typically display a constitutive core that expands and contracts in a lineage-specific manner to fine-tune the expression of associated genes. Many newly identified intergenic HMRs, both constitutive and lineage specific, were enriched for factor binding sites with an implied role in genome organization and regulation of gene expression, respectively. Overall, our studies represent an important reference data set and provide insights into directional changes in DNA methylation as cells adopt terminal fates.

INTRODUCTION

Development and tissue homeostasis rely on the balance between faithful stem-cell self-renewal and the ordered, sequential execution of programs essential for lineage commitment. Under normal circumstances, commitment is thought to be unidirectional with repressive epigenetic marks stabilizing loss of plasticity (De Carvalho et al., 2010). However, certain differentiated mammalian cells can be reverted to an induced pluripotent state (iPSCs) through exogenous transduction of specific transcription factors (Takahashi and Yamanaka, 2006). Yet, even these reprogrammed cells retain a residual "memory" of their former fate, displaying DNA methylation signatures specific to their tissue of origin (Kim et al., 2010).

DNA methylation is critical for the self-renewal and normal differentiation of somatic stem cells. For example, within the hematopoietic compartment, impaired DNA methyltransferase function disrupts stem cell maintenance (Maunakea et al., 2010; Trowbridge and Orkin, 2010), and loss of DNMT1 leads to defective differentiation and unbalanced commitment to the myeloid and lymphoid lineages (Bröske et al., 2009; Trowbridge et al., 2009). These studies highlight the well-characterized hematopoietic compartment as a context in which to study the link between DNA methylation patterns and cell-fate specification.

Toward this end, DNA methylation profiles of murine hematopoietic progenitors through early stages of lineage commitment were recently compared with CHARM (Irizarry et al., 2008; Ji et al., 2010), which profiles a predefined set of CpG-dense intervals. Overall, CHARM revealed that early lymphopoeisis involves more global acquisition of DNA methylation than myelopoiesis and that DNMT1 inhibition skews progenitors toward the myeloid state. These data support earlier reports that DNMT1 hypomorphic hematopoietic stem and progenitor cells (HSPCs) show reduced lymphoid differentiation potential (Bröske et al., 2009). Importantly, regions identified to have differential methylation through sequential stages of differentiation most often did not correspond to CpG islands (CGIs) but instead lay adjacent in areas referred to as "shores."

Higher-resolution maps of DNA methylation with shotgun bisulfite sequencing have mainly been produced from cultured cells (Laurent et al., 2010; Lister et al., 2009) or mixed cell types (Li et al., 2010). Several unexpected findings emerged from these early studies including significant frequencies of cytosines methylated in a non-CpG context in human embryonic stem cells (ESCs), a characteristic previously thought to be restricted to plants. Other genome-wide studies have implicated DNA methylation in the regulation of alternative promoters and even RNA splicing patterns (Maunakea et al., 2010). These observations emphasize the need for complete, unbiased, and quantitative assessment of cytosine methylation and the establishment of reference methylomes from purified populations of primary cells.

Here, we performed whole-genome shotgun bisulfite sequencing on female human HSPCs, B cells, and neutrophils to

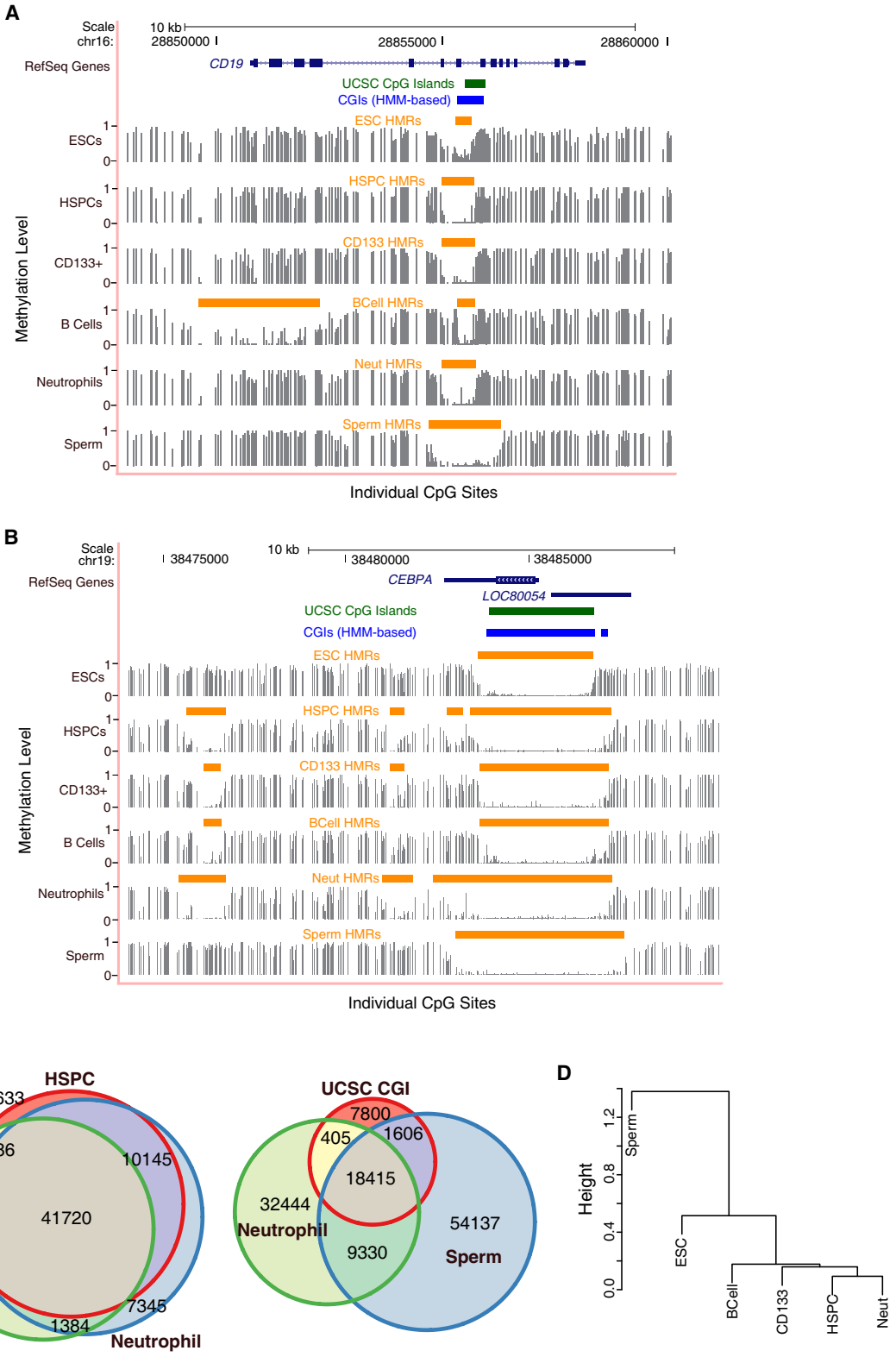


Figure 1. Features of Methylomes in Hematopoietic Cells

(A and B) Genome browser tracks depict methylation profiles across a lymphoid (A) and myeloid (B) specific locus in blood cells, ESCs, and sperm. Methylation frequencies, ranging between 0 and 1, of unique reads covering individual CpG sites are shown in gray with identified hypomethylated regions (HMRs) indicated

examine the relationships between the methylation states of multipotent blood-forming stem cells and two divergent derived lineages. This enabled us to probe directional changes in DNA methylation associated with cell-fate specification. Comparison of the three reference methylomes revealed a number of important principles of epigenetic regulation, in addition to providing insights into the dynamics of epigenetic changes during development.

RESULTS AND DISCUSSION

Lineage-Specific Hypomethylated Regions Extend beyond Annotated CGIs

We sought to generate reference, single nucleotide-resolution methylation profiles for several nodes within the human hematopoietic lineage using whole-genome bisulfite sequencing (see the Experimental Procedures). Therefore, we examined CD34+ CD38–Lin– HSPCs, CD19+ B cells, and granulocytic neutrophils from peripheral blood of pooled human female donors. These cell types represent one of the earliest self-renewing, multipotent populations, and two derived, mature cell types from the lymphoid and myeloid lineages, respectively. For comparison, we generated methylomes from HSPCs from male umbilical cord blood (CD133+CD34+CD38–Lin–) and compared to data sets created from primate sperm (Molaro et al., 2011) and embryonic stem cells (Laurent et al., 2010). In all cases, we achieved a median of 10× independent sequence coverage, sufficient to interrogate 96% of genomic CpG sites (Figure S1A and Table S1A available online). While this level of coverage is still subject to sampling error at individual sites (see discussion in Hodges et al., 2009), features such as transitions from high to low levels of methylation can still be identified with a resolution of the boundaries to within a few CpG sites.

In the genome as a whole, CpG dinucleotides have a strong tendency to be methylated (70%–80%) (Lister et al., 2009). Coincidently, CpGs are also underrepresented, perhaps because of their vulnerability to methylation-induced deamination and consequent loss over evolutionary time (Cooper and Krawczak, 1989; Gardiner-Garden and Frommer, 1987). Areas of increased CpG density, called CpG islands (CGIs) have a lower probability of being methylated and these or their adjacent regions (CGI shores) have been implicated as potential regulatory domains (Gardiner-Garden and Frommer, 1987; Irizarry et al., 2009a; Wu et al., 2010). Though CGIs have been defined computationally (Irizarry et al., 2009b), we developed an algorithm to identify hypomethylated regions (HMRs) empirically in bisulfite sequencing data sets, based on their methylation state alone (see Figures 1A and 1B).

Between 50,000 and 60,000 HMRs were identified from each hematopoietic profile (Table S1B), with neutrophils displaying the greatest number (~60,000), followed by HSPCs (~55,000) and B lymphocytes (~53,000) (Figure 1C). Interestingly, this was lower than the number in male germ cells (~80,000), perhaps because of the extensive repeat hypomethylation observed in sperm as compared to somatic cells.

Certainly, many annotated CGIs were contained within our set of functionally defined HMRs; however, CGIs appeared to fall short as a benchmark by which to define all HMRs with probable regulatory significance. Annotated CGIs accounted for fewer than half of the HMRs identified in any cell type (Figure 1C and Figure S1B). Moreover, many HMRs whose biological relevance is supported by lineage-specific methylation failed to meet the conservative CGI criteria.

Sequence tracks showing methylation levels for a lymphoid- (Figure 1A) or myeloid- (Figure 1B) specific gene illustrate several characteristics of HMRs. The locus for the B cell marker *CD19* displays a broad, cell type-specific HMR at its transcriptional start site (TSS), which does not overlap a predicted CGI. In contrast, "tidal" methylation at CGI shores characterizes several HMRs surrounding the myeloid transcription factor, *CEBPA*. The cores of these HMRs are shared among blood forming cells, but their widths differ, with neutrophils demonstrating the most expansive hypomethylation. In fact, shared HMRs often show variable widths, suggesting that the boundaries of HMRs fluctuate in a cell type-dependent manner. Due to the dynamic behavior of the HMRs, we were motivated to seek further validation of these characteristics as biological phenomena, rather than as technical artifacts of the methodology. Therefore, we focused on an independent data-set derived from chimpanzee. We reasoned that genic relationships to methylation dynamics should be preserved in closely related species. Indeed, HMRs show significant overlap between human and chimp, with chimp HMRs following very similar patterns of boundary fluctuations (Table S1C and Figure S2).

While a high proportion of identified HMRs ($\geq$70%) intersected all blood cell types studied, ~10-fold more HMRs were shared only between HSPCs and neutrophils than exclusively between HSPCs and B cells (Figure 1C). In contrast, ~45%–50% of HMRs identified in blood cells overlap sperm HMRs. Interestingly, the diversity of differentially expressed genes within the hematopoietic lineage has been reported to be similar to the complexity observed across human tissues (Novershtern et al., 2011). However, at the epigenetic level, HMR profiles easily distinguished closely related cell types (blood forming) from distantly related ones (Figure 1D), indicating that patterns of DNA methylation are strongly correlated within a lineage.

HMR Expansion Correlates with Differential Expression

Differentially methylated regions (DMRs) at promoters have been ascribed regulatory roles, with differential methylation being

by orange bars. UCSC predicted/annotated CpG islands (green bars) and HMM-based CpG islands (blue bars) (Irizarry et al., 2009b) are also displayed. Numbers (top) indicate base position along the chromosome.

(C) Venn diagrams depict the intersection between HMRs identified in blood as well as the overlap between blood-derived cells, sperm, and UCSC CpG islands. The size of the circles and the proportion of circle overlap reflect the relative number of HMRs identified as well as the degree of intersection between each set of HMRs.

(D) Dendrogram clusters cell-types according to their pearson correlations of individual CpG methylation levels within HMRs, both overlapping and nonoverlapping, across all tissues examined.

See also Figures S1 and S2 and Table S1.

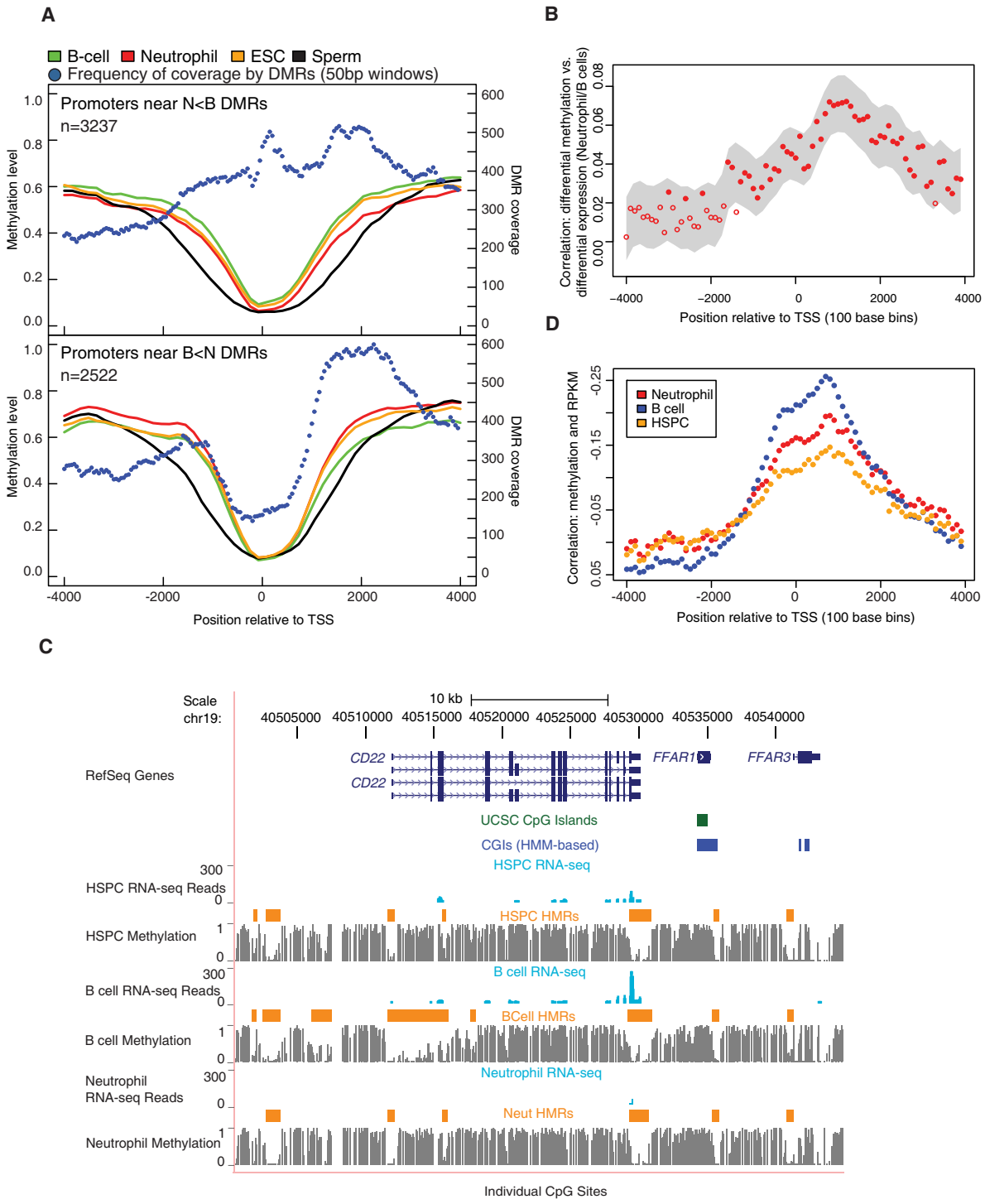


Figure 2. Promoter Differential Methylation and Gene Expression

(A) Average methylation levels across promoters of genes having a DMR within 4 kb of the TSS are shown. Two separate graphs display neutrophil hypomethylated promoter DMRs relative to B cells (N < B, top) and B cell hypomethylated promoter DMRs relative to neutrophils (B < N, bottom). The number of DMRs covering nonoverlapping 50 bp windows across the promoter is also shown.

(B) Correlations between differential methylation and differential expression between neutrophils and B cells as a function of position relative to the TSS are shown. The correlations were obtained by comparing log odds of differential methylation and log of RPKM. The probability for differential methylation at a given CpG is described in the Supplemental Experimental Procedures. The gray area displays the smoothed 95% confidence interval. The closed circles indicate correlation coefficients that are significantly different from 0.

linked to tissue-specific expression. Yet, HSPCs, B cells, and neutrophils mainly share promoter-associated HMRs at differentially expressed genes. Prior studies have associated changes in gene expression with changes in methylation states adjacent to constitutively hypomethylated CGIs, in so-called "CGI shores" (Irizarry et al., 2009a). Therefore, we examined correlations between the geography of promoter HMRs and changes in lineage-specific expression, focusing on a comparison of B cells and neutrophils.

Differential methylation often manifested as a broadening of TSS-associated HMRs in a specific lineage (Table S2A). The changes were asymmetric, with the greatest loss of methylation on the gene-ward side (Wilcoxon ranks sum: $p < 5e\text{-}60$, both DMR sets). Globally, these HMRs were broadest in sperm and constricted in ESCs (Figure 2A) (see also Molaro et al., 2011), widening again in a tissue-specific fashion. Thus, our analyses provide global support for "tidal" methylation changes at CGI shores.

For deeper analysis of these tidal patterns, we measured differential methylation in 50 base windows surrounding TSSs (Figure 2A). Moving 3′ toward B cell hypomethylated promoters (B < N), coverage by DMRs peaked between 1.5 Kbp and 2 Kbp downstream of the TSS. A slightly different pattern was observed for neutrophil hypomethylated promoters (N < B), with DMRs rising to a peak directly at the TSS. In both data sets, the greatest concentration of differential methylation occurred ~1–2 Kb downstream of the TSS, consistent with overall methylation being selectively reduced in the transcribed regions of genes with tissue-specific DMRs.

We next asked whether any element of DMR geography correlated with tissue-specific gene expression. We carried out RNA-seq and computed RPKM values for each cell type (Table S2B). We then computed the correlation between differential expression and differential methylation in 100 base windows surrounding the TSS (see the Experimental Procedures). This correlation was strongly asymmetric, peaking ~1,000 bases downstream of the TSS. Notably, this corresponded with the expansion of HMRs that contributes to tissue-specific promoter hypomethylation (Figure 2B).

CD22 provides a specific example of the general phenomena that we observed (Figure 2C). *CD22* is expressed in B cells, but not neutrophils. In each cell type its TSS is covered by an HMR, which in HSPCs and neutrophils extends ~500 bp and centered on the TSS. In B cells, the HMR begins at the same position upstream of the *CD22* TSS, but extends more than 4,300 bp into the transcribed region.

The properties noted for differentially expressed genes were extensible to the entire set of REFSEQ genes. Though hypomethylation was largely symmetric around REFSEQ TSSs, a strong correlation could be seen between RPKM and lower methylation levels peaking ~1.0 Kb downstream of the TSS (Figure 2D). This was true of all cell types examined, though the magnitude of the effect was lowest in HSPCs.

Our results are in accord with a recent study that revealed a unique chromatin signature surrounding the TSS of tissue-specific loci. Spreading of H3K4me2 into the 5′ untranslated region (UTR) was observed at tissue-specific genes, whereas it remained as a discrete peak at the TSS of ubiquitously expressed genes (Pekowska et al., 2010). To look for similar relationships between histone profiles and expanding promoter HMRs, we analyzed chromatin immunoprecipitation sequencing (ChIP-seq) data for H3K4me3, H3K4me1, and H3K27ac enrichment across eight different ENCODE cell lines (Bernstein et al., 2005; Birney et al., 2007). The ENCODE cell lines are derived from a variety of tissues and include GM12878, which is a lymphoblastoid cell line. First, we observe a strong enrichment for these histone marks at B cell promoters containing expanded HMRs. In addition, the greatest difference between the lymphoid cell line and the other cell lines appears upstream and downstream of the TSS compared to all promoters. Interestingly, the H3K4me3 differential enrichment is biased on the 3′ side of the TSS (Figure 3).

It has also been noted that for a subset of CGI-associated promoters, high CpG density extends downstream of the TSS and hypomethylation of the extended region is required for RNA polymerase II binding (Appanah et al., 2007). In fact, analysis of existing lymphoid ChIP-seq data of RNA polymerase II revealed a 3× enrichment in B cell expanded HMR regions compared to neutrophil-expanded regions (Table S2C) (Barski et al., 2010). This suggests that while core CGI promoters remain hypomethylated by default, expansion downstream of the TSS may be important for productive transcription.

Features of Shared and Lineage-Specific Intergenic HMRs

While REFSEQ gene promoters were often associated with an HMR, the majority of HMRs were not found at promoters (Figure S3). Nearly half of all identified HMRs were located in gene bodies. An additional quarter lay >10 Kb from the nearest annotated genes, and we defined this class as "intergenic HMRs."

Like promoter-associated HMRs, intergenic HMRs showed sequence conservation, suggesting that these are functional elements (Figure 4A). In fact, genome-wide comparisons of methylation states of orthologous sites in the corresponding cell types of chimpanzee supported concomitant conservation of constitutive and cell type-specific patterns of intergenic methylation (data not shown). Intergenic HMRs tended to be narrower than those found at promoters and were less likely to be shared among cell types. When they were shared, they displayed patterns of expansion and contraction very similar to what was observed for promoter-associated regions (Figure 4A), with their overall extent being widest in sperm.

(C) The browser image shows gene expression for CD22 in the form of mapped read profiles from RNA-seq data. Methylation profiles are also shown (as in Figure 1A) along with HMRs.

(D) Correlations between methylation levels and expression levels represented by RPKM values are shown as a function of position relative to the TSS. Correlation coefficients were averaged in 100 bp bins across regions between 4 kb upstream and downstream of the TSS. Y axis labels were reversed.

See also Figure S3 and Table S2.

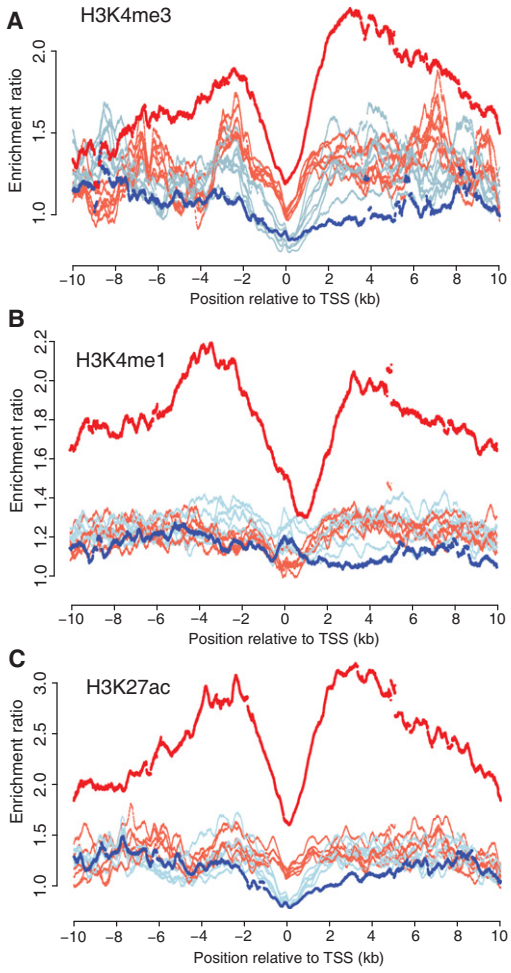


Figure 3. Histone Enrichment across Expanded HMRs

Read count enrichment ratios per 25 bp bins located 10 kb upstream and 10 kb downstream of the TSS were calculated for promoters overlapping HMRs included in Figure 2A for B cell HMRs (red lines) or neutrophil HMRs (blue lines) for H3K4me3 (A), H3K4me1 (B), and H3K27ac (C) by comparison of read counts across all REFSEQ annotated promoters. Data were obtained from ENCODE and include histone profiles for eight different cell lines. The lymphoblastoid cell line GM12878 is highlighted in darker shaded colors.

An early, pervasive view of DNA methylation proposed that germ cell profiles should represent a default state of hypomethylation in all potential regulatory regions (Gardiner-Garden and Frommer, 1987). This was based on the idea that hypomethylation in germ cells would prevent CpG erosion over evolutionary time spans. The high number of nonoverlapping HMRs in the adult somatic cell strongly argues against both of these notions (Figure 1C). However, the width of both genic and intergenic HMRs in sperm compared to somatic cells suggests that germ cells can define the ultimate boundaries of somatic HMRs.

Guided by the strong general enrichment for potential transcription factor binding sites in all HMRs (see Table 1), we searched for motifs in intergenic DMRs specific to neutrophils or B cells (Figure 4B). The strongest scoring motifs in the neutrophil-specific intergenic DMRs included those associated with C/EBP and ETS families, along with HLF and STAT motifs. This striking enrichment for C/EBP and ETS family binding sites is consistent with the functions of ETS factor PU.1 and several C/EBP factors as multipotent progenitors commit to become myeloblasts, which ultimately give rise to neutrophils (Nerlov and Graf, 1998). Because the ETS family contains a large number of transcription factors, we sought experimental support for their binding at HMRs. Therefore we probed existing ChIP-seq data of PU.1 from human HSPCs (Novershtern et al., 2011). We find numerous examples PU.1 enrichment in HMRs, several of which are provided in Figure S4. In contrast, the strongest scoring motifs in B cell-specific intergenic DMRs included the EBF motif, POU family motifs, E-boxes, a PAX motif, and those associated with NFκB and IRF. The simultaneous enrichment of EBF, E-box, and PAX motifs is consistent with the interacting roles of EBF, E2A (which binds E-boxes) and PAX5 as common lymphoid progenitors progress along the B cell lineage (Lin et al., 2010; Medina et al., 2004; Sigvardsson et al., 2002). The enrichment of NFκB and IRF motifs is consistent with the known roles for these factors in both activation and differentiation of lymphocytes (Hayden et al., 2006). Considered together, these analyses strongly suggest that at least a subset of intergenic DMRs can be engaged by tissue-specific transcription factors, leading to changes in chromatin organization that might have long-distance impacts on annotated genes or more local impacts on as yet unidentified ncRNAs. In fact, we do find evidence of transcriptional activity surrounding intergenic DMRs in our RNA-seq data sets, but we have not yet pursued this observation further (data not shown). Irrespective of the model, our results strongly support the biological relevance of tissue-specific intergenic HMRs.

We also probed the possible functions of shared intergenic HMRs. Prior studies had experimentally identified binding sites for the insulator protein, CTCF, by chromatin immunoprecipitation (Kim et al., 2007). These sites are strongly enriched (155-fold) in nonrepeat intergenic HMRs that are common to all cell types examined. In fact, ~90% (>500) of the nonrepeat, shared intergenic HMRs contain a CTCF site. This correlates with the known propensity of CTCF to bind unmethylated regions and suggests that many of the shared intergenic HMRs that we detect may function in the structural organization of chromosomes and nuclear domains.

Myeloid-Biased, Poised Methylation States Characterize HSPC Methylomes

For loci whose differential expression characterizes the lymphoid and myeloid lineages, we set out with a simple general expectation. Low methylation levels in stem and progenitor cells would be permissive for expression in either lineage, and an accumulation of methylation during differentiation would correlate with silencing of loci in the lineage in which they are not expressed.

To test this hypothesis, we selected lineage-specific HMRs arising from a comparison of neutrophils and B cells and examined their status in HSPCs. Both at the level of individual CpGs (Figure 5A) and at the level of overall methylation (Figure 5B), HSPCs showed intermediate methylation states at sites where B cells and neutrophils show opposing methylation patterns.

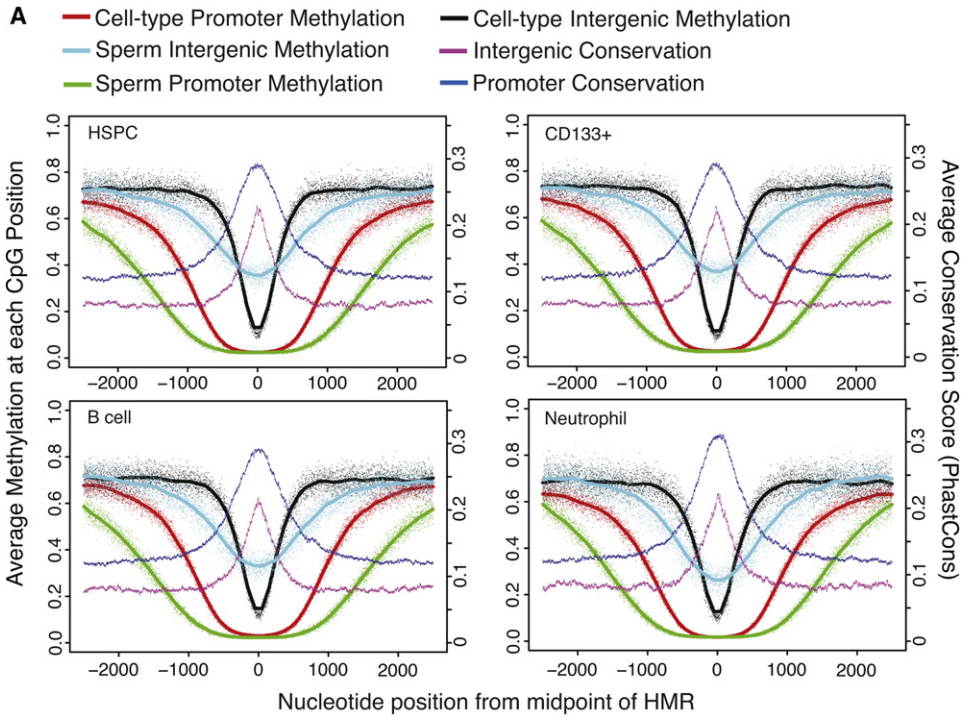


Figure 4. Features of Intergenic HMRs and DMRs

(A) Composite methylation profiles are plotted for individual CpG sites within HMRs. The x axes of the plots indicate genomic position centered on the midpoint of HMRs in the reference cell type labeled for each plot. Methylation profiles are given for the reference cell and sperm, separately for regions where the reference

This suggests that differentiation involves both gains and losses of DNA methylation at lineage-specific HMRs, an observation consistent with recent studies using other methodologies (Attema et al., 2007; Claus et al., 2005; Ji et al., 2010).

At the level of individual CpGs, HSPC patterns correlated better with those seen in neutrophils at myeloid HMRs than they did with B cell methylation patterns at nonoverlapping lymphoid HMRs (Figure 5A). Moreover, the median methylation level for B cells at B cell DMRs was more than twice as high as the median level at neutrophil specific DMRs (Figure 5B). This finding, along with the fact that B cells exhibited fewer total HMRs than either HSPCs or neutrophils, supported an earlier observation that lymphoid commitment in mice involves globally increased DNA methylation (Ji et al., 2010). As a whole, our results indicate that the HSPC methylome has more myeloid than lymphoid character. Many fewer DMRs were identified in comparisons of HSPC and neutrophil methylation profiles than of HSPCs and B cells (Figure S3). Such a myeloid bias is also consistent with prior studies, which point to the myeloid lineage as a default differentiation path for HSPCs (Månsson et al., 2007).

Regions that exhibit intermediate methylation occurred in two forms. The well-documented mode is allelic methylation that is characteristic of dosage compensated and imprinted genes. We detected such loci abundantly in our data sets, and these encompassed both known monoallelic genes and new candidates for monoallelic expression (data not shown). More prevalent were regions of intermediate methylation wherein each chromosome displayed different patterns of CpG modification with little correlation between the states of adjacent CpGs. Partially methylated regions were previously noted in ESCs (Lister et al., 2009), though they did not investigate whether these presented allelic versus stochastic and complex patterns.

To discriminate between allelic and complex patterns, we performed targeted conventional bisulfite PCR sequencing of individual clones from HSPCs across a selected set of myeloid loci and a known locus with allele-specific methylation (Figure 5C, Figure S5, and Table S3). This allowed detailed analysis of adjacent CpG methylation on individual molecules. As expected, for the allelic *XIST* locus on chromosome X, we observed uniform methylation profiles of adjacent CpG sites within individual clones representing two states that contributed nearly equally to the partial methylation observed. In contrast, the myeloid *AZU1* locus exemplified a stochastic pattern of methylation in HSPC. We cannot determine whether the complex states that we observed were in dynamic equilibrium or whether they were fixed in each chromosome that contributed to our analysis.

While the mechanisms underlying complex, partial methylation patterns in HSPCs are unclear, they are reminiscent of bivalent promoters that contain both repressive and active histone marks (Bernstein et al., 2006). Both during embryonic development and during stem cell differentiation, such poised promoters are converted to a determinate chromatin state by shifting the balance of histone marks. This has already been noted for lineage-specific genes in HSPCs (Attema et al., 2007), and our data indicate that this well-established property of chromatin may also extend to DNA methylation patterns.

Alternative explanations for our results must also be considered. Since we have used pooled individuals, each of the observed patterns could be specific to one donor, giving rise to a complex pool of clones; however, this seems unlikely as we also detect lower correlations between neighboring CpGs within single clones. Alternatively, complex states could represent heterogeneity within the isolated HSPC population (see Figure S6), with our data coming from a mixture of self-renewing and more committed cell types. To investigate this possibility, we searched within our RNA-seq data for expression patterns characteristic of each purified cell population. Transcriptional profiles revealed the top differentially expressed genes within the HSPC compartment to be highly enriched for signature gene markers associated with self-renewing hematopoietic stem cells (Figure 5D) and depleted for genes associated with committed progenitors. Collectively, these data suggest that the observed methylation patterns are likely derived from a highly enriched stem cell population, and indicate that those populations may naturally adopt complex, potentially dynamic, methylation patterns at lineage-specific HMRs.

Both the general trends of methylation loss along a lineage and the possibility of dynamic poised methylation states imply that demethylation, either passive or active, is a common event. In mammals, factors capable of promoting active demethylation have remained somewhat elusive (Ooi and Bestor, 2008). In vitro studies have demonstrated that MBD2, a methyl-CpG binding protein, can specifically demethylate cytosines, and components of the elongator complex and the cytidine deaminase, AID, have been implicated in demethylation during early development (Bhattacharya et al., 1999; Okada et al., 2010; Popp et al., 2010). Furthermore, in zebrafish, the coordinated activities of glycosylases, deaminases, and DNA repair proteins have been reported to cause differentiation defects when disrupted, and this has been posited as an effect of improper DNA methylation (Rai et al., 2010). Alternatively, demethylation could potentially be achieved through the action of hydroxymethylases (e.g., TET1-3), which have been proposed to execute an intermediate step toward methylation loss (Ito et al., 2010; Tahiliani et al., 2009; Zhang et al., 2010). Additional information will be necessary to resolve the relevance of any of these pathways to the transition in methylation states between HSPCs and mature neutrophils and B cells.

As a whole, our data not only provide insights into the global behavior of DNA methylation, both in individual cell types and along a well-characterized lineage, but also provide a critical

cell HMR spans a TSS and intergenic region (>10 Kbp from any RefSeq transcript; not overlapping a repeat). Average cross-species conservation scores from PhyloP probabilities derived from 44-way multiple alignments are plotted separately for promoter and intergenic HMRs.
(B) Transcription factor binding site motifs enriched in DMRs between neutrophils and B cells are shown. The top 20 most enriched motifs are shown separately for N < B and B < N DMRs, based on the motifclass tool in the CREAD package. See the Supplemental Experimental Procedures for details of enrichment calculations.
See also Figures S3 and S4.

Table 1. TFBS Enrichment in HMRs across Intergenic and Promoter Regions

Cell	Region	CGI?	HMR[a]	TFBS	Expected	Enrich-ment
N/A	promoter		34,257	244,998	91,570.8	2.7
	promoter	cgi	24,601	191,452	65,760.9	2.9
	promoter	nocgi	9,656	53,852	25,810	2.1
	intergenic	cgi	10,630	13,608	4,603.76	3.0
B Cell	all		53,834	339,943	76,196.1	4.5
	intergenic		5,849	16,150	3,779	4.3
	intergenic	cgi	1,670	4,802	1,194.97	4.0
	intergenic	nocgi	4,179	11,348	2,584.01	4.4
	promoter		13,650	212,644	36,548.3	5.8
	promoter	cgi	12,828	206,556	35,080	5.9
	promoter	nocgi	822	6,088	1,468.27	4.1
CD133	all		49,593	339,191	67,778.2	5.0
	intergenic		6,494	17,708	3,816.73	4.6
	intergenic	cgi	1,630	4,817	1,207.45	4.0
	intergenic	nocgi	4,864	12,891	2,609.26	4.9
	promoter		13,745	224,955	37,395.1	6.0
	promoter	cgi	12,965	219,407	36,309.9	6.0
	promoter	nocgi	780	5,548	1,085.18	5.1
ESC	all		40,476	318,377	65,062.3	4.9
	intergenic		3,768	11,220	2,404.28	4.7
	intergenic	cgi	1,151	3,295	882.802	3.7
	intergenic	nocgi	2,617	7,925	1,521.45	5.2
	promoter		13,098	222,654	36,332.4	6.1
	promoter	cgi	12,661	218,765	35,769.4	6.1
	promoter	nocgi	437	3,889	562.951	6.9
HSPC	all		55,984	352,574	77,671.2	4.5
	intergenic		6,154	17,619	3,972.1	4.4
	intergenic	cgi	1,663	4,775	1,222.27	3.9
	intergenic	nocgi	4,491	12,844	2,749.81	4.7
	promoter		13,820	222,635	37,830.8	5.9
	promoter	cgi	12,948	216,433	36,461.3	5.9
	promoter	nocgi	872	6,202	1,369.4	4.5
Neut.	all		60,594	362,074	82,427.7	4.4
	intergenic		6,422	18,515	4,212.75	4.4
	intergenic	cgi	1,626	4,760	1,243.88	3.8
	intergenic	nocgi	4,796	13,755	2,968.85	4.6
	promoter		13,862	224,621	38,503.6	5.8
	promoter	cgi	12,950	218,281	37,060.6	5.9
	promoter	nocgi	912	6,340	1,442.93	4.4
Sperm	all		81,446	440,856	201,006	2.2
	intergenic		2,616	14,903	3,158.15	4.7
	intergenic	cgi	865	6,181	1,307.11	4.7
	intergenic	nocgi	1,751	8,722	1,851.02	4.7
	promoter		14,051	270,798	63,641.3	4.3
	promoter	cgi	13,588	266,658	62,357.8	4.3
	promoter	nocgi	463	4,140	1,283.49	3.2

Enrichment of predicted transcription factor binding sites (TFBSs) in intergenic HMRs and HMRs that overlap promoters. For each set of HMRs, corresponding to a cell type, the TFBS enrichment (observed/expected site counts) is given for all HMRs, those overlapping promoters, those that are intergenic, separately according to whether the HMRs overlap CGIs. Data are presented for each of the following cell types: B cells, CD133 cord blood, HSPCs, ESCs, neutrophils, and sperm. For comparison, the TFBS enrichment in the full set of promoters (including those overlapping CGIs) is given, along with enrichment in the full set of intergenic CGIs.

[a] For the "N/A" group, the HMRs are simply the number of promoters or CGIs.

reference data set to enable detailed future studies of both the mechanisms that set somatic DNA methylation patterns and the consequences of those patterns for gene expression and genome organization.

EXPERIMENTAL PROCEDURES

Flow Cytometry and DNA Extraction

Peripheral blood was collected from six healthy female donors ages 25–35 and pooled. After isolation by Ficoll gradient, mononuclear cells were fixed in 1% paraformaldehyde (PFA) and stained with antibodies against the following human cell surface markers (eBiosciences): anti-CD34 (mucosialin) conjugated to PE-Cy7, anti-CD38 conjugated to APC, anti-CD45 conjugated to PE, anti-CD19 conjugated to PE, and anti-CD235a (Glycophorin) conjugated to PE. For lineage depletion, either a combination of PE-conjugated antibodies against CD45, CD19, and CD235a or a commercially available human hematopoietic lineage cocktail was used. CD34+CD38–Lin– hematopoietic stem cells and CD19+ B cells were purified with the FACSAriaII (Becton Dickinson). Neutrophils were purified according to their forward and side-scatter profile. FACS profiles are provided in Figure S6. Umbilical cord blood was collected from a single donor, and CD133+ cells were selected via magnetic separation on CD133+ microbeads (Milteny Biotec) according to instructions supplied by the manufacturer. Two column separations were performed for additional purity. All cells were collected in cell lysis buffer (50 mM Tris, 10 mM EDTA and 1% SDS), and PFA induced crosslinks were reversed with RNase A and a 65°C incubation overnight, after which residual proteins were digested with Proteinase K for 3 hr at 42°C. DNA was extracted with an equal volume of phenol:chloroform, followed by a single extraction with chloroform and ethanol precipitation. Human sperm was purified and sequenced according to methods described in Molaro et al. (2011).

Illumina Library Preparation for Bisulfite Sequencing

Bisulfite sequencing libraries were generated by previously described methods (Hodges et al., 2009) and on the manufacturer's instructions (Illumina) but with several additional modifications. In brief, after each enzymatic step, genomic DNA was recovered by phenol:chloroform extraction and ethanol precipitation. Adenylated fragments were ligated to Illumina-compatible paired-end adaptors synthesized with 5′-methyl-cytosine, and, when necessary, adaptors were diluted 100×–1000× to compensate for low-input libraries and maintain an approximate 10-fold excess of adaptor oligonucleotides. After ligation, DNA fragments were purified and concentrated on MinElute columns (QIAGEN). The standard gel purification step for size selection was excluded from the protocol. Fragments were denatured and treated with sodium bisulfite with the EZ DNA Methylation Gold kit according to the manufacturer's instructions (Zymo). Lastly, the sample was desulfonated and the converted, adaptor-ligated fragments were PCR enriched with paired-end adaptor-compatible primers 1.0 and 2.0 (Illumina) and the Expand High Fidelity Plus PCR system (Roche). Paired-end Illumina sequencing was performed on bisulfite converted libraries for 76–100 cycles each end.

RNA-Seq

For isolation of RNA from target cell populations, unfixed (live) cells were sorted as described above into Trizol-LS (Invitrogen), and RNA was purified

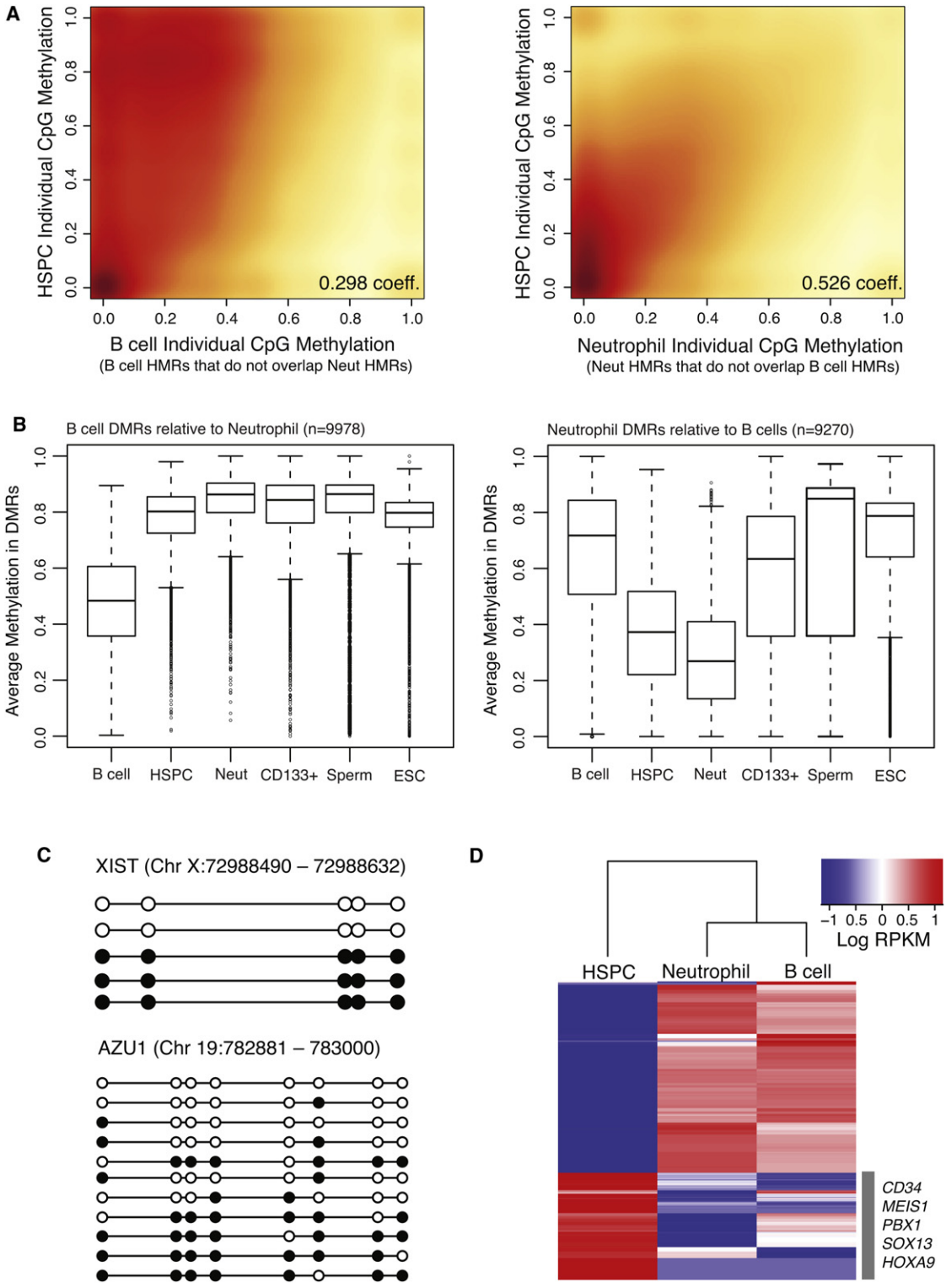


Figure 5. Methylation Dynamics during Lineage Selection

(A) Smoothed scatter plot heat maps showing the correlation between individual CpG methylation levels in HSPCs versus B cells (left) and HSPCs versus neutrophils (right) within B cell- and neutrophil-specific HMRs, respectively. Darker shading (red) indicates greater density of data points, while lighter (yellow) shading reflects lower density. Positive correlations between HSPCs and both B cells and neutrophils indicate an intermediate state for HSPCs.

according to the manufacturer's recommendations. Double-stranded complementary DNA (cDNA) libraries were generated with the Ovation RNA-seq system (Nugen). After reverse transcription and cDNA amplification, double-stranded cDNA fragments were phosphorylated, adenylated, and ligated to Illumina paired-end adaptors followed by 15 cycles of PCR amplification with Phusion HF PCR master mix (Finnzymes) according to the standard Illumina protocol for genomic libraries. Single-end sequencing was performed for 36 cycles.

Conventional Bisulfite Cloning and Sanger Sequencing

Genomic DNA isolated from pooled human HSPCs was bisulfite converted with the EZ DNA Methylation Gold kit (Zymo). For selection of specific regions for amplification, forward and reverse primers were designed with Methprimer (Li and Dahiya, 2002). Primer sequences are provided in the Table S3. The following PCR reaction components were combined in a total volume of 25 μl: 5 μl 5× Expand High Fidelity Plus buffer without $MgCl_2$, 1 μl 10 mM dNTPs, 1 μl 10 mM each forward and reverse primers, 2.5 μl 25 mM $MgCl_2$, 2 μl DNA template, and 11.5 μl nuclease-free water. Thermal cycling was performed as follows: 35 cycles each of denaturation at 94°C for 2 min, annealing at 60°C or 53°C for 1 min, and extension at 72°C for 30 s followed by 7 min at 72°C. The PCR products were purified on columns with a PCR purification kit (QIAGEN). PCR products were adenylated with Klenow exo– and purified. Purified amplicons were cloned and sequenced according to previously described methods (Hodges et al., 2009).

Computational Methods Summary

The Supplemental Experimental Procedures contain a detailed description of computational methods. Mapping bisulfite treated reads was done with methods described by Smith et al. (2009) with tools from the RMAP package (Smith et al., 2009). Hypomethylated regions (HMRs) were identified with a hidden Markov model as described in Molaro et al. (2011). DMRs were identified by (1) computation of probabilities of differential methylation at individual CpGs based on number of reads and frequencies of methylation, and (2) identification of peaks in these profiles after kernel smoothing. Cross-species conservation information was taken from UCSC MULTIZ 44-way vertebrate alignments and PhyloP profiles from these alignments.

ACCESSION NUMBERS

Data analyzed herein have been deposited in GEO with accession number GSE31971.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, six figures, and three tables and can be found with this article online at doi:10.1016/j.molcel.2011.08.026.

ACKNOWLEDGMENTS

We thank members of the McCombie lab and Michelle Rooks for help with experimental procedures, and Assaf Gordon, Luigi Manna, and the Cold Spring Harbor Laboratory and University of Southern California High Performance Computing Centers for computational support. Chimp blood was supplied by the New Iberia Research Center and the Southwest National Primate Center. This work was supported in part by grants from the National Institutes of Health and by a kind gift from Kathryn W. Davis (A.S., G.J.H.). The ENCODE ChIP-seq data were generated at the Broad Institute and in the Bradley E. Bernstein lab at the Massachusetts General Hospital/Harvard Medical School. Data generation and analysis was supported by funds from the National Human Genome Research Institute, the Burroughs Wellcome Fund, Massachusetts General Hospital, and the Broad Institute.

Received: May 20, 2011
Revised: July 19, 2011
Accepted: August 26, 2011
Published online: September 15, 2011

REFERENCES

Appanah, R., Dickerson, D.R., Goyal, P., Groudine, M., and Lorincz, M.C. (2007). An unmethylated 3′ promoter-proximal region is required for efficient transcription initiation. PLoS Genet. *3*, e27.

Attema, J.L., Papathanasiou, P., Forsberg, E.C., Xu, J., Smale, S.T., and Weissman, I.L. (2007). Epigenetic characterization of hematopoietic stem cell differentiation using miniChIP and bisulfite sequencing analysis. Proc. Natl. Acad. Sci. USA *104*, 12371–12376.

Barski, A., Chepelev, I., Liko, D., Cuddapah, S., Fleming, A.B., Birch, J., Cui, K., White, R.J., and Zhao, K. (2010). Pol II and its associated epigenetic marks are present at Pol III-transcribed noncoding RNA genes. Nat. Struct. Mol. Biol. *17*, 629–634.

Bernstein, B.E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D.K., Huebert, D.J., McMahon, S., Karlsson, E.K., Kulbokas, E.J., 3rd, Gingeras, T.R., et al. (2005). Genomic maps and comparative analysis of histone modifications in human and mouse. Cell *120*, 169–181.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315–326.

Bhattacharya, S.K., Ramchandani, S., Cervoni, N., and Szyf, M. (1999). A mammalian protein with specific demethylase activity for mCpG DNA. Nature *397*, 579–583.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., et al; ENCODE Project Consortium; NISC Comparative Sequencing Program; Baylor College of Medicine Human Genome Sequencing Center; Washington University Genome Sequencing Center; Broad Institute; Children's Hospital Oakland Research Institute. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799–816.

Bröske, A.M., Vockentanz, L., Kharazi, S., Huska, M.R., Mancini, E., Scheller, M., Kuhl, C., Enns, A., Prinz, M., Jaenisch, R., et al. (2009). DNA methylation protects hematopoietic stem cell multipotency from myeloerythroid restriction. Nat. Genet. *41*, 1207–1215.

Claus, R., Almstedt, M., and Lübbert, M. (2005). Epigenetic treatment of hematopoietic malignancies: in vivo targets of demethylating agents. Semin. Oncol. *32*, 511–520.

Cooper, D.N., and Krawczak, M. (1989). Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. Hum. Genet. *83*, 181–188.

(B) Box plots show the distribution of average methylation levels in regions of differential methylation (DMRs) between B cells and neutrophils. Whiskers represent minimum and maximum values, while boxes depict the interquartile range, with horizontal lines indicating the median value. Outliers are shown as open circles.
(C) Lollipop diagrams display the methylation status of HSPC-derived clones sequenced by conventional methods following bisulfite conversion and site-specific PCR amplification across an interval near the XIST gene (top) and the AZU1 gene (bottom). Filled and open circles represent methylated and unmethylated CpG sites, respectively.
(D) Heat map of log RPKM values show expression levels for the top 100 differentially expressed genes (rows), selected for high expression in one cell type compared to the other, in each cell population (columns). Signature marker genes found within the HSPC cluster are listed.
See also Figures S3, S5, and S6 and Table S3.

De Carvalho, D.D., You, J.S., and Jones, P.A. (2010). DNA methylation and cellular reprogramming. Trends Cell Biol. *20*, 609–617.

Gardiner-Garden, M., and Frommer, M. (1987). CpG islands in vertebrate genomes. J. Mol. Biol. *196*, 261–282.

Hayden, M.S., West, A.P., and Ghosh, S. (2006). NF-kappaB and the immune response. Oncogene *25*, 6758–6780.

Hodges, E., Smith, A.D., Kendall, J., Xuan, Z., Ravi, K., Rooks, M., Zhang, M.Q., Ye, K., Bhattacharjee, A., Brizuela, L., et al. (2009). High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing. Genome Res. *19*, 1593–1605.

Irizarry, R.A., Ladd-Acosta, C., Carvalho, B., Wu, H., Brandenburg, S.A., Jeddeloh, J.A., Wen, B., and Feinberg, A.P. (2008). Comprehensive high-throughput arrays for relative methylation (CHARM). Genome Res. *18*, 780–790.

Irizarry, R.A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., et al. (2009a). The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat. Genet. *41*, 178–186.

Irizarry, R.A., Wu, H., and Feinberg, A.P. (2009b). A species-generalized probabilistic model-based definition of CpG islands. Mamm. Genome *20*, 674–680.

Ito, S., D'Alessio, A.C., Taranova, O.V., Hong, K., Sowers, L.C., and Zhang, Y. (2010). Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. Nature *466*, 1129–1133.

Ji, H., Ehrlich, L.I., Seita, J., Murakami, P., Doi, A., Lindau, P., Lee, H., Aryee, M.J., Irizarry, R.A., Kim, K., et al. (2010). Comprehensive methylome map of lineage commitment from haematopoietic progenitors. Nature *467*, 338–342.

Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V.V., and Ren, B. (2007). Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell *128*, 1231–1245.

Kim, K., Doi, A., Wen, B., Ng, K., Zhao, R., Cahan, P., Kim, J., Aryee, M.J., Ji, H., Ehrlich, L.I., et al. (2010). Epigenetic memory in induced pluripotent stem cells. Nature *467*, 285–290.

Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C.T., Low, H.M., Kin Sung, K.W., Rigoutsos, I., Loring, J., and Wei, C.L. (2010). Dynamic changes in the human methylome during differentiation. Genome Res. *20*, 320–331.

Li, L.C., and Dahiya, R. (2002). MethPrimer: designing primers for methylation PCRs. Bioinformatics *18*, 1427–1431.

Li, Y., Zhu, J., Tian, G., Li, N., Li, Q., Ye, M., Zheng, H., Yu, J., Wu, H., Sun, J., et al. (2010). The DNA methylome of human peripheral blood mononuclear cells. PLoS Biol. *8*, e1000533.

Lin, Y.C., Jhunjhunwala, S., Benner, C., Heinz, S., Welinder, E., Mansson, R., Sigvardsson, M., Hagman, J., Espinoza, C.A., Dutkowski, J., et al. (2010). A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate. Nat. Immunol. *11*, 635–643.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315–322.

Månsson, R., Hultquist, A., Luc, S., Yang, L., Anderson, K., Kharazi, S., Al-Hashmi, S., Liuba, K., Thorén, L., Adolfsson, J., et al. (2007). Molecular evidence for hierarchical transcriptional lineage priming in fetal and adult stem cells and multipotent progenitors. Immunity *26*, 407–419.

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.

Medina, K.L., Pongubala, J.M., Reddy, K.L., Lancki, D.W., Dekoter, R., Kieslinger, M., Grosschedl, R., and Singh, H. (2004). Assembling a gene regulatory network for specification of the B cell fate. Dev. Cell *7*, 607–617.

Molaro, A., Hodges, E., Fang, F., Song, Q., McCombie, W.R., Hannon, G.J., and Smith, A.D. (2011). Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. Cell *146*, 1029–1041.

Nerlov, C., and Graf, T. (1998). PU.1 induces myeloid lineage commitment in multipotent hematopoietic progenitors. Genes Dev. *12*, 2403–2412.

Novershtern, N., Subramanian, A., Lawton, L.N., Mak, R.H., Haining, W.N., McConkey, M.E., Habib, N., Yosef, N., Chang, C.Y., Shay, T., et al. (2011). Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell *144*, 296–309.

Okada, Y., Yamagata, K., Hong, K., Wakayama, T., and Zhang, Y. (2010). A role for the elongator complex in zygotic paternal genome demethylation. Nature *463*, 554–558.

Ooi, S.K., and Bestor, T.H. (2008). The colorful history of active DNA demethylation. Cell *133*, 1145–1148.

Pekowska, A., Benoukraf, T., Ferrier, P., and Spicuglia, S. (2010). A unique H3K4me2 profile marks tissue-specific gene regulation. Genome Res. *20*, 1493–1502.

Popp, C., Dean, W., Feng, S., Cokus, S.J., Andrews, S., Pellegrini, M., Jacobsen, S.E., and Reik, W. (2010). Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. Nature *463*, 1101–1105.

Rai, K., Sarkar, S., Broadbent, T.J., Voas, M., Grossmann, K.F., Nadauld, L.D., Dehghanizadeh, S., Hagos, F.T., Li, Y., Toth, R.K., et al. (2010). DNA demethylase activity maintains intestinal cells in an undifferentiated state following loss of APC. Cell *142*, 930–942.

Sigvardsson, M., Clark, D.R., Fitzsimmons, D., Doyle, M., Akerblad, P., Breslin, T., Bilke, S., Li, R., Yeamans, C., Zhang, G., and Hagman, J. (2002). Early B-cell factor, E2A, and Pax-5 cooperate to activate the early B cell-specific mb-1 promoter. Mol. Cell. Biol. *22*, 8539–8551.

Smith, A.D., Chung, W.Y., Hodges, E., Kendall, J., Hannon, G., Hicks, J., Xuan, Z., and Zhang, M.Q. (2009). Updates to the RMAP short-read mapping software. Bioinformatics *25*, 2841–2842.

Tahiliani, M., Koh, K.P., Shen, Y., Pastor, W.A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L.M., Liu, D.R., Aravind, L., and Rao, A. (2009). Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science *324*, 930–935.

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.

Trowbridge, J.J., and Orkin, S.H. (2010). DNA methylation in adult stem cells: New insights into self-renewal. Epigenetics *5*, 189–193.

Trowbridge, J.J., Snow, J.W., Kim, J., and Orkin, S.H. (2009). DNA methyltransferase 1 is essential for and uniquely regulates hematopoietic stem and progenitor cells. Cell Stem Cell *5*, 442–449.

Wu, H., Caffo, B., Jaffee, H.A., Irizarry, R.A., and Feinberg, A.P. (2010). Redefining CpG islands using hidden Markov models. Biostatistics *11*, 499–514.

Zhang, H., Zhang, X., Clark, E., Mulcahey, M., Huang, S., and Shi, Y.G. (2010). TET1 is a DNA-binding protein that modulates DNA methylation and gene transcription via hydroxylation of 5-methylcytosine. Cell Res. *20*, 1390–1393.

Molecular Cell, *Volume 44*

Supplemental Information

Directional DNA Methylation Changes and Complex Intermediate States Accompany Lineage Specificity in the Adult Hematopoietic Compartment

Emily Hodges, Antoine Molaro, Camila O. Dos Santos, Pramod Thekkat, Qiang Song, Philip Uren, Jin Park, Jason Butler, Shahin Rafii, W. Richard McCombie, Andrew D. Smith, and Gregory J. Hannon

Contents

1 Supplemental Experimental Procedures

This supplement contains descriptions of many computational analysis methods used in the paper. A substantial portion of the methods have been used in Molaro et al. (2011) and therefore will not be described again here. Those methods include the basic pipeline for constructing the methylomes, the methodology for identifying HMRs, and the algorithm for measuring enrichment of one set of genomic intervals relative to another. We refer the reader to Molaro et al. (2011) for these details.

1.1 Hypomethylated regions

The method for identifying hypomethylated regions (HMRs) is described in Molaro et al. (2011) (in press). Briefly, HMRs were identified using a hidden Markov model (HMM) with 2 states: one for high methylation and one for low methylation. The data given to the HMM was the counts of methylated and unmethylated reads mapping over each CpG. These pairs of counts were modeled using a Beta-Binomial distribution. A cutoff for the minimum sum of posterior scores through an identified HMR was obtained by randomly permuting CpG sites and obtaining a size distribution for HMRs, and taking only those HMRs reaching the upper 1% of the scores obtained in the distribution obtained by randomization.

1.2 Differential methylation

Our general strategy for identifying differential methylation between two methylomes is to first calculate a differential methylation score for each individual CpG. Once the score has been calculated, we identify differentially methylated regions (DMRs) based on this score.

Our single-CpG differential methylation score is the probability that the CpG is methylated at a higher frequency in one methylome than the other. For CpG i, let m_i^a and u_i^a denote the number of methylated and unmethylated reads, respectively, in condition a. We assume p_i^a is the probability of methylation at CpG i in methylome a and that $p_i^a \sim$Beta(m_i^a, u_i^a). Given the observations of methylation at CpG i in conditions a and b, then we can use the exact formula for

$$\Pr(p_i^a > p_i^b) = f(m_i^a, u_i^a, m_i^b, u_i^b),$$

where the function f is as described by Altham (1969):

$$f(m_i^a, u_i^a, m_i^b, u_i^b) = \sum_{k=\max(m_i^b - u_i^a, 0)}^{m_i^b - 1} \frac{\binom{m_i^b + u_i^b - 1}{k} \binom{m_i^a + u_i^a - 1}{m_i^a + u_i^a - 1 - k}}{\binom{m_i^a + u_i^a + m_i^b + u_i^b - 2}{m_i^a + m_i^b - 1}}$$

This probability is symmetric and continuous, so for two methylomes a and b, we calculate either $\Pr(p_i^a > p_i^b) = 1 - \Pr(p_i^b > p_i^a)$.

Our algorithm for identifying differentially methylated regions consists of the following criteria:

- First the CpGs are partitioned into blocks such that no two consecutive CpGs is more than 500bp apart.
- Within each block, the differential methylation probabilities are smoothed using an Epanechnikov kernel with a bandwidth of 10 bases (not CpGs).
- After smoothing, we use a cutoff of 0.75 to identify peaks. For the opposite direction (since the scores are symmetric for the two methylomes being compared) we use 0.25.
- The set of candidate DMRs based on the cutoff are screened in two ways: (1) they must contain at least 10 CpGs, and (2) they must be at least 200bp in size.

We used randomization to indicate the cutoffs mentioned above. Briefly, the CpGs were randomly permuted within each block 100 times. For a given DMR size cutoff (both in terms of bases and CpGs) we obtain the number of DMRs identified, and compare this with the number identified in the real data. This provides a false discovery rate. We applied this method in several comparisons using an FDR of 0.05 to arrive at the cutoffs.

In Figure 3C we used differential methylation to correlate with differential expression. In this case we obtained the differential methylation probability $p_i(a, b) = \Pr(p_i^a > p_i^b)$ and then calculated the log-odds as $\log(p_i/(1 - p_i))$.

1.3 Measuring transcription factor binding site enrichment

In Figure 4B we measured transcription factor binding site (TFBS) enrichment inside the intergenic DMRs between neutrophils and B cells. The intergenic DMRs were selected for analysis using the following criteria:

1. Intergenic: residing at least 10Kbp from the nearest refGene.
2. Non-repeat: the DMRs must not overlap any annotated repeats, as downloaded through the UCSC Table Browser.
3. Non-CGI: the DMRs must not overlap an annotated CGI. The reason is that strong dinucleotide bias skews the motifs.

After applying these criteria, we ended with 1505 DMRs for N<B and 1175 for B<N. Before analyzing the sequences of these regions, they were expanded or contracted relative to their centers so that each sequence analyzed was 1Kbp in length.

The motif set used was a combination of known motifs from the JASPAR(Vlieghe et al., 2006) and TRANSFAC(Matys et al., 2006) databases. The total number of motifs in this data set was 775. For each motif we designated a family based on some shared binding property (*e.g.* motifs similar to CAGCTG were assigned the "EBOX2" class; motifs for all ETS family member TFs were designated "ETS").

We measured enrichment in the N<B DMRs relative to the B<N DMRs using the MOTIFCLASS program from the CREAD package(Smith et al., 2006). Briefly, MOTIFCLASS identifies the top scoring match

to each motif in each sequence from a foreground and background sequence set. We used the "binomial p-value" setting MOTIFCLASS, which for a given match score cutoff, calculates a p-value for enrichment in the foreground relative to the background from a binomial distribution. Let n be the total number of sequences with a match for the motif above a given cutoff, let k be the number of those from the foreground sequence set, and let p be the number of sequences in the foreground sequence set divided by the sum of the number of sequences in the foreground and the background. Then the p-value is for $\mathrm{Bin}(k; p, n)$. Finally, MOTIFCLASS optimizes the match score cutoff relative to this p-value. This procedure was applied once with N<B as foreground, and B<N as background; then the foreground and backgrounds were swapped and MOTIFCLASS was applied again. The p-values were not corrected for multiple hypothesis testing, as they were simply used to rank motifs; what is important in our results is the identity of the motifs landing in the top 20 for each DMR set, from among the 775 motifs evaluated in each case.

The method used to match motifs in sequences and identify the greatest match scores was described in (Hertz & Stormo, 1999) and implemented in the STORM program, also part of the CREAD package(Smith et al., 2006).

1.4 RNA-seq data processing

We used the RefSeq transcriptome as downloaded through the UCSC Table Browser (Karolchik et al., 2004). We mapped reads in two stages, first to sequences constructed from all RefSeq exons (with overlapping exons collapsed), and then to all possible junctions formed from all pairs of exons for the same gene. Mapping was done with RMAP (Smith et al., 2009) allowing up to 3 mismatches in 36 bases. Reads mapping ambiguously (including mapping to an exon and a junction) were discarded. For each RefSeq transcript we counted the number of reads whose mapping location was inside the transcript's exons (so a read can be counted for two transcripts, as long as the location is unique) or through one of the transcript's junctions. RPKM calculations discarded duplicate reads and corrected gene size for deadzones (portions of the transcripts to which no read can map uniquely).

Differential expression between two cells was computed using a 2×2 contingency table and chi-squared statistics or Fisher's exact test to obtain a p-value for differential expression. Briefly, the contingency tables contained, for each gene, the counts of reads inside the gene and outside the gene, for both cell types. We used Bonferroni correction for the p-values, and the remaining genes were called differentially expressed. These genes were then ranked based on RPKM ratios.

1.5 Cross species conservation

Cross species conservation was measured using phastCons scores as downloaded from the UCSC Genome Browser FTP server. These scores are posterior probabilities from a phylogenetic HMM (Siepel et al., 2005).

Supplemental References

Altham P (1969) Exact Bayesian analysis of a 2×2 contingency table, and Fisher's "exact" significance test. *Journal of the Royal Statistical Society. Series B (Methodological)* 31:261–269.

Hertz G, Stormo G (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 15:563–577.

Karolchik D, Hinrichs A, Furey T, Roskin K, Sugnet C, Haussler D, Kent W (2004) The UCSC Table Browser data retrieval tool. *Nucleic acids research* 32:D493.

Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E (2006) Transfac and its module transcompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 34:D108–10.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* 15:1034–1050.

Smith AD, Sumazin P, Xuan Z, Zhang MQ (2006) DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *Proc. Natl. Acad. Sci. USA* 103:6275–6280.

Vlieghe D, Sandelin A, De Bleser PJ, Vleminckx K, Wasserman WW, van Roy F, Lenhard B (2006) A new generation of jaspar, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res* 34:D95–7.

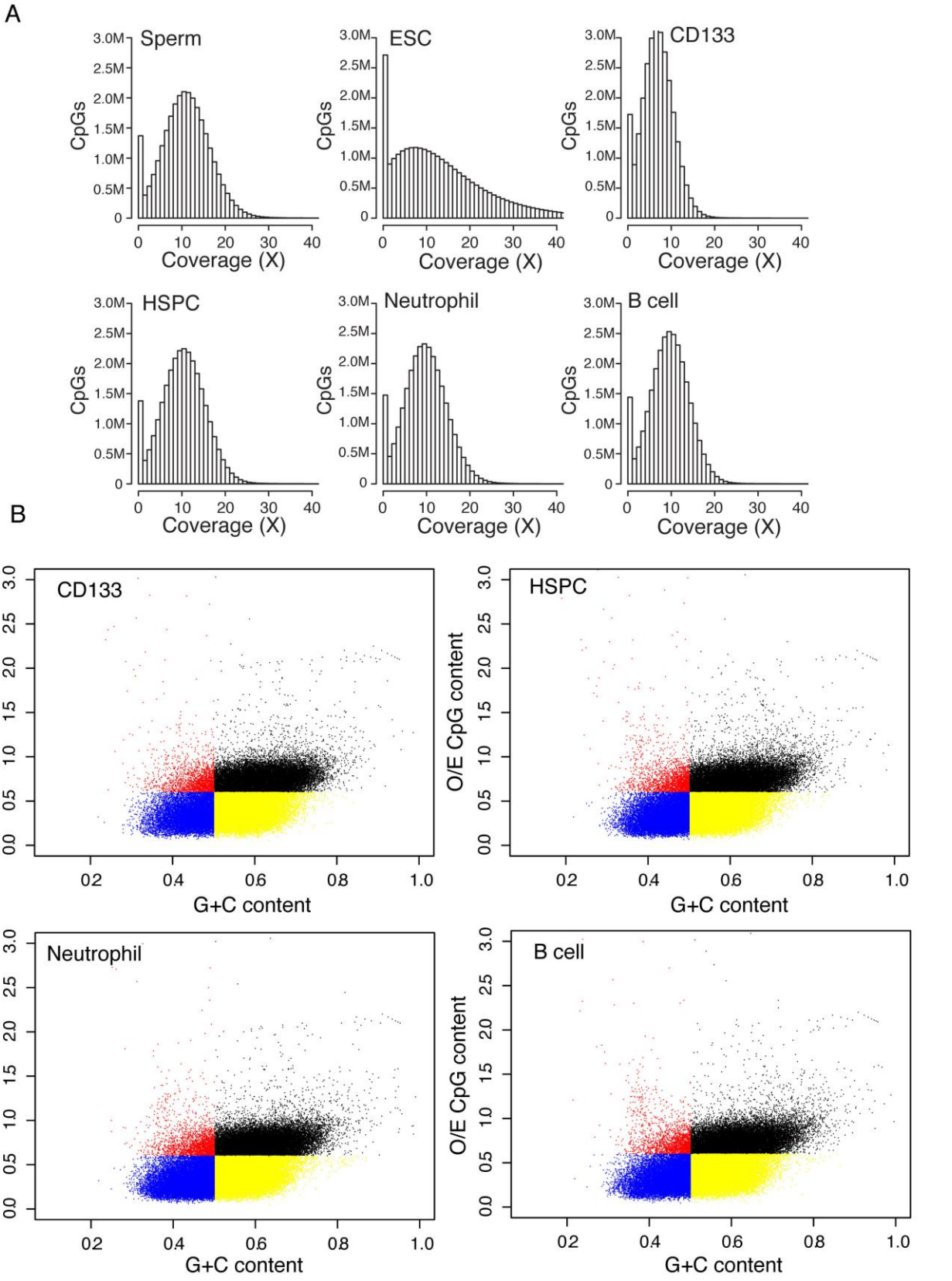

A
Sperm
ESC
CD133
HSPC
Neutrophil
B cell
CpGs
Coverage (X)
B
CD133
HSPC
Neutrophil
B cell
O/E CpG content
G+C content

Figure S1. CpG Mapping Coverage and HMR Characteristics, Related to Figure 1
(A) Distribution of read coverage for all CpG sites in the genome. Data is shown for human sperm, ESCs, CD133+ cord blood (stem cells), HSCs from peripheral blood, B cells and neutrophils. **(B)** G/C content and observed/expected CpG content for HMRs for each of the 6 cell types studied. Each HMR is a point, and colors indicate whether the HMR satisfies one or both of the sequence-based criteria described by Gardiner-Garden and Frommer and employed by UCSC to annotate CGIs genome-wide.

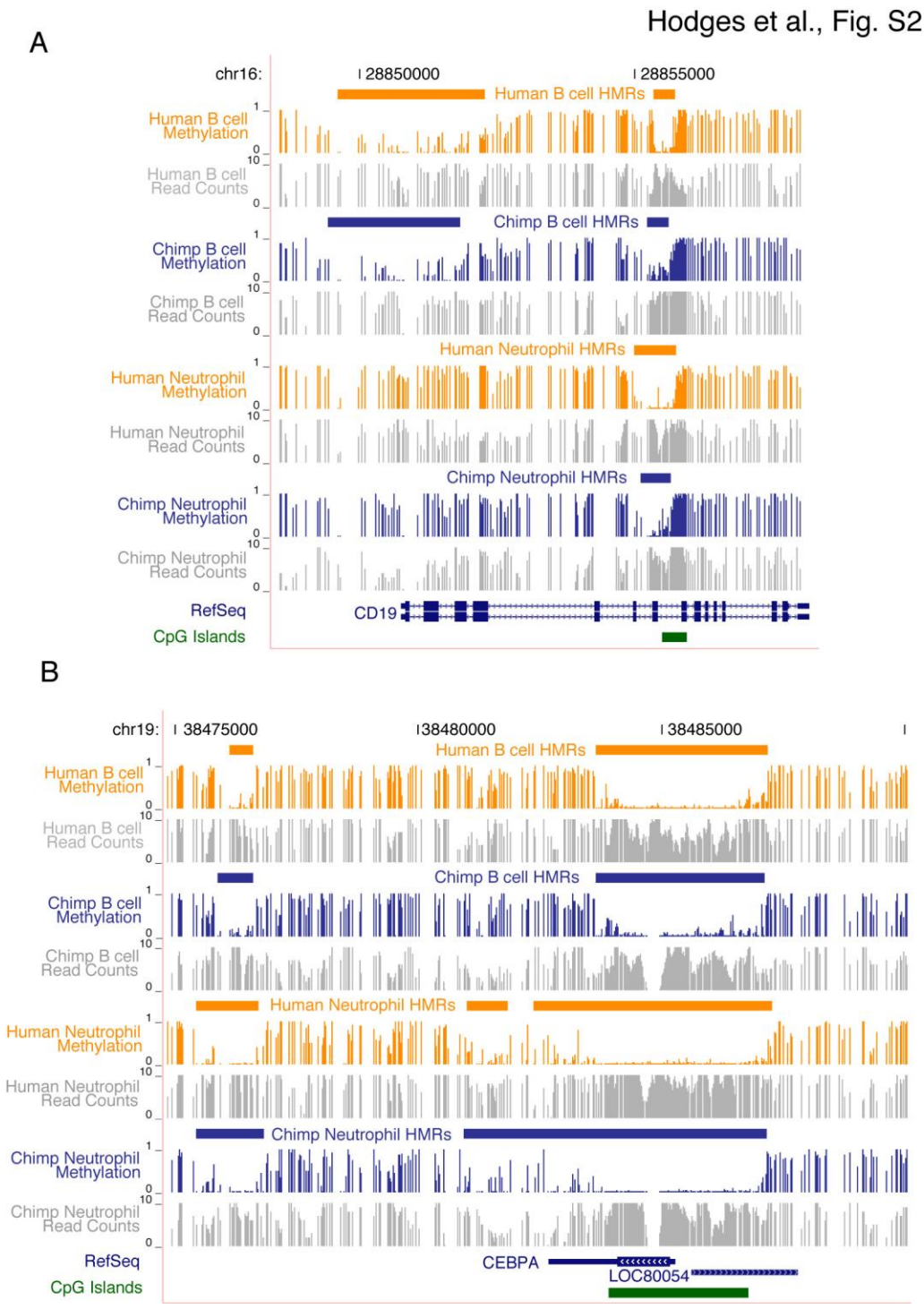


Figure S2. HMR Profiles Are Conserved between Chimp and Human, Related to Figure 1
Genome browser tracks depict methylation profiles across a lymphoid **(A)** and myeloid **(B)** specific locus in chimp and human blood cells. Methylation frequencies, ranging between 0 and 1, of unique reads covering individual CpG sites are shown in gray with identified hypomethylated regions (HMRs) indicated by orange bars. UCSC predicted/annotated CpG islands (green bars) as well as HMM-based CpG islands (blue bars) are also displayed. Numbers (top) indicate base position along the chromosome.

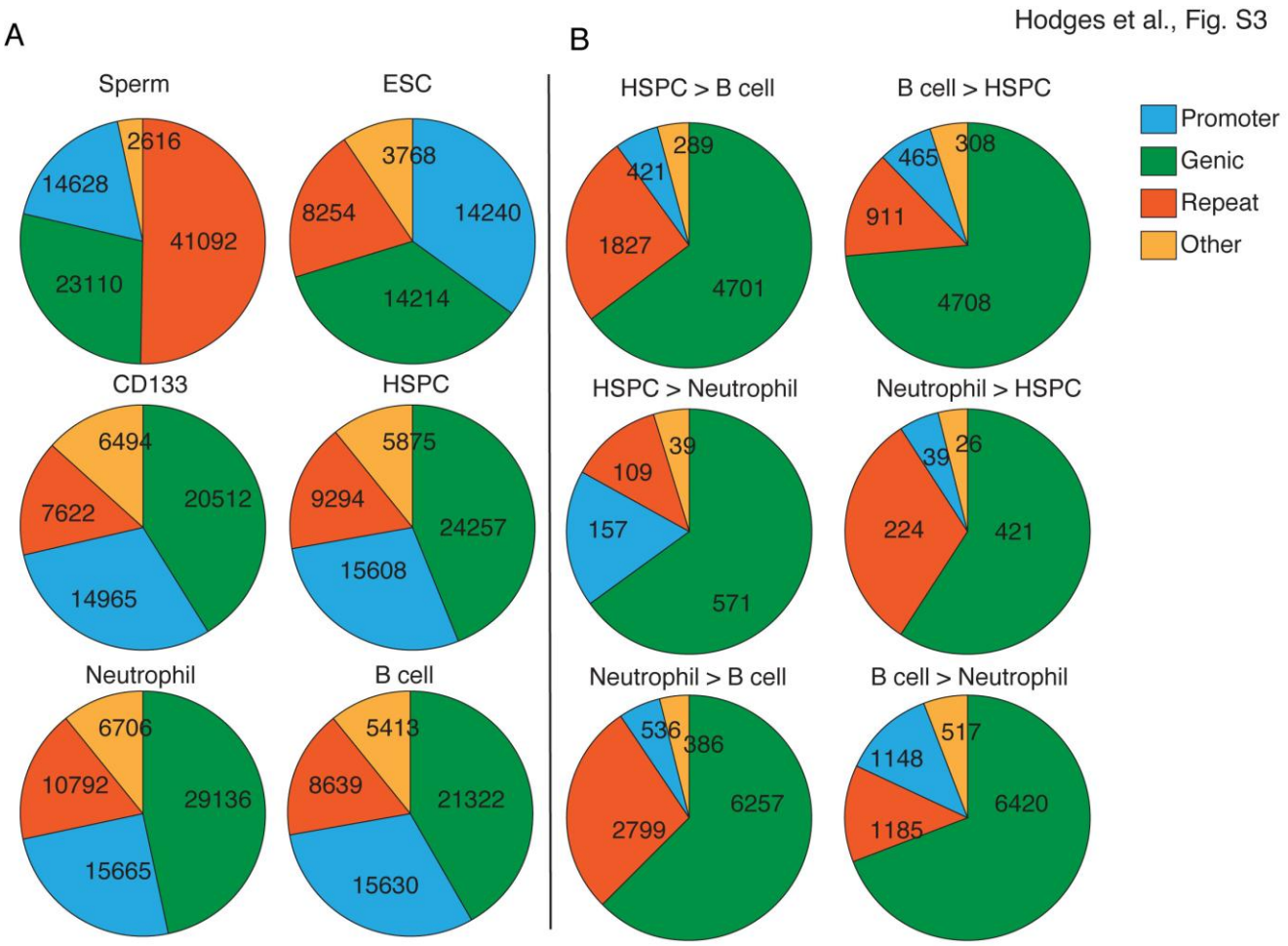
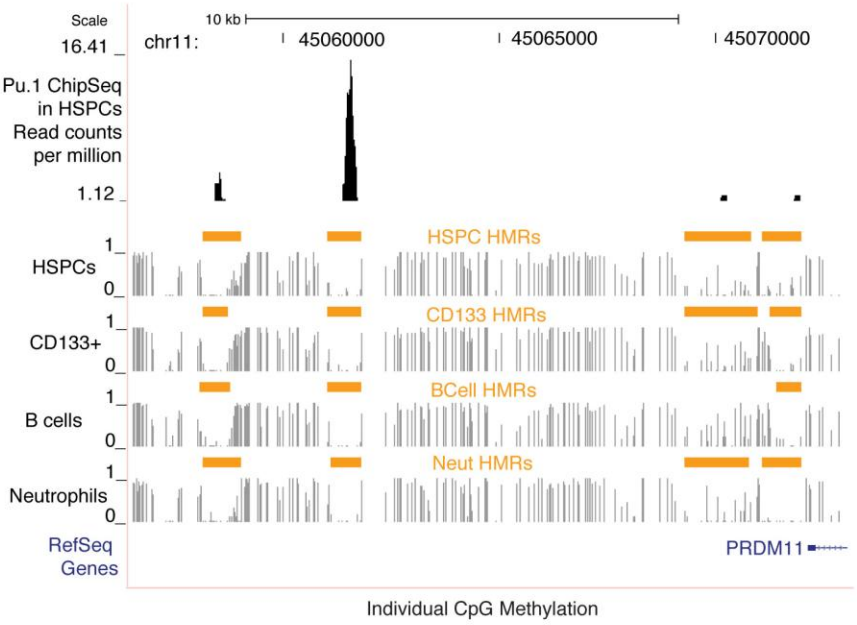


Figure S3. Distribution of HMRs and DMRs According to Genomic Annotations, Related to Figures 2, 4, and 5

The categories “promoter,” “genic,” “repeat” and “other” are exclusive, so first an HMR **(A)** or DMR **(B)** is checked for overlap with a promoter, the remainder are checked for overlap with a genic region, then the remainder are checked for overlap with annotated repeats (any class), and the “other” category is all those that remain.

A


Scale
10 kb
chr11:
45060000
45065000
45070000
16.41
Pu.1 ChipSeq
in HSPCs
Read counts
per million
1.12
HSPC HMRs
HSPCs
CD133 HMRs
CD133+
BCell HMRs
B cells
Neut HMRs
Neutrophils
RefSeq
Genes
PRDM11
Individual CpG Methylation

B

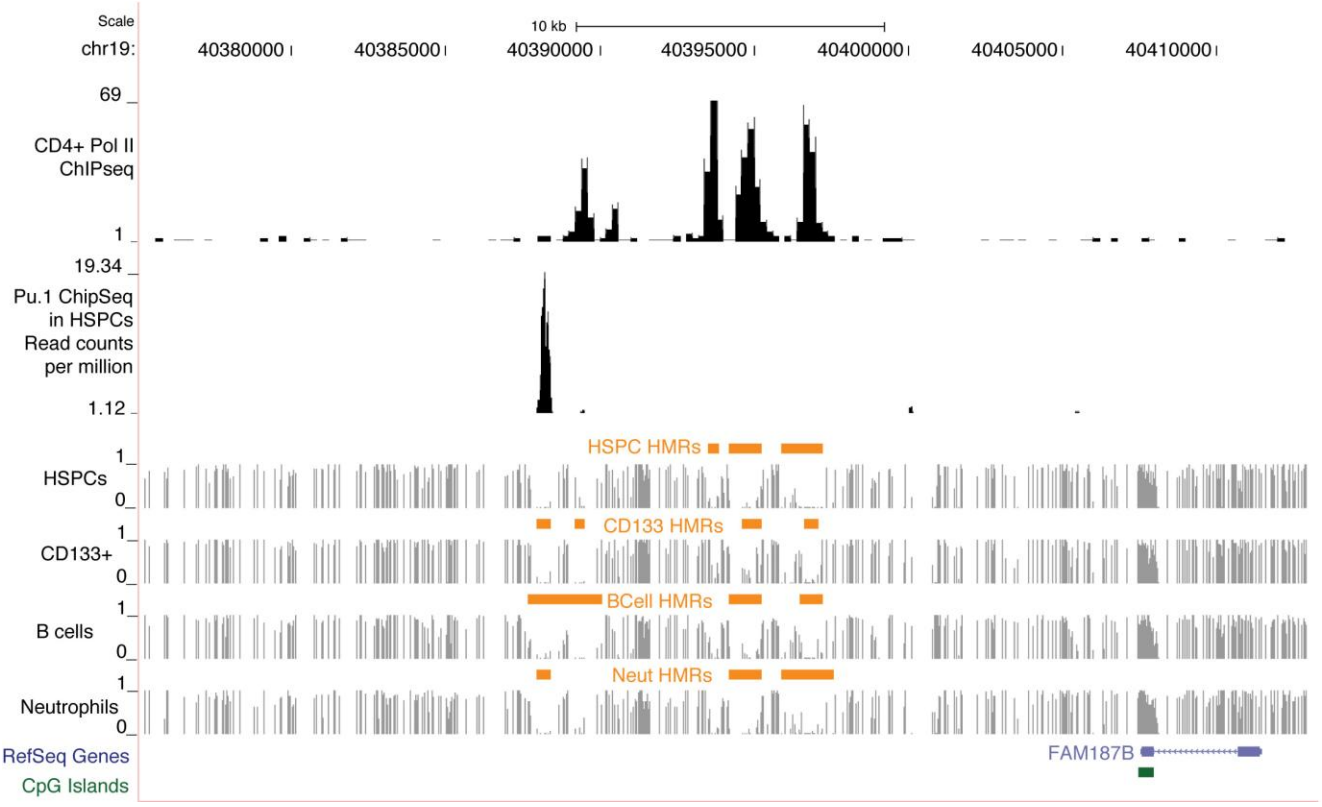

Scale
10 kb
chr19:
40380000
40385000
40390000
40395000
40400000
40405000
40410000
69
CD4+ Pol II
ChIPseq
1
19.34
Pu.1 ChipSeq
in HSPCs
Read counts
per million
1.12
HSPC HMRs
HSPCs
CD133 HMRs
CD133+
BCell HMRs
B cells
Neut HMRs
Neutrophils
RefSeq Genes
FAM187B
CpG Islands

Figure S4. PU.1 and RNA Polymerase II Enrichment in Intergenic HMRs, Related to Figure 4

Sequencing tracks of two loci **(A, B)** with ChIP-seq peaks derived from HSPCs enriched for PU.1 transcription factor (Novershtern et al., Cell. 2011 Jan 21;144(2):296-309) or RNA pol II overlapping intergenic HMRs. Peaks are displayed as read counts per million.

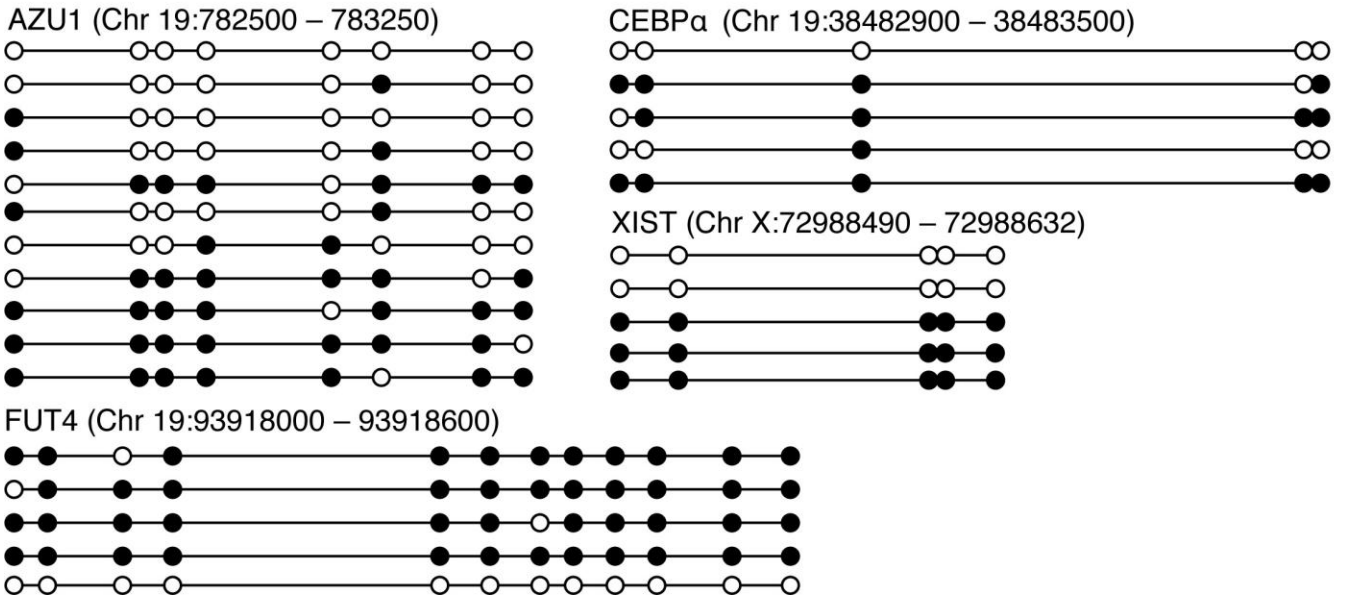


Figure S5. Bisulfite Sequencing of Clones Using Sanger Sequencing, Related to Figure 5

Lollipop diagrams show individual clones derived from HSCs across three myeloid specific genes and an allelicly methylated gene (see also Fig. 5).

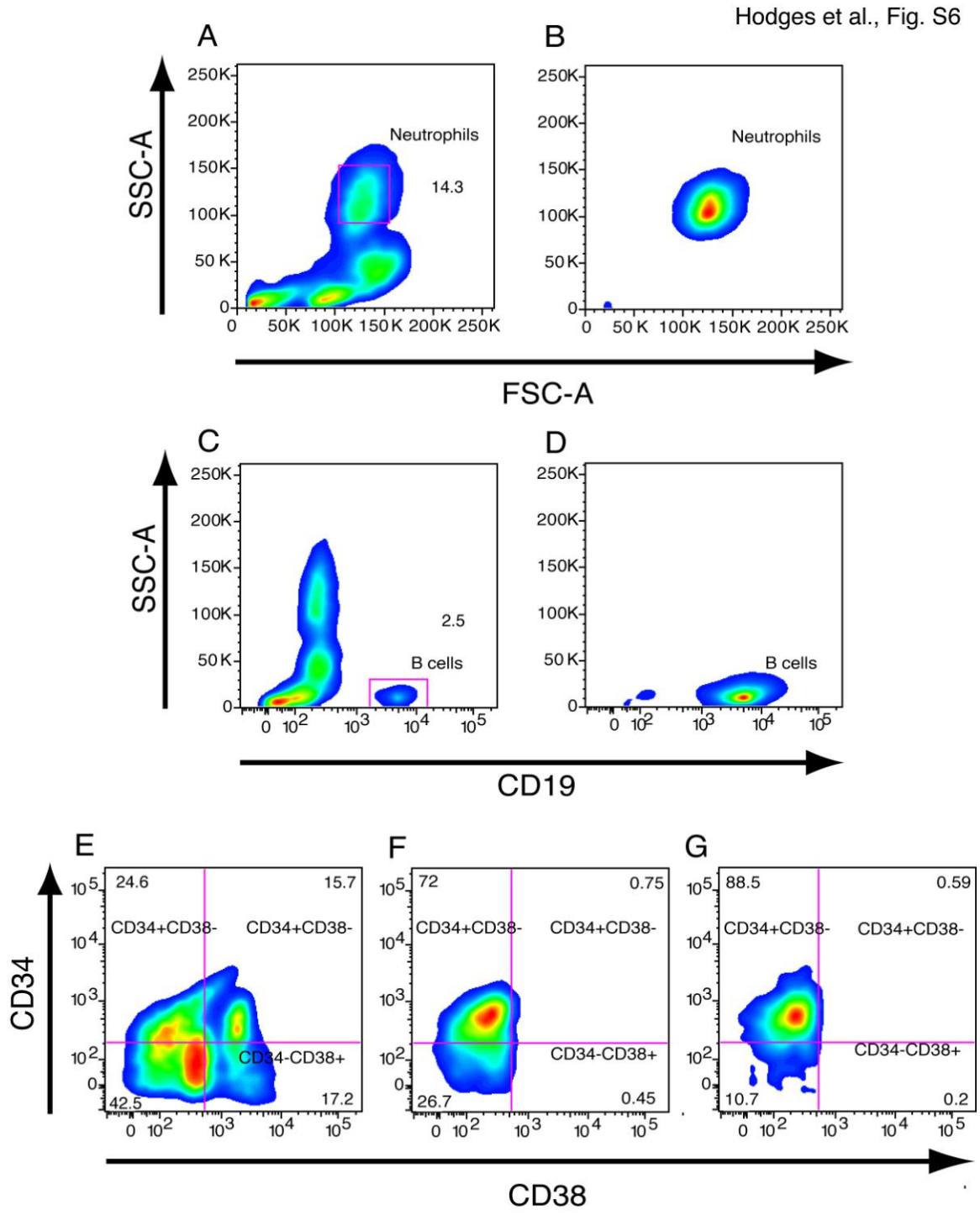


Figure S6. FACS profiles of purified blood cells, related to Figure 5. Peripheral blood mononuclear cells were purified according to the cell surface markers conjugated to the specified fluorophores. Displayed here are Neutrophils before **(A)** and after **(B)** sorting, B cells before **(C)** and **(D)** after sorting. Lineage depleted HSPCs **(E)** underwent two rounds of post-sorting to improve purity levels **(F, G)**.

Table S3. Primers Uses for Bisulfite PCR Cloning, Related to Figure 5

Chr19_CEBP Alpha

Forward – GGA AAG GGA GTT TTA GAT TTT TTT T
Reverse – CTA ACC TCT ATA CCC CAA CAA TAC CT

ChrX_XIST2

Forward – AAA AAG TGT AGA TAT TTT AGA GAG TGT AAT
Reverse – ACT TTA ATT TTT ATT TTT CTA ACC CAT C

Chr19_AZUI

Forward – GGG TTT GTG ATT TTT TAT GGA GTT
Reverse – CTT TAT TAC AAC CAA AAC CCC TCT A

Chr11_FUT4

Forward – GTG GTA TGG GTG GTG AGT TAT T
Reverse – CCA CTA TAT ACA AAA ACC CAA TTT C