DNA methylation in early mammalian development

by

Michelle M. Chan

B. Sc. Computer Science and Microbiology and Immunology
University of British Columbia, 2007

SUBMITTED TO THE PROGRAM OF COMPUTATIONAL AND SYSTEMS BIOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN COMPUTATIONAL AND SYSTEMS BIOLOGY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2013

Signature of Author:.....
Program in Computational and Systems Biology
May 15, 2013

Certified by:.....
Aviv Regev
Associate Professor of Biology
Thesis Supervisor

Accepted by:.....
Christopher B. Burge
Professor of Biology and Biological Engineering
Chair, Computational and Systems Biology Ph.D. Graduate Committee

DNA methylation in early mammalian development

by

Michelle M. Chan

Submitted to the Program in Computational and Systems Biology
on May 15, 2013 in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy in
Computational and Systems Biology

ABSTRACT

All the cells in the body contain the same genome yet showcase drastically different phenotypes. This is the result of different transcriptional programs, which are partly controlled by epigenetic modifications, including DNA methylation. In this thesis, I analyze genome-scale DNA methylation profiles across pre-implantation development to identify the targets and characterize the dynamics of global demethylation that lead to totipotency and the subsequent changes to embryonic specification. In Chapter 1, I validate and refine the decades old model for DNA methylation in mouse embryogenesis, identify many retrotransposons with active DNA methylation signatures at fertilization, and discover many, novel differentially methylated regions between the gametes that exist transiently during early development. Notably, the majority of epigenetic events unique to mammalian pre-implantation development are characterized in mouse. In Chapter 2, I describe the DNA methylation dynamics in human pre-implantation development and show that the regulatory principles that operate in mouse are conserved, though some of their targets are species-specific and define regions of local divergence. Finally, in Chapter 3, I compare DNA methylation dynamics of fertilization to an artificial reprogramming process, somatic cell nuclear transfer, in mouse, and find that most dynamics are conserved but occur at a smaller magnitude after artificial reprogramming. I conclude this thesis with a summary of the chapters and a brief discussion of ongoing and future work.

Thesis supervisor: Aviv Regev
Title: Associate Professor of Biology

Acknowledements

I gratefully acknowledge my thesis supervisor, Aviv Regev, my thesis advisory committee members, Chris Burge, and Mary Gehring, and co-authors for their contributions to this thesis, specifically Alexander Meissner and Zachary D. Smith. I would like to thank all past and present members of the Regev lab as well as members of the Meissner lab and the Burge lab. The computational and systems biology program has been wonderful and I would like to thank Alex, Tracy, Jonathan, Liraz, Albert, Yuanyuan, Naiyan, Emily, Jesse, Lawrence, and Danny, and also my other friends, Jackie, Ryan, Pritpaul, and Alan for making graduate school fun.

Table of Contents

List of Figures

List of Tables

Chapter 0

Introduction

Chapter 0:

Introduction

0.1 Overview

All the cells in the body contain the same genome yet showcase drastically different phenotypes. The epigenetic state of the cell determines its possible functional output by affecting cellular gene expression programs, which in turn, are responsible for overall phenotype[1-3]. Like gene expression patterns, the epigenome is dynamic – it changes between developmental phases and responds to environmental perturbations[4-6]. The extent to which epigenetic marks on the genome direct vs. reflect biological processes is an active area of research[6]. What is clear is that the epigenome has the ability to limit cellular potential through development though the exact mechanisms remain unknown. Conversely, which modifications must be changed or relocated to reprogram a cell to a different state? Moreover, it remains unknown whether a global reprogramming event is indiscriminate, erasing all modifications equally, or if some regions are targeted as part of ongoing regulation in the face of genome-wide erasure.

In this thesis, I analyze genome-scale DNA methylation profiles across pre-implantation development to identify the targets and characterize the dynamics of global demethylation and subsequent methylation that correspond to changes in cellular potency. I then compare the methylation changes that occur at fertilization to a process where totipotency is restored, specifically after reprogramming of the genome after somatic cell nuclear transfer. I begin this thesis with an overview of DNA methylation and its role in mammalian early development. I also introduce the reprogramming process, somatic cell nuclear transfer, and the DNA methylation profiling technology, reduced representation bisulfite sequencing. In Chapter 1, I

validate and refine the decades old model for DNA methylation in mouse embryogenesis at high resolution, identify many retrotransposons with active DNA methylation signatures at fertilization, and discover many, novel differentially methylated regions between the gametes that exist transiently during early development. Little is known about the global DNA methylation dynamics that occur over early human development. I show in Chapter 2 that global DNA methylation signatures are conserved between human and mouse but differences exist between the species specifically at loci, which show parentally-conferred methylation. The oocyte possesses the remarkable ability to reprogram the sperm genome upon fertilization, which is described in Chapter 1. In Chapter 3, I turn my attention to an artificial reprogramming process, somatic cell nuclear transfer, in mouse to compare DNA methylation dynamics in this process to its natural counterpart, fertilization and find that most dynamics are conserved but occur at a smaller magnitude after artificial reprogramming. I conclude this thesis in Chapter 4 with a summary of the chapters and a brief discussion of ongoing and future work.

0.2 DNA methylation distribution and mechanism

DNA methylation is a covalent modification restricted to cytosines largely in the context of CpG dinucleotides in mammals[2,5]. Non-CpG methylation, specifically in the nucleotide contexts CHG and CHH (where H = C, T, or A), is also common in plants and has been detected at low levels in some mammalian samples[8-10]. CpGs are not distributed equally across the mammalian genome and the CpG density of a region serves as a good predictor of methylation level[6]. CpGs present in high CpG density regions, representing ~1-2% of CpGs in mouse, are usually contained within CpG islands (CGIs) and are generally unmethylated while CpGs distributed elsewhere are largely methylated[3,6]. The relationship between CpG methylation and density may be evolutionarily coupled – methylated cytosines have increased mutability deaminating to

thymine which, if not repaired, will introduce mutations upon replication thus depleting CpG density where CpGs are methylated[11]. Most CGIs are found in gene promoters with the majority of the remainder existing within genic regions[12].

DNA methyltransferases (DNMTs) are responsible for establishing and maintaining CpG methylation. De novo methyltransferases add DNA methylation to the genome at specific developmental transitions while the maintenance methylatransferase, DNMT1, is responsible for mitotic inheritance of DNA methylation thereby conferring cellular memory[1]. The enzyme DNMT1 is targeted to hemi-methylated DNA by the co-factor UHRF1 (ubiquitin-like containing PHD and RING finger domains 1) during replication to ensure transmission of DNA patterns across cell divisions[13]. The de novo methyltransferases, DNMT3A and DNMT3B, are also required for maintenance methylation at some loci in certain cell types[14]. Targeting of the DNMT3s for de novo methylation is not yet well understood but interactions with PIWI-asscoiated RNAs (piRNAs) have been implicated in germ cells[15]. DNMT3L, a non-catalytic paralog of the DNMT3s, has been shown to interact with DNMT3A and unmethylated H3K4, and is required for proper establishment of imprint control regions and silencing of repeat elements[16].

Numerous demethylation mechanisms have been proposed for mammals and it is clear that some these mechanisms are not shared across different kingdoms. Active DNA demethylation in plants is carried out by DNA glycosylases, which catylze the removal of methylated cytosines through the base excision repair process[1,17]. Such glycosylases have not been identified in mammals but other pathways for demethylation have been proposed. The ten-eleven

translocation (TET) enzyme family oxidizes 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC)[18,19]. The role of 5hmC, a recently discovered epigenetic modification, is not yet defined but the prevailing hypothesis is that it may represent a transient intermediate during DNA demethylation[19,20]. In support of this hypothesis, 5hmC has been detected at low levels in mouse embryonic stem cells, and its levels have been observed to globally increase as levels of 5mC globally decrease after fertilization in the paternal genome[21-23]. To complete the demethylation process, 5hmC may be oxidized to 5-formylcytosine and 5-carboxylcytosine, which is then cleaved by the thymine DNA glycosylase (TDG) to produce unmethylated cytosine[18,19]. Alternatively, the catalytic conversion to 5hmC may destabilize the methylation maintenance machinery leading to passive methylation loss over replication[3]. This contrasts a fully catalytic demethylation process where cytidine deaminases may deaminate 5mC or 5hmC creating single basepair mismatches in DNA which are then recognized and repaired by the base-excision repair machinery[18,19].

0.3 DNA methylation functions

DNA methylation serves multiple critical functions in the cell and is traditionally considered a repressive mark. In this capacity, it is responsible for stably repressing promoters, silencing repeat regions, and establishing parental imprinting patterns. More recent studies have revealed potential, unexpected roles for DNA methylation in defining intragenic sequence features and splicing.

The correlation between DNA methylation and transcriptional repression implies the long held view that DNA methylation functions to regulate gene expression. Indeed, DNA methylation is important during cellular differentiation and lineage commitment partly due to its role in

silencing germline promoters and promoters on the inactive X chromosome[6,24]. However, even though DNA methylation is present at stably silenced promoters, it is rarely the initial directing modification, more often reinforcing repression that is initially established through chromatin modifications, such as H3K9 methylation which leads to heterochromatin formation[25-27]. DNA hypermethylation can occur passively *in cis*, through the loss of activating transcription factor binding, or can be activated *in trans* by the recruitment of transcriptional repressors[27-29]. H3K27 methylation is usually responsible for repressing CpG-island containing promoters, which are generally unmethylated *in vivo*[30,31]. Transcription factor binding as well as CXXC finger protein-1 (CFP1) binding, which leads to H3K4 methylation and consequently inhibits DNMT3 recruitment, help maintain the unmethylated status of CpG island promoters[29]. A further observation showing the weakness of DNA methylation as a regulatory repressor is that it is overcome by activating chromatin modifications at low CpG-density promoters[3].

DNA methylation silences repeat elements to maintain genomic integrity across cell divisions and generations[32]. Endogenous retrotransposons, namely long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and long terminal repeat (LTR) elements, make up ~40% of mammalian genomes and are constitutively methylated in somatic cells[33]. Many retrotransposons undergo demethylation during two major genome reprogramming events, germ cell and early pre-implantation development, where they show activity and must be targeted for repression[2,34]. piRNA-directed targeting has been suggested in germ cells but the mechanism during early development is unknown[15]. Studies in mouse embryonic stem cells (ESCs), where DNMT1 is insufficient to maintain methylation at LTRs across cell divisions, may provide some insight[35]. In ESCs, tripartite-motif-containing protein 28 (TRIM28) recruits

the H3K9 methyltransferase SETDB1 to LTRs through zinc finger proteins such as ZFP809, which is specific for proviral promoters[36,37]. Recruitment of DNMTs follows the repression of LTRs by H3K9 methylation though the mechanism remains unclear[38].

Imprinting control regions (ICRs) are portions of the genome that show monoallelic expression that is dependent on the parent-of-origin. DNA methylation is set during germ line development based on the gender of the individual, and withstands the global reprogramming that occurs in embryonic development persisting through many cell divisions in somatic tissues[2]. Similar to the mechanism for LTR silencing, TRIM28 is targeted to ICRs by a zinc finger protein, ZFP57[39]. DNMT3L is required for DNA methylation of imprints in the germ line as evidenced by the lack of imprints in female mice where both copies of DNMT3L have been disrupted[40]. Nevertheless, some maternal ICRs undergo stochastic methylation after fertilization suggesting that there may be epigenetic mechanisms that can rescue methylation defects[41]. There are approximately 20 well-characterized 'classic' ICRs covering ~100 genes in the mouse but recent work has identified many more loci in the brain[1,9]. It is unclear if these brain-specific ICRs are established through the same mechanism as classic ICRs since they have not been observed in the early embryo[42].

As more data about the distribution of DNA methylation across the genome is described, its sole function as a repressive epigenetic mark is being challenged. It has been known that while the promoters on the inactive X chromosome are hypermethylated, the global methylation present on the inactive X chromosome is less than on the active X chromosome suggesting a distinguishable difference between its local and chromosomal functions[43]. Analysis of whole genome bisulfite

sequencing data has revealed that DNA methylation may be a strong indicator of exons, and specifically exon-intron boundaries, even after accounting for the asymmetrical CpG density distributions inherent to exons and introns[44]. It is also enriched at splicing regulatory motifs and has been shown to distinguish between alternative and constitutive exons[45]. DNA methylation in splicing is an active field of research and insights into the possible mechanism of this novel, regulatory role of DNA methylation will be interesting to follow.

0.4 DNA methylation in early mammalian development

The DNA methylation landscape is relatively static across somatic tissues, following the canonical pattern of methylation that is dependent on CpG density[34]. The majority of methylated CpGs are pre-established and inherited through cell divisions[34]. Generally, only a small fraction of CpGs appears to switch their methylation levels as part of an orchestrated regulatory event[34]. The two exceptions are during primordial germ cell development, which will not be reviewed here, and pre-implantation development, where dramatic global demethylation of the genome occurs[2,3,34]. Studying DNA methylation during this unique phase provides an opportunity to understand how a genome is globally reprogrammed, identifying regions where DNA methylation is maintained, or removed and ultimately established again.

Pre-implantation development begins at fertilization, when the haploid sperm and oocyte fuse to form the zygote. The epigenetic states of the two gametes differ greatly[8,46,47]. The sperm genome is hypermethylated and mostly packaged in protamines, while the oocyte genome is less methylated and is arrested in metaphase II of meiosis[48,49]. In mouse, repeat elements retain the most methylation in oocytes, specifically intracisternal A particle (IAP) retrotransposons[50]. Shortly after fertilization, the paternal genome is repackaged with maternal histones forming the

paternal pronucleus and the oocyte completes meiosis, which gives rise to the second polar body[34]. The paternal pronucleus undergoes rapid genomewide conversion of 5mC to 5hmC in the zygote, while 5mC levels in the maternal pronucleus stay fairly constant[51,52]. The maternal-effect protein Stella has been shown to protect the maternal genome, as well as some paternally imprinted genes, from 5hmC conversion in the zygote[53]. Components of the replication elongator complex, specifically ELP3, have also been implicated in paternal genome demethylation[54]. 5hmC is passively depleted in a replication dependent manner, and 5mC may also be actively removed from the maternal genome, over the ensuing cleavage divisions resulting in a minimal methylation state in the blastocyst[21,47].

The blastocyst is the first stage of specification distinguishing between the trophectoderm and the inner cell mass (ICM)[52]. The ICM differentiates further to form the primitive endoderm and the epiblast[49]. Cells from the ICM and epiblast are pluripotent and both have been explanted to derive embryonic stem cell lines[55,56]. Unlike the hypomethylated ICM, the epiblast has a methylation profile similar to somatic tissues, its downstream lineages. Cells of the trohectoderm form extraembryonic tissues, such as the placenta, which is hypomethylated compared to somatic tissues[50,57]. After implantation, the embryo remethylates to somatic levels[34].

DNA methylation studies in early mouse development have been limited to either high resolution profiling by bisulfite sequencing of single loci or low resolution global immunofluorescence staining[22,47,51,52,58]. Consequently, the specific targets that either maintain methylation or are subject to demethylation remain unknown. Moreover, the regions affected by each phase of

paternal demethylation – the active demethyation upon fertilization, and a second passive demethylation across cleavage divisions – are also unknown? Finally, are there other parentally conferred imprinting control regions?

Eventhough a high resolution DNA methylation profile for mouse embryogenesis has yet to exist, it remains the best model for mammalian preimplantation development. The dynamics of DNA methylation across early human development is largely unexplored, presumably due to the scarcity of samples. There is some information detailing imprinting control region methylation in embryos, the transcriptional profile of oocytes and zygotes has been described, and the placental methylome was recently defined, but a systematic, global characterization of DNA methylation in normal embryos is absent[59-61]. A study of DNA methylation in human embryogenesis will not only produce a map of this epigenetic mark in an important organism but will also provide an opportunity to compare DNA methylation dynamics across two species to gain insight into the conservation of this process.

0.5 DNA methylation in somatic cell nuclear transfer

The ability to reprogram somatic cells to a pluripotent state has great potential in biomedical applications and basic research. Somatic cell nuclear transfer (SCNT), also known as nuclear cloning, demonstrates the incredible ability of the oocyte to reprogram a nucleus and remains the most effective method to convert somatic cells to a totipotent state[62]. In SCNT, a donor nucleus is injected into an enucleated oocyte[63]. The SCNT zygote is chemically activated and chromosomes assemble into two pseudo-pronuclei using mitotic spindles present in the ooplasm[64]. SCNT embryos are cultured *in vitro* to the blastocyst stage before transplantation into a surrogate female[63]. The efficiency of SCNT however is low - ~64% of SCNT embryos

develop to the blastocyst and <3% of transferred embryos develop to term[7,62,63]. One cause of the inefficiency is the incomplete reprogramming of epigenetic marks on the somatic genome[62,65]. Studies using immunofluorescence staining and bisulfite sequencing of selected loci show a severe lack of demethylation in the SCNT zygote as well as abnormal methylation profiles in several early developmental stages[47,66,67]. Inhibition of histone deacetylases increases the likelihood of full term development presumably through affecting chromatin modifications[62]. X-inactivation also poses a problem in SCNT zygotes[62,63]. The presence of Xist decreases X-linked expression in SCNT but when inhibited, development through implantation and post-implantation is increased greatly[62]. The epigenome after SCNT remains ill defined. Similar to mouse preimplantation development, there does not exist a genome wide high resolution DNA methylation profile for the SCNT process leaving the targets of demethylation unknown. A better understanding of the DNA methylation landscape of the somatic genome after reprogramming by SCNT may provide insight into the causes for the low efficiency of SCNT and may also reveal potential limitations of the oocyte's demethylation machinery.

0.6 Reduced representation bisulfite sequencing

There are several methods for measuring DNA methylation levels in the cell including immunofluorescence staining, methylation sensitive restriction digests, and bisulfite conversion of DNA. Immunofluorescence staining gives a global overview of DNA methylation levels but cannot provide locus specific information, while bisulfite conversion of DNA gives high-resolution information but typical studies only assay a small number of loci[47,51,52,58,67]. As genomic technology has become more common, these techniques have been modified to profile DNA methylation on a genome-wide scale. MeDIP-seq and MethylCap-seq utilize methylation specific antibodies or binding proteins to enrich for methylated regions prior to sequencing or

assaying on a microarray[68]. Bisulfite converted genomes are sequenced for single base pair resolution in whole genome bisulfite sequencing (WGBS)[69,70]. The technology used in this thesis is reduced representation bisulfite sequencing (RRBS) which will be the focus in this introduction[71].

As described earlier, CpG density is unevenly distributed in mammalian genomes. CpG-rich regions are generally classified as CpG islands (CGIs) and are mostly found in promoters[6]. For this reason, CGIs are particularly interesting because regulation of their methylation levels may have a direct impact on gene expression patterns in the cell. CpG-poor regions, on the other hand, are typically static across developmental transitions and are less frequently dynamic except as part of specific enhancer activation or repression[3]. RRBS uses an enzymatic digestion to enrich for CpG-rich regions, thereby focusing sequencing coverage on regions that may carry the most functional interest at a cost that facilitates the inclusion of many biological replicates that is not yet feasible using WGBS.

To make an RRBS library, genomic DNA is first isolated and subjected to an MspI digestion[72]. MspI is a restriction endonuclease that cuts at CCGG and thus, all end-sequenced fragments will contain CpG methylation information[72]. The 3'-terminal ends of the fragments are filled in, methylated adapter oligos are ligated, and the adapter-ligated DNA is run on an agarose gel for size selection[72]. For libraries using mouse samples, bands corresponding to ~28-300bp are excised, as fragments within this range cover most CGIs as predicted by digestion *in silico*[73]. The fragments then undergo either an extended round or two sequential rounds of bisulfite conversion to deaminate unmethylated cytosines before PCR amplification of the final

library[42,72]. Bisulfite treatment of DNA converts unmethylated cytosines to uracil, which is converted to thymine upon PCR amplification. Thus, bisulfite conversion creates a sequence disparity between methylated and unmethylated cytosines which can be used to estimate methylation levels. RRBS libraries are sequenced using Illumina and the resulting reads are aligned to a reference genome with software designed to handle bisulfite-converted libraries[42,72,74].

RRBS covers ~10% of the CpGs in mammalian genomes, including ~85% of CGIs[68]. It produces reproducible results with input as low as 0.5 ng of DNA and 1/10 of a HiSeq lane provides saturating coverage for one sample, though all samples used in this thesis were run in a single lane of an Illumina Genome Analyzer II[42]. RRBS can also be conducted in a 96 well format for high throughput genome-scale DNA methylation profiling[75,76]. In contrast, traditional WGBS requires at least 10-100ng of DNA and ~3 HiSeq lanes for adequate coverage, making the technology far more expensive than RRBS and less pragmatic for lower inputs[70,77]. Additionally, a high throughput protocol is not yet feasible given the extensive sequencing requirement. The obvious drawback of RRBS compared to WGBS is that it does not provide genome-wide information. Additionally, many reads start from the same position in the genome as a consequence of the enzymatic digestion making amplification artifacts more difficult to detect. Different methods for investigating amplification artifacts are included in the appendix of this thesis.

Bisulfite sequencing cannot distinguish between 5-methylcytosine and 5-hydroxymethylcytosine leaving both uncoverted after bisulfite treatment. An antibody specific to 5hmC exists and can

be used in immunofluorescence staining and enrichment sequencing methods[21,22,47]. Two protocols that convert 5hmC but not 5mC have been developed for single basepair resolution 5hmC profiling but both rely on a paired, traditional bisulfite sequenced library to identify regions where 5hmC exist[78,79]. These protocols have not yet been adapted to facilitate profiling of *in vivo* cell populations.

0.7 References

1 Feng, S., Jacobsen, S. E. & Reik, W. Epigenetic reprogramming in plant and animal development. *Science* **330**, 622-627, doi:10.1126/science.1190614 (2010).

2 Saitou, M., Kagiwada, S. & Kurimoto, K. Epigenetic reprogramming in mouse pre-implantation development and primordial germ cells. *Development* **139**, 15-31, doi:10.1242/dev.050849 (2012).

3 Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nature reviews. Genetics* **14**, 204-220, doi:10.1038/nrg3354 (2013).

4 Daxinger, L. & Whitelaw, E. Understanding transgenerational epigenetic inheritance via the gametes in mammals. *Nature reviews. Genetics* **13**, 153-162, doi:10.1038/nrg3188 (2012).

5 Feil, R. & Fraga, M. F. Epigenetics and the environment: emerging patterns and implications. *Nature reviews. Genetics* **13**, 97-109, doi:10.1038/nrg3142 (2011).

6 Suzuki, M. M. & Bird, A. DNA methylation landscapes: provocative insights from epigenomics. *Nature reviews. Genetics* **9**, 465-476, doi:10.1038/nrg2341 (2008).

7 Santos, F. *et al.* Epigenetic marking correlates with developmental potential in cloned bovine preimplantation embryos. *Current biology : CB* **13**, 1116-1121 (2003).

8 Tomizawa, S. *et al.* Dynamic stage-specific changes in imprinted differentially methylated regions during early mammalian development and prevalence of non-CpG methylation in oocytes. *Development* **138**, 811-820, doi:Doi 10.1242/Dev.061416 (2011).

9 Xie, W. *et al.* Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell* **148**, 816-831, doi:10.1016/j.cell.2011.12.035 (2012).

10 Ziller, M. J. *et al.* Genomic distribution and inter-sample variation of non-CpG methylation across human cell types. *PLoS genetics* **7**, e1002389, doi:10.1371/journal.pgen.1002389 (2011).

11 Cohen, N. M., Kenigsberg, E. & Tanay, A. Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. *Cell* **145**, 773-786, doi:10.1016/j.cell.2011.04.024 (2011).

12 Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes & development* **25**, 1010-1022, doi:10.1101/gad.2037511 (2011).

13 Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Reviews Genetics* **11**, 204-220, doi:Doi 10.1038/Nrg2719 (2010).

14 Chen, T., Ueda, Y., Dodge, J. E., Wang, Z. & Li, E. Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by Dnmt3a and Dnmt3b. *Molecular and cellular biology* **23**, 5594-5605 (2003).

15 Aravin, A. A. *et al.* A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. *Molecular cell* **31**, 785-799, doi:10.1016/j.molcel.2008.09.003 (2008).

16 Ooi, S. K. T. *et al.* DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature* **448**, 714-U713, doi:Doi 10.1038/Nature05987 (2007).

17 Gehring, M., Reik, W. & Henikoff, S. DNA demethylation by DNA repair. *Trends in genetics : TIG* **25**, 82-90, doi:10.1016/j.tig.2008.12.001 (2009).

18 Bhutani, N., Burns, D. M. & Blau, H. M. DNA demethylation dynamics. *Cell* **146**, 866-872, doi:10.1016/j.cell.2011.08.042 (2011).

19 Branco, M. R., Ficz, G. & Reik, W. Uncovering the role of 5-hydroxymethylcytosine in the epigenome. *Nature reviews. Genetics* **13**, 7-13, doi:10.1038/nrg3080 (2012).

20 Hackett, J. A. *et al.* Germline DNA demethylation dynamics and imprint erasure through 5-hydroxymethylcytosine. *Science* **339**, 448-452, doi:10.1126/science.1229277 (2013).

21 Inoue, A. & Zhang, Y. Replication-Dependent Loss of 5-Hydroxymethylcytosine in Mouse Preimplantation Embryos. *Science* **334**, 194-194, doi:Doi 10.1126/Science.1212483 (2011).

22 Iqbal, K., Jin, S. G., Pfeifer, G. P. & Szabo, P. E. Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. *P Natl Acad Sci USA* **108**, 3642-3647, doi:Doi 10.1073/Pnas.1014033108 (2011).

23 Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930-935, doi:10.1126/science.1170116 (2009).

24 Escamilla-Del-Arenal, M., da Rocha, S. T. & Heard, E. Evolutionary diversity and developmental regulation of X-chromosome inactivation. *Human genetics* **130**, 307-327, doi:10.1007/s00439-011-1029-2 (2011).

25 Ayyanathan, K. *et al.* Regulated recruitment of HP1 to a euchromatic gene induces mitotically heritable, epigenetic gene silencing: a mammalian cell culture model of gene variegation. *Genes & development* **17**, 1855-1869, doi:10.1101/gad.1102803 (2003).

26 Fuks, F., Hurd, P. J., Deplus, R. & Kouzarides, T. The DNA methyltransferases associate with HP1 and the SUV39H1 histone methyltransferase. *Nucleic acids research* **31**, 2305-2312 (2003).

27 Myant, K. *et al.* LSH and G9a/GLP complex are required for developmentally programmed DNA methylation. *Genome research* **21**, 83-94, doi:10.1101/gr.108498.110 (2011).

28 Brinkman, A. B. *et al.* Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. *Genome research* **22**, 1128-1138, doi:10.1101/gr.133728.111 (2012).

29 Thomson, J. P. *et al.* CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature* **464**, 1082-U1162, doi:Doi 10.1038/Nature08924 (2010).

30 Lindroth, A. M. *et al.* Antagonism between DNA and H3K27 methylation at the imprinted Rasgrf1 locus. *PLoS genetics* **4**, e1000145, doi:10.1371/journal.pgen.1000145 (2008).

31 Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553-560, doi:10.1038/nature06008 (2007).

32 Fedoroff, N. V. Presidential Address Transposable Elements, Epigenetics, and Genome Evolution. *Science* **338**, 758-767 (2012).

33 McCarthy, E. M. & McDonald, J. F. Long terminal repeat retrotransposons of Mus musculus. *Genome Biol* **5**, doi:Artn R14

Doi 10.1186/Gb-2004-5-3-R14 (2004).

34 Reik, W., Dean, W. & Walter, J. Epigenetic reprogramming in mammalian development. *Science* **293**, 1089-1093, doi:Doi 10.1126/Science.1063443 (2001).

35 Liang, G. *et al.* Cooperativity between DNA methyltransferases in the maintenance methylation of repetitive elements. *Molecular and cellular biology* **22**, 480-491 (2002).

36 Wolf, D. & Goff, S. P. TRIM28 mediates primer binding site-targeted silencing of murine leukemia virus in embryonic cells. *Cell* **131**, 46-57, doi:10.1016/j.cell.2007.07.026 (2007).

37 Wolf, D. & Goff, S. P. Embryonic stem cells use ZFP809 to silence retroviral DNAs. *Nature* **458**, 1201-1204, doi:10.1038/nature07844 (2009).

38 Leung, D. C. *et al.* Lysine methyltransferase G9a is required for de novo DNA methylation and the establishment, but not the maintenance, of proviral silencing. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 5718-5723, doi:10.1073/pnas.1014660108 (2011).

39 Li, X. J. *et al.* A Maternal-Zygotic Effect Gene, Zfp57, Maintains Both Maternal and Paternal Imprints. *Dev Cell* **15**, 547-557, doi:Doi 10.1016/J.Devcel.2008.08.014 (2008).

40 Bourc'his, D., Xu, G. L., Lin, C. S., Bollman, B. & Bestor, T. H. Dnmt3L and the establishment of maternal genomic imprints. *Science* **294**, 2536-2539, doi:10.1126/science.1065848 (2001).

41 Henckel, A. *et al.* Histone methylation is mechanistically linked to DNA methylation at imprinting control regions in mammals. *Human molecular genetics* **18**, 3375-3383, doi:10.1093/hmg/ddp277 (2009).

42 Smith, Z. D. *et al.* A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature* **484**, 339-344, doi:10.1038/nature10960 (2012).

43 Weber, M. *et al.* Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature genetics* **37**, 853-862, doi:10.1038/ng1598 (2005).

44 Gelfman, S., Cohen, N., Yearim, A. & Ast, G. DNA-methylation effect on cotranscriptional splicing is dependent on GC architecture of the exon-intron structure. *Genome research*, doi:10.1101/gr.143503.112 (2013).

45 Anastasiadou, C., Malousi, A., Maglaveras, N. & Kouidou, S. Human epigenome data reveal increased CpG methylation in alternatively spliced sites and putative exonic splicing enhancers. *DNA and cell biology* **30**, 267-275, doi:10.1089/dna.2010.1094 (2011).

46 Smallwood, S. A. *et al.* Dynamic CpG island methylation landscape in oocytes and preimplantation embryos. *Nature genetics* **43**, 811-814, doi:10.1038/ng.864 (2011).

47 Wossidlo, M. *et al.* Dynamic link of DNA demethylation, DNA strand breaks and repair in mouse zygotes. *Embo J* **29**, 1877-1888, doi:Doi 10.1038/Emboj.2010.80 (2010).

48 Egli, D., Birkhoff, G. & Eggan, K. Mediators of reprogramming: transcription factors and transitions through mitosis. *Nature reviews. Molecular cell biology* **9**, 505-516, doi:10.1038/nrm2439 (2008).

49 Hemberger, M., Dean, W. & Reik, W. Epigenetic dynamics of stem cells and cell lineage commitment: digging Waddington's canal. *Nat Rev Mol Cell Bio* **10**, 526-537, doi:Doi 10.1038/Nrm2727 (2009).

50 Popp, C. *et al.* Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. *Nature* **463**, 1101-U1126, doi:Doi 10.1038/Nature08829 (2010).

51 Oswald, J. *et al.* Active demethylation of the paternal genome in the mouse zygote. *Curr Biol* **10**, 475-478, doi:Doi 10.1016/S0960-9822(00)00448-6 (2000).

52 Santos, F., Hendrich, B., Reik, W. & Dean, W. Dynamic reprogramming of DNA methylation in the early mouse embryo. *Dev Biol* **241**, 172-182, doi:Doi 10.1006/Dbio.2001.0501 (2002).

53 Nakamura, T. *et al.* PGC7/Stella protects against DNA demethylation in early embryogenesis. *Nat Cell Biol* **9**, 64-U81, doi:Doi 10.1038/Ncb1519 (2007).

54 Okada, Y., Yamagata, K., Hong, K., Wakayama, T. & Zhang, Y. A role for the elongator complex in zygotic paternal genome demethylation. *Nature* **463**, 554-U177, doi:Doi 10.1038/Nature08732 (2010).

55 Evans, M. J. & Kaufman, M. H. Establishment in culture of pluripotential cells from mouse embryos. *Nature* **292**, 154-156 (1981).

56 Tesar, P. J. *et al.* New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature* **448**, 196-199, doi:10.1038/nature05972 (2007).

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

58 Mayer, W., Niveleau, A., Walter, J., Fundele, R. & Haaf, T. Embryogenesis - Demethylation of the zygotic paternal genome. *Nature* **403**, 501-502, doi:Doi 10.1038/35000656 (2000).

59 Huntriss, J. *et al.* Quantitative analysis of DNA methylation of imprinted genes in single human blastocysts by pyrosequencing. *Fertility and sterility* **95**, 2564-2567 e2561-2568, doi:10.1016/j.fertnstert.2011.04.035 (2011).

60 Noggle, S. *et al.* Human oocytes reprogram somatic cells to a pluripotent state. *Nature* **478**, 70-75, doi:10.1038/nature10397 (2011).

61 Schroeder, D. I. *et al.* The human placenta methylome. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 6037-6042, doi:10.1073/pnas.1215145110 (2013).

62 Narbonne, P., Miyamoto, K. & Gurdon, J. B. Reprogramming and development in nuclear transfer embryos and in interspecific systems. *Current opinion in genetics & development* **22**, 450-458, doi:10.1016/j.gde.2012.09.002 (2012).

63 Niemann, H., Tian, X. C., King, W. A. & Lee, R. S. Epigenetic reprogramming in embryonic and foetal development upon somatic cell nuclear transfer cloning. *Reproduction* **135**, 151-163, doi:10.1530/REP-07-0397 (2008).

64 Egli, D. *et al.* Reprogramming within hours following nuclear transfer into mouse but not human zygotes. *Nature communications* **2**, 488, doi:10.1038/ncomms1503 (2011).

65 Rideout, W. M., Eggan, K. & Jaenisch, R. Nuclear cloning and epigenetic reprogramming of the genome. *Science* **293**, 1093-1098, doi:Doi 10.1126/Science.1063206 (2001).

66 Bourc'his, D. *et al.* Delayed and incomplete reprogramming of chromosome methylation patterns in bovine cloned embryos. *Curr Biol* **11**, 1542-1546, doi:Doi 10.1016/S0960-9822(01)00480-8 (2001).

67 Dean, W. *et al.* Conservation of methylation reprogramming in mammalian development: Aberrant reprogramming in cloned embryos. *P Natl Acad Sci USA* **98**, 13734-13738, doi:Doi 10.1073/Pnas.241522698 (2001).

68 Bock, C. *et al.* Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nature biotechnology* **28**, 1106-1114, doi:10.1038/nbt.1681 (2010).

69 Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133**, 523-536, doi:10.1016/j.cell.2008.03.029 (2008).

70 Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315-322, doi:10.1038/nature08514 (2009).

71 Meissner, A. *et al.* Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic acids research* **33**, 5868-5877, doi:10.1093/nar/gki901 (2005).

72 Gu, H. *et al.* Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nature protocols* **6**, 468-481, doi:10.1038/nprot.2010.190 (2011).

73 Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766-770, doi:10.1038/nature07107 (2008).

74 Bock, C. Analysing and interpreting DNA methylation data. *Nature reviews. Genetics* **13**, 705-719, doi:10.1038/nrg3273 (2012).

75 Boyle, P. *et al.* Gel-free multiplexed reduced representation bisulfite sequencing for large-scale DNA methylation profiling. *Genome biology* **13**, R92, doi:10.1186/gb-2012-13-10-r92 (2012).

76 Smith, Z. D., Gu, H., Bock, C., Gnirke, A. & Meissner, A. High-throughput bisulfite sequencing in mammalian genomes. *Methods* **48**, 226-232, doi:10.1016/j.ymeth.2009.05.003 (2009).

77 Adey, A. & Shendure, J. Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing. *Genome research* **22**, 1139-1143, doi:10.1101/gr.136242.111 (2012).

78 Booth, M. J. *et al.* Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **336**, 934-937, doi:10.1126/science.1220671 (2012).

79 Yu, M. *et al.* Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149**, 1368-1380, doi:10.1016/j.cell.2012.04.027 (2012).

Chapter 1

A unique regulatory phase of DNA methylation in the early mouse embryo

This work was first published as:

Zachary D. Smith*, Michelle M. Chan*, Tarjei S. Mikkelsen, Hongcang Gu, Andreas Gnirke, Aviv Regev and Alexander Meissner. A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature* **484**(7394):339-44 Mar 2012. *These authors contributed equally to this work.

ZDS and AM conceived the study and ZDS, MMC, AM facilitated its design. ZDS collected samples and performed methylation profiling, MMC performed all analysis with assistance from ZDS and TSM. HG and AG provided critical technical assistance and expertise. ZDS, MMC, TSM, AR and AM interpreted the data. ZDS, MMC, and AM wrote the paper with the assistance of the other authors.

Chapter 1:

A unique regulatory phase of DNA methylation in the early mouse embryo

1.1 Abstract

DNA methylation is highly dynamic during mammalian embryogenesis. It is broadly accepted that the paternal genome is actively depleted of global cytosine methylation at fertilization, followed by passive depletion that reaches a minimum at the blastocyst stage. However, this model is based on limited data, and to date no base-resolution maps exist to support and refine it. Here, we generated genome-scale DNA methylation maps in mouse gametes and through post-implantation embryogenesis. We find that the oocyte already exhibits global hypomethylation, most prominently at specific families of LINE-1 and LTR-containing retro-elements, which are disparate between gametes and resolve to lower methylation values in zygote. Surprisingly, the oocyte contributes a unique set of Differentially Methylated Regions (DMRs), including many CpG Island promoter regions, that are maintained in the early embryo but are lost at the onset of embryonic specification and absent in somatic cells. In contrast, sperm contributed methylation includes retrotransposons that become completely methylated after the blastocyst stage. Our data provide a complete genome-scale, base-resolution timeline of DNA methylation in the pre-specified embryo, when this epigenetic modification is most dynamic and before returning to the canonical somatic pattern.

1.2 Introduction

Cytosine methylation in mammals is an epigenetic modification that is largely restricted to CpG dinucleotides and serves multiple critical functions including stable repression of target promoters, maintaining genomic integrity, establishing parent-specific imprinting patterns, and silencing endogenous retrotransposon activity[1,2]. In somatic tissues, CpG methylation exhibits global patterns based on relative CpG density: CpG islands at housekeeping or developmental promoters are largely unmethylated while non-regulatory CpGs distributed elsewhere in the genome are largely methylated[1,3]. This landscape is relatively static across all somatic tissues that have been examined to date where the majority of methylated CpGs are pre-established and inherited through cell divisions[4-6]. Generally, only a small fraction of CpGs appear to switch their methylation levels as part of an orchestrated regulatory event[7].

By contrast, DNA methylation is much more dynamic during mouse germ-cell and pre-implantation development. The classical model for DNA methylation during pre-implantation postulates that at fertilization, DNA methylation contributed by the paternal gamete must be "actively" removed by a targeted, though widespread, catalytic process. This erasure is believed to occur across the genome shortly after fertilization. Recent evidence implicates a demethylation mechanism that transitions through a hydroxymethylated (hmC) intermediate that is catalyzed by the Tet3 member of the Tet family[8,9]. However, only a portion of hydroxylated targets appears to be actively catalyzed to complete demethylation, and the identity of these targets remains unknown[10]. Following this dramatic change in the zygote, there appears to be a passive loss of global DNA methylation levels that continues until the blastocyst stage, where the inner cell mass (ICM) that gives rise to the embryo proper is first specified (reviewed by Ref.

11). Again, recent evidence suggests this passive depletion may be facilitated in part by targeted hydroxylation mediated by Tet enzymes, as the majority of accumulating hmC signal within the paternal genome is depleted in a division-dependent manner[10]. After specification of the ICM, the embryo implants into the uterine lining in concert with gastrulation, which is accompanied by a global remethylation of the genome that is believed to contribute to lineage restriction and loss of cellular potency[12,13].

Unfortunately, little is known on a quantitative, genome-wide scale about the specific dynamics of CpG methylation during these earliest developmental stages[14]. The classical model above is drawn from observations made using either global measurements, such as immunohistochemistry, or from limited analysis of one or several CpGs at individual loci using sodium bisulfite treatment, PCR, and clonal sequencing[11,12,15-22]. Key questions about DNA methylation patterns in early development remain open, including which genomic features are specifically targeted for demethylation as well as the identities of Differentially Methylated Regions (DMRs) inherited from either gamete beyond known Imprint Control Regions (ICRs). New genomic, high-resolution methylation profiling strategies[23,24] could contribute insight into the underlying mechanisms and regulatory principles of CpG methylation as it functions in early mammalian development.

1.3 Results

1.3.1 High quality genome-scale methylation maps of murine embryogenesis

To generate a global and high-resolution view of early mammalian DNA methylation dynamics, we collected oocytes and sperm, as well as zygote, 2-, 4-, and 8- cell cleavage stage embryos, the inner cell mass (ICM) and E6.5/7.5 post-implantation embryos (**Figs. 1-1a, 1-2** and **1-3**). All samples were extensively washed and purified to remove any somatic or meiotic contaminants. Potential maternal biasing from Meiosis I and II polar bodies (representing a 1x or 0.5x static genomic contaminant, respectively) was excluded by manually removing the polar bodies (**Fig. 1-2**) and assessing the paternal (129X1) to maternal (C57/B6xDBA) ratio of sequenced single nucleotide polymorphisms (SNPs) (**Figs. 1-4** and **1-5**). We generated reduced representation bisulfite sequencing (RRBS)[4] libraries from each stage, which were combined to provide a comprehensive timeline of genomic DNA methylation patterns during early mouse embryogenesis.

Compared to all other genome-wide profiling strategies currently available, RRBS is optimally suited for the low cell numbers that can be obtained from the embryonic stages in our study[23,24]. Within our range of 0.5-10ng genomic DNA, RRBS provides high sensitivity and reproducibility, and the expected genomic coverage (**Fig. 1-3**). On average, we obtained the methylation status of 953,606 CpGs for comparative analysis across our pre-implantation timeline. Different genomic features were equally well captured in libraries from the different stages (**Fig. 1-3**). Unfortunately, bisulfite sequencing cannot distinguish between methyl- and hydroxymethyl- cytosine (hmC), and current methods for global profiling of hmC lack the required sensitivity to investigate the pre-implantation stages in this study[9,25-30]; therefore, we

cannot make any definitive statements regarding base resolution hmC distribution during the pre-implanation stages. However, while hmC shows some unique distribution patterns in mES lines and brain tissue, it has not yet been linked to a regulatory mechanism other than to potentiate demethylation[31]. Given this ambiguity, regions of high methylation, especially those retained over multiple time points, could still be expected to function as if methylated.

1.3.2 Global CpG methylation in the early embryo does not resemble somatic patterns

The current model postulates a phase of global hypomethylation during mammalian pre-implantation development that reaches a minimum at the morula/blastocyst stage. However, this model does not specify which genomic regions are affected or whether the methylation patterns in this phase still follow the general correlation with sequence context that have been established for somatic cells[1]. To address these questions, we investigated the global dynamics of CpG methylation using 100bp tiles. We calculated the methylation level of each tile by averaging over all CpGs within it that were covered with at least 5 reads. Intriguingly, we found that oocytes are already dramatically and globally hypomethylated compared to sperm (0.32 median methylation in oocyte versus 0.85 in sperm for 100bp tiles, **Fig. 1-6**). We next examined the relative proportion of genomic regions at each stage falling into high ($\geq$0.8), intermediate (>0.2 and <0.8) or low ($\leq$0.2) methylated categories. Notably, oocyte methylation levels more closely resembled those of early embryonic time points than the levels in sperm, post-implantation embryos, or adult tissues (**Fig. 1-1b**). We also observed a gradual increase in the fraction of tiles that exhibit intermediate and low methylation values from oocytes to the ICM, which is consistent with loss of methylation over multiple cleavage divisions (**Fig. 1-1b**).

We observed a dramatic difference in CpG methylation levels as a function of both CpG density and developmental stage. Sperm and post-implantation embryos displayed the same strong inverse relationship between CpG density and methylation levels that is present in somatic cells. In oocyte and pre-implantation samples, this dependence was much weaker (**Fig. 1-1c,d**). More specifically, we examined the relative CpG density in 'methylated tiles' (>0.2 methylation, **Fig. 1-1e,** left) and 'hypomethylated tiles' (≤0.2 methylation, **Fig. 1-1e**, right). In both pre- and post-implantation embryos, methylated CpGs tend to be in tiles with low CpG density, thereby recapitulating the somatic pattern (**Fig. 1-1e,** left). In contrast, the average CpG densities in hypomethylated tiles were lower in pre-implantation embryos and oocytes than they were in sperm, post-implantation embryos or adult tissues (**Fig. 1-1e**, right). This suggests that CpG density-dependent methylation is primarily a post-implantation and adult pattern. In summary, pre-implantation development represents a unique developmental period where methylation is differentially positioned and regulated before being restored and maintained in a somatic fashion.

1.3.3 Two major transitions in methylation levels during early development

We next searched for substantial changes in regional DNA methylation and pinpointed the stage at which they occurred. For each pair of consecutive stages, we compared methylation levels of each 100bp tile and classified it as changed if the difference exceeded 0.2 and was significant according to a FDR-corrected t-test (**Fig. 1-7a**).

The most dramatic changes in DNA methylation occurred during two developmental transitions: between sperm and the zygote and between the ICM and the post-implantation embryo (**Fig. 1-7a**). At each of these transitions, the majority of changes were unidirectional (**Fig. 1-7b**): a gross

depletion of sperm hypermethylation upon fertilization (mean change = 0.48 decrease for 22% of regions examined) and massive remethylation from the ICM to post-implantation embryos (mean = 0.46 increase in methylation at 65% of tiles). Within our two post-implantation timepoints (E6.5 and E7.5), the methylation levels at the majority of assayed tiles were stable or increased only slightly (**Fig. 1-7b**). In addition, more subtle global changes, reflecting a gradual decrease in methylation, was observed from zygote/early cleavage through the 8-cell stage and into the ICM, where methylation levels reached their lowest observed values (**Fig. 1-1b,c**).

1.3.4 The oocyte defines the early methylation landscape

Active demethylation is expected to occur prior to pronuclear fusion or DNA synthesis and is generally described as complete after ~6 hours post fertilization[11,32]. Indeed, when we compare methylation patterns between sperm and zygote, the majority of regions in the genome show reduced methylation in the zygote with few additional changes from zygote to the 2-cell stage (**Fig. 1-7b**). Interestingly, the vast majority of tiles that are methylated at significantly different levels between gametes show higher methylation levels in sperm than in oocyte. In the zygote, their methylation is reduced to levels at or near those of the oocyte (**Fig. 1-7c, d**). Specifically, of 106,081 tiles that are significantly different between sperm and oocytes, 102,862 have higher methylation in sperm, but their levels more closely resemble oocyte values in the zygote stage. We confirmed this observation by tracking ~65 CpGs that fell within these tiles and could be assigned paternal or maternal specific values. Zygotes displayed a decrease in paternal methylation in contrast to maternally contributed CpGs, which remained unmethylated (**Fig. 1-7e**). It is worth noting that the isolated zygotes used here are likely in earlier stages of zygotic S phase, such that either a passive, replicative based mechanism could result in the synthesis of

unmethylated, nascent DNA or DNA methylation could be decreasing as a function of targeted catalysis[10,33,34]. The similarities in methylation levels between zygote and 2-cell, which has completed one full round of replication but not yet initiated another, argues that at least some of the measured decreases in methylation are a consequence of targeted removal, but distinguishing between these two models may be complicated by coupling of proposed Base-Excision repair mechanisms and DNA replication itself[34].

In contrast, the few regions (3,219 of 106,081) that are significantly hypermethylated in oocyte compared to sperm instead exhibit intermediate values in the zygote in a manner that suggests a more direct inheritance of the allelic methylation state (**Fig. 1-7d**). It is worth noting that developing PGCs and post-natal, maturing oocytes exhibit global, genome-wide hypomethylation followed by targeted remethylation, including that of Imprint Control Regions (ICRs)[35,36]. The disparity in the zygotic resolution of regions that are differentially methylated between the gametes indicates that the oocyte largely reflects the zygotic/pre-implantation methylome and describes its architecture (**Fig. 1-8**). Thus, the oocyte methylome, rather than the sperm methylome, appears to be more reflective of patterns in the early embryo.

1.3.5 Loss of methylation at fertilization is most prominent at specific repeat classes that are paternally hypermethylated

Consistent with a demethylation model, we confirmed that the vast majority (92%) of tiles that are hypermethylated in sperm in our data set become less methylated in the zygote. Moreover, the majority of this set already exhibits lower methylation in the oocyte, such that additive effects could also explain more subtle decreases in many regions. Interestingly, tiles exhibiting

the most extreme methylation changes during the sperm to zygote transition are enriched for Long Interspersed Elements (LINEs) ($P<4.7x10^{-184}$, FDR<0.05, hypergeometric enrichment) **(Fig. 1-9a)**. We therefore directly estimated the methylation level for individual LINEs surveyed by RRBS at each stage. We found that the degree of methylation changes in these elements is markedly bimodal during the sperm to zygote transition, with 18% of LINEs reducing their methylation values by over 0.45 **(Fig. 1-9a)**. By comparison, 12% of captured Long Terminal Retroelements (LTRs) exhibited similar levels of demethylation, but the distribution was not as clearly bimodal **(Fig. 1-9b)**. Short Interspersed Elements (SINEs) are generally less methylated in sperm than other repeat classes, as has been noted in human data[37], and exhibit dramatic shifts in their methylation values from sperm to early embryo as well, but again without the apparent bimodality observed for LINE elements **(Fig. 1-10)**.

Surprisingly, the LINEs that changed most dramatically during the sperm to zygote transition consisted nearly exclusively of two closely related families of L1 LINEs, notably L1Md_T and L1Md_Gf **(Fig. 1-9c,d,** $P<4.7x10^{-184}$; hypergeometric enrichment test)[38,39]. Repeats from these families had the strongest and most consistent decrease from hypermethylated values, while those from other equally represented families, such as L1Md_A elements, showed smaller changes in their methylation values upon fertilization and maintained higher methylation values in both oocyte and zygote **(Fig. 1-9e, Supplementary Fig. 8)**. Similarly, several LTR families exhibited discrete loss of methylation within the zygote **(Fig. 1-9f,g)**, while the Class II Intracisternal A-particles (IAPs, **Fig. 1-9h)** did not. The latter is consistent with the known retention of high methylation levels of IAPs throughout cleavage divisions[19].

Interestingly, over the more extended timeline, all retrotransposons resolved identically, reaching minimal values at ICM before increasing in overall methylation to levels observed within somatic cells by E6.5/7.5 **(Fig. 1-9i)**. Thus, repeat elements exist in a less methylated state primarily in the pre-implantation stages. Due to the technical limitations of bisulfite sequencing, we cannot measure whether all methylated cytosines at these repeats are converted to hmCs and only some are further modified to unmethylated cytosines while the remaining ones follow a passive trend. Alternatively, only a subset of mCs may be targeted via this intermediate for active demethylation with the remaining mC/hmC members losing their methylation passively over the cleavage stages. This would be consistent with recent metaphase immunostaining results[10].

1.3.6 Sperm and the oocyte contribute distinct genomic features as heritable DMRs

While targeted loss of methylation is widespread, the methylation state of some genomic elements must be differentially contributed from the two gametes. These include known ICRs that maintain their allele-specific methylation pattern throughout embryogenesis[40]. We systematically searched for novel DMRs inherited from either gamete and defined them as either oocyte- or sperm- contributed DMRs, respectively, based on the gamete where the region is hypermethylated. We applied linear regression to all 100bp tiles that had mean methylation greater than 0.75 in one gamete and mean methylation less than 0.25 in the other gamete to identify those tiles with an intermediate methylation level in the zygote. We identified 318 oocyte-contributed DMRs with intermediate methylation levels in the zygote ($P<0.04$, FDR<0.05, ANOVA; linear regression residual <0.29, FDR<0.1) **(Fig. 1-11a)** as well as 3,607 similar sperm-contributed DMRs **(Fig. 1-11c)**. We found that 4.5% and 4.3% of the regions that

were hypermethylated in sperm and oocytes reached intermediate values in pre-implantation embryos, respectively. Notably, oocyte-contributed DMRs primarily reside in gene bodies and CpG island-containing promoters (**Fig. 1-11b**), whereas sperm-contributed DMRs were predominantly retrotransposons or unannotated (**Fig. 1-11d**). The sperm- and oocyte-contributed DMRs also differed substantially in their relative CpG densities (**Fig. 1-12**).

We next focused specifically on oocyte-contributed promoter DMRs, in part due to their unusual enrichment for high CpG promoters (HCPs). While they had no clear functional enrichment, they did include several interesting genes that are not expressed in later stages of oogenesis, such as the somatic isoforms of Dnmt1 and Dnmt3b[41-43], suggesting a repressive function for at least some DMRs. The use of comprehensively genotyped strains enabled us to confirm that the allele specific zygote and ICM methylation proximal to the CpG island promoter of Copine VII (Cpne7), another putative DMR, was directly inherited from the oocyte (**Fig. 1-13a**). We next examined CpG methylation along oocyte-contributed promoter DMRs at each developmental stage (**Fig. 1-13b**). Intermediate methylation values around the TSS are retained from the zygote through the ICM, after which they resolve to hypomethylation (**Fig. 1-13b,c**). Thus, CpG island methylation is only transiently stabilized during cleavage divisions before re-establishing an unmethylated state around implantation. This unusual CpG island methylation in pre-implantation embryos provides another example of the distinct modes of DNA methylation regulation during early development. Unique methylation patterns during pre-implantation development are also observed in sperm-contributed DMRs, which retain intermediate methylation values through the ICM, before being hypermethylated post-implantation at typical somatic levels (**Fig. 1-13d**).

Notably, while RRBS is designed to enrich for CpG dinucleotdies (6-fold enrichment), it does capture non-CpG dinucleotdies at normal frequencies. Of the other three cytosine-containing dinucleotide combinations, CpA is the predominant target for methylation in mouse and human[44,45]. Consistent with previous locus-specific observations[46,47], we found that oocytes had the highest global CpA methylation level observed across pre-implantation development, and that this level decreased by ~50% in the zygote stage. This indicates that non-CpG methylation is inherited as part of the oocyte-contributed methylated alleles (**Fig. 1-14**). This pattern is stronger for novel DMRs, which display a mean CpA methylation of ~0.16 within the TSS boundary in the oocyte, and decrease to ~0.07 in the zygote (**Fig. 1-13b**).

1.4 Discussion

To better understand the regulation of methylation patterns during its most dynamic phase, we have generated genome-scale maps of DNA methylation in both gametes and through the complete pre-implantation timeline. Our observations greatly improve the scale and resolution of the existing model (reviewed in Ref. 11). Specifically, DNA methylation information contributed by sperm to the zygote is most dramatically altered in retro-elements of specific families, while other elements remain more protected and retain higher methylation levels throughout development (**Fig. 1-15**). Moreover, the general methylation status of the oocyte is a strong predictor of methylation values in the zygote stage, suggesting that the mechanism and targets of DNA demethylation during female gametogenesis could be the same as or very similar to those that dictate active demethylation targets of the paternal genome at fertilization[33]. Alternatively, regions that are already hypomethylated in the oocyte could explain the disparity

in global methylation between the early embryo and sperm. Regardless, the embryonic pattern most closely resembles that of the oocyte until the later stages of pre-implantation development where DNA methylation is further decreased.

In addition to classical ICRs, which exhibit parent-of-origin specific methylation that is maintained into adulthood, a substantial number of CpG island promoters are specifically hypermethylated in the oocyte in agreement with a recent study[36]. Surprisingly, these regions are retained at intermediate values indicative of differential allelic methylation before gradually decreasing through ICM specification and gastrulation, where somatic methylation patterns are generally re-established **(Fig. 1-15)**.

It remains to be investigated whether the diverse targets that exhibit low methylation levels retain their unique status during embryogenesis as a consequence of a single conserved regulatory mechanism. As a general principle, LINE and LTR activity in the early embryo is associated with some of the earliest transcriptional events during zygotic genome activation, and targeted depletion by antisense oligonucleotides of the L1Md_T class as well as certain LTRs have demonstrated a general requirement for retrotransposon transcription for progression through cleavage divisions[48,49]. These observations also fit with screens that have identified the transcription elongation factor ELP3 as a component of the DNA demethylation machinery and could explain a tight relationship between demethylation and transcription-associated complexes[50]. The notable interest in the global conversion of paternal methylcytosine to hydroxymethylcytosine as mediated by Tet3 (Refs 8,9) will over the coming years lead to technical improvements that might eventually allow investigation and dissection of mC and hmC

dynamics in early mammalian development. These technological advances will provide answers to questions regarding Tet3's universal necessity for conversion to unmethylated cytosines as well as address the effect this modification may have directly on Dnmt-mediated inheritance. In the absence of these technologies, we can make the following conclusions that are consistent with the published literature. Namely, Tet3's global conversion to hmC of the paternal genome does not appear to lead to equivalently dramatic conversion to unmethylated cytosine based on the retention of bisulfite-detected methylation. The feature-specific dynamics of DNA methylation at fertilization suggest that Tet3 and hmC may be required for targeted demethylation, as well as for driving a gradual hypomethylation over the progressive divisions of cleavage. However, other mechanisms might retain heritable methylation information at many sites in the genome because many targets display relative epigenetic stability from zygote onward. Importantly, many of these features exhibit embryogenesis-specific methylation patterns. Our data provide an improved understanding of the relationship and general targets of DNA demethylation at fertilization and could be used to refine current models regarding its mechanism.

In conclusion, the earliest stages of mammalian embryogenesis operate with a distinct methylation landscape that is not observed outside of pre-implantation or in the oocyte itself. This global hypomethylation is generally stabilized over multiple cleavage divisions for most genomic features, consistent with passive demethylation, until a minimum is reached at the ICM stage (**Figs. 1-1, 1-11 and 1-15**). Further experiments will be required to characterize the anticipated division-dependent demethylation in more detail, and expand it to regions with lower GC content that are under-represented in RRBS. Following these events, global methylation

patterns are rapidly reset and begin to resemble somatic tissues by the onset of gastrulation. Thus, more than a decade after the initial immunohistochemistry-based observation of pre-implantation DNA methylation dynamics, our representative genome-scale single base resolution data have begun to address many of the remaining open questions and set the stage for future epigenetic studies in early mouse development.

1.5 Methods

1.5.1 Preparation of Samples

Isolation of gametes, pre- and post-implantation embryos was performed using procedures described in detail elsewhere[2]. Briefly, 4-6 week old BDF1 female mice (Charles River) were injected with 5 IU of Pregnant Mare Gonadotropin (Sigma) followed 46h later by 5 IU Human Chorionic Gonadotropin (Sigma). Primed mice were then either directly used to collect oocytes or mated with 129X1 male mice (Jackson) to collect fertilized embryos. Twelve hours after final hormone injection, oocytes or zygotes were isolated from the ampulla under mineral oil and collected in hyaluronidase containing M2 medium (Millipore) drops to eliminate cumulus cells or spermatocyte contaminants. Oocytes were then depleted of somatic contaminants via progressive dilution through sequential drops of CO_2 buffered, amino acid supplemented KSOM medium (Millipore) until no somatic contaminants were observed.

Embryos were cultured in KSOM until collection at progressive cleavage stages with isolation occurring within 6 hours of the first observed cleavage event for that stage. Zygotes were screened for the presence of visible pronuclei and subjected to XY Clone (Hamilton Thorne) laser assisted polar body biopsy using a 8μm bore piezo pipette (Humagen, **Fig. 1-2**).

Clean cleavage stage embryos underwent an identical approach, with developmental progression unhindered by biopsy conducted at the 2-cell stage (**Fig. 1-2**). For each collection, batches of embryos were carefully screened to ensure each stage did not contain any abnormal embryos. Collection for zygotes was timed at ~10 hpf with fertilization assumed to occur 6-8 hrs after HCG injection, which was again confirmed by the relative synchronicity of the first cleavage division and by relative pronuclear stage. Biopsies were conducted in M2 media (Millipore) in batches of 5-10 embryos to reduce time on the micromanipulator stage. Before the final collection, cleaned and sorted samples were washed with Acid Tyrode's solution (Sigma) to eliminate the zona pellucida and to deplete any residual somatic contaminants or polar bodies through a short series of additional washes.

The inner cell mass was collected from blastocysts flushed from the uteri of naturally mated mice 3.5 days after fertilization using M2 or FBS supplemented DMEM followed by sequential washing in KSOM. The ICM itself was enriched from collected blastocysts by treating the embryo with rabbit anti-mouse serum (Sigma) before immunosurgical depletion of the trophectoderm using Guinea Pig Complement Serum (Sigma). Isolated ICMs were serially washed after isolation to remove contaminants (see **Figure 1-1**).

E6.5 and 7.5 embryos were isolated using mechanical dissection of the decidua from the uterine lining of mated mice. Samples were again serially washed and extra-embryonic tissues dissected from ICM derived tissues using fine glass capillaries (see **Figure 1-1**).

Swimming sperm was collected from the caudal epidinimis of male mice within 5 days of a successful natural mating scored by copulation plug.

All samples were then collected at minimal volume and either snap frozen or immediately resuspended in DNA lysis buffer.

1.5.2 Preparation of Reduced Representation Bisulfite-sequencing Libraries

RRBS libraries were generated as described[3-5]. Briefly, DNA was isolated from snap frozen embryos in DNA lysis buffer (100mM Tris-HCl (pH 8.5), 5mM EDTA, 0.2% SDS, 200mM NaCl) supplemented with 300μg/mL Proteinase K (Invitrogen) followed by Phenol:Chloroform extraction, Ethanol precipitation and resuspension in EB buffer. Isolated DNA was then subjected to MspI digestion (NEB), end repair using Klenow 3'-5' exo– supplemented with GTP, meCTP, and ATP in a 1:1:10 ratio to facilitate 3' A tailing, and ligation of standard adapters using ultraconcentrated 10^6 U T4 DNA ligase (NEB) and extended 20 hour ligation at 16°C. Size selection of 40-150 and 150-270bp fragments containing ligated adapter was conducted by extended gel electrophoresis using NuSieve 3:1 agarose (Lonza) and gel extraction (Qiagen) using 100ng dephosphorylated, sonicated *E. coli* DNA as a protective carrier and to increase overall yield. The isolated molecular weight fractions in a given RRBS library were then separately treated with Sodium Bisulfite using the Epitect® Bisulfite conversion and column purification sytem (Qiagen) with a modified conversion strategy as described[3]. Following clean up, optimal PCR cycle number to generate the final libraries was gauged using diagnostic PCR's for each library. Final libraries were then generated from the complete Bisulfite converted pool and purified through a second round of gel electrophoresis. High- and low- molecular weight fragments were then either sequenced separately or pooled at a 2:1 ratio by mass to assume an equimolar representation of both size ranges. Libraries were then sequenced on an Illumina Genome Analyzer II before alignment and analysis. The sequencing reads were aligned to the Mouse Genome Build 37 (mm9) using a custom computational pipeline

taking into account the strain background for each sample[5,6]. To supplement our data set we included sperm replicate 2 from Ref. 39 (SRA#ERP000689).

1.5.3 Estimating methylation levels

The methylation level of each sampled cytosine was estimated as the number of reads reporting a C, divided by the total number of reads reporting a C or T. Single CpG methylation levels were limited to those CpGs that had at least 10-fold coverage. For 100bp tiles, reads for all the CpGs that were covered more than 5-fold within the tile were pooled and used to estimate the methylation level as described for single CpGs. The CpG density for a given single CpG is the number of CpGs 50bp up and downstream of that CpG. The CpG density for a 100bp tile is the average of the CpG density for all single CpGs used to estimate methylation level in the tile. CpA methylation levels were estimated in the same way as for CpG methylation.

The methylation level reported for a sample is the average methylation level across replicates. A replicate will contribute to the average only if it meets the coverage criteria within the replicate.

1.5.4 Genomic features

High density CpG promoters (HCP), intermediate density CpG promoters (ICP), low density CpG promoters (LCP), transcription start sites (TSS), CpG island, and DMR annotations were taken from[6]. Promoters are defined as 1kb up and downstream of the TSS. LINE, LTR, and SINE annotations were downloaded from the UCSC browser (mm9) RepeatMasker tracks. Gene annotations were downloaded from the UCSC browser (mm9) refseq track. In each case, the methylation level of an individual feature is estimated by pooling read counts for all CpGs within the feature that are covered greater than 5-fold, and levels are only reported if a feature contains

at least 5 CpGs with such coverage (in contrast to 100bp tiles where no minimum number of CpGs is required). A tile is annotated as a genomic feature if any portion of the tile overlaps with the feature and may be annotated by more than one feature (e.g. the same can be annotated as both a promoter and a gene).

1.5.5 Identification of tiles with changing methylation levels and their enrichments

A tile is considered changing if it both has a methylation difference ≥ 0.2 between two stages and is significant in a two sample t-test with unequal variance after correction for multiple hypothesis testing (FDR < 0.05) using the Benjamini-Hochberg method[7]. Enrichment p-values are from the hypergeometric distribution where the background is the number of tiles that have a methylation difference ≥ 0.2 and are corrected for multiple hypotheses at FDR < 0.05, based on the number of gene sets tested.

1.5.6 Identification of enriched retrotransposon families

The same procedure for identifying changing tiles was applied to the methylation levels of retrotransposon elements to identify changing elements. Enrichment for families was done using annotations from the RepeatMasker track of the UCSC genome browser.

1.5.7 Novel DMR identification

100bp tiles where one gamete had a mean methylation greater than 0.75 and the other gamete had a mean methylation of less than 0.25 were flagged as potential DMRs. Linear regression was used to identify tiles which had methylation levels in zygote which were halfway between the methylation levels in oocyte and sperm. Only tiles that had two replicates present in each time point were considered. Residuals were calculated as the mean of the differences between the model predictions and the data taking into account missing values. ANOVA was used to assign a

p-value to each tile. A tile was considered a novel DMR if it had a residual in the tenth percentile of tiles tested and a significant p-value from ANOVA with a Benjamini-Hochberg FDR < 0.05. A residual in the tenth percentile corresponds to an FDR < 0.1 by a permutation test where zygote methylation values are shuffled for potential DMR tiles. In the pie charts (**Fig. 1-13b,d**), the genomic feature that covered the most novel tiles was reported first and then subtracted from the set before reporting the feature which covered the next largest number of tiles. This procedure was repeated until all tiles were categorized. The one exception was for Oocyte-contributed DMRs where promoters were taken out before genes.

1.5.8 Identification of SNPs

An initial set of SNPs between 129X1 and BDF1 (C57/B6 x DBA2/J) was taken from Mouse Genome Informatics[8]. The set was filtered such that SNPs that fell into the following categories were removed: (1) SNPs that had inconsistent entries for the same position, (2) SNPs not trackable by RRBS (C/T or A/G), (3) SNPs between C57/B6 and DBA2/J, and (4) SNPs that were not covered by X1 and BDF1 in an *in silico* digest. The log odds ratio [log2(X1 count +0.01/C57 count+0.01)] was calculated for each SNP that was covered in the data set (n=742). SNPs that had at least 10x coverage with an average log odds ratio across all replicates between -5 and 5 and a Sperm X1 log odds ratio greater than 2 were considered of stringent quality (n=578) and used to assess both maternal bias and to serve as a general quality control metric for all libraries incorporated into the data set.

1.5.9 Parent of origin methylation tracking

The 742 SNPs identified above corresponded to 1416 CpG dinucleotides and were used to track allelic single CpG methylation. Reads were segregated into either X1 or BDF1 according to SNP type, and CpG methylation levels were called in the same manner described above. SNP

normalized methylation values (**Fig. 1-5**) are calculated by pooling CpG counts across replicates for each strain and then ascribing allelic methylation values. The normalized methylation value is the average of the methylation values derived from each strain.

1.6 Figures

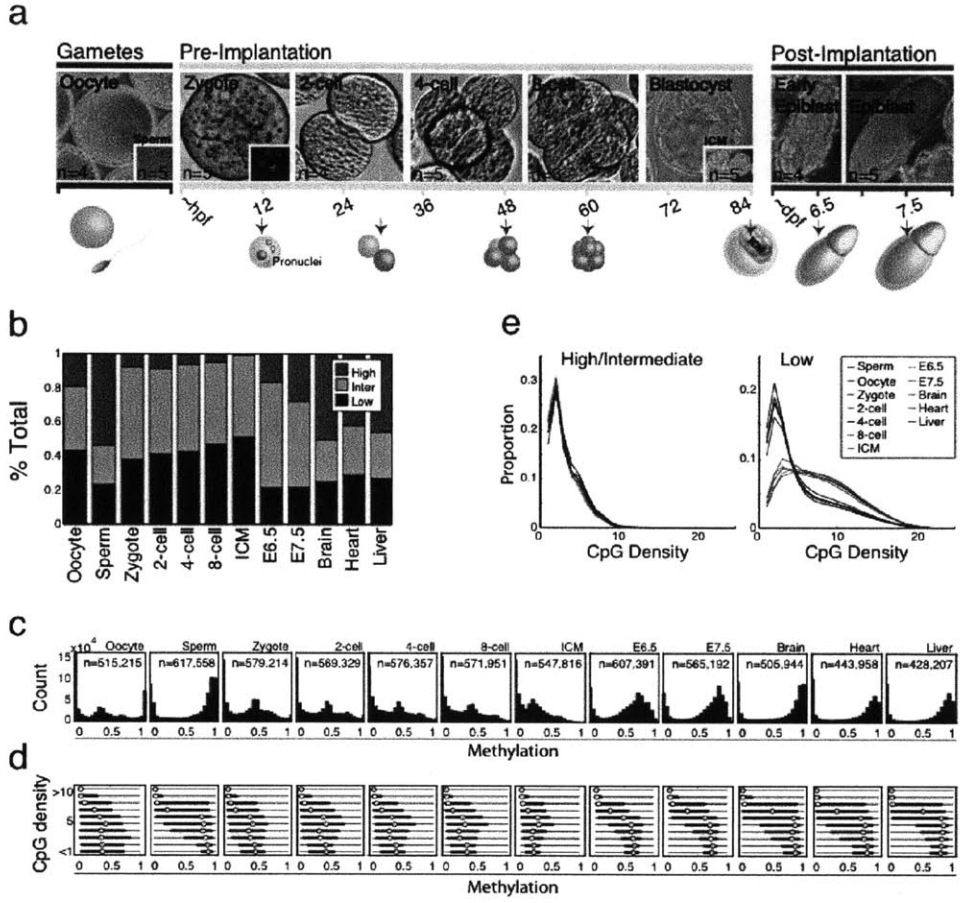

Figure 1-1: Global CpG methylation dynamics across early murine embryogenesis.
(a) Representative images of highly purified, polar body depleted oocyte, zygote (inlet; Hoechst stained pronuclei), 2-, 4- and 8- cell embryos (40x) (see **Fig. 1-2**). Sperm is shown as an inlet. The inner cell mass (ICM, inlet) was immunosurgically isolated from E3.5 blastocysts. *n* highlights replicate number. The schematic below describes the detailed timeline (hpf: hours post fertilization and dpf: days post fertilization) and red arrows point to the approximate time points when samples were isolated. **(b)** Fraction of 100bp tiles with High (≥0.8, red), Intermediate (Inter, >0.2 and <0.8, green) and Low (≤0.2, blue) methylation values. Brain, heart and liver tissue are shown for adult comparisons. **(c)** Histogram of methylation values (X-axis) across 100bp tiles, where each CpG within a tile must be sampled at least 5x for each timepoint. *n* is the number of tiles for each stage. **(d)** The distribution (as box plots, X-axis) of methylation values at different local CpG densities (defined as the average number of CpGs $^{+}$/- 50bp around each CpG sampled in the tile, Y-axis) highlight the difference between hypomethylated pre-implantation tissues and the canonical methylation architecture seen in sperm, post-implantation and somatic samples. Bulls-eye indicates the median, edges the 25th/75th percentile and whiskers the 2.5th/97.5th percentile. **(e)** Comparative density of >0.2 methylation (left panel) and ≤0.2 methylation (right panel) tiles in stages that display somatic patterning versus embryonic patterning (red and blue lines, respectively).

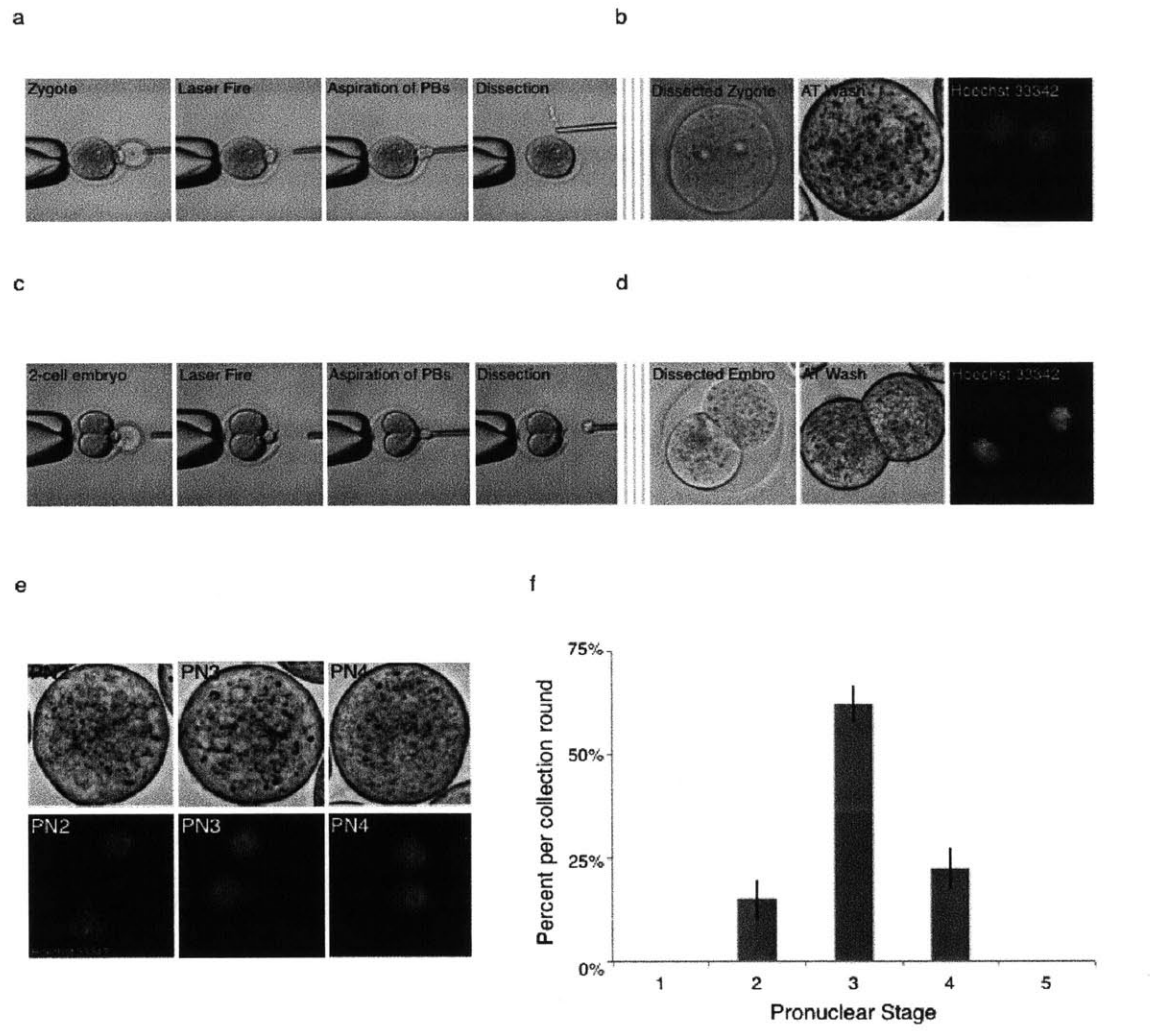

Figure 1-2: Isolation of samples and replicates for RRBS analysis. **(a)** Microdissection of polar body contaminants using laser ablation of the zona pellucida and aspiration with a piezo micromanipulator. Meiosis I and II polar bodies were mechanically removed in this fashion before collection of zygotic replicates. **(b)** After microdissection, embryos were washed in Acid Tyrode's solution and washed through serial microdrops to eliminate remaining contaminants. Staining using the intercalating dye Hoechst 33342 confirmed an absence of genomic contaminants and revealed the two pronuclei. **(c)** To facilitate normal development, both 2-cell and later cleavage stages were biopsied after first cleavage using a similar strategy as that for zygotes. **(d)** Acid Tyrode mediated ablation of the zona pellucida and Hoechst staining revealed no detectable contaminants using our described isolation strategy in cleavage stage embryos. **(e)** Zygote samples were isolated from natural matings using two visible pronuclei as a selective criteria. This restricted us to pre-syngamy embryos in which the pronuclei had not fused. Our scoring of each embryo followed the highlighted schematic adopted from Ref 1. **(f)** Pronuclear staging of triplicate replicates collected for RRBS analysis.

a

Sample	n	Input	Total CpGs (1x)	Total Unique (1x)	Mean Cov (1x)	Total CpGs (10x)	Total Unique (10x)	Mean Cov (10x)
Oocyte	4	~250 cells	54222070	1065085	53	53026780	754375	73
Sperm	5	~33 ng	49056060	1622074	30	46982560	1050991	44
Zygote	5	~63 cells	44368020	1657928	27	41760426	904473	45
2-cell	4	~100 cells	37411130	1624408	23	34903450	954615	36
4-cell	5	~170 cells	39351950	1706403	23	36791850	1029450	36
8-cell	3	~180 cells	41278790	1750977	24	38766400	1083324	36
ICM	5	~23 embryos	39284030	1535245	25	36594850	843261	36
E6.5	4	~6 embryos	59463460	1528915	39	58101030	1200623	48
E7.5	5	~6 embryos	62064200	1521215	41	61033030	1268909	48
Brain	2	50 ng	68372820	1627874	42	66613733	1120374	58
Heart	1	50 ng	57850729	1713221	34	55983685	1127397	50
Liver	2	50 ng	59559308	1545726	39	58036307	1126865	51

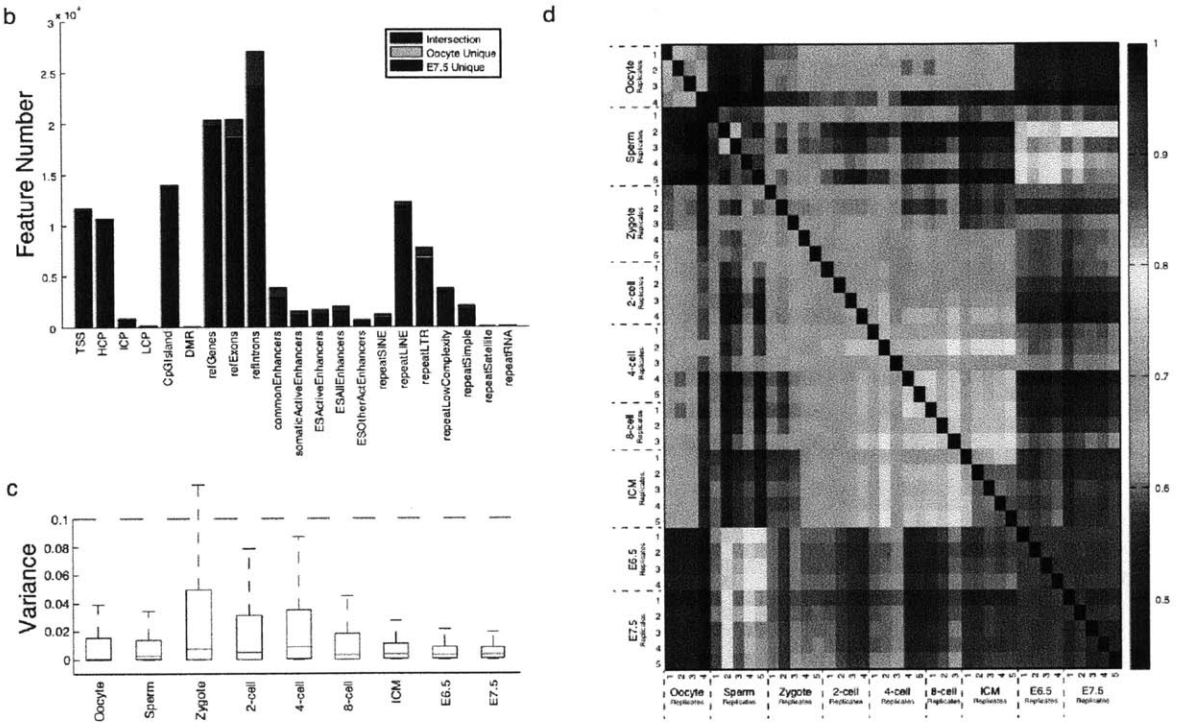

Figure 1-3: Comparison of RRBS performance across stages and between replicates. **(a)** CpG coverage captured by RRBS. Shown are the number of replicates (n) at each stage, the average input per timepoint, the numbers of total and unique CpGs, and their mean coverage (Cov) at 1X or 10X. **(b)** Shown is the correspondence in the sampled features for different genomic annotations (X-axis) between single replicates of low input oocyte and higher input E7.5 embryo libraries. Blue: features covered by both samples; green: features covered in oocyte only; red: features covered in E7.5 only. Overall, there is mostly similar coverage of sampled features per library. **(c)** Box plot of the variance (Y-axis) across replicates at each stage (X-axis). Red line indicates the median, edges the 25th/75th percentile and whiskers the 2.5th/97.5th percentile. **(d)** Pearson correlation coefficient heatmap between replicates and timepoints for all RRBS libraries used in these analyses. Correlation is generally higher within stages than between stages. Blue: correlation coefficient 0.45 or lower; bright red: 1.

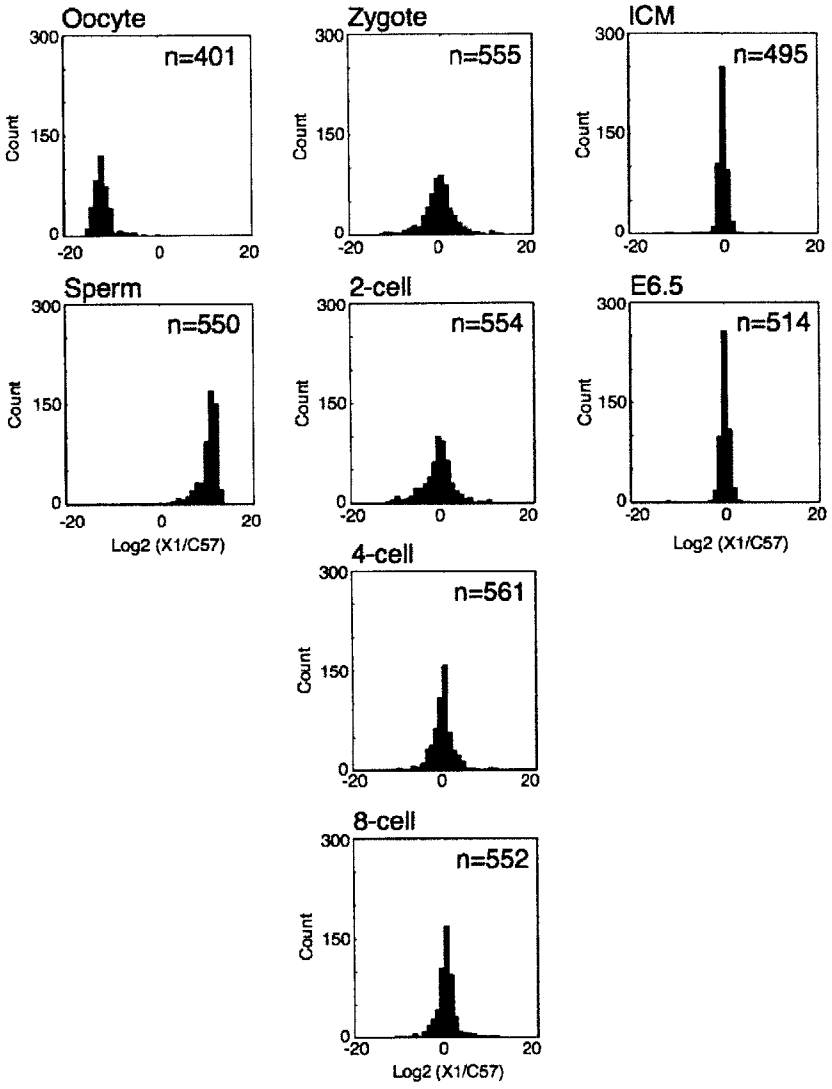

Figure 1-4: Parent of origin Single Nucleotide Polymorphism (SNP) distributions for isolated gametes and hybrid embryos. Log odds ratio histograms for allelic frequencies for BDF1 (C57/B6 based) maternal and 129X1 paternal SNPs demonstrate minimal maternal biasing throughout the cleavage stages. The number of SNPs captured (n) for each time point is highlighted.

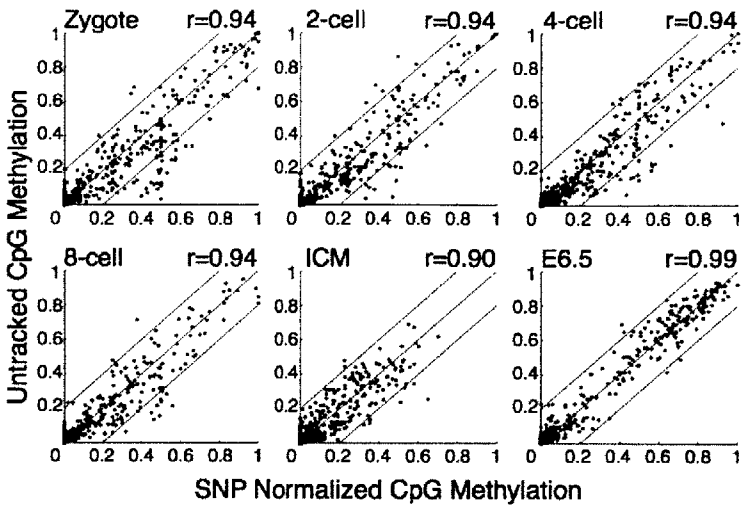

Figure 1-5: Reported methylation values reflect their contributions from paternal and maternal alleles. Scatterplots depicting untracked methylation values against SNP normalized CpG methylation values for each stage in which hybrid crosses were used. Red lines indicate the X=Y +/- 0.2.

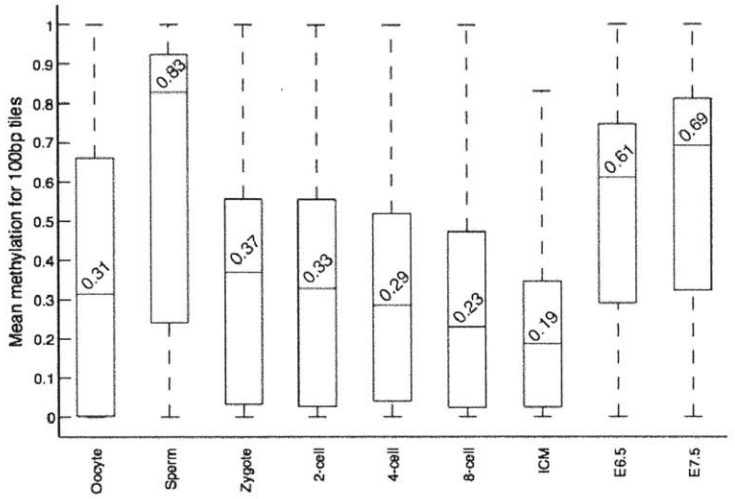

Figure 1-6: Methylation values for 100bp tiles across pre-implantation development. Box plots of the methylation value per tile (Y-axis) at each developmental stage (X-axis). Red line indicates the median, edges the 25th/75th percentile and whiskers the 2.5th/97.5th percentile. The median value is shown above its line.

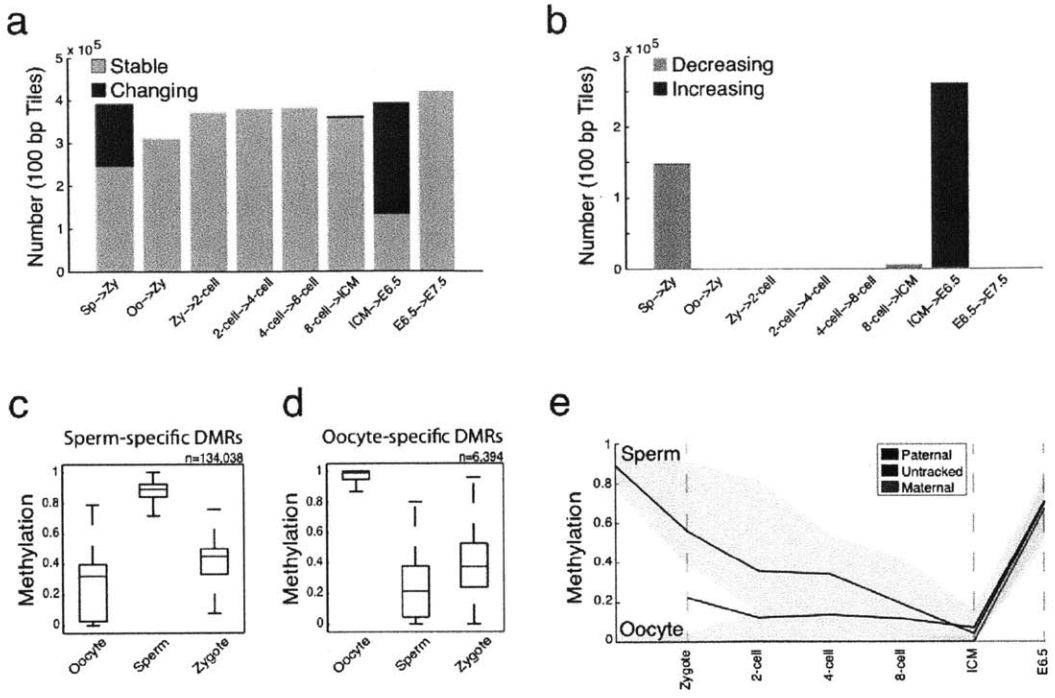

Figure 1-7: Major transitions in DNA methylation levels during early development. **(a)** Number of 100bp tiles available for pairwise comparison across consecutive embryonic stages (X-axis). Tiles that remain unchanged (stable) at the indicated transitions are shown in light blue. Tiles that change by greater than 0.2 and are significant by t-test are highlighted in dark blue. **(b)** Number of 100bp tiles with increasing (red) or decreasing (green) methylation levels at each consecutive transition (X-axis, as in **(a)**). For major transitions, 100bp tiles change their methylation status largely in one direction. **(c)** Distribution of methylation levels for the sperm-specific DMRs (n=102,862 tiles). Red line indicates the median, edges the $25^{th}/75^{th}$ percentile and whiskers the $2.5^{th}/97.5^{th}$ percentile. **(d)** Distribution of methylation levels for the oocyte-specific DMRs (n=3,219 tiles). Box plots as in **(c)**. **(e)** Approximately 65 CpGs within the sperm-specific DMR tiles **(c)** could be ascribed to paternal and maternal alleles and tracked across stages. Paternal CpG methylation values (blue line, median; colored space, $25^{th}/75^{th}$ percentile) exhibit marked decrease by the zygote stage while maternal CpG methylation (red line, median; colored space, $25^{th}/75^{th}$ percentile) remain unchanged. The untracked median methylation value for these CpGs describes intermediate values between parent-of-origin allelic values (black line).

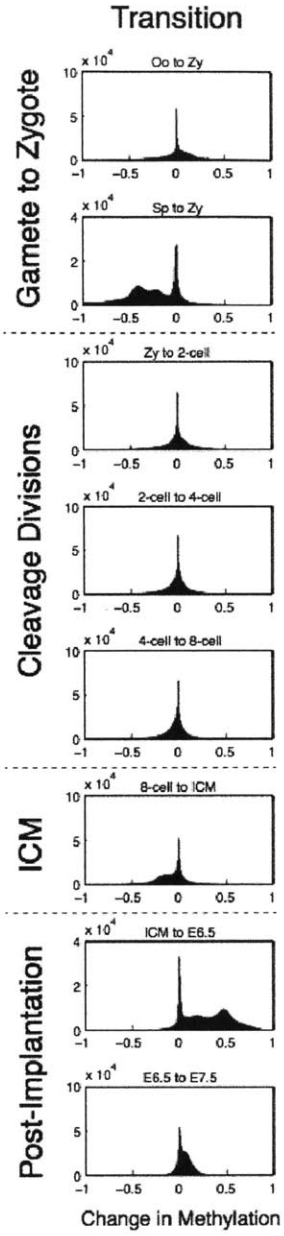

Figure 1-8: Distribution of change in methylation levels in 100bp tiles between consecutive stages. Histograms show the distribution of the difference between all methylation values available for comparison between consecutive stages. There are two major transitions: (1) sperm to zygote and (2) ICM to E6.5. For other transitions, the difference is less skewed and reveals minimal change for these steps with gradual hypomethylation observed in the later cleavage stages and through ICM specification.

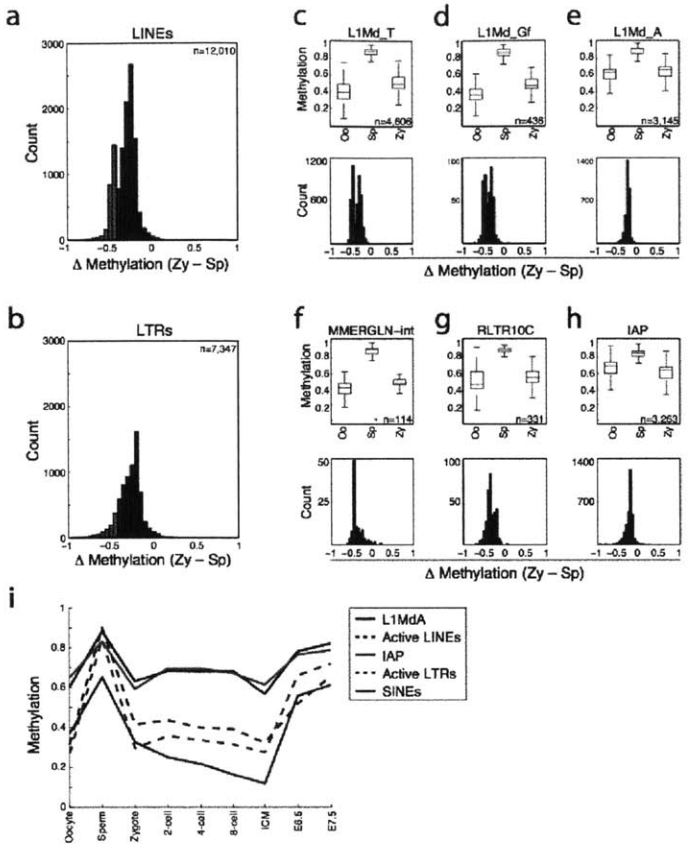

Figure 1-9: Specific families of LINE and LTR retroelements exhibit the most dramatic changes in the sperm to zygote transition. **(a)** Histogram of the difference in methylation levels (X-axis, negative values represent tiles decreasing from sperm to zygote) between sperm and zygote within LINE-1 elements. 74% of the elements have a significant difference ($P<0.038$,FDR<0.05; t-test). The distribution is bimodal with 18% of elements displaying a change in methylation status greater than 0.45 highlighted in red. **(b)** Differences in methylation between sperm and zygote within annotated LTR retroelements. Compared to LINE-1, a smaller fraction of elements appear regulated by DNA demethylation (41% significant, 12% of those sampled exhibiting changes greater than 0.45 highlighted in red). **(c-e)** Methylation levels in oocyte, sperm and zygote (top panels) as well as the distributions of change in methylation levels between sperm and zygote (bottom panels) for specific families of the LINE-1 class, including those that change **(c,d)** and those that stay comparatively hypermethylated **(e)**. Top panels: Red line indicates the median, edges the $25^{th}/75^{th}$ percentile and whiskers the $2.5^{th}/97.5^{th}$ percentile. Bottom panels: members of each family that are demethylated by greater than 0.45 are highlighted in red. **(f-h)** Methylation levels in oocyte, sperm and zygote (top panels) and the distributions of change in methylation levels between sperm and zygote (bottom panels) for specific families of LTR containing retroelements, including MMERGLN **(f)**, RLTR10C **(g)** and IAP elements **(h)**. Top and bottom panels as in **(c-e)**. **(i)** Mean methylation level (Y-axis) for all elements of the L1Md_A LINE (solid blue line) and IAP LTR class (solid red line) that do not dramatically change contrasted to LINEs (dashed blue line) and LTR elements (dashed red line) that show the greatest loss at fertilization. SINE elements (green line) are less methylated in sperm than other repeat elements and are generally demethylated to oocyte levels.

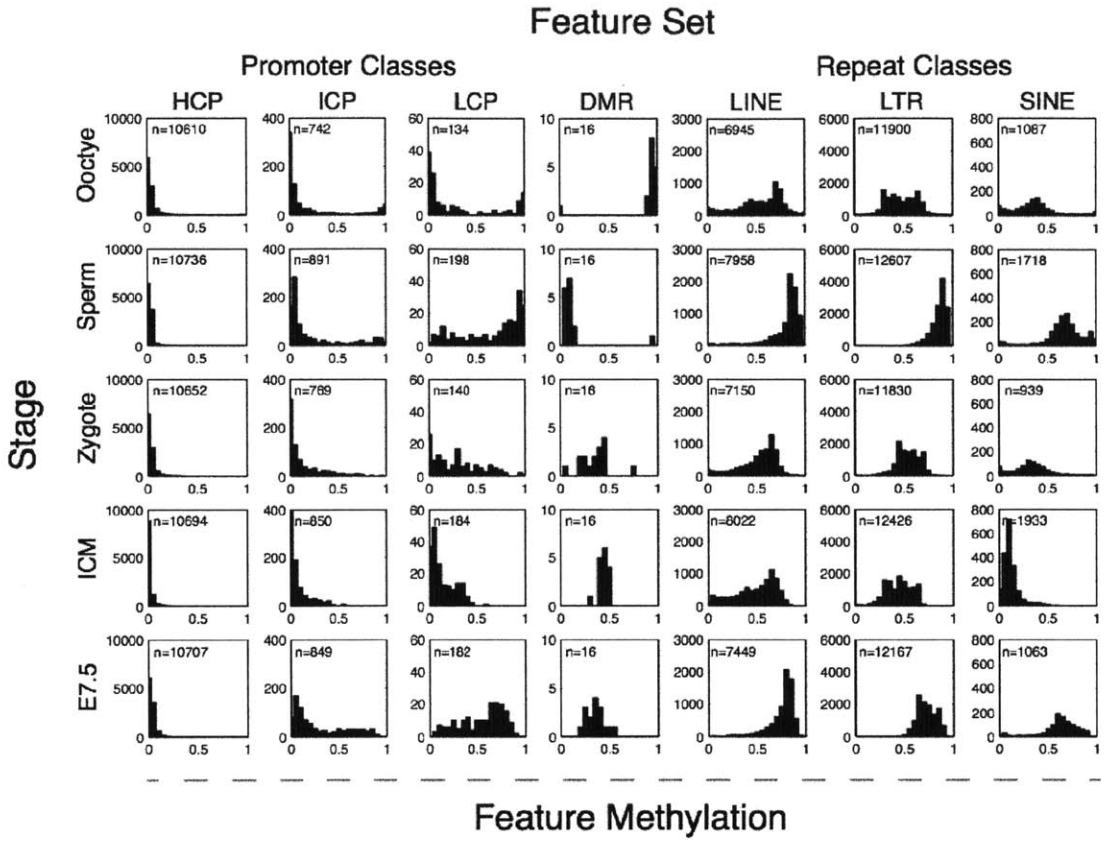

Figure 1-10: Methylation histograms for genomic feature annotations throughout pre-implantation development. Notable dynamics at fertilization, across pre-implantation and upon specification of the embryo proper occur across multiple genomic feature sets. n indicates the number of genomic features captured at each time point.

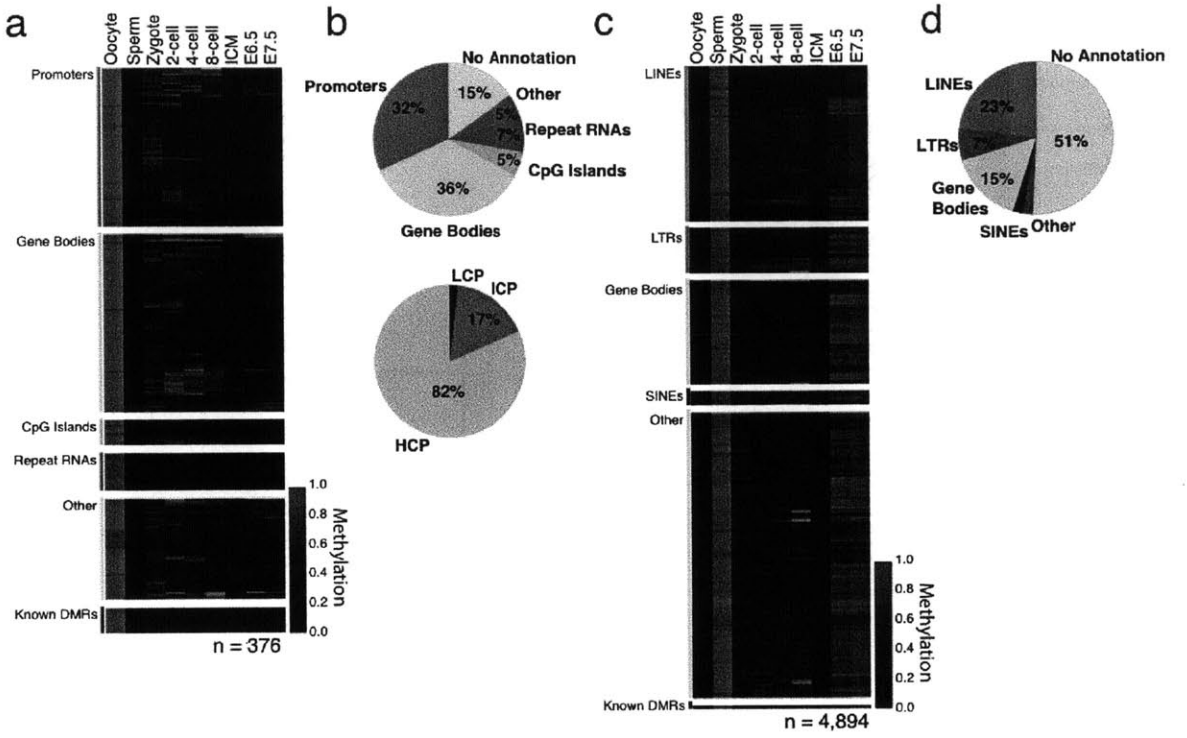

Figure 1-11: Differentially methylated regions represent discrete gamete specific feature classes. (a) Heatmap of methylation levels (black: 0; red: 1; grey: missing value) in 318 identified 100bp tiles (rows) that behave as oocyte-contributed DMRs in the zygote. Tiles are sorted by functional classes (labels, left) and clustered within each class. 15 known ICR regions, shown at the bottom, behave similarly in the early embryo and retain intermediate methylation through implantation. **(b)** Distribution of genomic features (top) and promoters of different CpG densities (bottom) in oocyte-contributed DMRs. Top: oocyte DMRs are enriched for promoter and gene body elements. Bottom: most of the 82 promoters that overlap tiles identified as oocyte-contributed DMRs are high CpG density promoters containing CpG Islands (HCPs, light blue). **(c)** Heatmap of methylation levels (black: 0; red: 1; grey: missing value) in 3607 identified 100bp tiles (rows) that behave as sperm-contributed DMRs in pre-implantation embryos. Tiles are sorted by functional classes (labels, left) and clustered within each class. Known DMRs contributed by sperm are at the bottom. **(d)** Distribution of genomic features in sperm-contributed DMRs are specifically enriched for retrotransposon elements (dark grey) and intergenic sequences and are not enriched for promoter regions.

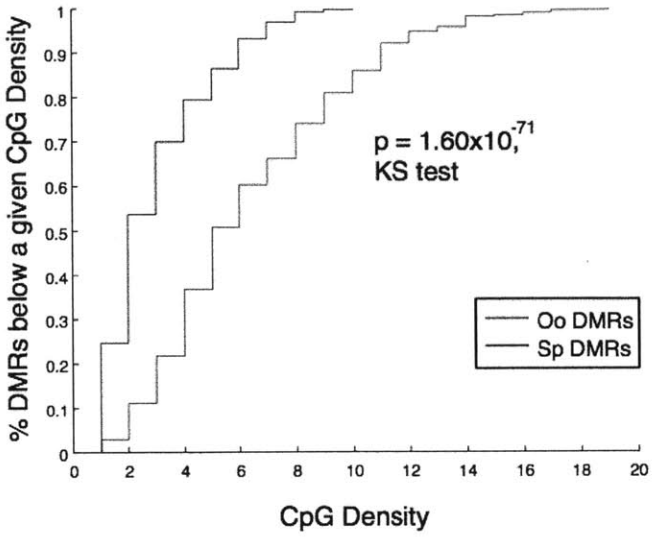

Figure 1-12: Cumulative distribution function plot of CpG densities for oocyte- and sperm-contributed DMRs. The CDF (cumulative distribution function) plot displays the fraction of CpGs (Y-axis) in the set that have a CpG density no greater than a certain value (X-axis) for oocyte- (red) and sperm- (blue) contributed DMRs. The median CpG density for each set is indicated by the X-axis value corresponding to y=0.5 (median CpG density is 6 and 2 for oocyte- and sperm- contributed DMRs, respectively). The CpG density dependence is significantly different between the two sets of DMRs ($P=3.94x10^{-50}$; KS-test).

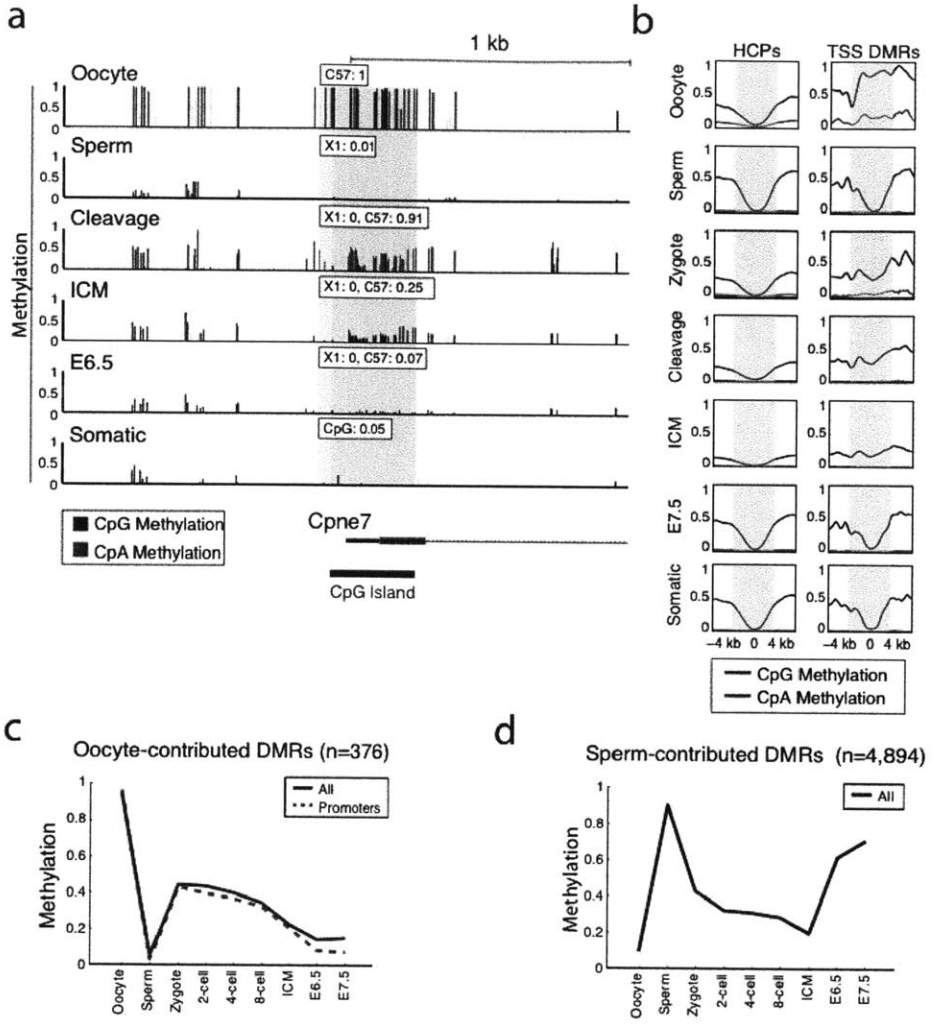

Figure 1-13: DMRs resolve after cleavage to univalent hyper- or hypo- methylated values in a gamete-of-origin specific fashion. **(a)** Methylation levels at single CpG resolution at the Cpne7 locus for each promoter CpG captured within 2kb of the TSS for both gametes and across embryonic development (rows). Dark gray bar highlights the targeted CpG island of the promoter. Blue bars represent CpG methylation and red bars highlight CpA methylation. A CpG proximal to this island can be tracked to a phase resolving SNP and this region is highlighted in light gray, with paternal (X1) and maternal (C57) methylation values discretized and highlighted as an inset for each trackable phase. **(b)** Composite plot of CpG (blue) and CpA (red) methylation for all HCPs (left) and for promoters that are specifically hypermethylated in oocytes (TSS DMRs, right) at each gamete and stage (rows). The region +/- 2kb of the TSS is marked in gray. Identified promoter DMRs contributed by the oocyte are hypermethylated around the periphery of the TSS and resolve to intermediate values throughout cleavage. An expected HCP methylation architecture is re-acquired for these DMRs around implantation. **(c)** Mean methylation levels (Y-axis) and stage (X-axis) for DMRs in promoters (red, dashed line) versus our complete set (red, solid line). **(d)** Sperm-contributed DMRs (solid blue line) resolve almost exclusively to hypermethylation.

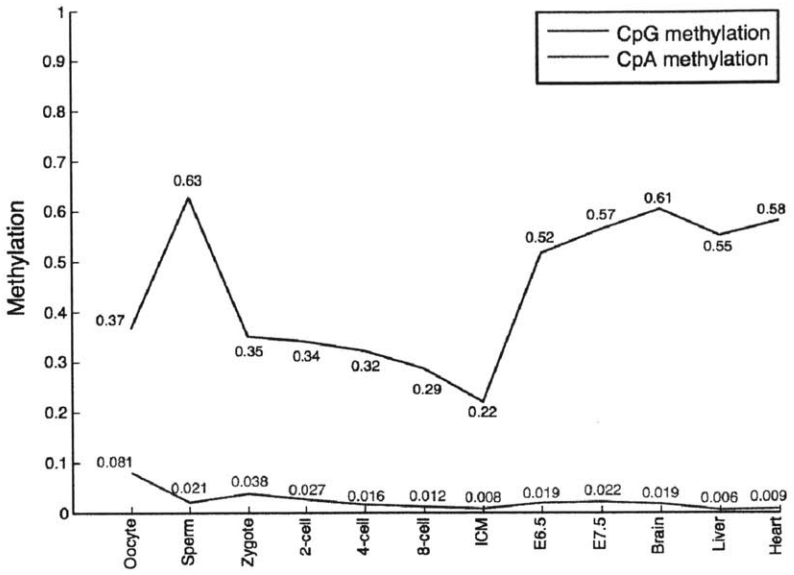

Figure 1-14: Global CpG and CpA methylation dynamics during pre-implantation development. Mean CpG (blue) and CpA (red) methylation (Y-axis) for 100bp tiles for each sample (X-axis). Mean values are marked and reveal that CpA methylation is highest in oocytes and depletes throughout cleavage.

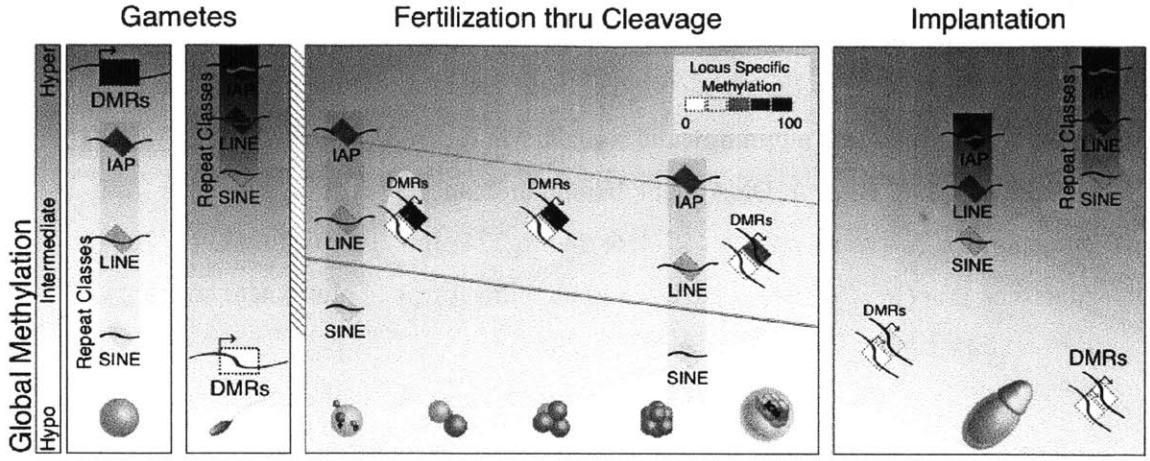

Figure 1-15: A model for DNA methylation dynamics during early embryogenesis. Globally, each gamete exhibits unique methylation patterns genome-wide, with oocyte levels closely mirroring those of the pre-implantation embryo. Global methylation as measured by bisulfite conversion is moderately stable throughout cleavage, but diminishes to minimal values at the blastocyst/ICM stage. Specific repeat classes exhibit disparate methylation levels that are either retained as high (in the case of IAPs) or demethylated to lower values at the zygote stage (in the case of certain L1Md families and SINEs). Methylation values for these elements are maintained through cleavage stages before the somatic pattern is restored at gastrulation. Some CpG island promoters exhibit hypermethylation in the oocyte and these putative DMR signatures show retention of maternal CpG methylation through to the end of the cleavage stages before resolving to expected hypomethylation in the embryo proper.

1.7 References

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

2 Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet* **33 Suppl**, 245-254 (2003).

3 Suzuki, M. M. & Bird, A. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* **9**, 465-476, doi:nrg2341 [pii] 10.1038/nrg2341 (2008).

4 Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766-770 (2008).

5 Weber, M. *et al.* Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet* **37**, 853-862 (2005).

6 Weber, M. *et al.* Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet* **39**, 457-466 (2007).

7 Ji, H. *et al.* Comprehensive methylome map of lineage commitment from haematopoietic progenitors. *Nature* **467**, 338-342, doi:nature09367 [pii] 10.1038/nature09367 (2010).

8 Wossidlo, M. *et al.* 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. *Nat Commun* **2**, 241, doi:ncomms1240 [pii] 10.1038/ncomms1240 (2011).

9 Gu, T. P. *et al.* The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. *Nature* **477**, 606-610, doi:nature10443 [pii] 10.1038/nature10443 (2011).

10 Inoue, A. & Zhang, Y. Replication-dependent loss of 5-hydroxymethylcytosine in mouse preimplantation embryos. *Science* **334**, 194, doi:science.1212483 [pii] 10.1126/science.1212483 (2011).

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

12 Kafri, T. *et al.* Developmental pattern of gene-specific DNA methylation in the mouse embryo and germ line. *Genes Dev* **6**, 705-714 (1992).

13 Borgel, J. *et al.* Targets and dynamics of promoter DNA methylation during early mouse development. *Nat Genet* **42**, 1093-1100, doi:ng.708 [pii] 10.1038/ng.708 (2010).

14 Meissner, A. Epigenetic modifications in pluripotent and differentiated cells. *Nat Biotechnol* **28**, 1079-1088, doi:nbt.1684 [pii] 10.1038/nbt.1684 (2010).

15 Razin, A. & Shemer, R. DNA methylation in early development. *Hum Mol Genet* **4**, 1751-1755. (1995).

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

17 Rougier, N. *et al.* Chromosome methylation patterns during mammalian preimplantation development. *Genes Dev* **12**, 2108-2113 (1998).

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

19 Lane, N. *et al.* Resistance of IAPs to methylation reprogramming may provide a mechanism for epigenetic inheritance in the mouse. *Genesis* **35**, 88-93 (2003).

20 Oswald, J. *et al.* Active demethylation of the paternal genome in the mouse zygote. *Curr Biol* **10**, 475-478 (2000).

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

22 Kim, S. H. *et al.* Differential DNA methylation reprogramming of various repetitive sequences in mouse preimplantation embryos. *Biochem Biophys Res Commun* **324**, 58-63, doi:S0006-291X(04)02067-4 [pii] 10.1016/j.bbrc.2004.09.023 (2004).

23 Bock, C. *et al.* Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nat Biotechnol*, doi:nbt.1681 [pii] 10.1038/nbt.1681 (2010).

24 Harris, R. A. *et al.* Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat Biotechnol*, doi:nbt.1682 [pii] 10.1038/nbt.1682 (2010).

25 Davis, T. & Vaisvila, R. High sensitivity 5-hydroxymethylcytosine detection in Balb/C brain tissue. *J Vis Exp*, doi:2661 [pii] 10.3791/2661 (2011).

26 Ficz, G. *et al.* Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* **473**, 398-402, doi:nature10008 [pii] 10.1038/nature10008 (2011).

27 Szulwach, K. E. *et al.* Integrating 5-hydroxymethylcytosine into the epigenomic landscape of human embryonic stem cells. *PLoS Genet* **7**, e1002154, doi:10.1371/journal.pgen.1002154 PGENETICS-D-11-00676 [pii] (2011).

28 Williams, K. *et al.* TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* **473**, 343-348, doi:nature10066 [pii] 10.1038/nature10066 (2011).

29 Wu, H. *et al.* Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. *Genes Dev* **25**, 679-684, doi:25/7/679 [pii] 10.1101/gad.2036011 (2011).

30 Xu, Y. *et al.* Genome-wide regulation of 5hmC, 5mC, and gene expression by Tet1 hydroxylase in mouse embryonic stem cells. *Molecular cell* **42**, 451-464, doi:S1097-2765(11)00283-8 [pii] 10.1016/j.molcel.2011.04.005 (2011).

31 Branco, M. R., Ficz, G. & Reik, W. Uncovering the role of 5-hydroxymethylcytosine in the epigenome. *Nat Rev Genet*, doi:nrg3080 [pii] 10.1038/nrg3080 (2011).

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

33 Hajkova, P. *et al.* Genome-wide reprogramming in the mouse germ line entails the base excision repair pathway. *Science* **329**, 78-82, doi:329/5987/78 [pii] 10.1126/science.1187945 (2010).

34 Wossidlo, M. *et al.* Dynamic link of DNA demethylation, DNA strand breaks and repair in mouse zygotes. *Embo J* **29**, 1877-1888, doi:emboj201080 [pii] 10.1038/emboj.2010.80 (2010).

35 Popp, C. *et al.* Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. *Nature* **463**, 1101-1105, doi:nature08829 [pii] 10.1038/nature08829 (2010).

36 Smallwood, S. *et al.* Dynamic CpG island methylation landscape in oocytes and preimplantation embryos. *Nat Genet.* **doi: 10.1038/ng.864** (2011).

37 Molaro, A. *et al.* Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. *Cell* **146**, 1029-1041, doi:S0092-8674(11)00942-1 [pii] 10.1016/j.cell.2011.08.016 (2011).

38 Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562 (2002).

39 Goodier, J. L., Ostertag, E. M., Du, K. & Kazazian, H. H., Jr. A novel active L1 retrotransposon subfamily in the mouse. *Genome Res* **11**, 1677-1685, doi:10.1101/gr.198301 (2001).

40 Edwardsa, C. & Ferguson-Smith, A. Mechanisms regulating imprinted genes in clusters. *Current Opinion in Cell Biology* **19**, 281-289 (2007).

41 Bestor, T. H. The DNA methyltransferases of mammals. *Hum Mol Genet* **9**, 2395-2402. (2000).

42 Hirasawa, R. *et al.* Maternal and zygotic Dnmt1 are necessary and sufficient for the maintenance of DNA methylation imprints during preimplantation development. *Genes Dev* **22**, 1607-1616, doi:22/12/1607 [pii] 10.1101/gad.1667008 (2008).

43 Lucifero, D. *et al.* Coordinate regulation of DNA methyltransferase expression during oogenesis. *BMC Dev Biol* 7, 36, doi:1471-213X-7-36 [pii] 10.1186/1471-213X-7-36 (2007).

44 Ramsahoye, B. H. *et al.* Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc Natl Acad Sci U S A* **97**, 5237-5242. (2000).

45 Ziller, M. J. *et al.* Genomic Distribution and Inter-Sample Variation of Non-CpG Methylation across Human Cell Types. *PLoS Genet* 7, e1002389, doi:10.1371/journal.pgen.1002389 PGENETICS-D-11-00694 [pii] (2011).

46 Haines, T., Rodenhiser, D. & Ainsworth, P. Allele-specific non-CpG methylation of the Nf1 gene during early mouse development. *Dev Biol.* **240**, 585-598 (2001).

47 Tomizawa, S. *et al.* Dynamic stage-specific changes in imprinted differentially methylated regions during early mammalian development and prevalence of non-CpG methylation in oocytes. *Development* **138**, 811-820, doi:dev.061416 [pii] 10.1242/dev.061416 (2011).

48 Beraldi, R., Pittoggi, C., Sciamanna, I., Mattei, E. & Spadafora, C. Expression of LINE-1 retroposons is essential for murine preimplantation development. *Mol Reprod Dev* **73**, 279-287, doi:10.1002/mrd.20423 (2006).

49 Kigami, D., Minami, N., Takayama, H. & Imai, H. MuERV-L is one of the earliest transcribed genes in mouse one-cell embryos. *Biol Reprod* **68**, 651-654 (2003).

50 Okada, Y., Yamagata, K., Hong, K., Wakayama, T. & Zhang, Y. A role for the elongator complex in zygotic paternal genome demethylation. *Nature* **463**, 554-558, doi:nature08732 [pii] 10.1038/nature08732 (2010).

Chapter 2

DNA methylation dynamics of the human pre-implantation embryo

Zachary D. Smith*, Michelle M. Chan*, Kate Humm*, Shila Mekhoubad, Aviv Regev, Kevin Eggan and Alexander Meissner. *These authors contributed equally to this work.

ZDS, MMC, KH, KE, and AM conceived and designed the study. KH and ZDS collected samples, and ZDS performed methylation profiling. MMC performed all analysis with assistance from ZDS. ZDS, MMC, KH, SM, AR, KE and AM interpreted the data. ZDS, MMC, and AM wrote the paper with the assistance of the other authors.

Chapter 2:

DNA Methylation dynamics of the human pre-implantation embryo

2.1 Abstract

During most stages of mammalian development, cell fate decisions are facilitated through the focal reconfiguration of chromatin states. The largely static epigenetic progression of later development is in dramatic contrast to fertilization and pre-implantation, where totipotency is rapidly established from specialized gametes. Notably, the majority of epigenetic events unique to early mammalian development are characterized in mouse, and only the behavior of DNA methylation has been described on the genomic scale. Here, we present the genome-scale dynamics of DNA methylation in early human development. We confirm a strong global erasure of the paternal genome as observed in mouse and find that the methylation of most features in the early embryo are consistent between the two species. The specific targets for persistent regulation through DNA methylation however are strikingly different. We identify thousands of regions that are likely exclusively methylated on the maternal genome and are stable as parent-specific methylation signatures only during this phase. These sequences often represent discrete targets from similar imprint-like regions found in mouse. Repetitive elements also show a broader range of class specific behaviors in the human embryo and a larger degree of escape in human sperm. Our data provide the first complete genome-scale, base-resolution DNA methylation profile of the early human embryo and extend the basic mammalian model for DNA methylation in pre-implantation development from mouse to human.

2.2 Introduction

In mammals, methylation of the cytosine base is largely restricted to CpG dinucleotides, where it is stably propagated across DNA replication as a classically defined epigenetic modification[1]. Comprehensive mapping of DNA methylation at base pair resolution confirms a genomic distribution that is largely static throughout life and broadly depends on local CpG density[2]. Most intergenic sequence is CpG poor and stably methylated, whereas high density CpG islands found at promoters remain constitutively unmethylated[2]. In keeping with this model, dynamic changes to CpG methylation generally occur focally, are cell type or lineage specific, and accompany the activating or repressive activities of DNA binding factors[3]. This stable global distribution, disrupted by punctate dynamics that reflect cell-type-specific trans-factor activity, represents a foundational principle of epigenetic regulation that can be extended to the behavior of numerous modifications during normal developmental processes.

In contrast to most transitions where chromatin changes are local and cellular potential is progressively restricted, two periods of mammalian development are characterized by dramatic, global epigenetic reprogramming: specification of primordial germ cells (PGCs) in the developing germ line and establishment of the totipotent embryo after fertilization[4]. Both processes co-occur with the reacquisition of molecular pluripotency and permit the establishment of *in vitro* propagating stem cells[3]. As opposed to PGC specification, where the majority of alleles are targeted equivalently, the epigenetic events of fertilization are characterized by unique asymmetries that reflect discrepancies between the two gametes. The oocyte genome exists in a chromatinized state with high transcriptional activity, while the sperm genome is nucleosome-depleted, protamine-compacted, and transcriptionally inert[4]. Upon fertilization, the condensed

paternal genome is rapidly rechromatinized by the maternal machinery to establish an epigenetically interpretable and independently functioning zygote[4].

While the absence of pre-existing nucleosomes necessitates genome-scale reprogramming, paternal DNA methylation is delivered *en masse* but is subsequently, and rapidly, erased. Using mouse as the classical model for early mammalian development, experiments conducted over the past several decades have quantified erasure kinetics and inferred specific targets[5,6]. However, genome-scale technologies have only recently permitted comprehensive description of the complete DNA methylation landscape during early embryogenesis[7,8]. Sequencing efforts have confirmed global disparities in the distribution and level of DNA methylation between the two gametes. The oocyte genome exhibits hypomethylation that more closely resembles the early embryonic landscape, while mouse sperm is comparatively somatic[7,8]. Additionally, the rate of demethylation is slower than previously expected with maternal methylation being highly stable throughout the first several divisions. After the dramatic paternal targeting events at fertilization, DNA methylation levels approach a global nadir in the early ICM, reaching a temporal blank slate before being globally re-established at implantation[8].

Though many methylated sequences are erased in the early embryo, several genomic features remain strikingly regulated through continued DNA methyltransferase activity[9-11]. Repetitive element classes show varying degrees of demethylation after fertilization but are continuously maintained at stable intermediate or high methylation levels while most of the landscape continues its passive decay[6,12]. For instance, several long terminal repeat (LTR) endogenous retroviral classes are strongly demethylated at fertilization and dynamically regulated upon

activation of the zygotic genome, whereas intracisternal-A type particles (IAPs), the prevailing LTR class II element in the murine genome, are robustly methylated throughout development[13-15]. Similarly, the transcriptional activity, retrotranspositional potential, and DNA demethylation level all correspond within different long interspersed nuclear element (LINE) classes. The broad spectrum of behaviors across different repetitive element classes represents a key feature of early embryogenesis, where repression counteracts promiscuous activity but new escape competent elements continually emerge.

Retained embryonic methylation can be exclusively maternal in origin, and indeed, the oocyte contributes several hundred differentially methylated regions (DMRs) that show retention in the early embryo before resolution to hypomethylation after implantation[7,8,16]. These oocyte-specific methylation signatures appear similar to the well characterized imprint control regions (ICRs) but rarely show parent-specific regulation after implantation[8]. In mouse, a large fraction of these pre-implantation specific maternal DMRs occur at CpG island containing promoters, possibly reflecting alternate transcription start site (TSS) activity and elongation through the embryonic promoter, as seen for *Dnmt1*[7,8]. While the current functions of these imprint-like elements remain unknown, they offer an expanded catalog to investigate the genomic attributes of oocyte-specific methylation and, possibly, the unique features specific to true ICRs that permit their retention through adulthood.

Continued dissection of the murine system has refined the general model for global DNA demethylation and pre-implantation development in mammals, but extension of these observations to other species has been limited to date. It cannot be immediately inferred if all

rules described in mice will be extendable to other species. For instance, the extensive activity of IAPs presumably led to the coevolution of counteractive targeting systems to ensure their constitutive hypermethylation, which is mediated by zinc finger protein-recruitment of Trim28 and the H3K9 methyltransferase SETDB1[17,18]. Given the divergence of endogenous retroelement sequences across species, the relationship between DNA methylation in the early embryo and repetitive element class cannot be easily transferred without direct investigation. Additionally, while many classical ICRs show high conservation across species, the conservation of maternal methylation at transient imprints is completely unknown[19]. While general observations from immunohistochemical or targeted bisulfite sequencing have suggested similarities across placental mammals, they have lacked the resolution necessary to directly ascribe regulatory equivalence of DNA methylation dynamics[20-22].

To address these questions, we have produced genome-scale, single basepair resolution maps of DNA methylation in human sperm and the pre-implantation embryo to characterize the human pre-implantation embryo methylome and to evaluate the conservation of developmental dynamics as they have been defined mouse. We find conservation of global hypomethylation and similar instances of repetitive element maintenance, though sperm-specific hypomethylation for recently emerging LTR sequences is exceedingly more common in human. Additionally, we identify many regions of presumed maternally contributed methylation in the early embryo but a larger fraction resolve to hypermethylation than seen for similar imprint-like regions in mouse. Moreover, the identities of CpG island containing maternally contributed regions that resolve to hypomethylation are not conserved across species.

2.3 Results

2.3.1 The human pre-implantation embryo is globally hypomethylated

To generate a genome-scale view of early human DNA methylation dynamics, we collected human sperm, cleavage stage embryos, and blastocysts and carried out reduced representation bisulfite sequencing. We thawed and screened for morphologically normal E3 cleavage stage embryos and E6 blastocysts, assembling duplicate and triplicate data, respectively, using the minimum necessary input required for high reproducibility. For blastocysts, sample inputs ranged from 3-5 morphologically normal embryos while, for the cleavage stage, we collected 19 embryos per replicate using embryos ranging from 4-12 cells that largely exhibited pre-compaction morphology. We collected motile sperm from 4 healthy donors (age 30-34) to produce our gametic timepoint. Methylation profiles of fetal tissues from the NIH Epigenomics Roadmap Initiative were used in order to compare to somatic cells representing the three germ layers. This time series is as comprehensive a representation of human pre-implantation development as can be feasibly assembled. On average, replicates showed a high reproducibility, capturing 1,753,958 CpGs of methylation data at 10x Coverage, and are sufficient for comparative analysis to our pre-implantation development timeline in mouse.

The global methylation profiles of human sperm and fetal tissues largely exhibit somatic methylation architectures, with strong bimodality as a function of local CpG density **(Fig. 2-1a)**. Both the cleavage stage and blastocyst embryos have hypomethylated profiles where low CpG density regions no longer predict high methylation levels. These DNA methylation profiles strongly resemble corresponding profiles in mouse, where sperm is similar to somatic tissues and pre-implantation embryos are characterized by diffuse hypomethylation across the majority of

sequences and feature classes. Compared to both human somatic tissue and mouse sperm, human sperm show far more intermediate methylation, which is also apparent within single replicates. This may represent regions of spermatozoan population heterogeneity.

2.3.2 DNA methylation dynamics are largely conserved between human and mouse

Similarities between the two species extend into methylation dynamics across developmental stages. The majority of paternal methylation is dramatically erased by the 8-cell stage, with the transition from late cleavage to blastocyst showing comparative stability (**Fig. 2-1b**). In the transition from early to late pre-implantation development, only a subtle decay is observed in the most highly methylated regions while the majority of blastocyst stage methylation values appear pre-established by late cleavage. Upon exit from pre-implantation, the vast majority of methylation is restored (**Fig. 2-1b**). Thus, from the global perspective, both human and mouse pre-implantation embryos exist in a state of transient global hypomethylation that is unique to this stage.

To gain insight into the regulatory behavior of DNA methylation across early human development, we segregated our data according to genomic context and clustered the methylation profiles of 100bp tiles that did not fall in repetitive elements (**Fig. 2-2**). Most paternally hypermethylated regions are completely demethylated in the early embryo (78% of 169,474 hypermethylated tiles), while a minority retain methylation throughout this stage (**Fig. 2-2a, clusters 1-4**). Alternatively, most paternally hypomethylated regions are constitutively maintained throughout development (**Fig. 2-2a, cluster 13**). Instances of sperm hypomethylated tiles that gain methylation are observed, and are either transiently methylated during

preimplantation exclusively, acquire sustained methylation during preimplantation, or gain methylation *de novo* after embryonic specification (**Fig. 2-2a, clusters 8-12**). Regardless of the early embryonic methylation state, most tiles have the same methylation levels in sperm and the somatic tissues (**Fig. 2-2a, horizontal bar chart**).

We examined each cluster for genomic feature enrichment and were able to infer several rules for where DNA methylation is present or dynamic in the early embryo. Regions that are hypermethylated in both sperm and somatic cells but unmethylated during preimplantation, i.e. transiently demethylated loci, represent a large proportion of tiles and are enriched for intergenic or intronic sequences, suggesting a background state of early embryonic hypomethylation for non-functional CpGs (**Fig. 2-2b, cluster 2**). Retained hypermethylation throughout pre-implantation, on the other hand, is more frequently intragenic and exonic, potentially indicating persistent maintenance as an attribute of gene regulation (**Fig. 2-2b, cluster 1**). Nearly all dynamics where sperm is hypomethylated enrich for CpG islands, though this enrichment is less apparent for clusters that show some methylation throughout life (**Fig. 2-2b, clusters 9-12**). Paternally hypomethylated regions that are intermediately methylated in the early embryo exist in two distinct scenarios: when intermediate methylation is retained in somatic cells as in imprinting control regions (ICRs), or when this signature is only transiently stable and lost upon embryonic specification (**Fig. 2-2a, clusters 10 and 11**). Intermediately methylated regions in sperm are mostly demethylated, suggesting that the impact of these regions as distinguishable, but variable, epialleles may be infrequent or cannot be captured from the analysis of embryo populations (**Fig. 2-2a, cluster 5-7**).

To assess the similarity of methylation changes to those observed in mouse, we compared the dynamics of regions falling within different clusters to their homologous murine loci. We used euclidean distance between methylation profiles of the two species to look for deviations in dynamic behavior (**Fig. 2-2c**). In general, regions that share methylation states in sperm and somatic cells, which represent the majority of tiles, have conserved dynamics (**Fig. 2-2c, clusters 1, 2, and 13**). Alternatively, regions specifically hypermethylated in human sperm and hypomethylated in somatic cells show less similarity, a phenomenon largely attributable to disparities between human and mouse sperm where human sperm are more methylated (**Fig. 2-2cd, clusters 3 and 4**). These sperm-specific hypermethylated regions, which are lost during early embryogenesis, are not only tissue-specific but also appear to be more frequently species-specific. Species-specific embryonic methylation is most prominent in regions that are hypomethylated in both sperm and somatic cells, but have intermediate methylation in the early embryo (**Fig. 2-2d, cluster 11**). These regions are selectively enriched for CpG islands and, to some extent, high CpG density promoters (HCPs), suggesting a divergence in subsets of promoters that function as transient maternal imprints (dissected in greater detail below).

2.3.3 Retrotransposon dynamics are class and family specific

The DNA methylation dynamics that occur during human preimplantation demonstrate strong global similarity and local conservation to mouse, with divergence predominantly limited to species-specific methylation patterns that appear to be gametic in origin. However, direct comparison cannot be made for repetitive elements, which are mostly species-specific. To address the equivalence of repetitive element regulation, we compared methylation dynamics between similar classes (**Fig. 2-3**). The pervasive intermediate methylation found in human

sperm is largely confined to SINE elements (**Fig. 2-3**). While SINEs are slightly more methylated in mouse, they are less methylated than LINEs and LTRs in their respective species. In human, SINEs decay to minimal values during preimplantation before resolving to hypermethylation in somatic cells – dynamics identical to those in mouse (**Fig. 2-3**). Unlike mouse where the majority of LTRs and LINEs are hypermethylated, methylation of these elements in sperm is surprisingly multimodal (**Fig. 2-3**). To investigate if sperm-specific DNA methylation escape is restricted to coherent subsets, we separated the elements in these retrotransposon classes into their families.

By and large, different behaviors are contained within families of LTRs. ERVL and MalR families belonging to the Class III LTRs are analogs in human and mouse (**Fig. 2-4**). They are predominantly hypermethylated in sperm, go to low methylation in the early embryo, and are methylated in somatic tissues (**Fig. 2-4**). Elements of the MalR family are rapidly expressed as part of zygotic genome activation in mouse[13,14]. Their similar methylation dynamics in human suggest they might be some of the first transcripts expressed in human as well. Human class II ERVK elements consist of at least two populations that can be distinguished by their demethylation kinetics, with one population appearing to demethylate more slowly through development (**Fig. 2-4**). Passive demethylation from cleavage to the blastocyst stage is also observed in the mouse ERVK family though they remain more highly methylated overall. In contrast to the Class II and III families, the Class I ERV1 family exhibits a bimodal methylation signature that includes the majority of unmethylated repetitive elements apparent in human sperm (**Fig. 2-4**). The methylation escape of ERV1 elements in sperm may represent a unique characteristic of human development as it is not observed in mouse, indicating that, while LTR

methylation shares some conservation between human and mouse pre-implantation development, there is also considerable divergence. In human, the fraction of LTR elements that resolve to complete hypomethylation in the early embryo is higher, suggesting that either more elements are capable of escaping the DNA methylation machinery or that fewer elements are actively regulated via DNA methylation during gametogenesis and pre-implantation development. Similar to regions that fall in non-repetitive genomic contexts, the species-specific difference observed in LTRs appears to originate in the gamete.

2.3.4 Evidence of adaptive escape within L1PA lineages

The DNA methylation landscape of human LINE elements in sperm is complex, consisting of several, distinct populations **(Fig. 2-3)**. We clustered LINEs according to their methylation profiles across embryonic development and observed striking separation according to class and family **(Fig. 2-5a)**. The methylation levels of elements in the L2 class were concordant in sperm and somatic tissue with similar behavior to background sequence over pre-implantation development **(Fig. 2-5a)**. This is consistent with their characterization as degenerate, ancient, and fixed **(Fig. 2-5c, clusters 1,3-4)**. Alternatively, more recently emerging L1 elements maintained higher levels throughout pre-implantation development, implying maintained repression **(Fig. 2-5ab)**[23].

The L1PA family is primate specific and emerged as a temporally linear phylogeny, with one element selectively radiating in evolutionary history before being usurped by an advantaged, adapted progeny **(Fig. 2-5d)**[24]. Multiple models have been posited to explain this mode of radiation, including the competitively advantaged utility of shared retrotranspositional machinery,

gain of transcriptional enhancers, or loss of repressive targeting[24]. Given the strong similarities in the promoter structure of historically adjacent elements, the L1PA phylogeny provides the opportunity to explore the relationship between DNA methylation and repetitive element behavior in a straightforward evolutionary context[24].

Two groups of L1PAs show DNA methylation escape during pre-implantation development (**Fig. 2-5e**). The more evolutionarily ancient of these groups, consisting of two closely related subfamilies L1PA10 and 8A, exhibits appreciable gametic escape, as seen in LTRs of the ERV1 class. This escape may either suggest exaptation of these elements after fixation or adaptation to escape detection by host genome defense mechanisms during both male gametogenesis and pre-implantation. Early development methylation escape also emerges in the transition from L1PA4 to L1PA3, with more ancient elements showing higher maintained methylation levels than those after L1PA3, including contemporary, human specific elements (L1HS) (**Fig. 2-5e**). Within the L1 class, we see a linear progression in pre-implantation regulation with more recent elements showing a greater degree of embryonic escape, supporting models where loss of repressive motifs may contribute to selective element radiation. This dynamic is analogous to that observed in mouse, where emerging L1Md_T and L1Md_Gf classes show more robust demethylation and zygotic transcription, as well as slower silencing kinetics upon implantation, in comparison to evolutionarily older LINE classes[8,25].

2.3.5 Maternal imprint-like regions represent equivalent features but divergent targets

Regions that are differentially methylated between the gametes are of particular interest because they may lead to allele-specific expression patterns that persist throughout life, as they do for canonical imprint control regions (ICRs)[19]. Differentially methylated regions (DMRs) that confer maternal-specific methylation also exist transiently through pre-implantation development and are enriched in CpG island containing promoters, which are usually hypomethylated[7,8,26]. To identify regions that likely contribute maternal-specific methylation, we searched for tiles with lower methylation in sperm than in the cleavage or blastocyst stage embryo. While this strategy cannot identify true maternal methylation, the limited *de novo* methylation observed during pre-implantation development supports their likely maternal origin[8]. We identified over 2800 100bp regions that demonstrated more than a 0.25 methylation difference between the early embryo and sperm (**Fig. 2-6a**). This method identified most known human ICRs[19].

We clustered the methylation profiles of these tiles to explore their resolution in somatic cells. Similar to mouse, the majority of these DMRs exist transiently in the pre-specified embryo and resolve to either hyper- or hypomethylation. In total, 22% are hypomethylated across all somatic tissues, 54% are completely hypermethylated, and 15% resolve to intermediate methylation, a set that includes many true ICRs (**Fig. 2-6a**). Tiles that resolve to hypermethylation are exclusively hypomethylated in sperm and represent of methylation pattern that is far more common in human than mouse. While this may be a consequence of human sperm's comparatively hypomethylated genomic profile, it does intimate a sizable discrepancy between the species where the oocyte, and not sperm, recapitulates constitutive somatic patterns for these regions in human. These two classes of transient imprint-like regions represent distinct types of regulation: a maternal-specific

methylation signature exclusive to the oocyte that resolves to hypomethylation, and a sperm-specific hypomethylation, where methylation in the oocyte and somatic tissue is concordant.

Regions in the two sets show a compelling asymmetry in their genomic location. Even though all DMRs predominantly fell in intragenic sequence, sperm-specific hypomethylated regions show a sizable repetitive element proportion while DMRs resolving to hypomethylation are more frequently found in CpG islands (CGIs), particularly at transcription start sites (**Fig. 2-6a**). Maternally contributed methylation at CGI-containing promoters are also present in mouse. To investigate the possible conservation of these regions, we directly searched for maternal-specific methylation in CGIs using the same strategy applied to tiles and identified 282 CGI DMRs, which included 4 known ICRs (**Fig. 2-6b**). Examination of these CGIs falling either at TSSs or within annotated genes shows only limited functional ontology, but there is an enrichment for alternative splicing and splice variants (DAVID, Benjamini P-value=0.0098)[27]. In mouse, a maternally contributed, transiently stabilized DMR overlaps with the promoter of the somatic isoform of Dnmt1, which presumably stabilizes the switch to an oocyte-specific isoform, Dnmt1-o[7,8]. It is unknown if this is a common function of transient maternal imprints in mouse, but the possibility that a conserved function may exist even in lieu of conserved targets is appealing.

We projected the CGI DMRs to mouse and found that 196 align (106 captured by mouse RRBS) and 85 of them (71 captured) overlap mouse CGIs. While patterns of somatic methylation are well correlated between the species, only ~23% of human maternally contributed CGIs are DMRs in mouse suggesting that these regions are largely species-specific (**Fig. 2-6b**). Sperm-specific hypomethylated CGIs are also not shared, with the majority displaying a consistent

methylation pattern between the gametes in mouse **(Fig. 2-6b)**. Interestingly, the sperm-specific hypomethylated CGIs, which have methylation in the oocyte, early embryo, and somatic tissues, align to fewer mouse CGIs possibly revealing the effect of methylation on genomic stability and CG density.

Methylated cytosines are more likely to be mutated than other bases because they can be deaminated to thymine leading to a mismatch that can be propagated after replication[28]. To investigate the effect of methylation on genomic stability, we compared the CpG density of CGIs for discrete sets that exhibit different patterns of methylation through development. There is a general trend showing that CGIs with some level of methylation have lower CpG density, supporting the hypothesis that CpG methylation increases the potential for stable mutation **(Fig. 2-6c)**[28]. Unexpectedly, somatically hypermethylated CGIs have lower CpG density than DMRs, even though DMRs are methylated in the early embryo, where mutation is more likely to be carried over through the ensuing germ line. However, in mouse, oocyte-specific methylation is acquired late in maturation after arrest in meiotic prophase, such that pDMRs that resolve to hypomethylation are only present for several divisions, limiting the window where mutations can be acquired[29]. Alternatively,, sperm-specific hypomethylated pDMRs show greater CpG density than constitutively methylated CGIs, possessing modest protection against mutation through maintained hypomethylation in the male germ line.

Sequences that are genetically unstable will likely show higher variability in the population. Using single nucleotide polymorphisms (SNPs) annotated from the 1000 Genomes project, we measured the frequency of SNPs for different embryonic dynamics and discovered a trend

correlating methylation behavior to SNP density (**Fig. 2-6d**)[30]. The relationship between SNP density and methylation is inversely correlated to the relationship between CpG density and methylation, with somatically hypermethylated CGIs again showing the highest levels of mutation. Therefore, the relationship between genomic stability and embryonic methylation is subtle but consistent across two independent assessment methods. The presence of embryonic methylation appears to have some effect on mutation, but is not as detrimental to the genome as regions that are methylated throughout life. This may represent either the protection from deaminating agents inherent to an unmethylated state or refined fidelity of DNA repair machinery in the gametes.

While the general behavior of maternal methylation suggests a high level of divergence, intriguing instances of conserved gametic methylation patterns are observed. For instance, Dnmt1 is a maternally contributed, transiently methylated DMR that is conserved between human and mouse, which strongly suggests conservation of an alternative oocyte specific isoform in human, though it has not yet been annotated[7,8]. In another case, the presence of DMRs are divergent, but co-occur in similar gene sets, as observed for the human DMR, GLIS3, and its mouse counterpart, Neurog3, a gene regulated downstream of GLIS3 activity[31,32]. Thus, while the apparent conservation of embryonic DMRs is complex, and the extent to which targets are shared is not especially high across species, many general attributes suggest that the regulatory mechanisms that lead to imprint-like signatures are conserved and function equivalently, as noted by the overlap found at several canonical targets, the similar enrichments for CpG island-containing TSSs, and the developmental restriction of these DMRs to the oocyte and early embryo.

2.4 Discussion

We generated genome-scale maps of DNA methylation in human sperm, cleavage stage, and blastocyst embryos to gain insight into the regulation of methylation patterns during human pre-implantation development. We find that the regulatory principles that define DNA methylation in mouse pre-implantation development are globally conserved though species-specific targets define regions of local divergence that often appear to be gametic in origin. This general rule applies to both unique or repetitive sequence context.

Human repetitive element methylation dynamics are largely consistent within classes and families, and exhibit both conserved and divergent behaviors when compared to equivalent classes in mouse. Human LTRs show a wide range of methylation kinetics in early development that is apparent at the class level. Moreover, element families within these classes exhibit unique regulatory behaviors including rapid demethylation (LTR13), passive demethylation (MER11B), static hypomethylation (LTRC12C), and static hypermethylation (subset of LTR12C and MER52A). The diversity of behaviors provides an opportunity to search for characteristics that correspond to these distinct dynamics, such as CpG density, the presence or absence of regulatory motifs, and other controlling features that may be present in the underlying sequence. Successful identification of sequence features that correspond to embryonic or gametic methylation levels will provide a refined understanding of DNA methylation escape and may lead to new insights into the mechanism of *de novo* methylation targeting and, potentially, how these mechanisms adapt to recognize emerging, active retrotranspons. The ERV1 element LTR12C is particularly interesting, displaying both gametic and embryonic DNA methylation escape. Its coverage in our genome-scale library implies that it is prevalent in the genome in a

context that may not be extremely depleted for CpGs[33]. Its prevalence may reflect its activity, corresponding to its successful evasion of gametic methylation targeting and continued propogation, or it may be exapted. In support of the latter hypothesis, LTR12C has been shown to affect splicing, producing an isoform of the CHM gene that is unique to cancer tissue[34]. The exploration of its conservation in the human population and other species, as well as a refined dissection of sequence divergence as it relates to embryonic methylation within this family, may provide clues to answer this question.

Similar to the human LTRs, the LINE L1PA family showcases a range of methylation dynamics during early development, including two groups that exhibit early embryonic methylation escape. The structure of the L1PA phylogeny makes it conducive to exploring the relationship between DNA methylation, retrotransposon activity, and evolution[24]. Evasion of embryonic DNA methylation emerges at a distinct branch in the phylogeny with more ancient elements showing higher methylation than evolutionarily younger ones. A cursory examination of consensus 5' UTRs of L1PA elements revealed a ~120bp deletion in the "young" group when compared to the "ancient" group. The sequence deletion may potentially contain repressor motifs that targets methylation to ancient elements, and warrants explicit examination.

The phenomenon of maternally contributed differentially methylated regions (DMRs) that are transiently stabilized during early development appears to be conserved between human and mouse, though the identity of targets are mostly species-specific. Moreover, the functional relationship between these regions has been difficult to elucidate, though a large proportion of human DMRs were annotated as splice variants. Indeed, the conserved DMR, Dnmt1, is an

example of alternative splicing regulation through this mechanism[7,8]. Intriguingly, targeting for this methylation signature may not occur at the level of genes but may reflect higher levels of cellular and genomic organization, as evidenced by the regulation of the mouse DMR, Neurog3, by the human DMR, GLIS3. The seeming necessity for pathways to be regulated through this transient methylation mechanism strongly implies a coherent function. Proper methylation of the imprinting control regions is important for development of the extra-embryonic tissues[35]. Additionally, the human DMR, GLIS3, is imprinted in placenta[36]. The potential role of DMRs in extra-embryonic tissues should be explored.

Our study of DNA methylation dynamics in early human development extends the mammalian model from mouse to human for this interesting phase. Much of the regulatory phenomena was conserved between the species but we still understand surprisingly little about the mechanism and function of these methylation patterns. We have now identified the targets of DNA methylation during early embryogenesis and can move forward in understanding the consequences of this regulation.

2.5 Figures

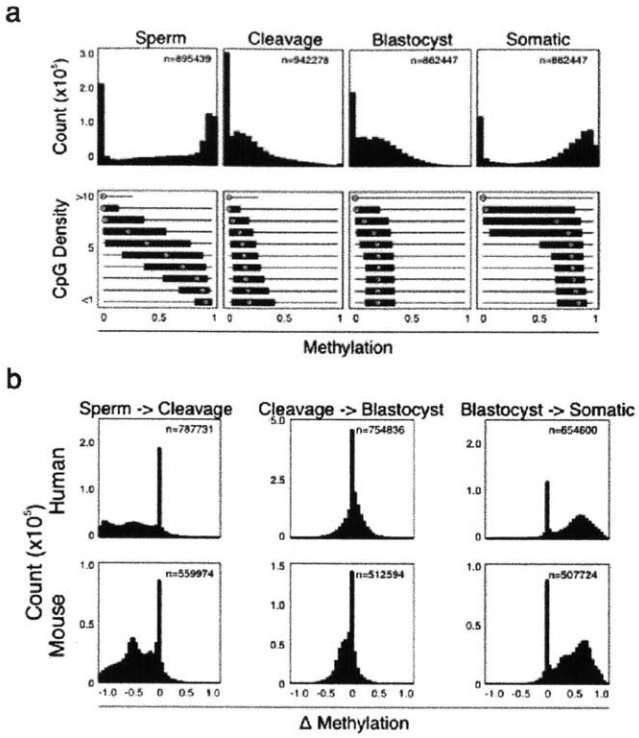

Figure 2-1: DNA methylation distribution of human pre-implantation embryos. **(a)** Histogram of DNA methylation values (X-axis) across 100bp tiles for human sperm, pre-implantation embryos, and somatic cells. Sperm exhibit a canonical bimodal pattern that is globally lost and reacquired as embryos transition through pre-implantation (top). The distribution (as box plots) of methylation values at different local CpG densities. The somatic sample shows the canonical distribution of DNA methylation that is largely reflective of local CpG density (bottom). Bulls-eye indicates the median, edges the $25^{th}/75^{th}$ percentile and whiskers the $2.5^{th}/97.5^{th}$ percentile. **(b)** Histograms of the change in methylation (**Δ**methylation) for each captured 100bp region over sequential developmental transitions, with comparable mouse data included to highlight the similarities of these dynamics. Namely, the paternal genome is largely demethylated by late cleavage, after which DNA methylation is largely stable into the blastocyst. Embryonic somatic methylation pattern is reset upon exit from pre-implantation.

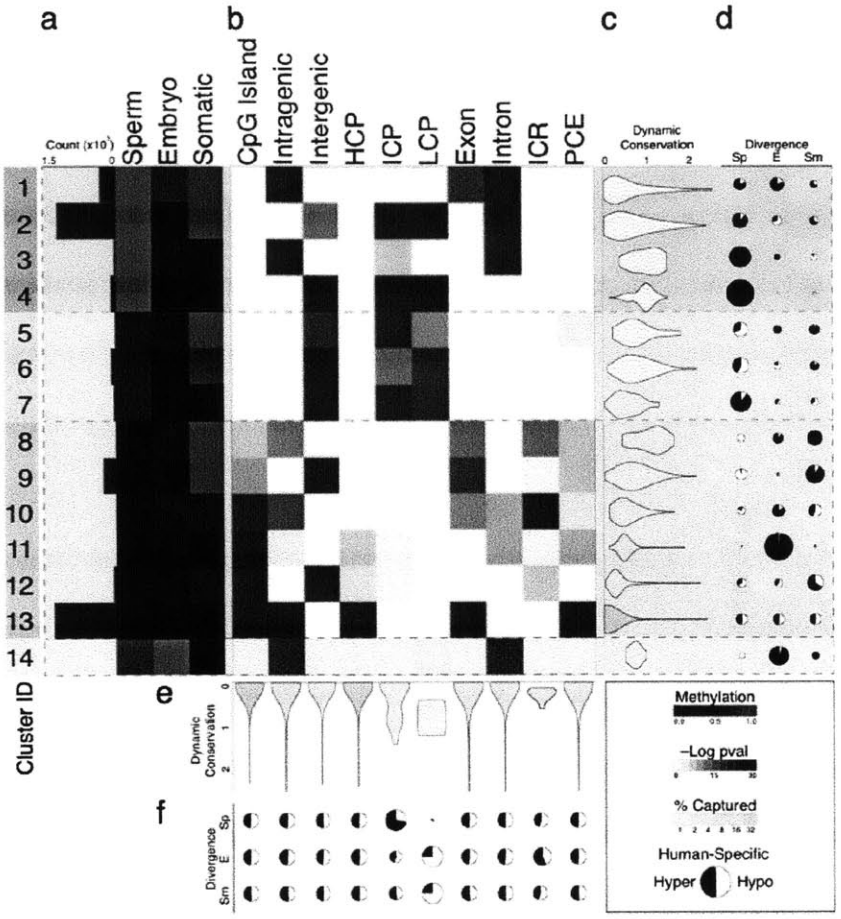

Figure 2-2: DNA methylation dynamics for non-repetitive sequence in early human development. **(a)** Heatmap of methylation patterns in early development ordered according to sperm methylation. The number of 100bp tiles that fall into each cluster is represented by the bar chart left of the heatmap. The majority of tiles fall in clusters 2 and 13 where sperm and somatic tissues have the same methylation levels. **(b)** Genomic feature enrichment for clusters. The –log p-value from the hypergeometric enrichment test. **(c)** Violin plot of the Euclidean distance distribution for clusters between human and mouse methylation profiles. Values close to one indicate conservation of methylation dynamics between the species. Color indicates the proportion of human tiles that could be aligned in mouse and are captured by RRBS. **(d)** Pie charts representing the methylation differences between human and mouse for each developmental stage. The radius of the pie chart is scaled according to the proportion of tiles in the cluster that show the largest difference between the species in that developmental stage. For example, a large pie chart in Sperm indicates that the divergence in (c) can be explained by the difference between sperm samples. Black indicates the fraction of tiles that are more highly methylated in human than mouse, and white indicates the fraction of tiles that are more highly methylated in mouse than human. **(e)** Violin plot of Euclidean distance distribution for genomic features. **(f)** Pie charts as in (e).

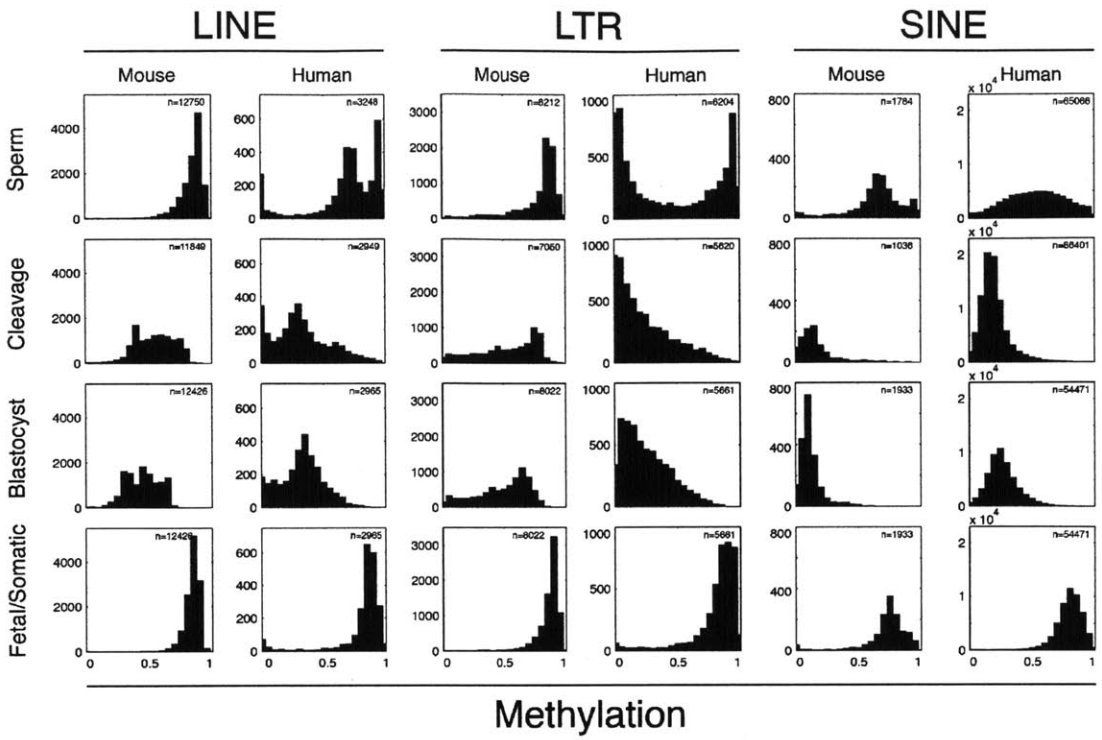

Figure 2-3: Retrotransposon DNA methylation across human and mouse development show conserved and divergent profiles. Histogram of methylation values across repetitive elements for long interspersed nuclear elements (LINEs), long terminal repeats (LTRs), and short interspersed nuclear elements (SINEs). In general, retrotransposons are highly methylated in the Sperm before being demethylated to varying degrees in the early embryo. All classes remethylate to high levels in somatic tissues.

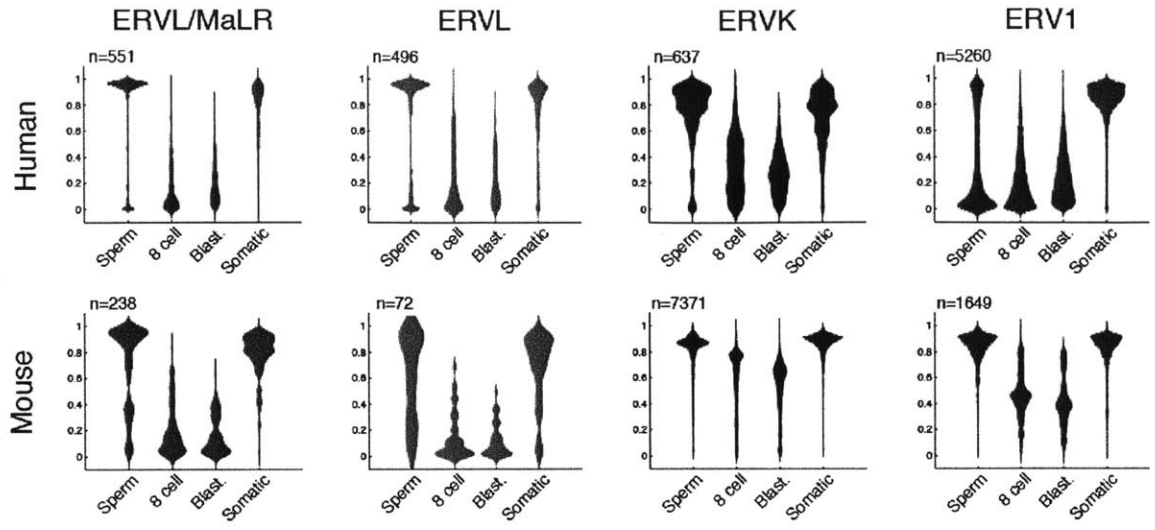

Figure 2-4: DNA methylation dynamics for LTR classes in human and mouse. Violin plots of methylation distributions for different LTR classes in human and mouse. ERVL/MaLR and ERVL are class III LTRs and have similar dynamics in both human and mouse. The human class II ERVK LTRs likely consist of at least two populations that have different demethylation kinetics over development. Mouse ERVKs show only modest demethylation through pre-implantation embryos. Class I ERV1 display divergent methylation dynamics between the species. A large proportion of human ERV1 elements are hypomethylated in sperm and exhibit gametic and embryonic DNA methylation escape that may indicate activity.

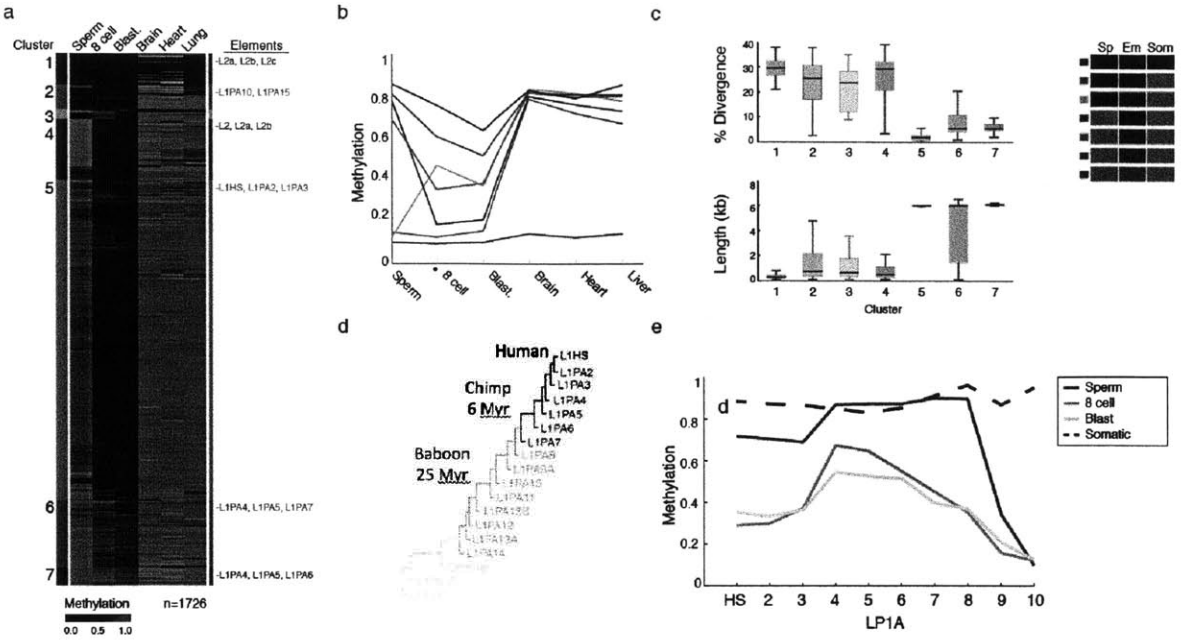

Figure 2-5: Human LINE methylation dynamics follow class and family classifications. **(a)** Heatmap of clustered methylation profiles for LINEs show a strong relationship between class and family classification and DNA methylation profiles. L2 elements consist of clusters 1 and 3 (top of heatmap), and L1 elements are in the remaining clusters. **(b)** Line plot of average DNA methylation profile for each cluster. The clusters have varying levels of methylation in the pre-specified embryos (8 cell, Blast) but display high methylation in somatic cells. The one exception is cluster 1 which represents L2 elements which are ancient, degenerate, and fixed. **(c)** Distribution (as a box plot) of the % divergence from the consensus sequence for the elements in each cluster (top). Distribution (as a boxplot) of the element lengths in each cluster (bottom). Elements that show high divergence from the consensus and have short length are likely older and inactive. **(d)** Phylogeny of the LINE L1PA family (adapted from ref 24). **(e)** Line plot of average DNA methylation profile for each L1PA subfamily. There are two groups of embryonic DNA methylation escape, in (L1PA8a, L1PA10) and (L1PAHS, L1PA2, L1PA3).

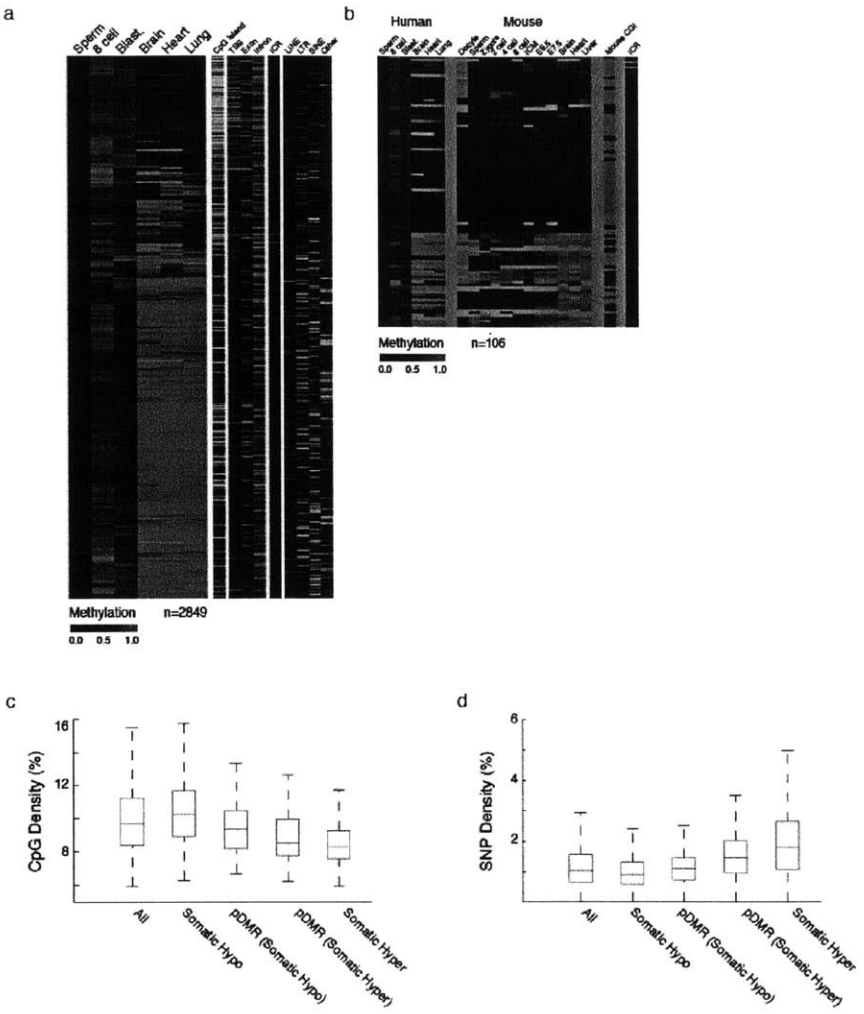

Figure 2-6: The majority of maternally contributed methylation is transiently stable only during pre-implantation. (a) Heatmap of clustered methylation profiles for 100bp tiles in human that were more lowly methylated in sperm than in the early embryos, indicating likely maternally contributed methylation information. Tiles that resolved to hypomethylation were enriched for CpG islands, a methylation signature that is also present in mouse. **(b)** Heatmap of clustered methylation profiles for maternally contributed methylation CpG islands in human (left) which can be aligned to mouse (right). The two columns on the far right indicate overlap of the aligned mouse region to an annotated mouse CpG island, and annotation as a known imprinting control region, respectively. **(c)** Box plot of CpG density for discrete sets of DNA methylation dynamics in early development show a relationship between the presence of methylation during development and lower CpG density. **(d)** Box plot of single nucleotide polymorphism (SNP) density for sets in (c).

2.6 References

1 Feng, S., Jacobsen, S. E. & Reik, W. Epigenetic reprogramming in plant and animal development. *Science* **330**, 622-627, doi:10.1126/science.1190614 (2010).

2 Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes & development* **25**, 1010-1022, doi:10.1101/gad.2037511 (2011).

3 Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nature reviews. Genetics* **14**, 204-220, doi:10.1038/nrg3354 (2013).

4 Reik, W., Dean, W. & Walter, J. Epigenetic reprogramming in mammalian development. *Science* **293**, 1089-1093, doi:Doi 10.1126/Science.1063443 (2001).

5 Santos, F., Hendrich, B., Reik, W. & Dean, W. Dynamic reprogramming of DNA methylation in the early mouse embryo. *Dev Biol* **241**, 172-182, doi:Doi 10.1006/Dbio.2001.0501 (2002).

6 Wossidlo, M. *et al.* Dynamic link of DNA demethylation, DNA strand breaks and repair in mouse zygotes. *Embo J* **29**, 1877-1888, doi:Doi 10.1038/Emboj.2010.80 (2010).

7 Smallwood, S. A. *et al.* Dynamic CpG island methylation landscape in oocytes and preimplantation embryos. *Nature genetics* **43**, 811-814, doi:10.1038/ng.864 (2011).

8 Smith, Z. D. *et al.* A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature* **484**, 339-344, doi:10.1038/nature10960 (2012).

9 Ooi, S. K. T. *et al.* DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature* **448**, 714-U713, doi:Doi 10.1038/Nature05987 (2007).

10 Dong, K. B. *et al.* DNA methylation in ES cells requires the lysine methyltransferase G9a but not its catalytic activity. *Embo J* **27**, 2691-2701, doi:10.1038/emboj.2008.193 (2008).

11 Leung, D. C. *et al.* Lysine methyltransferase G9a is required for de novo DNA methylation and the establishment, but not the maintenance, of proviral silencing. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 5718-5723, doi:10.1073/pnas.1014660108 (2011).

12 Iqbal, K., Jin, S. G., Pfeifer, G. P. & Szabo, P. E. Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. *P Natl Acad Sci USA* **108**, 3642-3647, doi:Doi 10.1073/Pnas.1014033108 (2011).

13 Beraldi, R., Pittoggi, C., Sciamanna, I., Mattei, E. & Spadafora, C. Expression of LINE-1 retroposons is essential for murine preimplantation development. *Molecular reproduction and development* **73**, 279-287, doi:10.1002/mrd.20423 (2006).

14 Kigami, D., Minami, N., Takayama, H. & Imai, H. MuERV-L is one of the earliest transcribed genes in mouse one-cell embryos. *Biology of reproduction* **68**, 651-654 (2003).

15 Lane, N. *et al.* Resistance of IAPs to methylation reprogramming may provide a mechanism for epigenetic inheritance in the mouse. *Genesis* **35**, 88-93, doi:10.1002/gene.10168 (2003).

16 Proudhon, C. *et al.* Protection against de novo methylation is instrumental in maintaining parent-of-origin methylation inherited from the gametes. *Molecular cell* **47**, 909-920, doi:10.1016/j.molcel.2012.07.010 (2012).

17 Rowe, H. M. *et al.* KAP1 controls endogenous retroviruses in embryonic stem cells. *Nature* **463**, 237-240, doi:10.1038/nature08674 (2010).

18 Wiznerowicz, M. *et al.* The Kruppel-associated box repressor domain can trigger de novo promoter methylation during mouse early embryogenesis. *The Journal of biological chemistry* **282**, 34535-34541, doi:10.1074/jbc.M705898200 (2007).

19 Woodfine, K., Huddleston, J. E. & Murrell, A. Quantitative analysis of DNA methylation at all human imprinted regions reveals preservation of epigenetic stability in adult somatic tissue. *Epigenetics & chromatin* **4**, 1, doi:10.1186/1756-8935-4-1 (2011).

20 Beaujean, N. *et al.* Non-conservation of mammalian preimplantation methylation dynamics. *Current biology : CB* **14**, R266-267, doi:10.1016/j.cub.2004.03.019 (2004).

21 Dean, W. *et al.* Conservation of methylation reprogramming in mammalian development: Aberrant reprogramming in cloned embryos. *P Natl Acad Sci USA* **98**, 13734-13738, doi:Doi 10.1073/Pnas.241522698 (2001).

22 Young, L. E. & Beaujean, N. DNA methylation in the preimplantation embryo: the differing stories of the mouse and sheep. *Animal reproduction science* **82-83**, 61-78, doi:10.1016/j.anireprosci.2004.05.020 (2004).

23 Bannert, N. & Kurth, R. The evolutionary dynamics of human endogenous retroviral families. *Annual review of genomics and human genetics* **7**, 149-173, doi:10.1146/annurev.genom.7.080505.115700 (2006).

24 Khan, H., Smit, A. & Boissinot, S. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome research* **16**, 78-87, doi:10.1101/gr.4001406 (2006).

25 Fadloun, A. *et al.* Chromatin signatures and retrotransposon profiling in mouse embryos reveal regulation of LINE-1 by RNA. *Nature structural & molecular biology* **20**, 332-338, doi:10.1038/nsmb.2495 (2013).

26 Bourc'his, D., Xu, G. L., Lin, C. S., Bollman, B. & Bestor, T. H. Dnmt3L and the establishment of maternal genomic imprints. *Science* **294**, 2536-2539, doi:10.1126/science.1065848 (2001).

27 Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* **4**, 44-57, doi:10.1038/nprot.2008.211 (2009).

28 Cohen, N. M., Kenigsberg, E. & Tanay, A. Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. *Cell* **145**, 773-786, doi:10.1016/j.cell.2011.04.024 (2011).

29 Saitou, M., Kagiwada, S. & Kurimoto, K. Epigenetic reprogramming in mouse pre-implantation development and primordial germ cells. *Development* **139**, 15-31, doi:10.1242/dev.050849 (2012).

30 Genomes Project, C. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:10.1038/nature09534 (2010).

31 Kim, Y. S. *et al.* Glis3 regulates neurogenin 3 expression in pancreatic beta-cells and interacts with its activator, Hnf6. *Molecules and cells* **34**, 193-200, doi:10.1007/s10059-012-0109-z (2012).

32 Yang, Y. *et al.* The Kruppel-like zinc finger protein GLIS3 transactivates neurogenin 3 for proper fetal pancreatic islet differentiation in mice. *Diabetologia* **54**, 2595-2605, doi:10.1007/s00125-011-2255-9 (2011).

33 Meissner, A. *et al.* Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic acids research* **33**, 5868-5877, doi:10.1093/nar/gki901 (2005).

34 Jung, Y. D. *et al.* Quantitative analysis of transcript variants of CHM gene containing LTR12C element in humans. *Gene* **489**, 1-5, doi:10.1016/j.gene.2011.09.001 (2011).

35 Niemann, H., Tian, X. C., King, W. A. & Lee, R. S. Epigenetic reprogramming in embryonic and foetal development upon somatic cell nuclear transfer cloning. *Reproduction* **135**, 151-163, doi:10.1530/REP-07-0397 (2008).

36 Barbaux, S. *et al.* A genome-wide approach reveals novel imprinted genes expressed in the human placenta. *Epigenetics : official journal of the DNA Methylation Society* **7**, 1079-1090, doi:10.4161/epi.21495 (2012).

Chapter 3

The mouse ooplasm confers context-specific reprogramming capacity

This work was first published as:

MM Chan*, ZD Smith*, D Egli*, A Regev, A Meissner. Mouse ooplasm confers context-specific reprogramming capacity. *Nature Genetics*, **44**(9):978-80, Sep 2012. *These authors contributed equally to this work.

MMC, ZDS, DE, and AM conceived and designed the study. DE performed SCNT, ZDS performed methylation profiling, and MMC performed all analysis. MMC, ZDS, AR and AM interpreted the data. MMC, ZDS and AM wrote the paper with input from the other authors.

Chapter 3:

Mouse ooplasm confers context-specific reprogramming capacity

3.1 Abstract

Enucleated oocytes have the remarkable ability to reprogram somatic nuclei back to totipotency. Here we investigate genome-scale DNA methylation patterns after nuclear transfer and compare them to the dynamics at fertilization. We identify specific targets for DNA demethylation after nuclear transfer such as germ-line associated promoters, as well as unique limitations that include certain repetitive element classes.

3.2 Results

Mammalian DNA methylation generally shows limited global dynamics except during pre-implantation and primordial germ cell development. The observed global demethylation of the paternal genome upon fertilization is mediated by the oocyte and is critical for establishment of totipotency and developmental competence[1]. An enucleated oocyte can reprogram the epigenetic and transcriptional identity of somatic cells through a procedure known as somatic cell nuclear transfer (SCNT), though successful reprogramming occurs at low efficiency and is likely affected in part by the retention of somatically-conferred epigenetic lesions[2,3]. To date, the oocyte's intrinsic capacity to erase somatic DNA methylation patterns to a true zygotic signature remains incompletely characterized. Here, we generated genome-scale single basepair resolution maps of DNA methylation from donor fibroblasts and SCNT reconstructed mouse embryos using reduced representation bisulfite sequencing (RRBS) [4] (**Fig. 3-1a**). We completed two independent rounds of nuclear transfer into enucleated BDF1 (C57/DBA F1) oocytes, each consisting of two biological replicates. We used both BDF1 x Cast hybrid and 129X1 inbred tail tip fibroblasts (referred to as Cast and X1, respectively) to serve as controls for low input RRBS by capturing over 10,000 hybrid SNPs and to ensure that dynamics observed were consistent across strain identity (**Fig. 3-1b-e**). Using our stringently collected samples and the first genome-scale measurement in any nuclear transfer experiment, we detected a low level (~15%) of the host oocyte genome (see **Methods**). This affected ~35% of loci in the X1 samples and likely ~13% of loci in the embryos using the hybrid donors. Given the complexity inherent to the protocol and the number of cells required, the presence of residual host DNA may be unavoidable; however, the low frequency with which it is sampled is unlikely to be consistent

across experiments. Accordingly, we present the SCNT embryo data without compensation for residual host DNA methylation.

We compared methylation profiles between the donor cells and reconstructed embryos to the methylation dynamics observed during fertilization [5]. DNA methylation patterns of donor fibroblasts and sperm exhibit a conventional somatic bimodality that depends upon relative CpG density. After nuclear transfer, a shift in the fibroblast methylation landscape resembles the demethylation that occurs within the paternal genome upon fertilization (**Fig. 3-2a**). Though many regions are affected in both processes, demethylation occurs at a smaller magnitude after SCNT (**Fig. 3-2b**). Globally, reconstructed embryos more closely resemble donor fibroblasts than the pre-implantation embryo or estimates of the paternal methylation pattern after fertilization (**Fig. 3-1f**). Importantly, SCNT embryos are more similar to each other regardless of experimental round or donor strain than they are to either fibroblasts or the early embryo, suggesting the majority of methylation changes conferred by SCNT are consistent across experiments (**Fig. 3-1g**). The difference in reprogramming response appears to be partly due to genomic context. For instance, we find that different repetitive element classes change by different magnitudes (**Fig. 3-2c**, top). While SINE elements appear similarly demethylated in both processes, LINEs and LTRs only slightly decrease or do not change after nuclear transfer (**Fig. 3-2c,** bottom, **Fig. 3-3a**). Upon closer inspection of LINE families, we found that methylation at L1Md_A elements remains almost completely static whereas the evolutionarily younger L1Md_T/Gf families appear slightly more dynamic [6,7] (**Fig. 3-2d**). The predominance of LINE and LTR element classes in the genome and their recalcitrance to demethylation could explain the striking retention of somatic methylation patterns after SCNT. Taken together, we conclude that repetitive elements, which represent a large proportion of demethylation events

during fertilization, appear more resistant to change when the ooplasm is confronted with a somatic nucleus. By the nature of the experiment we cannot rule out the possibility that some of the global dynamics observed in SCNT embryos may be due to the presence of residual host oocyte DNA. However, we observed similar demethylation at CpGs associated with Cast alleles compared to C57 in our hybrid fibroblast experiments, which can only result from reprogramming of the donor fibroblast genome (**Fig. 3-3b**). It is technically not possible to extend SNP analysis to repetitive elements, but the reasonable association between Cast exclusive demethylation dynamics and the C57 haplotype argue that demethylation is in fact an observable and measurable event during nuclear transfer.

We then compared methylation dynamics in promoters during both processes. Given the likely stochastic models for demethylation during SCNT and the potential effect of residual host DNA, we applied stringent criteria for identifying changing promoters such that hypermethylated promoters (>0.5) were required to change by > 0.2 in at least three replicates, change by >0.2 in donor-normalized X1 replicates, and could not have contradictory dynamics in Cast-tracked CpGs (see **Methods**). This strategy identifies the most consistently changing promoters during the nuclear transfer procedure but excludes promoters that change either less efficiently or are targeted less frequently for DNA demethylation. We then classified promoter dynamics as being either SCNT specific, shared with fertilization, or unique to fertilization (**Fig. 3-4a**). We identified a stringent set of 15 SCNT specific promoters, which include several genes that function during meiosis and are typically hypomethylated in both gametes and the early embryo (**Fig. 3-4b**). The specific demethylation observed in the SCNT embryos suggests the presence of defined, targeting factors within the ooplasm that ensure the unmethylated status of these sites (**Fig. 3-4c**) [8]. We observed that approximately two thirds of fertilization specific demethylated

targets retained considerable DNA methylation after nuclear transfer, indicating that only regions in certain contexts are equivalently demethylated in both processes (**Fig. 3-3c,d**). These artifacts likely represent somatically retained epigenetic information that could affect embryonic development. We also investigated 102 previously identified differentially methylated promoters that are hypermethylated in the oocyte and transiently remain methylated on the maternal allele during pre-implantation (**Fig. 3-3e**) [5]. The set is unmethylated in fibroblasts and remains so after nuclear transfer, representing another example of the inequivalence between SCNT reconstructed and true embryos in which maternally encoded regulatory information is not conferred.

3.3 Discussion

The demethylation events observed after nuclear transfer are similar to those at fertilization, though often at lower magnitudes, which may indicate less efficient or stochastic targeting. In turn, this may affect the number of SCNT embryos that successfully reactivate sufficient developmental loci upon zygotic activation[9]. Certain repeats, such as LTRs and L1Md_As, appear completely refractory to demethylation in nuclear transfer and may be protected within the epigenetic context of somatic cells. Demethylation of the paternal genome at fertilization is accompanied by global rechromatinization[1], which may provide a unique window of opportunity for parasitic genomic elements to either initiate demethylation or escape methylation machinery. It has been observed, for instance, that cytoplasmic injection of chromatinized round spermatids generates live embryos at equivalent rates to injection of terminally differentiated, protamine-compacted spermatozoa but does not co-occur with an apparently equivalent global erasure of DNA methylation[10]. Our genome-scale profiling strategy confirms previous locus specific bisulfite sequencing and global immunostaining data which

have shown that DNA methylation after nuclear transfer is not as dramatic as that observed in the paternal genome[7]. Presumably, the interplay between histone exchange, transcription, and active demethylation may limit the magnitude and targets for DNA demethylation observed after nuclear transfer or lead to aberrant signatures[11]. Tet3 mediated hydroxymethylation may also be more robustly targeted upon fertilization than after nuclear transfer, where it is not specifically recruited to a single pronucleus but rather distributed at restricted levels across the entire diploid genome[12,13]. It is interesting that demethylated loci after nuclear transfer resemble the genomic features and promoter classes enriched for hydroxymethylation within mouse embryonic stem cells, namely repetitive LINE families and germ-line gene promoters, which suggests that at least a portion of the demethylation events involved with induced pluripotency may be specifically targeted[14,15]. Technical improvements to other genomic profiling strategies should soon identify the targets and dynamics of histone deposition, as well as the distribution of hydroxymethylcytosine[15], and should comprehensively define the full reprogramming capacity of the mammalian ooplasm.

3.4 Methods

3.4.1 Preparation of samples and genome-scale libraries

Reduced representation and alignment was performed essentially as described [3]. Gamete and zygotic data sets were taken from a parallel study where they contribute to a complete pre-implantation timeline [2] but in this study were used exclusively as an *in vivo* counterpart to examine the global epigenetic consequences of NT on somatic cells.

Our comparative oocyte-based reprogramming data set consisted of donor fibroblasts and those same donor fibroblasts submitted to nuclear transfer. Donor fibroblasts were isolated of tail tips from two BDF x CAST/EiJ or two 129X1/SvJ donor males. BDF1 oocytes for NT were obtained 13 h post hCG injection and the spindle was removed under 5 μg/ml cytochalasin B. Donor fibroblasts were injected into enucleated oocytes and the meiosis I polar body was eliminated as a 1x genomic contaminant via manual XY Clone Infra-red laser assisted biopsy (Hamilton Thorne), as were meiosis I/II polar bodies in fertilized zygotes. Activation was done 1-3 hours post injection by 1h incubation in 1 mM $SrCl_2$ in Ca-free MZCB supplemented with cytochalasin B, followed by 4h incubation in cytochalasin B only. Embryos were then cultured in KSOM-AA (Chemicon) and collected 12-14 hours after the onset of $SrCl_2$ (hours post activation, hpa) in accordance with later zygotic stages, and before completion of the first interphase. Zygotes used for comparison were staged within pronuclear stages 2 and 4, after the reported completion of active DNA demethylation by immunohistochemical detection of methylated cytosine[4].

3.4.2 Analysis of RRBS data

For 100 bp tiles, reads for the CpGs that were covered more than 5x were pooled and used to estimate the methylation level by taking the number of reads reporting a C, divided by the total

number of reads reporting a C or T. Outliers were removed using Dixon's Q-test with a confidence level of 90% for tiles that had more than 3 replicate values before calculating mean methylation values for a sample. The CpG density for a 100 bp tile is the average number of CpGs within 50 bp of contributing CpGs.

High, intermediate, and low CpG density promoter (HCP, ICP, LCP), and ICR annotations were taken from [1]. LINE, LTR, and SINE annotations were downloaded from the UCSC browser (mm9) RepeatMasker tracks. The methylation level of an individual feature is estimated by pooling read counts for CpGs within the feature that are covered more than 5x, and levels are reported if a feature contains ≥5 CpGs with such coverage.

3.4.3 Calculation of paternal methylation levels in zygote

Zygote methylation is the mean methylation of the maternal and paternal genomes:

Zygote=0.5(Oocyte)+0.5(Sperm)

Assuming methylation on the oocyte genome does not change in the zygote:

Zy(Sp)=(Zygote-0.5(Oocyte))/0.5= 2*Zygote – Oocyte

The variance for Zy(Sp) is then:

Var(Zy(Sp)) = 4*Var(Zygote) + Var(Oocyte)

Methylation values for Zy(Sp) are estimated using the mean methylation value of the zygote and oocyte samples. Predicted values <0 and >1 are set to their respective boundaries if that boundary is within one standard deviation for that locus.

We are able to track 565 methylation values of individual CpGs using SNPs at 10x paternal coverage in BDF1xX1 zygotes. The Pearson correlation coefficient between the inferred sperm in zygote value and the SNP tracked paternal methylation value is 0.8.

The inferred paternal methylation values in 2-cell embryos are calculated in the same manner by swapping 2-cell methylation values for zygote methylation values.

3.4.4 Estimation of residual host oocyte DNA in NT embryos

We used the proportion of SNPs that showed any recipient oocyte allele as an estimate for the proportion of loci affected by contamination in 129X1 NT embryos. SNPs affected by contamination are represented by the left peak of the bimodal log odds distribution and contain 40% and 32% of the SNPs for each replicate respectively (**Fig. 3-1c**). To estimate the magnitude of contamination, we binned all SNPs that showed any contamination by their coverage into 5x intervals. As coverage increases, we should achieve a better estimate of the true contamination level and indeed, we see contamination leveling out around 16% and 14% for each replicate, respectively. This works out to 0.4*0.16=0.064 total contamination. We believe the true level of contamination is lower as this estimate is confounded by amplification noise. Estimating contamination in BDF1/Cast NT embryos is more difficult as the recipient BDF1 oocyte cannot be readily distinguished from the donor genome. To estimate the proportion of loci affected by contamination, we assume that the distribution for an ideal, uncontaminated sample follows the same shape but is shifted such that its peak is over a log odds ratio of zero. Then, the estimate of loci affected is the density of SNPs that remain with a log odds ratio less than zero after the density of the ideal distribution has been subtracted. This is ~13% for both BDF1/Cast NT embryo replicates. To estimate the average magnitude of contamination for these samples, we take the weighted mean level of contamination for the affected loci, which is ~33% for both replicates. This results in an overall contamination of 0.13*0.33=0.043.

3.4.5 Analysis of dynamic genomic feature sets

In order to equalize the statistical power across feature sets that had different numbers of elements, we subsampled the feature set with replacement 1000 times and took the average p-value from the Wilcoxan signed rank test on the mean value across replicates using 150 elements, which is the approximate coverage of SINE elements. We used the Bonferroni correction with FDR=0.05 to identify significantly changing features in each comparison. We considered accounting for potential oocyte DNA contamination by modifying the fibroblast methylation levels to include the appropriate proportion of oocyte methylation (thereby creating the appropriate background to measure NT embryo dynamics against) but decided that both overestimation or underestimation of contamination levels would lead to biased results.

3.4.6 Analysis of dynamic promoters

Promoters are considered dynamic if they show a change >0.2 between the means of the replicates in fertilization (Sp – Zy(Sp)) or NT (Fib– NT(Fib)). A promoter is robustly changing during fertilization if the p-value from a one-tail t-test comparing the Zy(Sp) to the sperm values is significant after multiple hypothesis test correction using the Benjamini-Hochberg method. The samples in NT are paired such that each donor has a matching NT embryo except in the case of the two 129X1 replicates where the fibroblasts from the same animal were used for both NT embryos (X1 tail tip replicate 1). Nonetheless, we assume that the two replicates are independent because the samples were cultured independently and subjected to separate rounds of nuclear transfer. Promoters must satisfy three criteria in order to be considered robustly changing in NT. First, promoters must change >0.2 in all replicates measured and must be present in at least three replicates when the change is calculated on raw values. For a given replicate, a change >0.2 corresponds to an empirical p-value of 0.0372 - 0.0479 when the null distribution is determined

by scrambling the relationship between promoters in each donor-NT embryo pair. The null distribution for each pair was generated 1000 times and the average proportion of promoters >0.2 was taken to determine the empirical p-value. The requirement of change in three replicates corresponds to a p-value of 9.7*10^-5. Unfortunately, this would not be significant after multiple hypothesis test correction but we did not want to limit our results due to missing values. This results in 49 promoters. Secondly, we were concerned that oocyte contamination would bias results towards demethylation so we require promoters to change >0.2 in contamination-normalized 129X1 replicates. They must change in all replicates measured and must be present in one of the replicates. To normalize the data for comparison, we modified all fibroblast promoter methylation levels to include an appropriate amount of oocyte contamination (16% and 14% in each replicate respectively). For example, Fibroblast(normalized)= (1-0.16)*Fibroblast + 0.16*Oocyte.

An empirical p-value for change >0.2 in contamination-normalized comparisons is 0.0344-0.0412. We now have 31 promoters. Finally, if Cast allelo-typed CpG methylation data at 5x exists for the promoter, then it must also show the same dynamic. This supported two promoters and eliminated two other promoters. Promoters were then assigned to three categories (NT specific, fertilization specific, or shared) depending on their dynamics. In order to be shared, a promoter must be robustly changing in both NT and fertilization. DAVID was used to ascribe functional enrichment [5].

3.4.7 Identification of SNPs

Known SNPs between 129X1 and BDF1 (C57BL/6N × DBA/2) and Cast and BDF1 were taken from mouse genome resources[6,7]. The sets were filtered such that SNPs that fell into the following categories were removed: (1) SNPs that had inconsistent entries for the same position;

(2) SNPs not trackable by RRBS (C/T or A/G); (3) SNPs between C57BL/6N and DBA/2; and (4) SNPs that were not covered by X1 and BDF1, or Cast and BDF1 in an *in silico* digest. The log odds ratio [$\log_2$(X1 count +0.01/C57 count+0.01)] was calculated for each SNP that was covered in the data set.

3.4.8 Parent-of-origin methylation tracking

Reads were segregated into either Cast or BDF1 according to SNP type, and CpG methylation levels were called in the same manner described above. SNP normalized methylation values (**Fig. 3-1d**) are the average of the methylation values derived from each haplotype.

3.5 Figures

a

Sample	n	Input	CpGs (1x)	Unique (1x)	Mean Cov (1x)	CpGs (10x)	Unique (10x)	Mean Cov (10x)	r
Fib (Cast)	2	50 ng	39884177	1856379	21.5	37341677	1187179	31.4	0.95
Fib (X1)	2	20 ng	36205062	1590138	22.8	34473576	1198195	28.8	0.95
NT_{Fib} (Cast)	2	100 NT embryos	28560428	1780280	16.1	25054769	859503	29.1	0.81
NT_{Fib} (X1)	2	100 NT embryos	26202353	1393633	18.8	23698820	799314	29.7	0.83
Oocyte	4	275 gametes	54146487	1064353	53.4	52951207	753664	73.3	0.73
Sperm	5	50 ng	48968469	1621643	30.5	46895112	1050372	44.3	0.90
Zygote	5	60 embryos	44325454	1657098	26.8	41718426	903862	45.5	0.62
2-cell	5	100 cells	37394204	1623550	22.9	34888729	954075	36.2	0.73

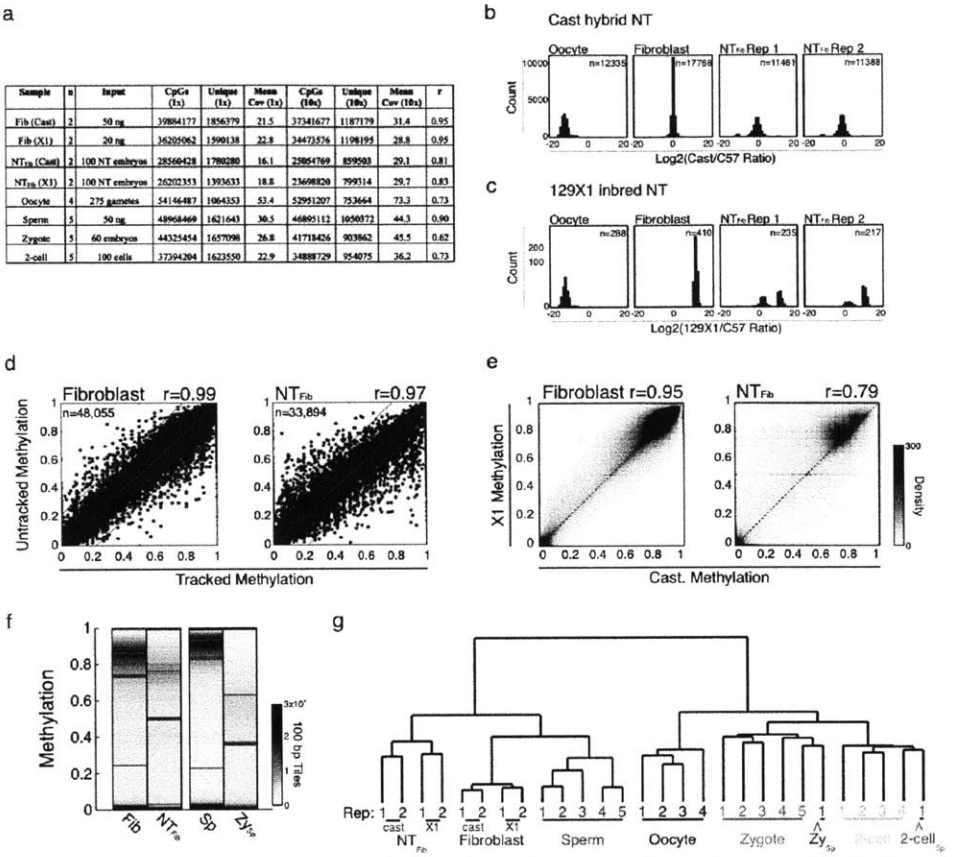

Figure 3-1: Genome-scale methylation profiling of nuclear transfer embryos. **(a)** CpG coverage captured by reduced representation bisulfite sequencing (RRBS). Mean replicate Pearson correlation coefficient (r) is included for each sample. Oocyte, sperm, zygote, and 2-cell data is from Ref 2. **(b)** Log odds ratio histograms for allelic frequencies for C57/B6 and CAST/EiJ SNPs for experimental round 1 with BDF1xCast hybrid tail tip fibroblasts. The shift of the log odds ratio away from 0 suggests possible recipient oocyte contamination (~13%) in the NT_{Fib} sample. n is the number of SNPs captured for each sample. **(c)** Log odds ratio histograms for allelic frequencies for C57/B6 and 129X1 SNPs for experimental round 2 using 129X1 inbred tail tip fibroblasts. The bimodal distribution shows recipient oocyte contamination affected 40% and 32% of loci in each NT_{Fib} replicate, respectively. The level of contamination is low (~15%, see Methods). **(d)** Methylation values of single CpGs that can be allelotyped show a high concordance between the untracked calculation and the allele-normalized methylation value. **(e)** Scatterplot of methylation values for 100 bp tiles between donor fibroblasts and NT embryos. **(f)** Boxplot overlaid on a density heatmap showing the distribution of methylation values for 100 bp tiles for donor fibroblast, SCNT reconstructed embryos, sperm and the paternal value in zygote (Zy_{sp}). There is retention of more methylation after nuclear transfer than during fertilization. **(g)** Sample clustering using Euclidean distance on methylation values for 100 bp tiles shows that NT embryos most closely resemble, but are distinct from, donor fibroblasts. This trend holds regardless of the experimental round. Raw zygote values are used in clustering in addition to inferred Zy_{sp} values used in analysis. 2-cell and inferred paternal values for 2-cell embryos (2-cell_{sp}) cluster with their respective biological samples.

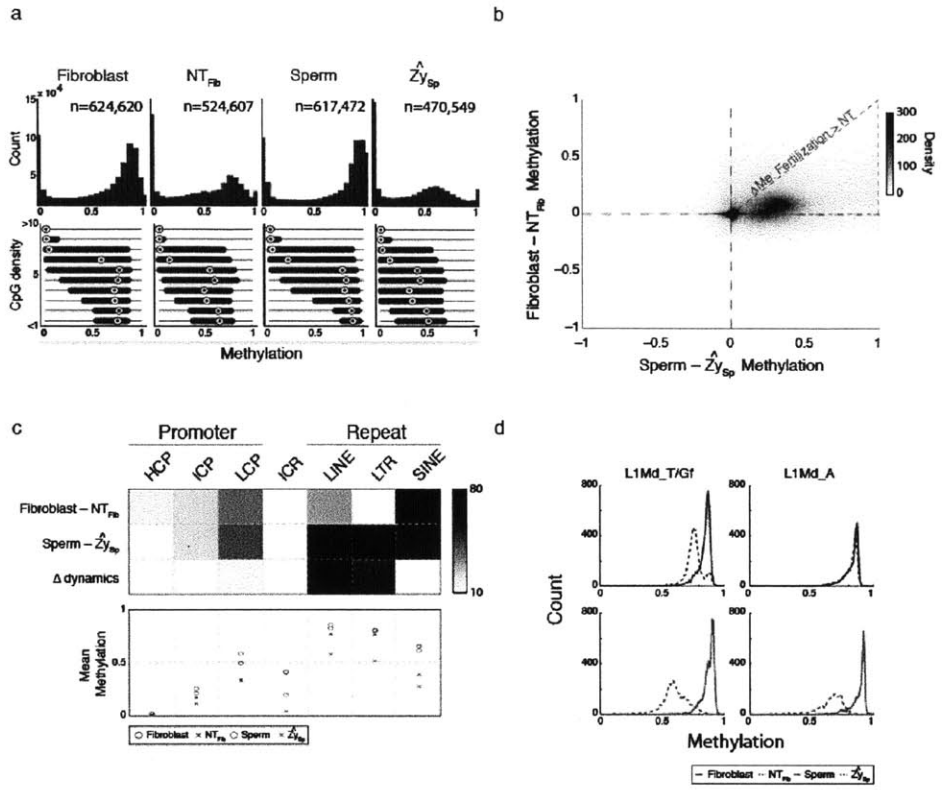

Figure 3-2: Classifying common and distinct DNA methylation dynamics during fertilization and nuclear transfer. **(a)** Histogram of methylation (top) and boxplots of methylation by CpG density (bottom) for 100 bp tiles in fibroblast, nuclear transfer reconstructed embryos (NT_{Fib}, mean of 4 replicates), sperm, and the inferred sperm value in zygote (Zy_{Sp}; see **Supplementary Methods**). Fibroblasts and sperm show a global methylation pattern typical of somatic tissues while SCNT embryos (NT_{Fib}) exhibit a similar global shift as Zy_{Sp}. Bulls-eye indicates the median, edges the 25th/75th percentile and whiskers the 2.5th/97.5th percentile. **(b)** Scatterplot comparing global methylation dynamics between nuclear transfer (Fibroblast – NT_{Fib}) and fertilization (Sperm – Zy_{Sp}). While demethylated regions appear to occur in common sites (upper right quadrant), the magnitude of demethylation is larger during fertilization as indicated by the dense cloud below the diagonal. The red triangle outlines the region where demethylation after SCNT is smaller than during fertilization. **(c)** Heatmap (top) depicting genomic features that significantly change (dark) in Fibroblast-NT_{Fib} or Sperm-Zy_{Sp} transitions, and the comparison of changes between the two transitions (Δ dynamics). Promoters are partitioned to high (H), intermediate (I) or low (L) CpG density and known imprint control regions (ICRs) are included as a control set. Color is the –log p-value. The mean methylation value for each feature set is shown in the bottom panel. Most features appear to change similarly in both processes with the exception of LINE and LTR features, which are comparably resistant to change after nuclear transfer. **(d)** Histogram of methylation for elements in the L1Md_T and L1Md_Gf (left) and L1Md_A (right) families of the LINE-1 class. Both families are dynamic during fertilization, but only the L1Md_T/Gf families, which show larger demethylation during fertilization, show any detectable change after nuclear transfer.

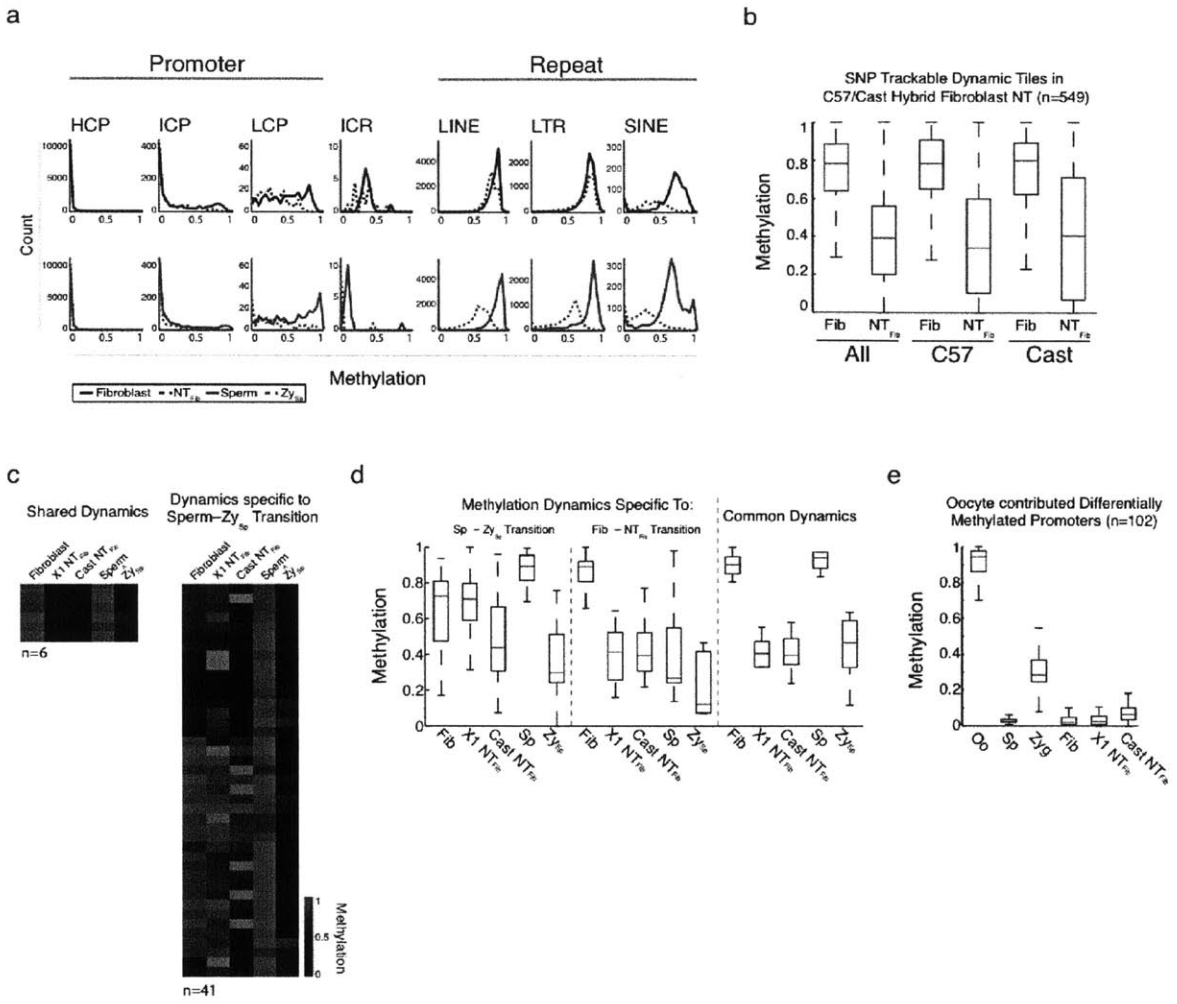

Figure 3-3: Genomic and promoter dynamics in nuclear transfer. **(a)** Histogram of methylation values for genomic features. Features show very similar global dynamics during fertilization and NT except for LINE and LTR repetitive elements. **(b)** Boxplots for SNP trackable CpGs (n=549) falling into dynamic 100 bp tiles (>0.2 change) after NT in BDF1xCast hybrid fibroblast experiments. "All" refers to the SNP normalized methylation value before segregation into either C57 or Cast allelotypes. **(c)** Heatmap of promoter dynamics that are shared in fertilization and NT and those that are fertilization specific. Rounds of NT using different background strains are shown separately to highlight the fidelity of these trends. Grey represents missing values. **(d)** Boxplot of promoter methylation for the dynamic sets. Promoters specific to the sperm-zygote transition are mostly hypermethylated in fibroblasts and retained higher methylation levels after NT. Similarly, promoters specific to NT are predominantly hypomethylated in sperm. **(e)** The early embryo receives methylation information from the oocyte genome in the form of 102 hypermethylated promoter DMRs. These methylated regions are pre-implantation specific and erased upon embryonic specification. They remain hypomethylated in both fibroblasts and in NTFib experiments.

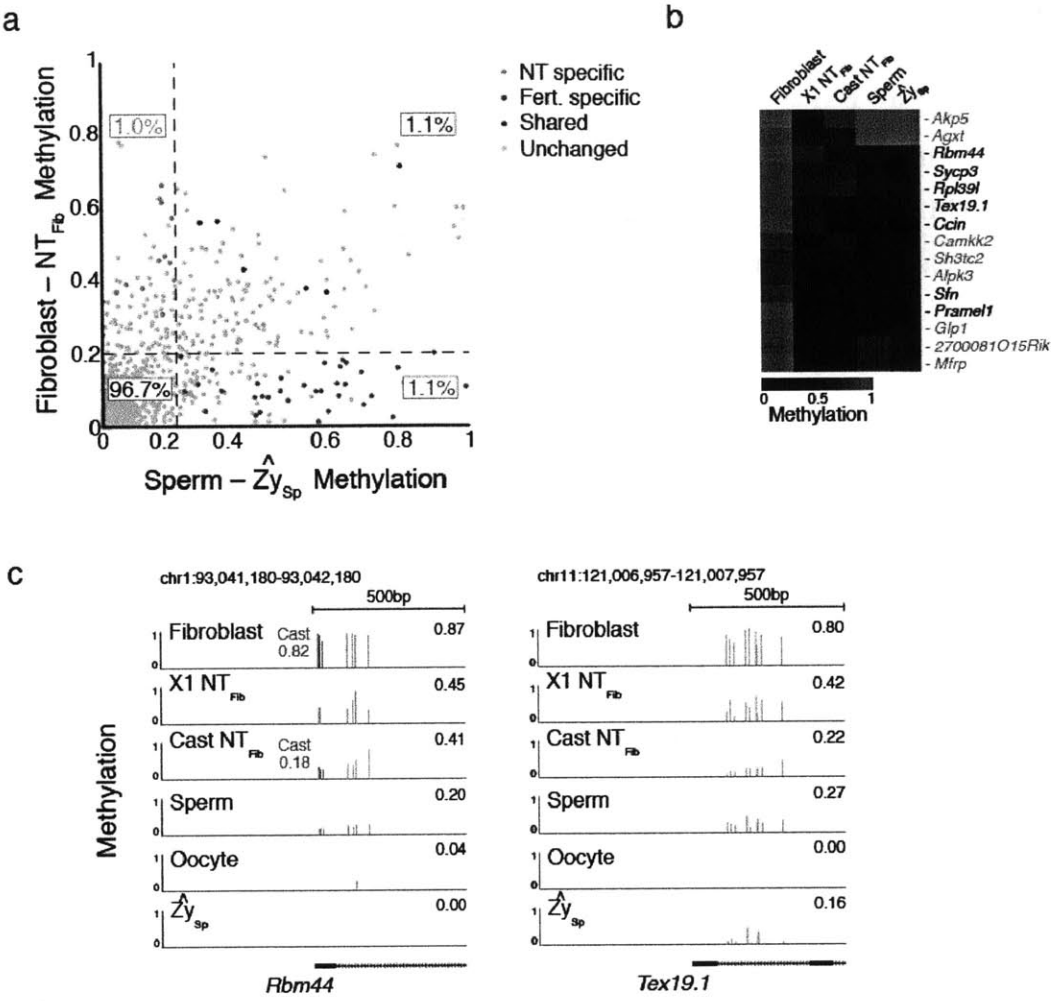

Figure 3-4: Promoter dynamics during SCNT include demethylation of gamete-specific genes. (a) Scatterplot of promoter dynamics between donor fibroblasts and SCNT embryos compared to those observed at fertilization. The majority of promoters are unchanged in either process, with 1.0%, 1.1% and 1.1% of promoters dynamic in either NT specific, shared, or fertilization specific contexts. Colored dots refer to promoters that were observed to change consistently across SCNT experiments and/or during demethylation of the paternal genome after fertilization. "Other" includes promoters that either did not change or did not pass the stringent criteria for being called as dynamic. **(b)** Dynamics specific to the Fibroblast-NT transition include several promoters that function specifically in gametes and are already hypomethylated in sperm and oocyte. **(c)** The germ-line associated genes RNA binding motif protein 44 (*Rbm44*) and Testes expressed gene 19.1 (*Tex19.1*) are hypomethylated and expressed during gametogenesis, the early embryo, and pluripotent cell lines but *de novo* methylated upon gastrulation/differentiation [4]. In fibroblasts, both gene promoters are hypermethylated and show a strong demethylation after either NT round. The level of demethylation suggests erasure in a large proportion of transplanted cells. Blue bars highlight single CpGs that are captured in all stages, red bars highlight 3 CpGs that can be associated with the Cast allele, with the mean allele-specific methylation value highlighted in red.

3.6 References

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

2 Rideout, W. M., 3rd, Eggan, K. & Jaenisch, R. Nuclear cloning and epigenetic reprogramming of the genome. *Science* 293, 1093-1098. (2001).

3 Wakayama, T., Perry, A. C., Zuccotti, M., Johnson, K. R. & Yanagimachi, R. Full-term development of mice from enucleated oocytes injected with cumulus cell nuclei. *Nature* 394, 369-374, doi:10.1038/28615 (1998).

4 Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* 454, 766-770 (2008).

5 Smith, Z. D. *et al.* A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature*, doi:10.1038/nature10960 (2012).

6 Goodier, J. L., Ostertag, E. M., Du, K. & Kazazian, H. H., Jr. A novel active L1 retrotransposon subfamily in the mouse. *Genome Res* 11, 1677-1685, doi:10.1101/gr.198301 (2001).

7 Wossidlo, M. *et al.* Dynamic link of DNA demethylation, DNA strand breaks and repair in mouse zygotes. *Embo J* 29, 1877-1888, doi:emboj201080 [pii] 10.1038/emboj.2010.80 (2010).

8 Egli, D. *et al.* Reprogramming within hours following nuclear transfer into mouse but not human zygotes. *Nat Commun* 2, 488, doi:ncomms1503 [pii] 10.1038/ncomms1503 (2011).

9 Dean, W. *et al.* Conservation of methylation reprogramming in mammalian development: aberrant reprogramming in cloned embryos. *Proc Natl Acad Sci U S A* 98, 13734-13738. (2001).

10 Polanski, Z., Motosugi, N., Tsurumi, C., Hiiragi, T. & Hoffmann, S. Hypomethylation of paternal DNA in the late mouse zygote is not essential for development. *Int J Dev Biol* 52, 295-298, doi:072347zp [pii] 10.1387/ijdb.072347zp (2008).

11 Ng, R. K. & Gurdon, J. B. Epigenetic memory of an active gene state depends on histone H3.3 incorporation into chromatin in the absence of transcription. *Nat Cell Biol* 10, 102-109, doi:ncb1674 [pii]

10.1038/ncb1674 (2008).

12 Gu, T. P. *et al.* The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. *Nature* 477, 606-610, doi:nature10443 [pii] 10.1038/nature10443 (2011).

13 Wossidlo, M. *et al.* 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. *Nat Commun* 2, 241, doi:ncomms1240 [pii] 10.1038/ncomms1240 (2011).

14 Branco, M. R., Ficz, G. & Reik, W. Uncovering the role of 5-hydroxymethylcytosine in the epigenome. *Nat Rev Genet*, doi:nrg3080 [pii] 10.1038/nrg3080 (2011).

15 Booth, M. J. *et al.* Quantitative Sequencing of 5-Methylcytosine and 5-Hydroxymethylcytosine at Single-Base Resolution. *Science*, doi:10.1126/science.1220671 (2012).

Chapter 4

Conclusion

Chapter 4:

Conclusion

4.1 Summary

In this thesis, I analyzed genome-scale DNA methylation maps of mammalian embryogenesis and somatic cell nuclear transfer embryos. In mouse pre-implantation development, my results uncovered novel signatures of this well characterized epigenetic mark identifying regions that retained methylation in the face of global methylation erasure as well as discovering loci with transient, parentally conferred methylation. I described a model for DNA methylation in early human development and found that many dynamics were shared across species, but targets that exhibited transient, maternally conferred methylation were species-specific. Finally, I examined the reprogramming capacity of the oocyte and highlighted limitations of the DNA demethylation machinery in an artificial reprogramming process, which may explain its low efficiency of developmental success.

In Chapter 1, I investigated DNA methylation dynamics across pre-implantation development in mouse. The sperm genome methylation profile resembled a classic somatic pattern while the oocyte genome was surprisingly hypomethylated, displaying a DNA methylation signature that was also observed in the zygote. Loss of hypermethylation continued over the cleavage divisions reaching a minimum at the inner cell mass (ICM) of the blastocyst. Global hypomethylation is thought to be associated with broad developmental potential, and appropriately reflected the ICM's pluripotent state[1]. By the early post-implantation embryo, the genome was already remethylated to resemble a somatic pattern. Reprogamming of the paternal genome at fertilization is a dramatic chromatin remodeling event that includes global

demethylation[2]. Observations from immunohistochemistry suggested that the paternal genome is completely demethylated but the similarity between our oocyte and zygote profiles implied that the paternal genome is actually demethylated to maternal levels. One class of genomic features which showed surprising demethylation is retrotransposons, specifically an evoluntionarily young family of long interspersed nuclear elements (LINE)[3]. Retrotransposons are generally highly methylated and thereby silenced in cells, but have been shown to be some of the earliest transcripts present after fertilization and necessary for successful progress through the cleavage divisions[4,5]. We found that the oocyte and sperm contributed differentially methylated regions (DMRs) to the zygote in addition to the known imprinting control regions (ICRs). Oocyte-contributed DMRs, which were highly methylated in the oocyte, were enriched for promoters, while sperm-contributed DMRs mainly consisted of repeat elements. Surprisingly, both classes were transient DMRs meaning that they only exhibited parent-specific methylation before establishing their expected methylation levels (low methylation for promoters and high methylation for repeats) by the post-implantation stage. The purpose of these transient DMRs remains unclear.

The mammalian model for DNA methylation during early development is predominantly based on studies in mouse and it is unknown if the regulatory principles extend to human. To gain insight into the unique and conserved DNA methylation dynamics during embryogenesis, we produced RRBS maps of human development in Chapter 2. We find that methylation patterns are globally conserved between human and mouse but species-specific targets define regions of local divergence that appear to be gametic in origin. Repetitive elements cannot be directly compared between species, but comparison of methylation dynamics between equivalent classes

shows both conserved and divergent behavior. Regardless of their similarity to murine developmental dynamics, repetititve elements from the same class or family show coherent methylation signatures. Evolutionarily younger members of the L1PA family exhibit embryonic DNA methylation escape compared to ancestral members – a pattern reminiscent of LINE elements in mouse. The presence of maternally contributed differentially methylated regions that exist transiently in the pre-specified embryo were shared between human and mouse, but the specific identities of the targets were not. Intriguingly, targeting for this methylation signature may not occur at the level of genes but at higher levels of cellular organization, as evidenced by the regulation of the mouse DMR, Neurog3, by the human DMR, GLIS3. The seeming necessity for pathways to be regulated through this transient methylation mechanism strongly implies a coherent function, which may involve splicing regulation, like the conserved DMR, DNMT1, or may be related to the development of extraembryonic tissue. Our study of DNA methylation dynamics in early human development extends the mammalian model from mouse to human for this interesting phase.

Our work in early mouse development showcased the power of the ooplasm at reprogramming the paternal genome during fertilization. In Chapter 3, we turned our attention to reprogramming in an artificial system, namely after somatic cell nuclear transfer (SCNT). By comparing the DNA methylation profiles of the donor genome to the SCNT genome, we explicitly characterized methylation dynamics caused by the ooplasm and compared them to those observed in the paternal genome upon fertilization. The SCNT genome is less methylated compared to the donor genome, which is reflective of a methylation profile for a cell with greater differentiation potential[1]. The methylation dynamics during SCNT are very similar, albeit at a

smaller magnitude, to those during natural fertilization. The major differences in global dynamics are in repetitive elements, which appear to be more resistant to demethylation in SCNT than fertilization. In contrast to the somatic nucleus, whose DNA is packaged into chromatin, the paternal genome contains protamines and undergoes rechromatinization during fertilization[2]. This may provide a unique window of access to the ooplasm's demethylation machinery that does not exist in SCNT and may explain the higher retention of DNA methylation in the SCNT genome. The smaller magnitude of demethylation observed after SCNT may reflect heterogeneity in the population with some SCNT zygotes exhibiting demethylation changes comparable to those at fertilization. Presumably, it is these SCNT zygotes that have a higher likelihood of developing to term after implantation into a surrogate female. Advancements in single cell technology coupled with the application of global profiling methods may help explain the inefficiency of SCNT and characterize the attributes of successfully reprogrammed SCNT genomes. SCNT is often considered the gold standard of reprogramming but I described barriers that exist in SCNT which are not present during natural development.

4.2 Future Perspectives

Great strides have been made in the past few years in understanding the distribution and mechanism of DNA methylation during early mammalian development, and we now have a foundation to explore its role in this interesting phase. One path of future exploration may be to systematically dissect the demethyation process to gain a better understanding of its targets, dynamics, and mechanism. A first step would be to increase our ability to distinguish between the maternal and paternal genomes. In our study of murine pre-implantation development, we used natural mating between two relatively divergent strains but by utilizing intracytoplasmic

sperm injection (ICSI) instead, we can cross more divergent strains which will give us >20x more SNPs to track parent specific methylation. With higher resolution paternal genome tracking, we can more precisely classify the regions that undergo active demethylation at fertilization from those that decay passively over the cleavage divisions. We can also investigate the effect of different mutants, such as Tet-knockout mice on DNA demethylation dynamics. The Tet enzymes convert 5-methylcytosine to 5-hydroxmethylcytosine which appears to be an intermediate to demethylation[6,7]. Many pathways to mammalian DNA demethylation have been proposed and many may be used in the cell[6,7]. Identifying which Tet enzymes may be responsible for DNA demethylation, and the targets and phase of demethylation it affects will be an important contribution.

Several groups have identified the transient, maternally-contributed differentially methylated regions (DMRs) that are imprinted until implantation but their biological function has yet to be described[8-10]. Moreover, they do not appear to share functional annotation as a set. Genes in maternal imprinting control regions (ICRs) regulate pathways important for the establishment of the feto-maternal interface in early development[11]. Perhaps these transient DMRs have a similar role. In support of this hypothesis, when the extra-embryonic lineage of SCNT embryos, which lack the transient, maternally-contributed methylation, is replaced with one generated through *in vitro* fertilization, the frequency of full-term development of cloned mice increases ~6-fold[12]. Exploring the potential function of transient DMRs will produce insight into the early developmental process, and may also provide new information about the formation of extraembryonic tissues.

We can apply our new knowledge of early development to gain a better understanding of *in vitro* derived embryonic stem cells (ESCs). Human and mouse ESCs are derived from the inner cell mass (ICM) of the blastocyst, and the mouse epiblast can also be used to generate a stem cell line, termed EpiSCs[13-15]. Interestingly, hESCs are more similar to mEpiSCs than mESCs[16]. None of these ESCs however recapitulate the hypomethylated profile of the ICM, instead having DNA methylation patterns reminiscent of somatic cells[17,18]. Developments to the mESC derivation protocol have produced more defined culture conditions and a pluripotent cell line that closely resembles the ICM[19]. This condition, "2i", utilizes MEK and Gsk3b inhibitors which lead to downregulation of Dnmt3a and Dnmt3b[20,21]. We have profiled DNA methylation in both human and mouse ICMs, mouse epiblast, hESCs, and mESCs from multiple derivation conditions and now have the opportunity to explore the methylation differences that define different states and separate them from those that delineate species.

DNA methylation is one of the best-characterized epigenetic modifications yet there remains much to be discovered. In this thesis, I have described its distribution and dynamics in two processes – early mammalian development and somatic cell nuclear transfer. The functional consequences of DNA methylation in these phases however are still unclear. Unraveling its regulatory role at these times will not only lead to a greater understanding of multicellular development but may also provide fundamental clues to the reprogramming process.

4.3 References

1 Reik, W., Dean, W. & Walter, J. Epigenetic reprogramming in mammalian development. *Science* **293**, 1089-1093, doi:Doi 10.1126/Science.1063443 (2001).

2 Polanski, Z., Motosugi, N., Tsurumi, C., Hiiragi, T. & Hoffmann, S. Hypomethylation of paternal DNA in the late mouse zygote is not essential for development. *The International journal of developmental biology* **52**, 295-298, doi:10.1387/ijdb.072347zp (2008).

3 Goodier, J. L., Ostertag, E. M., Du, K. & Kazazian, H. H., Jr. A novel active L1 retrotransposon subfamily in the mouse. *Genome research* **11**, 1677-1685, doi:10.1101/gr.198301 (2001).

4 Beraldi, R., Pittoggi, C., Sciamanna, I., Mattei, E. & Spadafora, C. Expression of LINE-1 retroposons is essential for murine preimplantation development. *Molecular reproduction and development* **73**, 279-287, doi:10.1002/mrd.20423 (2006).

5 Kigami, D., Minami, N., Takayama, H. & Imai, H. MuERV-L is one of the earliest transcribed genes in mouse one-cell embryos. *Biology of reproduction* **68**, 651-654 (2003).

6 Bhutani, N., Burns, D. M. & Blau, H. M. DNA demethylation dynamics. *Cell* **146**, 866-872, doi:10.1016/j.cell.2011.08.042 (2011).

7 Branco, M. R., Ficz, G. & Reik, W. Uncovering the role of 5-hydroxymethylcytosine in the epigenome. *Nature reviews. Genetics* **13**, 7-13, doi:10.1038/nrg3080 (2012).

8 Proudhon, C. *et al.* Protection against de novo methylation is instrumental in maintaining parent-of-origin methylation inherited from the gametes. *Molecular cell* **47**, 909-920, doi:10.1016/j.molcel.2012.07.010 (2012).

9 Smallwood, S. A. *et al.* Dynamic CpG island methylation landscape in oocytes and preimplantation embryos. *Nature genetics* **43**, 811-814, doi:10.1038/ng.864 (2011).

10 Smith, Z. D. *et al.* A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature* **484**, 339-344, doi:10.1038/nature10960 (2012).

11 Niemann, H., Tian, X. C., King, W. A. & Lee, R. S. Epigenetic reprogramming in embryonic and foetal development upon somatic cell nuclear transfer cloning. *Reproduction* **135**, 151-163, doi:10.1530/REP-07-0397 (2008).

12 Narbonne, P., Miyamoto, K. & Gurdon, J. B. Reprogramming and development in nuclear transfer embryos and in interspecific systems. *Current opinion in genetics & development* **22**, 450-458, doi:10.1016/j.gde.2012.09.002 (2012).

13 Evans, M. J. & Kaufman, M. H. Establishment in culture of pluripotential cells from mouse embryos. *Nature* **292**, 154-156 (1981).

14 Tesar, P. J. *et al.* New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature* **448**, 196-199, doi:10.1038/nature05972 (2007).

15 Thomson, J. A. *et al.* Embryonic stem cell lines derived from human blastocysts. *Science* **282**, 1145-1147 (1998).

16 Hanna, J. *et al.* Human embryonic stem cells with biological and epigenetic characteristics similar to those of mouse ESCs. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 9222-9227, doi:10.1073/pnas.1004584107 (2010).

17 Bock, C. *et al.* Reference Maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. *Cell* **144**, 439-452, doi:10.1016/j.cell.2010.12.032 (2011).

18 Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766-770, doi:10.1038/nature07107 (2008).

19 Nichols, J. & Smith, A. Pluripotency in the embryo and in culture. *Cold Spring Harbor perspectives in biology* **4**, a008128, doi:10.1101/cshperspect.a008128 (2012).

20 Leitch, H. G. *et al.* Naive pluripotency is associated with global DNA hypomethylation. *Nature structural & molecular biology* **20**, 311-316, doi:10.1038/nsmb.2510 (2013).

21 Yamaji, M. *et al.* PRDM14 ensures naive pluripotency through dual regulation of signaling and epigenetic pathways in mouse embryonic stem cells. *Cell stem cell* **12**, 368-382, doi:10.1016/j.stem.2012.12.012 (2013).

Appendix A

Reduced representation bisulfite sequencing and the detection of amplification artifacts

Appendix A:

Reduced representation bisulfite sequencing and the detection of amplification artifacts

A.1 Overview

Many reads are expected to start at the same position in RRBS libraries due to the Msp1 digestion of the genome. This poses a problem in judging data quality, which is more easily assessed when duplicate reads, which may or may not arise from the same amplicon, can be thrown out. There is not yet a gold standard for gauging data quality in RRBS libraries with most quality control criteria relying on similarity between biological replicates. One intersample independent measure that has been used is looking at the methylation levels in imprint control regions (ICRs). We have added a second measure, investigating the distribution of single nucleotide polymorphism (SNP) ratios, as well as explored possible amplification biases in other ways. In the somatic cell nuclear transfer experiment, we were able to use the presence of recipient DNA to assess amplification bias because we used different strain backgrounds for the donor genome and recipient oocyte. Finally, in collaboration with Zack Smith, Hongcang Gu, Andreas Gnirke, and Alex Meissner, we barcoded amplicons for varying amounts of input DNA to directly determine amplification bias.

A.2 Results

A.2.1 Methylation in imprint control regions

Imprint control regions (ICRs) are differentially methylated according to their parent-of-origin with one parental copy being fully methylated and the other copy unmethylated. Their methylation patterns are established during primordial germ cell development and are maintained throughout early development and differentiation[1]. The expected methylation levels for these regions are thus 0.5 and this is generally the value seen in our samples (**Fig. A-1**). The major problem with this quality control measure is that it relies on good annotation of the genome. ICRs are well annotated in mouse but less well characterized in human[2,3]. Our human data presents an opportunity to improve this annotation but consequently, the identified regions can no longer be used as an indicator of data quality.

A.2.2 Single nucleotide polymorphism ratios

A single nucleotide polymorphism (SNP) is a base that is different between two genomes at the same position. By breeding mice that come from different strain backgrounds and have sequenced genomes, a catalog of SNPs between the two strains can be assembled[4]. With the positions and identities of the SNPs known, the number of reads that originate from the paternal genome vs. the maternal genome can be compared. Since each parent should provide one copy of an allele, the expected ratio of maternal to paternal reads is 0.5 (**Fig. A-2**). Smaller amounts of input DNA are expected to be more sensitive to amplification bias as a consequence of reduced input and complexity and, possibly, extended PCR amplification. Indeed, samples that start with lower amounts of input DNA show larger amplification biases, which is represented by a larger spread around the mean ratio (**Fig. A-2**). The methylation value of CpGs that occur on

reads that can be assigned to a parent-of-origin can be normalized. For most samples, the correlation between unnormalized and normalized methylation values is high, though this is partly a reflection of methylation values being typically either 0 or 1 irrespective of parental origin (**Fig. A-3**). Because there are so few data quality measures that can be used for RRBS, an effort should be made to utilize hybrid mouse strains that possess the highest level of genetic diversity.

Unlike the case in mouse, the genotypes of our human samples are unknown. Nonetheless, we can continue to use SNP ratios as a measure of data quality when a sample originates from a single person, which is the case for cell lines or single blastocysts. Rather than using de novo SNP finding against the human reference genome which requires fairly high sequencing coverage and would be sensitive to sample sequencing errors, we take SNP positions from the 1000 genomes project[5]. Deviating from the strategy in mouse, we ignore the minor allele SNP identity and simply count the number of reads that match the reference genome vs. those that do not. Homozygous positions are filtered out and the remaining SNP ratios are plotted with the expectation that they will be 0.5 (**Fig. A-4**). In addition, two sets of siblings are known to comprise our five single human blastocyst samples. The genotypes of siblings should be more similar to each other than non-siblings, and indeed, this is reflected in our data set (**Fig. A-5**). While genotype similarity does not provide specific information about amplification artifacts, it lends general assurance that amplification bias is not so large as to distort true biological signal.

A.2.3 Presence of recipient oocyte DNA after somatic cell nuclear transfer

In somatic cell nuclear transfer (SCNT), the nucleus of the recipient oocyte is removed before the genome of the donor is injected; thus, in a perfect world, there should be no recipient oocyte DNA detectable in the sequenced SCNT sample[6]. Unfortunately, this is not the case and may be the result of imperfect oocyte enucleation or may represent genomic DNA that exists outside of the metaphase plate, where oocyte genomic DNA is retained in an arrested, condensed state. If we assume that the amount of recipient DNA remaining in a single enucleated oocyte is small and the likelihood of any particular region to be left behind is equal, then we can assume that any recipient DNA region that exists in the SCNT sample of ~100 zygotes to be in a single copy. With these assumptions, the proportion of reads from the recipient genome at a given locus can be used as an indicator of amplification error. This strategy can only be applicable when different strain backgrounds are used for the donor genome and recipient oocyte.

To estimate the amplification artifact, we binned all SNPs that showed any level of recipient DNA by their coverage into 5x intervals. As coverage increases, we should achieve a better estimate of the mean amplification bias and it does appear to level out around 15% (**Fig. A-6**). This strategy for detecting amplification artifacts is not ideal as we cannot distinguish between amplification and multiple copies of recipient DNA in the sample. Nevertheless, it gives an upper limit estimate of the average amplification artifact.

A.2.4 RRBS barcoded amplicon libraries: the gold standard

To better understand the amplification bias that might exist in RRBS libraries, we worked in collaboration with Zack Smith, Hongcang Gu, Andi Gnirke, and Alex Meissner to add 8mer barcodes to fragments before adapter ligation to RRBS libraries. We varied the starting amount of DNA (from 0.06 ng or ~20 cells to 37.5ng or ~100 000 cells) and varied the PCR cycle number for some inputs (**Table A-1**).

By tagging each amplicon with a barcode, we can directly assess amplification artifacts since a unique read can be identified using its barcode and mapped position. More input DNA produces more unique fragments in the library and as expected, the proportion of unique reads increases as the input DNA amount increases. The proportion of unique reads decreases with increased PCR cycles, which is not surprising as more PCR amplification might be expected to increase amplification artifacts (**Fig. A-7**). Looking at the amplification bias (total reads/unique reads) at each position shows generally the same trend with higher input libraries showing less amplification artifacts than lower input libraries (**Fig. A-8**). Note that a large amplification bias at a few positions may have a dramatic effect on the unique proportion of reads reported above since the unique proportion does not explicitly consider position. For the highest input library 625k-15, the proportion of reads that showed some amplification bias is 25% with the median magnitude of the bias for reads that had an amplification artifact being 1.2. For comparison, the lowest input library 100-27, the proportion of reads that showed some amplification bias is 44% with the average magnitude of the bias for reads that had an amplification artifact being 2.5. These estimates agree with the amplification biases we inferred in our somatic cell nuclear transfer experiment.

To investigate directly if the artifacts were the result of a single amplicon being amplified repeatedly, which we term "amplicon takeover", or many amplicons being amplified a small number of times, we looked at the proportion of reads that correspond to the maximum covered amplicon out of the total number of reads for a position. Positions that correspond to amplicon takeover will have a proportion close to 1. Samples with higher inputs did not show behavior consistent with amplicon takeover, but samples with lower input showed some degree of amplicon takeover **(Fig. A-9)**.

With barcoded RRBS libraries we can directly assess amplification artifacts and found that low input libraries show some degree of amplification bias that increases as input decreases or as PCR cycles increase. For low input libraries, the amplification bias is not equal across unique reads but instead, one amplicon for a position may represent the majority of reads for that position for a minor fraction of positions in the library. It is worthwhile to note that the correlation coefficient does not increase greatly after barcode-normalization except in the case of extremely low input libraries **(Fig. A-10)**. This is likely due to the binary nature of methylation, meaning that amplification artifacts make no difference if a position is completely unmethylated or methylated.

A.3 Figures

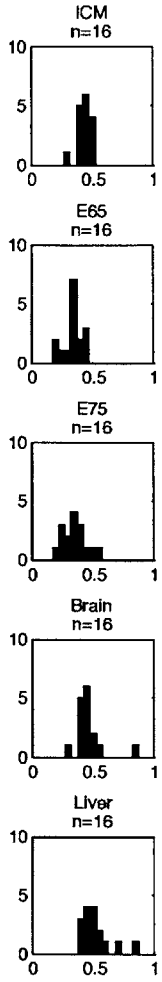

Figure A-1: Methylation histograms for imprinting control regions in mouse. Imprinting control regions (ICRs) are differentially methylated according to their parent-of-origin and are thus, expected to be ~0.5.

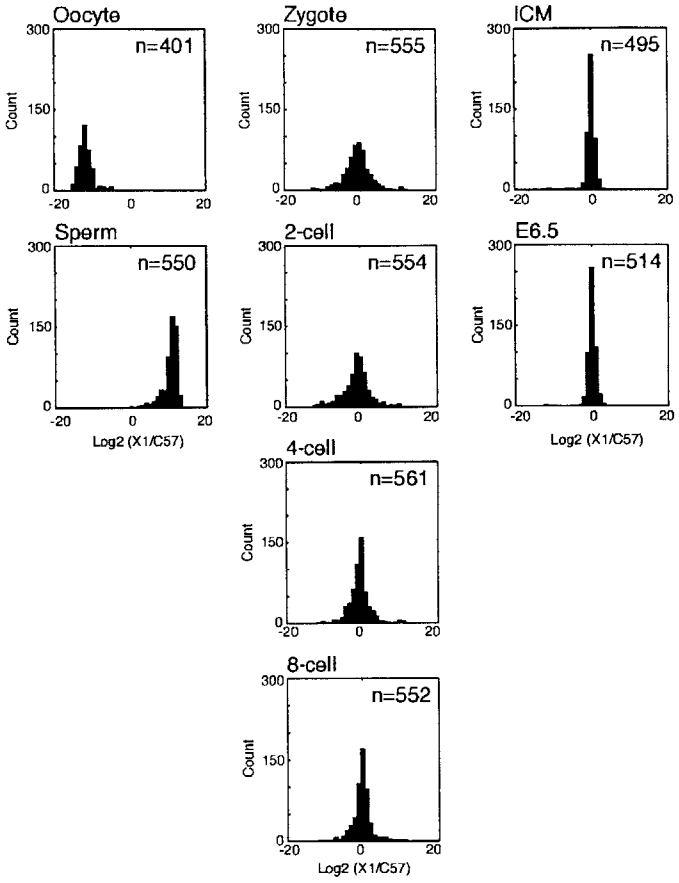

Figure A-2: Single nucleotide polymorphism (SNP) distributions for gametes and hybrid embryos. Log odds ratio histograms of allelic frequencies for BDF1 (C57/B6 based) maternal and 129X1 paternal SNPs. The spread around the mean indicates possible amplification bias and decreases as the input amount of DNA increases. The amount of input DNA increases from zygote through 8-cell. This figure was first published in [ref 7].

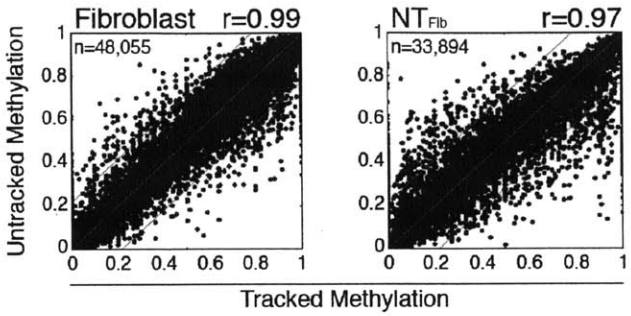

Figure A-3: Scatterplot of untracked vs. tracked methylation values. Methylation values of CpGs that occur on reads that can be assigned to a parent-of-origin can be normalized by parental coverage. The correlation between untracked and tracked, i.e. normalized, methylation is high. This figure was first published in [ref 8].

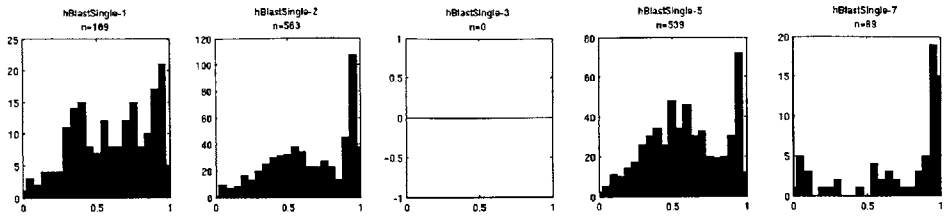

Figure A-4: SNP ratio histograms for single human blastocyst samples. The SNP ratio (reference allele/other allele) should be 0.5 for all positions in single blastocysts since there is one genotype in the sample. Deviations from 0.5 may indicate amplification artifacts. Homozygous positions have been removed and the peak near 1 likely represents sequencing errors. n is the number of SNP positions with >=10x coverage and are represented in the histogram. The sample 'hBlastSingle-3' has no SNPs that meet this threshold.

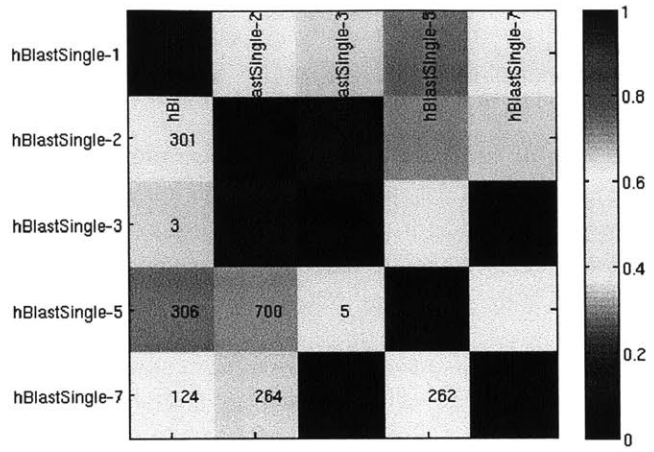

Figure A-5: Distance matrix between genotypes of single human blastocyst genotypes. Samples were genotyped according to the following scheme: AA=ratio>0.8; Aa=0.2>=ratio>=0.8; aa=ratio<0.2, and each genotype was assigned a number. SNPs with consistent genotypes across all samples were excluded and the average hamming distance was calculated. Blastocyst samples 1 and 2 are siblings, as are samples 3, 5, and 7. The genetic relationship between samples is reflected in the distance matrix where siblings have more similar genotypes than non-siblings. The numbers in the lower left triangle represent the number of SNP positions used to calculate the average hamming distance between two samples. SNPs were included if they had >=5x coverage.

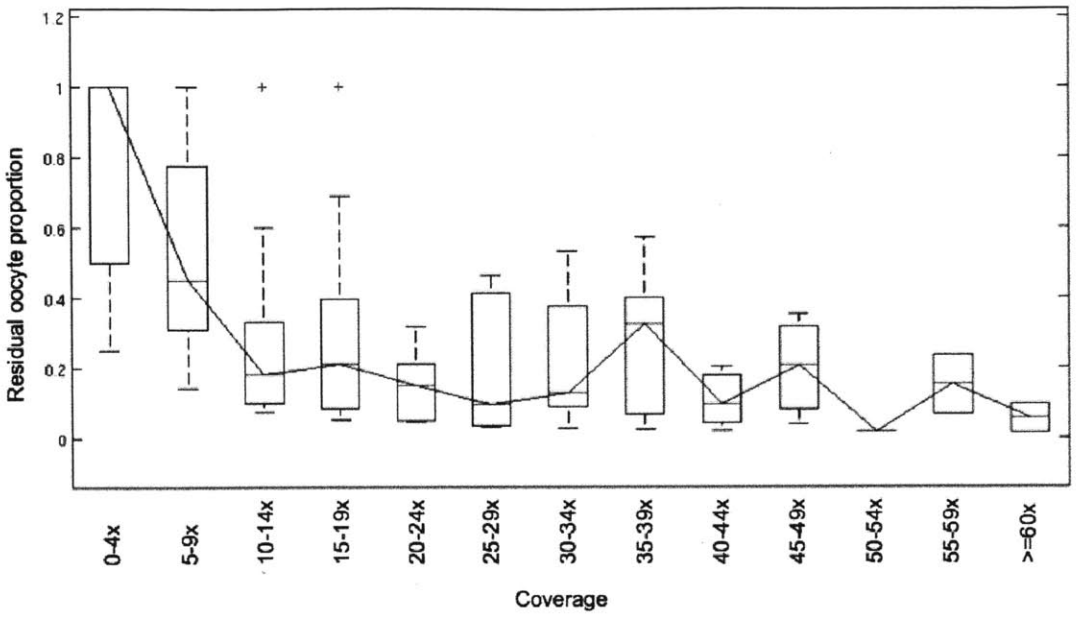

Figure A-6: Boxplot of residual oocyte DNA read proportion for increasing read coverage levels. Residual oocyte DNA is distinguished from the donor nucleus genome using SNPs. Assuming the residual oocyte DNA exists in a single copy, then the proportion of oocyte DNA can be used to estimate possible amplification error. Only loci where residual host DNA is detected is plotted.

Sample name	Input DNA (ng)	Approx. # sperm	PCR cycles
625K-15	37.5	12 500	15
625K-19	37.5	12 500	19
12.5K-18	7.5	2500	18
2.5K-21	1.5	500	21
500-23	0.3	100	23
500-27	0.3	100	27
100-27	0.06	20	27

Table A-1: Amount of input DNA and number of PCR cycles for barcoded RRBS libraries.

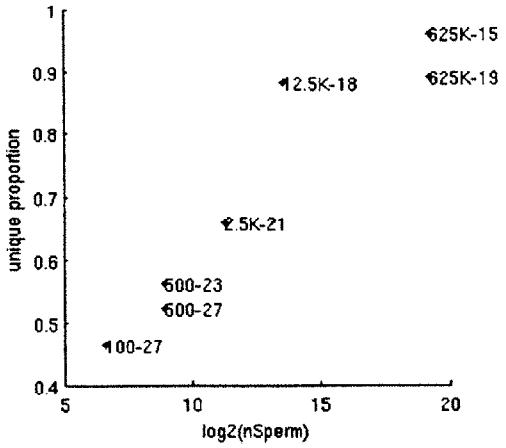

Figure A-7: Scatterplot of the unique proportion of reads against the number of cells in a sample. The unique proportion of reads increases as the amount of input DNA increases, meaning that duplicate reads are less common in samples of high input DNA. The number of PCR cycles (15 cycles vs 19 cycles in 625K-15 and 625K-19, respectively) also affects the duplication rate.

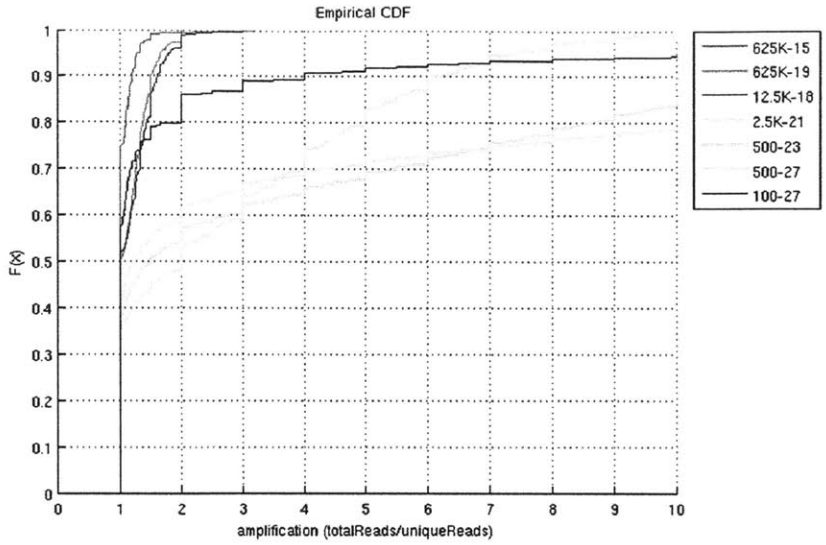

Figure A-8: Cumulative distribution function plot of amplification effect for each mapped position. The amplification error (total reads/unique reads) for each mapped position was calculated and plotted for comparison. High input libraries have fewer duplicate reads and the magnitude of duplication is less than for low input libraries.

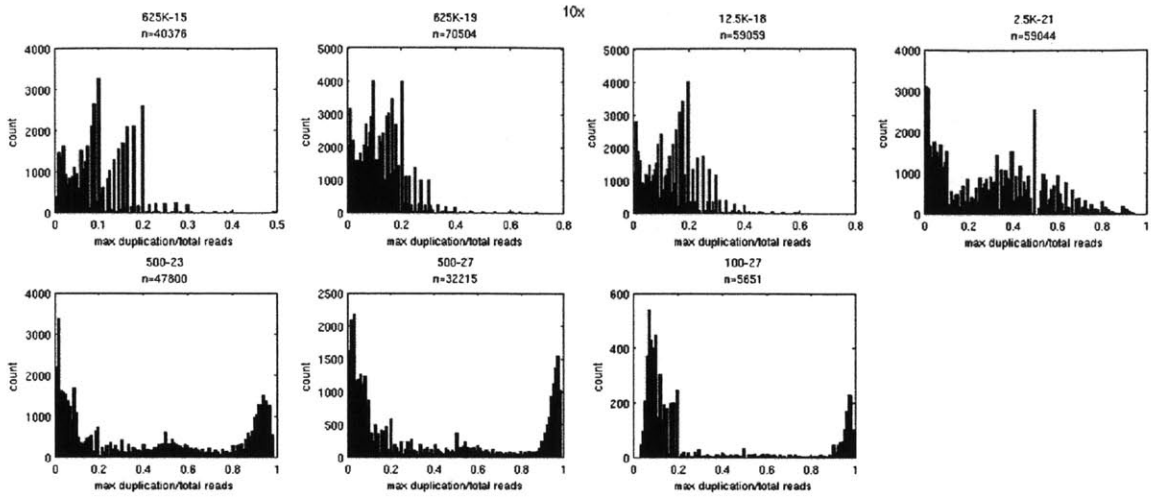

Figure A-9: Maximum duplicate amplicon proportion histograms. A high duplicate proportion, or inversely, low unique proportion, in a sample may be caused by many amplicons with some duplicates or few amplicons with many duplicates. The proportion of reads from the amplicon with the most duplicates for a position is plotted here and reveals that low input samples suffer from the latter effect.

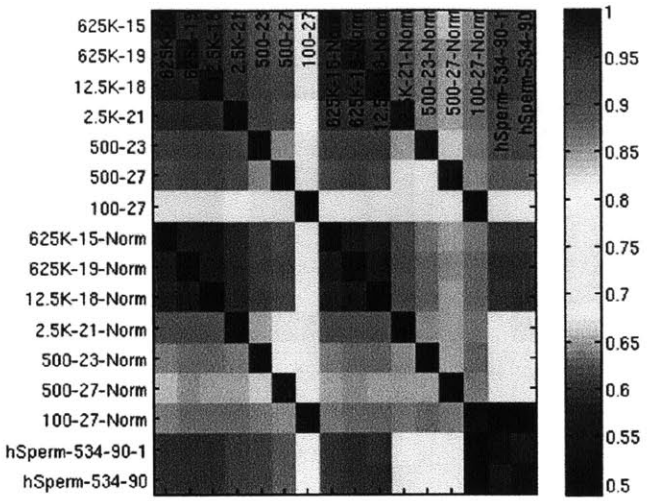

Figure A-10: Pearson correlation heatmap for unnormalized and barcode-normalized methylation. The correlation coefficient across different DNA input amounts is high regardless of sample normalization. This is likely due to the binary nature of DNA methylation where amplification artifacts will not affect loci that are either 0 or 1. The one exception is the lowest input sample, 100-27, where normalization improves its similarity to other samples. hSperm-534-90 and hSperm-534-91 are un-barcoded RRBS libraries from the same biological sample included for comparison.

A.4 References

1 Saitou, M., Kagiwada, S. & Kurimoto, K. Epigenetic reprogramming in mouse pre-implantation development and primordial germ cells. *Development* **139**, 15-31, doi:10.1242/dev.050849 (2012).

2 Daxinger, L. & Whitelaw, E. Understanding transgenerational epigenetic inheritance via the gametes in mammals. *Nature reviews. Genetics* **13**, 153-162, doi:10.1038/nrg3188 (2012).

3 Woodfine, K., Huddleston, J. E. & Murrell, A. Quantitative analysis of DNA methylation at all human imprinted regions reveals preservation of epigenetic stability in adult somatic tissue. *Epigenetics & chromatin* **4**, 1, doi:10.1186/1756-8935-4-1 (2011).

4 Blake, J. A. *et al.* The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. *Nucleic acids research* **39**, D842-848, doi:10.1093/nar/gkq1008 (2011).

5 Genomes Project, C. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:10.1038/nature09534 (2010).

6 Niemann, H., Tian, X. C., King, W. A. & Lee, R. S. Epigenetic reprogramming in embryonic and foetal development upon somatic cell nuclear transfer cloning. *Reproduction* **135**, 151-163, doi:10.1530/REP-07-0397 (2008).

7 Smith, Z. D. *et al.* A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature* **484**, 339-344, doi:10.1038/nature10960 (2012).

8 Chan, M. M., Smith, Z. D., Egli, D., Regev, A. & Meissner, A. Mouse ooplasm confers context-specific reprogramming capacity. *Nature genetics* **44**, 978-980, doi:10.1038/ng.2382 (2012).