Application of an allele-specific pipeline to study DNA methylation inheritance and dynamics in the early embryo

by

Julien Richard Albert

B.Sc., Concordia University, 2013

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Medical Genetics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

April 2020

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Application of an allele-specific pipeline to study DNA methylation inheritance and dynamics in the early embryo.

submitted by Julien Richard Albert in partial fulfillment of the requirements for

the degree of Doctor of Philosophy

in Medical Genetics

Examining Committee:

Matthew Lorincz, Medical Genetics
Supervisor

Louis Lefebvre, Medical Genetics
Supervisory Committee Member

Paul Pavlidis, Neuroscience
University Examiner

Wan Lam, Pathology and Laboratory Medicine
University Examiner

Additional Supervisory Committee Members:

Inanç Birol, Medical Genetics
Supervisory Committee Member

Dixie Mager, Medical Genetics
Supervisory Committee Member

Abstract

Classic genomic imprinted loci display parent-of-origin transcription in adult cells. Such allele-specific expression is thought to be driven by epigenetic marks, including DNA methylation (DNAme) and histone modifications, established in the gametes. However, the extent of monoallelic transcription in early mammalian development has remained relatively unexplored. As chromatin is highly dynamic before implantation, parent-of-origin directed transcription may be particularly high at this stage.

To identify parent-of-origin controlled transcription, including during early embryogenesis, I helped develop MEA, a Methylomic and Epigenomic Allele-specific analysis pipeline designed to integrate whole genome bisulphite (WGBS)-, RNA- and ChIP-sequencing datasets. To demonstrate the utility and sensitivity of MEA, I processed existing mouse and human datasets and uncovered classical as well as novel candidate imprinted genes.

Subsequently, I applied MEA to analyze WGBS and ChIP-seq data generated from adult gametes and preimplantation mouse embryos, which yielded female/maternal and male/paternal epigenomic profiles. Surprisingly, despite global DNAme loss following fertilization, I uncovered several dozen CpG island promoters that are *de novo* methylated on the paternal genome in 2-cell embryos, coincident with H3K4me3 loss on the same allele. A subset of these loci is hypermethylated in

androgenetic blastocysts but hypomethylated in parthenogenetic blastocysts, confirming the paternal genome is susceptible to post-fertilization DNAme activity. Notably, this zygotic paternal DNAme gain was ablated following genetic depletion of maternal DNMT3A. Parental DNAme levels at these loci are harmonized in the post-implantation embryo and beyond: thus, unlike classic imprinted regions, these novel DMRs are established on the paternal genome by maternal stores of DNMT3A in the zygote and likely only transiently impart allelic transcription in the embryo. Indeed, expression analysis of DNMT3A maternal KO preimplantation embryos revealed that genes normally gaining paternal DNAme following fertilization are prematurely activated from the paternal allele in 4C embryos.

Taken together, the results presented in my thesis illustrate the benefits of increasing the range and sensitivity of allele-specific analyses, and uncover zygotic *de novo* DNAme activity at CpG island promoters on the paternal genome against a backdrop of acute global demethylation.

Lay Summary

Similar to Indiana Jones, where a sandbag is cleverly swapped for a Mayan fertility figure, 1980s researchers realized they can swap DNA from one cell and replace it with DNA from another (with more success than Professor Jones). Using this method, researchers cloned a sheep by swapping DNA from a skin cell with DNA from a fertilized egg. Subsequently, researchers asked: “do we (mammals) need a mother and a father for conception?” In other words, can we artificially inseminate eggs with DNA from another egg instead of DNA from a sperm? Conversely, can we artificially inseminate an egg with two sperm and remove the egg DNA? Despite the success of the swapping method, motherless and fatherless embryos died before birth, indicating sperm and egg DNA are each required for conception. This thesis deepens our understanding of why these embryos die and extends our knowledge of mammalian reproduction and inheritance.

Preface

The candidate was the lead investigator and major contributor to data analysis, interpretation and manuscript writing. In cases where data collection was required, collaborations with international experts were established and are indicated below. Parts of this thesis have been published in peer-reviewed journals.

Chapter 2 has been published in *BMC Genomics*.

Julien Richard Albert, Tasuku Koike, Hamid Younesy, Richard Thompson, Aaron B. Bogutz, Mohammad M. Karimi and Matthew C. Lorincz. Development and application of an integrated allele-specific pipeline for methylomic and epigenomic analysis (MEA). *BMC Genomics* 19, 463 (2018).

JRA participated in study design, developed code, performed all analyses, generated the figures, and wrote the manuscript. TK developed code and performed allele-specific WGBS analysis related to Figures 2.4-2.6. HY and MKK participated in study design. RT and ABB ran bug tests and wrote the user guide. MCL participated in study design and wrote the manuscript with JRA.

MEA is free to use and available at: https://github.com/julienrichardalbert/MEA/

Chapter 3 has been submitted for peer-review.

Julien Richard Albert, Wan Kin Au Yeung, Keisuke Toriyama, Hisato Kobayashi, Ryutaro Hirasawa, Julie Brind'Amour, Aaron Bogutz, Hiroyuki Sasaki & Matthew Lorincz. Maternal DNMT3A-dependent de novo methylation of the zygotic paternal

genome inhibits gene expression in the early embryo.

JRA performed data analysis with the assistance of JBD and ABB, generated all the figures, conceived the project, and wrote the manuscript with MCL. WKAU, KT and HS carried out the experiments (WGBS and RNA-seq) on *Dnmt3a* matKO mice. HK performed the uniparental embryo WGBS experiments. RH and HS performed the zygotic DNMT3A IF analysis.

This project is available as a preprint:

https://www.biorxiv.org/content/10.1101/2020.03.26.009977v1

Table of Contents

List of Tables

List of Figures

List of Abbreviations

5mC	5-methyl-cytosine
AS	Allele-specific
bp	Base pair
BER	Base excision repair
C	Cytosine
CGI	CpG island
ChIP	Chromatin immunoprecipitation
ChIP-seq	Chromatin immunoprecipitation followed by sequencing
Chr	Chromosome
CpG	Cytosine-Guanine dinucleotide
IG-gDMR	Intergenic gametic differentially methylated region
DMR	Differentially methylated region
DNA	Deoxyribonucleic acid
DNMT1	DNA methyltransferase 1
DNMT3	DNA methyltransferase 3
DNAme	DNA methylation
Dr.	Doctor
E	Embryonic days *post coitum*
E2C	Early 2 cell
FDR	False discovery rate
G	Guanine

GVO	Germinal vesicle oocyte
H3	Histone H3
H3.3	Histone H3 variant H3.3
H3K4me1	Histone H3 Lysine 4 monomethylation
H3K4me2	Histone H3 Lysine 4 dimethylation
H3K4me3	Histone H3 Lysine 4 trimethylation
H3K9ac	Histone H3 Lysine 9 acetylation
H3K9me3	Histone H3 Lysine 9 trimethylation
H3K27ac	Histone H3 Lysine 27 acetylation
H3K27me3	Histone H3 Lysine 27 trimethylation
H3K36me3	Histone H3 Lysine 36 trimethylation
HTS	High throughput sequencing
ICM	Inner cell mass
INDEL	Insertion and/or deletion
Kb	kilobase
KO	Knock-out
KTMase	Lysine methyltransferase
L2C	Late 2 cell
Mat	Maternal
MatKO	Maternal knock-out
Mb	megabase
MEA	Pipeline for allele-specific methylomic and epigenomic analysis

MI	Metaphase I of meiosis
MII	Metaphase II of meiosis
MNase	Micrococcal nuclease
Pat	Paternal
PDA	Paternal DNA methylation acquisition
PTM	Post-translational modification
RNA	Ribonucleic acid
RNA-seq	Ribonucleoic acid sequencing
Seq	Sequencing
SNV	Single nucleotide variant
Sperm	Spermatozoa
SVs	Structural Variants
TF	Transcription Factor
TSS	Transcription start site
VCF	Variant call format

Acknowledgements

I have been fortunate to receive financial support for my graduate studies: a Medical Genetics Graduate Support Initiative Award, a Four Year Fellowship and a Killam Doctoral Scholarship from the University of British Columbia, a Doctoral scholarship from the Natural Sciences and Engineering Research Council of Canada, and a Master's scholarship from the Canadian Institutes of Health Research.

I am eternally grateful to Dr. Paul Joyce and Dr. Judith Kornblatt and their labs (Paul, Judy, Pam, Prit, Matthew, Esther, Gabby and Sabi) for spoon-feeding my first taste of basic science research, Dr. Reginal Storms' lab (Greg, Lina, Susan, Yun, Edith, Rebecca, Shaghayegh, Simon and Naïm) for an amazing summer of fungal research and Dr. Malcolm Whiteway's lab (Malcolm, Faïza, Pierre, Nicholas, Hannah, Tuana, Amjad, Jaideep, Hui, Walters, Adam, Yuan, Yaolin and Yisha) for treating me like family. To say Malcolm and Faïza inspired me to pursue a graduate degree would be an understatement – I continue to emulate them in both my professional and personal life. I am grateful for Dr. Adnane Sellam's mentoring in Dr. Mike Tyers' lab and exemplifying how to balance a demanding research/family lifestyle (and for the Farhat sandwiches).

I wish to thank my supervisory committee members Dr. Dixie Mager, Dr. Louis Lefebvre and Dr. Inanç Birol for their time, guidance, wisdom and encouragement. A big thank you to Lillian Jackson and Cheryl Bishop for their love and dedication to trainees. I am also grateful to Dr. Carolyn Brown for advice and opening her home to students.

I am forever indebted to my supervisor Dr. Matthew Lorincz – his lab members punch well above their weight class because of him. He exudes passion for research and is an inspiration for all those lucky enough to surround him. I am truly grateful for his role in shaping my scientific career. Thank you to his lab members Matt, Julie, Carol, Irina, Peter, Preeti, Sheng, Mehdi, Kris, Kenjiro, Kentaro and Aaron for their support, insightful discussions and training. A special thank you to our Japanese collaborators Drs. Hiroyuki Sasaki, Hisato Kobayashi, Donald Au Yeung and Task Koike for without their generosity and expertise this work would be impossible.

Thanks to Sidney, Aaron, Julie, Kris, Carol, Tom, Oscar, Ben, Grace, Hilary, Amanda and the Sams for sitting through my jokes and for their love and friendship. Special thanks to Kris for being the best roommate ever – you will be sorely missed. Thank you to Ryan, Seb and Brent for making me feel like I've never left home.

Finally, I would be remiss if I did not acknowledge my mother Sandra, for not only giving me life, but for imparting in me all the best life principles, like working hard and being forthright and compassionate with others. Words cannot express my gratitude to my big brother and father for being lifelong mentors. While I like to think I obtain my strength and inspiration from within, it would be dishonest not to acknowledge that I am constantly striving to fill two big pairs of shoes.

Dedication

To Meaghan.

Why a woman of such beauty and compassionate devotion

entrusts her life's happiness in me

eludes all understanding.

Chapter 1: Introduction

1.1 Genetic inheritance

Our understanding of the laws of human inheritance are deeply rooted in experiments using model organisms. In 1865, Gregor Mendel described the fundamental laws of inheritance by conducting experiments in his monastic garden on pea plants with distinct pea colour, shape and size. To generate such genetically distinct pea plant varieties, Mendel harnessed the pea plant's ability to self-pollinate (sexually reproduce with itself), which yielded lineages with reduced genetic variation. In this context of classical genetic inheritance, variation is measured as the extent of sequence variation between alleles; different forms of the same gene encoded on the maternal and paternal homologous chromosomes. During self-pollination, where the individual is the sole parent, half of the alleles inherited in the next generation are identical. In other words, through sequential rounds of self-pollination combined with trait selection, more and more alleles of the diploid pea plant genome become identical (homozygous).

Ultimately, reaching full homozygosity creates a pea plant "**strain**", where both parental alleles within an individual are isogenic, and all offspring are genetically identical to their parents. This process requires careful breeding and selection of parental traits (such as pea colour and shape) over time and is not always successful due to the expression of recessive lethal alleles. This potential lethality likely explains why most flowering plants evolved methods to prevent self-pollination (self-

incompatibility) and why animals reproduce sexually. Interestingly, the evolution of separate male and female germlines resulted in gametes (pollen/sperm and eggs) with distinct physiological features such as dramatically different cell size and DNA compaction levels as well giving rise to genomic imprinting in flowering plants and mammals.

While the fundamental laws of inheritance described by Mendel apply to mammals as well as plants, employing these laws to identify genes involved in human development and disease required a mammalian model organism. By the early 1900s, mouse breeders (also called mouse fanciers) in New England had domesticated mice and used sibling mating (akin to self-pollination) to create an array of mouse strains with unique characteristics, including coat colour and susceptibility to diseases such as cancers and ocular degeneracy (Taft et al. 2006). These strains were later adopted in 1901 (without credit to the fanciers) as the first mammalian model for genetic studies. Notably, mating distinct mouse strains generates F1 hybrid mice that carry one copy of each parental chromosome in each cell (resembling Mendel's pea strain crosses). These F1 hybrid mice are useful models for studying factors that regulate gene expression in *cis*, as discussed below, as polymorphisms in the parental alleles enable the direct assessment of the influence of allelic variation on allele-specific gene expression (Pastinen 2010; Wang and Clark 2014). In the century of genetic research that followed, researchers extended the laws of inheritance to mammals and humans and uncovered additional, "**non-Mendelian**", modes of inheritance. These exceptions to

Mendel's laws of inheritance include so called "parent-of-origin" effects, whereby the expression of a phenotype depends on whether the affected allele is inherited from the mother or the father, despite being identical.

1.2 The discovery of genomic imprinting and epigenetics

An increasing number of studies of human disease highlighted the importance of "parent-of-origin" inheritance in human development. For example, a deletion on human chromosome 15q11.2 manifests into two distinct developmental disorders depending on the parental allele affected: while maternal deletion results in Angelman syndrome, paternal deletion results in Prader-Willi syndrome (Horsthemke and Wagstaff 2008). Additionally, a deletion on chromosome 11p15.5 causes Beckwith-Wiedemann syndrome only when inherited from the mother, and is characterized by childhood overgrowth (Choufani et al. 2010). Thus, the parental alleles within a given individual are functionally non-equivalent at a subset of loci, but measuring the full extent of such parent-of-origin effects, such as at non-syndromic loci, and the mechanism by which they are established required modelling in a non-human experimental organism.

Artificial activation of mouse oocytes in the 1970s showed that parthenogenetic embryos failed to develop, suggesting parental DNA is functionally non-equivalent. However, since these uniparental embryos were never fertilized, such studies could not delineate whether failure to develop was due to functional differences between parental DNA or as consequence of artificial activation (Kaufman 1973). In 1984, Surani, Barton

and Norris harnessed the emerging nuclear transfer technique to control for fertilization and conclusively demonstrate that parthenogenetic (fatherless) and androgenetic (motherless) embryos failed to develop beyond embryonic day 6.5 (E6.5) (Surani et al. 1984; McGrath and Solter 1984; Barton et al. 1985). Intriguingly, these uniparental conceptuses had different phenotypes; while parthenogenetic embryos showed limited extraembryonic development and normal embryos of reduced size, androgenetic embryos showed extraembryonic overgrowth and poor embryonic development reminiscent of human hydatidiform moles, which are rare human androgenetic embryos characterized by normal placental growth and the lack of an embryo proper (Kajii and Ohama 1977). Together, these results demonstrated that, despite identical in genomic sequence, a "genomic imprint" is retained on parental genomes following fertilization and this memory imparts control over distinct embryonic developmental programmes.

Less than a decade later, the first "imprinted gene", a gene transcribed from only one of the two parental alleles, was discovered. Using a combination of a genetic model for imprinting disease and F1 hybrid embryos, Barlow and colleagues determined that the gene *Igf2r* is expressed exclusively from the maternal genome, and inheritance of an *Igf2r* deletion results in embryonic lethality only when inherited from the maternal genome (Barlow et al. 1991). This study and others also shed light on the question of why diploid genomes restrict gene expression to a single allele, which incidentally doubles the chances of expressing recessive mutations and therefore would reduce individual fitness. The most compelling evidence comes from the molecular function of

imprinted genes, which generally regulate embryonic growth (Barlow and Bartolomei 2014). For example, the insulin growth factor receptor gene *Igf2r*, which is expressed exclusively from the maternal genome, negatively controls organism size (Barlow et al. 1991; Haig and Graham 1991). Conversely, the paternally expressed gene *Igf2* promotes embryo growth (Ferguson-Smith et al. 1991). Since the majority of imprinted genes display such reciprocal growth phenotypes depending on the affected parental allele, the prevailing theory behind the evolution of genomic imprinting is the parental conflict or kinship hypothesis, which suggests that parental genomes are locked in a tug-of-war competition that mediates embryo size by controlling maternal resource transfer to the developing embryo (Haig 2000; Wilkins and Haig 2003). Another related driving factor behind the evolution of imprinted genes could be to prevent the generation of parthenotes, as these would result in a dramatic decrease in genetic variation, as discussed above.

Regardless, while the mechanism underlying parental-genome specific expression of imprinted genes, and transcription regulation in general, was elusive at the time, a few essential characteristics of genomic imprints were posited. As outlined by Barlow, the putative “genomic imprint” has four key features: 1. imparts allele-specific expression, as discussed above, 2. is established during gametogenesis, when the maternal and paternal genomes are separated 3. is maintained in a allele-specific manner following fertilization and 4. is reversed in the germline of the next generation (Barlow 1993). Recognition of these essential features, along with the parallel

development of the field of epigenetics (literally: on top of genes), converged in the discovery of the mechanism governing the monoallelic expression of several imprinted genes. The term “epigenetics” loosely encapsulates all mechanisms for gene-regulation that do not change the underlying DNA sequence (Surani 2001). Generally, epigenetic “marks” on chromatin have three key features: 1. the ability to regulate transcription without changing the underlying DNA sequence, enabling isogenic cells to differentiate into a myriad of cell types with distinct transcriptional programs (Waddington 1942), 2. maintenance of the epigenetic mark across mitotic divisions, conferring an epigenetic memory to cells and propagating their transcriptional programme, and 3. reversibility, for resetting cellular commitments.

1.3 DNA methylation and histone post translational modifications (PTMs) in mammals play an instructive role in epigenetic gene regulation

The most well studied epigenetic mark is DNA methylation (DNAme), a reversible chemical modification to DNA found almost exclusively at 5’-cytosine-guanine-3’ (CpG) dinucleotides in mammals (**Figure 1.1**) (Ramsahoye et al. 2000). Incidentally, global DNAme levels differ dramatically between the oocyte and sperm genomes. Thus, DNAme was the prime candidate for the “genomic imprinting” that occurs in gametes, which results in parental genome-specific gene regulation and likely explains the morbidity of uniparental embryos (Barlow 1993; Reik et al. 1993).

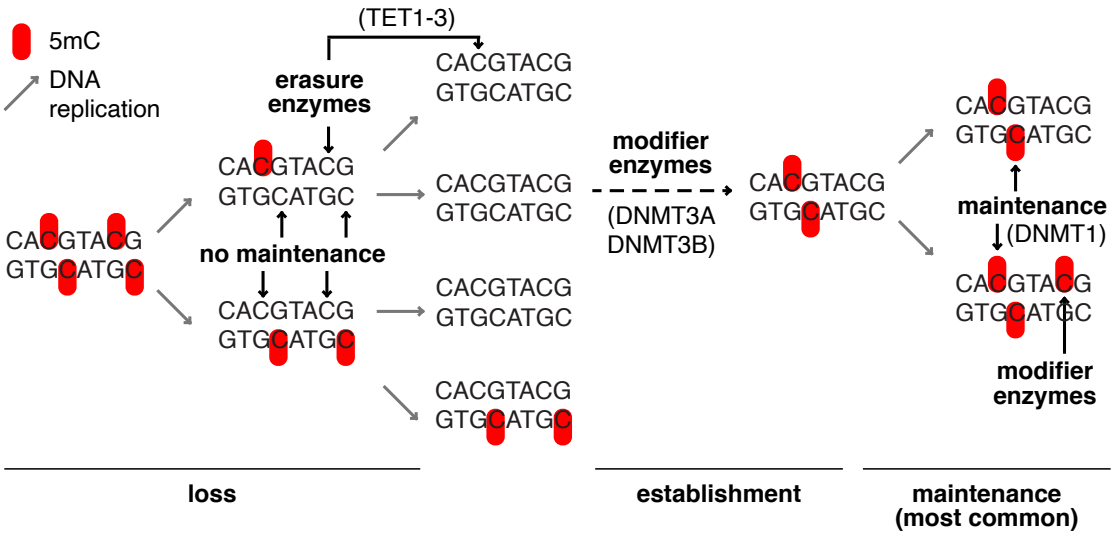

Figure 1.1 A basic model for DNAme establishment, maintenance and erasure. 5mC deposited by modifier enzymes can positively or negatively regulate transcription. Since DNA polymerase does not recognize nor deposit 5mC, a hemimethylated state is created following DNA synthesis (methylated at only one CpG dinucleotide on the Watson and Crick strands), which can be recognized by maintenance enzymes that propagate 5mC to the newly synthesized daughter strand. Finally, either active conversion of 5mC to C or passive dilution of DNAme by inhibition of maintenance enzymes coupled with DNA replication resets the epigenetic mark. Adapted from (Holliday and Pugh 1975) and (Sontag et al. 2006).

Since its conceptualization as an epigenetic mark, DNAme has been experimentally validated as being strongly associated with mammalian transcriptional regulation (Robertson 2005; Greenberg and Bourc'his 2019). Classic examples include X-chromosome inactivation (XCI) (Riggs and Pfeifer 1992; Beard et al. 1995), transposable element silencing (McClelland 1981; Walsh et al. 1998) and **genomic imprinting** (Reik et al. 1993; Li et al. 1993), each of which implicate DNAme in

epigenetic silencing of transcription. For example, once XCI occurs, all daughter cells inherit the same densely methylated, transcriptionally inactive X (Riggs and Pfeifer 1992). In addition, DNAme participates in restraining the proliferation of endogenous retroviruses through transcriptional silencing of their promoter elements (Bourque et al. 2018). Furthermore, during genomic imprinting, differentially DNA methylated regions inherited from gametes are maintained in an allele-specific manner and control allele-specific gene expression in the adult. For example, *Mest* is a paternally expressed gene in the mouse mesoderm with repression of *Mest* on the maternal allele by DNAme inherited from the oocyte over its genic promoter (Lefebvre et al. 1998; Brind'Amour et al. 2018) (**Figure 1.2**). In other words, DNAme marks, or imprints, established before fertilization can be faithfully maintained in an allele-specific manner in the adult and result in monoallelic gene expression. In the context of F1 hybrid strains, where both parental alleles are exposed to the same nuclear environment, allelic variation of DNAme levels at genomic imprints can be directly assessed using a within-sample control – the other allele (Wittkopp et al. 2004; Pastinen 2010).

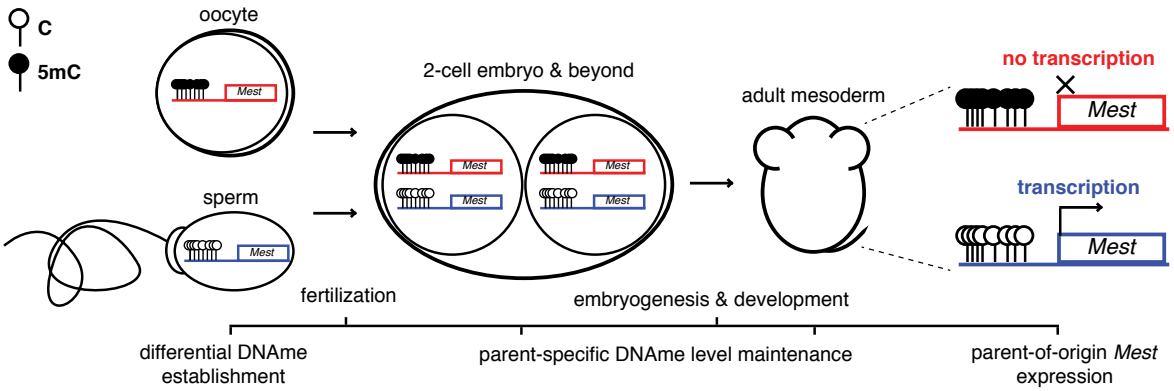

Figure 1.2 Genomic imprinting of the paternally expressed gene *Mest*.
The *Mest* promoter element is a CpG rich sequence that is densely methylated in oocytes and hypomethylated in sperm. Following fertilization, parent-specific DNAme levels are maintained in the embryo and beyond, including the adult mouse mesoderm. In association with allele-specific promoter DNAme levels, monoallelic transcription of *Mest* is observed from the hypomethylated paternal allele. As a consequence of this genomic imprinting, maternal transmission of a *Mest* deletion shows no abnormal phenotype while paternal transmission results in embryonic growth retardation and abnormal maternal behaviour (Lefebvre et al. 1998). The maternally inherited alleles are shown in red, paternal alleles in blue.

The two major mechanisms behind DNAme-mediated transcriptional silencing are by 1. steric inhibition of transcription factors (TFs)/RNA polymerase from chromatin (through methyl-binding domain proteins or independently) and 2. directly modifying TF binding motifs (Yin et al. 2017). In line with the role of DNAme in stably silencing transcription, fluctuations in DNAme levels have been linked to oncogene expression and cancer progression (Feinberg and Vogelstein 1983; Robertson 2005; Jones and Baylin 2007), genome instability (Xu et al. 1999) and imprinting disorders (Robertson 2005). As such, a general consensus emerged that DNAme is a “stable” epigenetic mark and that fluctuations in DNAme levels are detrimental.

In addition to DNAme, studies in the 1960s revealed that histone tail post-translational modifications (PTMs), including methylation and acetylation, could reduce or promote transcription (Allfrey et al. 1964; Lewis 1978). In the following years,

numerous studies revealed that methylation of DNA and specific histone residues generally inhibit transcription by reducing chromatin accessibility to TFs and the transcriptional machinery (Rose and Klose 2014). For example, trimethylation of histone 3 lysine 9 (H3K9me3) is frequently found at transcriptionally silent telomeres and centromeres in association with Heterochromatin Protein 1, which plays an important role in chromatin compaction (Rose and Klose 2014). Importantly, DNAme and H3K9me3 are also found at genic promoters of transcriptionally silent genes in euchromatin regions (Karimi et al. 2011; Auclair et al. 2016). Of note, while histone PTMs have also been associated with transcription regulation (Fischle et al. 2003), including at imprinted genes (Umlauf et al. 2004; Weaver and Bartolomei 2014; Inoue et al. 2017), the mechanism behind mitotic inheritance of these epigenetic marks is less well characterized (discussed below).

1.4 DNA methylation dynamics during development

As outlined in **Figure 1.1**, DNA methylation-mediated spatial and temporal gene expression requires, in addition to stable DNAme maintenance, mechanisms for regulated establishment and erasure. The difficulty in studying these two critical aspects of DNAme homeostasis is the need of a system in which cellular differentiation can be modelled. A critical system for studying the dynamics of DNA methylation is the developing mammalian germline and early embryo, where global DNAme levels are established and erased in two “waves” of genome-wide DNAme remodeling (**Figure**

1.3). The preimplantation embryo (E0-3.5) in particular is the primary model system used in this thesis.

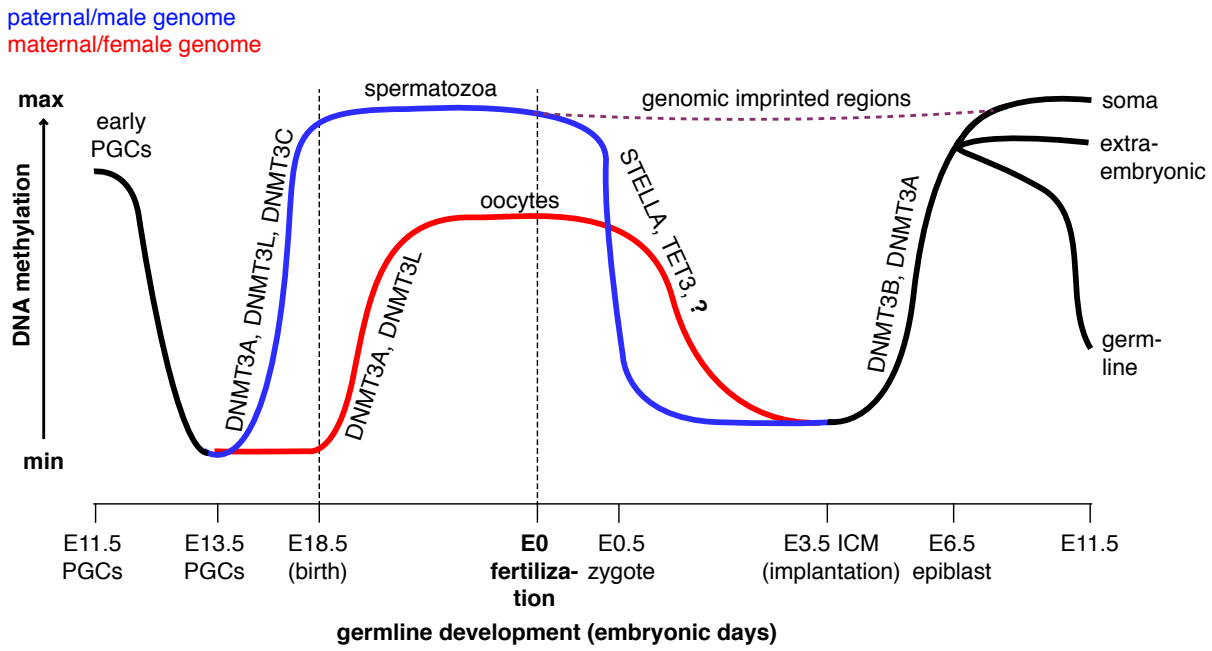

Figure 1.3 Global DNAme levels are dynamic in the mouse germline.
Primordial germ cell (PGC) genomes specified in the epiblast gradually lose DNAme as they proliferate and colonize the genital ridge, with DNAme reaching a low point ~E13.5. Once DNAme levels are reset in PGCs, de novo DNAme establishment (see section 1.4.1) occurs in a sex-specific manner, resulting in hypermethylated sperm and intermediately methylated oocyte genomes. Immediately following fertilization, the paternal genome undergoes active demethylation in the E0.5 zygote (see section 1.4.3). More recent evidence suggests the maternal genome also undergoes active demethylation **(Wang et al. 2014)**. Following the first S-phase of the zygote, the maternal and paternal genomes undergo passive DNAme loss, reaching a nadir in the inner cell mass (ICM) of the E3.5 blastocyst (see section 1.4.2). De novo DNAme establishment following implantation is largely complete by E6.5 in the epiblast, which differentiates into somatic cells of the embryo proper. Of note, while the

hypermethylated allele of genomic imprinted regions does not undergo DNA demethylation following fertilization, the hypomethylated allele must also be protected from de novo DNAme activity in the epiblast, maintaining parental differences in DNAme levels in the embryo and beyond (Proudhon et al. 2012). Global DNAme levels are relatively lower in extra-embryonic tissues derived from the trophectoderm of the blastocyst such as the placenta. Finally, around E6.25, the germline of the subsequent generation is specified from the epiblast. Adapted from (Smallwood and Kelsey 2012; Lee et al. 2014; Greenberg and Bourc'his 2019).

1.4.1 DNA methylation establishment

In mouse, DNA methyltransferases DNMT3A, DNMT3B and the recently discovered DNMT3C convert unmodified cytosine to 5mC de novo. A fourth member of the Dnmt3 family, DNMT3L, lacks a functional methyltransferase domain (**Figure 1.4**), yet clearly enhances DNMT3A activity in both the male and female germline (Hata et al. 2002; Bourc'his et al. 2001; Webster et al. 2005). As an aside, DNMT2 was renamed TRDMT1 as it methylates cytosine on transfer RNA and does not display DNA methyltransferase activity (Tuorto et al. 2015). While the catalytically active de novo DNA methyltransferases are structurally similar, their roles in DNAme establishment are non-redundant as they have divergent temporal expression profiles and genomic targets (Bestor and Bourc'his 2004; Edwards et al. 2017).

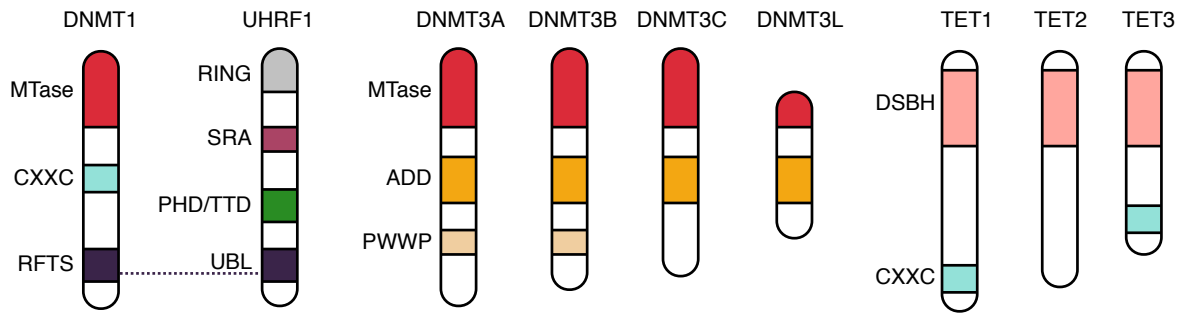

Figure 1.4 Structural representation of key DNAme homeostasis enzymes. **DNMT1**: maintenance enzyme that is recruited to replication forks and UHRF1 binding sites through its replicating targeting foci sequence (RFTS). A CXXC domain targets unmodified CpGs and a methyltransferase (MTase) domain actively converts cytosine to 5-methylcytosine. **UHRF1**: maintenance enzyme co-factor that specifically recognizes hemimethylated DNA via its SET- and RING- associated (SRA) domain. The RING domain ubiquitinylates H3, a modification recognized by DNMT1's RFTS domain. The ubiquitin-like (UBL) domain also recruits DNMT1, and the TUDOR (TTD) and plant homeodomain (PHD) regions recognize H3K9me2/3. **DNMT3A & DNMT3B**: de novo DNA methyltransferases containing a MTase domain, an ATRX-DNMT3-DNMT3L (ADD) domain that recognizes unmodified H3K4 and a Pro-Trp-Trp-Pro (PWWP) domain that recognizes H3K36me3. **DNMT3C**: lacks a PWWP domain. **DNMT3L**: lacks a PWWP and functional MTase domain. **TET1 & TET3**: DNA methylation "erasure" enzymes that oxidate 5mC via their double-stranded beta-helix domain (DBSH) and target DNA via a CXXC domain. **TET2**: lacks a CXXC domain. Adapted from (Barau et al. 2016; Jeltsch and Jurkowska 2016; Edwards et al. 2017; Greenberg and Bourc'his 2019).

The impact of this differential expression pattern of de novo DNMTs is best illustrated by the two waves of global DNAme establishment that occur during mouse development, the first in the developing germline and the second in the post-

implantation epiblast (**Figure 1.3**). The first wave occurs in the developing germline in a sex-specific manner and results in distinct methylomes in mature gametes (oocyte and sperm). While the mouse male germline initiates de novo DNA methylation one week before birth (embryonic day 16.5, E16.5), the female germline initiates de novo DNA methylation following ovulation (post-natal day 21 (P21) and beyond). While *Dnmt3L, Dnmt3a* and to a lesser extent *Dnmt3c* are all required for the proper establishment of DNAme of the developing male germline (Bourc'his and Bestor 2004; Kaneda et al. 2004; Barau et al. 2016), only *Dnmt3L* and *Dnmt3a* are required for de novo DNAme in the female germline (Bourc'his et al. 2001; Kaneda et al. 2004) (notably, *Dnmt3b* gene products are not expressed in either germline). Full-body *Dnmt3L* and *Dnmt3c* knock-out (KO) males are viable but infertile as they do not produce spermatozoa, whereas full-body *Dnmt3a* KO mice die shortly after birth. Interestingly, in addition to the morbidity of *Dnmt3a* KO, stillbirth males show impaired spermatogenesis (Okano et al. 1999). In contrast, female juvenile oocytes lacking *Dnmt3a, Dnmt3c* or *Dnmt3L* (maternal gene knock-out, matKO) develop into physiologically normal and fertile mature oocytes (Bourc'his et al. 2001; Barau et al. 2016). However, embryos generated from *Dnmt3a* or *Dnmt3L* KO oocytes die around E9.5 (Bourc'his et al. 2001; Kaneda et al. 2004). Thus, unlike the male germline, where proper DNAme establishment is essential for gamete production (Bourc'his and Bestor 2004), DNAme in the female germline does not affect meiosis or oocyte viability. Nevertheless, the lethality of *Dnmt3a* and *Dnmt3L* matKO embryos clearly indicates maternal DNAme levels are maintained to some extent following fertilization and impact the expression of genes

critical for normal development. Indeed, imprinted genes *Peg3* and *Snrpn* show biallelic expression (ectopic activation of the maternal allele) in embryos generated from *Dnmt3a* or *Dnmt3L* oocytes, but no increase in expression levels in oocytes themselves (Kaneda et al. 2004). Interestingly, while maternal stores of DNMT3A are detected in the fertilized zygote to the 8-cell stage (Hirasawa et al. 2008), global DNAme levels are reduced during this period (**Figure 1.3**, discussed in detail below). Notably, due to the low-throughput nature of the assays available to study DNAme at the time, the extent of genome-wide demethylation and associated ectopic transcription of genes and transposable elements in *Dnmt3* KO preimplantation embryos remained uncharacterized.

The second wave of global DNAme establishment occurs in the post-implantation embryo during the transition from blastocyst to epiblast (between E6.5 and E9.5) (Borgel et al. 2010). Of note, while both DNMT3A and DNMT3B are functional at this stage, only DNMT3B is required for the proper methylation and transcriptional repression of a specific subset of genes expressed exclusively in the germline, suggesting that DNMT3A and DNMT3B have divergent target specificities (Borgel et al. 2010) (see section 1.4.4). Importantly, while DNAme establishment at this stage resolves parental DNAme level differences at “transient” genomic imprints (defined below), protection from de novo DNAme in the post-implantation embryo plays a critical role in maintaining parent-specific DNAme levels at classical imprinted loci (Proudhon et al. 2012). The DNAme pattern established during the transition from blastocyst to

epiblast is largely reflected in the adult soma (Wang et al. 2014). Therefore, perturbations to DNAme homeostasis during this developmental stage have potentially life-long consequences (Greenberg et al. 2016).

Interestingly, a recent study employing immunofluorescence and ultra-sensitive liquid-chromatography-mass-spectrometry revealed that the paternal genome is unexpectedly subject to low levels of de novo DNAme immediately following fertilization (Amouroux et al. 2016). This observation contradicts the dogma that paternal DNAme are globally removed at this stage (**Figure 1.3**) (Santos et al. 2002; Mayer et al. 2000; Oswald et al. 2000). Unfortunately, neither of the methods employed are suitable for identifying the specific loci subject to de novo DNAme. Furthermore, the role that such enigmatic de novo DNAme of the paternal genome following fertilization plays in the regulation of affected genes was not explored in this study.

1.4.2 DNA methylation maintenance

DNAme is often referred to as a stable epigenetic mark because it can be maintained with high fidelity across cell divisions in a clonal cell population (Lorincz et al. 2002). This mitotic inheritance of 5mC relies on the symmetry of CpG dyads; as shown in **Figure 1.1**, following genome replication, methylated CpG dyads become hemimethylated on daughter strands. DNA methyltransferase 1, (DNMT1), the first DNMT discovered in mammals (Bestor et al. 1988), deposits 5mC on the unmethylated strand at hemimethylated CpGs, restoring 5mC patterns inherited from the parental cell

(Song et al. 2011). UHRF1, a DNMT1 co-factor (Sharif et al. 2007), aids DNMT1 catalytic activity by “flipping-out” the unmodified cytosine at hemimethylated CpGs from the inside of the double helix (Hashimoto et al. 2008). Notably, UHRF1 is preferentially recruited to chromatin enriched for histone 3 lysine 9 methylation (H3K9me2-3) moieties, an epigenetic mark associated with transcription repression (Rothbart et al. 2012). As a result of this interplay, H3K9me3 and DNAme are often co-localized over transcriptionally silent heterochromatic regions (Rose and Klose 2014; Liu et al. 2014). Recent studies reveal that UHRF1-mediated ubiquitylation of histone H3 also plays a critical role in DNMT1 recruitment and catalytic activity (Ishiyama et al. 2017; DaRosa et al. 2018) (see section 1.4.4 for more information on the interplay between histone PTMs and DNAme). Together, DNMT1 and UHRF1 show high fidelity of DNAme maintenance.

The importance of DNAme maintenance is highlighted by the phenotype of *Dnmt1* KO mice, which display significantly reduced global levels of 5mC in E10.5 embryos, undergo abnormal embryogenesis, and die early in gestation (E9.5-10.5) (Li et al. 1992). On the transcription level, these non-viable conceptuses are characterized by loss of monoallelic expression of imprinted genes (Li et al. 1993) and ectopic expression of transposable elements (Walsh et al. 1998). *Uhrf1* KO embryos have a phenotype indistinguishable from that of *Dnmt1* KOs (Sharif et al. 2007). Based on these observations, a general consensus has emerged that failure to maintain DNAme patterns during early embryogenesis directly results in the misexpression of developmentally important genes such as imprinted genes, which in turn result in

embryo death. However, the extent to which aberrant DNAme outside of imprinted regions affect the phenotypic consequences of developing KO embryos remains unclear. Recently, the advent of inexpensive high-throughput sequencing technology has enabled the study of DNAme dynamics genome-wide with single nucleotide resolution (discussed below).

In addition to its role in DNAme maintenance, DNMT1 was recently shown to have de novo DNAme activity in non-dividing cells, including the growing oocyte (Maenohara et al. 2017; Li et al. 2018; Han et al. 2019). Further, maternal factor STELLA protects the oocyte genome from such de novo DNAme activity by sequestering UHRF1 and DNMT1 to the cytoplasm (Li et al. 2018). Following fertilization, UHRF1 sequestration from the nucleus is observed until at least the blastocyst stage (Maenohara et al. 2017), which may explain the disparate observations that DNAme levels gradually decrease in a replication-dependent manner following fertilization (**Figure 1.3**) (Smith et al. 2012), despite the presence of DNMT1 and UHRF1 transcripts and proteins at this stage. Interestingly, 2-cell embryos generated from *Stella* deficient oocytes have a hypermethylated genome compared to wild-type. Further, these abnormal embryos show dramatic developmental failure beyond the early 2-cell stage associated with failure to undergo zygotic genome activation (ZGA), presumably as a failure to transcribe aberrantly densely methylated gene promoters (Li et al. 2018). Unfortunately, the lack of cell-type-specific gene KO constructs (conditional alleles) that enable the deletion of a gene immediately following fertilization precludes

the isolation of the role of STELLA in global DNA demethylation in preimplantation embryos from its role in protecting the oocyte genome from de novo DNAme activity, i.e. we do not know whether the lethal phenotype of *Stella* KO 2-cell embryos is the result of abnormal DNAme patterns established in the oocyte or from abnormal DNAme maintenance following fertilization.

Regardless, despite inhibition of DNMT1 activity by maternal stores of STELLA, a subset of regions maintains parental DNAme levels (inherited from the gametes) following fertilization. Loci refractory to DNAme loss include genomic imprints and intracisternal particle (IAP) endogenous retroviral elements (Lane et al. 2003; Smith et al. 2012; 2014), both of which, as mentioned above, are marked with H3K9me3 (Wang et al. 2018). Since a subset of these IAP elements are still capable of retrotransposition, persistent DNAme at these loci is likely critical for their repression and in turn, genome stability (Walsh et al. 1998). Thus, the complexities of DNAme level dynamics in the early embryo, including the enzymes responsible for DNAme level homeostasis over specific genomic loci, are beginning to unravel.

1.4.3 DNAme erasure

In addition to enzymatic machinery for the establishment and maintenance of DNAme, pathways for the removal of DNAme are also clearly operating during early embryonic and germline development.

Two mechanisms are involved in DNAme erasure; “passive” and “active”. Passive DNAme erasure is the result of inhibition of maintenance DNAme (i.e. DNMT1 activity) following DNA replication, resulting in a dilution of DNAme levels. Active DNAme erasure on the other hand is characterized by sequential oxidation of the methyl moiety followed by conversion to cytosine (see below). Of note, these mechanisms are not mutually exclusive. For example, oxidized versions of 5mC at hemimethylated CpGs are poorly recognized by maintenance DNMT1, which can also lead to passive erasure (Shen and Zhang 2013). Further, as mentioned above, active demethylation is observed on the paternal zygotic genome before the first S-phase (Mayer et al. 2000) followed immediately by DNAme level dilution of both alleles after each cell division, with DNAme levels reaching a nadir in the ICM of the blastocyst (Smith et al. 2012).

Enzymes that oxidize 5mC, effectively acting as “DNAme erasers”, were only recently identified (Tahiliani et al. 2009). These enzymes, named after a common Ten-Eleven Translocase (TET1-3) mutation frequently observed in blood cancers, sequentially oxidize 5mC to 5-hydroxymethylC, 5-formylC, and 5-carboxylC (Ito et al. 2010; 2011). Oxidized products are then converted to cytosine by base-excision repair enzymes (Branco et al. 2011; Bochtler et al. 2017). While TET1 and TET3 enzymes directly recognize CpG-dense regions, so-called “CpG islands”, through their zinc-finger-CXXC domain, TET2 lacks a DNA binding motif and is likely recruited to DNA via interaction with specific transcription factors (Sardina et al. 2018) (**Figure 1.4**).

While TET3, the sole TET enzyme expressed during embryogenesis, was implicated by a number of groups as a critical factor in the wave of active DNA demethylation of the paternal genome in the zygote (Hajkova et al. 2010; Gu et al. 2011a; Guo et al. 2014; Tsukada et al. 2015), subsequent studies revealed that this TET protein is responsible for only a fraction (~8%) of the roughly 40% reduction in paternal DNA demethylation levels (Peat et al. 2014), however, reports of the contribution of TET3 on active demethylation of the paternal and maternal genomes vary (Wang et al. 2014; Guo et al. 2014). Another set of enzymes that can theoretically remove 5mC are DNA deaminases, which directly convert 5mC to thymine, creating a T-G mismatch which is then recognized and corrected by the base excision repair (BER) pathway. However, widespread DNA deaminase directed paternal DNA demethylation would be surprising, as introducing millions of T-G mismatches in the zygote is likely highly mutagenic. As such, two important questions remain unanswered: 1. what is the identity of the enzyme(s) responsible for the dramatic loss of 5mC on the paternal genome in the zygote, and 2. why is DNAme loss in the zygote restricted to the paternal genome (Messerschmidt 2016).

Finally, during germ cell migration and proliferation (~E8.5-E11.5 in mice), global DNAme levels are again dramatically reduced. Indeed, by E12.5, almost all regions, including imprinted regions, are hypomethylated (genome-wide average 7-14%) (Szabo and Mann 1995; Seisenberger et al. 2012). Loss of methylation at this stage is likely

due to downregulation of the de novo DNAme machinery and sequestration of UHRF1 in the cytoplasm (Seisenberger et al. 2013b). Unlike in the blastocyst, where imprinted regions are maintained, DNAme at imprinted loci is erased in PGCs, at least in part by the activity of TET1 and TET2 (Hackett et al. 2013). Following DNAme erasure, reestablishment of DNAme levels, including at genomic imprinted regions, occurs in a sex-specific manner; in prospermatogonia in male and growing oocytes in female. For a more complete review of DNAme erasure and establishment in gametes, see (Trasler 2006; Reik 2007; Seisenberger et al. 2013a; Meyenn and Reik 2015).

While TET enzymes “actively erase” DNA, it is difficult to distinguish their oxidative activity from their role in inhibiting DNMT activity. Indeed, a recent study using the hematopoietic system elegantly demonstrates that, in addition to converting 5mC to cytosine, TET1 sterically inhibits DNMT3A-mediated DNAme deposition (Gu et al. 2018). Regardless, TET enzymes ultimately reduce DNAme levels, and nicely fit within the putative DNAme “erasers” model, although other DNAme “eraser” enzymes likely exist.

Altogether, these three classes of enzymes, implicated in maintenance, establishment and erasure, compete and cooperate to maintain DNAme levels. The importance of DNAme level homeostasis is highlighted by mutations in these enzymes, which are commonly found in human diseases such as in cancers, immunodeficiency,

centromeric instability and facial anomalies (ICF1) and Tatton-Brown-Rahman syndromes (Tatton-Brown et al. 2014; Robertson 2005).

1.4.4 The interplay between DNA methylation and histone H3 methylation

As discussed previously, UHRF1 guides DNMT1 to hemimethylated CpGs to promote maintenance DNAme, but what confers this specificity? As shown in **Figures 1.4 & 1.5**, the SET- and RING- Associated (SRA) domain of UHRF1 specifically recognizes hemimethylated DNA, while its Ubiquitin-like (UBL) domain is recognized by the replication foci targeting sequence (RFTS) domain of DNMT1. In addition, the TUDOR (TTD) and plant homeodomain (PHD) domains of UHRF1 recognize histone PTM H3K9me2-3 moieties, which further promotes DNAme maintenance at these sites. *In vivo*, while mouse primordial germ cells are globally hypomethylated, as discussed above, low levels of DNAme persist at certain H3K9me3-enriched repeats, such as IAP elements in mouse primordial germ cells (Lees-Murdock et al. 2003; Lane et al. 2003; Seisenberger et al. 2012; Liu et al. 2014). In addition, DNAme levels over H3K9me3-enriched regions are refractory to the wave of global DNAme loss in the early embryo (Wang et al. 2018). Human fetal germ cells also show increased DNAme levels over evolutionarily young long interspersed nuclear elements (LINEs) LINE-1 and LINE-2 (Guo et al. 2017), likely in association with H3K9me3. Such H3K9me3-associated DNAme may confer transgenerational epigenetic inheritance, although evidence for this phenomenon in mammals is lacking (Perez and Lehner 2019).

De novo DNMTs are also guided to chromatin by the methylation status of specific residues on the histone H3 tail. Mechanistically, two domains confer site specificity to DNMT3A and DNMT3B: an ATRX-DNMT3-DNMT3L (ADD) domain and a Pro-Trp-Trp-Pro (PWWP) domain (**Figures 1.4 & 1.5**). The ADD domain interacts with unmethylated H3K4 histones (H3K4me0) (Ooi et al. 2007; Rose and Klose 2014), which directs de novo DNAme to genomic regions lacking H3K4 methylation (Greenberg and Bourc'his 2019). The exclusivity of H3K4 and DNA methylation is particularly striking at CpG-islands (CGIs, defined in detail below) (Erkek et al. 2013). While CGI promoters of actively transcribed genes show DNA hypomethylation and H3K4me3 enrichment, transcriptionally inactive CGIs are hypermethylated and devoid of H3K4me3 (Deaton and Bird 2011).

On the other hand, the PWWP domain has been shown to interact with histone 3 specifically when methylated on lysine 36. DNMT3A for example interacts with H3K36 trimethylation (H3K36me3), and to a lesser extent H3K36me2 *in vitro* (Dhayalan et al. 2010) and in embryonic stem cells (Weinberg et al. 2019) via the PWWP domain. Furthermore, DNMT3B selectively binds to actively transcribed gene bodies in embryonic stem cells through direct interaction of the PWWP domain and H3K36me3 (Baubec et al. 2015). As H3K36me3 is deposited over actively transcribed loci by the lysine methyltransferase (KTMase) SETD2 (Kizer et al. 2005), there is strong correlation between DNAme levels, active transcription and H3K36me3 enrichment in embryonic stem cells and mature oocytes (Kobayashi et al. 2012; Smallwood et al. 2011; Baubec

et al. 2015; Brind'Amour et al. 2018; Xu et al. 2019). Indeed, DNAme establishment is dramatically impaired in SETD2 KO oocytes (Xu et al. 2019). Further, gain-of-function point mutations that abolish the PWWP domain in mouse and human DNMT3A result disruption of H3K36me3 interactions and in abnormal DNAme patterns characterized by hypermethylation at bivalent chromatin, resulting in severe growth retardation (Heyn et al. 2019; Sendzikaite et al. 2019). In contrast, haploinsufficient mutations that abolish DNMT3A function entirely are associated with overgrowth in patients with Tatton-Brown-Rahman syndromes (Tatton-Brown et al. 2014).

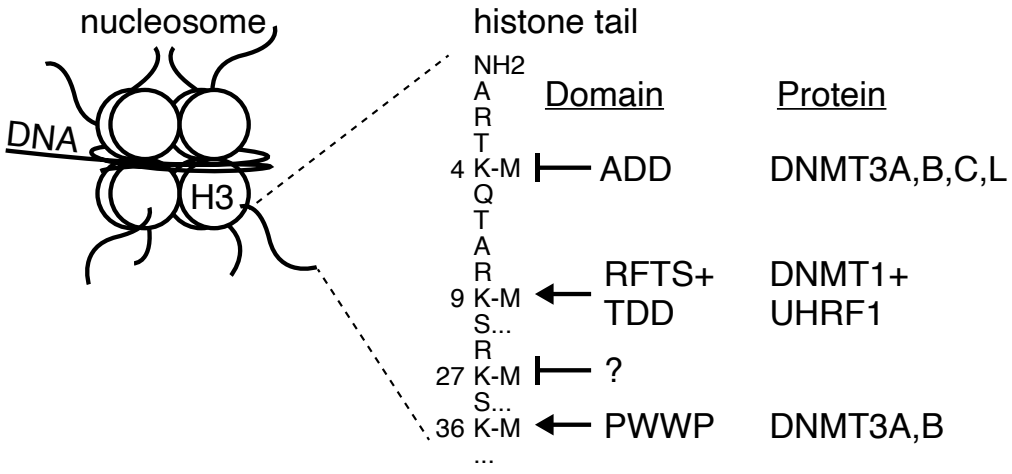

Figure 1.5 Specific histone H3 tail modifications inhibit or promote DNMT activity. **Left:** the nucleosome is the basic unit of chromatin and consists of ~147 bp DNA wrapped around two copies of each core histone protein: histone 2A (H2A), H2B, H3 and H4. **Right:** Methylation of lysine (K) residues on the H3 tail can enhance (H3K36) or inhibit (H3K4) *de novo* DNMT activity while H3K9me3 promotes maintenance DNAme activity by recruiting UHRF1 and, subsequently, DNMT1 via it's Replicating Target Foci Sequence (RFTS) domain. Specific DNMT domains (**Figure 1.4**) involved in such interactions are included. Adapted from (Fischle et al. 2003).

DNMT3B plays a predominant role in de novo DNAme in the epiblast, resulting in ~70-80% global DNAme levels. Interestingly, site-specific targeting of DNMT3B to the promoters of germline genes is associated with H3K9me3 KMTases and the non-canonical polycomb-group (PcG) complex PRC1.6 via transcription factors E2F6 and MGA (Endoh et al. 2017; Tatsumi et al. 2018), likely independent of the PWWP domain. Additionally, patients with recessive mutations in the MTase domain of *DNMT3B* show chromosomal instability in association with aberrant hypomethylation of juxtacentromeric satellite repeats (Xu et al. 1999; Okano et al. 1999). Thus, DNMT3A and DNMT3B have both overlapping and unique target specificities yet the sequence specific factors that guide these DNMTs to specific genomic loci beyond the CpG rich promoters of germline genes remain largely unknown.

In contrast to H3K9me3, H3K27me3, another “repressive” histone PTM (Simon and Kingston 2013), generally marks CGI promoters that lack DNA methylation (Boulard et al. 2015). Indeed, studies of DNMT mutants reveal that DNAme likely inhibits the deposition of H3K27me3 (Heyn et al. 2019). This relationship is likely driven by the fact that subunits of both PRC1 and PRC2 PcG complexes have affinity for unmethylated CpG-rich sequences (Blackledge et al. 2014). Notably, the complement of H3K4me0, H3K36me3 and H3K27me3 regions do not account for the entirety of the methylome, suggesting that additional chromatin factors may play a role in recruitment of DNMTs to specific genomic regions. While a multitude of histone PTMs are proposed to play a role in DNAme and transcription (Fischle et al. 2003; Sood et al. 2019), this thesis focuses

exclusively on methylation of lysines in histone H3 (reviewed here (Rose and Klose 2014)).

Interestingly, the adult sperm genome is globally hypermethylated except at H3K4me-enriched regions, suggesting that DNMT3A acts promiscuously in the male germline (Hammoud et al. 2009; Erkek et al. 2013). While DNMT3C targets young repetitive elements in the male germline for de novo DNAme, it lacks a PWWP domain and is likely guided to target sequences independent of H3K36 methylation (**Figure 1.4**) (Barau et al. 2016). As a result of differential DNAme establishment by DNMT3A and DNMT3L in the male and female germlines, the sperm genome is ~85% methylated (similar to somatic cells) while the oocyte genome is roughly 50% methylated, almost exclusively over actively transcribed DNA (Kobayashi et al. 2012; Veselovska et al. 2015; Brind'Amour et al. 2018; Greenberg and Bourc'his 2019).

As discussed above, classical genomic imprints arise from differences in DNAme levels in adult gametes (gametic DMRs) and are maintained following fertilization, imparting parent-of-origin transcription in the embryo and the adult. However, the role of DNAme differences that clearly exist between sperm and oocytes outside of classical imprinted regions remains to be determined. While numerous genetic studies reveal that DNMTs are essential for proper embryonic development and establishment/maintenance of DNAme over transposable element and genomic imprints, we do not fully understand the molecular basis for the morbidity of *Dnmt* KO

mice. Indeed, little is known about the effect of DNMT loss on transcription, DNAme and histone PTM levels at regions outside of classical genomic imprints. Recent work demonstrated that methylation inherited from the oocyte regulated non-imprinted germline gene transcription in the preimplantation embryo (Borgel et al. 2010; Rutledge et al. 2014). Interestingly, a recent study focused on the post-implantation trophoblast lineage (placenta precursor cells) showed that DNAme inherited from the oocyte silences non-imprinted gene expression, indicating DNAme can indeed be inherited outside of imprinted regions and regulate transcription in the post-implantation embryo (Branco et al. 2016).

1.5 Classical-, transient- and placental-specific genomic imprinting

As discussed previously, parental DNAme levels can be maintained following fertilization. Such maintenance of parental DNAme levels is pronounced at **classical genomic imprints**, differentially methylated regions (DMRs) established in the gametes (gametic DMRs, gDMRs) that instruct allele-specific gene expression in a parent-of-origin pattern in adult cells. Hence, gDMRs must survive the wave of DNAme removal during preimplantation development (**Figure 1.2**) (Seisenberger et al. 2013a; Smith et al. 2012). These gDMRs regulate the expression of well characterized imprinted transcripts such as *Igf2*, *Igf2r*, *Mest*, *Snrpn*, *Grb10*, *H19*, *Kcnq1ot1*, etc. (White et al. 2016). Interestingly, the majority of genomic imprints are of maternal origin (hypermethylated in the oocyte) in both mouse and human (additional details below). A feature of classical maternal gDMRs is that they are CpG islands, short stretches of

DNA (about 1kb) with normal CpG content (1/16 nucleotides) that interrupt an otherwise depleted “ocean” of CpG-poor (1/100) regions, and frequently overlap transcription start sites. Recently, the advent of NGS technology revealed that non-imprinted CGIs are hypomethylated throughout the mammalian life cycle, including the early embryo and germline, which may explain how they have avoided the inherent mutability of 5mC and maintained normal CpG content over evolutionary timescales.

In mouse and human, 24 and 36 gDMRs govern parent-of-origin expression of surrounding genes, respectively (Court et al. 2014; White et al. 2016). While 17 of these gDMRs were shown to be conserved between these species, others require experimental validation (White et al. 2016). In other words, the imprinting status of several gDMRs is contentious. An emerging method for identifying biologically relevant gDMRs takes advantage of high-throughput sequencing (HTS) data to measure allele-specific DNA methylation levels and parent-of-origin gene expression genome-wide. However, this approach has been employed in surprisingly few mammalian species (generally limited to mouse and human) as it requires the generation of a reference genome. Comparing gDMRs and parent-of-origin gene expression in additional mammalian species, such as rats, rabbits, chimps, etc., which shared a last common ancestor ~90 million years ago (Kumar et al. 2017) is now possible with emerging reference genomes, and will elucidate which imprinted regions are conserved and likely impart developmentally important allele-specific gene expression, while those that arose in specific lineages will increase our understanding of the diversity of imprinted genes.

Another set of genomic imprints are defined as “somatic” as they are established during post-implantation development and result in parent-of-origin gene expression in adult cells (John and Lefebvre 2011). In other words, somatic imprinted regions are not differentially methylated between the adult gametes (hence: somatic). Since these post-implantation DMRs arise near gametic DMRs as a consequence of classical imprinted gene expression, somatic imprints are generally considered “secondary” imprints. Regardless, somatic imprints remain an important model for understanding how allele-specific DNAme level differences can arise when both genomes are exposed to the same nuclear milieu.

Recently, a new class of early-embryo-specific genomic imprints, referred to as “transient imprints” were identified. Such transient imprints are defined as gametic DMRs that are resolved by the implantation stage (Kobayashi et al. 2012; Proudhon et al. 2012; Tucci et al. 2019). Interestingly, while transient imprinted DMRs are resolved in the post-implantation embryo, the allele-specific transcription regulation they impart on genes in the early embryo can have life-long effects (Greenberg et al. 2016). Unfortunately, few transient imprints have been characterized in mouse and human, primarily due to the relative difficulty in studying this stage in development (discussed below).

Genomic imprints can be further classified by whether they confer parent-of-origin monoallelic gene expression in embryonic or extra-embryonic derived tissues. In line with the seminal findings that uniparental embryos show defects in extraembryonic development in mouse and human (discussed above), many genes are imprinted specifically in the placenta (Monk 2015). However, the extent and mechanism underlying placental-specific imprints is divergent between both species. For example, while mouse *Sfmbt2*, *Zfp64, Slc38a4* and *Gab1* show paternal expression in placenta, these genes are biallelically expressed in the human placenta (Monk 2015). Likewise, *Ascl2* and *Cd81* are maternally expressed in mouse but biallelically expressed in human placenta. Conversely, there are numerous putative imprinted genes in human, including *DNMT1* and *LIN28B*, that are biallelically expressed in mouse. Additionally, differences in human and mouse imprinting includes the primate-specific gene *ZNF331*, which shows placental imprinted expression and has no mouse homologue (Monk 2015). One important caveat in measuring imprinted gene expression in the placenta is the potential for maternal decidua contamination, which results in the overestimation of maternally expressed imprinted genes (Okae et al. 2012), and may explain the high number of placental imprinted genes identified in human, where embryonic collection is difficult. Intriguingly, many mouse placental imprints display DNAme-independent gene regulation, including H3K9 and H3K27 methylation-mediated silencing (Umlauf et al. 2004; Inoue et al. 2017), whereas human placental imprinted genes are generally located near gDMRs (Monk 2015).

Thus, *identifying non-classical imprints and measuring the extent to which they contribute to parent-of-origin allele-specific gene expression in the early embryo in mouse and other mammals will reveal the extent to which allele-specific transcription in the early stages of development has lasting effects on transcription at later stages in life.*

1.5.1 The interplay between allele-specific maintenance of DNAme and H3K9 methylation at genomic imprints

As discussed above, DNAme and repressive histone PTMs at CGI promoters negatively regulate gene expression. However, to impart allele-specific gene expression, including of imprinted genes, these epigenetic marks must be maintained in an allele-specific manner. Allele-specific DNAme maintenance at gametic DMRs is conferred by the methyl-sensitive TF ZFP57, which specifically recognizes the methylated state of its motif (TGCmCGCN) (Strogantsev et al. 2015; Liu et al. 2012) and recruits KAP1, a scaffold protein that mediates heterochromatin formation by in turn recruiting chromatin modifying enzymes, including the H3K9 KMTase SETDB1, which deposits H3K9me3 at these loci (Schultz et al. 2001; Rowe et al. 2010; Quenneville et al. 2011). As discussed above, H3K9me3 promotes DNAme maintenance. While ZFP445 has also been implicated in the maintenance of genomic imprints, its genomic targets largely overlap those of ZFP57 (Takahashi et al. 2019). Therefore, transcription factors such as ZFP57 and ZFP445 impart site- and allele-specific DNAme maintenance at genomic imprints, including during early embryogenesis when global DNAme levels are reset.

While genomic imprinting is generally thought to be imparted by allele-specific DNAme levels established in the gametes (Barlow 1993; Li et al. 1993), PcG protein-established H3K27me3 was recently shown to also play a role, particularly at placental imprinted genes (Inoue et al. 2017). While adult gametes have numerous differentially H3K27me3-marked regions, the mechanism behind the inheritance of this PTM in an allele-specific manner remains unknown.

1.6 Bioinformatic analysis of high-throughput sequencing data generated from early embryos

Embryogenesis provides a dynamic *in vivo* model to test the association between allele-specific DNAme and transcription, including at classical genomic imprints. Unfortunately, scant research has been conducted at this stage due to limited availability of samples. Typical HTS experiments require at least one million cells to achieve high-quality datasets, necessitating the collection of half a million 2-cell embryos to study this stage. Therefore, longstanding questions regarding the extent of DNAme-regulated transcription during embryogenesis could only be answered in the context of assays that required fewer cells for analysis, such as immunofluorescence (IF) or blotting of repetitive DNA. And while the global DNAme level dynamics during germline and embryo development measured by IF have been highly informative, they may not necessarily accurately represent DNAme levels at lower-copy functional loci such as CpG islands and imprinted regions.

Only recently have protocols with increased resolution progressively reduced the number of cells required for high-throughput sequencing assays (Lister et al. 2009; Lister and Ecker 2009; Brind'Amour et al. 2015), removing the primary roadblock in conducting HTS experiments on rare cell populations. Indeed, the first single-nucleotide resolution genome-wide maps of DNAme levels in the preimplantation embryo were only published in 2014 (Wang et al. 2014).

1.6.1 Most bioinformatic tools are agnostic to allele-specific phenomena

Despite the development of HTS technology and the generation of data from early embryos, the development of robust and standardized analysis software to process these datasets is forthcoming. Specifically, conventional HTS analysis software relies on a single **haploid reference genome** for read alignment and subsequent analyses such as DNAme and transcription level quantitation (Li and Durbin 2009; Langmead and Salzberg 2012; Krueger and Andrews 2011). This limitation has precluded the identification of parental-genome specific DNAme, histone PTM and transcription dynamics, such as at imprinted regions, in a genome-wide scale.

1.6.2 Inbred mouse strains and the reference genome

Inbred mouse strains have a critical advantage over human and other outbred individuals because all individuals are genetically identical (discussed above). As such, experiments conducted by mouse geneticists from different parts of the globe and eras

can control for genetic variation and their results can be directly compared. Further, the assembled genomic sequence of one individual represents the genome of all individuals of the same isogenic strain and is therefore termed the “reference genome”. As the most widely studied mouse strain, C57BL/6J (“C57 black 6 Jax”) was selected as the mouse reference genome. Indeed, the vast majority of the genetic experiments conducted in the cited literature in this thesis used the C57BL/6J mouse as a reference. Currently, the *Mus musculus* reference genome is on revision 10 (mm10), has one of the highest contiguity (N50) scores, and only a few (44) unplaced contigs (2019). In addition, mm10 has genome annotation resources on the location of protein coding genes, CpG islands, transposable elements, retrogenes, etc.

While such reference genomes are adopted globally, they neglect critical genetic variation between strains or outbred individuals. In other words, the major advantage of employing isogenic mouse strains for research comes with a major downfall; the lack of genetic variation information. To balance these two parameters, 18 commonly used lab mouse strains and 4 strains recently derived from the wild have been deeply sequenced (Keane et al. 2011). Reads generated from these genetically distinct mouse strains were aligned to the reference genome to create comprehensive maps of naturally occurring genetic variation. These maps include the location and identify of single nucleotide variants (SNVs), short insertions and deletions (INDELs), polymorphic transposable elements and large structural variants (Keane et al. 2011; Nellåker et al. 2012).

Despite the availability of this genetic variant information, HTS reads generated from genetically diverse individuals are typically aligned to a reference genome, introducing significant alignment bias that alters downstream calculations such as allele-specific transcription level quantification (Degner et al. 2009). This limitation applies to all types of HTS alignment software, including those designed for gapped read alignment (RNA-seq), short read alignment (ChIP-seq) and bisulphite-converted DNA alignments (WGBS). Therefore, due to shortcomings in disparate HTS analysis software, the full potential of HTS datasets generated by laboriously collecting preimplantation embryos has not yet been realized. *In this thesis, I leverage naturally occurring genetic variation (SNVs and INDELs) between parental chromosomes in F1 hybrid mice to identify maternal- and paternal-specific reads in HTS datasets with minimal reference alignment bias.*

1.6.3 Using allele-specific analysis of HTS data to study locus-specific parental DNAme dynamics in the embryo and germline

The developing embryo offers numerous loci that display dynamic DNAme levels from which we can study the interplay between parental-genome specific DNAme, histone PTMs levels and transcription. However, many of the changes in DNAme that occur during this stage (**Figure 1.3**) may not be biologically relevant, as regions undergoing such DNAme level remodeling are located in gene deserts with little or no known functional significance (Edwards et al. 2017). Indeed, the vast majority of promoter CpG

islands, which are generally transcriptionally inert when hypermethylated (Deaton and Bird 2011), tend to remain hypomethylated throughout these phases of DNAme resetting (Weber et al. 2007). Conversely, specific families of young transposable elements tend to remain hypermethylated (Smith et al. 2012). Only very recently have high-throughput sequencing datasets been generated to fully explore DNAme dynamics over various functional genomic loci in the early embryo (Li and Li 2019; Greenberg and Bourc'his 2019).

In this thesis, I address these shortcomings by combining the F1 hybrid mouse genetic model with novel software that enables integrated analysis of WGBS, ChIP-seq and RNA-seq data with allele-specific resolution to measure the inheritance and dynamics of DNAme and transcriptional programmes in the developing mouse embryo.

1.7 Thesis Goals

The goal of my thesis work was to revisit global DNAme dynamics during embryogenesis using allele-specific HTS data to measure the extent of DNAme inheritance from gametes, including at classical and transient genomic imprints, using the mouse as a model system. Additionally, I aimed to integrate this allele-specific DNAme data with histone PTM and transcription data to better understand the interplay between these marks and gene expression in the early embryo.

To this end, as discussed in **Chapter 2,** I helped develop MEA, a bioinformatic software that leverages known genetic variation, including SNVs and INDELs, to identify maternal- and paternal-genome specific reads in any given HTS dataset, including WGBS, ChIP- and RNA-seq. In **Chapter 3**, I identified CpG island promoters that are de novo DNA methylated specifically on the paternal allele of zygotes. Interestingly, this post-fertilization establishment of differentially methylated promoters is mediated by maternally deposited DNMT3A and suppresses the premature expression of blastocyst-stage genes. Since parent-specific DNAme levels at these loci are harmonized following implantation, these results suggest transient genomic imprinting can also be established following fertilization. Finally, in **Chapter 4,** I discuss the overall significance of the research presented in this thesis and the potential for future discoveries using MEA.

Chapter 2: Development and application of an integrated allele-specific pipeline for methylomic and epigenomic analysis (MEA)

2.1 Introduction

HTS-based approaches for genome-wide analysis of RNA, histone PTMs, DNAme and chromatin conformation are now routinely conducted on both model organisms and human samples. Such studies have yielded many insights into the interplay between chromatin structure and transcription, including the surprising observation that allele-specific phenomena may be more widespread than previously believed (Holliday 1990; Pinheiro and Heard 2017). Unfortunately, while such datasets, including RNA-seq, ChIP-seq, WGBS, are theoretically amenable to allele-specific profiling, HTS analysis software generally does not discriminate between parental alleles from diploid genomes. Indeed, popular read aligners depend on alignment to a single reference genome, essentially considering the sequencing reads generated from autosomes (and the X-chromosome in the case of females) as originating from isogenic rather than outbred individuals. In merging both parental alleles into a single measurement, these aligners neglect allele-specific phenomena, such as genomic imprinting (Holliday 1990), X-chromosome inactivation (Pinheiro and Heard 2017) and sequence-dependent *cis*-regulatory effects (Goncalves et al. 2012).

To overcome this shortcoming, a number of software packages have recently been developed that assign HTS sequencing reads to a specific parental allele. For example, MMSEQ (Turro et al. 2011), QuASAR (Harvey et al. 2015), MBASED (Mayba et al. 2014) and SCALE (Jiang et al. 2017) were designed to analyze RNA-seq data, while MethPipe (Song et al. 2013), epiG (Vincent et al. 2017) and BSPAT (Hu et al. 2015) were designed to process DNAme data. Several independent custom scripts for allele-specific analyses have also been reported (Wang et al. 2014; Leung et al. 2015; Zhang et al. 2016), but the details required for implementing them were not included. Pipelines such as Allelome.PRO (Andergassen et al. 2015), WASP (van de Geijn et al. 2015) and ALEA, a toolbox developed in our lab (Younesy et al. 2014), accommodate both RNA- and ChIP-seq datasets, yet no pipeline offers the additional capability of processing DNAme data. The lack of a universal allele-specific pipeline has precluded robust integration of allele-specific transcription, histone PTMs and DNAme profiles. Importantly, while such pipelines can be applied in parallel to analyze distinct epigenomic features, installation and implementation of multiple software packages can be time consuming, even for experienced bioinformaticians. Additionally, comparing allelic results generated using different software can introduce confounding factors, as the strategies used to process reads depend on multiple parameters, including read trimming, alignment mismatch scoring and read alignment filtering (mapping quality, PCR duplicate reads). For example, several allele-specific analysis packages rely on reference genome alignment followed by variant calling (Song et al. 2013; Andergassen et al. 2015; Hu et al. 2015), while others leverage publicly available SNV data to derive

a diploid genome for read alignment (van de Geijn et al. 2015; Harvey et al. 2015; Younesy et al. 2014). This "pseudogenome" strategy is a significant improvement over the former as it enables alignment over loci with high levels of genetic variation. However, current pipelines exclude INDELs for pseudogenome reconstruction, as they modify reference chromosome sequence lengths and annotated gene coordinates required for downstream analyses. Given the relative abundance of INDELs, this shortcoming may lead to the omission of a significant fraction of informative allelic reads. Indeed, analysis of high quality genotyping information for mouse strains reveals that, exclusive of large insertions or deletions (structural variants), short INDELs compose up to 20% of genetic variation (Adams et al. 2015). Thus, an INDEL-aware allele-specific pipeline that considers both SNVs and INDELs for pseudogenome reconstruction would represent a significant improvement over existing software.

Here, I present MEA, an “all-in-one” bioinformatic toolbox that exploits both SNVs and INDELs to enable allele-specific analyses of RNA-seq and ChIP-seq as well as WGBS datasets generated using short-read sequencing technology **(Figure 2.1a)**.

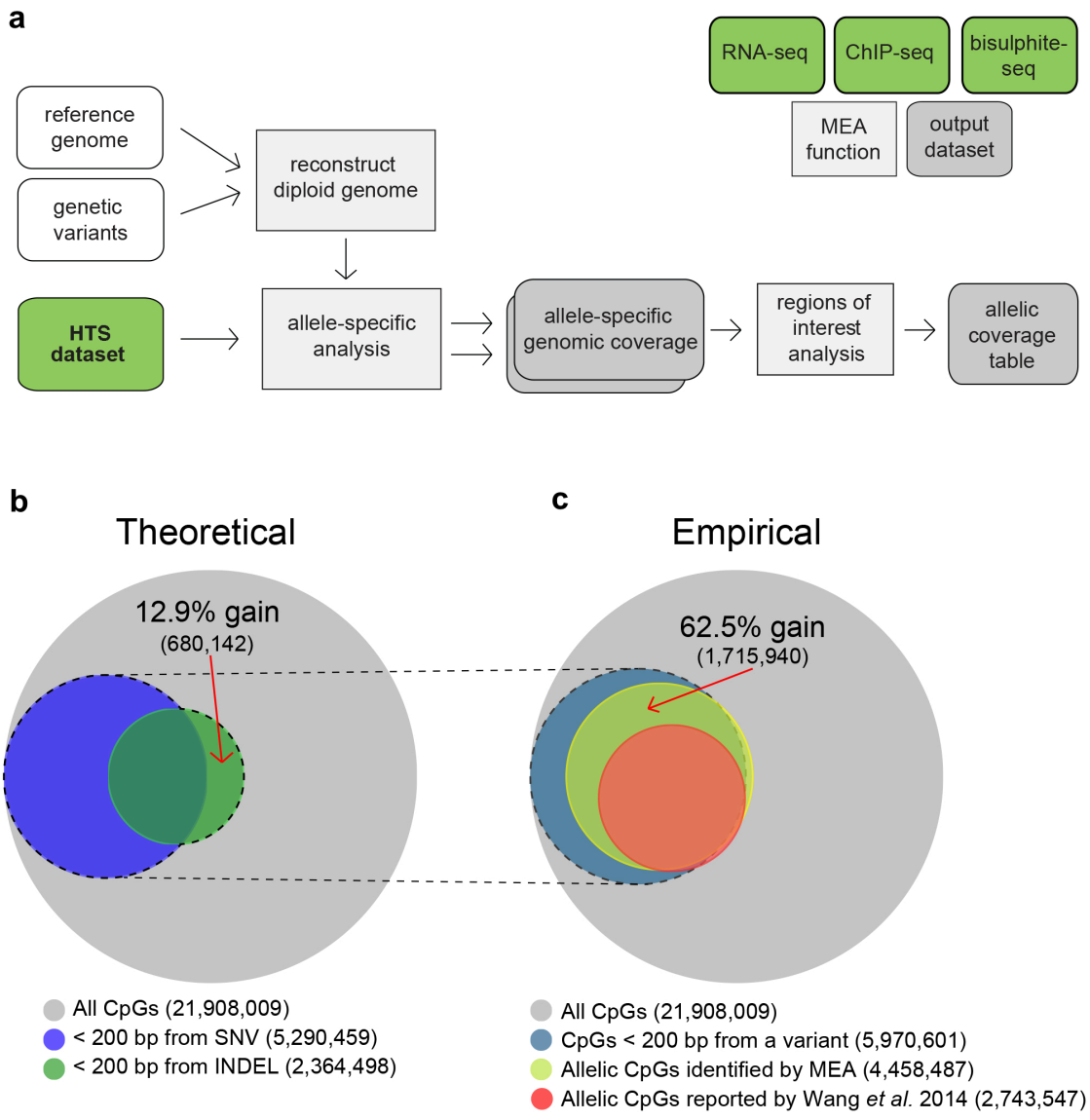

Figure 2.1 A bioinformatic toolkit for allele-specific epigenomic analysis.

(a) MEA pipeline flow chart. Supplied with a reference genome assembly and relevant genetic variants, MEA first reconstructs a diploid pseudogenome. Subsequently, with user provided gene expression (RNA-seq), histone PTM (ChIP-seq) or DNAme (WGBS) data in FASTQ format, allele-specific analysis is performed. MEA calculates allelic imbalance values using the resulting allele-specific genomic coverage files and generates a tab-delimited table for the user-defined regions of interest. Mouse and human exon, gene body and transcription start site coordinates are provided to facilitate

analyses of such regions. **(b)** Venn diagram showing the theoretical number of CpG dinucleotides for which allele-specific DNAme levels can be calculated using C57BL/6J and DBA/2J SNVs (blue) or INDELs (green) alone. CpGs for which allelic information can theoretically be extracted are defined as those that fall within 200 bp (an insert size typical of WGBS libraries) of a genetic variant. **(c)** Venn diagram showing the observed number of C57BL/6J-specific CpG dinucleotides for which allele-specific DNAme levels were calculated using MEA (yellow) versus an INDEL-agnostic contemporary allele-specific DNAme script (Wang et al. 2014) using the same dataset (red).

MEA is shipped in a Docker container, enabling one step installation of all dependencies independent of operating system type. After providing a reference genome assembly (e.g. hg19 or mm10) and a VCF file containing the relevant genetic variants, users simply input an HTS dataset in FASTQ format. MEA will then automatically generate allele-specific genomic coverage files in BigWig format and allele-specific analyses over user-defined regions of interest in a tab-delimited table. To benchmark the performance of our INDEL-aware software, I present both theoretical and real-world evidence for improved allele-specific DNAme analysis relative to an INDEL-agnostic pipeline. Furthermore, to highlight the utility of MEA, I investigate DNAme data processed in parallel with RNA- and ChIP-seq data from mouse hybrid embryos and uncover novel DMRs. Additionally, using human brain cell data, I observe the expected H3K27ac and DNAme enrichment at known imprinted genes and uncover novel monoallelically expressed genes, further demonstrating the power of integrating epigenetic and expression analyses in a unified workflow. The MEA toolbox harmonizes HTS read processing, with all dependencies consolidated in a Docker container,

includes pan-species compatibility, maximizing its utility for allele-specific profiling of model organisms as well as human samples.

2.2 Materials and Methods

2.2.1 Samples used in this study.

I validated our tool using previously published bisulphite-seq data generated from inner cell mass (ICM) cells from an F1 hybrid between mouse strains C57BL/6J and DBA/2J (Wang et al. 2014). DBA/2J differs from the reference strain (C57BL/6J) by 5,126,997 SNVs (roughly 1 SNV/530 bp) and 1,019,400 INDELs, comparable to other commonly used lab mouse strains (Keane et al. 2011) (see Discussion). ICM bisulphite-sequencing (GSM1386023) was complemented with RNA-seq (GSM1845307-8) as well as ChIP-sequencing data for H3K4me3 (GSM1845274-5) and H3K27me3 (GSM2041078-9), permissive and repressive histone post-translation modifications respectively, from ICM cells isolated from C57BL/6J x PWK/PhJ F1 mice (20,626,644 SNVs and 3,044,259 INDELs) (see **Table 2.1**). RNA-seq data from C57BL/6J x DBA/2J ICM (GSM1625868) was used to test allele-specific alignment performance of contemporary RNA-seq aligner software. Bisulphite sequencing datasets from C57BL/6J MII oocytes (GSM1386019) and DBA/2J spermatozoa (GSM1386020) were analyzed to directly measure false-positive allele-specific alignment rates. Processed fully grown oocyte (DRX001583) and sperm (DRX001141-9) bisulphite-seq were used for visualization. Processed human sperm and oocyte WGBS was obtained from JGAS00000000006. Adult human brain datasets were obtained as part of the Canadian

Epigenetics, Environment and Health Research Consortium (CEEHRC) Network. A detailed description of all datasets can be found in **Table 2.1**. The results reported here are in part based upon data generated by The Canadian Epigenetics, Epigenomics, Environment and Health Research Consortium (CEEHRC) initiative funded by the Canadian Institutes of Health Research (CIHR), Genome BC, and Genome Quebec. Information about CEEHRC and the participating investigators and institutions can be found at http://www.cihr-irsc.gc.ca/e/43734.html.

Table 2.1 Publicly available datasets used in this chapter and their source.

Dataset name	Source experiment	SRA code	Sequencing type	Reference
Mouse datasets				
MII oocyte DNA methylation	GSM1386019	GSM1386020	101 PE	Wang et al. 2014
SpermDNA methylation	GSM1386020	GSM1386020	101 PE	Wang et al. 2014
ICM DNA methylation	GSM1386023	GSM1386020	101 PE	Wang et al. 2014
E7.5 DNA methylation	GSM1386025	GSM1386020	101 PE	Wang et al. 2014
FGO DNA methylation	DRA000570	DRX001583	100 SE	Shirane et al. 2013
Sperm DNA methylation	DRA000484	DRX001141-9	76 PE	Kobayashi et al. 2012
BD ICM RNAseq	GSE71434	GSM1845307-8	126 PE	Zhang et al. 2016
BP ICM RNAseq	GSE66390	GSM1625868	126 PE	Wu et al. 2016
ICM H3K4me3 ChIPseq	GSE71434	GSM1845274-5	101 PE	Zhang et al. 2016
ICM H3K27me3 ChIPseq	GSE76687	GSM2041078-9	101 PE	Zheng et al. 2016
Human datasets				
Oocyte DNA methylation	JGAS00000000006	JGAS00000000006	101 SE	Okae et al. 2014
Sperm DNA methylation	JGAS00000000006	JGAS00000000006	101 SE	Okae et al. 2014
Brain RNA-seq	CEEHRC	EGAD00001001402	75 PE	CEEHRC
Brain DNA methylation	CEEHRC	EGAD00001001312	125 PE	CEEHRC
Brain H3K27ac ChIPseq	CEEHRC	EGAD00001001403	75 PE	CEEHRC
Brain ChIPseq input control	CEEHRC	EGAD00001001409	75 PE	CEEHRC

2.2.2 Implementation.

To generate a harmonized workflow for processing of DNAme, RNA-seq and ChIP-seq datasets, we developed a universal strategy for detecting allele-specific reads. Further, to maximize the number of experimental reads that can be assigned to a specific allele for each data type, MEA was designed to exploit underlying genetic variation by incorporating both SNVs and INDELs during pseudogenome construction. For each data type, allelic reads are captured by constructing an *in silico* pseudogenome comprised of both parental genomes followed by HTS read alignment. Aligning reads simultaneously to both haplotype sequences of a diploid genome facilitates the appropriate alignment of reads that map to heterozygous loci onto their cognate allele, reads which otherwise would be discarded due to "sequencing errors". Such reads are thus extracted and can be used to de-convolute allelic phenomena.

2.2.3 *in silico* diploid genome reconstruction.

MEA constructs a diploid pseudogenome using a reference sequence (.fasta) and known genetic variants (.vcf) including SNVs and INDELs, as shown previously (Younesy et al. 2014). For samples requiring genotype phasing, MEA utilizes SHAPEIT2 (Delaneau et al. 2008) and a publicly available reference panel of haplotypes provided by the 1000 Genomes Project (McVean et al. 2012) to output phased haplotypes. These steps generate an *in silico* diploid genome containing two copies of each chromosome, one for each parental genome. Aligning HTS reads to a

diploid genome enables the extraction of uniquely aligned allele-specific reads, which are separated into parent-of-origin read alignment files. An automatically-generated index file (.refmap) enables reversal of coordinate alterations in non-reference allelic read alignments caused by differential parental INDEL lengths. This allows projection of allelic genomic tracks back onto the original reference genome for consistent data visualization in genome browsers (which are built around reference genomes) and downstream analyses over coordinate-based regions of interest.

2.2.4 MEA exploits widely used HTS alignment software.

In order to detect allele-specific reads from RNA-, ChIP-seq and WGBS data, we designed MEA to align reads using an *in silico* pseudogenome and extract uniquely mapped reads. This approach allows allele-specific alignment of reads containing sequencing errors, which is critical for datasets with long (100+ bp) reads commonly sequenced on Illumina sequencers, which have approximately 0.26-0.80% sequence error rates (Quail et al. 2012). This pipeline modification assures adoption and operation of our tool well into the future as sequencing technologies continue to extend read lengths without necessarily improving error rates.

2.2.5 Special considerations for allele specific DNAme analysis.

DNAme levels can be accurately measured genome-wide using sodium bisulphite conversion of unmethylated cytosines to thymines followed by whole genome sequencing (bisulphite-seq). To measure allele-specific DNAme levels, MEA detects

allelic reads and calculates the proportion of cytosines and thymines at CpG dinucleotides. To do so, MEA aligns bisulphite-seq reads to the *in silico* diploid genome using the popular aligner and methylation caller Bismark (Krueger and Andrews 2011). Unlike ChIP- or RNA-seq aligners, Bismark considers cytosine to thymine mutations (introduced during sodium bisulphite conversion) in order to accurately align reads to a genomic sequence. Allele-specific DNAme levels therefore reflect both genetic and epigenetic effects: users can retroactively delineate both effects using their original list of genetic variants.

2.2.6 UCSC track hubs for allelic track visualization.

UCSC Track Hubs are a hierarchical file organization system that allow combining multiple genomic tracks into one for convenient data visualization and interpretation (Kent et al. 2002; Raney et al. 2014). MEA automatically normalizes allele-specific tracks by sequencing depth and generates corresponding track hub database files. Using UCSC binaries (hubCheck), we ensure the integrity of MEA-generated track hubs for standardized visualization experiences. Additionally, we provide scripts for the automatic computation of allelic RNA- and ChIP-seq coverage over user-defined regions of interest (for example: transcription start sites, genes, enhancers, etc.), outputting a tab-delimited table. While RPKM- and coverage-calculating software already exist, confounding variables are inherent to allelic analyses, requiring custom scripting. For example, calculating allelic RPKM values using conventional tools is complicated by the variability in SNV and INDEL density between regions of interest. To

account for such effects, MEA's default output includes allelic read coverage for both alleles (to calculate allelic imbalance) and total RPKM (to filter for enrichment). Users can easily interpret allelic imbalance calculations with the combination of these two metrics (allelic read coverage and total RPKM) over their regions of interest. In this study, VisRseq (Younesy et al. 2015) was used to plot allelic read coverage for RNA-seq data from human brain.

2.2.7 Consolidation of tool dependencies into self-sufficient pipeline.

Packaging MEA into a Docker Container allows the one-step installation of all 15 dependencies, significantly reducing the work required by the end-users. Simply, the Docker container is a text file containing instructions for installing a virtual system and setting environment variables, followed by standardized installation of each bioinformatic dependency. Once installed through the Docker container, MEA is immediately operational.

2.2.8 Software tool requirements.

Users are encouraged to install MEA through Docker. Alternatively, manual installation requires the following software (with specific versions used during development of MEA): Java v-1.6, Python v-2.4, Bismark v-0.15.0, Bowtie2 v-2.2.3, Bwa v-0.7.10, STAR v-2.5.1b, Tophat2 v-1.1, Samtools v-0.1.16, Bedtools v-2.22.1, VCFtools v-0.1.10, SHAPEIT2, bgzip v-1.1, bedGraphToBigWig v-1.1, wigToBigWig v-4 & hubCheck.

2.3 Results

2.3.1 An allele-specific DNA methylation pipeline.

To establish a pipeline for allele-specific DNAme analysis, I began by incorporating Bismark (Krueger and Andrews 2011), a widely adopted bisulphite-seq read aligner and methylation caller, into ALEA, our previously developed tool for allele-specific analyses of RNA-seq and ChIP-seq datasets (Younesy et al. 2014). I first quantified the hypothetical increase in the percentage of informative CpG sites from which I can infer allelic information by incorporation of INDELs in addition to SNVs during pseudogenome reconstruction. As high-quality genetic variation information of inbred mouse strains is available (Keane et al. 2011), I constructed a pseudogenome from two mouse strains, namely DBA/2J and the reference strain C57BL/6J (build mm10), incorporating known genetic variants (SNVs and/or INDELs). By counting CpGs within 200 bp (an insert size typical of WGBS libraries) of an INDEL or SNV, I found that INDEL incorporation leads to a theoretical increase in the number of informative CpGs (i.e. CpGs for which DNAme differences between alleles can be deduced) of 12.9% for this pseudogenome (**Figure 2.1b**). Notably, a subset of genomic regions with associated INDELs are entirely devoid of SNVs and therefore include nearby CpGs that theoretically can only be assessed by pipelines that are “INDEL-aware”.

2.3.2 MEA is informative for significantly more CpGs than an INDEL-agnostic script.

To test whether the inclusion of INDELs increases the number of informative CpGs for which allelic methylation state can be calculated in practice, I processed raw reads from a previously published WGBS dataset from C57BL/6J x DBA/2J mouse F1 inner cell mass (ICM) cells (Wang et al. 2014). Applying the same filtering parameters allowed us to directly compare results obtained with the MEA pipeline to those of the Bismark-based INDEL-agnostic custom script employed by Wang *et al.* (Wang et al. 2014). MEA yielded a 62.5% increase in the number of CpGs covered by at least 5 allele-specific C57BL/6J reads (**Figure 2.1c**). Importantly, informative CpGs gained using MEA overlapped almost exclusively with CpGs within 200 bp of an INDEL or SNV, as expected. This gain is likely the result of an increase in the number of informative heterozygous sites (quantified in **Figure 2.1b**) as well as efficacious alignment of reads to the non-reference genome over regions with high INDEL density.

Reads from regions with high INDEL density were presumably excluded by the pipeline from Wang *et al*. as “sequencing errors”, rather than assigned as allelic variants. To confirm that MEA increases the alignment rate of non-reference reads, I repeated the alignment of C57BL/6J x DBA/2J F1 WGBS reads to a reference genome or the MEA-constructed diploid pseudogenome (composed of the reference and DBA/2J genomes) and determined the number of reads that aligned to each genome 0, 1 or >1 time (**Figure 2.2a-b**). Alignment to a pseudogenome increased the overall alignment

rate by 1.25% (80.83 to 82.08%), most likely due to alignment of non-reference-originating reads at loci that show significant genetic divergence (high SNV and INDEL density) from the reference. As expected, the majority of reads aligned uniquely to the haploid reference genome aligned at least twice to the pseudogenome, except over regions containing genetic variants. This crucial distinction allowed the uniquely aligned reads to be extracted and assigned to their cognate parental genomes, with 8.8% and 8.2% of all aligned reads specific to C57BL6J and DBA/2J strains, respectively (**Figure 2.2c**). By capturing a greater number of sites at which I can measure allelic DNAme levels, a higher proportion of experimental reads can be assigned to a specific parental haplotype, thus enabling the evaluation of allelic differences in DNAme levels for a higher fraction of the genome.

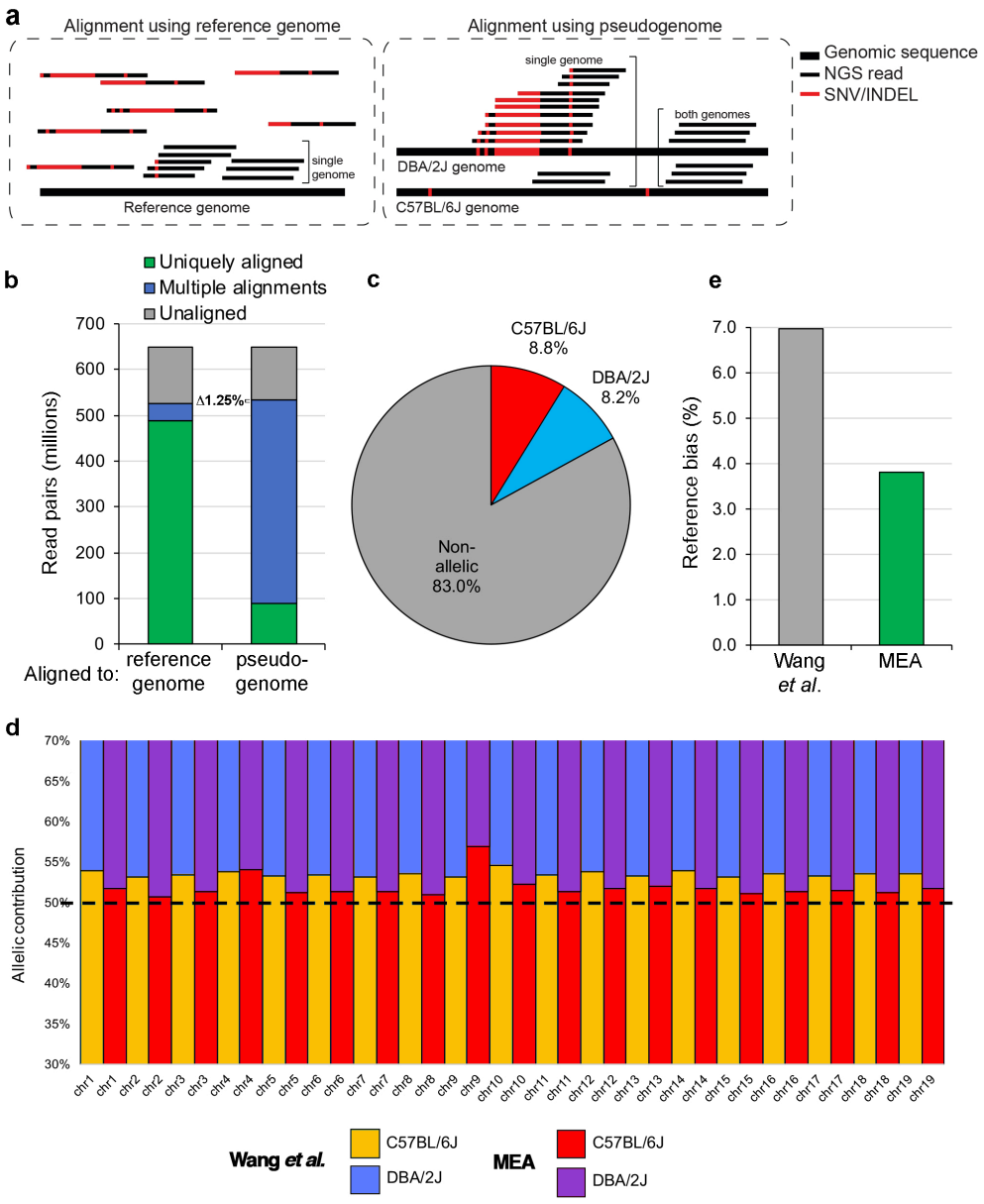

Figure 2.2 Empirical benchmarking of allele-specific read alignment reveals reduced reference bias.

(a) Graphical representation of MEA's unified strategy for detecting allele-specific reads from RNA-, ChIP-seq and WGBS datasets. Aligning F1 hybrid reads to a pseudogenome enables alignment to their cognate genome of reads originating from highly variable loci. **(b)** Paired-end WGBS reads (101 bp) from a previously published dataset of C57BL/6J x DBA/2J ICM cells (Wang et al. 2014) were aligned using the

Bismark aligner to the (haploid) reference genome (mm10 build) and a MEA-constructed diploid pseudogenome. When using MEA, multiple (2 or more) alignments reflect non-allelic reads, while uniquely aligned reads are allele-specific. Reads aligning uniquely to the pseudogenome were extracted and retroactively assigned to their parental haplotype. **(c)** The percentages of allele-specific reads called for each parental haplotype and the number of aligned reads that did not overlap with a genetic variant (non-allelic) is shown. **(d)** Allelic contribution of read alignments to each parental haplotype (C57BL/6J or DBA/2J) on each autosome. Relative to the script employed by Wang *et al.*, MEA displays about half the reference bias on the majority of autosomes. **(e)** Global reference bias for each pipeline is shown.

2.3.3 MEA significantly reduces reference genome alignment bias.

A major concern when exploring allele-specific data is the potential for reference bias caused by differences in genomic sequence quality between the reference and non-reference genomes, caused by preferential alignment of reads to the former and artefactual allelic imbalance results (Degner et al. 2009). For example, using an INDEL-agnostic pipeline similar to that employed by Wang *et al*. (Wang et al. 2014), Keown *et al,* reported a reference bias of 15.4% in their study of allele-specific DNAme in C57BL/6J x SPRET/EiJ cells (Keown et al. 2017) (SPRET/EiJ has > 5 times the number of SNVs relative to C57BL/6J than does DBA/2J (Keane et al. 2011)). To determine the extent of reference bias in our MEA pipeline, I benchmarked the observed parental contribution to allelic read alignment for each autosome from the C57BL/6J x DBA/2J ICM WGBS dataset generated by Wang *et al.* **(Figure 2.2d)**. Notably, MEA yielded an

alignment reference bias on all autosomes of 3.81%, only ~54% of that reported by the INDEL-agnostic pipeline (6.98%, **Figure 2.2e**). This reduction in alignment bias is consistent with the increased fraction of allele-specific reads aligned to the non-reference genome.

2.3.4 Estimation of allele-specific alignment error rate using isogenic mice.

False positives caused by erroneous allelic read alignment at regions devoid of true genetic variation can lead to an underestimation of reference bias in allele-specific experiments. To quantify the false positive allelic alignment rate of our pipeline, I processed pure C57BL/6J WGBS data using the C57BL/6J x DBA/2J pseudogenome described above and determined the parental contribution to allelic read alignment (**Figure 2.3a**). Curiously, 0.8% of all aligned reads (5.13% of allelic reads) were scored as DBA/2J-specific, indicating that MEA has an FDR of ~5%. When calculating the parental contribution to allelic read alignment over each autosome, I found that the majority of false-positive ("DBA/2J-specific") allelic read alignments clustered on chromosomes 2 and 9 (**Figure 2.3b**). Closer inspection revealed that these regions are annotated by RepeatMasker as Satellite DNA **(Figure 2.3c)**. Such allele-specific calls at sites lacking genetic variants are the result of Bismark's mapping quality algorithm, which calculates an erroneously high mapping score at these highly repetitive regions. Analysis of processed WGBS data from pure DBA/2J spermatozoa without black-listing of repetitive regions revealed a C57BL/6J-specific alignment rate of 3.80% **(Figure 2.3 d-e)**, indicating that a global false positive rate of ~5% may be expected when using the

MEA pipeline for analysis of WGBS data without excluding repetitive regions. Interestingly, false-positive C57BL/6J read alignments spanned large chromosomal segments (**Figure 2.3f**), hinting that there may have been interbreeding between experimental mouse strains. Since satellite DNA is generally omitted in studies of the transcriptome or epigenome, I excluded reads aligned to annotated satellite repeats (0.19% of the mappable genome) and recalculated the false-positive rate for the C57BL/6J dataset, which dropped to 1.62% of allelic reads, with no specific chromosome enriched **(Figure 2.3g)**. Thus, when applying the MEA pipeline, the majority of false positive read alignments can likely be removed by black-listing satellite repeats.

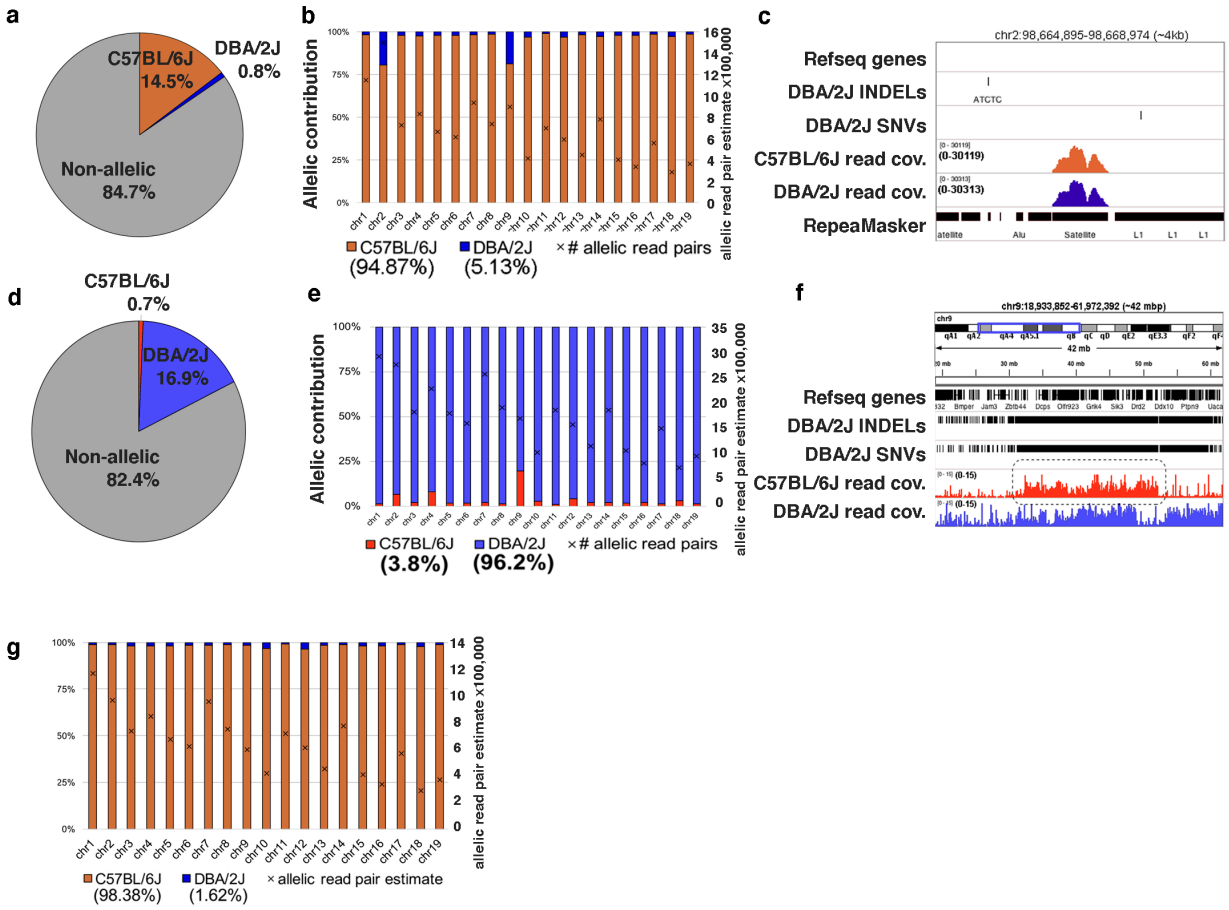

Figure 2.3 Quantifying allele-specific alignment error rates.

(a-c) To estimate the rate of false-positive errors for allelic analysis of DNAme data, WGBS reads generated from C57BL/6J mice (Wang et al. 2014) were aligned to the MEA-generated C57BL/6J x DBA/2J pseudogenome, and the percentage of DBA/2J-specific read alignments was scored. The expected allelic contribution from C57BL/6J is 100%, as these cells are of C57BL/6J origin. **(a)** The percentages of reads aligning uniquely to the C57BL/6J and DBA/2J (false-positive) pseudogenomes, as well as the number of aligned reads that did not overlap with a genetic variant (non-allelic) is shown. **(b)** The false-positive alignment rate for each autosome, along with the total number of aligned allelic read pairs, is shown. **(c)** Genome browser screenshot of a locus that displays a high rate of false-positive allele-specific alignment to a repeat annotated as Satellite DNA by RepeatMasker and devoid of genetic variants. **(d-f)** To

estimate the rate of false-positive errors for WGBS analyses, raw data generated from DBA/2J mice (Wang et al. 2014) was aligned to the MEA-generated C57BL/6J x DBA/2J pseudogenome and the percentage of C57BL/6J-specific read alignments was scored. The expected allelic contribution from C57BL/6J is 0%, as these cells are of DBA/2J origin. **(d)** The percentage of reads aligning to C57BL/6J (false-positive) and DBA/2J as well as the number of aligned reads that did not overlap with a genetic variant (non-allelic) is shown. **(e)** The false-positive alignment rate for each autosome, along with the number of aligned allelic read pairs, is shown. **(f)** Genome browser screenshot of a representative false-positive locus. C57BL/6J-specific reads aligned in large stretches of false-positive alignment regions, suggesting that the parental strain DBA/2J from this study was not isogenic. Indeed, when manually inspecting these stretches of false-positive read alignments, experimental reads perfectly matched the reference sequence over known DBA/2J SNVs and INDELs, again suggesting that "DBA/2J" mice analyzed by Wang *et al*. (Wang et al. 2014) contained C57BL/6J sequence. **(g)** To assess the false-positive rate exclusive of repetitive Satellite DNA, allele-specific read alignments over these Repbase annotated repetitive sequences, as recognized by RepeatMasker, were culled and the rate of false-positive allele-specific alignments recalculated over each autosome as in **(b)**.

2.3.5 MEA reports the expected allelic imbalance in DNA methylation at known gametic DMRs (gDMRs).

To establish the accuracy of calculating allele-specific DNAme levels using the MEA pipeline, I measured allele-specific DNAme levels over known imprinted gDMRs. Such regions are densely methylated on one allele and unmethylated on the other as a result of parent-of-origin dependent differences in methylation established in the gametes, representing a unique resource for benchmarking allele-specific DNAme calling. Of the

24 known mouse gDMRs, 9 harbor SNVs and/or INDELs between the C57BL/6J and DBA/2J genomes and can therefore be assessed for allele-specific DNAme levels. For consistency, I directly compared our allele-specific results over these regions with those reported by Wang *et al*. **(Figure 2.4a)**. For most gDMRs, MEA yielded average allelic DNAme levels similar to those reported by the INDEL-agnostic pipeline. However, MEA consistently yielded allele-specific information over a greater number of CpGs (mean±SD: 72±24 vs 38±21 CpGs on either allele), increasing the statistical power of allelic imbalance calculations. For example, MEA detected a total of 68 CpGs informative for allelic methylation state at the *Dlk1-Meg3* IG-gDMR, nearly three times greater than the number reported by Wang *et al*. **(Table 2.2)**.

Table 2.2 Allele-specific DNA methylation level analysis over the *Dlk1-Meg3* IG-gDMR in ICM cells.

Pipeline	Allelic call	CpGs covered	Mean Methylation (%)
MEA	-	129	30.24
	C57BL/6J	31	1.66
	DBA/2J	37	58.55
	Total allelic informative	68	30.11
Wang *et al*.	-	129	30.63
	C57BL/6J	12	1.59
	DBA/2J	12	48.09
	Total allelic informative	24	24.84

As expected, when calculated over the same 129 CpGs covered by at least five reads in the gDMR, DNAme levels calculated by the two pipelines independent of allelic calling were nearly identical (30.2% vs 30.6%). However, the discordance between the percentage of methylation calculated for the CpGs that are informative at an allelic level was significantly lower using the MEA pipeline (0.13% vs 5.8%), indicating that the accurate determination of allelic DNAme levels at specific loci can be adversely impacted by sampling errors. Furthermore, as expected, only the MEA pipeline yields informative results for CpGs proximal to INDELs at the *Dlk1-Meg3* IG-gDMR locus **(Figure 2.4b)**, confirming the benefit of incorporating the latter during pseudogenome reconstruction. Taken together, these analyses demonstrate that MEA outperforms an INDEL-agnostic pipeline.

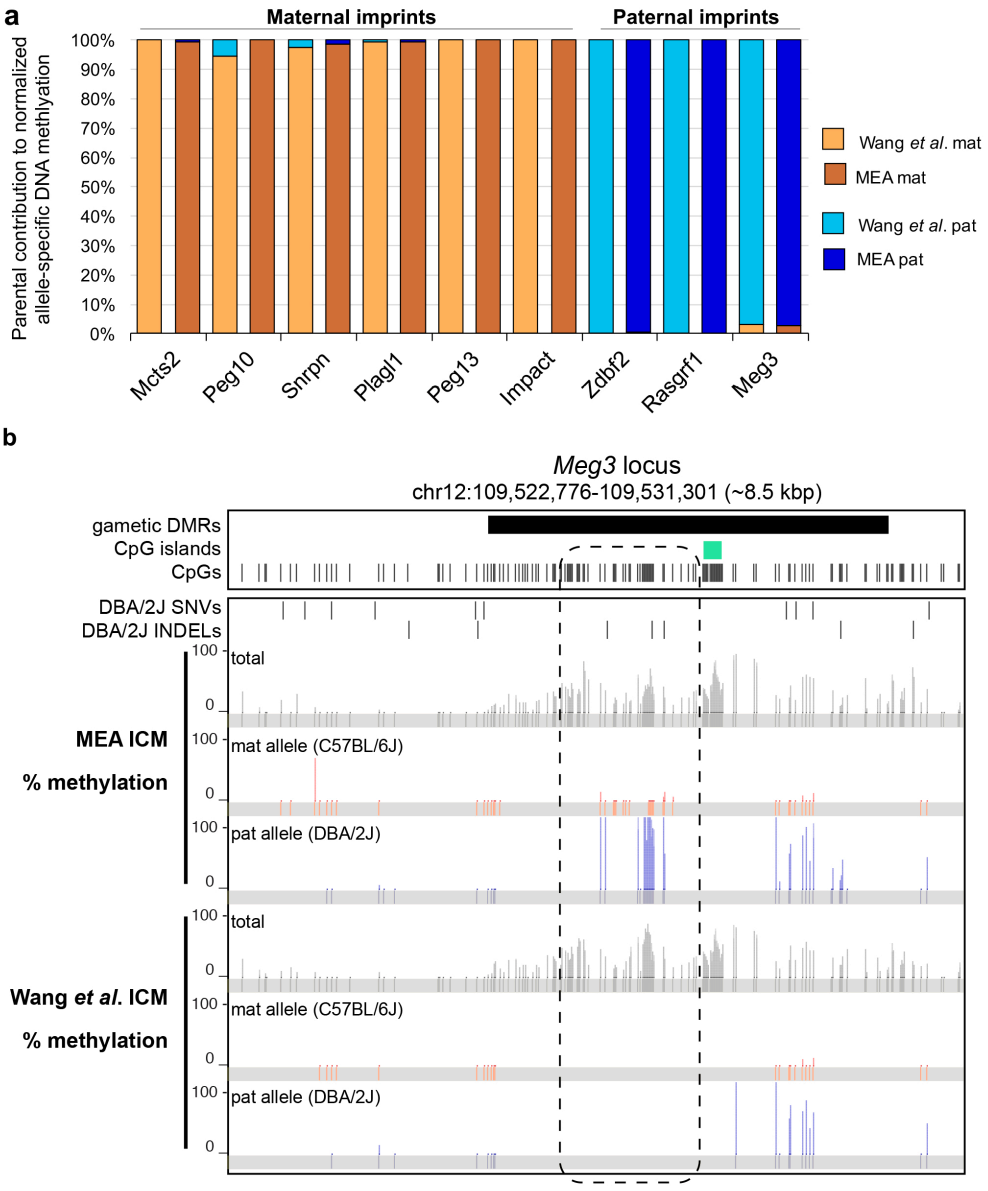

Figure 2.4 Validation of allele-specific DNA methylation level calculations over known gDMRs.

C57BL/6J x DBA/2J ICM WGBS reads were processed in parallel with MEA and a published pipeline (Wang et al. 2014) using identical parameters. **(a)** Allelic methylation levels over 9 known gDMRs are shown for both pipelines. **(b)** UCSC genome browser screenshot of the *Dlk1-Meg3* IG-gDMR including the allele-agnostic percentage of DNAme calculated using each pipeline (total) as well as allelic calls for each informative

CpG. The location of each informative CpG for each pipeline (highlighted in grey) is also included. Only MEA detects allele-specific reads in a region within the gDMR that lacks SNVs but contains several INDELs (dashed box). A summary of the total number of allelic CpG counts and DNAme levels over this locus is included in **Table 2.2**.

2.3.6 MEA uncovers novel putative transient DMRs at annotated transcription start sites (TSSs).

A recent study employing MeDIP on genomic DNA isolated from early mouse embryos revealed the presence of maternally-methylated DMRs that are resolved during post-implantation development (Proudhon et al. 2012). While these "transient DMRs" may have important biological functions during pre-implantation development (Proudhon et al. 2012; Greenberg et al. 2016), the extent of transient imprinting remains unclear. To determine whether MEA can be used to identify novel DMRs, I assayed the subset of informative regions gained using our refined pipeline, namely loci exclusively overlapping INDELs, using the aforementioned WGBS data from C57BL/6J x DBA/2J ICM cells. As expected for preimplantation cells, which are characterized by globally low DNAme levels (Smith et al. 2012), hypomethylation of both parental alleles was generally observed over such informative regions, including at those with high CpG density (**Figure 2.5a**). Importantly, analysis agnostic to allelic alignment also revealed hypomethylation across such regions (for example, see **Figure 2.6**). However, focusing on regions within 200 bp of annotated transcription start sites (TSSs) reveals that a subset shows clear asymmetric DNAme levels (**Figure 2.5b**), with either maternal or paternal bias.

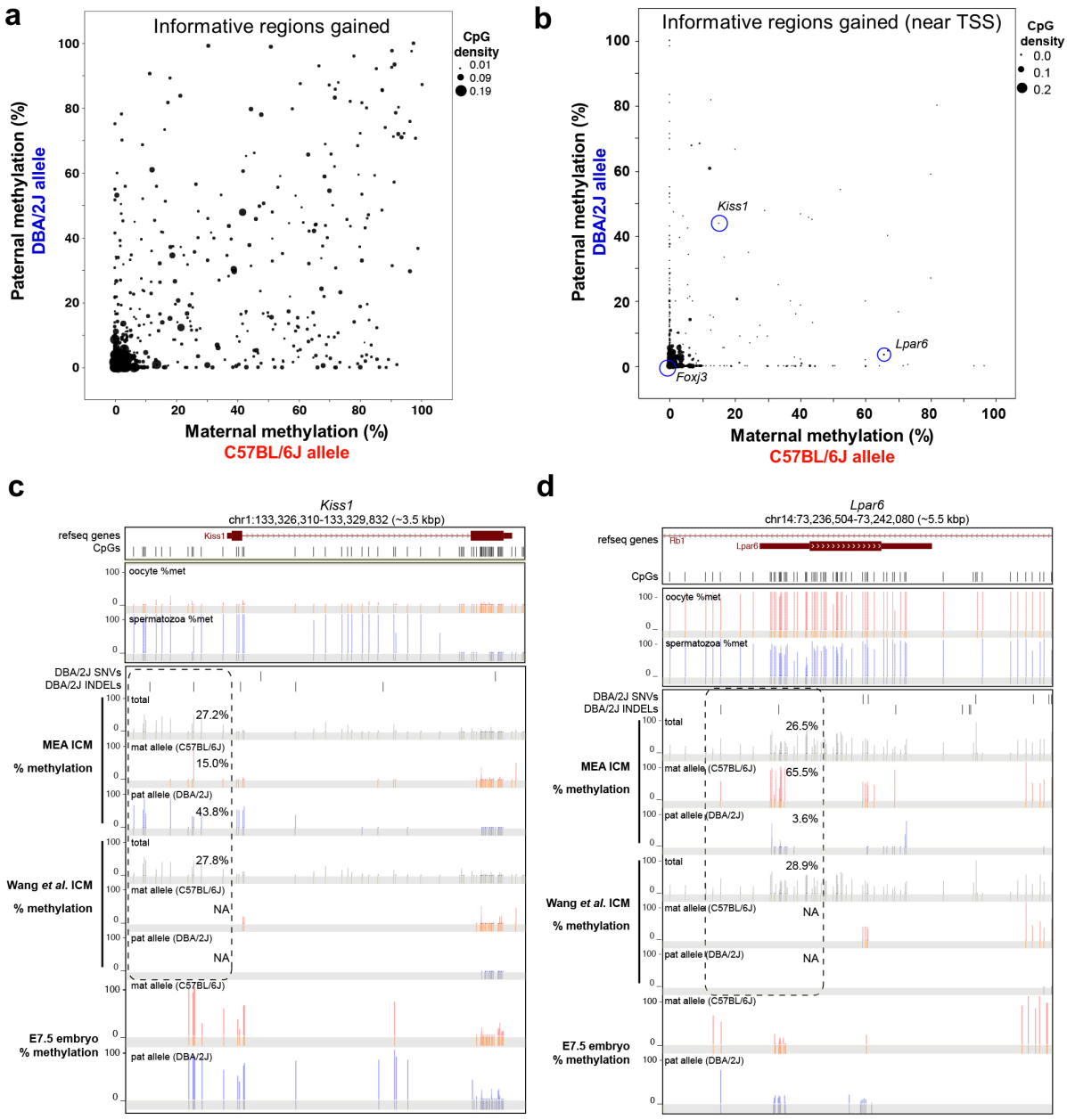

Figure 2.5 Identification of novel DMRs using the MEA pipeline.

Allele-specific DNAme levels were calculated over 133,065 regions containing INDELs but lacking SNVs (representing novel informative regions gained using MEA) using C57BL/6J x DBA/2J ICM WGBS data (Wang et al. 2014). **(a)** Maternal versus paternal DNAme levels and CpG density (data point size) are plotted for informative regions overlapping with at least 10 CpGs from which allele-specific DNAme levels can be

ascertained (746 data points). **(b)** CpG density (data point size) and allele-specific DNAme levels are shown, as in (a) over the subset of novel informative regions +/- 200bp from annotated TSSs (with at least five informative CpGs on both alleles). Representative novel informative regions for which screenshots are provided are circled in blue. **(c-d)** UCSC genome browser screenshots of differentially methylated regions (dashed boxes) near the promoters of the *Kiss1* and *Lpar6* genes. Tracks from Wang *et al.* are included to illustrate differences in pipeline sensitivity. DNAme tracks of sperm and oocytes (Kobayashi et al. 2012; Shirane et al. 2013) as well as E7.5 embryos (Wang et al. 2014) are also shown, along with the location of informative CpGs (highlighted in grey).

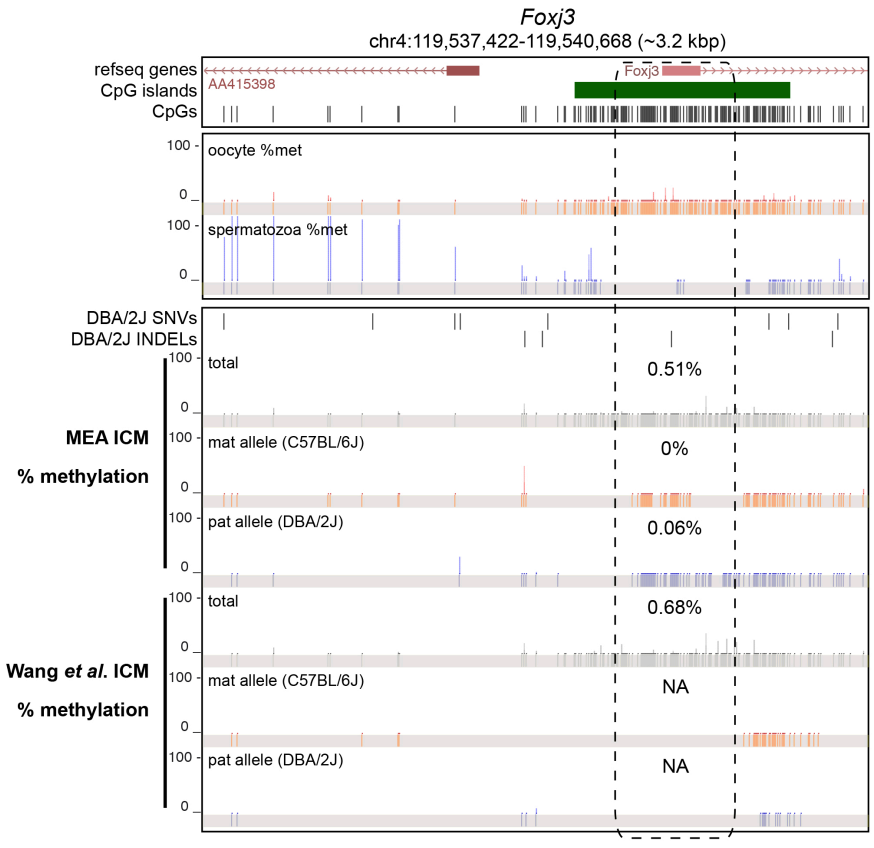

Figure 2.6 DNA methylation dynamics over the *Foxj3* CpG island promoter.

UCSC genome browser screenshot of a representative region (*Foxj3* CGI promoter) over which an allele-agnostic pipeline calculated a total DNAme level of <1% (dashed box). Accordingly, the levels of allele-specific DNAme on both parental alleles, as calculated by MEA, are <1%. DNAme tracks of male and female germ cells (Kobayashi et al. 2012; Shirane et al. 2013) are also shown, as well as a track indicating the location of each informative CpG (highlighted in grey).

UCSC genome browser screen shots of two putative TSS proximal DMRs, including the apparently paternally methylated *Kiss1* (a suppressor of metastasis) and maternally methylated *Lpar6* (a lysophosphatidic acid receptor) genes, are shown in **Figures 2.5c** and **2.5d**. Using the MEA pipeline, 15 and 34 CpGs respectively, are informative on either allele at these loci. Importantly, the absolute methylation levels reported by the allele-agnostic pipeline (27.2 and 26.5%) are similar to those of the mean allele-specific methylation (29.4 and 34.6%), consistent with the observation that methylation at these loci is allele-specific. Moreover, intersection of these ICM data with WGBS data from mature gametes (Shirane et al. 2013; Kobayashi et al. 2012) reveals that paternal DNAme at the *Kiss1* gene in the former is likely the result of methylation already present in spermatozoa (**Figures 2.5c-d**), indicating that this locus potentially protected from the wave of genome-wide DNA demethylation that occurs early in mouse embryonic development (Leitch et al. 2013). Parental asymmetry at the *Kiss1* locus is resolved by E7.5, when the maternal allele gains DNAme coincident with the wave of global *de novo* DNAme that occurs during early post implantation development (Borgel et al. 2010). On the other hand, the short, intron-less gene *Lpar6* is hypermethylated in

both mature oocytes and spermatozoa, indicating that the paternal but not the maternal allele is susceptible to the global wave of DNAme erasure that takes place after fertilization. Parental asymmetry of DNAme is resolved by loss of maternal DNAme in the E7.5 post-implantation embryo, revealing that the allelic bias in DNAme at this locus is also transient but involves sequential loss of DNAme on the paternal followed by the maternal allele. Whether these non-canonical DNAme dynamics are driven by genetic or parent-of-origin effects, and their contribution to the development of the early embryo, remains to be tested. Regardless, the novel DMRs identified proximal to the *Kiss1* and *Lpar6* TSSs exemplify the merit of increasing the number of allelic reads extracted from experimental datasets and underscores the potential for future discoveries using this approach.

2.3.7 Comparison of RNA- and ChIP-seq read aligners using the MEA pipeline.

In order to integrate epigenomic and transcriptomic-based datasets, alignment to the same genomic sequence is required. Transcriptomic data presents a unique challenge when aligning to a genome, as processed messenger RNA contains many gaps (introns) relative to the template DNA sequence. In our previously published pipeline ALEA (Younesy et al. 2014), RNA-seq alignment was carried out using the short-read aligner BWA, which does not allow alignment of intron-spanning reads. Thus, to enable integration of transcriptomic and epigenomic datasets, gapped read alignment is essential. Tophat2 (Kim et al. 2013) and STAR (Dobin et al. 2013), two widely used aligners that incorporate this feature, were recently shown to perform well in short-read

RNA-seq alignment (Križanovic et al. 2018). To determine which of the two shows superior allele-specific gapped read alignment, I carried out a side by side comparison of these aligners, as well as the non-gapped read aligner BWA, using a published RNA-seq dataset from C57BL/6J x DBA/2J F1 ICM cells. STAR clearly outperformed both Tophat2 and BWA (**Figure 2.7a**), likely due to its advanced gapped read alignment algorithm (Dobin et al. 2013) and ability to properly assign paired-end reads associated with the same DNA molecule (if a read aligns to a region including a genetic variant, its mate is also identified as allelic regardless of whether it overlaps a genetic variant). Thus, analysis of paired-end sequencing data using the STAR aligner and MEA pipeline increases the fraction of regions showing relatively high sequence conservation over which allele-specific HTS reads can be aligned, an improvement over using flanking regions as a proxy. Based on these observations, we currently recommend the STAR aligner, but MEA's flexibility in incorporating new HTS aligners facilitates its adoption for analyzing epigenomic and expression datasets using alternative/next generation aligners, such as those that can accommodate increased read lengths.

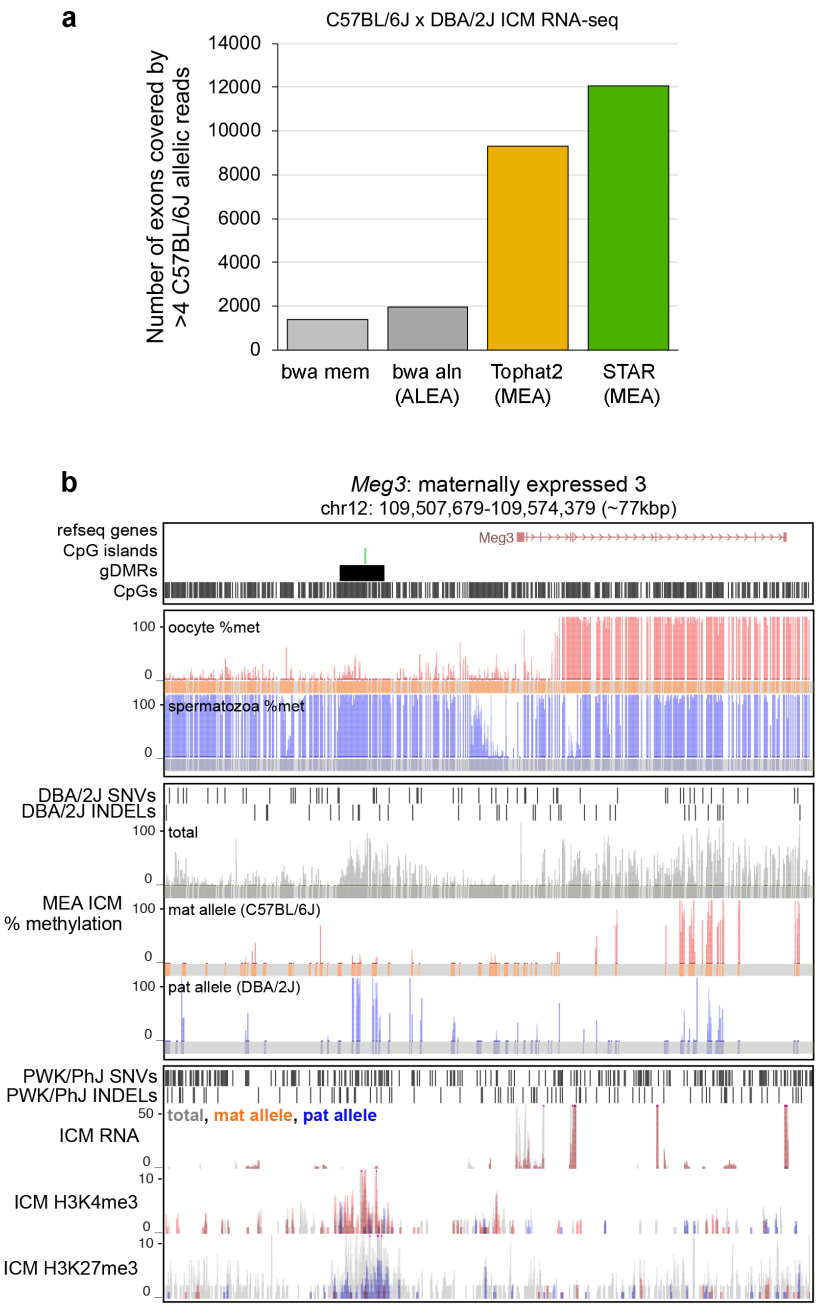

Figure 2.7 Validation of allele-specific transcription level calculations and integration with ChIP-seq and WGBS datasets at allelic resolution.

MEA was extended to accommodate contemporary RNA-seq aligners and to automatically organize allelic and total genomic tracks into UCSC Track Hubs to aid data visualization and interpretation**. (a)** The number of annotated genic exons covered

by allelic reads using BWA, Tophat2 and STAR aligners is shown for an RNA-seq dataset generated from C57BL/6J x DBA/2J ICM cells (Wu et al. 2016). **(b)** UCSC genome browser screenshot of the *Dlk1-Meg3* IG-gDMR and downstream gene using the default MEA output for visualization of allelic (WGBS, RNA- and ChIP-seq) data. MEA automatically generates composite tracks containing total (allele-agnostic, grey), reference/maternal (red) and non-reference/paternal (blue) genomic tracks for visualization of allelic RNA- and ChIP-seq datasets. Bottom three tracks show MEA output from previously published C57BL/6J x PWK/PhJ F1 ICM ChIP-seq data (Zhang et al. 2016; Zheng et al. 2016).

In our previously published pipeline ALEA (Younesy et al. 2014), allele-specific alignment of ChIP-seq datasets was limited to the BWA-aln algorithm, which was designed to process short (<100bp) HTS reads (Li and Durbin 2009). To enhance MEA's flexibility, I incorporated another popular ChIP-seq aligner Bowtie2. To compare the performance of BWA-aln and Bowtie2 for allele-specific ChIP-seq alignment, I processed H3K4me3 ChIP-seq data generated form pure C57BL/6J and PWK/PhJ gametes (Zhang et al. 2016). While both alignment algorithms yield a low false-positive alignment rate of ~0.2-4.8%, BWA-aln clearly reports more allele-specific read alignments than Bowtie2 (**Figure 2.8a-d**). Thus, while users can choose between BWA-aln and Bowtie2, we recommend the former for allele-specific analysis of ChIP-seq data using MEA.

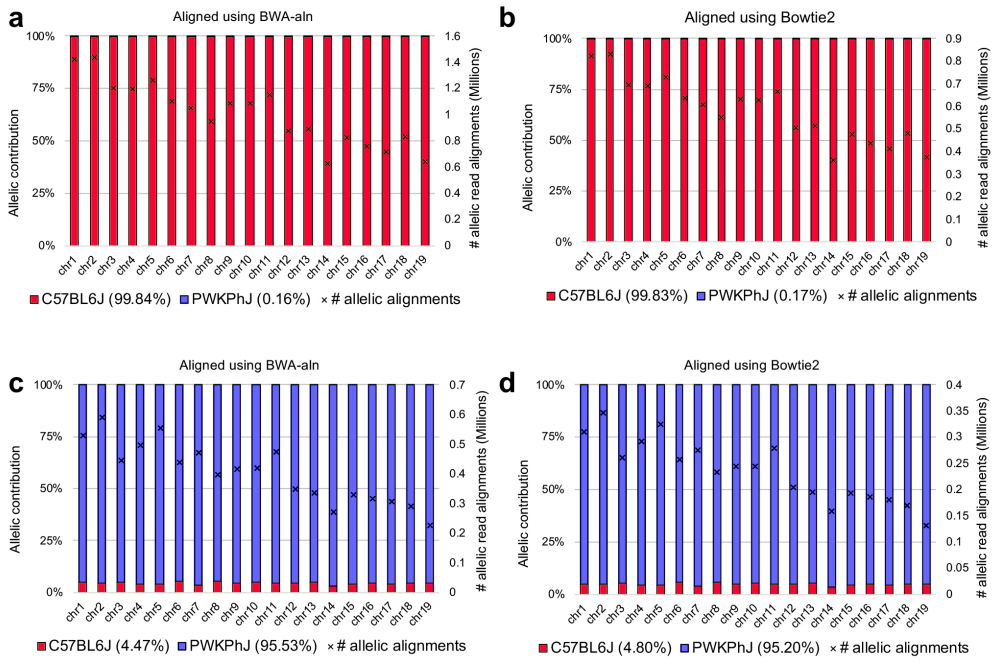

Figure 2.8 Comparison of ChIP-seq software for allele-specific read alignment.
To estimate the rate of allele-specific read alignments and false-positive errors for ChIP-seq analyses, raw H3K4me3 ChIP-seq data generated from C57BL/6J (fully grown oocytes) and PWK/PhJ (spermatozoa) mice (Zhang et al. 2016) was aligned to the MEA-generated C57BL/6J x PWK/PhJ pseudogenome and the number of C57BL/6J- and PWK/PhJ-specific read alignments was scored. The number of reads aligning to C57BL/6J and PWK/PhJ as well as the total number of allele-specific alignments on each autosome is shown for each analysis. **(a-b)** The expected allelic contribution for C57BL/6J is 100%, as these cells are of C57BL/6J origin. **(a)** Allele-specific alignment using the BWA-aln algorithm. **(b)** Allele-specific alignment using Bowtie2. **(c-d)** The expected allelic contribution for C57BL/6J is 0%, as these cells are of PWK/PhJ origin. **(c)** Allele-specific alignment using the BWA-aln algorithm. **(d)** Allele-specific alignment using Bowtie2.

2.3.8 Integration of WGBS, RNA-seq and ChIP-seq datasets using the MEA pipeline.

Dissecting the interplay between epigenetic marks and transcription was greatly facilitated by the advent of HTS-based approaches for measuring RNA levels and the genome-wide distribution of DNAme and histone PTMs. However, as such datasets are commonly processed using different pipelines, integrating and visualizing allelic information embedded therein is non-trivial. To automate dataset integration, MEA processes WGBS, RNA- and ChIP-seq alignment data using the same allele-specific read identification strategy, yielding standardized allele-specific genomic tracks. This unification of file types allows simultaneous visualization of each datatype (in BigWig format) using popular genome browsers. Further, to automate the process of reporting allelic imbalance, MEA generates a tab-delimited table containing allelic imbalance measurements over user-defined regions of interest, such as transcription start sites, genic exons or gene bodies.

This approach solves two important considerations in the presentation of allele-specific data. First, allelic genomic tracks, i.e. those displaying only read coverage that is informative for allelic alignment, are inherently sparse, especially at regions devoid of genetic variants. To delineate signal from noise, allele-specific genomic track visualization should be considered in the context of all aligned reads and the position of the genetic variant sites. Second, allele-specific enrichment is greatest at sites of genetic variation and therefore does not necessarily coincide with the profiles generated

from all reads agnostic of allelic assignment. For example, while reads derived from H3K4me3 ChIP-seq datasets are enriched over active TSSs, allelic H3K4me3 reads may align anywhere within the set of allele-agnostic peaks. Thus, allelic reads aligning at the edge of a region of H3K4me3 enrichment that is devoid of genetic variants at its center may be incorrectly discarded as noise.

The MEA pipeline standardizes such integrated track visualization by organizing genomic tracks into a UCSC Track Hub (Raney et al. 2014). These hubs agglomerate multiple colour-coded data tracks, enabling the concurrent visualization of allele-specific and "total" (allele-agnostic) alignment profiles, and in turn interpretation of allelic imbalance. Variant files used for pseudogenome reconstruction can also be directly visualized as UCSC custom tracks. The utility of this approach is illustrated using the *Meg3* gene and its governing intergenic gDMR as a representative locus (**Figure 2.7b**). Imprinting is simultaneously displayed in four independent datasets generated from two distinct F1 hybrid crosses. The *Dlk1-Meg3* IG-gDMR is paternally methylated and weakly enriched for both permissive (H3K4me3) and repressive (H3K27me3) histone PTMs (grey). Interestingly, H3K4me3 and H3K27me3 asymmetrically mark the maternal and paternal alleles, respectively, as expected for the promoter of a gene expressed exclusively from the maternal allele. Notably, each dataset is consistent with paternal imprinting, with repressive marks associated with the paternal allele and active marks with the expressed maternal allele. Profiles of the maternally methylated imprinted (and paternally expressed) *Snrpn* and *Impact* loci reveal similar patterns **(Figure 2.9**). Note

that for the *Impact* locus, a single genetic variant in the F1 hybrid analyzed is sufficient to score DNAme asymmetry between parental alleles. The observed enrichment of both H3K4me3 and H3K27me3 at imprinted DMRs is consistent with a previous report (McEwen and Ferguson-Smith 2010), and evidence of H3K4me3 and H3K27me3 enrichment asymmetry on active and repressed alleles has been documented for individual genes (Maupetit-Méhouas et al. 2016). Thus, the allele-specific genomic tracks and dataset integration employed by MEA enhances the visualization of allelic differences between epigenetic marks and transcription across the genome.

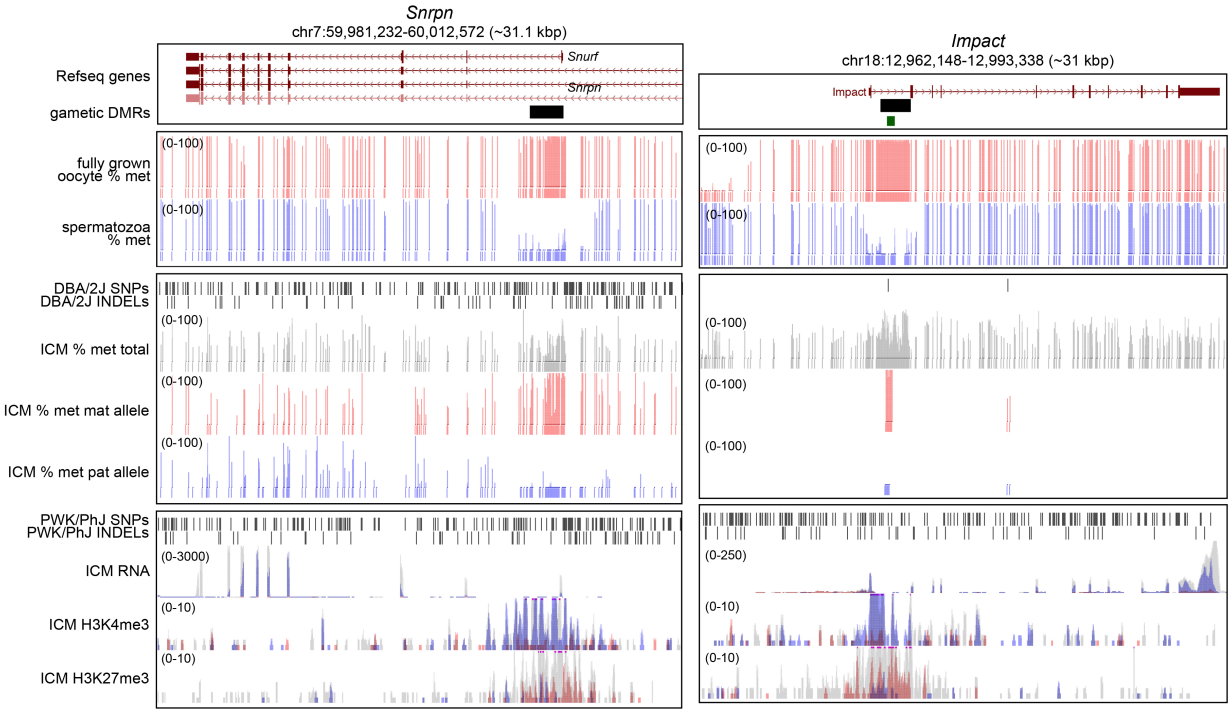

Figure 2.9 Integration of WGBS with allele-specific RNA- and ChIP-seq over the maternally methylated imprinted genes *Snrpn* and *Impact*.

UCSC genome browser screenshot of the *Snrpn* and *Impact* gDMRs and downstream genes using the default MEA output as presented in **Figure 2.7b**. Notably, only the

expressed paternal allele is enriched for H3K4me3 while the inactive maternal allele is enriched for H3K27me3 and DNAme. The *Impact* locus demonstrates that a single genetic variant is apparently sufficient to score DNAme level asymmetry between parental alleles in an F1 hybrid.

2.3.9 Application of the MEA pipeline to human WGBS, RNA-seq and ChIP-seq datasets.

To demonstrate the utility of MEA for the study of HTS datasets from human samples, I used the STAR aligner to analyze an RNA-seq dataset generated from human brain tissue. For individuals whose parental genomic sequences are unavailable, MEA uses Shape-IT (Delaneau et al. 2008) to phase individual genetic variants into inferred haplotypes. For each annotated gene, the haplotype-specific contribution to allelic read alignment was calculated using MEA. As expected, human imprinted genes (White et al. 2016) such as *MEST, MEG3, PEG3* and *PEG10* display monoallelic expression (**Figure 2.10a**), confirming the suitability of MEA for the analysis of RNA-seq data from human samples.

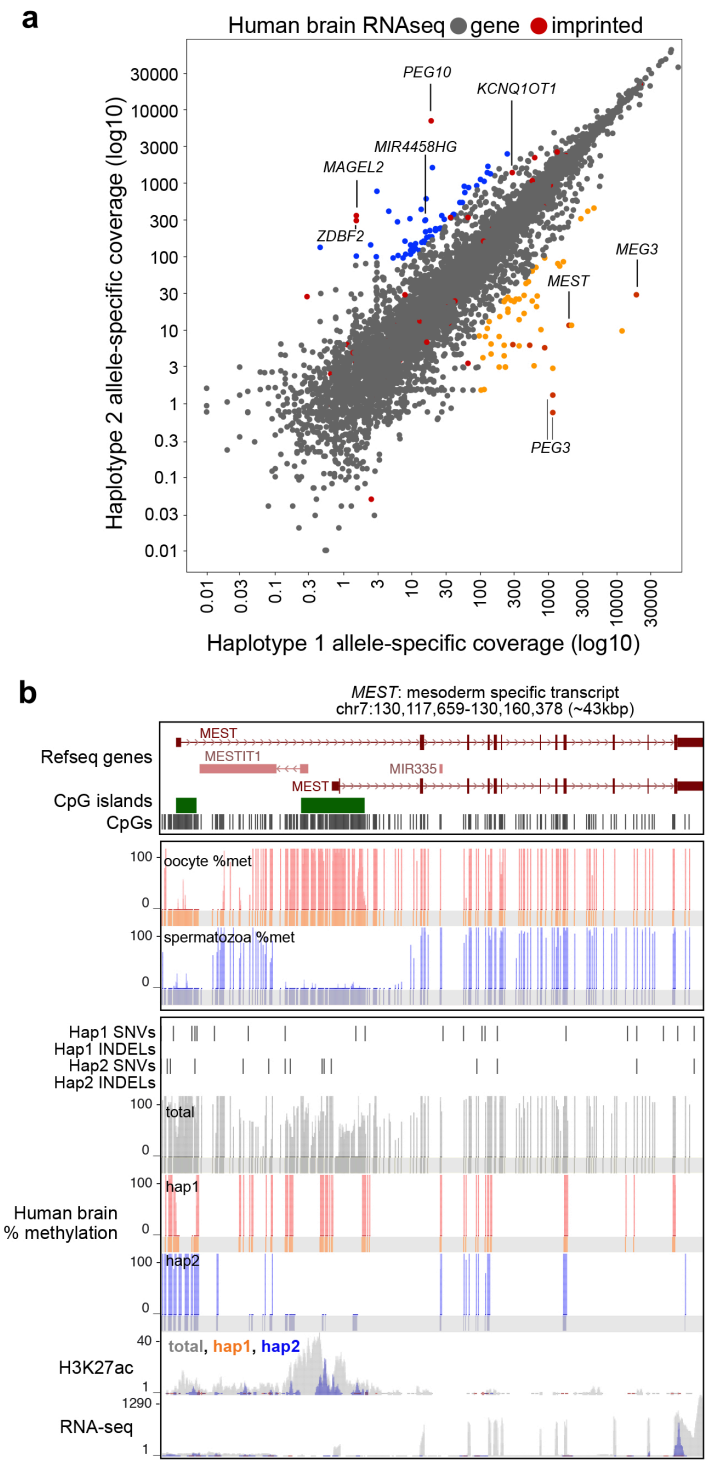

Figure 2.10 Allelic integration of RNA-, ChIP-seq and WGBS datasets from human brain.

(a) Analysis of allele-specific gene expression using RNA-seq data from adult human brain. Imprinted genes are highlighted in red and monoallelically expressed genes (defined by total expression (RPKM >1), allele-specific coverage (mapped reads >100) and expression bias (>90% of transcript levels from one allele)) are highlighted in blue and orange. *MEST*, an imprinted gene, is highly expressed in brain and shows the expected allelic bias. **(b)** UCSC genome browser screenshot of the *MEST* locus showing allele-agnostic (total) and allele-specific (blue and red) DNAme levels in adult brain. DNAme levels in gametes (oocyte & spermatozoa) are also shown (Okae et al. 2014). RNA-seq and H3K27ac ChIP-seq data from human brain were integrated using MEA and allele-agnostic (total) as well as allele-specific coverage is shown for each. Note that only the expressed allele (hap2) is unmethylated and enriched for H3K27ac.

I next generated UCSC Track Hubs to visualize the RNA-seq data analyzed above, as well as matched DNAme (WGBS) and histone PTM (cross-linked ChIP-seq) data from human brain and focused on imprinted genes that include genetic variants in their exons and respective DMRs. Thirteen known imprinted genes were expressed (RPKM >1) and had at least 10 allele-specific mapped read coverage on either allele, 6 of which show >80% expression from one allele. A screen shot of the imprinted *MEST* gene, which is paternally expressed in somatic tissues, is shown in **Figure 2.10b**. As expected, analysis of sperm and oocyte WGBS data from unrelated individuals reveals a DMR at the *MEST* TSS that is methylated exclusively in the oocyte and shows ~50% methylation across the annotated DMR in adult brain cells. MEA output reveals one allele with dense methylation in this region (haplotype 1) and the other with very low methylation (haplotype 2). Importantly, only the latter, which is transcriptionally active, shows enrichment of H3K27ac, a histone modification associated with active genes.

Based on allele-specific DNAme, transcription and histone PTM patterns, we surmise that haplotypes 1 and 2 of the *MEST* locus were inherited from the proband's mother and father, respectively. Taken together, these results reveal that MEA successfully integrates allele-specific RNA-seq data with WGBS and ChIP-seq data for identification and visualization of human loci harbouring genetic variants.

To determine whether H3K27ac shows allele-specific enrichment in the promoter regions of genes exhibiting allele-specific transcription, I identified all genes that harbor genetic variants over annotated exons and the TSS and calculated their allelic ratios (**Figure 2.11a**). While the correlation between expression and H3K27ac allele-specific ratios is low (Pearson $r^2 = 0.29$), many genes displaying strong allele-specific expression bias (over two standard deviations from the mean) are also enriched for H3K27ac on the active allele (χ^2 test p values for bias towards haplotype 1 = 1.38_E -24 and haplotype 2 = 4.8_E -38), as expected. Moreover, manual inspection of a subset of genes displaying monoallelic expression and biallelic H3K27ac reveals that transcription originates at alternative promoters. To further quantify the relationship between allele-specific H3K27ac and transcription, I categorized genes based on allele-specific transcription bias and measured the distribution of allele-specific H3K27ac at TSSs (**Figure 2.11b**). Notably, while allele-specific H3K27ac was positively correlated with transcriptional activity, the ChIP-seq input (control) dataset also showed a higher level of enrichment on the active allele for each haplotype. This observation is consistent with previous studies demonstrating that the promoter regions of active genes are inherently

more sensitive to sonication than inactive genes (Song and Crawford 2010; Buenrostro et al. 2013). That this bias also applies to individual genes exhibiting allelic differences in expression/PTMs reiterates the importance of input-correction of ChIP-seq material and highlights the sensitivity of the MEA pipeline for quantifying allele-specific differences in enrichment.

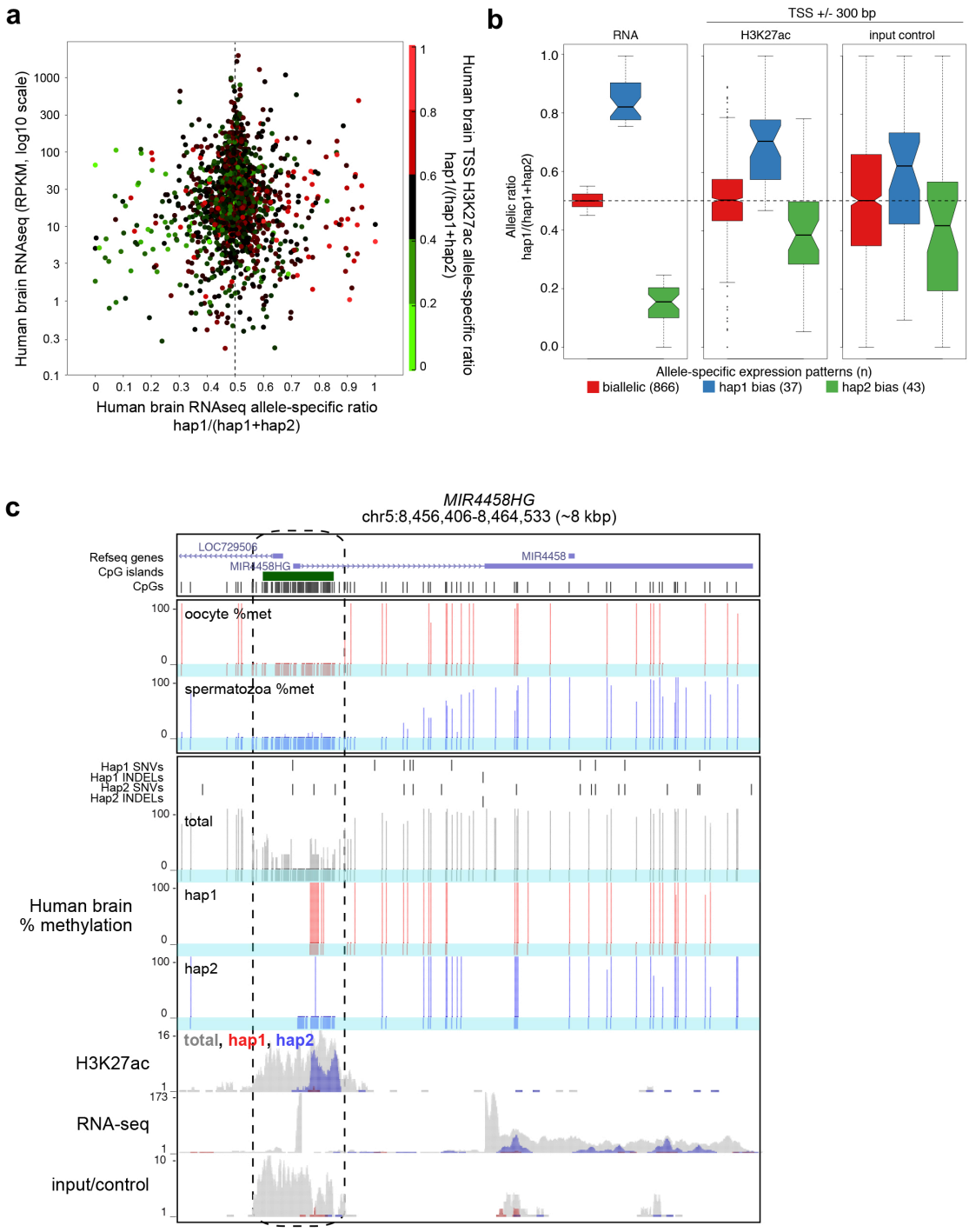

Figure 2.11 Allele-specific transcription, H3K27ac and DNA methylation at the *MIR4458HG* locus.

(a) Integration of allele-specific gene expression and promoter H3K27ac enrichment using human brain RNA-seq and matched ChIP-seq datasets. Only transcripts with

informative allele-specific RNA-seq coverage over exons and ChIP-seq coverage over TSSs (+/- 300 bp) are shown (n = 1,759). **(b)** Distribution of H3K27ac and input/control allelic ratios at TSSs of transcripts expressed from one or both alleles. Note the allelic ratio bias even in the input control **(c)** UCSC genome browser screenshot of the *MIR4458HG* locus. Only the expressed allele (hap2) is enriched for H3K27ac and hypomethylated at the CpG island promoter.

To determine whether MEA can be employed to identify novel monoallelically expressed transcripts in human samples, I revisited the brain RNA-seq data described above. Applying thresholds for total expression (RPKM >1), allele-specific coverage (mapped reads >100) and expression bias (>90% of transcript levels from one allele), I identified 222 monoallelically expressed transcripts (**Figure 2.10a**). Ten of these 222 transcripts showed sufficient H3K27ac ChIP-seq coverage for allele-specific calling (total RPKM >1 and allele-specific CpGs on each allele. While seven of these transcripts (*PIK3R3*, *ZNF662*, *PSMC1*, *LOC145784*, *CYP4F24P*, *C19orf48* and *ZNF805)* showed biallelic or minor allele-specific bias in H3K27ac, perhaps indicative of allele-specific post-transcriptional regulation, three (*MEST*, *MIR4458HG* and *PCDHA5*) showed strong H3K27ac bias toward the active allele (>90% allelic reads). Importantly, the latter represent known and candidate novel imprinted genes. *PCDHA5* belongs to a large gene family of protocadherins, complicating allelic interpretation. However, analysis of the previously described imprinted gene *MEST* (**Figure 2.10b**) and the uncharacterized non-coding RNA gene *MIR4458HG* (**Figure 2.11c**), revealed H3K27ac enrichment and intermediate methylation at their TSSs. As described above for the

MEST gene, allelic deconvolution at the *MIR4458HG* promoter using MEA reveals H3K27ac enrichment and the absence of DNAme exclusively on the active allele. Furthermore, analysis of published WGBS data from gametes reveals hypomethylation of the *MIR4458HG* TSS in both sperm and oocyte, indicating that the allelic gain of DNAme at this locus occurs in somatic tissues. Thus, using MEA to integrate complementary RNA-, ChIP-seq and DNAme datasets allows for the allele-specific resolution of epigenetic states at the regulatory regions of both known and novel monoallelically expressed genes.

2.3.10 Consolidation of all dependencies into a Docker container.

The proper installation and configuration of bioinformatics dependencies is a major hurdle for both new and experienced users. To address this challenge, we packaged MEA into a Docker Container, an open-source software packaging and distribution system (see Materials and Methods). The self-contained nature of the container allows one-step installation of all 15 bioinformatic dependencies (STAR, bwa, Bedtools, Bowtie2, Tophat2, Bismark, Java, etc.), providing a consistent user experience independent of operating system (Windows, MacOS, Linux, etc.). Furthermore, the consolidation of all MEA tool installation steps will greatly facilitate future incorporation of alternative HTS aligners.

2.4 Discussion

The surge of publicly available HTS epigenomic and expression datasets generated by international consortia, including IHEC and PHANTOM, has outpaced the development and dissemination of bioinformatic pipelines that can be used to analyze disparate epigenomic datasets at allelic resolution. To address this need, we developed a universal pipeline that generates integrated allele-specific genomic tracks for DNA methylation (WGBS or Reduced Representation Bisulphite Sequencing (RRBS)), expression (RNA-seq) and histone modification (ChIP-seq) data. Using a unique strategy that incorporates INDELs in addition to SNVs during pseudogenome reconstruction, MEA increases the quality of non-reference genomic sequences, yielding a reduction in reference genome alignment bias. Additionally, in the case of mouse datasets, false positive allele-specific alignments can be minimized by excluding satellite repeats from post-alignment analysis. By considering INDELs and SNVs, MEA captures significantly more allelic CpGs than an INDEL-agnostic script and in turn increases the sensitivity of allele-specific, parent-of-origin DNAme level calculations. Furthermore, by implementing RNA-seq aligners developed specifically to address spliced read alignment, such as STAR (Dobin et al. 2013), MEA reports allele-specific expression over a greater proportion of the transcriptome relative to other aligners.

The fraction of the genome for which allele-specific state can be calculated is a function of several experimental variables, including the choice of parental strains in the case of F1 hybrid studies in model organisms. I was able to measure allele-specific

DNA levels over 20.4% of all CpGs in C57BL/6J x DBA/2J F1 hybrid mice. The DBA/2J strain is quite similar genetically to the reference C57BL/6J, containing on average one SNV per 530 bp (0.19%), at the lower limit of the optimal sequence divergence range of 0.1 to 5% for genome-wide allelic analysis (Wang and Clark 2014). Wild and inbred mouse strains such as PWK/PhJ, CAST/EiJ or SPRET/EiJ are up to eight times more divergent than commonly used strains, such as DBA/2J, 129S1/SvImJ and C3H/HeJ (Keane et al. 2011). Thus, when crossed with any other strain, such F1 hybrids will yield a significant increase in the fraction of informative reads. Regardless of parental genome diversity, the incorporation of INDELs in addition to SNVs during pseudogenome reconstruction, as implemented in MEA, significantly increases the number of regions over which allele-specific methylation can be discerned. For strains with available SNV and INDEL annotations, such as those provided by the Sanger Institute's Mouse Genomes Project (Adams et al. 2015), the average genetic variant frequency between parental genomes can easily be calculated, and in turn, the fraction of the genome likely to be informative for discriminating allele-specific reads determined *a priori*.

By increasing the number of allele-specific reads extracted from HTS datasets of outbred individuals, including F1 hybrid model organisms as well as humans, MEA enables the identification of novel DMRs in WGBS data, allelic-specific gene expression from RNA-seq data and the discrimination of histone marks showing parent-of-origin specific patterns from true bivalent marks by ChIP-seq. As this toolbox was developed

to process high throughput sequencing reads regardless of experiment type, MEA can also be used to analyze additional chromatin features with allelic resolution. For example, to map chromatin accessibility at an allelic level, DNase I hypersensitivity site-sequencing (DNase-seq, (Song and Crawford 2010)) or transposase-accessible chromatin followed by high-throughput sequencing (ATAC-seq, (Buenrostro et al. 2013)) datasets can be interrogated and the results integrated with the data types described above. Importantly, if allele-specific resolution is desirable, previously generated datasets using any of these approaches can be revisited using MEA.

While MEA can be applied to datasets generated from any diploid organism, there are several important limitations that must be considered for clinical studies. As each individual has a unique diploid genome (except in the case of monozygotic twins), pseudogenome reconstruction is essential. While MEA exploits publicly available whole genome sequencing datasets from the Sanger Institute's Mouse Genomes Project (Adams et al. 2015) and the human-focused 1000 genomes project (McVean et al. 2012), additional genotyping and variant-calling steps will be required for haplotypes not covered by these population level sequencing projects. Nevertheless, large-scale efforts such as The Cancer Genome Atlas (TCGA) project that harmonize various cancer-related dataset types, including genotype information, may be analyzed using MEA to deconvolute complex relationships that may operate at an allele-specific level. For example, a recent publication combined genetic, DNAme and gene expression variation to explain aberrant gene regulatory networks in thyroid carcinoma samples (Chen et al.

2017). Given the high frequency of heterozygous somatic mutations in many cancer types, MEA may be applied to directly measure the effect of these mutations on DNAme and gene expression levels on the same allele by using the other allele as a control, potentially allowing for the identification of additional driver mutations. Since *in silico* diploid genome sequences are twice as large as their respective reference assemblies, such population-based studies (encompassing thousands of individuals) will require extensive computational infrastructure. These technical restrictions limit the number of unique individuals that can be practically evaluated. Therefore, for studies encompassing large outbred populations, an alternative approach that combines genotyping and allele-specific read calling is more suitable (Cheung et al. 2017). Nevertheless, for smaller scale epigenomic studies, such as those involving trios, MEA can be applied to study the role of genetics in epigenetic variation, and in turn, to facilitate the discovery or validation of variants of interest, complementing epigenome-wide association studies (EWAS) (Pastinen 2010).

2.5 Conclusion

To our knowledge, MEA is the first software package to provide integrated allele-specific analysis of DNA methylation, histone modification and expression data. Exploiting both SNV and INDEL information, this pipeline increases the sensitivity and specificity of allelic analyses relative to an INDEL-agnostic approach. MEA automates diploid pseudogenome reconstruction, allele-specific read detection and haplotype-resolved genomic track agglomeration for intuitive data visualization and allelic

imbalance detection. With one-step installation and user-friendly file outputs, MEA can be applied without relying on extensive bioinformatic expertise. Intersection of epigenomic and transcriptomic datasets using this novel toolbox will facilitate studies of parent-of-origin effects as well as the interplay between genomic sequence, the epigenome and transcriptional regulation in both humans and model organisms.

Chapter 3: Maternal DNMT3A-dependent de novo methylation of the zygotic paternal genome inhibits gene expression in the early embryo

3.1 Introduction

Male germ cell development in mammals is characterized by widespread de novo DNA methylation and compaction of DNA via histone-to-protamine exchange. As DNAme is largely maintained throughout spermatogenesis, the genomes of mature spermatozoa harbor characteristically high levels of DNAme (Miller et al. 2010). Following fertilization, the paternal genome undergoes another profound change in chromatin state, including replacement of protamines with histones and a global reduction in DNAme before the first S-phase (Mayer et al. 2000; Santos et al. 2002). Subsequently, DNAme levels on both parental genomes are progressively reduced with each DNA replication cycle (Smith et al. 2012). While passive demethylation in the early embryo is likely explained by sequestration of the maintenance DNA methyltransferase DNMT1 in the cytoplasm (Maenohara et al. 2017; Li et al. 2018; Han et al. 2019), the mechanism of active demethylation remains controversial. Indeed, while TET3-mediated oxidation followed by BER has been implicated in this process, a TET-independent mechanism is also clearly involved (Hajkova et al. 2010; Gu et al. 2011b; Guo et al. 2014; Peat et al. 2014; Tsukada et al. 2015; Amouroux et al. 2016; Kweon et al. 2017). Regardless, WGBS analyses reveal that DNAme levels on both parental genomes reach a nadir in ICM cells of E3.5 mouse blastocysts, followed by widespread de novo DNAme during post-

implantation development (Wang et al. 2014; 2018). This wave of genome-wide demethylation followed by remethylation is conserved in human embryonic development, albeit with slower kinetics (Smith et al. 2014; Eckersley-Maslin et al. 2018; Zhu et al. 2018). Notably, disruption of the machinery required for the establishment or maintenance of DNAme result in infertility and/or embryonic lethality in mice, revealing the importance of DNAme homeostasis in early mammalian development (Okano et al. 1999; Bourc'his et al. 2001; Hata et al. 2002; Bourc'his and Bestor 2004; Kaneda et al. 2004; Hirasawa et al. 2008; Tsukada et al. 2015; Li et al. 2018; Han et al. 2019).

In contrast, the vast majority of CGIs are hypomethylated in mature male and female germ cells, as well as in early embryonic development (Hammoud et al. 2009; Brykczynska et al. 2010; Erkek et al. 2013; Shirane et al. 2013; Qu et al. 2017; Edwards et al. 2017). In addition, CGIs retain nucleosomes in spermatozoa, foregoing the exchange for protamines (Hammoud et al. 2009; Brykczynska et al. 2010; Erkek et al. 2013). These retained nucleosomes include canonical H3 or its variant H3.3, which are enriched for di- and/or tri-methylation on lysine 4 (H3K4me2/3) (Erkek et al. 2013; Siklenka et al. 2015; Yamaguchi et al. 2018). Although the extent of histone PTM maintenance following fertilization is controversial (Siklenka et al. 2015; Zhang et al. 2016; Xu et al. 2019), the persistence and/or rapid deposition of H3K4 methylation at CGI promoters may protect these regions against de novo DNAme (Ooi et al. 2007) in the developing male germline as well as the early embryo. Indeed, most CGI promoters are enriched for H3K4me3 and remain hypomethylated on both parental genomes

throughout early embryonic development and in adult tissues (Smallwood et al. 2011; Wang et al. 2014). H3K4me2/3 also likely facilitate the initiation of transcription from the paternal genome during zygotic gene activation (Hammoud et al. 2009; Siklenka et al. 2015), which marks the transition between oocyte and embryonic transcriptional programmes (Eckersley-Maslin et al. 2018). As in sperm, CGIs in oocytes are generally hypomethylated and harbor nucleosomes enriched for H3K4me3 (Zhang et al. 2016) and/or H3K27me3 (Zheng et al. 2016; Inoue et al. 2017). However, a subset of CGIs, including genic promoters and maternal gDMRs, are de novo methylated in growing oocytes (Stewart et al. 2015). These exceptional CGIs are enriched for H3K36me3 and methylated by the DNMT3A/DNMT3L complex, which is highly expressed in oocytes (Brind'Amour et al. 2018; Xu et al. 2019). Notably, maternal DNMT3A is also clearly detected in both parental pronuclei in mouse zygotes (Hirasawa et al. 2008; Amouroux et al. 2016). Furthermore, a recent study employing IF and ultrasensitive liquid chromatography/mass spectrometry revealed that the zygotic paternal genome is subject to de novo DNAme (Amouroux et al. 2016). However, the genomic regions subject to such **paternal DNAme acquisition (PDA)** in the early embryo and relevance to transcription was not addressed.

To determine which loci gain DNAme immediately following fertilization, I employed MEA (Chapter 2) to carry out an allele-specific analysis of WGBS data from 2-cell (2C) F1 hybrid embryos (Wang et al. 2014) and identified specific genomic regions, including CGI promoters, that show PDA. Corroborating these enigmatic

findings, I observe PDA of an overlapping set of CGI promoters in androgenetic but not in parthenogenetic blastocysts. Allele-specific analysis of ChIP-seq data from 2C embryos reveals that PDA is accompanied by loss of H3K4me3 over the same regions, consistent with the conjecture that such DNAme may inhibit transcription from the paternal allele in early embryonic development. Indeed, I show that PDA is lost in the absence of maternal DNMT3A and a subset of hypomethylated genes are concomitantly upregulated specifically from the paternal allele in the resulting mutant 4C embryos. Taken together, these experiments reveal that the role of DNMT3A as a maternal effect gene extends beyond maternal imprinting, as it methylates a subset of genes on the paternal genome in the zygote, which inhibits their expression in preimplantation development.

3.2 Materials and Methods

3.2.1 Ethical approval for animal work.

All animal experiments were performed under the ethical guidelines of Kyushu University and Tokyo University of Agriculture.

3.2.2 Isolation of androgenetic blastocyst.

Diploid androgenones were prepared as described previously (Kono et al. 1993). Oocyte and spermatozoa were isolated from B6D2F1/Jcl and C57BL/6NJcl mice (Clea Japan, Tokyo, Japan), respectively. Briefly, enucleated oocytes were *in vitro* fertilized and zygotes with two male pronuclei were cultured for 4 days in KSOM medium at 37^{o}C

and 5% CO_2 (Obata et al. 2000). 5 blastocyst pools were collected in duplicate for WGBS library construction.

3.2.3 *Dnmt3a* maternal KO embryo culture and genotyping.

Dnmt3a KO oocytes were generated using *Dnmt3a*^{2lox} and *Zp3*-cre C57BL/6J mice as described previously (Kaneda et al. 2004; de Vries et al. 2000). Superovulation was induced using PMSG/hCG and MII oocytes were collected from oviducts. *Dnmt3a*^{2lox};*Zp3*-cre MII oocytes were artificially inseminated with DBA/2J spermatozoa. Cumulus cells were removed using hyaluronidase after insemination and embryos were cultured in KSOM at 37^{o}C and 5% CO_2. Early-mid 2C embryos were collected at 22 hours (WGBS) and 24 hours (RNA-seq), 4C embryos at 36 hours and blastocysts at 96 hours. Zona pellucida and polar bodies of 2C embryos were removed (WGBS). ICM cells were purified by immunodissection using anti-mouse IgG (Cedarlane) and guinea pig complement (Rockland) (Solter and Knowles 1975). Genotyping was performed by PCR using primers for *Dnmt3a* (CTGTGGCATCTCAGGGTGATGAGCA and GCAAACAGACCCAACATGGAACCCT) and the *Zp3*-cre transgene (GCAGAACCTGAAGATGTTCGCGAT and AGGTATCTCTGACCAGAGTCATCC).

3.2.4 DNMT3A immunofluorescence.

Dnmt3a^{2lox};*Zp3*-cre females were mated with JF1 males and IF was performed on one-cell zygotes. DNMT3A was detected using the IMG-268 IMGENEX antibody, and DNA

was counterstained using propidium iodide, as described previously (Hirasawa et al. 2008).

3.2.5 WGBS and RNA-seq library construction and sequencing.

Lysates of androgenetic blastocysts were spiked with 0.1 ng lambda phage DNA and subjected to WGBS library construction according to the PBAT protocol for single-read sequencing (Kobayashi et al. 2013). DNA from 20-30 pooled 2C embryos per replicate was purified and spiked with 1% unmethylated lambda phage DNA, and WGBS libraries were generated by PBAT with 4 cycles of library amplification (Au Yeung et al. 2019). All WGBS libraries had >99% bisulphite conversion rates. Total RNA was extracted from 20-40 pooled 2C embryos, 5-10 4C embryos and 2-7 blastocysts per replicate using Trizol reagent. Strand-specific RNA-seq libraries were generated using NEBNext: rRNA Depletion Kit, RNA First Strand Synthesis Module, Ultra Directional RNA Second Strand Synthesis Module, and Ultra II DNA Library Prep Kit. Libraries were sequenced on HiSeq 1500 or HiSeq 2500 (WGBS: HCS v2.2.68 and RTA v1.18.66.3) (Toh et al. 2017). See **Table 3.1** for full sequencing and alignment statistics.

3.2.6 HTS data processing.

Reads were trimmed using Trimmomatic v0.32 (Bolger et al. 2014) and processed for total and allele-specific alignments using MEA v1.0 (Richard Albert et al. 2018) using default parameters and the mm10 reference genome. Random primer extension sequences (the first 4 bases from the 5' end) were removed for all PBAT sequences. All

publicly available HTS data (**Table 3.2**) was reprocessed as above, with the exception of WGBS datasets from sperm (Kubo et al. 2015), oocytes (Shirane et al. 2013), parthenotes (Brind'Amour et al. 2018) and primordial germ cells (Kobayashi et al. 2013), which were previously processed and filtered using identical parameters as in this study.

3.2.7 WGBS data analysis.

DNAme levels over individual CpGs with ≥5x coverage (including allele-specific) were scored. For allele-specific alignments of WGBS datasets generated in this study (WT and *Dnmt3a* matKO 2C embryos), CpGs with ≥1x coverage were scored. DNAme levels were calculated over CGI promoters using Bedops v2.4.27 (Neph et al. 2012) and visualized using VisRseq v0.9.12 (Younesy et al. 2015). Only CGI promoters that overlapped at least 2 informative CpGs separated by the maximum sequencing read length of the library were kept. Genome-wide 2- and 20-kb bins were generated using Bedtools, and bins covered by at least 4 CpGs separated by over 1 read length in each dataset were used, and a random subset of 1,000 bins were visualized as parallel coordinate plots using VisRseq. Hypomethylated regions in sperm were defined on the basis of WGBS data from C57BL/6J (Kubo et al. 2015) and DBA/2J (Wang et al. 2014) strains. CpGs with <20% DNAme were scored as hypomethylated, those within 500 bp of one another were grouped, and regions overlapping at least 5 grouped CpGs were analyzed. This strategy yielded a list of 45,259 regions of on average 1250 bp in length and 4.01% average DNAme in both datasets. Of these, 11,110 had sufficient

information to calculate paternal DNAme levels in both allele-specific DBA/2J sperm and 2C WGBS data from Wang et al. (Wang et al. 2014).

3.2.8 ChIP-seq data analysis.

Raw sequencing reads were reprocessed as described above into total and allele-specific genomic tracks. RPKM values were calculated over TSSs (+/- 300bp) using VisRseq on the basis of sequencing-depth normalized genomic tracks. ChromHMM v1.12 (Ernst and Kellis 2012) was employed to define distinct chromatin states (LearnModel, k=6) on the basis of filtered BAM files (BinarizeBam) using default parameters.

3.2.9 RNA-seq data processing.

Raw sequencing reads were reprocessed as described above into total and allele-specific genomic tracks. Gene expression (RPKM) values over genic exons were calculated using VisRseq (NCBI Refseq). K-means clustering was performed on the basis of log10 transformed RPKM values from developing and mature spermatozoa (Gaysinskaya et al. 2018), oocyte and preimplantation embryo (Wu et al. 2016) data using the Hartigan-Wong algorithm. The mean RPKM value for each state was then plotted on a parallel-coordinate plot using VisRseq. Correlograms were generated using Morpheus (https://software.broadinstitute.org/morpheus) on the basis of log2 transformed values. For genome browser visualization, biological replicates were merged and total (allele-agnostic) and allele-specific genomic tracks were organized

into UCSC Track Hubs as described previously (Richard Albert et al. 2018). Total, paternal and maternal-specific changes in gene expression were calculated using DESeq2 (Love et al. 2014). Genes with a ≥2-fold change in expression and a Benjamini-Hochberg corrected p-value ≤ 0.01 were considered differentially expressed. Transcription initiation sites for *Dnmt3a* matKO and wild-type embryos were determined using StringTie v1.3.5 (Pertea et al. 2015).

3.2.10 Motif analysis.

Known and de novo DNA motif discovery was conducted using HOMER findMotifs.pl v4.11.1 (Heinz et al. 2010) and the MEME suite (Bailey et al. 2009). The sequences of CGI promoters showing PDA were input in fasta format, default parameters were used and CGIs showing persistent paternal hypomethylation (n=4,315) were used as background sequences. Both C57BL/6J and DBA/2J sequences were tested.

3.2.11 Statistical tests.

T-tests of two samples assuming unequal variances were performed when comparing the distribution of DNAme or H3K4me3 levels between CGI promoters that show PDA or persistent hypomethylation.

3.2.12 Data availability.

Datasets generated in this study have been deposited in GEO under the accession number GSE141877. See **Tables 3.1 and 3.2** or the full list of data analyzed for this study.

Table 3.1 Datasets generated in this study.

Stage	Strain	Dnmt 3a	Data type		Read length	Rep.	Sequenced reads (pairs)	Uniquely aligned reads	Average coverage	Bisulphite conversion rate
			WGBS	RNA-seq						
2C	C57BL/6N x DBA2J	WT			108 SE	rep 1	310,677,626	121,850,552	3.7	99.51%
2C	C57BL/6N x DBA2J	WT			108 SE	rep 2	135,334,173	48,417,043	1.4	99.50%
2C	C57BL/6N x DBA2J	KO			108 SE	rep 1	305,628,174	111,290,069	3.3	99.53%
2C	C57BL/6N x DBA2J	KO			108 SE	rep 2	126,869,065	41,450,542	1.2	99.54%
2C	C57BL/6N x DBA2J	WT			108 PE	rep 1	95,469,627	31,940,689		
2C	C57BL/6N x DBA2J	WT			108 PE	rep 2	108,069,613	45,989,770		
2C	C57BL/6N x DBA2J	KO			108 PE	rep 1	103,727,387	43,946,497		
2C	C57BL/6N x DBA2J	KO			108 PE	rep 2	108,981,430	52,568,580		
4C	C57BL/6N x DBA2J	WT			108 PE	rep 1	10,618,215	3,219,907		
4C	C57BL/6N x DBA2J	WT			108 PE	rep 2	40,825,340	10,266,357		
4C	C57BL/6N x DBA2J	KO			108 PE	rep 1	53,248,652	12,572,929		
4C	C57BL/6N x DBA2J	KO			108 PE	rep 2	33,401,187	7,693,336		
ICM	C57BL/6N x DBA2J	WT			108 PE	rep 1	43,229,550	15,439,045		
ICM	C57BL/6N x DBA2J	WT			108 PE	rep 2	39,091,478	14,082,043		
ICM	C57BL/6N x DBA2J	KO			108 PE	rep 1	50,179,669	12,560,866		
ICM	C57BL/6N x DBA2J	KO			108 PE	rep 2	89,343,385	29,776,235		
Andro. blasto.	C57BL/6N	WT			101 SE	rep 1	133,484,009	78,733,582	2.4	99.34%
Andro. blasto.	C57BL/6N	WT			101 SE	rep 2	110,941,095	63,639,012	1.9	99.38%

Table 3.2 Datasets mined in this study and their source.

Stage	Strain	Data type			Accession(s)	Reference
		WGBS	ChIP-seq	RNA-seq		
Spermatozoa	DBA/2J				GSM1386020	Wang et al. 2014
MII oocyte	C57BL/6J				GSM1386019	Wang et al. 2014
2C	C57BL/6JxDBA/2J				GSM1386021	Wang et al. 2014
4C	C57BL/6JxDBA/2J				GSM1386022	Wang et al. 2014
ICM	C57BL/6JxDBA/2J				GSM1386023	Wang et al. 2014
E6.5 epiblast	C57BL/6JxDBA/2J				GSM1386024	Wang et al. 2014
E7.5 epiblast	C57BL/6JxDBA/2J				GSM1386025	Wang et al. 2014
E10.5 PGC male	C57BL/6N				DRS001891	Kobayashi et al. 2013
E10.5 PGC female	C57BL/6N				DRS001892	Kobayashi et al. 2013
E13.5 PGC male	C57BL/6N				DRS001893	Kobayashi et al. 2013
E13.5 PGC female	C57BL/6N				DRS001894	Kobayashi et al. 2013
E16.5 PGC male	C57BL/6N				DRS001895	Kobayashi et al. 2013
E16.5 PGC female	C57BL/6N				DRS001896	Kobayashi et al. 2013
GV oocyte	C57BL/6N				DRS001541	Shirane et al. 2013
GV oocyte Dnmt3a matKO	C57BL/6N				DRS001543	Shirane et al. 2013
Parthenogenetic blastocyst	C57BL/6N				DRA006679	Brind'Amour et al. 2018
Spermatozoa	C57BL/6		H3K4me3		GSM1046832-3	Erkek et al. 2013
Spermatozoa	C57BL/6		H3K27me3		GSM1046834-5	Erkek et al. 2013
Spermatozoa	C57BL/6		Mnase		GSM1046827-8	Erkek et al. 2013
Spermatozoa	C57BL/6		H3K4me2		GSM1337514-5	Siklenka et al. 2015
Spermatozoa	C57BL/6		MNase		GSM1845266-7	Siklenka et al. 2015
MII oocyte	C57BL/6N		H3K4me3		GSM1845262-3	Zhang et al. 2016
PN3 zygote	DBA/2N		H3K4me3		GSM2101161	Zhang et al. 2016
PN5 zygote	C57BL/6N x PWK		H3K4me3		GSM1845264-5	Zhang et al. 2016
E2C	C57BL/6N x PWK		H3K4me3		GSM1845266-7	Zhang et al. 2016
L2C	C57BL/6N x PWK		H3K4me3		GSM1845268-9	Zhang et al. 2016
4C	C57BL/6N x PWK		H3K4me3		GSM1845270-1	Zhang et al. 2016
8C	C57BL/6N x PWK		H3K4me3		GSM1845272-3	Zhang et al. 2016
ICM	C57BL/6N x PWK		H3K4me3		GSM1845274-5	Zhang et al. 2016
ICM	C57BL/6N x PWK				GSM1845307-8	Zhang et al. 2016

Spermatozoa	C57BL/6				DRA002477	Kubo et al. 2015
MII oocyte	C57BL/6N		H3K27me3		GSM2041070-1	Zheng et al. 2016
PN5 zygote	C57BL/6N x PWK		H3K27me3		GSM2041072-3	Zheng et al. 2016
E2C	C57BL/6N x PWK		H3K27me3		GSM2041074	Zheng et al. 2016
L2C	C57BL/6N x PWK		H3K27me3		GSM2041075-6	Zheng et al. 2016
ICM	C57BL/6N x PWK		H3K27me3		GSM2041078-9	Zheng et al. 2016
Spermatogonia	C57BL/6				PRJNA326117	Gaysinskaya et al. 2018
Preleptotene	C57BL/6				PRJNA326117	Gaysinskaya et al. 2018
Leptotene	C57BL/6				PRJNA326117	Gaysinskaya et al. 2018
Zygotene	C57BL/6				PRJNA326117	Gaysinskaya et al. 2018
Pachytene	C57BL/6				PRJNA326117	Gaysinskaya et al. 2018
Diplotene	C57BL/6				PRJNA326117	Gaysinskaya et al. 2018
Spermatozoa	C57BL/6				PRJNA326117	Gaysinskaya et al. 2018
Spermatozoa	C57BL/6 x DBA2		H3K9me3		GSM2588560-1	Wang et al. 2018
MII oocyte	C57BL/6 x DBA2		H3K9me3		GSM2588563-4	Wang et al. 2018
1C zygote	C57BL/6 x DBA2		H3K9me3		GSM2588656-7	Wang et al. 2018
E2C	C57BL/6 x DBA2		H3K9me3		GSM2588659-60	Wang et al. 2018
L2C	C57BL/6 x DBA2		H3K9me3		GSM2588662-4	Wang et al. 2018
ICM	C57BL/6 x DBA2		H3K9me3		GSM2588666-7	Wang et al. 2018
MII oocyte	C57BL/6		H3K36me3		GSM3084624-5	Xu et al. 2019
1C zygote	C57BL/6N x PWK		H3K36me3		GSM3084627-8	Xu et al. 2019
L2C	C57BL/6N x PWK		H3K36me3		GSM3084629-30	Xu et al. 2019
ICM	C57BL/6N x PWK		H3K36me3		GSM3084633-4	Xu et al. 2019
MII oocyte	C57BL/6				GSM1933935-6	Wu et al. 2016
1C zygote	C57BL/6 x DBA2				GSM1625860-1	Wu et al. 2016
E2C	C57BL/6 x DBA2				GSM1933937-8	Wu et al. 2016
L2C	C57BL/6 x DBA2				GSM1625862-3	Wu et al. 2016
4C	C57BL/6 x DBA2				GSM1625864-5	Wu et al. 2016
8C	C57BL/6 x DBA2				GSM1625866-7	Wu et al. 2016
ICM	C57BL/6 x DBA2				GSM1625868-9, GSM1625872	Wu et al. 2016

3.3 Results

3.3.1 De novo DNAme of the paternal genome following fertilization

To trace parent-specific DNAme levels following fertilization and throughout embryonic development with single-nucleotide resolution, I first processed publicly available WGBS data derived from primordial germ cells (PGCs), spermatozoa and MII oocytes (Kobayashi et al. 2013; Kubo et al. 2015; Shirane et al. 2013). I then applied our recently developed allele-specific pipeline MEA (Richard Albert et al. 2018) to WGBS data generated from 2C, 4C, ICM, E6.5 and E7.5 F1 hybrid embryos (Wang et al. 2014). This integrated analysis yielded female/maternal and male/paternal DNAme profiles (**Figure 3.1a-b**). Consistent with previous IF data (Mayer et al. 2000; Santos et al. 2002), comparison of pre- and post-fertilization DNAme levels specifically in mature gametes and 2C embryos (55X coverage) reveals an overall decrease on the maternal and paternal genomes of 8% and 43%, respectively (**Figure 3.1c**). Surprisingly however, coincident with global DNAme loss across the paternal genome, robust DNAme gain (defined as an increase of ≥30%) was detected at ~2% of all hypomethylated regions in sperm, totalling 389 kbp of the mappable genome (**Figure 3.1d**). De novo DNAme of the paternal genome is consistent with the presence of maternal DNMT3A in the zygotic paternal pronucleus (**Figure 3.1e**).

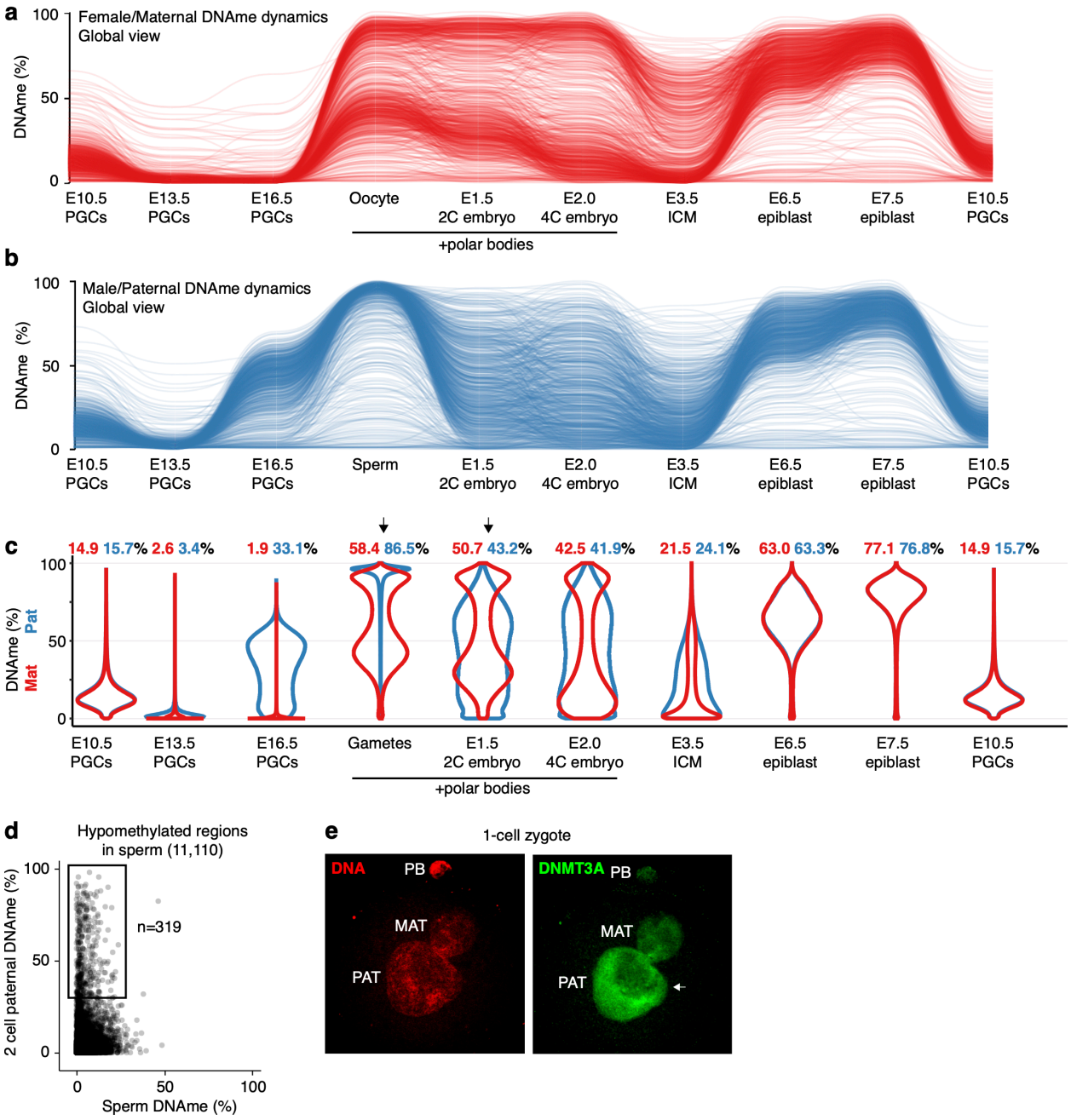

Figure 3.1 Female/maternal and male/paternal DNAme level dynamics during gametogenesis and embryonic development.

(a-b) Parallel coordinate plots illustrating average DNAme levels over 2 kb genomic windows during **(a)** female/maternal and **(b)** male/paternal gametogenesis and embryonic development. Allele-specific analysis of WGBS data yielded 176,240

windows (roughly 15% of the mouse genome) overlapping at least 5 informative CpGs for which I could infer both maternal and paternal DNAme levels in all datasets. A random set of 1,000 windows on chromosome 19 are shown. MII oocyte, 2C and 4C embryo datasets include polar bodies. **(c)** Distribution of global DNAme levels during gametogenesis and embryonic development. Mean DNAme percentages for each stage (red: female/maternal, blue: male/paternal) are shown above each violin plot. **(d)** 2D scatterplot showing post-fertilization paternal DNAme dynamics at genomic regions that are hypomethylated in sperm (mean DNAme=4%, mean length=1,250 bp, n=11,110). **(e)** IF analysis of DNMT3A in a representative 1 cell zygote. Zygotic DNA is counterstained with propidium iodide (PI) PB; polar body. The white arrow indicates DNMT3A staining in the paternal pronucleus.

Remarkably, regions showing clear evidence of paternal DNAme acquisition (PDA) include a number of annotated genic transcription start sites (TSSs) (**Figure 3.2**). Other regions include enhancers and endogenous retroviral elements, but neither of these classes of elements were enriched for PDA (data not shown). While PDA is not restricted to CpG-rich regions, I focused our analyses on CGI promoters, as DNAme is reported to have the strongest impact on transcription of this class of promoters (Deaton and Bird 2011; Weber et al. 2007).

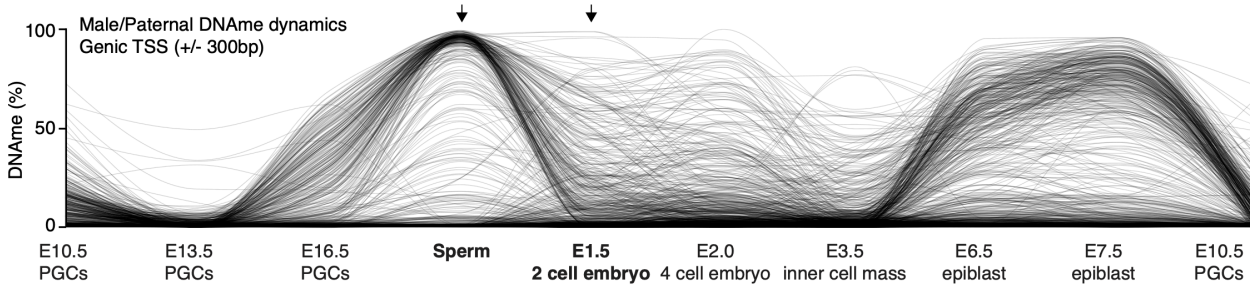

Figure 3.2 Paternal DNAme dynamics over genic promoters throughout germline and embryonic development

Parallel coordinate plot showing male/paternal DNAme level dynamics at genic promoters (TSS +/- 300 bp) during male gametogenesis and embryogenesis. Each line represents mean DNAme level over a promoter overlapping at least 2 informative CpGs (covered by 5X reads and separated by >1 sequencing read length) at all stages (n=3,925). A random set of 1,000 promoters are shown, and the two arrows mark the developmental stages at which we measured post-fertilization paternal DNAme acquisition.

To generate a curated list of CpG-rich hypomethylated TSSs in sperm, I first categorized promoters by CpG density (high, intermediate and low), as described for the human genome (Weber et al. 2007). As expected, DNAme levels and CpG density are anti-correlated (**Figure 3.3a**). To minimize the potential confounding effects of strain-specific differences when comparing DNAme levels of parental genomes, I determined the variation in DNAme levels in C57BL/6J versus DBA/2J sperm using published WGBS datasets (Wang et al. 2014; Kubo et al. 2015). Methylation profiles from these strains show a strong correlation (r^2=0.98, **Figure 3.3b**), with 20,163 promoters showing consistent DNAme levels (high or low) in both strains. Promoters showing strain-specific DNAme, including 185 and 82 showing DBA/2J and C57BL/6J specific hypomethylation, respectively, were excluded from further analysis. Taking advantage of the fact that DNAme and H3K4me3 are anticorrelated in soma and germ cells, I further refined our list of CGI promoter TSSs using publicly available H3K4me3 ChIP-seq data from C57BL/6J sperm (Erkek et al. 2013). As expected, H3K4me3

enrichment levels show a positive correlation with CpG density and negative correlation with DNAme (**Figure 3.3a**).

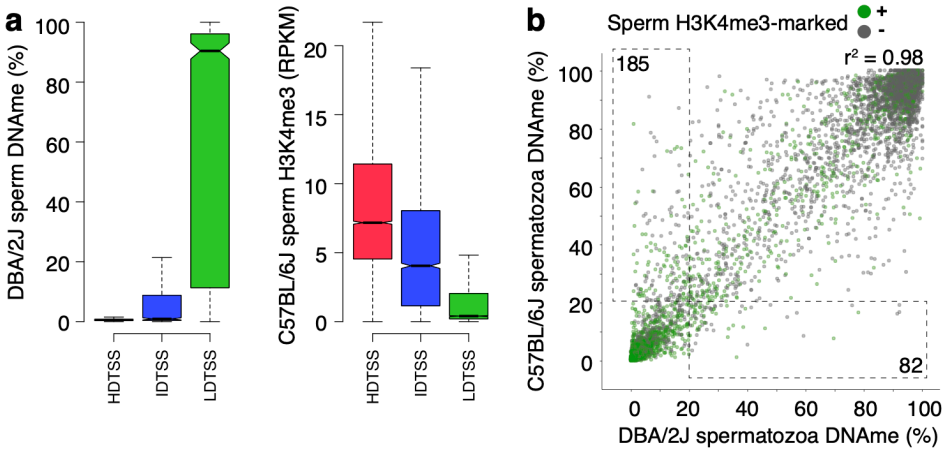

Figure 3.3 Defining hypomethylated CGI promoters in sperm.

(a) Promoters (TSSs +/- 300 bp) were classified by CpG density as described previously (Weber et al. 2007), and the distribution of DNAme and H3K4me3 levels for each promoter category is shown. HD: high CpG density, ID: intermediate CpG density, LD: low CpG density. **(b)** 2D scatterplot illustrating the correlation of DNAme levels between C57BL/6J and DBA/2J sperm. Data points are coloured by H3K4me3 enrichment in C57BL/6J sperm. Only promoters for which I calculated DNAme levels in both datasets (at least 2 CpGs with 5X coverage separated by >1 sequencing read length) are shown (n=20,583). 20,163 promoters had consistent DNAme levels (difference <20%) between strains.

Selecting TSSs that show H3K4me3-enrichment (RPKM ≥1) and intermediate to high CpG density (CGI promoters, CpG ratio ≥0.12) yielded a list of 12,253 CGI promoters, the vast majority of which are hypomethylated in sperm (mean % DNAme; C57BL/6J=2.1, DBA/2J=2.0). Of these, 4,434 harbor a SNV or INDEL and had sufficient

WGBS coverage (5X allele-specific read coverage over ≥2 CpGs separated by >1 sequencing read length) to score DNAme levels in both sperm and the paternal genome of C57BL/6J x DBA/2J F1 2C embryos (**Figure 3.4a**). To unequivocally identify targets of post-fertilization de novo DNAme, I focussed on these CGI promoters.

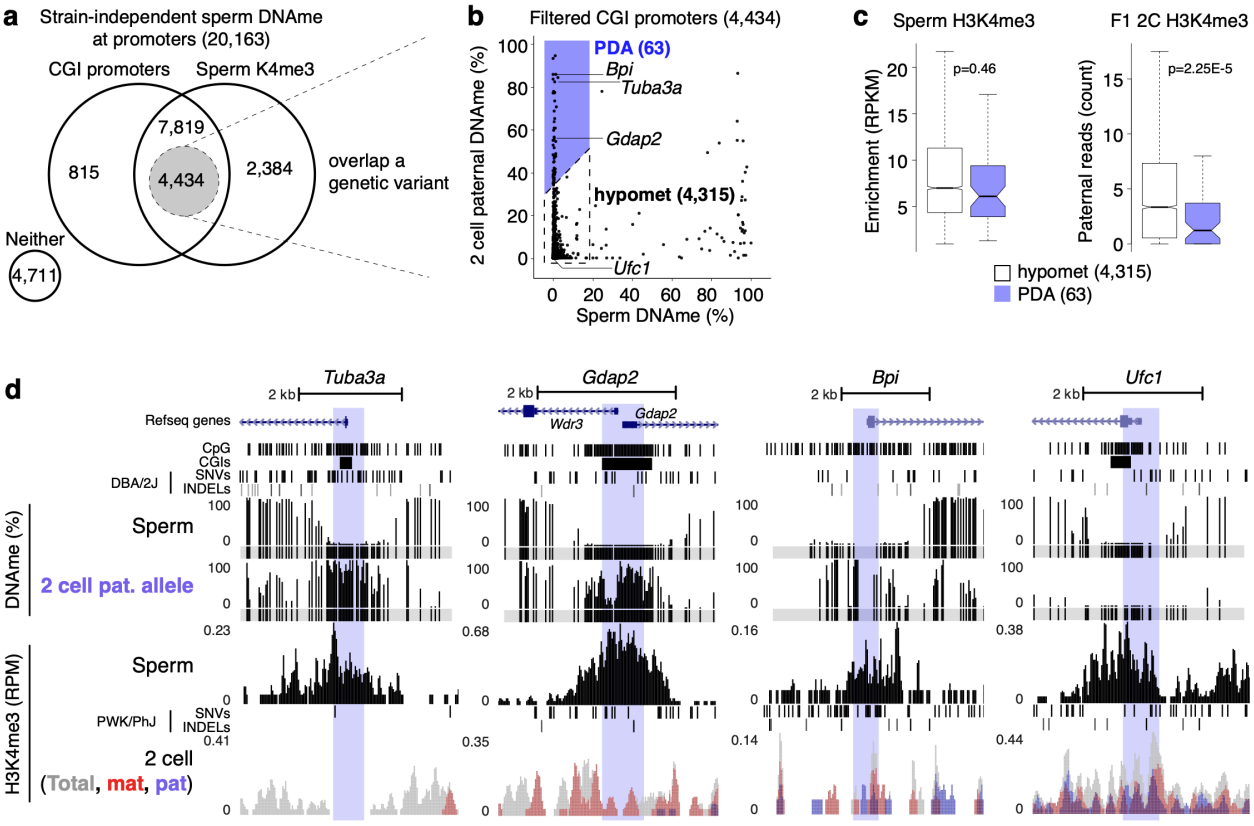

Figure 3.4 Paternal DNAme acquisition at CpG-rich promoters following fertilization.

(a) Venn diagram including all annotated autosomal promoters with sperm DNAme levels consistent between C57BL/6J and DBA/2J strains. The proportion of such promoters with high/intermediate CpG density (CGIs) and enriched for H3K4me3 in sperm is depicted. The subset overlapping a genetic variant and at least 2 informative CpGs in sperm and 2C WGBS datasets are highlighted. **(b)** CGI promoters showing

PDA (≥30% gain in paternal DNAme) from sperm to the 2C stage (n=63) versus persistent hypomethylation (hypomet, n=4,315) are shown. **(c)** Distribution of H3K4me3 levels at PDA or hypomet CGI promoters is shown for sperm (left) and on the paternal allele in 2C embryos (right). **(d)** UCSC genome browser screenshots of the promoter regions of *Tuba3a*, *Gdap2*, *Bpi* and *Ufc1*. TSS regions (+/- 300 bp) are highlighted in blue and the location of informative CpGs (5X coverage) for each WGBS dataset are highlighted in grey below each WGBS dataset. The genomic locations for NCBI Refseq genes, all CpG dinucleotides, CGIs and genetic variants used in our allele-specific analyses (SNVs and INDELs) are also included. 2C H3K4me3 data is represented as a composite track containing total (allele-agnostic, grey), maternal (red) and paternal (blue) genomic tracks.

While the vast majority of CGI promoters hypomethylated in sperm maintain low paternal DNAme levels in 2C embryos, 63 showed clear evidence of PDA (defined as a ≥30% gain, **Figure 3.4b** and **Table 3.3**).

Table 3.3 List of genes that show PDA at their CGI promoters.

Genomic coordinates - mm10	Gene name	Genomic coordinates (mm10)	Gene name
chr1:38987513-38988114	*Pdcl3*	chr3:32817325-32817926	*Usp13*
chr1:39720688-39721289	*Rfx8*	chr3:100162102-100162703	*Wdr3*
chr1:135765784-135766385	*Phlda3*	chr3:100162162-100162763	*Gdap2*
chr1:180903973-180904574	*Pycr2*	chr3:100969362-100969963	*Ttf2*
chr1:192136597-192137198	*Gm10516*	chr3:105704298-105704899	*Fam212b*
chr10:82047826-82048427	*Zfp873*	chr4:134396019-134396620	*Pafah2*
chr10:86498595-86499196	*Syn3*	chr5:124112037-124112638	*Ogfod2*
chr11:87616863-87617464	*Hsf5*	chr5:142960054-142960655	*Fscn1*
chr11:117780382-117780983	*Tmc6*	chr5:147400188-147400789	*Flt3*

chr14:101199768-101200369	*Prr30*	chr6:65042366-65042967	*Smarcad1*
chr15:73839370-73839971	*Mroh5*	chr6:115994704-115995305	*Plxnd1*
chr16:91010948-91011549	*4930404I05Rik*	chr6:125285741-125286342	*Tuba3a*
chr16:91011007-91011608	*Synj1*	chr6:136661592-136662193	*Plbd1*
chr17:24209086-24209687	*Ntn3*	chr7:4844395-4844996	*Shisa7*
chr17:26561211-26561812	*Ergic1*	chr7:16221731-16222332	*Dhx34*
chr17:33929593-33930194	*Rgl2*	chr7:18910103-18910704	*Ccdc61*
chr17:34031511-34032112	*Rxrb*	chr7:31150749-31151350	*Gramd1a*
chr17:35393798-35394399	*H2-Q5*	chr7:40898023-40898624	*A230077H06Rik*
chr18:34651435-34652036	*Cdc23*	chr7:44383825-44384426	*Syt3*
chr18:37706928-37707529	*Pcdhga6*	chr7:97696356-97696957	*Clns1a*
chr19:43612024-43612625	*Nkx2-3*	chr7:101011752-101012353	*P2ry2*
chr19:47014397-47014998	*Ina*	chr8:14911362-14911963	*Arhgef10*
chr19:61228117-61228718	*Csf2ra*	chr8:72160899-72161500	*Rab8a*
chr2:91236845-91237446	*A330069E16Rik*	chr8:83607874-83608475	*Dnajb1*
chr2:104816395-104816996	*Qser1*	chr8:85492316-85492917	*Gpt2*
chr2:104849549-104850150	*Prrg4*	chr8:116801111-116801712	*4930563M20Rik*
chr2:105224041-105224642	*Them7*	chr9:35175686-35176287	*Dcps*
chr2:119029092-119029693	*Ccdc32*	chr9:44798915-44799516	*Tmem25*
chr2:119617997-119618598	*Nusap1*	chr9:66124585-66125186	*Snx1*
chr2:119618204-119618805	*Oip5*	chr9:90270468-90271069	*Tbc1d2b*
chr2:158257940-158258541	*Bpi*	chr9:95406969-95407570	*Chst2*
chr2:164562415-164563016	*Wfdc2*		

Notably, the distribution of H3K4me3 levels in sperm was insufficient to explain the differences between these CGIs and those that showed no gain in DNAme ($p=0.46$). However, relative to CGIs that remain unmethylated, PDA CGI promoters show a significantly greater decrease of H3K4me3 levels over the paternal allele in 2C embryos

(p=2.25E-5, **Figure 3.4c**). For example, *Tuba3a*, *Gdap2* and *Bpi* show PDA across 28, 58 and 7 CpGs proximal to their TSSs (+/- 300bp), respectively, coincident with loss of H3K4me3 on the paternal allele (**Figure 3.4d**), while the control gene *Ufc1* displays persistent DNA hypomethylation and H3K4me3 enrichment on the paternal genome both pre- and post-fertilization. Though I cannot discriminate between active demethylation of H3K4 and histone H3 turnover, these results reveal that de novo DNAme of CGIs on the paternal genome is accompanied by a reduction of H3K4me3. Surprisingly, only 11 of the PDA genes are methylated (≥20% DNAme) in MII oocytes (**Figure 3.5a**) and none gain DNAme on the maternal genome in 2C embryos. In contrast, maternally methylated imprinted genes, such as *Impact* and *Snurf*, show maternal allele-specific DNA hypermethylation in the ICM and paternal allele-specific enrichment of H3K4me3 throughout early embryonic development (**Figure 3.5b**), as expected. Taken together, these WGBS and ChIP-seq data reveal that a specific subset of CGI promoters are de novo DNA methylated exclusively on the paternal genome following fertilization, concomitant with reduced H3K4me3.

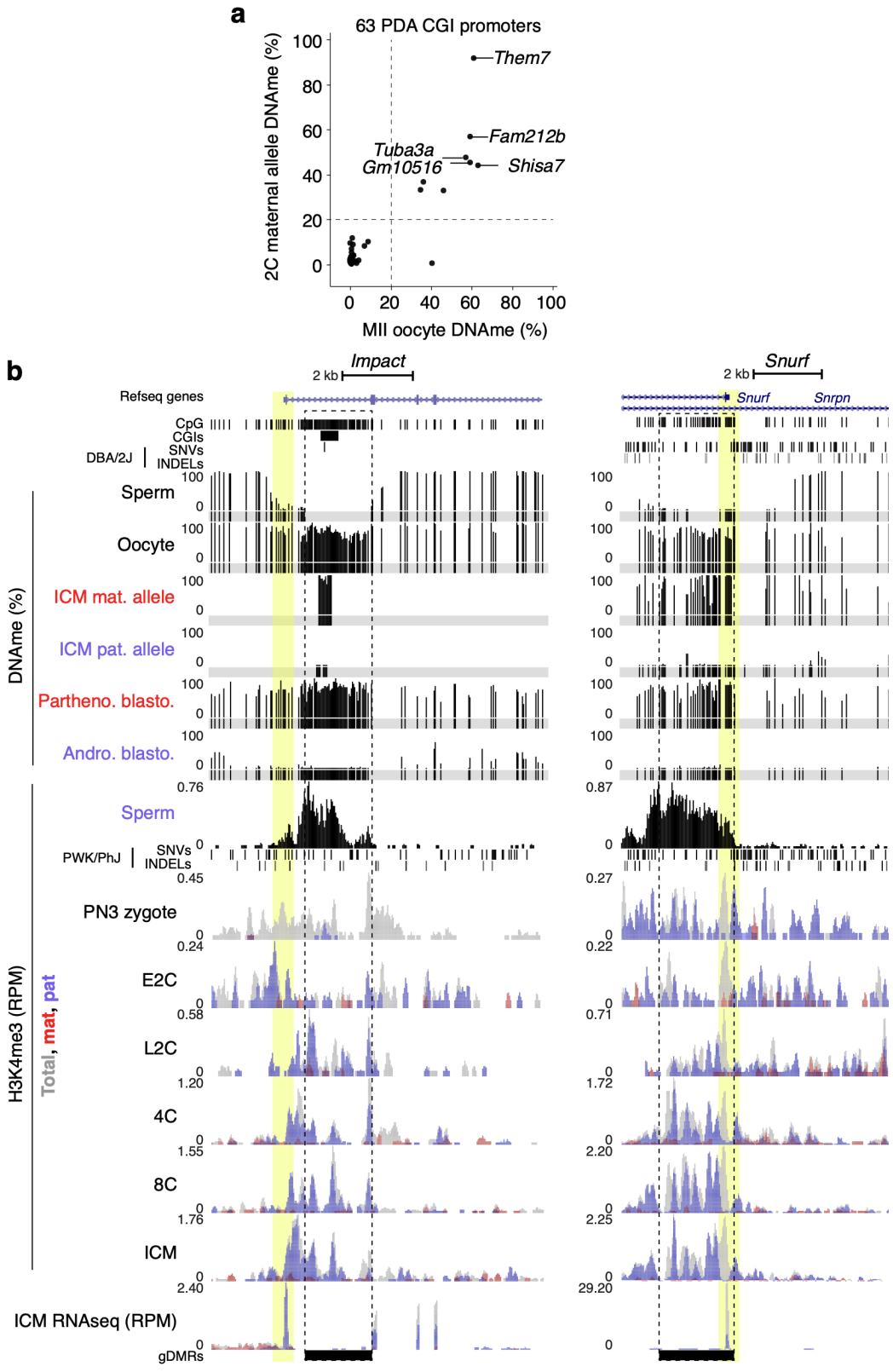
a
63 PDA CGI promoters
Them7
Fam212b
Tuba3a
Gm10516
Shisa7
2C maternal allele DNAme (%)
MII oocyte DNAme (%)
b
Impact
Snurf
2 kb
Refseq genes
Snurf
Snrpn
CpG CGIs
DBA/2J
SNVs
INDELs
Sperm
Oocyte
ICM mat. allele
ICM pat. allele
Partheno. blasto.
Andro. blasto.
DNAme (%)
PWK/PhJ
PN3 zygote
E2C
L2C
4C
8C
ICM
H3K4me3 (RPM)
Total, mat, pat
ICM RNAseq (RPM)
gDMRs

Figure 3.5 The maternal genome does not gain DNAme following fertilization. **(a)** Maternal allele DNAme levels over CGI promoters that show PDA (both datasets include polar bodies, (PB)). **(b)** UCSC genome browser screenshots of the *Impact* and *Snurf* paternally expressed imprinted genes as presented in **Figure 3.4d**. Promoters are highlighted in yellow and gametic DMRs by a dashed box. Mat.: maternal, Pat.: paternal, Partheno.: parthenogenetic, Andro.: androgenetic, Blasto.: blastocyst. PN3: pronuclear stage 3, RPM: reads per million aligned.

3.3.2 DNAme at many PDA loci is maintained through the blastocyst stage

To determine whether DNAme at PDA sites persists through the wave of global DNAme erasure, I scored paternal DNAme levels at these loci using WGBS data from F1 hybrid ICM cells (Wang et al. 2014). Relative to TSSs that remain hypomethylated following fertilization, genic promoter regions that show PDA retain higher DNAme on the paternal allele in ICM cells (p=4.97E-6, **Figure 3.6a-b**), albeit at lower levels than observed in 2C embryos. In contrast, 19 CGI promoters that show PDA, including *Gdap2/Wdr3* (**Figure 3.6c)**, are hypomethylated (mean <5%) in blastocysts. Thus, while DNAme at a subset of PDA genes is transient, other CGIs showing PDA either resist DNA demethylation or are reiteratively de novo methylated in early embryonic development.

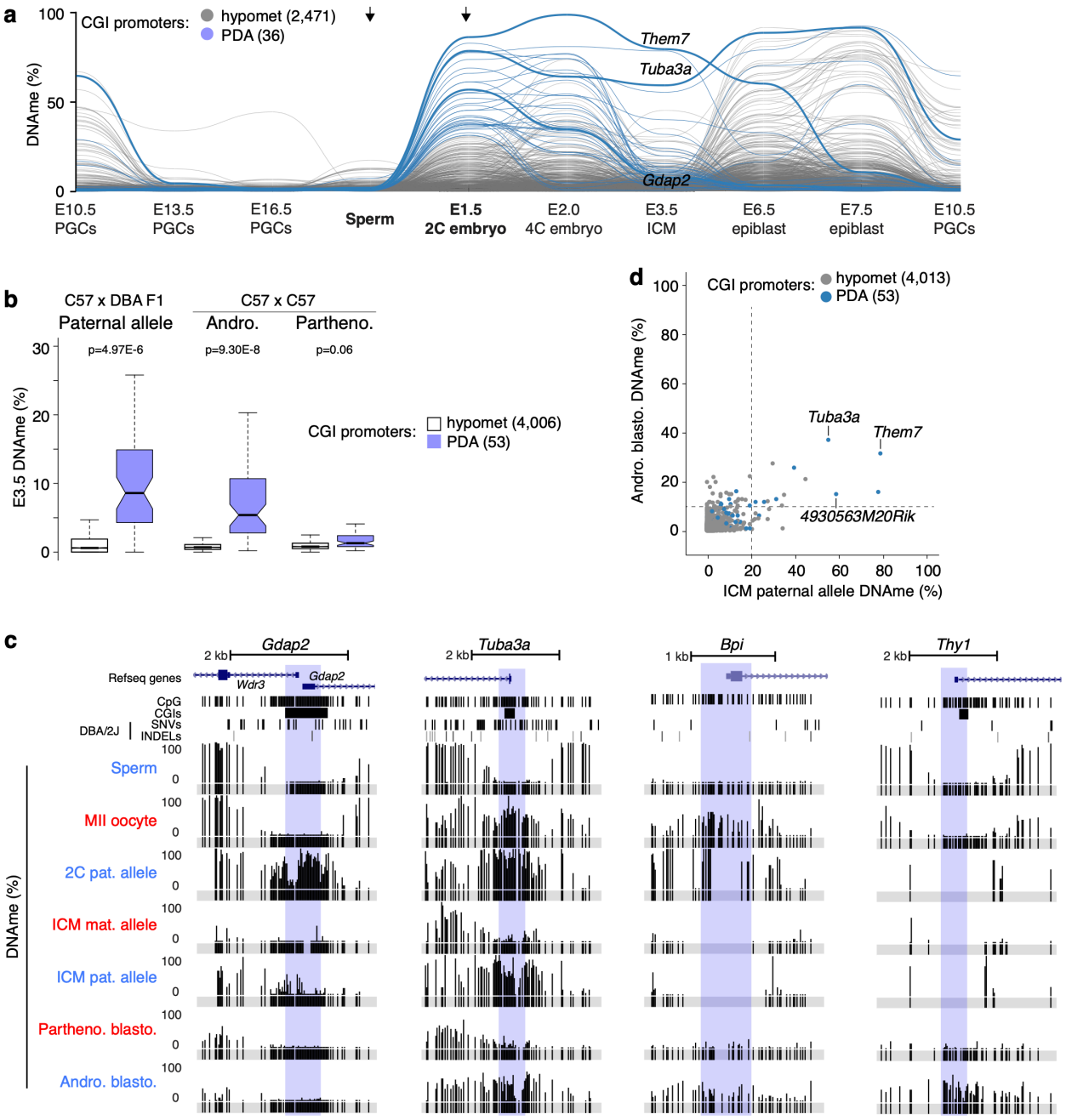

Figure 3.6 Paternal DNAme levels at many PDA sites are maintained in normal and androgenetic blastocysts.

(a) Parallel coordinate plot as in **Figure 3.2** showing DNAme dynamics at CGI promoters. Those that remain hypomethylated (n=2,471) or gain paternal DNAme (n=36) following fertilization are shown. **(b)** Distribution of CGI promoter DNAme levels

in E3.5 C57xDBA F1 ICM, C57 androgenetic blastocysts and C57 parthenogenetic blastocysts. Outliers not shown. T-tests of two samples assuming unequal variances were performed, and the p-values are indicated in the graph. **(c)** Screenshots of the *Gdap2/Wdr3*, *Tuba3a*, *Bpi* and *Thy1* CGI promoters, presented as in **Figure 3.4d**. **(d)** 2D scatterplot showing paternal DNAme levels over individual CGI promoters in normal ICM cells and androgenetic blastocysts.

To confirm the persistence of paternal allele-specific DNAme and expand upon the number of loci at which PDA is likely occurring in the early embryo, we conducted WGBS on isogenic (C57BL/6NJcl) androgenetic blastocysts. Compared to CGI promoters that remain hypomethylated in 2C embryos, those that show PDA exhibited a significantly greater level of DNAme in such bipaternal blastocysts (p=9.30E-8, **Figure 3.6b**). Furthermore, the majority of CGI promoters that show persistence of DNAme (≥20%) on the paternal allele in F1 ICM, including 8 PDA genes, show ≥10% DNAme in androgenetic blastocysts, while the vast majority of CGIs hypomethylated in F1 2C embryos show <10% DNAme in these cells (**Figure 3.6d**). In contrast, analysis of our previously published WGBS data from parthenogenetic blastocysts (Brind'Amour et al. 2018), in which both genomes are maternally derived, reveals that DNAme remains low at all of the CGI promoters showing zygotic PDA (**Figure 3.6b**), with the exception of 6 of the 11 that are already hypermethylated in MII oocytes. In summary, loci showing PDA are de novo methylated in the early embryo exclusively when at least one paternal genome is present.

Given that androgenetic and parthenogenetic blastocysts are uniparental, I was able to extend our parental genome-specific DNAme analysis from 18 to ~90% of all annotated autosomal CGI promoters. As expected, maternally methylated imprinted CGI promoters, such as *Peg3* and *Zdbf2*, which are hypermethylated exclusively on the maternal allele in normal embryos, remain hypermethylated in parthenogenetic blastocysts (**Figure 3.7a)**. In contrast, the paternally methylated *Dlk1-Meg3* intergenic-gDMR (analyzed in Chapter 2) shows 56% DNAme in androgenetic and 1% DNAme in parthenogenetic blastocysts, as expected (data not shown in this TSS-centric analysis). However, only one CGI promoter (*Ccdc114*) shows a >20% DNAme gain in these cells relative to germinal vesicle (GV) oocytes. In contrast, 28 CGI promoters show such a gain in androgenetic blastocysts relative to sperm, confirming that the paternal genome is the preferred target for such post-fertilization de novo DNAme (**Figure 3.7b**).

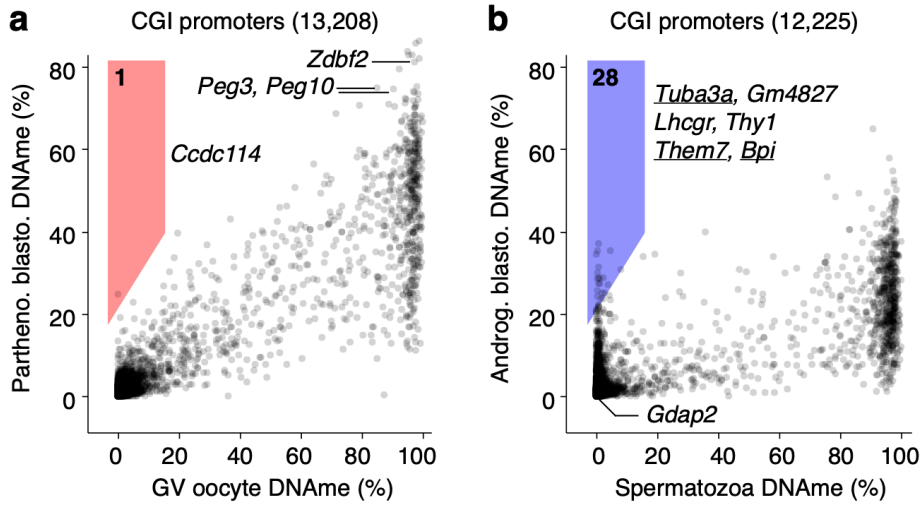

Figure 3.7 A subset of CGI promoters gain DNAme following fertilization specifically on the paternal genome.

(a) Parthenogenetic blastocysts versus GV oocyte (not including polar bodies; n=13,208) and **(b)** androgenetic blastocysts versus sperm (n=12,225). Promoters that gain >20% DNAme are highlighted and genes previously identified using our allele-specific DNAme analysis pipeline are underlined.

Importantly, 6 of these loci (*Tuba3a, Bpi, Them7, Shisa7, Syn3 and A230077H06Rik*) were also identified as PDA genes in our allele-specific analysis of F1 hybrid embryos, revealing that this phenomenon occurs independent of DBA/2J-specific variants and is thus a *bona fide* parent-of-origin effect at these loci. Of the remaining 22 CGI promoters, 5 fell into the "persistent hypomethylation" category, but 4 of these (*H1fnt, Dbx2, Tbx4* and *Prss39*) showed a gain in DNAme of 9-25% in normal 2C embryos, suggesting that while they did not meet our stringent >30% gain cutoff, these CGI promoters also likely gain DNAme following fertilization, albeit at lower levels. The remaining 17 genes that gain DNAme in androgenetic blastocysts (**Figure 3.7b)**, including *Thy1* (**Figure 3.6c**), do not harbor a genetic variant in their TSS and could therefore not be assessed in the F1 hybrid datasets. The generation of PBAT data from androgenetic blastocysts also enabled us to determine the paternal DNAme level status of the 10 PDA genes that could not be ascertained from F1 blastocyst WGBS data due to low sequencing depth in these regions. Relative to sperm, 3 of these loci, including *Bpi, Syn3 and Shisa7*a, showed a ≥20% gain of DNAme. As these promoters do not gain DNAme in parthenogenetic blastocysts, it is likely that the paternal allele of these genes is also methylated in normal blastocysts.

3.3.3 Relationship between histone PTMs and PDA

To determine whether promoter regions showing PDA share a common DNA motif that may render them susceptible to de novo DNAme I performed de novo DNA motif discovery using HOMER (Heinz et al. 2010) and MEME (Bailey et al. 2009). However, I did not detect any common DNA motif around the TSSs (+/- 300 bp) of genes that show PDA. Therefore, I focussed on the relationship between histone PTMs and paternal DNAme acquisition and/or subsequent DNAme maintenance at these sites. I analyzed H3K4me2 (Siklenka et al. 2015), H3K4me3 (Erkek et al. 2013; Zhang et al. 2016), H3K9me3 (Wang et al. 2018), H3K27me3 (Erkek et al. 2013; Zheng et al. 2016) and H3K36me3 (Xu et al. 2019) ChIP-seq data from sperm, oocytes, zygote (1C), early 2C (E2C), late 2C (L2C) and ICM cells using ChromHMM (Ernst and Kellis 2012). No single histone PTM or combination thereof in sperm predicted PDA at CGI promoters versus those that remain hypomethylated (**Figure 3.8a-b**). However, a striking enrichment of H3K9me3 was observed in the zygote at PDA sites (**Figure 3.8b**) and those that show persistent DNAme in the ICM show an even greater enrichment of H3K9me3 in the zygote and 2C embryo (**Figure 3.8c**). Indeed, a positive correlation between paternal 2C H3K9me3 levels and paternal ICM DNAme levels is observed at many CGI promoters (**Figure 3.8d**), including the PDA gene *Tuba3a* (**Figure 3.8d-e**). Thus, this PTM may protect regions showing PDA against loss of paternal DNAme in the preimplantation embryo, consistent with previous reports indicating that H3K9me3 plays a role in promoting maintenance of DNAme (Liu et al. 2014; Du et al. 2015).

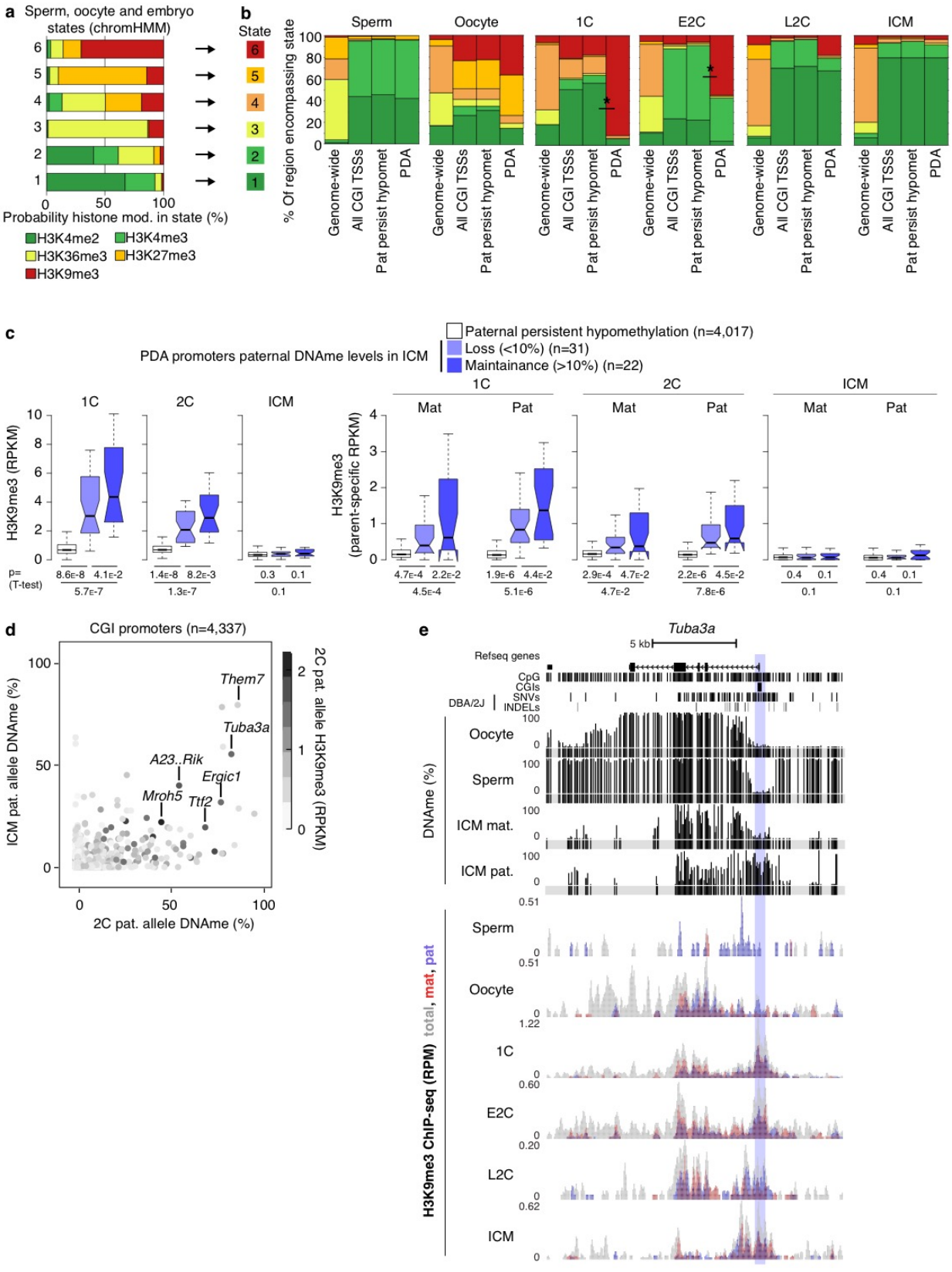
a
Sperm, oocyte and embryo states (chromHMM)
Probability histone mod. in state (%)
H3K4me2
H3K4me3
H3K36me3
H3K27me3
H3K9me3
b
State
% Of region encompassing state
Sperm
Oocyte
1C
E2C
L2C
ICM
Genome-wide
All CGI TSSs
Pat persist hypomet
PDA
c
PDA promoters paternal DNAme levels in ICM
Paternal persistent hypomethylation (n=4,017)
Loss (<10%) (n=31)
Maintainance (>10%) (n=22)
H3K9me3 (RPKM)
H3K9me3 (parent-specific RPKM)
Mat
Pat
p= (T-test)
d
CGI promoters (n=4,337)
ICM pat. allele DNAme (%)
2C pat. allele DNAme (%)
2C pat. allele H3K9me3 (RPKM)
Them7
Tuba3a
A23..Rik
Ergic1
Mroh5
Ttf2
e
Tuba3a
5 kb
Refseq genes
CpG CGIs
SNVs INDELs
DBA/2J
DNAme (%)
Oocyte
Sperm
ICM mat.
ICM pat.
H3K9me3 ChIP-seq (RPM) total, mat, pat
1C
E2C
L2C
ICM

Figure 3.8 Relationship between histone PTMs and PDA.

(a) Six distinct chromatin states were identified using ChromHMM and ChIP-seq data (H3K4me2, H3K4me3, H3K36me3, H3K27me3 and H3K9me3) from spermatozoa, MII oocytes, 1C, early 2C, late 2C and blastocyst-stage embryos. The probability that a histone modification is found within each state is shown. Based on these probabilities, a colour code was assigned to each state (right). **(b)** The relative enrichment of each chromatin state genome-wide and over genomic regions of interest, including all autosomal CGI promoters (n=13,342), those that show persistent paternal DNA hypomethylation following fertilization (n=4,315) and those that show PDA (n=63), is shown. **(c)** The distribution of total (left), and parent allele-specific H3K9me3 (right) levels over CGI promoters in 1C, E2C and blastocyst-stage embryos is shown. CGI promoters were categorized by paternal DNAme dynamics, including persistent hypomethylation following fertilization (n=4,015), PDA genes showing loss of paternal DNAme by the blastocyst stage (n=32), and PDA genes showing persistence of paternal DNAme in the blastocyst (n=22). **(d)** 2D scatterplot illustrating the association between H3K9me3 and DNAme levels on the paternal allele at the 2C stage and maintenance of DNAme on the paternal allele in ICM cells. **(e)** UCSC genome browser screenshot of the *Tuba3a* locus as in **Figure 3.4d**. H3K9me3 levels are illustrated as composite maternal (red) and paternal (blue) and total (grey) genomic tracks.

3.3.4 Many PDA target genes are expressed during spermatogenesis but silenced in the early embryo

To study the temporal expression profiles of genes with CGI promoters showing PDA, I analyzed existing RNA-seq datasets from multiple stages in spermatogenesis, oocytes and preimplantation development using k-means clustering (**Figure 3.9a**). Autosomal genes were binned into 5 clusters based on temporal expression levels, which I categorize as: constitutive high (High, n=4,022 e.g. *Actb*), constitutive intermediate (Int,

n=4,415 e.g. *Wiz*), spermatogenesis specific (sperm, n=4,418 e.g. *Bpi*), oocyte high, spermatogenesis low (oocyte, n=2,125 e.g. *Nlrp5*) and constitutive low (n=7,605 e.g. *Myod1*). Intersecting these expression clusters with the list of genes that show PDA (**Figure 3.9b**) reveals that PDA genes are enriched within the spermatogenesis category (40% of all PDA genes), relative to all autosomal genes (20%) or those that show persistent hypomethylation following fertilization (19%). This enrichment suggests that de novo DNAme following fertilization may play a role in silencing of male germline genes during early embryonic development.

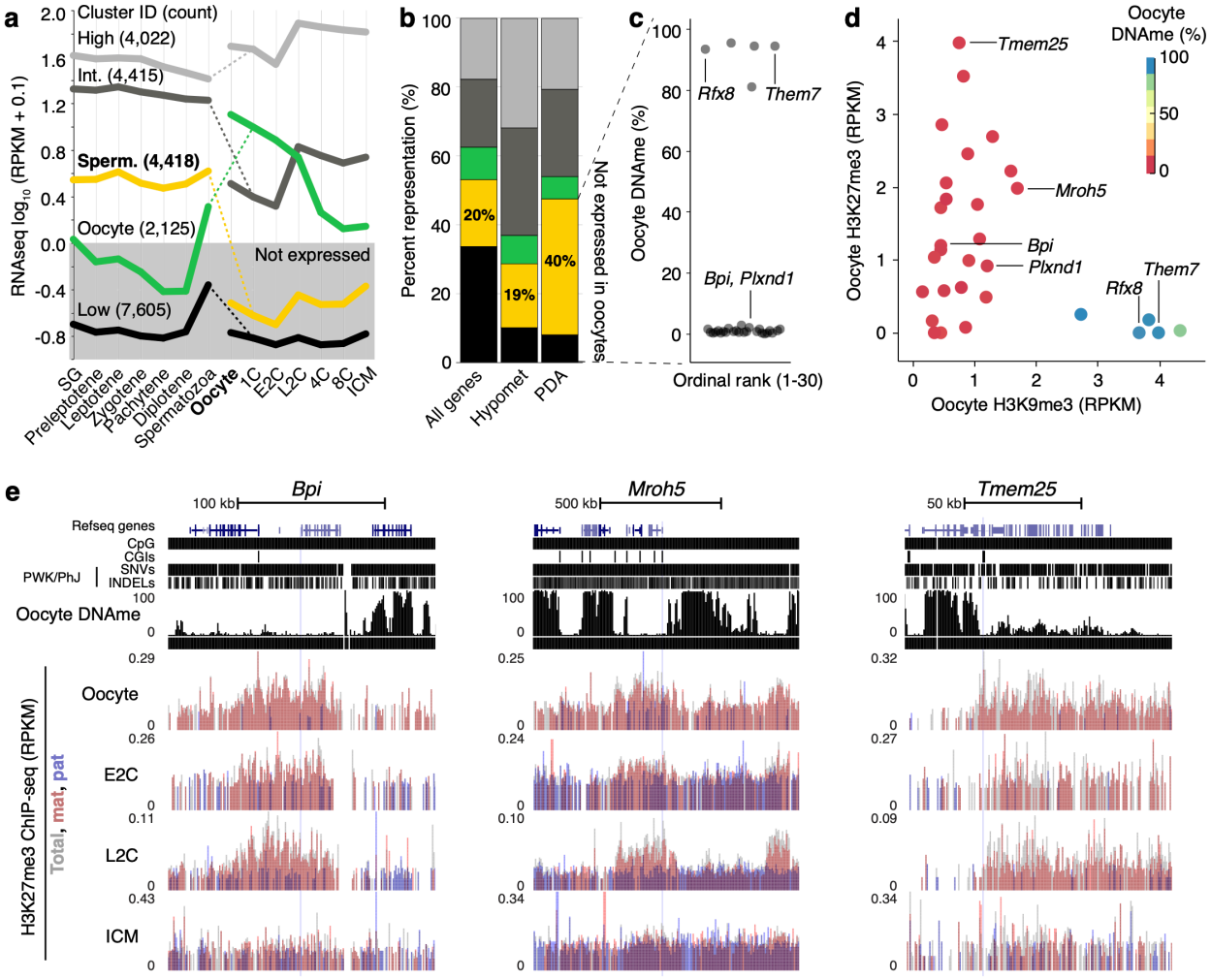

Figure 3.9 PDA genes not expressed in oocytes are enriched for repressive histone marks.

(a) Parallel coordinate plot showing the temporal expression pattern of all autosomal genes in the male germline, MII oocytes and the early embryo. SG: spermatogonia. Five temporal expression categories were calculated by k-means clustering of RNA-seq data, and the average expression level within each cluster is plotted. High: constitutive high, Int.: constitutive intermediate, sperm: expressed exclusively in spermatogenesis, Oocyte: expressed in oocyte but low in spermatogenesis, Low: constitutive low. The dashed line represents the union of spermatozoa and oocyte transcriptomes in the zygote. The threshold for scoring gene clusters as not expressed (average RPKM <1.0) is highlighted in grey. **(b)** Fraction of all genes (n=22,585), genes that remain

hypomethylated (hypomet, n=4,315) and PDA genes (n=63) falling in each cluster identified in **a**. **(c)** Mean DNAme levels in GV oocytes (data excludes polar bodies) over CGI promoters that show PDA and are not expressed during early embryogenesis. **(d)** Enrichment of H3K9me3 and H3K27me3 in MII oocytes over the same CGI promoters as in **c**, with heat map of DNAme levels in GV oocytes. **(e)** Browser screenshots of *Bpi*, *Mroh5* and *Tmem25* loci, including MII oocyte DNAme and H3K27me3 ChIP-seq tracks, presented as in **Figure 3.8e**.

Analysis of the expression levels of the 63 PDA genes in oocytes reveals that 33 and 30 are expressed and silent, respectively. Surprisingly, 25 of the latter are hypomethylated in oocytes (**Figure 3.9c**). To determine whether these genes harbor histone marks associated with transcriptional repression, I integrated previously published H3K27me3 and H3K9me3 ChIP-seq datasets derived from oocytes and F1 hybrid embryos (Zheng et al. 2016; Wang et al. 2018). Consistent with a role for H3K9me3 in promoting DNAme maintenance, all 5 hypermethylated PDA loci are enriched for this histone PTM in oocytes (**Figure 3.9d**). In contrast, of the 25 hypomethylated CGI promoters, 17 are embedded within H3K27me3-enriched (RPKM >0.9) domains (**Figure 3.9d**). The TSSs of *Bpi*, *Mroh5* and *Tmem25* for example, are all embedded within extended (50-200kb) H3K27me3 domains in oocytes and this mark persists at each of these loci until at least the L2C stage (**Figure 3.9e**). Notably, H3K27me3 was recently implicated in transcriptional silencing of genes subject to non-canonical maternal imprinting in the mouse (Inoue et al. 2017; Matoba et al. 2018). Taken together, these results reveal that de novo DNAme of the paternal allele and H3K27me3 established on the maternal allele in the oocyte may play complementary

roles in transcriptional repression of a subset of PDA genes during early embryonic development.

3.3.5 Maternal DNMT3A is required for PDA

DNMT3A and its co-factor DNMT3L are required for de novo DNAme in both the female (Shirane et al. 2013) and male germlines (Bourc'his and Bestor 2004; Kaneda et al. 2004; Kato et al. 2007; La Salle et al. 2007). To determine whether maternal DNMT3A is responsible for PDA, we crossed oocyte-specific homozygous *Dnmt3a* knock-out (matKO) C57BL/6J females with WT DBA/2J males and performed WGBS on F1 hybrids at the early-mid 2C stage (**Figure 3.10a**). As expected, global maternal DNAme levels in 2C matKO embryos mirror those in *Dnmt3a* matKO GV oocytes (**Figure 3.10b**). Furthermore, while the paternal gDMRs *H19*, *Dlk1-Meg3* and *Rasgrf1* show no decrease, DNAme at maternal gDMRs is dramatically reduced in 2C matKO embryos relative to control (**Figure 3.10c**). Importantly, a positive correlation was observed when comparing paternal DNAme levels of CGI promoters with sufficient allele-specific coverage in our control dataset (mean sequencing coverage 5X) with WGBS data from Wang *et al*. (Wang et al. 2014), with the majority of PDA genes showing relatively high DNAme on the paternal allele in both (**Figure 3.10d**). In the absence of maternal DNMT3A however, DNAme is lost on the paternal allele of all 23 CGI promoters showing PDA for which allelic methylation can be deduced (**Figure 3.10e**). At the *Tuba3a*, *Gdap2* and *Snx1* promoters for example, while total (allele-agnostic) and paternal (allele-specific) DNAme levels are concordant between wild-type (WT) 2C

datasets, PDA is lost in matKO embryos across all CpGs in the promoter region (**Figure 3.10f**). While the remaining 40 PDA loci lack allele-specific coverage in our matKO samples, analysis of total DNAme levels reveals that 33 are clearly hypomethylated (<4%), as shown for *Bpi* (**Figure 3.10f**). Taken together, these results demonstrate that DNAme acquisition on the paternal allele immediately following fertilization is dependent upon maternal DNMT3A.

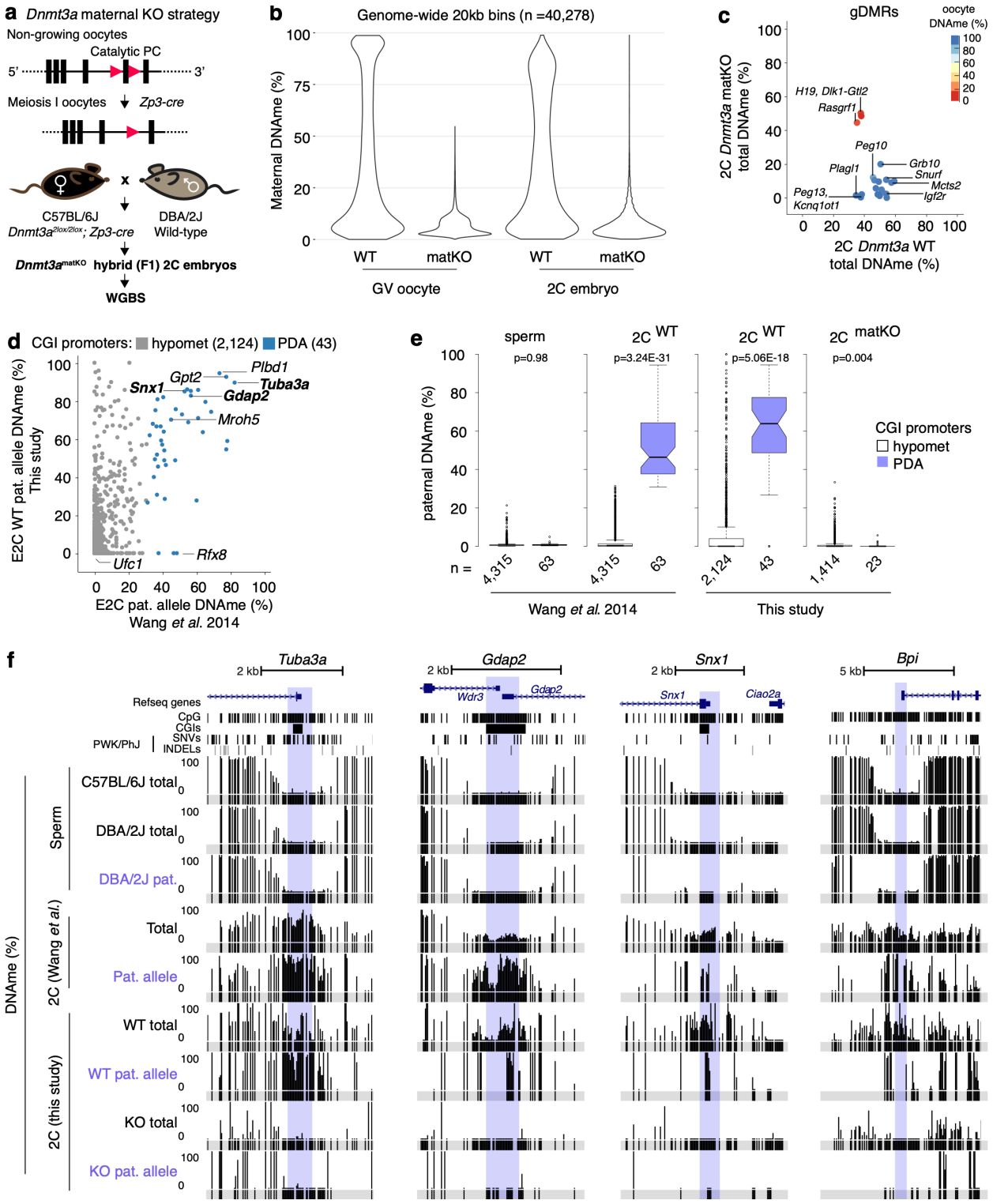

Figure 3.10 Paternal DNAme acquisition is mediated by maternal DNMT3A.

(a) Maternal *Dnmt3a* knock-out (KO) strategy, including *lox*P sites (red triangles) and deleted exon 18. Oocytes from C57BL/6J matKO, mice were *in vitro* fertilized using sperm from wild-type (WT) DBA/2J males and PBAT data was generated from two independent biological replicates. PC: catalytic proline-cysteine dipeptide of the methyltransferase domain. **(b)** Distribution of maternal DNAme levels over 20kbp genomic bins in WT and *Dnmt3a* matKO GV oocytes and 2 cell embryos. 20kb bins overlapping >3 informative CpGs (covered by 5X reads and separated by >1 sequencing read length) in all 4 datasets are reported (n=40,278). GVO data from (Shirane et al. 2013). **(c)** Scatterplot showing the average total (allele-agnostic) DNAme levels over gDMRs overlapping >1 informative CpG in *Dnmt3a* WT and matKO 2C embryos, with heat map of DNAme levels in GV oocytes. **(d)** CGI promoter paternal DNAme level correlation between wild-type 2C datasets generated in this study and Wang *et al*. Paternal DNAme levels over CGI promoters were reported if they were covered by >4 informative CpGs (covered by 1X read and separated by >1 sequencing read length) in our 2C WGBS datasets. **(e)** Paternal DNAme level distribution over CGI promoters in sperm (left), normal 2C embryos (middle), and *Dnmt3a* matKO 2C embryos (right). The number of CGI promoters represented is indicated below each plot. **(f)** Screenshots of the *Tuba3a*, *Gdap2/Wdr3*, *Snx1* and *Bpi* loci as presented in **Figure 3.4d**. Total (allele-agnostic) and paternal allele-specific DNAme levels are shown. KO: *Dnmt3a* matKO.

3.3.6 Loss of PDA results in ectopic expression from the paternal allele

Hypermethylation of CGI promoters is associated with transcriptional silencing (Weber et al. 2007; Deaton and Bird 2011). To determine whether DNAme of genes showing PDA impacts their expression specifically from the paternal allele, we conducted strand-specific RNA-seq on early-mid 2C as well as 4C and blastocyst-stage F1 matKO embryos (**Figure 3.11a**). Analysis of the splicing profile at the *Dnmt3a* locus revealed

efficient deletion of the targeted exon (**Figure 3.11b**). Transcriptome analysis revealed strong correlation between *Dnmt3a* matKO and WT 2C, 4C and blastocyst-stage embryos as well as previously published WT transcriptomes of the same developmental stages (**Figure 3.11c**), indicating that maternal depletion of DNMT3A does not disrupt global transcriptional programming. This is consistent with previous observations that *Dnmt3a* matKO embryos develop normally until E8.5 (Kaneda et al. 2004; 2010). I next determined whether genes known to be regulated by maternally established genomic imprints show loss of imprinting (LOI) in *Dnmt3a* matKO embryos. Indeed, an increase in expression from the maternal allele of at least 2-fold was observed in 2C, 4C or ICM for several *bona fide* imprinted genes, including *Mcts2*, *Plagl1*, *Snurf*, *Peg10* & *Zdbf2* (**Figure 3.11d**), consistent with their loss of DNAme (**Figure 3.10c**). Taken together, these results validate our *Dnmt3a* matKO model, and demonstrate that, in the absence of DNAme, some maternally methylated imprinted genes may be regulated by additional repressive factors such as H3K27me3.

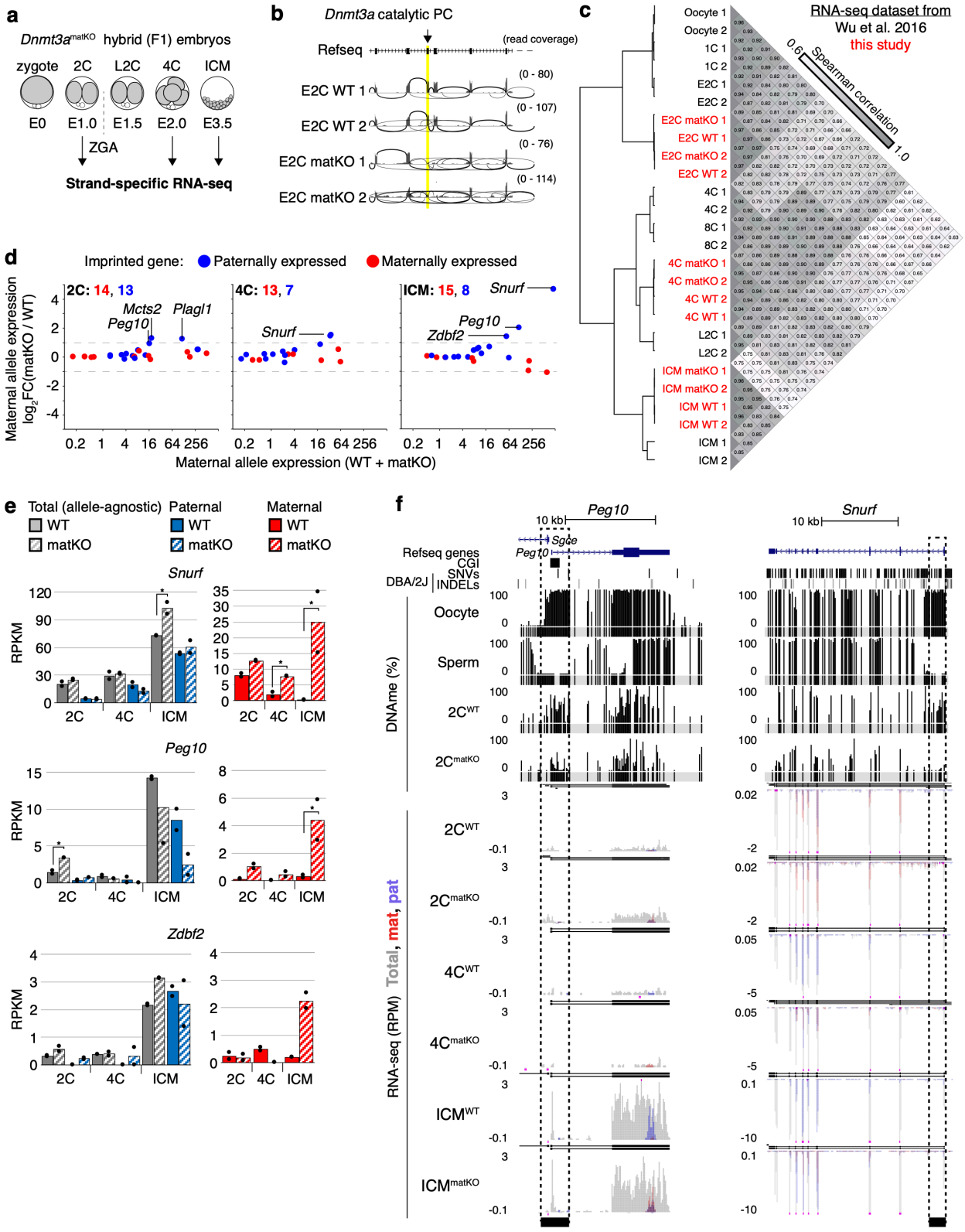
a
Dnmt3a^matKO hybrid (F1) embryos
zygote 2C L2C 4C ICM
E0 E1.0 E1.5 E2.0 E3.5
ZGA
Strand-specific RNA-seq
b
Dnmt3a catalytic PC
Refseq
(read coverage)
E2C WT 1
E2C WT 2
E2C matKO 1
E2C matKO 2
(0 - 80)
(0 - 107)
(0 - 76)
(0 - 114)
c
RNA-seq dataset from
Wu et al. 2016
this study
Spearman correlation
0.6
1.0
Oocyte 1
Oocyte 2
1C 1
1C 2
E2C 1
E2C 2
E2C matKO 1
E2C WT 1
E2C matKO 2
E2C WT 2
4C 1
4C 2
8C 1
8C 2
4C matKO 1
4C matKO 2
4C WT 2
4C WT 1
L2C 1
L2C 2
ICM matKO 1
ICM matKO 2
ICM WT 1
ICM WT 2
ICM 1
ICM 2
d
Imprinted gene: Paternally expressed Maternally expressed
2C: 14, 13
4C: 13, 7
ICM: 15, 8
Mcts2
Peg10
Plagl1
Snurf
Peg10
Zdbf2
Maternal allele expression log2FC(matKO / WT)
Maternal allele expression (WT + matKO)
e
Total (allele-agnostic) Paternal Maternal
WT matKO
Snurf
Peg10
Zdbf2
RPKM
2C 4C ICM
f
Peg10
Snurf
10 kb
Sgce
Peg10
Refseq genes
CGI
DBA/2J SNVs INDELs
DNAme (%)
Oocyte
Sperm
2C^WT
2C^matKO
RNA-seq (RPM) Total, mat, pat
2C^WT
2C^matKO
4C^WT
4C^matKO
ICM^WT
ICM^matKO

Figure 3.11 Transcriptomic validation of maternal *Dnmt3a* KO embryos.
(a) RNA-seq libraries were generated in biological duplicates for wild-type and *Dnmt3a* matKO early F1 2C, 4C and ICM embryos. **(b)** 2 cell strand-specific RNA-seq sashimi plot illustrating deletion of the catalytic exon of *Dnmt3a*. **(c)** Spearman correlation of autosomal gene expression (log2(RPKM+1) in datasets mined from (Wu et al. 2016) (black) and generated in this study (red, bold). **(d)** 2D scatterplots showing change in expression on the maternal allele of known imprinted genes. Data points are coloured by whether they are normally paternally (blue) or maternally (red) expressed in somatic cells. **(e)** Bar plots illustrating differential expression of select imprinted genes from both alleles (total, grey), the paternal genome (blue) and the maternal genome (red) in wild-type (filled bars) and *Dnmt3a* matKO (striped bars) embryos. Each bar represents the mean expression value of two biological replicates (dots) in RPKM. Statistically significant differences (Benjamini-Hochberg adjusted P value ≤0.01) are indicated by an asterisk. **(f)** UCSC genome browser screenshots of the paternally expressed imprinted genes *Peg10* and *Snurf* illustrating loss of imprinting. Presented as in **Figure 3.4d**. Gametic DMRs are indicated by a dashed box and strand-specific RNA-seq data is represented as a composite track of biological duplicates containing total (allele-agnostic, grey), maternal (red) and paternal (blue) genomic tracks. *De novo* assembly of transcripts is included above each RNAseq dataset.

Of the genes for which I could infer paternal CGI promoter DNAme dynamics (**Figure 3.4a**), 17, 19 and 16 are upregulated ≥2-fold from the paternal allele in 2C, 4C and ICM embryos, respectively (**Figure 3.12a-b**). Applying DESeq2 (Love et al. 2014) to measure genes upregulated from the paternal allele (Wald Chi-squared test, Benjamini-Hochberg adjusted P-value ≤0.01) yields only 2 significantly upregulated genes in 2C matKO embryos, *Gdap2* and *Wdr3*, which are PDA genes with overlapping

promoter regions. Further, of the 10 significantly upregulated genes in 4C matKO embryos, 3, *Gdap2*, *Plbd1* and *Snx1*, are PDA genes. In contrast, none of the 9 genes upregulated in ICM are PDA genes and thus likely represent either loci that are methylated post-fertilization on distal regulatory elements or indirect effects of maternal DNMT3A loss. Indeed, 4 of these genes are also significantly upregulated from the maternal allele. Importantly, transcription of all upregulated PDA genes initiates from the aberrantly hypomethylated CGI promoter, excluding alternative promoter usage as an explanation for the increased expression from the paternal allele (**Figure 3.12 b-c**).

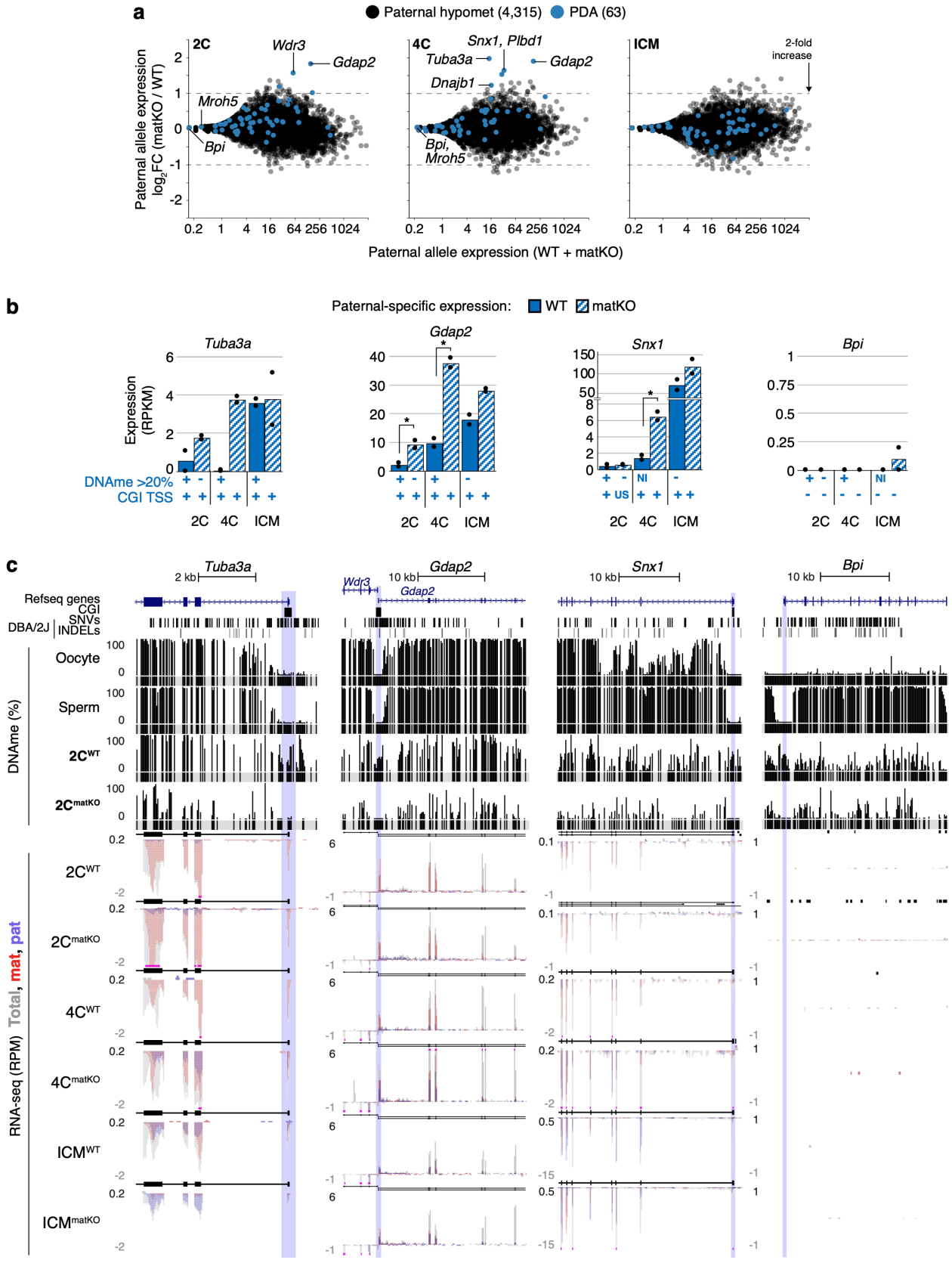
a
Paternal hypomet (4,315)
PDA (63)
2C
4C
ICM
Wdr3
Gdap2
Mroh5
Bpi
Snx1, Plbd1
Tuba3a
Dnajb1
2-fold increase
Paternal allele expression log2FC (matKO / WT)
Paternal allele expression (WT + matKO)
b
Paternal-specific expression: WT matKO
Tuba3a
Gdap2
Snx1
Bpi
Expression (RPKM)
DNAme >20%
CGI TSS
c
Refseq genes
CGI
DBA/2J SNVs INDELs
DNAme (%)
Oocyte
Sperm
2C^WT
2C^matKO
RNA-seq (RPM) Total, mat, pat
4C^WT
4C^matKO
ICM^WT
ICM^matKO

Figure 3.12 Impact of maternal DNMT3A deletion on expression from the paternal allele.

(a) 2D scatterplots showing average versus differential paternal-allele expression for 2C and 4C embryos as well as ICM cells. Genes are coloured by paternal DNAme dynamics following fertilization. **(b)** Bar charts illustrating differential expression of select genes from the paternal genome in WT and matKO embryos, as presented in **Figure 3.11e**. Maintenance of paternal DNAme levels (>20%) over the CGI promoter and whether transcription of each gene initiates from the CGI promoter is indicated below each bar. NI: no information, US: upstream. **(c)** Screenshots of *Tuba3a, Gdap2* and *Snx1* as presented in **Figure 3.11f**.

Intriguingly, none of the PDA genes categorized as spermatogenesis-specific (**Figure 3.9a**) including *Bpi*, showed increased transcription from the paternal allele in matKO embryos, which may reflect the absence of the relevant transcriptional activators at these early stages. Furthermore, no upregulation of PDA genes was observed from the maternal allele, consistent with DNAme-independent silencing at this stage (**Figure 3.9d**), or when total (allele-agnostic) transcript levels were analyzed (**Figure 3.13a-c**). Taken together, these results reveal that in the absence of maternal DNMT3A, a subset of genes normally de novo methylated on the paternal genome show transient ectopic expression.

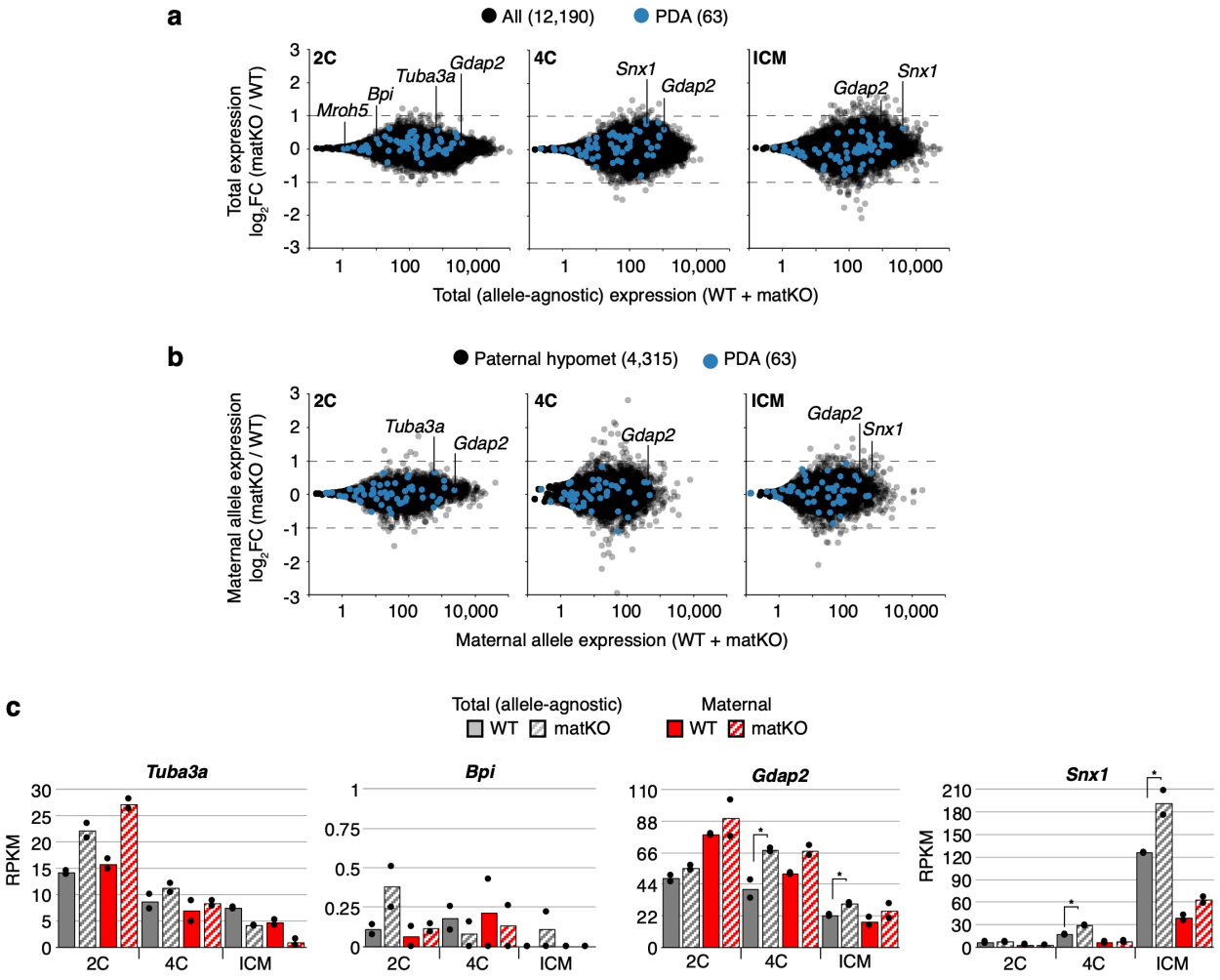

Figure 3.13 Allele-agnostic and maternal-allele analysis of the change in CGI promoter gene expression in Dnmt3a matKO embryos.

(a) 2D scatterplots showing average (x-axis) and differential (y-axis) total (allele-agnostic) expression between wild-type WT and matKO 2C, 4C embryos and ICM cells. Genes that show PDA are coloured in blue (n=63). **(b)** 2D scatterplots showing average (x-axis) and differential (y-axis) maternal-allele expression between wild-type WT and matKO 2C, 4C embryos and ICM cells. Genes are coloured by paternal DNAme dynamics immediately following fertilization (black; paternal hypomet, n=4,315, blue; PDA, n=63). The outlier gene *Rps3a1* is omitted **(c)** Bar charts illustrating differential expression of select genes from both alleles (grey) and the maternal genome (red) as presented in **Figure 3.11e**.

Due to the density of naturally occurring polymorphisms and depth of WGBS sequencing coverage, paternal DNAme dynamics could be measured in our earlier F1 hybrid analysis over only 4,434 of the 12,253 filtered autosomal CGI promoters (**Figure 3.4a**). Of the additional 2,903 autosomal genes with CGI promoters that include at least 1 exonic genetic variant between parental strains and are expressed from the paternal allele in our RNA-seq data, 5 were significantly upregulated from the paternal genome in matKO 4C embryos, including *Npm3*, *Eif3b*, *2610524H06Rik*, *Crk* & *Pim1* (**Figure 3.14a-b**).

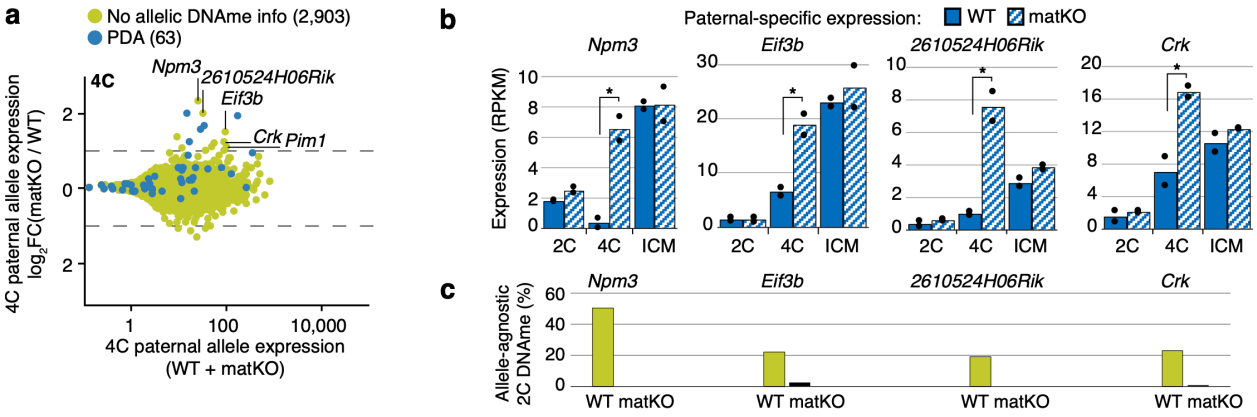

Figure 3.14 Additional CGI promoters showing ectopic expression from the paternal allele in *Dnmt3a* matKO 4C embryos.

(a) 2D scatterplot showing average and differential paternal-allele expression in WT versus matKO 4C embryos. Genes for which allelic DNAme levels could not be inferred over the CGI promoter due to a lack of genetic variants and/or insufficient sequencing depth are shown (n=2,903) along with genes showing PDA (n=63). **(b)** Bar charts illustrating differential expression of *Npm3, Eif3b*, *2610524H06Rik* and *Crk* from the paternal genome in wild-type and *Dnmt3a* matKO embryos as in **Figure 3.11e (c)** Total

(allele-agnostic) DNAme levels in WT and Dnmt3a *matKO* 2C embryos of the CGI promoters of genes shown in **b**.

Importantly, none of these genes show a change in expression from the maternal allele (**Figure 3.15a-b**). Furthermore, as for *bona fide* PDA loci, these genes are not upregulated in matKO ICM. Analysis of DNAme levels agnostic to parental allele clearly shows that these loci are methylated in WT but not in matKO 2C embryos, indicating that these CGI promoters are direct targets of maternal DNMT3A (**Figure 3.14c & Figure 3.15b**).

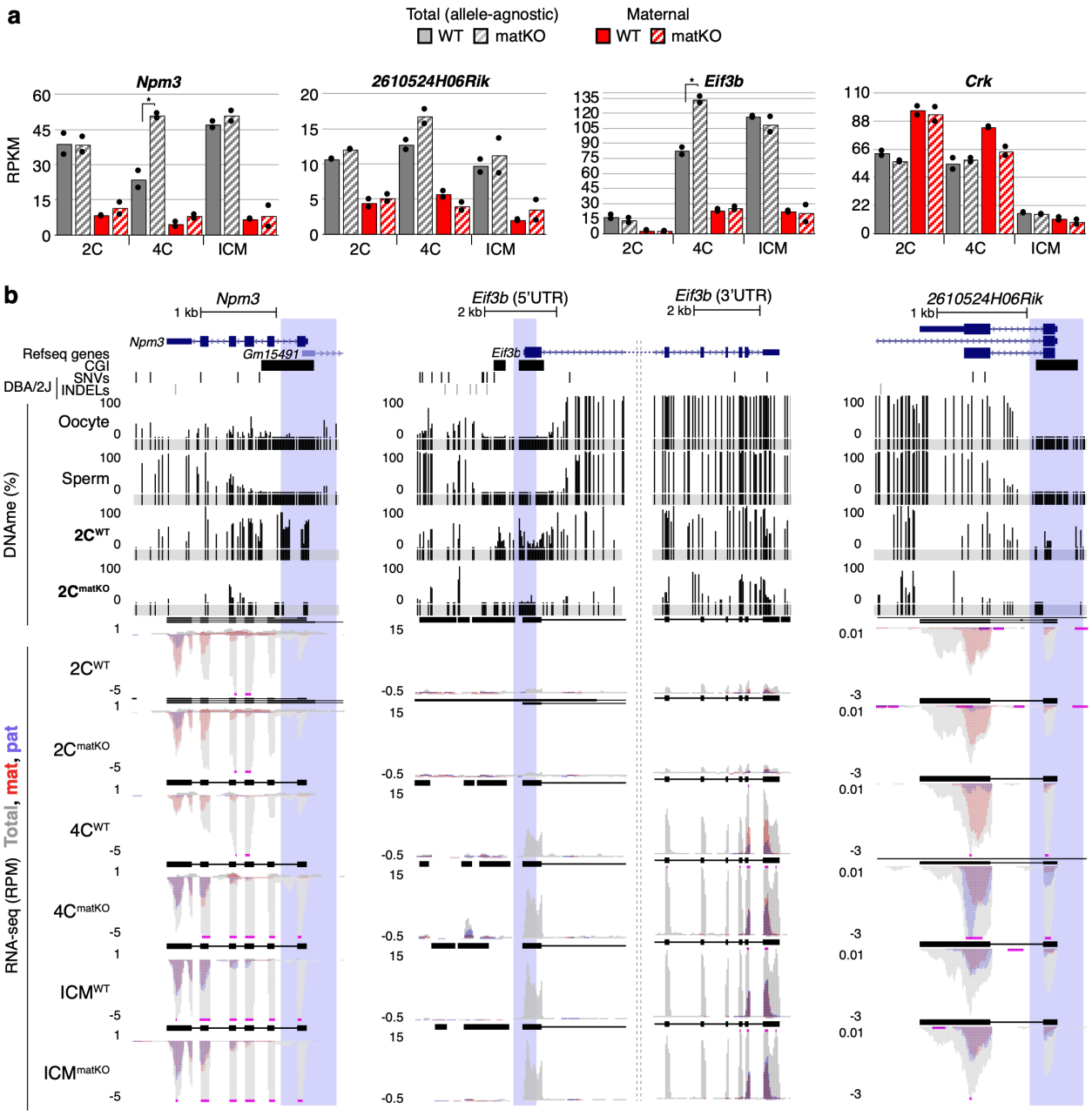

Figure 3.15 Additional CGI promoters that are activated in *Dnmt3a* matKO embryos show allele-agnostic DNAme gain at their CGI promoters in 2C embryos. **(a)** Bar charts illustrating differential expression of select genes as presented in **Figure 3.11e**. **(b)** Screenshots of *Npm3, Eif3b* and *2610524H06Rik*, as in **Figure 3.11f.**

Thus, in addition to the PDA genes described above, *Npm3*, *Eif3b*, *2610524H06Rik*, *Crk* & *Pim1* are likely de novo methylated on the paternal allele in 2C embryos, and repressed by this mark in 4C embryos.

3.4 Discussion

In this study I exploited allele-specific analyses and uniparental embryos to identify the genomic regions subject to de novo DNAme during early embryogenesis. Using this approach, ~2% of the mappable genome including at least 63 CGI promoters were identified as de novo methylated on the paternal genome by the 2C stage. WGBS analysis of androgenetic blastocysts revealed 86 CGI promoters showing a ≥10% gain in DNAme relative to sperm, 15 of which overlap with the PDA genes identified in normal 2C embryos. These observations are particularly surprising, given that the hypermethylated sperm genome is globally demethylated in the mouse zygote. Indeed, following polyspermic fertilization, where methylation of the maternal genome remains relatively constant, up to five paternal genomes are demethylated in the zygote (Santos et al. 2002). While TET3 was initially believed to be responsible for the replication-independent loss of paternal DNAme in the zygote (Iqbal et al. 2011; Shen et al. 2014; Peat et al. 2014; Gu et al. 2011b; Guo et al. 2014), it was recently shown that this methylcytosine dioxygenase likely acts specifically at regions on the paternal genome that are de novo methylated (Amouroux et al. 2016). Although zygotic de novo DNAme of the maternal genome has been reported for the TKZ751 transgene and *Igf2* (Oswald et al. 2000), ETn retroelements (Wossidlo et al. 2010) and H3K9me2-enriched regions

(Au Yeung et al. 2019), to our knowledge, this is the first report to characterize the specific regions of the paternal genome subject to de novo DNA methylation in the mouse zygote, supporting the findings of Amouroux et al.

To measure parental-epigenome dynamics in the early embryo, previous studies relied on IF-based assays, which are inherently of low resolution and therefore uninformative with respect to specific genomic loci (Mayer et al. 2000; Santos et al. 2002; Gu et al. 2011b). Further, while such studies have revealed large-scale epigenomic dynamics in the early embryo, dissecting the interplay between local DNAme, histone PTMs and transcription is not possible using IF. Our integrated allele-specific analysis reveals that loss of H3K4me3 and PDA occur shortly following fertilization, consistent with the observation that DNMT3A recognizes the unmethylated state of H3K4 (Ooi et al. 2007; Otani et al. 2009). Notably, persistence of H3K4 methylation on the paternal genome was previously reported to play an important role in gene regulation in 2C embryos (Siklenka et al. 2015). Our results reveal that at a distinct subset of loci, loss of H3K4 methylation may be a pre-requisite for de novo DNAme, and in turn transcriptional silencing.

How DNMT3A is targeted to PDA loci remains to be determined. While I did not uncover any common DNA motifs, it is possible that multiple different DNA binding factors bind to such regions to promote PDA. Interestingly, all but 3 PDA sites concomitantly gain H3K9me3 (RPKM $\geq$1) in the zygote, raising the possibility that this

mark may promote and/or maintain acquired DNAme. As KRAB-ZFPs have been shown to promote H3K9me3 deposition and potentially de novo DNAme at gDMRs (Li et al. 2008; Strogantsev et al. 2015; Quenneville et al. 2011), repetitive elements (Jacobs et al. 2014) and genic promoters (Yang et al. 2017), binding of yet to be characterized KRAB-ZFPs, in complex with TRIM28, may be responsible for sequence-specific targeting of DNMT3A to regions showing PDA. Alternatively, previous studies revealed that a specific set of CpG-rich germline gene promoters harbor E2F6 and E-box motifs that promote binding of the non-canonical PRC1 complex PRC1.6 and de novo DNAme in post-implantation embryos (Velasco et al. 2010; Endoh et al. 2017). Furthermore, such de novo DNAme occurs in conjunction with H3K9me3 acquisition (Auclair et al. 2016; Tatsumi et al. 2018). However, de novo DNAme at these germline gene promoters is DNMT3B-dependent and the E2F6 motif is present in the promoter region of only 4 of the loci showing PDA. Additional studies will be required to determine whether specific transcription factors and/or chromatin features are required for de novo DNA methylation of the paternal genome.

Dnmt3a maternal KO embryos die during post-implantation development around E8.5 (Kaneda et al. 2004; 2010). While it is tempting to speculate that aberrant expression of PDA genes from the paternal allele plays a role, our temporal analysis reveals that this is a transient phenomenon, with mRNA levels of these genes indistinguishable from wild type by the blastocyst stage. Alternatively, aberrant expression of imprinted genes as a consequence of failure of de novo DNAme in the

oocyte may be responsible for such embryonic lethality (Bourc'his et al. 2001; Kono et al. 1996). However, I find that several maternally methylated imprinted alleles are already expressed from the hypomethylated maternal allele in *Dnmt3a* matKO ICM cells. Thus, as for genes showing PDA, aberrant expression of at least a subset of maternally methylated imprinted genes occurs well before the gross phenotypic effects are manifest.

3.5 Conclusion

In summary, this study reveals that maternal DNMT3A is required for de novo DNAme of specific regions on the paternal genome immediately following fertilization, and in turn silencing of a subset of methylated genes on the paternal allele in early mouse embryonic development. Whether maternal DNMT3A methylates the paternal genome following fertilization in other mammals remains to be determined.

Chapter 4: Discussion

4.1 DNAme in mammals plays an instructive role in epigenetic gene regulation

As the functional relevance of genomic imprinting was emerging from nuclear transfer experiments (see Chapter 1), a seminal paper by Holliday and Pugh envisioned a distinct function for DNAme in vertebrates from its role in viral defense described in bacteria (Coulondre et al. 1978; Arber et al. 1963). They based their theory on the observation that every cell within a given multicellular organism contains the same DNA sequence, yet only a fraction of genes are transcribed in any given cell (Bird 2002; Lee and Young 2013). As cells are characterised by the complement of genes that are expressed, and, perhaps more importantly, the genes not expressed (Surani 2001), Holliday and Pugh reasoned that a system where information set upon genes that can reversibly repress transcription without altering the underling DNA sequence likely exists (Holliday and Pugh 1975). They postulated that reversible covalent modifications to DNA such as 5mC at CpGs could be the molecular basis of such an epigenetic system that regulates the spatial and temporal expression of genes (Holliday and Pugh 1975). The fact that DNAme is conserved in almost all vertebrates despite its inherent mutability strongly suggests it plays an important role. Whether that role is predominantly in regulating transcriptional initiation, as predicted, or other nuclear processes such as transcriptional elongation (Lorincz et al. 2004), chromosome stability (Xu et al. 1999) or TE suppression (Zemach et al. 2010), remains a central question in the epigenetics field and this thesis.

Holliday's theory was experimentally confirmed using genomic imprinting as a case study for DNAme-regulated gene expression (Li et al. 1993). These classical studies took advantage of reciprocal F1 hybrid crosses, which leverage known variants between two isogenic strains to enable allele-specific analysis of DNAme levels and gene expression. This remains the most elegant and powerful approach for studying epigenetic regulation of transcription as each gene has an internal control – namely the other allele (Pastinen 2010; Wang and Clark 2014). Unfortunately, as HTS technologies/genome-wide analyses were only available in the last decade, DNAme-regulated gene expression was generally tested on a gene-by-gene basis, usually in adult tissues. Regardless, these studies confirmed Holliday's theory by revealing that CpG rich promoters are reactivated upon genetically induced loss of DNA demethylation (Deaton and Bird 2011).

4.2 Importance of investigating DNAme in the developing mouse embryo

The propagation of successive generations involves two critical stages in the mammalian life cycle, both of which are associated with global chromatin remodeling: germline specification and fertilization. During germline specification, DNAme and histone PTMs are broadly removed and re-established in a sex-specific manner. Following fertilization, paternal chromatin is completely reset, including global DNAme loss, coincident with the transition from a maternal to an embryonic developmental programme (Eckersley-Maslin et al. 2018). This transition results in a totipotent cell,

which has the capacity to become every embryonic- and extra-embryonic cell type. Fully understanding this process, which begins with the fusion of two specialized, unipotent gametes and ends in the generation of a totipotent embryo, has immense clinical potential for assisted reproductive technology, cloning and stem cell therapy.

Largely due to a lack of input material and software capable of discriminating HTS reads in an allele-specific manner, few studies have measured the extent of chromatin remodelling on each parental allele following fertilization. Indeed, despite the well documented differences between sperm and oocyte chromatin, little is known about whether these differences are inherited in the early embryo and consequently, whether they regulate transcription of non-imprinted regions on the paternal or maternal genomes.

As such, fully understanding the extent of chromatin remodelling (including DNAme and histone PTMs) in the embryo, and how disruption of this process leads to ectopic gene expression, loss of cellular identity, and developmental arrest, is one of the major focuses of current biological research. Of course, other mechanisms for transcriptional regulation exist, such as chromosome looping (Murrell et al. 2004) and, more recently described, liquid phase separation (McSwiggen et al. 2019). Further, post-transcriptional regulation of transcript levels, translation itself and post-translational modifications to proteins all ultimately contribute to final gene product levels and activity.

4.3 Major findings of this thesis

Here, by developing and employing a novel bioinformatic pipeline for integrated analysis of HTS data with allele-specific resolution, I quantified the dynamics of DNAme on both parental genomes in the pre- and post-implantation mouse embryo. This strategy enabled me to confirm parent-specific DNAme and associated monoallelic transcription at imprinted loci. Surprisingly, fine-grained analysis of parental methylome dynamics uncovered novel candidate imprinted genes (Chapter 2) and antithetical paternal DNAme acquisition (PDA) at several dozen CGI promoters (Chapter 3). Interestingly, DNAme levels at a subset of such CGI promoters regulate paternal-specific gene expression in the early mouse embryo.

Chapter 2 outlines the development of a pipeline for the analysis of allele-specific WGBS, ChIP- and RNA-seq data (MEA). While other allelic pipelines use known SNVs to perform allele-specific read alignment, MEA also takes advantage of INDELs, which in mouse represents a significant 20% increase in the number of genetic variants used (Keane et al. 2011), which in turn expands the range of genomic loci that can be analyzed with allele-specific resolution. Additionally, MEA processes all types of HTS data, providing a universal toolkit for the analysis of the interplay between DNAme, histone PTMs and transcription with allelic resolution.

Applying MEA to existing data enabled me to uncover novel loci that show monoallelic gene expression in association with allele-specific DNA hypomethylation

and H3K4me3 enrichment in mouse (*Kiss1* and *Lpar6*) and human (*MIR4458HG*). These novel regions are termed “candidate” imprinted loci because only two aspects of genomic imprinting were measured: monoallelic expression in association with promoter hypomethylation of the same allele. In order to qualify as a canonical imprinted gene, these regions must also: 1) be differentially methylated between gametes 2) overlap ZFP57 or other KRAB-ZFP binding sites, 3) be protected from *de novo* DNAme during post-implantation development (Proudhon et al. 2012) and 4) lose DNAme during germ cell specification (Barlow and Bartolomei 2014). While the TSS of *Kiss1* showed sperm-specific hypermethylation, *Lpar6* is hypermethylated in both gametes; it remains to be determined how differential methylation of the *Lpar6* promoter is established following fertilization and whether these candidate imprinted regions overlap ZFP57 binding sites. Further contesting the imprinting status of *Kiss1*, subsequent analysis revealed the existence of two upstream TSSs (defined by the alternate gene model set GENCODE VM23), both of which are hypomethylated on both alleles in ICM cells and may therefore contribute to *Kiss1* expression independent of the TSS overlapping the novel DMR described here. In human samples, demonstrating imprinted gene expression is more difficult because reciprocal crosses are impossible, confounding parent-of-origin and genetic cis-effects (reviewed here (Wang and Clark 2014)). To determine whether these regions are truly imprinted in human, additional WGBS, RNA- and ChIP-seq must be conducted on individuals with different genotypes as genetic variation can lead to epigenetic differences *in cis* (Pastinen 2010). In other words, “maternal-specific transcription” and “haplotype 1-specific expression” can be delineated using additional,

non-related samples. Finally, whether parent-of-origin expression of such genes underlies the etiology of uniparental disomy or other mouse or human imprinting disorders remains to be determined.

MEA offers several incremental improvements over existing software including decreased reference alignment bias, increased range of allelic loci, and ease of use. As such, MEA enables researchers to extract the most out of expensive, labour intensive data generated from F1 hybrid embryos. Fortunately, my doctoral studies were conducted in a time when HTS datasets generated for a publication (including those derived from F1 hybrid embryos) have been made accessible to researchers worldwide through public servers (Gene Expression Omnibus, DNA Data Bank of Japan, etc.). While HTS data is usually generated to test a single hypothesis, the quality and quantity of information in each dataset can be harnessed to test other hypotheses. Further, consortia such as ENCODE and FANTOM generate HTS data without any specific hypothesis in mind in order to provide genomic resources to the scientific community. As such, the reanalysis of existing HTS data, whether it be using novel improved software like MEA or simply by formulating distinct hypotheses, is currently a quick and cost-effective method for producing genomic insights.

For example, in Chapter 3, I applied MEA to existing ultra-deep F1 hybrid WGBS data to trace maternal and paternal DNAme dynamics in the early mouse embryo. Initially, this project was focused on quantifying DNAme differences on each parental

genome throughout mouse embryogenesis in order to test MEA and complement old IF data (Mayer et al. 2000; Oswald et al. 2000). Then, emerging IF evidence suggested low levels of *de novo* DNAme occur on the zygotic paternal allele (Amouroux et al. 2016), yet this finding was highly controversial as it is in direct contrast with the current dogma of "global" loss of DNAme from the paternal genome at this stage. Fine-toothed analysis of parent-specific methylomes generated using MEA uncovered antithetical paternal DNAme acquisition at ~2% of the genome, including several dozen CGI promoters in the 2C embryo, a subset of which maintained such paternal DNAme to the blastocyst stage.

As mentioned above, reciprocal F1 hybrid crosses enable the delineation of genetic and parent-of-origin effects. Alternatively, comparing isogenic uniparental embryos, where both genomes are of maternal or paternal origin, can identify parent-of-origin effects. To confirm that *de novo* DNAme activity following fertilization is specific to the paternal genome, rather than a consequence of strain-specific genetic variants, our collaborator Dr. Hisato Kobayashi conducted WGBS on androgenetic blastocysts and I compared CGI promoter DNAme levels from these bi-paternal cells with those from sperm. Using this strategy, I confirmed that at several CGI promoters, identified as PDA in my allelic analysis, do indeed acquire paternal DNAme following fertilization, and uncovered additional CGI promoters that show PDA. In contrast, parthenogenetic blastocysts do not show DNAme gain relative to oocytes, suggesting that the paternal genome is indeed targeted by *de novo* DNMTs following fertilization. While these results

indicate the paternal genome is a bona fide *de novo* DNMT target, I did not detect any unique histone modifications (H3K4me2, H3K4me3, H3K9me3, H3K27me3, or H3K36me3) or DNA motifs at PDA CGI promoters. As such, I do not know whether histone modifications inherited from sperm or if sequence-specific TFs guide DNMTs to PDA sites (or, alternatively, protect non-PDA sites from *de novo* DNAme activity). As Kenjiro Shirane, postdoc in the Lorincz lab, recently found that H3K36me2 directs de novo DNAme during prenatal male germ line development, it would be interesting to see if this histone PTM is involved in PDA at the loci I identified. Another mark of interest is H2A and H4 arginine methylation (H2A/H4R3me2), which are repressive modifications found at transcriptionally silent TEs in globally hypomethylated cells such as PGCs and the preimplantation embryo (Kim et al. 2014). Whether these marks are functionally linked to DNAme, and whether they are enriched over our CGI promoters of interest remains to be determined.

In collaboration with Dr. Hiroyuki Sasaki, I demonstrated that the maternal factor DNMT3A localizes to the paternal pronucleus immediately following fertilization and, using conditional *Zp3-cre* genetic knock-out mice, that maternal DNMT3A mediates PDA. This is in line with recent IF evidence that suggested low levels of zygotic paternal DNAme is mediated by DNMT3A in a DNMT3L-independent manner (Amouroux et al. 2016). In addition to identifying the maternal factor that mediates PDA, maternal DNMT3A WT and KO embryos provided us with a system in which we could test the consequences of a lack of PDA on embryonic transcription *in vivo*.

Maternal DNMT3A KO oocytes fail to undergo proper DNAme establishment in oocytes (Shirane et al. 2013). As such, maternally methylated imprinted genes (*Mest*, *Peg3* & *Snrpn*) are ectopically expressed from the maternal allele of *Dnmt3a* matKO embryos, doubling transcript levels (Kaneda et al. 2004). While the overexpression of maternally methylated imprinted genes is generally thought to underlie the morbidity of *Dnmt3a* matKO post-implantation embryos, other regions such as CGI promoters that undergo PDA may be involved. For example, it was recently shown that non-imprinted genes in the trophoblast lineage are silenced by DNAme established in the oocyte, and their misregulation may contribute to the phenotype of *Dnmt3a* matKO embryos (Branco et al. 2016).

Interestingly, while maternally methylated imprinted genes (including *Snrpn*) are indeed overexpressed in *Dnmt3a* matKO embryos, loss of maternal DNMT3A also results in the premature expression of blastocyst-stage genes from the paternal allele of 4C embryos, and, importantly, those that show the greatest upregulation are PDA genes (*Gdap2*, *Tuba3a* & *Snx1*). Thus, I uncovered a previously unappreciated post-fertilization role of DNMT3A in allele-specific regulation of non-imprinted genes in the early embryo. The nature of these PDA genes, which are conserved from human to zebrafish, and the potential consequences on their premature expression is discussed below.

Taken together, the results reported in this thesis demonstrate the power of analyzing HTS datasets with allelic resolution, and that DNAme dynamics following fertilization are more nuanced than previously thought. Indeed, zygotic paternal DNAme acquisition at CGI promoters specifically silences genic transcription on the paternal allele in 2 and 4C embryos. In contrast, the maternal allele is hypomethylated but marked by H3K27me3 at several of these candidate regions. Such bi-modal silencing of both parental alleles reinforces the biological importance of silencing these CGI promoters during early development. How maternal DNMT3A is guided to its targets, and whether the aberrant expression of genes that show PDA at their promoters underlies the morbidity of *Dnmt3a* matKO or UPD embryos, remains to be determined.

4.4 Future directions

As discussed above, the breadth and quality of HTS datasets allow more than one hypothesis to be tested. While I successfully identified 63 CGI promoters that undergo PDA using allele-aware WGBS analysis software, other functional genomic regions also showed PDA. I focused Chapter 3 on CGI promoters because densely methylated regulatory DNA generally inhibits gene expression (Deaton and Bird 2011). Indeed, in our RNA-seq analysis, a subset of genes identified as PDA showed increased expression in DNMT3A matKO embryos. Other regions that show PDA include enhancer elements, intergenic CGIs and endogenous retroviral elements.

As with all scientific endeavours, many questions were raised during our quest to answer the simple question “can I use an allele-aware bioinformatics pipeline (MEA) to identify loci that undergo *de novo* DNAme immediately following fertilization?”. Whether regions that show PDA also regulate transcription of genes, retroelements or long non-coding RNAs can be directly tested using the same data generated for this project. As Chapter 3 focused almost exclusively on the paternal allele, a study focused on the effects of ablating maternal DNMT3A on the maternal genome, including transcription and the redistribution of histone PTMs, is worth conducting on the same datasets.

As mentioned above, another important unanswered question raised in Chapter 3 is the identity of the site-specificity factor(s) that guide maternal DNMT3A to specific regions, including CGI promoters of the zygotic paternal genome (**Figure 4.1a**).

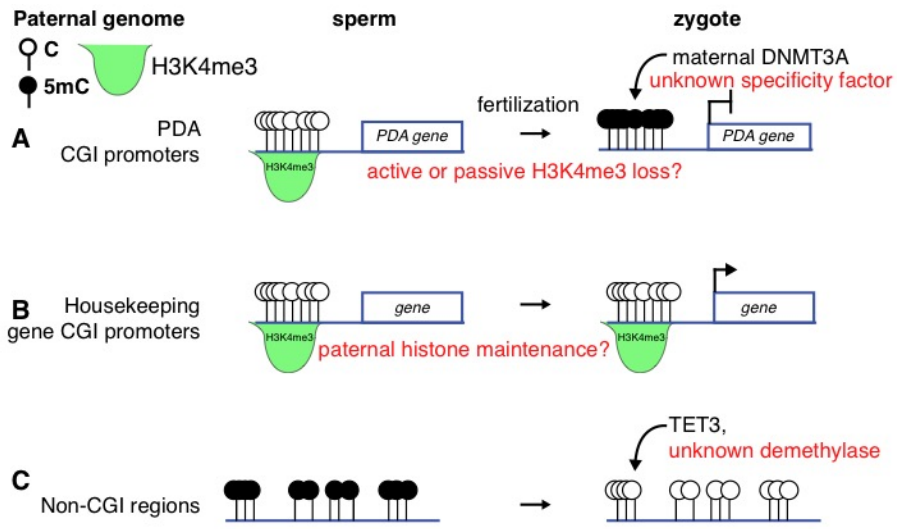

Figure 4.1 Several factors involved in paternal chromatin remodeling following fertilization remain undiscovered.

(a) 63 CGI promoters show paternal DNAme acquisition following fertilization concomitant with loss of H3K4me3. Whether specific H3K4 lysine demethylases or sperm-inherited nucleosomes are passively lost in the zygote remains unknown. Additionally, DNMT3A is likely recruited to PDA by an unknown specificity factor. **(b)** CGI promoters of housekeeping genes are marked by H3K4me3 in sperm and on the paternal genome in the early embryo, concomitant with early activation during zygotic genome activation. Whether H3K4me3 is maintained or quickly removed and replaced remains controversial (Saitou and Kurimoto 2014; Siklenka et al. 2015; Zhang et al. 2016). **(c)** Non-CGI regions, such as intergenic DNA, are hypermethylated in sperm and lose such DNAme before the first zygotic S-phase. While TET3 contributes to this demethylase activity, another unknown factor is clearly also involved.

Two scenarios are possible: DNMT3A can be guided by a specific histone PTM (or combination thereof) inherited from spermatozoa, or by a TF expressed in the oocyte/zygote. While I did not detect a histone PTM in sperm that is specific to PDA sites, other untested histone PTMs may nevertheless promote DNAme, including H3K36me2, which was recently shown to contribute to *de novo* DNAme in mouse ESCs (Weinberg et al. 2019). While H3K36me2 and H3K4me3 generally do not colocalize, there is evidence that both histone PTMs mark promoters and enhancers in oocytes (personal communication, Lorincz lab). Therefore, assuming the nucleosome content of sperm permits such “bivalent” histones, H3K36me2 inherited from sperm, combined with H3K4me3 loss in the zygote, may direct maternal DNMT3A to the zygotic paternal genome. As such, H3K36me2 ChIP-seq on mature sperm chromatin would be highly

informative. Alternatively, a TF or multiple different TFs could mediate site-specificity of DNMT3A to PDA sites. While no common DNA motifs were detected at PDA sites, several overlapping TFs may be acting in concert. Notably, several PDA sites show H3K9me3 gain on both alleles in the zygote, which is a hallmark of KRAB-ZFP binding (KRAB-ZFP proteins recruit KAP1 and in turn the H3K9 KMTase *Setdb1* to their targets) (Rowe et al. 2010; Jacobs et al. 2014; Wolf et al. 2015; Yang et al. 2017). Given that KRAB-ZFPs are the largest TF family in mammals (Looman et al. 2002; Imbeault et al. 2017), and that the DNA binding motif of their C2H2 ZFs are notoriously difficult to predict (Najafabadi et al. 2015), the identification of such TFs is non-trivial.

Another unanswered question is whether the ectopic expression of PDA genes from the paternal allele in *Dnmt3a* matKO 4C embryos is a direct or indirect result of DNAme loss at their CGI promoters. To determine whether loss of promoter DNAme at the 2C stage directly results in ectopic expression, a targeting system can be employed to specifically erase DNAme levels at single targets. For example, the endonuclease domains of Cas9 can be genetically inactivated, creating a protein that targets DNA without modifying its sequence (Gilbert et al. 2013; 2014). Tethered to an epigenetic modifier such as TET1 or DNMT3A, catalytically dead Cas9 (dCas9) enables site-directed DNAme erasure and establishment, respectively (Liu et al. 2016). Employing such epigenome modifying approaches to study the effects of *de novo* DNAme acquisition on the paternal genome will provide insights into whether transcriptional repression is a direct consequence of PDA.

4.5 Outstanding questions

While I identified loci that undergo PDA and the enzyme responsible for such activity, I do not know “why” PDA occurs, or whether the ectopic transcription of genes has any bearing on the viability of *Dnmt3a* maternal KO embryos.

It is possible that PDA is simply a by-product of the paternal genome being exposed to high maternal stores of DNMT3A. The paternal genome may be particularly susceptible to de novo DNAme, as chromatin on the paternal genome is rapidly remodelled following fertilization, and nascent histone H3 (predominantly H3.3) is thought to be almost completely devoid of H3K4me3 and H3K27me3, both of which inhibit DNMT3A binding (Torres-Padilla et al. 2006; Erkek et al. 2013; Hammoud et al. 2009). While CGIs retain H3K4me3-modified nucleosomes in sperm, several groups have reported that such modified nucleosomes are replaced following fertilization (Burton and Torres-Padilla 2010; Zhang et al. 2016). However, whether paternal nucleosomes are replaced or retained following fertilization is a controversial topic (Saitou and Kurimoto 2014; Siklenka et al. 2015; Zhang et al. 2016), and as such, whether PDA depends on histone PTM loss or maintenance remains unknown (**Figure 4.1b**). Live cell imaging of individual nucleosomes or histones during fertilization will be likely resolve this controversy and others in the field of intergenerational epigenetic inheritance.

The developmental arrest of *Dnmt3a* maternal KO embryos at ~E8.5 is generally attributed to the ectopic expression of maternally methylated imprinted genes. However, the full etiology of this mortality, including transcriptional changes of non-imprinted genes, remains elusive. While I confirmed that maternally methylated imprinted genes are indeed ectopically activated from the maternal allele, and identified additional upregulated genes from the paternal allele, these genes show changes in expression as early as the 2C stage. Since *Dnmt3a* matKO embryos die around E8.5, it is unlikely that the mortality of these embryos is attributable to changes in genic expression in the 2C embryo (~E2.0). Nevertheless, changes in expression can engender the incorrect establishment of epigenetic marks at these genes, which can result in negative effects later in development. For example, a recent report at the maternally methylated imprinted gene *Zdbf2* shows that incorrect inactivation of the long isoform of this gene in the blastocyst results in the failure to activate the canonical isoform in the adult mouse brain, in turn causing growth defects (Greenberg et al. 2016). So, while increased levels of transcripts in the 2-4C embryo likely does not contribute to embryonic lethality, premature transcription of PDA genes such as *Tuba3a*, *Gdap2* or *Snx1* may lead to downstream effects. For example, *Tuba3a* is a sperm-specific isoform of tubulin, which may lead to cytoskeletal defects in the post-implantation embryo if improperly activated. Further, *Gdap2* is a differentiation-association protein, the timing of which its expression may be critical for proper development. Finally, *Snx1* is a sorting nexin protein involved in cellular trafficking, which when genetically inactivated along with its homologue *Snx2* results in lethality at midgestation (Schwarz et al. 2002). However, since *Snx1* and *Snx2*

are redundant and *Snx2* expression is not affected in *Dnmt3a* matKO ICM cells, the aberrant expression of *Snx1* likely does not interfere with normal development.

The identity of the factor(s) responsible for the global loss of paternal DNAme before the first zygotic S-phase remain unknown (**Figure 4.1c**). While TET3 clearly plays a role, it is only responsible for ~8% of the observed ~40% reduction in DNAme (Peat et al. 2014). What factor demethylates the paternal genome before the first S-phase? Since 5mC is lost before 5hmC is detected (Amouroux et al. 2016), the putative enzyme would function using a novel demethylase mechanism or direct base removal. The elusive protein is likely expressed in the oocyte/zygote, may have a 5mC recognition domain, and may directly convert 5mC to C. Identifying such a factor and the basis of its demethylase activity is of clear interest in the field.

Finally, as the vast majority of genomic imprinting studies have been conducted in mouse and human adult tissues, the evolutionary conservation of this phenomena is relatively unknown. Further, since DNAme levels are dynamic during embryogenesis, differential parental DNAme levels in the embryo impart (albeit transient) parent-of-origin gene regulation in the early embryo, as shown in Chapter 3. Using another mammalian model organism such as the rat or rabbit would enable the measurement of conservation of imprinting, transient or otherwise.

4.6 Conclusion

In summary, this thesis describes the development and application of a novel bioinformatic pipeline for integrated analysis of disparate HTS datasets. Using this pipeline, I demonstrate that DNAme dynamics in the early embryo are more nuanced than previously thought, and determine the effects of zygotic DNAme gain on embryonic transcription. The results reported deepen our appreciation for how, shortly after fertilization, the unique maternal and paternal chromatin unite and initiate the first stages of life. Further, I believe employing MEA on existing and emerging datasets will help future clinical research in fertility, artificial reproduction, cloning and stem cell biology.

Bibliography

Adams DJ, Doran AG, Lilue J, Keane TM. 2015. The Mouse Genomes Project: a repository of inbred laboratory mouse strain genomes. *Mamm Genome* **26**: 403–412.

Allfrey VG, Faulkner R, Mirsky AE. 1964. Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. *Proceedings of the National Academy of Sciences* **51**: 786–794.

Amouroux R, Nashun B, Shirane K, Nakagawa S, Hill PWS, D'Souza Z, Nakayama M, Matsuda M, Turp A, Ndjetehe E, et al. 2016. De novo DNA methylation drives 5hmC accumulation in mouse zygotes. *Nat Cell Biol* **18**: 225–233.

Andergassen D, Dotter CP, Kulinski TM, Guenzl PM, Bammer PC, Barlow DP, Pauler FM, Hudson QJ. 2015. Allelome.PRO, a pipeline to define allele-specific genomic features from high-throughput sequencing data. *Nucleic Acids Research* **43**: e146.

Arber W, Hattman S, Dussoix D. 1963. On the host-controlled modification of bacteriophage λ. *Virology* **21**: 30–35.

Au Yeung WK, Brind'Amour J, Hatano Y, Yamagata K, Feil R, Lorincz MC, Tachibana M, Shinkai Y, Sasaki H. 2019. Histone H3K9 methyltransferase G9a in oocytes is essential for preimplantation development but dispensable for CG methylation protection. *Cell Rep* **27**: 282–293.

Auclair G, Borgel J, Sanz LA, Vallet J, Guibert S, Dumas M, Cavelier P, Girardot M, Forné T, Feil R, et al. 2016. EHMT2 directs DNA methylation for efficient gene silencing in mouse embryos. *Genome Research* **26**: 192–202.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research* **37**: W202–208.

Barau J, Teissandier A, Zamudio N, Roy S, Nalesso V, Hérault Y, Guillou F, Bourc'his D. 2016. The DNA methyltransferase DNMT3C protects male germ cells from transposon activity. *Science* **354**: 909–912.

Barlow DP. 1993. Methylation and imprinting: from host defense to gene regulation? *Science* **260**: 309–310.

Barlow DP, Bartolomei MS. 2014. Genomic imprinting in mammals. *Cold Spring Harb Perspect Biol* **6**: a018382.

Barlow DP, Stöger R, Herrmann BG, Saito K, Schweifer N. 1991. The mouse insulin-like growth factor type-2 receptor is imprinted and closely linked to the Tme locus. *Nature* **349**: 84–87.

Barton SC, Adams CA, Norris ML, Surani MA. 1985. Development of gynogenetic and parthenogenetic inner cell mass and trophectoderm tissues in reconstituted blastocysts in the mouse. *J Embryol Exp Morphol* **90**: 267–285.

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

Beard C, Li E, Jaenisch R. 1995. Loss of methylation activates Xist in somatic but not in embryonic cells. *Genes & Development* **9**: 2325–2334.

Bestor T, Laudano A, Mattaliano R, Ingram V. 1988. Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells: The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. *Journal of Molecular Biology* **203**: 971–983.

Bestor TH, Bourc'his D. 2004. Transposon Silencing and Imprint Establishment in Mammalian Germ Cells. *Cold Spring Harb Symp Quant Biol* **69**: 381–388.

Bird A. 2002. DNA methylation patterns and epigenetic memory. *Genes & Development* **16**: 6–21.

Blackledge NP, Farcas AM, Kondo T, King HW, McGouran JF, Hanssen LLP, Ito S, Cooper S, Kondo K, Koseki Y, et al. 2014. Variant PRC1 complex-dependent H2A ubiquitylation drives PRC2 recruitment and Polycomb domain formation. *Cell* **157**: 1445–1459.

Bochtler M, Kolano A, Xu G-L. 2017. DNA demethylation pathways: Additional players and regulators. *Bioessays* **39**: 1–13.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Borgel J, Guibert S, Li Y, Chiba H, Schübeler D, Sasaki H, Forné T, Weber M. 2010. Targets and dynamics of promoter DNA methylation during early mouse development. *Nat Genet* **42**: 1093–1100.

Boulard M, Edwards JR, Bestor TH. 2015. FBXL10 protects Polycomb-bound genes from hypermethylation. *Nat Genet* **47**: 479–485.

Bourc'his D, Bestor TH. 2004. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature* **431**: 96–99.

Bourc'his D, Xu GL, Lin CS, Bollman B, Bestor TH. 2001. Dnmt3L and the establishment of maternal genomic imprints. *Science* **294**: 2536–2539.

Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvák Z, Levin HL, Macfarlan TS, et al. 2018. Ten things you should know about transposable elements. *Genome Biology* **19**: 199.

Branco MR, Ficz G, Reik W. 2011. Uncovering the role of 5-hydroxymethylcytosine in the epigenome. *Nat Rev Genet* **13**: 7–13.

Branco MR, King M, Perez-Garcia V, Bogutz AB, Caley M, Fineberg E, Lefebvre L, Cook SJ, Dean W, Hemberger M, et al. 2016. Maternal DNA Methylation Regulates Early Trophoblast Development. *Developmental Cell* **36**: 152–163.

Brind'Amour J, Kobayashi H, Richard Albert J, Shirane K, Sakashita A, Kamio A, Bogutz A, Koike T, Karimi MM, Lefebvre L, et al. 2018. LTR retrotransposons transcribed in oocytes drive species-specific and heritable changes in DNA methylation. *Nature Communications* **9**: 3331.

Brind'Amour J, Liu S, Hudson M, Chen C, Karimi MM, Lorincz MC. 2015. An ultra-low-input native ChIP-seq protocol for genome-wide profiling of rare cell populations. *Nature Communications* **6**: 6033.

Brykczynska U, Hisano M, Erkek S, Ramos L, Oakeley EJ, Roloff TC, Beisel C, Schübeler D, Stadler MB, Peters AHFM. 2010. Repressive and active histone methylation mark distinct promoters in human and mouse spermatozoa. *Nat Struct Mol Biol* **17**: 679–687.

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Chem Biol* **10**: 1213–1218.

Burton A, Torres-Padilla M-E. 2010. Epigenetic reprogramming and development: a unique heterochromatin organization in the preimplantation mouse embryo. *Brief Funct Genomics* **9**: 444–454.

Chen Y-C, Gotea V, Margolin G, Elnitski L. 2017. Significant associations between driver gene mutations and DNA methylation alterations across many cancer types. *PLoS Comput Biol* **13**: e1005840.

Cheung WA, Shao X, Morin A, Siroux V, Kwan T, Ge B, Aïssi D, Chen L, Vasquez L, Allum F, et al. 2017. Functional variation in allelic methylomes underscores a strong

genetic contribution and reveals novel epigenetic alterations in the human epigenome. *Genome Biology* **18**: 50.

Choufani S, Shuman C, Weksberg R. 2010. Beckwith-Wiedemann syndrome. *Am J Med Genet C Semin Med Genet* **154C**: 343–354.

Coulondre C, Miller JH, Farabaugh PJ, Gilbert W. 1978. Molecular basis of base substitution hotspots in Escherichia coli. *Nature* **274**: 775–780.

DaRosa PA, Harrison JS, Zelter A, Davis TN, Brzovic P, Kuhlman B, Klevit RE. 2018. A bifunctional role for the UHRF1 UBL domain in the control of hemi-methylated DNA-dependent histone ubiquitylation. *Molecular Cell* **72**: 753–765.

de Vries WN, Binns LT, Fancher KS, Dean J, Moore R, Kemler R, Knowles BB. 2000. Expression of Cre recombinase in mouse oocytes: a means to study maternal effect genes. *genesis* **26**: 110–112.

Deaton AM, Bird A. 2011. CpG islands and the regulation of transcription. *Genes & Development* **25**: 1010–1022.

Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, Pritchard JK. 2009. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**: 3207–3212.

Delaneau O, Coulonges C, Zagury J-F. 2008. Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics* **9**: 540.

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

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

Du J, Johnson LM, Jacobsen SE, Patel DJ. 2015. DNA methylation pathways and their crosstalk with histone methylation. *Nat Rev Mol Cell Biol* **16**: 519–532.

Eckersley-Maslin MA, Alda-Catalinas C, Reik W. 2018. Dynamics of the epigenetic landscape during the maternal-to-zygotic transition. *Nat Rev Mol Cell Biol* **19**: 436–450.

Edwards JR, Yarychkivska O, Boulard M, Bestor TH. 2017. DNA methylation and DNA methyltransferases. *Epigenetics Chromatin* **10**: 23.

Endoh M, Endo TA, Shinga J, Hayashi K, Farcas A, Ma K-W, Ito S, Sharif J, Endoh T, Onaga N, et al. 2017. PCGF6-PRC1 suppresses premature differentiation of mouse embryonic stem cells by regulating germ cell-related genes. *eLife* **6**: e21064.

Erkek S, Hisano M, Liang C-Y, Gill M, Murr R, Dieker J, Schübeler D, Vlag JVD, Stadler MB, Peters AHFM. 2013. Molecular determinants of nucleosome retention at CpG-rich sequences in mouse spermatozoa. *Nat Struct Mol Biol* **20**: 868–875.

Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat Chem Biol* **9**: 215–216.

Feinberg AP, Vogelstein B. 1983. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature* **301**: 89–92.

Ferguson-Smith AC, CATTANACH BM, Barton SC, Beechey CV, Surani MA. 1991. Embryological and molecular investigations of parental imprinting on mouse chromosome 7. *Nature* **351**: 667–670.

Fischle W, Wang YM, Allis CD. 2003. Binary switches and modification cassettes in histone biology and beyond. *Nature* **425**: 475–479.

Gaysinskaya V, Miller BF, De Luca C, van der Heijden GW, Hansen KD, Bortvin A. 2018. Transient reduction of DNA methylation at the onset of meiosis in male mice. *Epigenetics Chromatin* **11**: 15.

Gilbert LA, Horlbeck MA, Adamson B, Villalta JE, Chen Y, Whitehead EH, Guimaraes C, Panning B, Ploegh HL, Bassik MC, et al. 2014. Genome-scale CRISPR-mediated control of gene repression and activation. *Cell* **159**: 647–661.

Gilbert LA, Larson MH, Morsut L, Liu Z, Brar GA, Torres SE, Stern-Ginossar N, Brandman O, Whitehead EH, Doudna JA, et al. 2013. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**: 442–451.

Goncalves A, Leigh-Brown S, Thybert D, Stefflova K, Turro E, Flicek P, Brazma A, Odom DT, Marioni JC. 2012. Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. *Genome Research* **22**: 2376–2384.

Greenberg MVC, Bourc'his D. 2019. The diverse roles of DNA methylation in mammalian development and disease. *Nat Rev Mol Cell Biol* **20**: 590–607.

Greenberg MVC, Glaser J, Borsos M, Marjou FE, Walter M, Teissandier A, Bourc'his D. 2016. Transient transcription in the early embryo sets an epigenetic state that programs postnatal growth. *Nat Genet* **49**: 110–118.

Gu T, Lin X, Cullen SM, Luo M, Jeong M, Estecio M, Shen J, Hardikar S, Sun D, Su J, et al. 2018. DNMT3A and TET1 cooperate to regulate promoter epigenetic landscapes in mouse embryonic stem cells. *Genome Biology* **19**: 88.

Gu T-P, Guo F, Yang H, Wu H-P, Xu G-F, Liu W, Xie Z-G, Shi L, He X, Jin S-G, et al. 2011a. The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. *Nature* **477**: 606–610.

Gu T-P, Guo F, Yang H, Wu H-P, Xu G-F, Liu W, Xie Z-G, Shi L, He X, Jin S-G, et al. 2011b. The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. *Nature* **477**: 606–610.

Guo F, Li X, Liang D, Li T, Zhu P, Guo H, Wu X, Wen L, Gu T-P, Hu B, et al. 2014. Active and passive demethylation of male and female pronuclear DNA in the mammalian zygote. *Cell Stem Cell* **15**: 447–458.

Guo H, Hu B, Yan L, Yong J, Wu Y, Gao Y, Guo F, Hou Y, Fan X, Dong J, et al. 2017. DNA methylation and chromatin accessibility profiling of mouse and human fetal germ cells. *Cell Res* **27**: 165–183.

Hackett JA, Sengupta R, Zylicz JJ, Murakami K, Lee C, Down TA, Surani MA. 2013. Germline DNA demethylation dynamics and imprint erasure through 5-hydroxymethylcytosine. *Science* **339**: 448–452.

Haig D. 2000. The kinship theory of genomic imprinting. *Annual Review of Ecology and Systematics* **31**: 9–32.

Haig D, Graham C. 1991. Genomic imprinting and the strange case of the insulin-like growth factor II receptor. *Cell* **64**: 1045–1046.

Hajkova P, Jeffries SJ, Lee C, Miller N, Jackson SP, Surani MA. 2010. Genome-wide reprogramming in the mouse germ line entails the base excision repair pathway. *Science* **329**: 78–82.

Hammoud SS, Nix DA, Zhang H, Purwar J, Carrell DT, Cairns BR. 2009. Distinctive chromatin in human sperm packages genes for embryo development. *Nature* **460**: 473–478.

Han L, Ren C, Zhang J, Shu W, Wang Q. 2019. Differential roles of Stella in the modulation of DNA methylation during oocyte and zygotic development. *Cell Discov* **5**: 9.

Harvey CT, Moyerbrailean GA, Davis GO, Wen X, Luca F, Pique-Regi R. 2015. QuASAR: quantitative allele-specific analysis of reads. *Bioinformatics* **31**: 1235–1242.

Hashimoto H, Horton JR, Zhang X, Bostick M, Jacobsen SE, Cheng X. 2008. The SRA domain of UHRF1 flips 5-methylcytosine out of the DNA helix. *Nature* **455**: 826–829.

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

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell* **38**: 576–589.

Heyn P, Logan CV, Fluteau A, Challis RC, Auchynnikava T, Martin C-A, Marsh JA, Taglini F, Kilanowski F, Parry DA, et al. 2019. Gain-of-function DNMT3A mutations cause microcephalic dwarfism and hypermethylation of Polycomb-regulated regions. *Nat Genet* **51**: 96–105.

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

Holliday R. 1990. Genomic imprinting and allelic exclusion. *Development* **108**: 125–129.

Holliday R, Pugh JE. 1975. DNA modification mechanisms and gene activity during development. *Science* **187**: 226–232.

Horsthemke B, Wagstaff J. 2008. Mechanisms of imprinting of the Prader-Willi/Angelman region. *Am J Med Genet A* **146A**: 2041–2052.

Hu K, Ting AH, Li J. 2015. BSPAT: a fast online tool for DNA methylation co-occurrence pattern analysis based on high-throughput bisulfite sequencing data. *BMC Bioinformatics* **16**: 220.

Imbeault M, Helleboid P-Y, Trono D. 2017. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**: 550–554.

Inoue A, Jiang L, Lu F, Suzuki T, Zhang Y. 2017. Maternal H3K27me3 controls DNA methylation-independent imprinting. *Nature* **547**: 419–424.

Iqbal K, Jin S-G, Pfeifer GP, Szabó PE. 2011. Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. *Proc Natl Acad Sci USA* **108**: 3642–3647.

Ishiyama S, Nishiyama A, Saeki Y, Moritsugu K, Morimoto D, Yamaguchi L, Arai N, Matsumura R, Kawakami T, Mishima Y, et al. 2017. Structure of the Dnmt1 reader module complexed with a unique two-mono-ubiquitin mark on histone H3 reveals the basis for DNA methylation maintenance. *Molecular Cell* **68**: 350–360.

Ito S, D'Alessio AC, Taranova OV, Hong K, Sowers LC, Zhang Y. 2010. Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* **466**: 1129–1133.

Ito S, Shen L, Dai Q, Wu SC, Collins LB, Swenberg JA, He C, Zhang Y. 2011. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333**: 1300–1303.

Jacobs FMJ, Greenberg D, Nguyen N, Haeussler M, Ewing AD, Katzman S, Paten B, Salama SR, Haussler D. 2014. An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature* **516**: 242–245.

Jiang Y, Zhang NR, Li M. 2017. SCALE: modeling allele-specific gene expression by single-cell RNA sequencing. *Genome Biology* **18**: 74.

John RM, Lefebvre L. 2011. Developmental regulation of somatic imprints. *Differentiation* **81**: 270–280.

Jones PA, Baylin SB. 2007. The epigenomics of cancer. *Cell* **128**: 683–692.

Kajii T, Ohama K. 1977. Androgenetic origin of hydatidiform mole. *Nature* **268**: 633–634.

Kaneda M, Hirasawa R, Chiba H, Okano M, Li E, Sasaki H. 2010. Genetic evidence for Dnmt3a-dependent imprinting during oocyte growth obtained by conditional knockout with Zp3-Cre and complete exclusion of Dnmt3b by chimera formation. *Genes to Cells* **15**: 169–179.

Kaneda M, Okano M, Hata K, Sado T, Tsujimoto N, Li E, Sasaki H. 2004. Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. *Nature* **429**: 900–903.

Karimi MM, Goyal P, Maksakova IA, Bilenky M, Leung D, Tang JX, Shinkai Y, Mager DL, Jones S, Hirst M, et al. 2011. DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs. *Cell Stem Cell* **8**: 676–687.

Kato Y, Kaneda M, Hata K, Kumaki K, Hisano M, Kohara Y, Okano M, Li E, Nozaki M, Sasaki H. 2007. Role of the Dnmt3 family in de novo methylation of imprinted and

repetitive sequences during male germ cell development in the mouse. *Human Molecular Genetics* **16**: 2272–2280.

Kaufman MH. 1973. Parthenogenesis in the mouse. *Nature* **242**: 475–476.

Keane TM, Goodstadt L, Danecek P, White MA. 2011. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**: 289–294.

Keown CL, Berletch JB, Castanon R, Nery JR, Disteche CM, Ecker JR, Mukamel EA. 2017. Allele-specific non-CG DNA methylation marks domains of active chromatin in female mouse brain. *Proc Natl Acad Sci USA* **114**: E2882–E2890.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* **14**: R36.

Kim S, Günesdogan U, Zylicz JJ, Hackett JA, Cougot D, Bao S, Lee C, Dietmann S, Allen GE, Sengupta R, et al. 2014. PRMT5 protects genomic integrity during global DNA demethylation in primordial germ cells and preimplantation embryos. *Molecular Cell* **56**: 564–579.

Kizer KO, Phatnani HP, Shibata Y, Hall H, Greenleaf AL, Strahl BD. 2005. A novel domain in Set2 mediates RNA polymerase II interaction and couples histone H3 K36 methylation with transcript elongation. *Molecular and Cellular Biology* **25**: 3305–3316.

Kobayashi H, Sakurai T, Imai M, Takahashi N, Fukuda A, Yayoi O, Sato S, Nakabayashi K, Hata K, Sotomaru Y, et al. 2012. Contribution of intragenic DNA methylation in mouse gametic DNA methylomes to establish oocyte-specific heritable marks. *PLoS Genet* **8**: e1002440.

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

Kono T, Obata Y, Yoshimzu T, Nakahara T, Carroll J. 1996. Epigenetic modifications during oocyte growth correlates with extended parthenogenetic development in the mouse. *Nat Genet* **13**: 91–94.

Kono T, Sotomaru Y, Sato Y, Nakahara T. 1993. Development of androgenetic mouse embryos produced by in vitro fertilization of enucleated oocytes. *Mol Reprod Dev* **34**: 43–46.

Križanovic K, Echchiki A, Roux J, Šikic M. 2018. Evaluation of tools for long read RNA-seq splice-aware alignment. *Bioinformatics* **34**: 748–754.

Krueger F, Andrews SR. 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**: 1571–1572.

Kubo N, Toh H, Shirane K, Shirakawa T, Kobayashi H, Sato T, Sone H, Sato Y, Tomizawa S-I, Tsurusaki Y, et al. 2015. DNA methylation and gene expression dynamics during spermatogonial stem cell differentiation in the early postnatal mouse testis. *BMC Genomics* **16**: 624.

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: A resource for timelines, timetrees, and divergence times. *Mol Biol Evol* **34**: 1812–1819.

Kweon S-M, Zhu B, Chen Y, Aravind L, Xu S-Y, Feldman DE. 2017. Erasure of Tet-oxidized 5-methylcytosine by a SRAP nuclease. *Cell Rep* **21**: 482–494.

La Salle S, Oakes CC, Neaga OR, Bourc'his D, Bestor TH, Trasler JM. 2007. Loss of spermatogonia and wide-spread DNA methylation defects in newborn male mice deficient in DNMT3L. *BMC Dev Biol* **7**: 104.

Lane N, Dean W, Erhardt S, Hajkova P, Surani A, Walter J, Reik W. 2003. Resistance of IAPs to methylation reprogramming may provide a mechanism for epigenetic inheritance in the mouse. *genesis* **35**: 88–93.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.

Lee HJ, Hore TA, Reik W. 2014. Reprogramming the methylome: erasing memory and creating diversity. *Cell Stem Cell* **14**: 710–719.

Lee TI, Young RA. 2013. Transcriptional regulation and its misregulation in disease. *Cell* **152**: 1237–1251.

Lees-Murdock DJ, De Felici M, Walsh CP. 2003. Methylation dynamics of repetitive DNA elements in the mouse germ cell lineage. *Genomics* **82**: 230–237.

Lefebvre L, Viville S, Barton SC, Ishino F, Keverne EB, Surani MA. 1998. Abnormal maternal behaviour and growth retardation associated with loss of the imprinted gene Mest. *Nat Genet* **20**: 163–169.

Leitch HG, McEwen KR, Turp A, Encheva V, Carroll T, Grabole N, Mansfield W, Nashun B, Knezovich JG, Smith A, et al. 2013. Naive pluripotency is associated with global DNA hypomethylation. *Nat Struct Mol Biol* **20**: 311–316.

Leung D, Jung I, Rajagopal N, Schmitt A, Selvaraj S, Lee AY, Yen C-A, Lin S, Lin Y, Qiu Y, et al. 2015. Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature* **518**: 350–354.

Lewis EB. 1978. A gene complex controlling segmentation in Drosophila. *Nature* **276**: 565–570.

Li E, Beard C, Jaenisch R. 1993. Role for DNA methylation in genomic imprinting. *Nature* **366**: 362–365.

Li E, Bestor TH, Jaenisch R. 1992. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69**: 915–926.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Li X, Ito M, Zhou F, Youngson N, Zuo X, Leder P, Ferguson-Smith AC. 2008. A maternal-zygotic effect gene, Zfp57, maintains both maternal and paternal imprints. *Developmental Cell* **15**: 547–557.

Li Y, Li J. 2019. Technical advances contribute to the study of genomic imprinting. *PLoS Genet* **15**: e1008151.

Li Y, Zhang Z, Chen J, Liu W, Lai W, Liu B, Li X, Liu L, Xu S, Dong Q, et al. 2018. Stella safeguards the oocyte methylome by preventing de novo methylation mediated by DNMT1. *Nature* **564**: 136–140.

Lister R, Ecker JR. 2009. Finding the fifth base: genome-wide sequencing of cytosine methylation. *Genome Research* **19**: 959–966.

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo Q-M, et al. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**: 315–322.

Liu S, Brind'Amour J, Karimi MM, Shirane K, Bogutz A, Lefebvre L, Sasaki H, Shinkai Y, Lorincz MC. 2014. Setdb1 is required for germline development and silencing of H3K9me3-marked endogenous retroviruses in primordial germ cells. *Genes & Development* **28**: 2041–2055.

Liu XS, Wu H, Ji X, Stelzer Y, Wu X, Czauderna S, Shu J, Dadon D, Young RA, Jaenisch R. 2016. Editing DNA methylation in the mammalian genome. *Cell* **167**: 233–247.

Liu Y, Toh H, Sasaki H, Zhang X, Cheng X. 2012. An atomic model of Zfp57 recognition of CpG methylation within a specific DNA sequence. *Genes & Development* **26**: 2374–2379.

Looman C, Abrink M, Mark C, Hellman L. 2002. KRAB zinc finger proteins: an analysis of the molecular mechanisms governing their increase in numbers and complexity during evolution. *Mol Biol Evol* **19**: 2118–2130.

Lorincz MC, Dickerson DR, Schmitt M, Groudine M. 2004. Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. *Nat Struct Mol Biol* **11**: 1068–1075.

Lorincz MC, Schübeler D, Hutchinson SR, Dickerson DR, Groudine M. 2002. DNA Methylation Density Influences the Stability of an Epigenetic Imprint and Dnmt3a/b-Independent De Novo Methylation. *Molecular and Cellular Biology* **22**: 7572–7580.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**: 550.

Maenohara S, Unoki M, Toh H, Ohishi H, Sharif J, Koseki H, Sasaki H. 2017. Role of UHRF1 in de novo DNA methylation in oocytes and maintenance methylation in preimplantation embryos. *PLoS Genet* **13**: e1007042.

Matoba S, Wang H, Jiang L, Lu F, Iwabuchi KA, Wu X, Inoue K, Yang L, Press W, Lee JT, et al. 2018. Loss of H3K27me3 Imprinting in Somatic Cell Nuclear Transfer Embryos Disrupts Post-Implantation Development. *Cell Stem Cell* **23**: 343–354.

Maupetit-Méhouas S, Montibus B, Nury D, Tayama C, Wassef M, Kota SK, Fogli A, Cerqueira Campos F, Hata K, Feil R, et al. 2016. Imprinting control regions (ICRs) are marked by mono-allelic bivalent chromatin when transcriptionally inactive. *Nucleic Acids Research* **44**: 621–635.

Mayba O, Gilbert HN, Liu J, Haverty PM, Jhunjhunwala S, Jiang Z, Watanabe C, Zhang Z. 2014. MBASED: allele-specific expression detection in cancer tissues and cell lines. *Genome Biology* **15**: 405.

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

McClelland, M. The effect of sequence specific DNA methylation on restriction endonuclease cleavage. 1981. *Nucleic Acids Research* **9,** 5859–5866.

McEwen KR, Ferguson-Smith AC. 2010. Distinguishing epigenetic marks of developmental and imprinting regulation. *Epigenetics Chromatin* **3**: 2.

McGrath J, Solter D. 1984. Completion of mouse embryogenesis requires both the maternal and paternal genomes. *Cell* **37**: 179–183.

McSwiggen DT, Mir M, Darzacq X, Tjian R. 2019. Evaluating phase separation in live cells: diagnosis, caveats, and functional consequences. *Genes & Development* **33**: 1619–1634.

McVean GA, Altshuler Co-Chair DM, Durbin Co-Chair RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, et al. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56–65.

Messerschmidt DM. 2016. A twist in zygotic reprogramming. *Nat Cell Biol* **18**: 139–140.

Meyenn von F, Reik W. 2015. Forget the Parents: Epigenetic Reprogramming in Human Germ Cells. *Cell* **161**: 1248–1251.

Miller D, Brinkworth M, Iles D. 2010. Paternal DNA packaging in spermatozoa: more than the sum of its parts? DNA, histones, protamines and epigenetics. *Reproduction* **139**: 287–301.

Monk D. 2015. Genomic imprinting in the human placenta. *Am J Obstet Gynecol* **213**: S152–62.

Murrell A, Heeson S, Reik W. 2004. Interaction between differentially methylated regions partitions the imprinted genes Igf2 and H19 into parent-specific chromatin loops. *Nat Genet* **36**: 889–893.

Najafabadi HS, Mnaimneh S, Schmitges FW, Garton M, Lam KN, Yang A, Albu M, Weirauch MT, Radovani E, Kim PM, et al. 2015. C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nature Biotechnology* **33**: 555–562.

Nellåker C, Keane TM, Yalcin B, Wong K, Agam A, Belgard TG, Flint J, Adams DJ, Frankel WN, Ponting CP. 2012. The genomic landscape shaped by selection on transposable elements across 18 mouse strains. *Genome Biology* **13**: R45–21.

Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E, Maurano MT, Vierstra J, Thomas S, et al. 2012. BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**: 1919–1920.

Obata Y, Ono Y, Akuzawa H, Kwon OY, Yoshizawa M, Kono T. 2000. Post-implantation development of mouse androgenetic embryos produced by in-vitro fertilization of enucleated oocytes. *Hum Reprod* **15**: 874–880.

Okae H, Chiba H, Hiura H, Hamada H, Sato A, Utsunomiya T, Kikuchi H, Yoshida H, Tanaka A, Suyama M, et al. 2014. Genome-wide analysis of DNA methylation dynamics during early human development. *PLoS Genet* **10**: e1004868.

Okae H, Hiura H, Nishida Y, Funayama R, Tanaka S, Chiba H, Yaegashi N, Nakayama K, Sasaki H, Arima T. 2012. Re-investigation and RNA sequencing-based identification of genes with placenta-specific imprinted expression. *Human Molecular Genetics* **21**: 548–558.

Okano M, Bell DW, Haber DA, Li E. 1999. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**: 247–257.

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

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

Otani J, Nankumo T, Arita K, Inamoto S, Ariyoshi M, Shirakawa M. 2009. Structural basis for recognition of H3K4 methylation status by the DNA methyltransferase 3A ATRX-DNMT3-DNMT3L domain. *EMBO reports* **10**: 1235–1241.

Pastinen T. 2010. Genome-wide allele-specific analysis: insights into regulatory variation. *Nat Rev Genet* **11**: 533–538.

Peat JR, Dean W, Clark SJ, Krueger F, Smallwood SA, Ficz G, Kim JK, Marioni JC, Hore TA, Reik W. 2014. Genome-wide bisulfite sequencing in zygotes identifies demethylation targets and maps the contribution of TET3 oxidation. *Cell Rep* **9**: 1990–2000.

Perez MF, Lehner B. 2019. Intergenerational and transgenerational epigenetic inheritance in animals. *Nat Cell Biol* **21**: 143–151.

Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* **33**: 290–295.

Pinheiro I, Heard E. 2017. X chromosome inactivation: new players in the initiation of gene silencing. *F1000Research* **6**: 344.

Proudhon C, Duffié R, Ajjan S, Cowley M, Iranzo J, Carbajosa G, Saadeh H, Holland ML, Oakey RJ, Rakyan VK, et al. 2012. Protection against de novo methylation is

instrumental in maintaining parent-of-origin methylation inherited from the gametes. *Molecular Cell* **47**: 909–920.

Qu J, Hodges E, Molaro A, Gagneux P, Dean MD, Hannon GJ, Smith AD. 2017. Evolutionary expansion of DNA hypomethylation in the mammalian germline genome. *Genome Research* **28**: 145–158.

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**: 341.

Quenneville S, Verde G, Corsinotti A, Kapopoulou A, Jakobsson J, Offner S, Baglivo I, Pedone PV, Grimaldi G, Riccio A, et al. 2011. In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions. *Molecular Cell* **44**: 361–372.

Ramsahoye BH, Biniszkiewicz D, Lyko F, Clark V, Bird AP, Jaenisch R. 2000. Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proceedings of the National Academy of Sciences* **97**: 5237–5242.

Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita PA, Wang T, Nguyen N, Paten B, Zweig AS, Karolchik D, et al. 2014. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics* **30**: 1003–1005.

Reik W. 2007. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* **447**: 425–432.

Reik W, Romer I, Barton SC, Surani MA, Howlett SK, Klose J. 1993. Adult phenotype in the mouse can be affected by epigenetic events in the early embryo. *Development* **119**: 933–942.

Richard Albert J, Koike T, Younesy H, Thompson R, Bogutz AB, Karimi MM, Lorincz MC. 2018. Development and application of an integrated allele-specific pipeline for methylomic and epigenomic analysis (MEA). *BMC Genomics* **19**: 463.

Riggs AD, Pfeifer GP. 1992. X-chromosome inactivation and cell memory. *Trends in Genetics* **8**: 169–174.

Robertson KD. 2005. DNA methylation and human disease. *Nat Rev Genet* **6**: 597–610.

Rose NR, Klose RJ. 2014. Understanding the relationship between DNA methylation and histone lysine methylation. *Biochim Biophys Acta* **1839**: 1362–1372.

Rothbart SB, Krajewski K, Nady N, Tempel W, Xue S, Badeaux AI, Barsyte-Lovejoy D, Martinez JY, Bedford MT, Fuchs SM, et al. 2012. Association of UHRF1 with methylated H3K9 directs the maintenance of DNA methylation. *Nat Struct Mol Biol* **19**: 1155–1160.

Rowe HM, Jakobsson J, Mesnard D, Rougemont J, Reynard S, Aktas T, Maillard PV, Layard-Liesching H, Verp S, Marquis J, et al. 2010. KAP1 controls endogenous retroviruses in embryonicstem cells. *Nature* **463**: 237–240.

Rutledge CE, Thakur A, O'Neill KM, Irwin RE, Sato S, Hata K, Walsh CP. 2014. Ontogeny, conservation and functional significance of maternally inherited DNA methylation at two classes of non-imprinted genes. *Development* **141**: 1313–1323.

Saitou M, Kurimoto K. 2014. Paternal nucleosomes: are they retained in developmental promoters or gene deserts? *Developmental Cell* **30**: 6–8.

Santos F, Hendrich B, Reik W, Dean W. 2002. Dynamic reprogramming of DNA methylation in the early mouse Embryo. *Developmental Biology* **241**: 172–182.

Sardina JL, Collombet S, Tian TV, Gómez A, Di Stefano B, Berenguer C, Brumbaugh J, Stadhouders R, Segura-Morales C, Gut M, et al. 2018. Transcription factors drive Tet2-mediated enhancer demethylation to reprogram cell fate. *Cell Stem Cell* **23**: 727–741.

Schultz DC, Friedman JR, Rauscher FJ. 2001. Targeting histone deacetylase complexes via KRAB-zinc finger proteins: the PHD and bromodomains of KAP-1 form a cooperative unit that recruits a novel isoform of the Mi-2alpha subunit of NuRD. *Genes & Development* **15**: 428–443.

Schwarz DG, Griffin CT, Schneider EA, Yee D, Magnuson T. 2002. Genetic analysis of sorting nexins 1 and 2 reveals a redundant and essential function in mice. *Mol Biol Cell* **13**: 3588–3600.

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

Seisenberger S, Peat JR, Hore TA, Santos F, Dean W, Reik W. 2013a. Reprogramming DNA methylation in the mammalian life cycle: building and breaking epigenetic barriers. *Philos Trans R Soc Lond, B, Biol Sci* **368**: 20110330.

Seisenberger S, Peat JR, Reik W. 2013b. Conceptual links between DNA methylation reprogramming in the early embryo and primordial germ cells. *Current Opinion in Cell Biology* **25**: 281–288.

Sendzikaite G, Hanna CW, Stewart-Morgan KR, Ivanova E, Kelsey G. 2019. A DNMT3A PWWP mutation leads to methylation of bivalent chromatin and growth retardation in mice. *Nature Communications* **10**: 1884.

Sharif J, Muto M, Takebayashi SI, Suetake I, Iwamatsu A, Endo TA, Shinga J, Mizutani-Koseki Y, Toyoda T, Okamura K, et al. 2007. The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. *Nature* **450**: 908–912.

Shen L, Inoue A, He J, Liu Y, Lu F, Zhang Y. 2014. Tet3 and DNA replication mediate demethylation of both the maternal and paternal genomes in mouse zygotes. *Cell Stem Cell* **15**: 459–471.

Shen L, Zhang Y. 2013. 5-Hydroxymethylcytosine: generation, fate, and genomic distribution. *Current Opinion in Cell Biology* **25**: 289–296.

Shirane K, Toh H, Kobayashi H, Miura F, Chiba H, Ito T, Kono T, Sasaki H. 2013. Mouse oocyte methylomes at base resolution reveal genome-wide accumulation of non-CpG methylation and role of DNA methyltransferases. *PLoS Genet* **9**: e1003439.

Siklenka K, Erkek S, Godmann M, Lambrot R, McGraw S, Lafleur C, Cohen T, Xia J, Suderman M, Hallett M, et al. 2015. Disruption of histone methylation in developing sperm impairs offspring health transgenerationally. *Science* **350**: aab2006.

Simon JA, Kingston RE. 2013. Occupying chromatin: polycomb mechanisms for getting to genomic targets, stopping transcriptional traffic, and staying put. *Molecular Cell* **49**: 808–824.

Smallwood SA, Kelsey G. 2012. De novo DNA methylation: a germ cell perspective. *Trends in Genetics* **28**: 33–42.

Smallwood SA, Tomizawa S-I, Krueger F, Ruf N, Carli N, Segonds-Pichon A, Sato S, Hata K, Andrews SR, Kelsey G. 2011. Dynamic CpG island methylation landscape in oocytes and preimplantation embryos. *Nat Genet* **43**: 811–814.

Smith ZD, Chan MM, Humm KC, Karnik R, Mekhoubad S, Regev A, Eggan K, Meissner A. 2014. DNA methylation dynamics of the human preimplantation embryo. *Nature* **511**: 611–615.

Smith ZD, Chan MM, Mikkelsen TS, Gu H, Gnirke A, Regev A, Meissner A. 2012. A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature* **484**: 339–344.

Solter D, Knowles BB. 1975. Immunosurgery of mouse blastocyst. *Proceedings of the National Academy of Sciences* **72**: 5099–5102.

Song J, Rechkoblit O, Bestor TH, Patel DJ. 2011. Structure of DNMT1-DNA complex reveals a role for autoinhibition in maintenance DNA methylation. *Science* **331**: 1036–1040.

Song L, Crawford GE. 2010. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc* **2010**: pdb.prot5384.

Song Q, Decato B, Hong EE, Zhou M, Fang F, Qu J, Garvin T, Kessler M, Zhou J, Smith AD. 2013. A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PLoS ONE* **8**: e81148.

Sontag LB, Lorincz MC, Georg Luebeck E. 2006. Dynamics, stability and inheritance of somatic DNA methylation imprints. *Journal of Theoretical Biology* **242**: 890–899.

Sood AJ, Viner C, Hoffman MM. 2019. DNAmod: the DNA modification database. *J Cheminform* **11**: 30.

Stewart KR, Veselovska L, Kim J, Huang J, Saadeh H, Tomizawa S-I, Smallwood SA, Chen T, Kelsey G. 2015. Dynamic changes in histone modifications precede de novo DNA methylation in oocytes. *Genes & Development* **29**: 2449–2462.

Strogantsev R, Krueger F, Yamazawa K, Shi H, Gould P, Goldman-Roberts M, McEwen K, Sun B, Pedersen R, Ferguson-Smith AC. 2015. Allele-specific binding of ZFP57 in the epigenetic regulation of imprinted and non-imprinted monoallelic expression. *Genome Biology* **16**: 112.

Surani MA. 2001. Reprogramming of genome function through epigenetic inheritance. *Nature* **414**: 122–128.

Surani MA, Barton SC, Norris ML. 1984. Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. *Nature* **308**: 548–550.

Szabo PE, Mann JR. 1995. Biallelic expression of imprinted genes in the mouse germ line: implications for erasure, establishment, and mechanisms of genomic imprinting. *Genes & Development* **9**: 1857–1868.

Taft RA, Davisson M, Wiles MV. 2006. Know thy mouse. *Trends in Genetics* **22**: 649–653.

Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, Agarwal S, Iyer LM, Liu DR, Aravind L, et al. 2009. Conversion of 5-methylcytosine to 5-

hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**: 930–935.

Takahashi N, Coluccio A, Thorball CW, Planet E, Shi H, Offner S, Turelli P, Imbeault M, Ferguson-Smith AC, Trono D. 2019. ZNF445 is a primary regulator of genomic imprinting. *Genes & Development* **33**: 49–54.

Tatsumi D, Hayashi Y, Endo M, Kobayashi H, Yoshioka T, Kiso K, Kanno S, Nakai Y, Maeda I, Mochizuki K, et al. 2018. DNMTs and SETDB1 function as co-repressors in MAX-mediated repression of germ cell-related genes in mouse embryonic stem cells. *PLoS ONE* **13**: e0205969.

Tatton-Brown K, Seal S, Ruark E, Harmer J, Ramsay E, Del Vecchio Duarte S, Zachariou A, Hanks S, O'Brien E, Aksglaede L, et al. 2014. Mutations in the DNA methyltransferase gene DNMT3A cause an overgrowth syndrome with intellectual disability. *Nat Genet* **46**: 385–388.

The Genome Reference Consortium. 2019. https://www.ncbi.nlm.nih.gov/grc.

Toh H, Shirane K, Miura F, Kubo N, Ichiyanagi K, Hayashi K, Saitou M, Suyama M, Ito T, Sasaki H. 2017. Software updates in the Illumina HiSeq platform affect whole-genome bisulfite sequencing. *BMC Genomics* **18**: 31.

Torres-Padilla M-E, Bannister AJ, Hurd PJ, Kouzarides T, Zernicka-Goetz M. 2006. Dynamic distribution of the replacement histone variant H3.3 in the mouse oocyte and preimplantation embryos. *Int J Dev Biol* **50**: 455–461.

Trasler JM. 2006. Gamete imprinting: setting epigenetic patterns for the next generation. *Reprod Fertil Dev* **18**: 63–7.

Tsukada Y-I, Akiyama T, Nakayama KI. 2015. Maternal TET3 is dispensable for embryonic development but is required for neonatal growth. *Sci Rep* **5**: 15876.

Tucci V, Isles AR, Kelsey G, Ferguson-Smith AC, Group TEI, Tucci V, Bartolomei MS, Benvenisty N, Bourc'his D, Charalambous M, et al. 2019. Genomic imprinting and physiological processes in mammals. *Cell* **176**: 952–965.

Tuorto F, Herbst F, Alerasool N, Bender S, Popp O, Federico G, Reitter S, Liebers R, Stoecklin G, Gröne H-J, et al. 2015. The tRNA methyltransferase Dnmt2 is required for accurate polypeptide synthesis during haematopoiesis. *EMBO J* **34**: 2350–2362.

Turro E, Su S-Y, Goncalves A, Coin LJM, Richardson S, Lewin A. 2011. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biology* **12**: R13.

Umlauf D, Goto Y, Cao R, Cerqueira F, Wagschal A, Zhang Y, Feil R. 2004. Imprinting along the Kcnq1 domain on mouse chromosome 7 involves repressive histone methylation and recruitment of Polycomb group complexes. *Nat Genet* **36**: 1296–1300.

van de Geijn B, McVicker G, Gilad Y, Pritchard JK. 2015. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat Chem Biol* **12**: 1061–1063.

Velasco G, Hube F, Rollin J, Neuillet D, Philippe C, Bouzinba-Segard H, Galvani A, Viegas-Pequignot E, Francastel C. 2010. Dnmt3b recruitment through E2F6 transcriptional repressor mediates germ-line gene silencing in murine somatic tissues. *Proceedings of the National Academy of Sciences* **107**: 9281–9286.

Veselovska L, Smallwood SA, Saadeh H, Stewart KR, Krueger F, Maupetit-Méhouas S, Arnaud P, Tomizawa S-I, Andrews S, Kelsey G. 2015. Deep sequencing and de novo assembly of the mouse oocyte transcriptome define the contribution of transcription to the DNA methylation landscape. *Genome Biology* **16**: 209.

Vincent M, Mundbjerg K, Skou Pedersen J, Liang G, Jones PA, Ørntoft TF, Dalsgaard Sørensen K, Wiuf C. 2017. epiG: statistical inference and profiling of DNA methylation from whole-genome bisulfite sequencing data. *Genome Biology* **18**: 38.

Waddington CH. 1942. Canalization of development and the inheritance of acquired characters. *Nature* **150**: 563–565.

Walsh CP, Chaillet JR, Bestor TH. 1998. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat Genet* **20**: 116–117.

Wang C, Liu X, Gao Y, Yang L, Li C, Liu W, Chen C, Kou X, Zhao Y, Chen J, et al. 2018. Reprogramming of H3K9me3-dependent heterochromatin during mammalian embryo development. *Nat Cell Biol* **20**: 620–631.

Wang L, Zhang J, Duan J, Gao X, Zhu W, Lu X, Yang L, Zhang J, Li G, Ci W, et al. 2014. Programming and inheritance of parental DNA methylomes in mammals. *Cell* **15**: 979–991.

Wang X, Clark AG. 2014. Using next-generation RNA sequencing to identify imprinted genes. *Heredity (Edinb)* **113**: 156–166.

Weaver JR, Bartolomei MS. 2014. Chromatin regulators of genomic imprinting. *Biochim Biophys Acta* **1839**: 169–177.

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

Webster KE, O'Bryan MK, Fletcher S, Crewther PE, Aapola U, Craig J, Harrison DK, Aung H, Phutikanit N, Lyle R, et al. 2005. Meiotic and epigenetic defects in Dnmt3L-knockout mouse spermatogenesis. *Proceedings of the National Academy of Sciences* **102**: 4068–4073.

Weinberg DN, Papillon-Cavanagh S, Chen H, Yue Y, Chen X, Rajagopalan KN, Horth C, McGuire JT, Xu X, Nikbakht H, et al. 2019. The histone mark H3K36me2 recruits DNMT3A and shapes the intergenic DNA methylation landscape. *Nature* **573**: 281–286.

White CR, MacDonald WA, Mann MRW. 2016. Conservation of DNA Methylation Programming Between Mouse and Human Gametes and Preimplantation Embryos. *Biology of Reproduction* **95**: 61.

Wilkins JF, Haig D. 2003. What good is genomic imprinting: the function of parent-specific gene expression. *Nat Rev Genet* **4**: 359–368.

Wittkopp PJ, Haerum BK, Clark AG. 2004. Evolutionary changes in cis and trans gene regulation. *Nature* **430**: 85–88.

Wolf G, Yang P, Füchtbauer AC, Füchtbauer E-M, Silva AM, Park C, Wu W, Nielsen AL, Pedersen FS, Macfarlan TS. 2015. The KRAB zinc finger protein ZFP809 is required to initiate epigenetic silencing of endogenous retroviruses. *Genes & Development* **29**: 538–554.

Wossidlo M, Arand J, Sebastiano V, Lepikhov K, Boiani M, Reinhardt R, Schöler H, Walter J. 2010. Dynamic link of DNA demethylation, DNA strand breaks and repair in mouse zygotes. *EMBO J* **29**: 1877–1888.

Wu J, Huang B, Chen H, Yin Q, Liu Y, Xiang Y, Zhang B, Liu B, Wang Q, Xia W, et al. 2016. The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature* **534**: 652–657.

Xu GL, Bestor TH, Bourc'his D, Hsieh CL, Tommerup N, Bugge M, Hulten M, Qu XY, Russo JJ, Viegas-Pequignot E. 1999. Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. *Nature* **402**: 187–191.

Xu Q, Xiang Y, Wang Q, Wang L, Brind'Amour J, Bogutz AB, Zhang Y, Zhang B, Yu G, Xia W, et al. 2019. SETD2 regulates the maternal epigenome, genomic imprinting and embryonic development. *Nat Genet* **51**: 844–856.

Yamaguchi K, Hada M, Fukuda Y, Inoue E, Makino Y, Katou Y, Shirahige K, Okada Y. 2018. Re-evaluating the localization of sperm-retained histones revealed the modification-dependent accumulation in specific genome regions. *Cell Rep* **23**: 3920–3932.

Yang P, Wang Y, Hoang D, Tinkham M, Patel A, Sun M-A, Wolf G, Baker M, Chien H-C, Lai K-YN, et al. 2017. A placental growth factor is silenced in mouse embryos by the zinc finger protein ZFP568. *Science* **356**: 757–759.

Yin Y, Morgunova E, Jolma A, Kaasinen E, Sahu B, Khund-Sayeed S, Das PK, Kivioja T, Dave K, Zhong F, et al. 2017. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**: eaaj2239.

Younesy H, Möller T, Heravi-Moussavi A, Cheng JB, Costello JF, Lorincz MC, Karimi MM, Jones SJM. 2014. ALEA: a toolbox for allele-specific epigenomics analysis. *Bioinformatics* **30**: 1172–1174.

Younesy H, Möller T, Lorincz MC, Karimi MM, Jones SJM. 2015. VisRseq: R-based visual framework for analysis of sequencing data. *BMC Bioinformatics* **16 Suppl 11**: S2.

Zemach A, McDaniel IE, Silva P, Zilberman D. 2010. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328**: 916–919.

Zhang B, Zheng H, Huang B, Li W, Xiang Y, Peng X, Ming J, Wu X, Zhang Y, Xu Q, et al. 2016. Allelic reprogramming of the histone modification H3K4me3 in early mammalian development. *Nature* **537**: 553–557.

Zheng H, Huang B, Zhang B, Xiang Y, Du Z, Xu Q, Li Y, Wang Q, Ma J, Peng X, et al. 2016. Resetting epigenetic memory by reprogramming of histone modifications in mammals. *Molecular Cell* **63**: 1066–1079.

Zhu P, Guo H, Ren Y, Hou Y, Dong J, Li R, Lian Y, Fan X, Hu B, Gao Y, et al. 2018. Single-cell DNA methylome sequencing of human preimplantation embryos. *Nat Genet* **50**: 12–19.