GENETIC REGULATION OF EPIGENETIC PROCESSES IN MOUSE: DNA METHYLATION AND X CHROMOSOME INACTIVATION

John Douglas Calaway

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Curriculum in Genetics and Molecular Biology.

Chapel Hill
2014

Approved by:

Fernando Pardo-Manuel de Villena

Corbin Jones

Leonard McMillan

Mihai Niculescu

Terry Magnuson

William Valdar

ABSTRACT

John Douglas Calaway: GENETIC REGULATION OF EPIGENETIC PROCESSES IN MOUSE: DNA METHYLATION AND X CHROMOSOME INACTIVATION
(Under the direction of Fernando Pardo-Manuel de Villena)

Epigenetics is the study of inheritance not encoded by primary DNA sequence. In mammals, epigenetic processes are required for proper development, gene regulation, chromosome function (*e.g.,* X-chromosome inactivation (XCI)), and genome stability. Misregulation of epigenetic processes is typically a hallmark of disease. Epigenetic marks vary depending on genomic position, cell type, environment, time, sex, and even between individuals within a population. Genetic variation is one source of epigenetic variability that has only recently been appreciated. It is unknown how prevalent and to what extent underlying genetic variation influences epigenetic variability, and furthermore, how this epigenetic variability contributes to phenotypic variation within a population. It has been postulated that epigenetic variation between individuals may help solve the 'missing heritability' problem.

In an attempt to address these questions and further characterize the influence of genetics on epigenetics, I demonstrate that DNA sequence variation in *cis* affects two epigenetic processes, DNA methylation and XCI. In the first section, I performed a genome- wide allele-specific methylation survey in the mouse brain to show widespread loci that influence nearby DNA methylation at CpGs. These differentially methylated CpGs tend to reside near transcription start sites and may serve a functional role. We estimate that there are roughly 13,000 of these loci genome-wide. Additionally, I show

that these strain-specific *cis*-acting loci also influence a parent-of-origin differentially methylated region in the 3'UTR of the *Actn1* gene, which suggests that genetic variation might also influence highly conserved imprinted regions as well.

In the second section, I mapped a *cis*-acting locus called the *X-chromosome controlling element* (*Xce*) that influences XCI choice in mouse. I reduced the *Xce* candidate interval to a 176 kb region located approximately 500 kb proximal to *Xist*. I extensively characterized the genetic architecture of the new candidate interval in over 300 inbred and wild-caught mice. I conclude that each mouse taxa examined has a different functional *Xce* allele and there is no sharing. I identified two new *Xce* alleles (Xce^{e} and Xce^{f}) that bring the number to six functional alleles in *Mus*. I propose that structural variation of segmental duplications within this interval explains the presence of multiple functional *Xce* alleles.

Overall these results provide new insights into the genetic regulation of epigenetic processes in mouse. Furthermore, this work creates a foundation for future work to untangle the molecular mechanisms behind differential DNA methylation and X-chromosome inactivation choice.

To my grandmother, Wilma Geraldine Street

ACKNOWLEDGMENTS

I would like to thank:

My thesis advisor Fernando Pardo-Manuel de Villena for his unyielding support, generosity and guidance throughout the course of my doctoral research

My family and friends for their love and support

My thesis committee for their thoughtful guidance and encouragement

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BAC	BACTERIAL ARTIFICIAL CHROMOSOME
BP	BASE PAIR(S)
CNV	COPY NUMBER VARIATION
CpG	CYTOSINE PHOSPHODIESTER GUANINE
DMR	DIFFERENTIALLY METHYLATED REGION
µg	MICROGRAM
MDA	MOUSE DIVERSITY ARRAY
MSNP	METHYLATION-SENSITIVE SINGLE NUCLEOTIDE POLYMORPHISM ANALYSIS
MS-RFLP	METHYLATION-SENSITIVE RESTRICTION FRAGMENT LENGTH POLYMORPHISM
PCA	PRINCIPAL COMPONENT ANALYSIS
PCR	POLYMERASE CHAIN REACTION
PoO	PARENT-OF-ORIGIN
eQTL	EXPRESSED QUANTITATIVE TRAIT LOCUS
SD	SEGMENTAL DUPLICATION
SDP	STRAIN DISTRIBUTION PATTERN
SNP	SINGLE NUCLEOTIDE POLYMORPHISM
eSNP	EXPRESSED SINGLE NUCLEOTIDE POLYMORPHISM
UTR	UNTRANSLATED REGION
Xce	*X CHROMOSOME CONTROLLING ELEMENT*
XCI	X CHROMOSOME INACTIVATION
Xist	*X-INACTIVE SPECIFIC TRANSCRIT*

CHAPTER I: INTRODUCTION

Epigenetics is the study of inheritance that is not encoded by primary DNA sequence. It encompasses a wide-range of biological processes and involves a diverse set of molecular players that include DNA methylation, histone post-translational modification, microRNA, long non-coding RNA, and prions [5-8]. In mammals, epigenetics is required for a single cell to give rise to a complex, multicellular organism through dynamic regulation of gene expression. Furthermore, epigenetics has additional roles in determining and maintaining the functional architecture of genomes and chromosomes (*e.g.,* X chromosome inactivation (XCI)) [9, 10]. Accordingly, aberrant epigenetic regulation is associated with many human diseases, including cancer [8].

Epigenetics is context specific and varies depending on genomic position, cell type, environment, time, sex, and between individuals within a population. This poses experimental challenges and technological obstacles that have severely hampered its study. Yet despite these difficulties, significant progress has been made to characterize these different sources of epigenetic variability and determine how they impact phenotype. To date, however, the majority of work has focused on characterizing epigenetics within different functional regions of the genome (*i.e.,* promoters and genes) [11]. More recently, studies have been published that attempt to catalog cell-specific epigenetic differences [12]; tease apart environmental factors that influence epigenetics [13]; and characterize the changing epigenome with age [14, 15]. Over the past three decades, epigenetics has

impacted most major fields of biology including practical applications in both medicine and agriculture [16-19].

The genetic regulation of epigenetics

Like two sides of a coin, the functional role of mammalian genomic DNA is incomplete without also considering its epigenetic component and *vice versa*. Both DNA and epigenetic players act in concert to bring about the complex genic regulation required for proper development and homeostasis [7]. Thus it stands to reason that changes in one may elicit a change in the other. Mutations in DNA sequence encoding epigenetic machinery have drastic and obvious effects, for example, deletions of methyl transferases have genome-wide effects on CpG methylation [20-22]. On the other hand, redundant epigenetic mechanisms are in place to maintain the structural integrity of the genome [23, 24]. Without such mechanisms, transposable elements would proliferate unchecked [25]. In fact, these epigenetic mechanisms guard the genome during chromosome segregation, recombination, and double strand break repair and are key for faithful transmission of the genome from one generation to the next [26]. Although these examples provide evidence of the critical link between DNA sequence and epigenetics, they do not provide insight into how changes in underlying DNA sequence influence epigenetics. In other words, understanding the role genetic variation plays in epigenetic variability within a population. In its broadest sense, this work aims to investigate how DNA sequence variability impacts epigenetics by means of two epigenetic processes: DNA methylation and XCI. Some important questions for the genetic regulation of epigenetics are: To what extent does genetic variability within a population influence epigenetics and impact epigenetic variability? And ultimately, what effect does epigenetic variability have on phenotypic variability?

In the first section (**Chapters II and III**), I demonstrate that differentially methylated regions (DMRs) are determined by genetic variation in *cis*. The DMRs are found genome-wide in both genic and intergenic regions. These results suggest that the impact of local

sequence variation on DNA methylation in the mouse is pervasive with an estimated 13,000 differentially methylation CpGs genome-wide. Furthermore, there is an enrichment of DMRs found near transcription start sites that may indicate a functional role in differential gene expression. Based on my results using two inbred mouse lines (129S1/SvlmJ and PWK/PhJ) that capture only a small fraction of the genetic diversity in *Mus musculus* (~1/3). I conclude that local DNA sequence variation contributes to substantial genome-wide DNA methylation variation in mouse than previously thought [27]. However, it is clear that the functional significance of these DMRs needs to be determined before conclusions can be drawn regarding their contribution to true phenotype variation. I further demonstrate that local genetic variation affects differential methylation at a maternally methylated region in the 3'UTR of *Actn1* in mouse. I show that parent-of-origin DMRs are influenced by local DNA sequence and speculate that genetic variation within DMRs at imprinting control regions may in fact alter expression at imprinted regions.

In the second section (**Chapters IV and V**), I examine the genetic regulation of XCI choice by mapping and characterizing a *cis*-acting locus called *the X-chromosome controlling element* (*Xce*). By using a combination of historical phenotyping data and new mouse genetic resources [1, 2], I narrowed the *Xce* candidate interval 10-fold to a region that lies 500 kb proximal to *Xist,* thereby excluding *Xite, Xist,* and *Tsix* as *Xce* candidates.

It is thought that *Xce* serves as a *trans*-factor binding site that determines which X chromosome will undergo XCI, a multistep epigenetic process that functionally inactivates an entire X chromosome [28]. I show that the new *Xce* candidate interval contains a series of segmental duplications and an inversion in the C57BL/6J reference assembly. I postulate that the different functional alleles of *Xce* in *Mus* can be explained by structural variation within the segmentally duplicated regions.

Furthermore, I investigated the genetic architecture of the new *Xce* candidate interval in over 300 individual mice, including classical and wild-derived inbred and wild-caught

mice. I conclude that each species or subspecies of mouse appears to have its own functional *Xce* allele. XCI skewing is common in the laboratory mouse, and the degree of XCI skewing, determined by genetics alone, might reach complete skewing in favor of one X chromosome over another.

CHAPTER II: GENOME-WIDE DIFFERENTIAL METHYLATION PATTERNS IN INTERSUBSPECIFIC HYBRID MICE[1]

BACKGROUND AND INTRODUCTION

DNA methylation is an epigenetic process that covalently binds a methyl group to a DNA base [30]. At the molecular level, methylated DNA may affect the binding of cellular machinery [30, 31], influence the positioning of nucleosomes [32], and even change the shape of DNA itself [33]. There are examples of DNA methylation utilized in all three major branches of life [34]. Mammalian DNA methylation is predominantly 5-methylcytosine within the context of a CpG dinucleotide motif (mCpG). 5-methylcytosine does exist outside of the CpG motif, albeit limited and primarily restricted to early development [35]. mCpG is a key epigenetic mark that plays a critical role in development and cell differentiation [32, 36], XCI [31, 37], tr[1]ansposon silencing [38], tumorigenesis [39], and overall genomic stability [40]. As with other critical epigenetic marks, disruption of normal methylation patterns has severe phenotypic consequences [41-43].

[1] The following chapter describes work done in collaboration with Dr. Hyuna Yang, Dr. Elena de la Casa-Esperon, Dr. David L. Aylor, Dr. Leonard McMillan, Dr. Gary A. Churchill, and Dr. Fernando Pardo-Manuel de Villena. I significantly contributed to the sample preparation, design and implementation of molecular phenotyping assays, and data analysis. The expression data used to examine DMR functionality is from a previously reported study [29]. I also significantly contributed to the manuscript preparation including writing and figure design.

DNA methylation is variable

DNA methylation is dynamic and varies widely during development with global changes in methylation during the pre-implantation stages [44, 45]. Changes in DNA methylation have also been associated with aging [14, 46]. Furthermore, multiple environmental stimuli might also lead to changes in mCpG and there is a growing interest in the intersection of epigenetics and toxicology [13, 46]. It is well established that local variation in methylation along chromosomes plays a significant role in functional regulation of gene expression [47-49]. In some cases, such is XCI, gene bodies are hypermethylated on the active X chromosome while promoter elements are hypermethylated on the inactive X chromosome [50, 51]. Methylation is also strongly associated with genomic imprinting and the methylation status at the same CpG in two homologous chromosomes from a single cell may vary depending on their parent-of-origin [52].

Local DNA effects on cytosine methylation

Variation in mCpG is observed among individuals from a population [27, 53]. Studies of monozygotic and dizygotic twins reveal that genetic variation is a major driver of mCpG variation among individuals [15, 54-56], although there is contention of whether DNA sequence or environmental factors play a larger role [57]. In a previous study by Schilling and coauthors [27], C57BL/6J and BALB/cJ reciprocal F1 hybrid mice were used to detect allele-specific methylation. Importantly, allele-specific methylation in F1 hybrids requires *cis*-acting DNA sequence or epigenetic differences (**Figure 2-1**). The authors discovered that differentially methylated regions (DMRs) are primarily genetic-driven (strain effect) and act in *cis*. Furthermore, functional analysis demonstrated that strain-specific DMRs influence nearby gene expression levels. In fact, their results indicate that inter-individual variation in epigenetic marks may contribute to phenotypic variation and also help explain missing heritability [58]. Ultimately, the identification and functional annotation of variable methylation that is heritable would be important to understand the evolution of epigenetics

[59, 60]. And yet, despite the increased interest in population variation of DNA CpG methylation, the genetic regulation of epigenetics remains in its infancy.

This study aims to address some basic questions including: How prevalent is the role that local genetic variation plays in variation of mCpG; how much mCpG variation can be ascribed to genetic variation (strain effects) versus imprinting (parent-of-origin effects); and is methylation variable depending on sex?

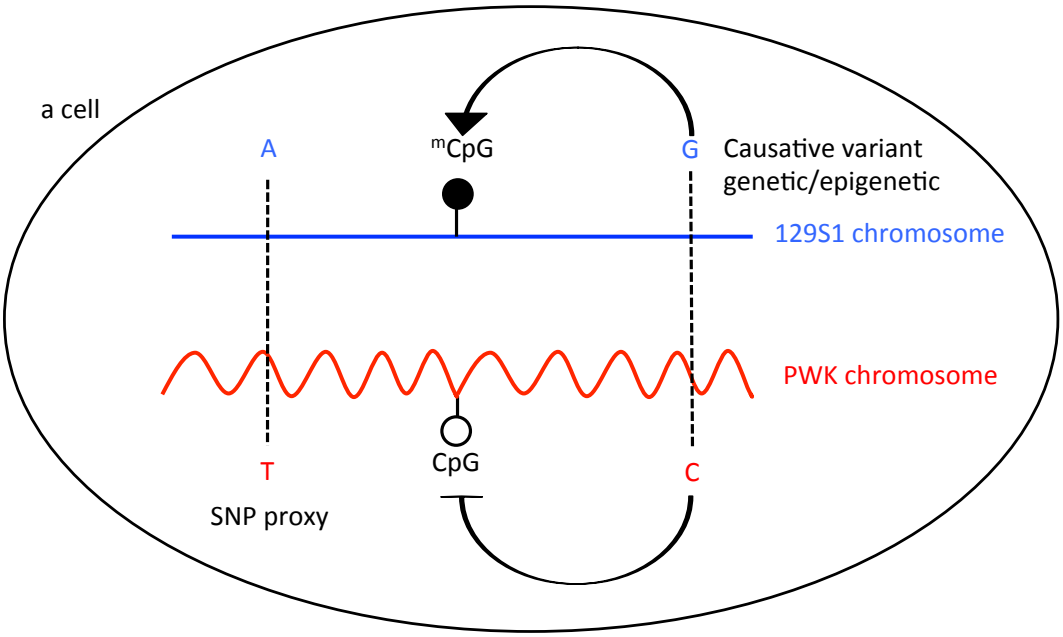


Figure 2-1. F1 hybrids as a tool for discovering *cis*-acting variants that direct DNA methylation. Shown is a cell from an F1 hybrid between mouse strains 129S1/SvImJ and PWK/PhJ. Both parental chromosomes (blue, 129S1 and red, PWK) are exposed to the same cellular environment that includes *trans*-factors. Therefore, differential methylation requires a *cis* causative variant (genetic or epigenetic) that distinguishes the two chromosomes from one another.

Experimental Design

Any study aimed to answer these questions will confront several early decisions that might impact the conclusions reached including: the platform used to examine mCpG, the selection of organism, experimental design, and tissue(s) and/or developmental stage. In the following paragraphs we briefly outline the rationale behind our choices.

Ideally, one would like to select a platform that estimates quantitative methylation at individual CpGs within a large dynamic range in an allele-specific manner. This latter requirement is critical to partition methylation levels according the genotype and the parent-of-origin that requires the presence of a closely linked informative variant to the corresponding CpG. The platform would interrogate as many CpGs as possible evenly distributed across the genome, and residing within or near functional elements. Lastly, the platform would be cost effective to allow the analysis of many biological replicates. A technique was developed that satisfies many of these requirements called Methylation-sensitive Single Nucleotide Polymorphism analysis (MSNP) [53, 61]. MSNP was described and applied to Affymetrix human genotyping arrays that use endonucleases to fragment genomic DNA followed by PCR amplification of those fragments to create a genomic library. MSNP exploits the amplification step by first introducing a methylation-sensitive endonuclease digestion. To determine allele-specific methylation, this method compares buffer treated samples to *Hpa*II treated samples and looks for SNPs that shift from heterozygosity to homozygosity in favor of the methylated allele [61]. Additionally, *Msp*I digestion is used a positive control. Our laboratory has developed a high-density Affymetrix genotyping array for the mouse, the Mouse Diversity Array (MDA) [62]. The MDA is particularly well suited to extend MSNP to the mouse because of the high density and uniform distribution of SNP probes sets (>600K) that target the genetic variation of laboratory mouse stocks [1]. MDA has the added value of the presence of over 900K exon probes that cover known transcribed regions of the genome. Although we cannot link allele-

specific information to these probes, we can estimate the level of mCpG associated with these genomic targets.

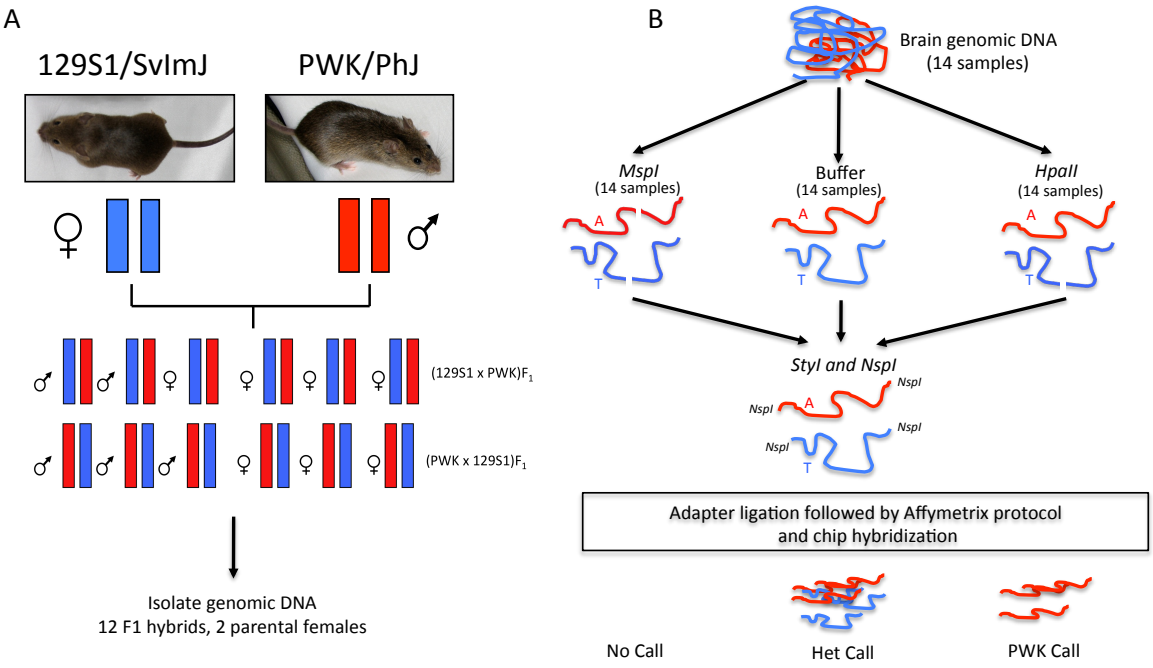


Figure 2-2. Experimental design. Panel A shows the breeding scheme to generate the mice used in this study. Blue bars and red bars represent 129S1 and PWK genomes, respectively. The order of the colored bars in the F1 progeny denotes the parent-of-origin, maternal on the left and paternal on the right. The sex of the progeny is listed to the left of the colored bars. **Panel B** shows each MSNP experimental condition and the predicted outcome. The example shown is a PWK-specific methylated CpG near an A/T SNP.

Mouse as a model for the genetic regulation of cytosine methylation

Given our focus on mammalian epigenetics, the mouse is an obvious model organism. An often-cited advantage of the laboratory mouse is the availability of a large collection of well-characterized inbred strains with a wide range of genetic diversity in pair wise comparisons. By crossing two inbred strains, researchers can generate the desired number of reciprocal F1 hybrids that for the autosomes only differ in the parental origin of each pair of homologues. The ability of replicate genomes can be combined with

environmental control to facilitate the characterization of the genetic contribution to mCpG variability. Reciprocal F1 hybrid mice are particularly attractive as experimental subjects because heterozygous SNPs between the two parental strains can be utilized to tag allele-specific methylation and thus identify strain-dependent and parent-of-origin effects. Furthermore, every strain-dependent difference in methylation identified in F1 hybrid mice necessary requires the presence of a local (*cis*) causative variant between parental strains (**Figure 2-1**).

Recent advances in genotyping and whole genome sequencing (WGS) [1, 2, 63] greatly facilitates the selection of parental strains. Using WGS and imputation, one can select the parental strains to have the desired level and distribution of genetic variation. On the other hand, the known MDA genotypes of the parental strains can be leveraged to determine the number and distribution of putatively informative CpG sites using MSNP analysis. Finally, these studies can be placed in the desired evolutionary context thanks to the recent assignment of every genomic region of every laboratory strain to one of the three major subspecies of the house mouse [1].

Based on these data, we selected the 129S1/SvImJ and PWK/PhJ strains to generate reciprocal F1 hybrids. These parental strains have been sequenced [2, 64], are highly genetically divergent genome-wide, fully inbred, readily available, and easy to breed. They are also of interest to the wider scientific community because of their common use and their inclusion in new mouse resources such as the Collaborative Cross (CC) and Diversity Outbred (DO) populations [64, 65]. It is important to note that 129S1/SvImJ is mostly derived from *Mus musculus domesticus* while PWK/PhJ is mostly derived from *Mus musculus musculus* [1]. Although the F1 hybrids used in this study are unlike most mice found in natural populations, they provide an excellent platform to determine the effect of genetic diversity on epigenetic variation. It is also possible to take advantage of the fraction of the genome that originates from the minority subspecies in each parental strain to

estimate the effect of genetic variation on methylation variation within a species**. Figure 2-2** provides detailed information about the experimental design. In contrast with the constancy of genotype within an individual, the epigenome will vary depending on the tissue analyzed. Here we investigate allele-specific methylation in the adult mouse brain. We chose this tissue because imprinting is common in the mouse brain and thus can serve as a positive control for the identification of allele-specific methylation [66]. The fact that the brain is a heterogeneous tissue poses some challenges but has also advantages. Among the advantages is the fact that by using a heterogeneous tissue, we will increase the number of potential allele-specific and parent-of-origin CpG sites. On the negative side we may be unable to detect cell-type specific effects because of noise. Ideally, we would like to investigate allele-specific methylation in every cell type within the brain, but the current cost would be prohibitive.

Here we demonstrate that MSNP analysis with the MDA platform is an effective method for surveying genome-wide allele-specific mCpG. We identified overall differences in DNA methylation between sexes. We discovered that strain-specific DNA methylation is far more pervasive that parent-of-origin DNA methylation genome-wide.

RESULTS

Global MSNP analysis

A brief description of the 14 biological samples and 42 Affymetrix arrays is provided in **Figure 2-2**. We classified SNP and exon probes as informative for CpG methylation by either stringent (CCGGI probes) or liberal criteria (CCGGI + CCGGII probes) (see **Materials and Methods**). According to the liberal criteria, of the 623,054 SNP probes and 597,245 exon probes on the MDA, we identified 340,828 SNP and 465,921 exon probes as having one or more *Msp*I restriction sites internal to its corresponding Affymetrix amplification fragment (**Figure 2-2 and Figure 2-3**).

For each probe, we defined the possible methylation status of nearby CpGs based on intensity differences between *Hpa*II, *Msp*I and buffer treated samples. We classified each SNP and exon probe sets as partially methylated, fully methylated, and unmethylated (see **Materials and Methods**). In addition, 309,278 SNP and 418,769 exon probes sets cannot be classified in any of these three categories and are ignored in subsequent analyses (denoted as unclassified in **Figure 2-3**). We reported previously that off-target variants within MDA probe binding sites affect hybrization performance [67]. To determine if off-target variants influence our ability to classify SNP and exon probe sets, we compared the number of off-target variants between the four probe classes (**Table S2-1**). We found a significant enrichment of probes with off-target variants in the unclassified methylation class. This result suggests that our ability to assign methylation state to SNP and exon probes is hindered by off-target variation in the probe binding site.

Among the 31,550 SNP and 47,152 exon probe sets that can be classified, 7,613 SNP and 38,185 exon probes are associated with fully methylated CpG(s), 3,337 SNP and 8,260 exon probes are associated with unmethylated CpG(s), and 497 SNP and 707 exon probes are associated with partially methylated CpG(s). The methylation state associated

with SNP and exon probes was mapped back to the genome to create a global methylation map (liberal analysis, **Figure 2-4**; stringent analysis, **Figure S2-1**).

We then tested whether there are global sex-dependent methylation differences. Because there were four (129S1xPWK)F1 female samples, but only two (129S1xPWK)F1 male samples, we randomly chose two out of the four females to have a balanced sample size between the two sexes. Obviously, when female and male X chromosomes were analyzed separately, we observed striking differences in the degree of X chromosome methylation (**Figure 2-4 and Figure S2-1**). While sex differences in X chromosome methylation are well documented [51, 68], the existence and direction of autosomal differences is controversial. Therefore, we compared the number of methylated, unmethylated, and partially methylated probes in male and female autosomes. Overall, females have approximately 9% higher levels of autosomal methylation (males and females have 41,643 and 45,790 SNP and exon probes associated with fully methylated CpG(s), respectively. *p*-value = 1.55e-27). To examine the differences at a local level, we mapped the methylation state associated with each probe set back to the genome to create sex-specific methylation maps (**Figure S2-2**). Although sex specific maps are similar, there are areas of apparent sex-specific methylation. For example, proximal chromosome 15 is heavily methylated in females, but not males.

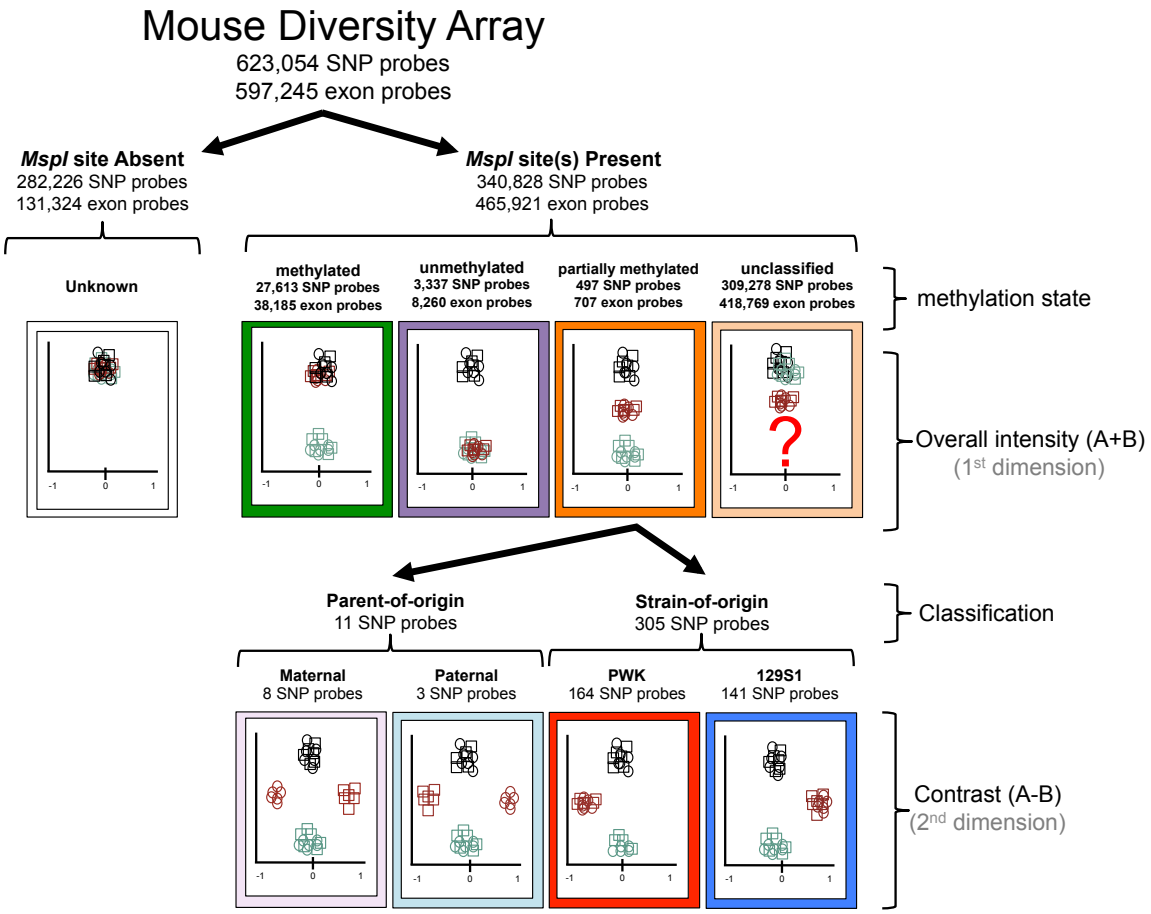


Figure 2-3. Classification of MDA probes according to the CpG methylation status of tagged *Hpa*II sites. This Figure shows the method used to classify the MDA SNP and exon probe sets into methylation-informative subsets. The number of probe sets in each classification was determined by liberal analysis (CCGGI, see **Materials and Methods**). Under each classification is a simulated two-dimensional plot to illustrate the hybridization pattern expected for each class. The y-axis represents the overall intensity and is used throughout the entire analysis. The x-axis represent the contrast between the two allelic probes at heterozygous SNPs and is used only in the strain and parent-of-origin classes. Each plot show hybridization data for a single probe set for the 14 experimental samples used in this study. Each sample is represented three times according to whether they were subject to *Hpa*II (maroon), *Msp*I (light green) and buffer (black) treatments. Circles and squares represent the two types of reciprocal F1 hybrids. We assigned seven different colors to seven methylation classes with consistent results. Consistently methylated CpGs class is green, consistently unmethylated CpGs class is purple, partially methylated CpGs class is orange, PWK-specific methylation class is red, 129S1 specific methylation class is blue, maternal specific methylation class is pink and paternal specific methylation is light blue. These colors are used throughout the chapter.

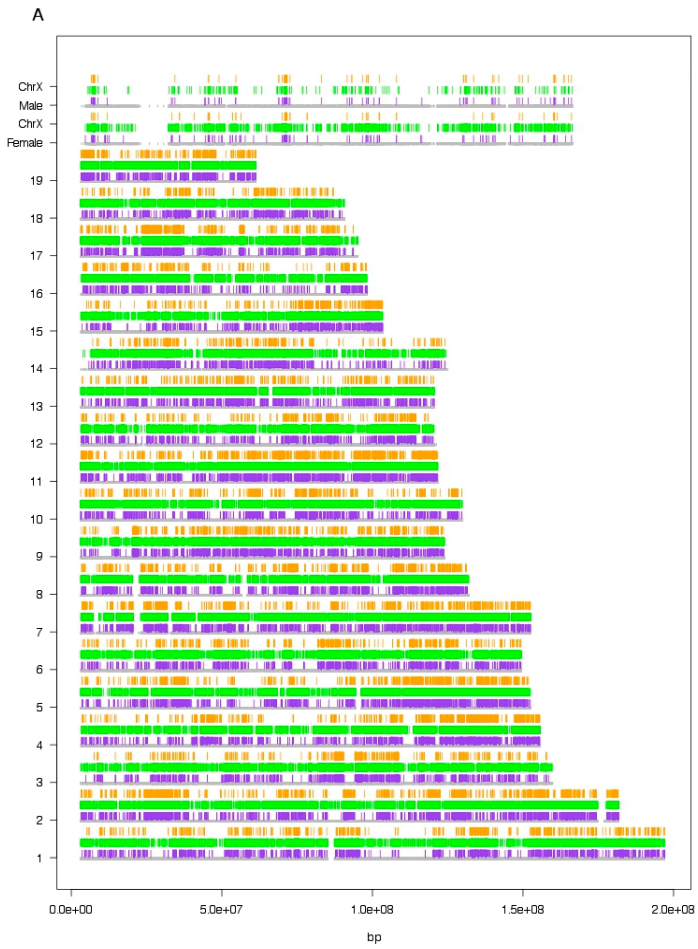


Figure 2-4. Global and allele-specific maps of methylation patterns in the mouse brain. Each tick mark represents a probe set located at its corresponding chromosome position. SNP and exon probes associated with fully methylated CpGs are shown in green, probe sets associated with unmethylated CpGs are shown in purple, and probe sets associated with partially methylated CpGs are shown in orange. Results for X chromosome probe sets in males and females are plotted separately. All probe sets are plotted in light grey.

Allele-specific analysis of mouse brain DNA reveals strain and parent-of-origin differences in methylation.

To determine the extent and localization of allele-specific methylation we restricted our analysis to SNPs that were heterozygous in our F1 hybrids and were classified as partially methylated in the global analysis. These conditions ensure that our analysis focuses on the SNPs with potential strain-specific and parent-of origin information. Of the 497 partially

methylated SNP probes, we identified 305 SNPs displaying a strain dependent methylation patterns and 11 SNPs displaying a parent-of-origin dependent methylation patterns (**Figure 2-3 and Table S2-2**). We mapped both classes of probes back to the genome to create strain-of-origin and parent-of-origin global differential methylation maps (**Figure S2-3**). One hundred eighty-one partially methylated sites do not conform to this simple strain and parental origin partition.

	strain			parent-of-origin		
	129S1 (CCGGI)	**PWK (CCGGI)**	**Total**	**maternal (CCGGI)**	**paternal (CCGGI)**	**Total**
Chr1	9 (1)	7 (0)	16 (1)	0 (0)	1 (0)	1 (0)
Chr2	16 (5)	11 (2)	27 (7)	4 (0)	0 (0)	4 (0)
Chr3	10 (1)	6 (0)	16 (1)	0 (0)	0 (0)	0 (0)
Chr4	4 (1)	20 (3)	24 (4)	0 (0)	0 (0)	0 (0)
Chr5	8 (2)	10 (2)	18 (4)	0 (0)	0 (0)	0 (0)
Chr6	5 (0)	9 (0)	14 (0)	0 (0)	0 (0)	0 (0)
Chr7	8 (2)	11 (1)	19 (3)	1 (1)	1 (0)	2 (1)
Chr8	10 (1)	8 (3)	18 (4)	0 (0)	0 (0)	0 (0)
Chr9	8 (3)	6 (0)	14 (3)	0 (0)	0 (0)	0 (0)
Chr10	3 (0)	7 (0)	10 (0)	0 (0)	0 (0)	0 (0)
Chr11	11 (0)	13 (1)	24 (1)	0 (0)	0 (0)	0 (0)
Chr12	4 (0)	10 (3)	14 (3)	2 (0)	0 (0)	2 (0)
Chr13	5 (1)	8 (2)	13 (3)	0 (0)	0 (0)	0 (0)
Chr14	4 (0)	5 (2)	9 (4)	0 (0)	0 (0)	0 (0)
Chr15	4 (0)	4 (0)	8 (0)	0 (0)	1 (0)	1 (0)
Chr16	6 (2)	5 (1)	11 (3)	0 (0)	0 (0)	0 (0)
Chr17	6 (2)	7 (0)	13 (2)	1 (0)	0 (0)	1 (0)
Chr18	12 (3)	7 (1)	19 (4)	0 (0)	0 (0)	0 (0)
Chr19	8 (1)	10 (3)	18 (4)	0 (0)	0 (0)	0 (0)
ChrX	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Total	141 (25)	164 (24)	305 (49)	8 (1)	3 (0)	11 (1)

Table 2-1. Allele-specific DMRs. The table shows the breakdown of allele-specific DMRs into strain, parent-of-origin, and number per chromosome. The numbers inside parenthesis are from the stringent analysis (CCGGI).

Approximately half of the 305 strain dependent sites are associated with consistent methylation of the allele of one of the parental strains while the other allele is consistently unmethylated (*i.e.,* 164 sites with consistent methylation of the PWK allele and 141 sites with consistent methylation of the 129S1 allele). As expected there is a significant deficit of strain dependent sites on the X chromosome (see **Discussion**). Both types of strain effects

are represented in each autosome and the distribution is uniform. To our knowledge, the strain-dependent methylation loci have not been reported before.

The parent-of-origin analysis identified SNPs tagging either maternal or paternal-specific methylation (eight and three SNPs, respectively; **Table S2-2**). All but two of the SNPs associated with parent-of-origin methylation are located within or near (less 500 kb) known clusters of imprinted genes [69-77]. The remaining two SNPs (rs32640412 and rs32641208) tag two *Hpa*II sites within the same Affymetrix amplicon on chromosome 12 (**Figure S2-3**). Although the allele-specific information is not independent, the methylation status of the two *Hpa*II sites must be consistent given the observed parent-of-origin effect. These CpGs are maternally methylated and in the last intron of *Actn1* that codes for a-Actinin, a microfilament protein that interacts dynamically with Actin (see **Chapter III**). Note that the methylation statuses of the relevant *Hpa*II sites have not been reported previously. For example, the imprinted methylation status of the site associated with SNP rs31991512 on chromosome 7 was previously unknown.

Functional relevance of strain-specific DMRs

DNA methylation may directly or indirectly affect gene expression (*i.e.,* by influence the binding of transcription regulatory elements or by altering the local chromatin landscape) [78-80]. Typically, transcription regulatory elements reside near the transcription start sites (TSS) of genes they control. To explore the possible functional role of the strain-specific DMRs, we began with a simple test to determine if the strain-specific DMRs are significantly closer to gene TSS than expected by chance. For each of the 305 strain-specific DMRs, we calculated the distance to the immediate proximal and distal gene TSS. We then compared the average of these distances to the average distance between all informative SNP probes with at least one *Hpa*II/*Msp*I cut site (**Figure 2-3**). We rejected the null hypothesis that the DMRs are not significantly closer to TSS than the informative SNP average (**Figure 2-5A**, *p*-value = $1.96e^{-44}$).

Next, we tested each of the 610 TSS for differential expression between the two parental strains. We used publically available eQTL data from liver expression of Pre-Collaborative Cross mice (See **Materials and Methods**) [29]. Of the 610 TSS nearby strain-specific DMRs, we found 157 of the TSS are associated with eQTLs. We plotted the frequency and direction of the differential expression and compared that to the frequency of all genes with eQTLs (**Figure 2-5B**). The bimodal distribution of DMRs associated with eQTLs is similar to that of all genes with eQTLs.

Lastly, we determined if the direction of allele-specific expression was correlated to the direction of allele-specific methylation. To do so, we overlaid the methylation data with the eQTL effect to determine if the methylation patterns correlated with the eQTL effect direction (**Figure 2-5C**). Of the 170 DMRs associated with eQTLs, we found all four possible methylation and expression combinations: 43 showed high PWK methylation and high PWK expression, 61 shows high PWK methylation and low PWK expression, 30 showed high 129S1 methylation and low 129S1 expression, and 36 showed high 129S1 methylation and high 129S1 expression (**Figure 2-5D**).

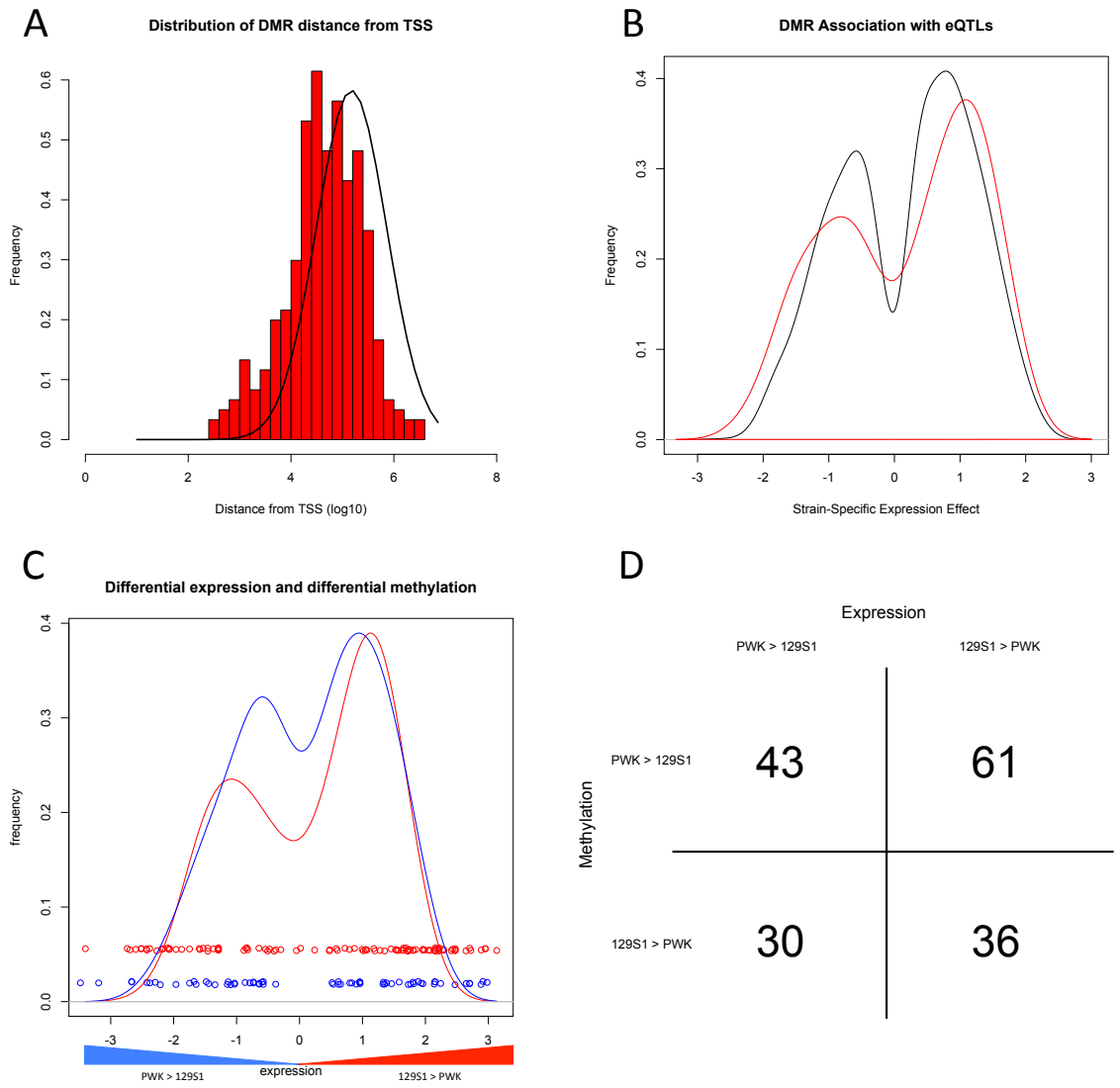


Figure 2-5. Functional analysis of strain-specific DMRs. Panel A shows the distribution of distances (log10(base pairs)) between the 305 strain-specific DMRs and TSSs immediately proximal and distal (red histograms). The superimposed density curve (black) is the distribution of distances between all MDA methylation informative SNPs (between PWK and 129S1) and their respective proximal and distal TSSs. **Panel B** shows the distribution of DMRs associated with eQTLs. The red line shows the distribution of strain-specific DMRs that are associated with differentially expressed genes and whether the gene is upregulated (positive effect) or downregulated (negative effect). The black line shows the distribution of all genes that are differentially expressed. **Panel C** shows the distribution of the direction of differential methylation and the direction of differential expression. Each circle represents a DMR and its color denotes the direction of methylation (blue, 129S1 and red, PWK), while its position on the X axis represents the eQTL effect. **Panel D** shows the relationship between methylation and expression at the 107 DMRs with nearby eQTLs.

DISCUSSION

Allele-specific variation in epigenetic marks, including DNA methylation, is an emerging field of great interest in basic biology and in the animal modeling of human diseases. Here we report an analysis that combines MSNP with a high-density genotyping array in the mouse. We identified tens of thousands of consistently methylated and consistently unmethylated CpG sites distributed across the mouse genome (**Figures 2-3 and 2-4**). Equally important, we were able to examine the relationship between genetic variation and epigenetic variation at hundreds of partially methylated CpGs. This extensive new catalog of methylation status at CpG sites in laboratory mice can be mined to develop assays to survey methylation variation during development, in different genetic backgrounds and environmental conditions and healthy and diseased mice. The catalog of SNP and exon probe sets with methylation information and the CpGs tagged by them can be found on the UNC System Genetics webpage (http://csbio.unc.edu/CCstatus/index.py)

Although we were inspired by previous reports describing MSNP in humans [53, 61], our study has several important differences. First, experimental design takes full advantage of inbred mouse strains to disentangle relative contribution of sex, strain and parental origin in a simple and elegant manner. Second, the high density of the MDA greatly increases the number of surveyable mCpG's with allele-specific information. We have also overcome some of the limitations of previous analyses by determining methylation status based on probe intensities instead of genotype calls. We have shown previously that this approach has significant advantages in phylogenetic analyses, reduction of ascertainment biases and accuracy of haplotype reconstruction [1, 64, 67, 81]. An added advantage of intensity-based analysis is that it can be extended to uninformative (*i.e*., homozygous) SNP probes and to invariant genomic probes such as the exon probes of the MDA.

An attractive feature of identifying allele-specific effects in F1 hybrids is that there is an absolute requirement for strain specific variation in *cis* (*i.e*., local variation is necessary to observed allele-specific effects, **Figure 2-1**). This applies to gene expression as well as to DNA methylation. Thus, the work reported here provides the foundation to a genetic approach to dissect an important epigenetic mark. Briefly, F1 hybrids can be used to identify strain-dependent effects but provide only a rough localization of the *cis* genetic variant causative of mCpG variation. In a second step one can take advantage of new mouse resource populations such as the CC and DO to finely map the genetic variation driving strain specific DMRs [64, 81]. Depending on the density of recombination, it may be possible to localize the *cis* variants, propose molecular mechanisms, and identify sequence motifs.

Despite the success of our modified MSNP approach, there is room for improvement in two key areas: the number and type of enzyme used in the fragmentation steps prior to library preparation and the number and type of methylation-sensitive endonuclease used to estimate methylation levels. With the availability of the mouse reference assembly and whole genome sequence of commonly used laboratory strains [2, 3] a bioinformatics approach could be used to maximize the methylation information collected, including both presence/absence and allele-specific methylation.

Among the unanticipated conclusions reached in this study is the evidence of female-specific global autosomal hypermethylation (**Figure S2-2**). Our results are in conflict with previous reports of hypermethylation in male autosomes [82, 83]. A possible explanation for these discrepancies is the differences in number, location and identity of the CpG sites surveyed. MSNP uses SNP and exons probes to tag the methylation status at one or a few nearby CpGs. Therefore, each probe set provides an independent estimate of CpG methylation at specific locations of the genome. Overall, MDA-based MSNP targets simultaneously thousands of localized CpGs distributed across the genome. In contrast

previous studies examine either only a few loci [82] or global methylation determined by Southern blot analysis after digestion with *Hpa*II [84]. We believe that MSNP analysis better reflects sex-specific differences in autosomal mCpG methylation at many localized regions of the genome. Whether these sex-specific methylation differences can be generalized to other conditions (for example other tissues, backgrounds and species) and whether they contribute to explain sex-specific phenotypic differences are open questions. Nevertheless, these findings highlight the importance of including both sexes in epigenetic studies.

We observe a deficit of X-linked strain-specific methylation. We expected such results because only females can have informative SNPs on the X chromosome and therefore, provide allele-specific information in our experimental design. In addition, we identify strain-specific methylation by consistent hypermethylation of one allele and hypomethylation of the other. However, mCpG plays a significant role in maintaining X inactivation and therefore our analysis of X-linked mCpG is dependent on the X inactivation status of a female [85]. Females are expected to have equal number of cells with an inactive maternal or paternal X chromosome. However, genetic (*Xce* genotype), parent-of-origin and stochastic factors can contribute to X inactivation skewing and this process is highly variable in mouse (see **Chapter IV**). Although the genetic component (*Xce*) in these reciprocal crosses between 129S1 (Xce^a) and PWK (Xce^e, see **Chapter IV**) should lead to minimal mean XCI skewing, stochastic variation and parent-of-origin effects can contribute to large variability within genetically identical mice and significantly mask strain effects.

In this study, we identified 600 partially methylated CpGs. In our analysis, "partially methylated" denotes CpGs that have intermediate levels of methylation that are consistently observed in F1 hybrids. The number of loci subject to strain dependent effects is an order of magnitude larger than the loci subject to parent-of-origin effects. Of the 305 strain-specific DMRs, 164 are PWK-specific and 141 are 129S1-specific. This symmetry is expected under neutral, positive and negative selection scenarios. A strong asymmetry would require

consistent selection over long evolutionary periods for hypermethylation in one lineage and hypomethylation on the other lineage at many independent and uniformly distributed loci in the genome: an unlikely scenario. On the other hand, equal contribution to the DMRs of two inbred strains with highly different divergence from the mouse genome reference is in sharp contrast with the biases in strain effects found in microarray gene expression studies in highly divergent mouse strains [86]. The differential expression biases are due to the presence of genetic variants in the probes that preferentially reduce hybridization intensity in the most divergent strain [67]. We avoided such artifacts by limiting our allele-specific analysis to well performing probes that lack off target polymorphisms in the two parental strains [67].

An important use of our data is that they can be used to estimate the total number of CpGs that are subject to strain-specific methylation. Using the stringent criteria we identified 49 strain-specific autosomal differentially methylated CpGs. Given the limitations of the MSNP method, we surveyed only 4.9% of all *Hpa*II cut sites or 0.37% of all CpGs genome wide. Extrapolating from our results, we estimate that there are ~13,000 strain-specific differentially methylated CpGs. We acknowledge that there are obvious limitations in our approach including the fact that only two inbred strains have been surveyed (and more importantly only two of the three major house mouse subspecies) and that we require a high threshold to declare a CpG site subject to strain specific effects. We also have only analyzed a single tissue and developmental time. Nonetheless this estimate is remarkable because it predicts that strain effects on epigenetic variation is as prevalent as strain effects in gene expression even after accounting for the high correlation in methylation status between consecutive CpGs expected (**Figure S2-3**).

We tested the functional relevance of the strain-specific DMRs by using three different analyses. In mouse, CpG islands typically reside near TSS of genes they control [87, 88]. We therefore first determined the distribution of distances between known TSS and

the strain specific DMRs (**Figure 2-5A**). We found that there are more differentially methylated CpGs located near transcription start sites than expected by chance.

To further investigate the possible functional roles of the strain-specific DMRs, we analyzed the correlation between the DMRs and differentially expressed genes. We found that 28.1% (170 of 604 tested) of the DMRs are located near transcripts with differential gene expression between 129S1 and PWK. This percentage is only slightly higher than expected given the total number of differentially expressed genes compared to total genes expressed (23.3%, *p*-value of 0.006). One possible explanation for the low correlation between DMR and differential gene expression is the tissue type used. The MSNP experiment utilized genomic DNA extracted from whole brain, while the expression data was generated from 129S1xPWK F1 liver. Though the mismatch between tissue types is not ideal, the small number of DMRs that are associated with differentially expressed transcripts may represent DMRs that are consistent between all tissue types. These DMRs would likely represent epigenetic marks established very early during development, before tissue lineage was specified. Nevertheless, future experiments should have matching methylation and expression data in order to draw significant conclusions about DMR effects on nearby gene expression.

MATERIALS AND METHODS

Mice and tissues

Mice from the two parental strains (129S1/SvImJ, and PWK/PhJ) were originally obtained from The Jackson Laboratory. They were bred at UNC-Chapel Hill for multiple generations and interbred to generate reciprocal F1 hybrids. Mice were euthanized at eight-weeks of age and the right-hemispheres of the brain were dissected. All procedures were conducted in accordance with NIH guidelines for the care and use of experimental animals and based on protocols approved by the Institutional Animal Care and Use Committee of UNC-Chapel Hill.

Mouse Diversity Array (MDA) processing

Genomic DNA was purified from tissues according to a standard protocol of phenol/chloroform extraction followed by ethanol precipitation. Each genomic DNA sample was divided into three separate restriction digestion reactions containing *Hpa*II*, Msp*I*,* or reaction buffer only. For each reaction, 2.5 µg of DNA was digested in a total volume of 100 µl for three hours at 37°C, followed by heat inactivation. Pre-digested samples were then processed, from start to finish, according to the Affymetrix 6.0 genotyping protocol and hybridized to the MDA [62] at the UNC Functional Genomics Core Facility.

MSNP analysis

The array intensities were normalized using MouseDivGeno [67] and intensities from SNP and exon probe sets were used for further analysis. We ignored probes with restriction fragments longer than 2 kb because they yield a weak signal. For each probe set, we determined the number of *Msp*I restriction sites (CCGG) internal to the corresponding *Nsp*I or *Sty*I restriction fragment and classified probe sets with one as CCGGI. Probe sets with more than one *Msp*I site were classified as CCGGII. Note that CCGGI includes cases where there is one common *Msp*I site in both *Sty*I and *Nsp*I, and also cases where the *Sty*I (*Nsp*I) fragment has one *Mps*I site and the length of *Nsp*I (*Sty*I) fragment is longer than 2 kb. We used t-statistics to test whether any of the three digestion reactions conditions (no enzyme, *Hpa*II and *Msp*I) lead to differences in mean intensity level at each probe set. Since the number of probe sets is large and the sample size is small, we used the t-test with shrinkage variance, implemented in R/maanova [89]. Using a *p*-value of 10^{-10}, which corresponds to a false discovery rate (q-value) of 10^{-7} after multiple test correction we tested three hypotheses for each probe set:

i) Intensities from buffer condition > *Hpa*II condition > *Msp*I condition.

ii) Intensities from buffer condition = *Hpa*II condition > *Msp*I condition.

iii) Intensities from buffer condition > *Hpa*II condition = *Msp*I condition.

We denoted SNP or exon probe sets that reject first, second and third hypothesis as partially methylated, fully methylated, and unmethylated probe sets, respectively.
To test strain or parent-of-origin specific methylation, we restricted the analysis to 118,154 SNP probe sets, for which the genotypes of our F1 mice were heterozygous. These genotypes were obtained from the consensus call based on twelve undigested F1 samples and also predicted the genotype of F1 hybrids based on the genotypes of parental strains [1]. We evaluated the contrast between intensities for the two allelic probes at each informative SNP tagging partially methylated CpGs. In buffer treated F1 samples, both alleles are expected to have similar intensities and the contrast to be near zero (**Figure 2-3**, black circles and squares). To identify allelic effects we tested whether the contrast between the alternative alleles at each informative SNP probe set deviates from zero after *Hpa*II treatment. This test is analogous to the logic employed by Kerkel *et al.* and Yuan *et al.* with the advantage of using probe intensity data directly. If the contrast deviated from zero and the direction in both reciprocal crosses was consistent we classified this probe set as strain specific methylation. If the contrast deviated from zero and the direction in both reciprocal crosses was opposite we classified this probe set as parent-of-origin dependent methylation. We used a t-test with *p*-value 0.01, equivalent to false discovery rate 0.05 and silhouette score 0.5. **Figure S2-5** provides the contrast plots for a subset of the 305 SNPs with strain and parent-of-origin effect. All sequence analyses are based on the mouse genome assembly mm9, NCBI Build 37.

Expression analysis

To determine significant differential PWK and 129S1 expression, we analyzed previously reported expression quantitative trait locus (eQTL) data [29]. Briefly, mRNA was extracted from livers of 15-week-old Pre-Collaborative Cross mice and processed for hybridization to the Affymetrix GeneChip Mouse Gene 1.0 ST Array. Allele effects were estimated using partial correlation coefficients.

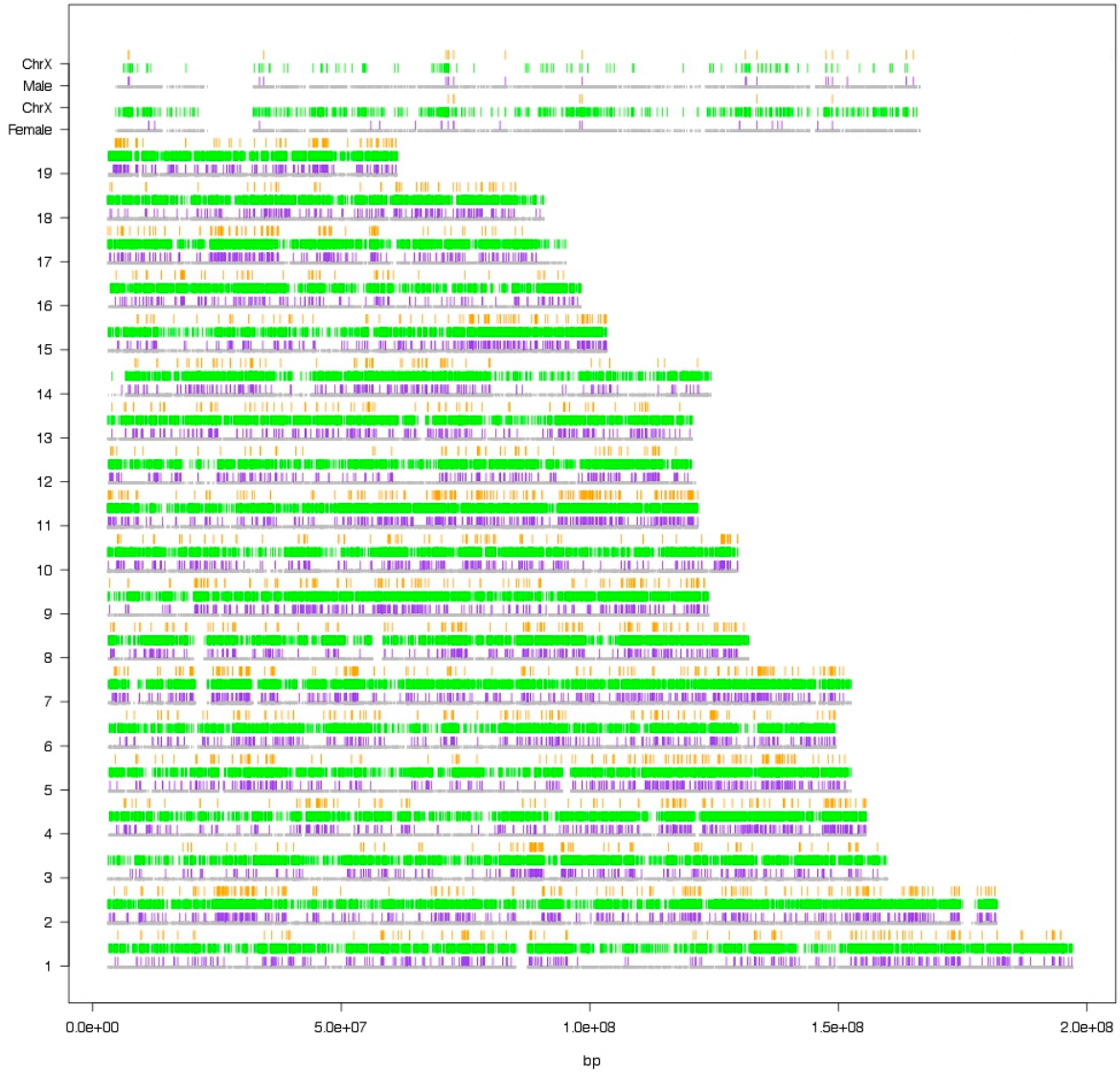


Figure S2-1. Global map of methylation patterns in the mouse brain according to stringent criteria. SNP and exon probes tagged to: fully methylated CpGs are shown in green, fully unmethylated CpGs are shown in purple, and partially methylated CpGs are shown in orange. Results for X chromosome probe sets in males and females are plotted separately. All probe sets are plotted in light grey.

Figure S2-2. Sex-specific methylation map. Each tick mark represents a probe set located at its corresponding chromosome position. SNP and exon probes associated with fully methylated CpGs are shown in dark green (male) or light green (female), probe sets associated with unmethylated CpGs are shown in dark purple (male) or light purple (female), and probe sets associated with partially methylated CpGs are shown in dark orange (male) or light orange (female).

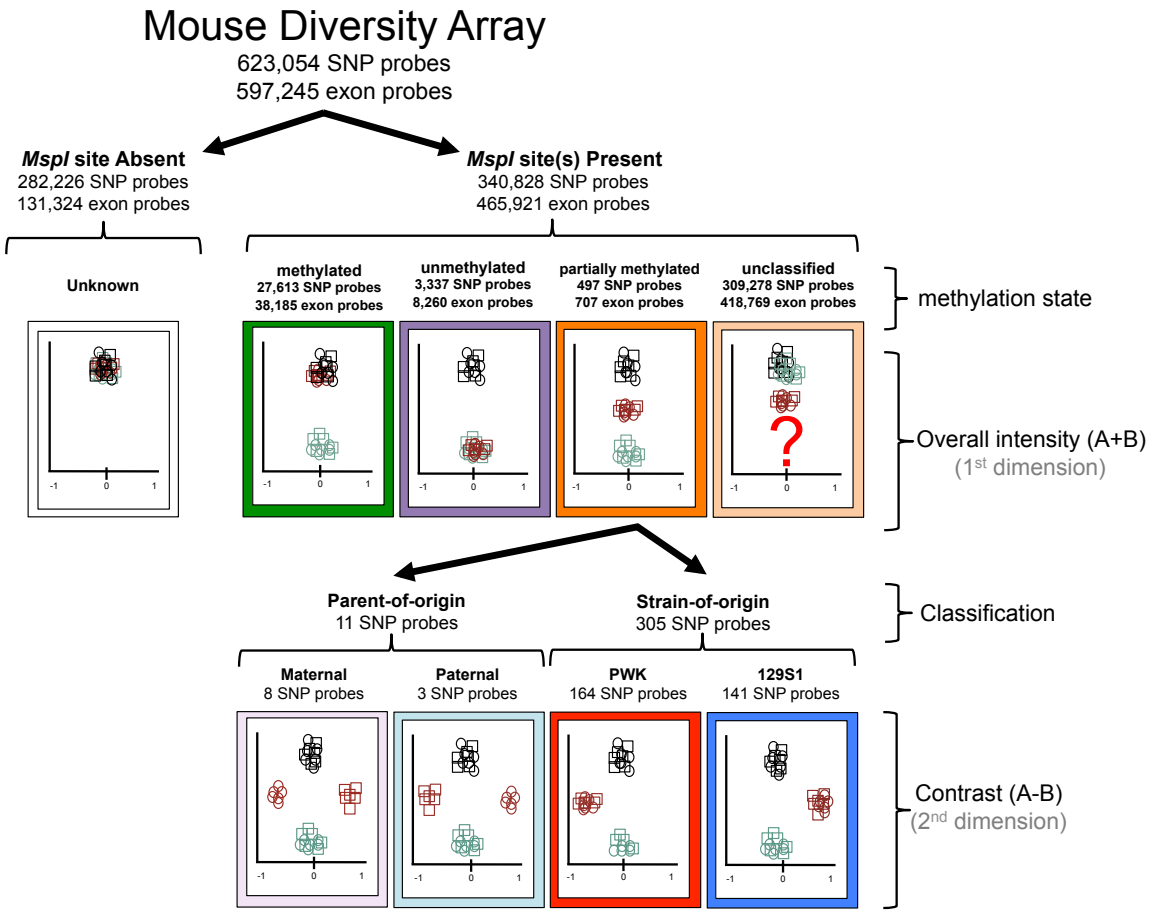


Figure S2-3. Strain-specific and parent-of-origin methylation map. The Figure shows partially methylated SNP probe sets with consistent allele-specific results. SNP probe sets associated with strain-specific CpG methylation are shown on top of each chromosome while SNP probe sets with parent-of-origin methylation are shown below each chromosome. SNPs tagging PWK-specific methylation class are shown in red, SNPs tagging 129S1 specific methylation class are shown in blue, maternal specific methylation are shown pink and paternal specific methylation are shown in light blue. Locations of known clusters of imprinted genes are shown in black.

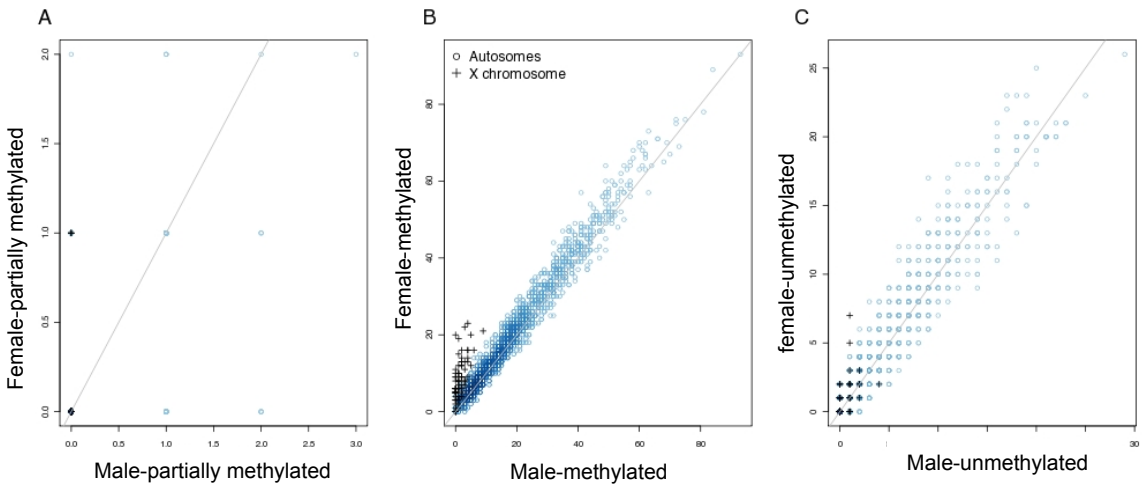


Figure S2-4. Pairwise comparisons of sex-specific methylation. Each panel compares male and female SNP and exon probes sets associated with partially methylated (A), methylated (B), and unmethylated (C) CpGs.

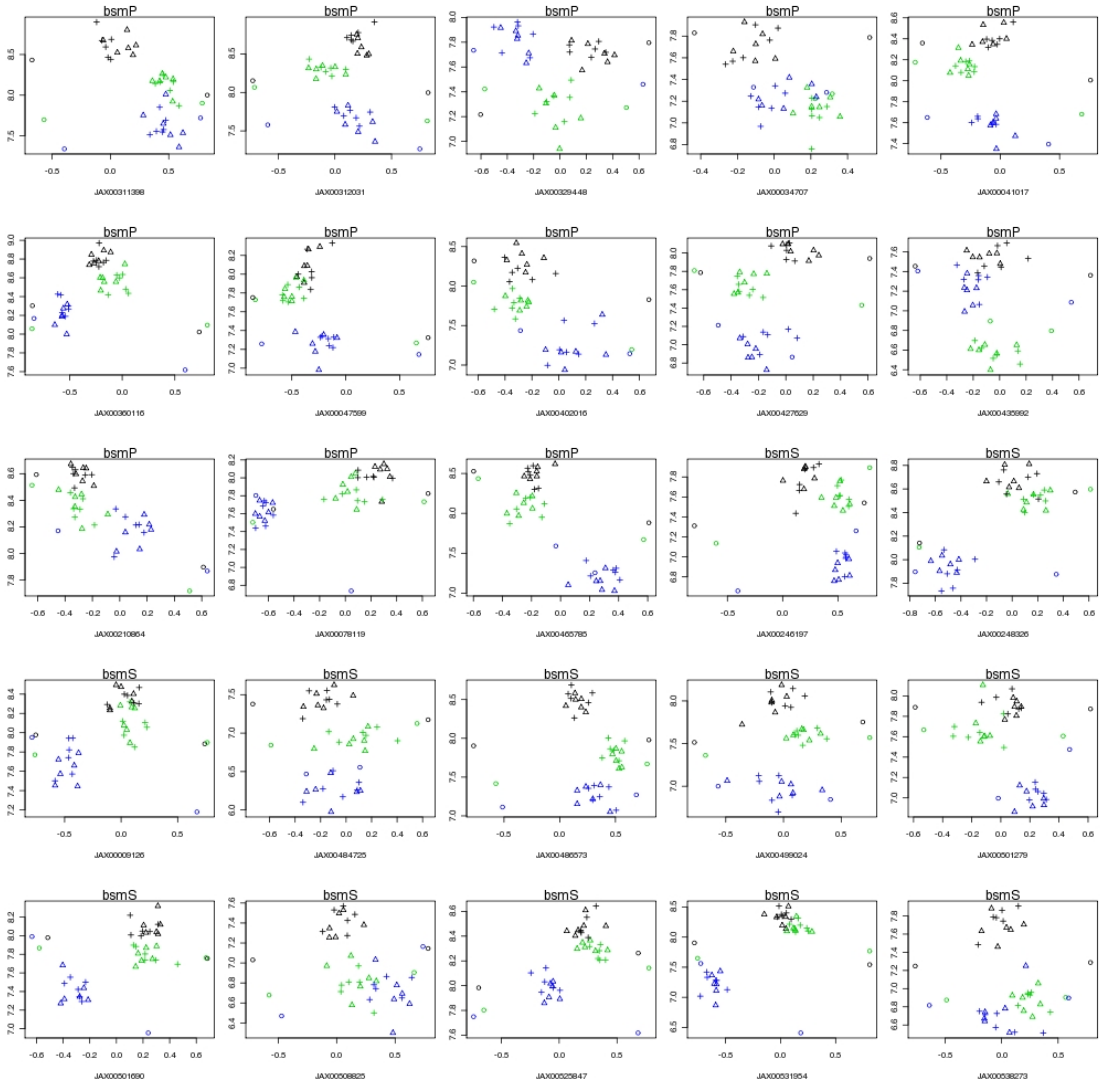


Figure S2-5. Contrast plot for SNP probes tagging strain-specific and parent-of origin differentially methylated CpGs. The y-axis represents the total hybridization intensity (probe A + probe B). The x-axis represents the contrast in probes (A-B)/(A+B). Samples treated with buffer only are black, *Hpa*II treated are green, and *Msp*I are blue.

	off target totals			
	Methylated	Unmethylated	partially methylated	unclassified
off-target	1562	160	29	23337
on-target	26051	3177	571	285941

	Expected			
	Methylated	Unmethylated	partially methylated	unclassified
off-target	2033	246	44	22766
on-target	25580	3091	556	286512

	off target strain			
	Methylated	Unmethylated	partially methylated	unclassified
129S1	189	17	5	3438
PWK	1373	143	24	19899

	Expected			
	Methylated	Unmethylated	partially methylated	unclassified
off-target	227	23	4	3394
on-target	1335	137	25	19943

	off target allele-specific		
	strain	PoO	unclassified
off-target	1	0	28
on-target	305	11	181

	Expected		
	strain	PoO	unclassified
off-target	17	1	12
on-target	289	10	197

Table S2-1. Contingency table with off-target SNP probes. Shown is the number of observed and expected off-target variants that may interfere with probe binding [67].

					Untreated Arrays		*HpaII* treated Arrays					
Chr	**Position(m37)**	**SNP**	**129S1 allele**	**PWK allele**	**(129S1xPWK)F1 mean call**	**(PWKx129S1)F1 mean call**	**(129S1xPWK)F1 mean call**	**(PWKx129S1)F1 mean call**	**associated gene**	**parent-of-origin methylation**	**MSNP captured CpG's that agree with previous differential methylation studies (m37)**	**Previous evidence of imprinting**
1	63,312,598	rs30012754	A	G	heterozygous	heterozygous	hemizygous (G)	hemizygous (A)	*Zdbf2*	*paternal*	*none*	imprinting[a]
2	157,387,678	rs27338074	G	A	heterozygous	heterozygous	hemizygous (G)	hemizygous (A)				
	157,387,807	rs27338072	T	C	heterozygous	heterozygous	hemizygous (T)	hemizygous (C)	*Nnat, Blcap*	*maternal*	*none*	imprinting[b]
	157,388,013	rs27338071	A	T	heterozygous	heterozygous	hemizygous (A)	hemizygous (T)				
2	174,125,510	rs6314659	C	T	heterozygous	heterozygous	hemizygous (C)	hemizygous (T)	*Gnasxl*	*maternal*	*174,124,504; 174,124,583; 174,124,643; 174,124,856; 174,124,929;174,124,965; 174,124,993; 174,125,154; 174,125,190; 174,125,398; 174,125,448; 174,125,552*[c]	imprinting[c]
7	69,085,111	rs31991512	A	C	heterozygous	heterozygous	hemizygous (A)	hemizygous (C)	*AK080655*	*maternal*	*N/A*	??
7	149,767,670	rs33821081	T	C	heterozygous	heterozygous	hemizygous (C)	hemizygous (T)	*Igf2/H19*	*paternal*	149,768,105;149,767,291 149,767,092;149,767047[f]	imprinting[d]
12	81,087,354	rs32640412	C	T	heterozygous	heterozygous	hemizygous (C)	hemizygous (T)	*Actn1*	*maternal*	*N/A*	??
	81,087,458	rs32641208	G	A	heterozygous	heterozygous	hemizygous (G)	hemizygous (A)				
15	73,007,162	rs31451869	A	G	heterozygous	heterozygous	hemizygous (G)	hemizygous (A)	*Eif2c2*	*paternal*	*N/A*	imprinting[b]
17	12,935,736	rs46625914	T	A	heterozygous	heterozygous	hemizygous (T)	hemizygous (A)	*Airn*	*maternal*	12,935,242[g];12,935,286[g]	imprinting[e]

[a]Kobayash et al. 2009; [b]Schulz et al. 2009; [c]Kelsey et al. 1999; [d]Bartolomei et al. 1991, Guillemot et al. 1995; [e]Stoger et al. 1993, Wutz et al. 2001, Seidl et al. 2006; [f]Bartolomei et al. 1993; [g]Birger et al. 1999, [b]Gregg et al. 2010

Table S2-2. Parent-of-origin table. Shown are the 11 SNPs that are associated with parent-of-origin methylation.

CHAPTER III: INTRONIC PARENT-OF-ORIGIN DEPENDENT DIFFERENTIAL METHYLATION AT THE *ACTN1* GENE IS CONSERVED IN RODENTS BUT IS NOT ASSOCIATED WITH IMPRINTED EXPRESSION[2]

RESULTS

A Novel Actn1 DMR has Preferential Maternal Methylation in Diverse Mouse Tissue

In a previous study, we performed a genome-wide methylation study of the mouse brain DNA by methylation-sensitive single nucleotide polymorphism (MSNP) analysis (Calaway *et al*. u[2]npublished results, **Chapter II**). This analysis was applied to brain DNA of F1 offspring of reciprocal crosses between 129S1 and PWK mice. Our study identified a novel parent-of-origin dependent DMR associated with two SNPs, rs32640412 and rs32641208, located in a CpG island and in the last intron of the *Actn1* gene (**Figure 3-1A**). Maternal-specific methylation of this DMR was confirmed by methylation-sensitive restriction fragment length polymorphism (MS-RFLP) analysis (Calaway *et al*. unpublished results, **Chapter II**).

[2] The following chapter describes work done in collaboration with Jose Ignacio Dominguez, Megan E. Hanson, Ezequiel C. Cambranis, Dr. Fernando Pardo-Manuel de Villena, and Dr. Elena de la Casa-Esperon. The purpose of this study was to further characterize a parent-of-origin DMR discovered in the 3'UTR of the *Actn1* gene from the genome-wide survey described in **Chapter II**. We quantified parent-of-origin and strain-specific methylation and attempted to determine the general size of the DMR. Furthermore, we explored the possible functions of the DMR by measuring expression of *Actn1* and surrounding genes. We found no allelic imbalance at the *Actn*1 gene or any nearby surrounding genes in seven different tissue types. I significantly contributed to sample preparation, design and implementation of molecular assays, and data analysis. I contributed to the writing of the manuscript, and the figures were of my design (with the exception of **Figures 3-3 and 3-4**). These results have been published in PLoS One (Calaway *et al*. 2012).

In this study, we have expanded the methylation analysis of the *Actn1* gene. First, we examined whether the *Actn1* DMR occurs in tissues other than brain. Genomic DNA isolated from whole brain, kidney, liver, spleen, testis, tail, and femoral muscle from four (PWK×129S1)F1 mice and four (129S1×PWK)F1 mice were subjected to MS-RFLP. In this technique, restriction digestion with methylation-sensitive endonucleases is performed prior to PCR amplification of the region under our study; consequently, only methylated restriction sites are preserved and, thus, amplified. In order to determine the methylation status of each allele, an additional digestion was performed after PCR and before electrophoresis with strain-specific endonucleases *Sty*I (which only digests the PWK allele) or *Ahd*I (specific for the 129S1 allele) (**Figure 3-1A**). Depending on the direction of the cross, the percent methylated maternal allele or paternal allele was calculated by the ratio of relative fragment densities of either *Sty*I (PWK) or *Ahd*I (129S1) digestions (see **Materials and Methods**). Both methylation measurements were correlated for each of the three methylation-sensitive enzymes used: *BsaA*I, *Eag*I and *Hpa*II (**Figure S3-1**). Examples of *BsaA*I MS-RFLP results for liver are shown in **Figure 3-1B**. **Figure 3-1C** represents the percent maternal methylation at a single CpG internal to the *BsaA*I cut site (chr12:81,269,613 (m37), **Figure 3-1A**) in diverse tissues. We observed that differential methylation at *Actn1* is not unique to brain. Similar results were obtained for both *Eag*I and *Hpa*II digestions (**Figure S3-2**).

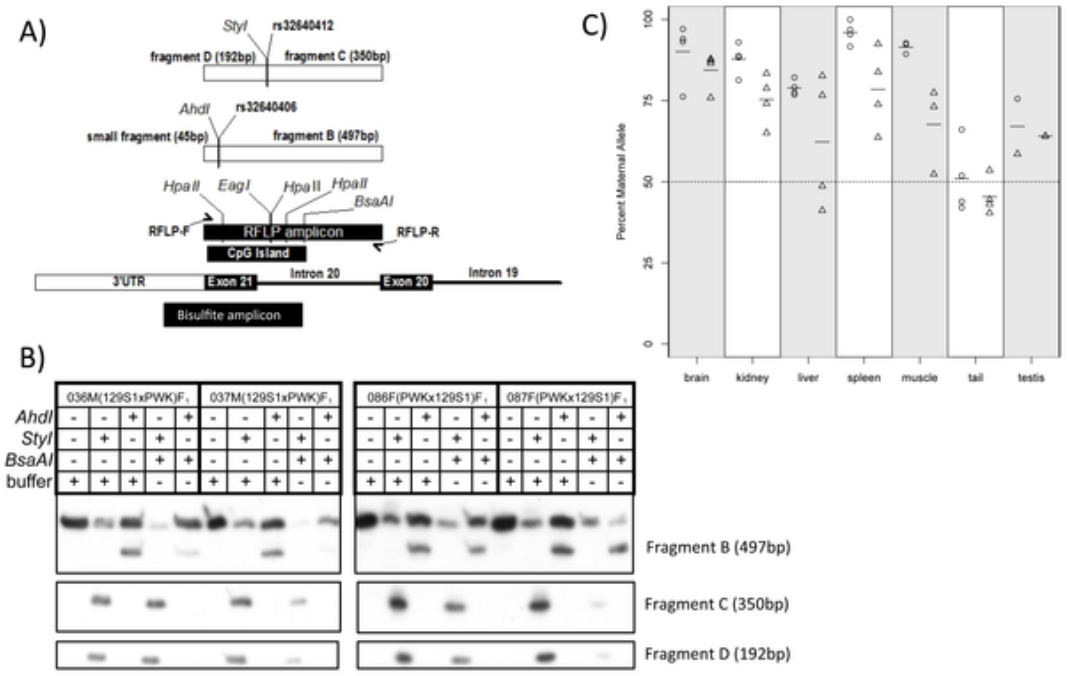


Figure 3-1. Maternal methylation of a novel DMR at the *Actn1* gene in diverse mouse tissues. (A) A detailed map of the novel maternal *Actn*1 DMR is shown in the lower part. The diagram directly above shows the design for the MS-RFLP and bisulfite sequencing validation assays. Also included in this diagram are the locations of the methylation-sensitive enzyme restriction sites tested with MS-RFLP (*BsaA*I*, Eag*I and *Hpa*II*)*, the strain-specific cut sites (*Ahd*I (present in 129S1 but not in PWK, due to SNP rs32640406) and *Sty*I (present in PWK but not in 129S1, due to SNP rs32640412)), and the strain-specific resulting restriction fragments (see Methods). (B) MS-RFLP results of four mouse liver samples. The matrix above the gel shows the different conditions for each individual lane. The plus sign (+) indicates addition, while the minus sign (-) indicated no addition of each corresponding endonuclease. (C) Percent maternal methylation of an individual CpG (targeted by the *BsaA*I endonuclease) within different tissues. Circles represent individual (PWKx129S1)F_1 mice, while triangles represent individual (129S1xPWK)F_1 mice. Horizontal bars represent percent maternal methylation averages.

Moreover, we observed differences in the mean percent maternal methylation at the *BsaA*I CpG site between tissue types (**Figure 3-1C**). Pair wise t-tests revealed significant differences in the percent maternal methylation between tail and other tissues ($\alpha < 0.05$ in both types of F1 mice, **Table S3-1**). In addition, a two-factor ANOVA test identified statistically significant differences not only between tissue types ($F = 16.733$, *p*-value = $7.639 \cdot 10^{-10}$), but also between reciprocal F1 hybrids ($F = 20.413$, *p*-value = $4.821 \cdot 10^{-5}$).

However, the varying degree of maternal methylation between tissues is not significantly different between reciprocal F1s.

Actn1 DMR Extent and Conservation in Murine Rodents

To determine if the *Actn1* DMR is unique to mice or, on the contrary, conserved in other mammalian species, we analyzed the orthologous regions in humans and rats. Located distally on chromosome 12 in mouse (81,268,534-81,361,303, NCBI37/mm9), *Actn1* is orthologous with a region on rat chromosome 6 (103,187,905–103,282,948, Baylor 3.4/rn4) and human chromosome 14 (69,341,075–69,359,000, GRCh37/hg19). We predicted the location of the human and rat orthologous DMRs based on the assumption that they are typically associated with regions of high CpG dinucleotide density (CpG islands) and their shores [41, 90]. We used the following criteria to define a CpG island: a GC content greater than 50% and an observed/expected (O/E) CpG ratio greater than 0.6 over a 200 bp minimum length. Both the mouse *Actn1* CpG island (27CpGs, 57.3% GC content over 302 bp, CpG O/E 1.10) and the rat *Actn1* CpG island (25CpG, 60.8% GC content over 265 bp, CpG O/E 1.03) span most of the last exon coding region and part of the last intron (intron 20 in reference sequences NM_134156.2 for mouse and NM_031005.3 for rat) (**Figure 3-1A**). In humans, the CpG island is larger (40 CpGs, 68.2% GC content, length 393 bp, CpG O/E 0.91) and includes a large portion of the 3′ UTR (reference sequence NM_001102.3).

We investigated the methylation status of multiple CpG sites at the *Actn1* CpG islands of these three species by sodium bisulfite treatment followed by PCR and sequencing analysis. (PWK×129S1)F1 mice displayed brain maternal hypermethylation and paternal hypomethylation, while (129S1×PWK)F1 mice showed weak maternal methylation and sporadic paternal methylation across the 19 CpG's sequenced (**Figure 3-2B**).

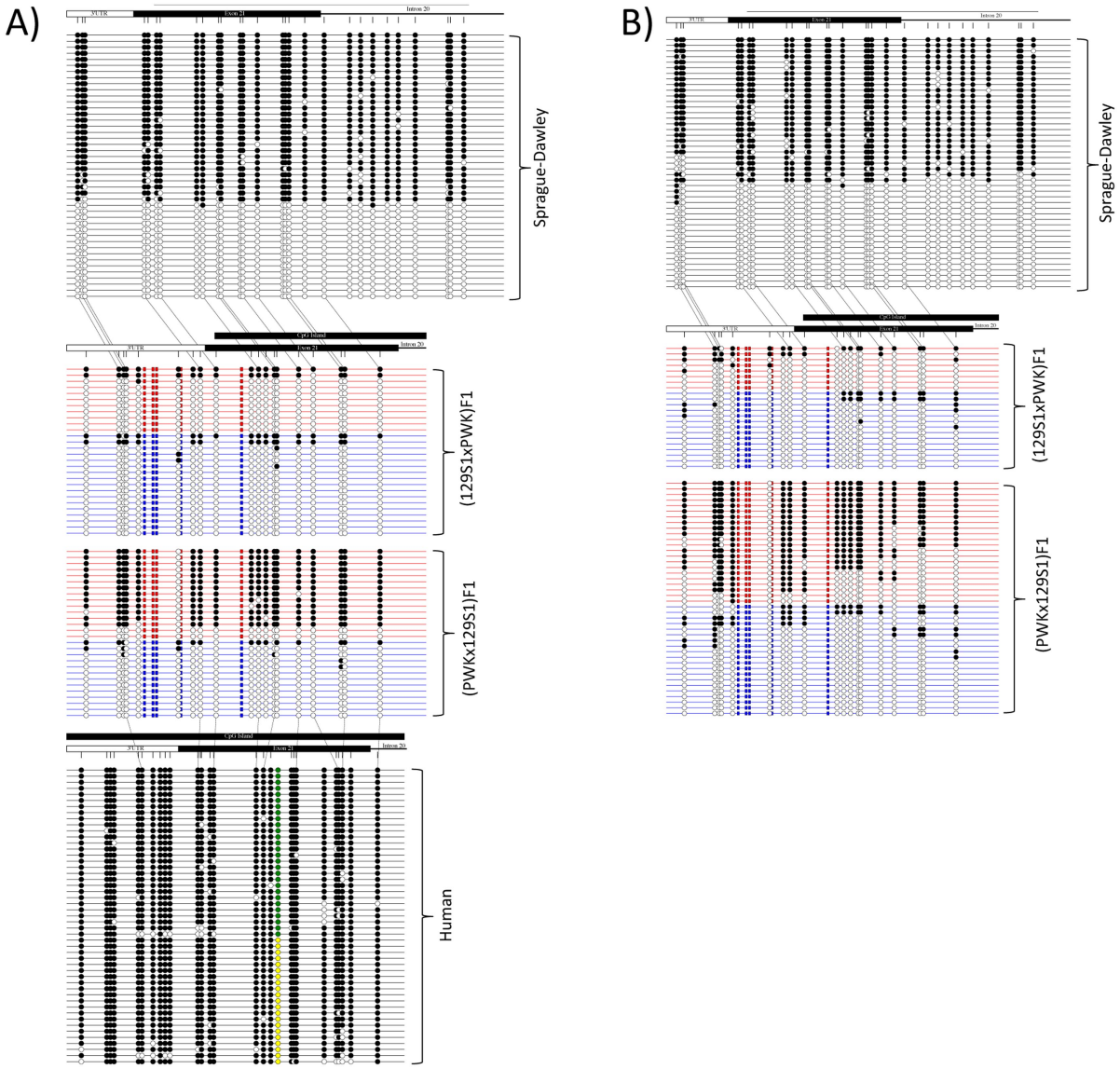


Figure 3-2. Bisulfite sequencing analysis of the *Actn1* DMR in mouse, rat and human tissues. Panel A shows bisulfite sequencing results from clones isolated from rat liver, mouse liver, and human hepatocytes. Each horizontal line represents a unique clone. Red and blue lines represent maternal and paternal parent-of-origin, respectively, based on five strain-specific variants. Open circles are unmethylated CpGs, while closed circles are methylated CpGs. Green and yellow circles shown in human hepatocyte clones represent variant rs11557769 and distinguish parental alleles, although parent-of-origin is unknown. Orthologous CpGs are connected by dotted lines (in relation to mouse). Panel B shows bisulfite sequencing results from clones isolated from rat right brain hemisphere (top) and mouse right brain hemispheres (bottom).

We found similar results in mouse liver DNA (**Figure 3-2A**). These data indicate that methylation at the *Actn1* DMR depends both on the parental and the strain origin (the

sequences in *cis*) of the CpG sites. They are also consistent with the MS-RFLP results of (129S1×PWK)F1 mice (**Figure 3-1C**), which showed more methylation variability and, on average, lower percent of maternal methylation than (PWK×129S1)F1 animals. In rat, we were unable to identify a polymorphism for establishing a parent-of-origin anchor within the 347 bp bisulfite amplicon, due to the limited genetic diversity between available rat strains. Nevertheless, we observed a strongly polarized population of hypermethylated or hypomethylated bisulfite amplicons suggestive of differential methylation in both rat brain and liver DNA (**Figures 3-2A and 3-2B**). In contrast, the human *ACTN1* DMR is consistently methylated at greater than 94% (false discovery rate of 0.68%) in hepatocytes (**Figure 3-2A**). The presence of a T → A transversion (rs11557769, at position 69,341,653 (GRCh37/hg19)) allowed us to conclude that both the maternal and paternal alleles are hypermethylated (**Figure 3-2A**). Therefore, in human hepatocytes, the orthologous region to the mouse *Actn1* DMR is not differentially methylated, while biased methylation is conserved in murine rodents.

We also examined the methylation upstream and downstream of the mouse *Actn1* DMR, by performing bisulfite treatment of liver DNA followed by PCR amplification of flanking sequences. Our assay design was constrained by the scarcity of informative SNPs between 129S1 and PWK and the profusion of homopolymers in the sequences surrounding the DMR. Nevertheless, we generated data for one region upstream (81,270,081-81,270,495, NCBI37/mm9) and one region downstream (81,263,196-81,263,520, NCBI37/mm9) from our previous DMR bisulfite assay (**Table S3-2** and **Figure S3-3**). Comparative analysis of the results of the reciprocal crosses shows that the preferential maternal methylation observed in the DMR does not extend to these neighboring regions (**Figure S3-3**). Therefore, in mouse liver DNA, the DMR appears to be confined to the vicinity of the last *Actn1* intron.

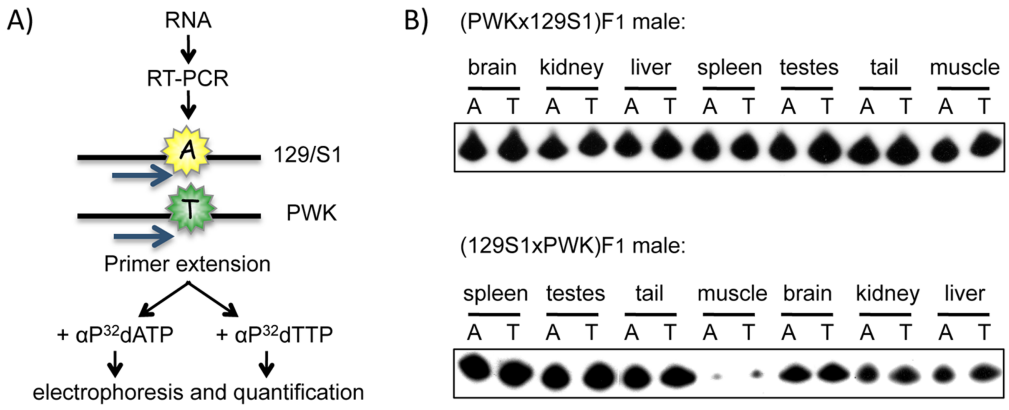


Figure 3-3. *Actn1* allelic expression analysis by SNuPE. A) Summary of the SNuPE (Single Nucleotide Primer Extension) method. B) Autoradiogram of SNuPE products after electrophoresis, showing biallelic expression of*Actn1* in all tissues analyzed.

Next, we tested if the *Actn1* parent-of-origin dependent DMR is associated with imprinted expression of nearby genes. To date, there have been no reports of imprinted expression of *Actn1*. To investigate if such is the case, we analyzed the expression of the mouse gene in RNA obtained from the same tissues and F1 individuals studied for DNA methylation purposes. In order to distinguish maternal from paternal expression, we sequenced the *Actn1* coding sequences and identified several SNPs between the 129S1 and PWK strains. The relative expression of 129S1 and PWK alleles was tested by Single Nucleotide Primer Extension (SNuPE) at a SNP located in chr12:81269902 (m37) (**Figure 3-3**) [91, 92]. In spite of the presence of the *Actn1* DMR, we always observed *Actn1* biallelic expression, finding no indication of allelic expression bias in any sex, F1 or tissue type (**Figure 3-3B** and **Figure S3-4A**). These results were validated by direct sequencing of the cDNAs generated in the SNuPE analysis (**Figure S3-4B**). We also confirmed biallelic expression of *Actn1* at other SNPs (chr12:81,284,503 and chr12:81,274,013 (m37)) by independent RT-PCR and sequencing analyses of F1 RNA samples (**Table S3-3** and **Figure S3-4B**).

Expression Studies of *Actn1* do not Reveal Imprinting Effects

Imprinted gene expression can be restricted to specific isoforms or developmental stages, being particularly common in placenta and embryonic tissues [58, 93-95]. In order to test if a DMR effect on *Actn1* transcription is restricted to prenatal stages, allelic expression analysis by RT-PCR and sequencing was applied to (129S1×PWK)F1 and (PWK×129S1)F1 E9.5 embryos and placentas of both sexes (**Table S3-3**).

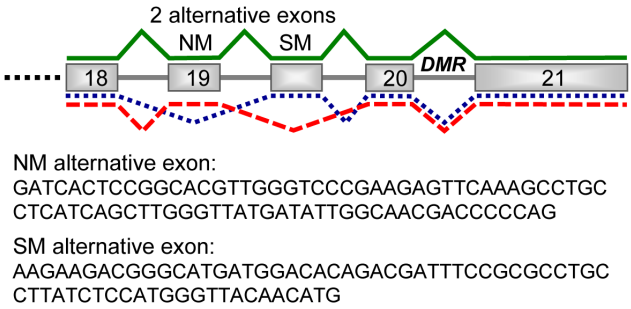


Figure 3-4. Mouse *Actn1* isoforms. They result from alternative splicing of two exons at the 3′ end of the gene. These exons are designated SM (smooth muscle) and NM (non-muscular) due to their homology to previously described rat alternative exons [96]. Exons are numerated 18–21 as on Ensembl transcript isoform ENSMUST00000021554. The position of the DMR in is indicated in the last intron (image not drawn at scale).

The results of this analysis showed no apparent allelic expression bias. We also tested if the DMR had an imprinted expression effect restricted to any specific *Actn1* isoform. In rat, three isoforms resulting from two alternatively spliced exons (NM ("non-muscle") and SM ("smooth muscle") exons) have been described of this gene [96]. We found these three *Actn1* isoforms are also present in mouse (**Figure 3-4**). Sequencing analysis of RT-PCR products with isoform-specific primers (**Table S3-3**) revealed that expression of the three isoforms is biallelic in (129S1×PWK)F1 and (PWK×129S1)F1 adult brain, E9.5 placentae and embryos of both sexes. Therefore, our results show that the *Actn1* parent-of-origin dependent DMR observed in F1 mice derived from PWK and 129S1 strains is not associated with *Actn1* imprinted expression in any of the sexes, tissues, developmental stages and isoforms analyzed.

DISCUSSION

During a genome-wide methylation study of the mouse brain DNA, we identified a novel parent-of-origin dependent DMR in the 3′ end of the *Actn1* gene (**Chapter II**). We have confirmed that this intronic DMR is maternally methylated in brain of F1 individuals derived from reciprocal crosses between 129S1 and PWK strains by MS-RFLP and bisulfite analyses. We have extended our mouse study to a tissue panel that is representative of all three germ layers: ectoderm (brain), mesoderm (kidney, spleen, muscle and testes) and endoderm (liver). All examined tissues (except for the tail, a body part of mixed origin [97]) display preferential maternal methylation of the *Actn1* DMR. These results suggest that the imprint was established very early during embryogenesis. Although this imprint persists through subsequent differentiation, the extent of maternal methylation varies significantly among tissue types, as well as between reciprocal crosses. Differences in allelic methylation levels among tissues, as well as interindividual variation, have also been observed in other DMRs, such as those associated with several imprinted genes [98-101].

Traditionally, parent-of-origin dependent DMRs have been identified due to their proximity to imprinted genes. In fact, they have been found even within imprinted gene sequences (*e.g.,* introns). Therefore, we examined the expression of *Actn1* in the same tissue panel as the methylation analyses. We found no indication of allelic imbalance in any of the adult tissues examined. We also explored the possibility that imprinted expression could be restricted to particular isoforms or to specific developmental stages (particularly embryonic and extraembryonic tissues) [58, 93, 94]. We found three isoforms of mouse *Actn1* that result from alternative splicing of two alternative exons. Nevertheless, none of them showed allelic expression bias in adult brain, E9.5 embryos or E9.5 placentas of both reciprocal crosses and sexes. Therefore, our results do not support an association of parent-of-origin dependent methylation at *Actn1* with imprinted expression of the same

gene. However, we cannot exclude the possibility that such imprinting could be restricted to a very specific cell type and/or developmental stage that have not been captured by our study.

We also tested if the DMR is involved in the imprinted expression of the next closest transcripts: *AK037382* and *Zfp36l1*, which are overlapping and close to the 3′ end of *Actn1*, respectively, as well as *Dcaf5 (Wdr22)*, a gene near to the 5′ end of *Actn1* (see **Materials and Methods**). However, we did not detect imprinted expression of these genes in any of the adult and embryonic mouse tissues analyzed (data not shown). In fact, the closest known imprinted genes are located as far as 29 Mb apart in the *Dlk1-Dio3* cluster (http://www.mousebook.org/catalog.php?catalog=imprinting).

From these results, we conclude that parent-of-origin dependent DMRs can be uncoupled from imprinted expression effects on nearby genes and, therefore, they are not perfect predictors of imprinted expression of genes located in their immediate proximity. This has important implications for large-scale searches for novel imprinted genes through the identification of parent-of-origin dependent epigenetic marks. In fact, recent genome-wide studies have also revealed the existence of novel parent-of-origin dependent DMRs outside known imprinted regions [102-105]. Although deeper analyses have allowed the association of several of these DMRs with imprinted genes, the role of other DMRs remains unclear. Some are located within introns (as the *Actn1* DMR), while others are in intergenic regions and far from gene sequences [103].

We have gone a step further and interrogated if the *Actn1* DMR is an oddity unique to the mice used in our study (*i.e.,* intersubspecific hybrids [1]), or if it is also present in other species. We have found that, while orthologous *Actn1* CpG islands exist in other mammals, differential methylation is conserved in murine rodents (mouse and rat) but absent in humans. Our findings open an interesting question: can parent-of-origin dependent DMRs have been evolutionarily selected due to a functional role other than imprinted expression

regulation? In other words: is the regulation of imprinted expression the only function of these DMRs? Several evidences indicate that DMRs and imprinted gene expression do not always go hand in hand. Within species, uncoupling of DMRs from imprinted expression can occur even in those typically associated with imprinted genes: for instance, paternal methylation of the imprinting control region of the *Rasgrf1* gene has been observed even in those tissues in which this gene is biallelically expressed [106]. This suggests that certain parent-of-origin dependent DMRs may have been selected for imprinting regulation and retained in all tissues throughout development, although imprinted expression would require tissue-specific factors in addition to differential methylation [107]. However, these selective pressures would be insufficient for the existence of other class of DMRs: those that are associated to imprinted expression in some species but not others. Such is the case of DMRs of the *IGF2R* gene, which is a gene that is imprinted in mice but not humans, while parent-of-origin differential methylation is present in both species [108-110]. Our finding adds an additional twist: DMR conservation in murine rodents in the absence of imprinted expression evidence.

A simple explanation for the *Actn1* DMR murine conservation is selection due to its necessary contribution to the regulation of chromosomal functions other than imprinted expression. In sexually reproducing organisms, parent-of-origin dependent epigenetic differences have been associated to phenomena as diverse as chromosome segregation or elimination and can affect replication, recombination and heterochromatinization of chromosomes in many sexually reproducing organisms [111, 112]. They have also been proposed to contribute to meiotic pairing and recombination and to DNA repair [111, 112]. From this broad perspective, large-scale studies of differentially methylated regions have the potential to unveil not only new imprinted genes, but also novel parent-of-origin dependent phenomena.

MATERIALS AND METHODS

Mouse Lines and Samples

Two mouse strains were obtained from the Jackson Laboratory: 129S1/SvImJ (abbreviated 129S1) and PWK/PhJ (abbreviated PWK). For MS-RFLP and expression analyses, we collected whole-brain, kidneys, spleen, liver, testes, femoral muscle, and tail from two female and two male (129S1xPWK)F1 mice, as well as two female and two male (PWKx129S1)F1 mice at 6-weeks of age. In all crosses, dams are listed first and sires last. Additionally, we isolated whole brain and liver from a 45-day-old, male Sprague Dawley rat (Harlan). Dissected tissues were immediately frozen in liquid nitrogen and DNA and RNA were extracted according to standard procedures. Expression studies were also performed in RNA extracted from pooled E9.5 whole embryos and from E9.5 placentas: two female and two male (129S1xPWK)F1 pools and two female and two male (PWKx129S1)F1 pools. All procedures were conducted in accordance with NIH guidelines for the care and use of experimental animals and based on protocols approved by the Institutional Animal Care and Use Committee of UNC-Chapel Hill. Human hepatocytes, harvested from subjects with various causes of death, were purchased from ADMET Technologies, Inc. (Durham, NC, USA).

Methylation-Sensitive Restriction Fragment Length Polymorphism (MS-RFLP) Analysis

Genomic DNA was first digested with *EcoR*I (New England Biolabs) to reduce structural complexity and ensure that the restriction site is accessible to subsequent endonucleases digestions [113]. Samples were then either digested with methylation-sensitive enzymes *BsaA*I, *Eag*I, *Hpa*II (NEB), or mock treated (buffer only). The cut sites of *BsaA*I, *Eag*I and *Hpa*II include one or more of the CpGs targeted for PCR amplification. Methylation-sensitive digested samples were then PCR amplified using a RFLP forward

primer and a RFLP reverse primer, and radiolabeled dCTP (**Figure 3-1A and Table S3-2**). PCR products were digested with either 129S1-specific *Sty*I, PWK-specific *Ahd*I, or mock treated. Samples were electrophoresed through 5% acrylamide denaturing gel and visualized by X-ray film.

For allelic ratio quantitation, X-ray films were scanned (Epson) and the Tiff images were imported into ImageJ [114] for densitometry. We arbitrarily named the undigested RFLP amplicon, "A" (542 bp); the fragment generated by *Ahd*I digestion, "B" (497 bp); the larger fragment from *Sty*I digestion, "C" (350 bp); and the smaller *Sty*I fragment, "D" (192 bp) (**Figure 3-1A**). The relative amount of each parental allele was determined by the ratio of the sum of the absolute density of allele-specific fragments (**Figure S3-5**) and to the total absolute density of all bands:

*Sty*I digestion: methylated PWK allele = (C+D)/(A+C+D) (direct measurement)
*Ahd*I digestion: methylated 129S1 allele = B/(A+B) (direct measurement)

This method for calculating percent methylated parental alleles gave an inflationary result for PWK and a deflationary result for 129S1 based on buffer-only controls. We, therefore, created a panel with diverse ratios of PWK and 129S1 genomic DNA and digested with *Sty*I or *Ahd*I to serve as a standard curve (PWK/129S1:0/100, 5/95, 25/75, 50/50, 75/25, 95/5, 100/0). This allowed us to interpolate "actual" PWK/129S1 allelic ratios from "observed" ratios (**Figure S3-5B**). We normalized all RFLP densitometry measurements by applying the respective interpolation equations. We utilized the R environment for conducting the two-factor ANOVA and t-tests for determining significant differences in maternal methylation between tissues and reciprocal crosses.

Sodium Bisulfite Sequencing

One microgram of genomic DNA from mouse (n = 2), rat (n = 1) or human (n = 1) tissues was treated with Zymo Research EZ DNA Methylation-Gold Kit according to the manufacturer's protocol. Species-specific primers were designed to flank and amplify the bisulfite converted DMR (**Table S3-2**). Purified PCR products were cloned and sequenced. The false discovery rate for methylated CpG's was calculated by the number of unconverted non-CpG cytosines divided by the total number of non-CpG cytosines across individual PCR reactions.

Expression analysis

Allele-specific expression of *Actn1* was analyzed by two independent methods: sequencing or Single Nucleotide Primer Extension (SNuPE) analysis of SNPs present in RT-PCR products. RNA of the above described mouse tissue samples was retrotranscribed (using *Actn1*-specific primers), followed by PCR (**Tables S3-2 and S3-3**), using the appropriate controls to avoid genomic DNA amplification. An informative SNP at position 12:81,269,902 (m37) was selected for analysis of the relative expression of alleles by Single Nucleotide Primer Extension (SNuPE) [91, 92] (**Figure 3-3**). Sanger sequencing of the same RT-PCR products was performed in order to verify the SNuPE results; *Actn1* allelic expression was determined by chromatogram inspection of three SNPs at positions 12:81,269,902, 12:81,269,896 and 12:81,269,456 (m37) (**Table S3-3** and **Figure S3-4B**). As additional confirmation, we performed RT-PCR of brain samples with a different set of primers (**Tables S3-2** and **S3-3**). The resulting products were subjected to Sanger-sequencing to test for allelic expression at SNPs located in positions 12:81,284,503 and 12:81,274,013 (m37) (**Figure S3-4B**). Allelic expression analysis of *Actn1* isoforms was also performed by sequencing of RT-PCR products, using isoform-specific primers Actn1-18SM, Actn1-NM20 and Actn1-NMSM (**Figure 3-4** and **Tables S3-2** and **S3-3**). Allelic expression of *Zfp36l1, AK037382* and *Dcaf5 (Wdr22)* in embryonic and adult mouse tissues was

determined by RT-PCR (see primers in **Table S3-2**), followed by Sanger-sequencing and chromatogram inspection of SNPs between 129S1 and PWK alleles described at http://www.sanger.ac.uk/cgi-bin/modelorgs/mousegenomes/snps.pl.

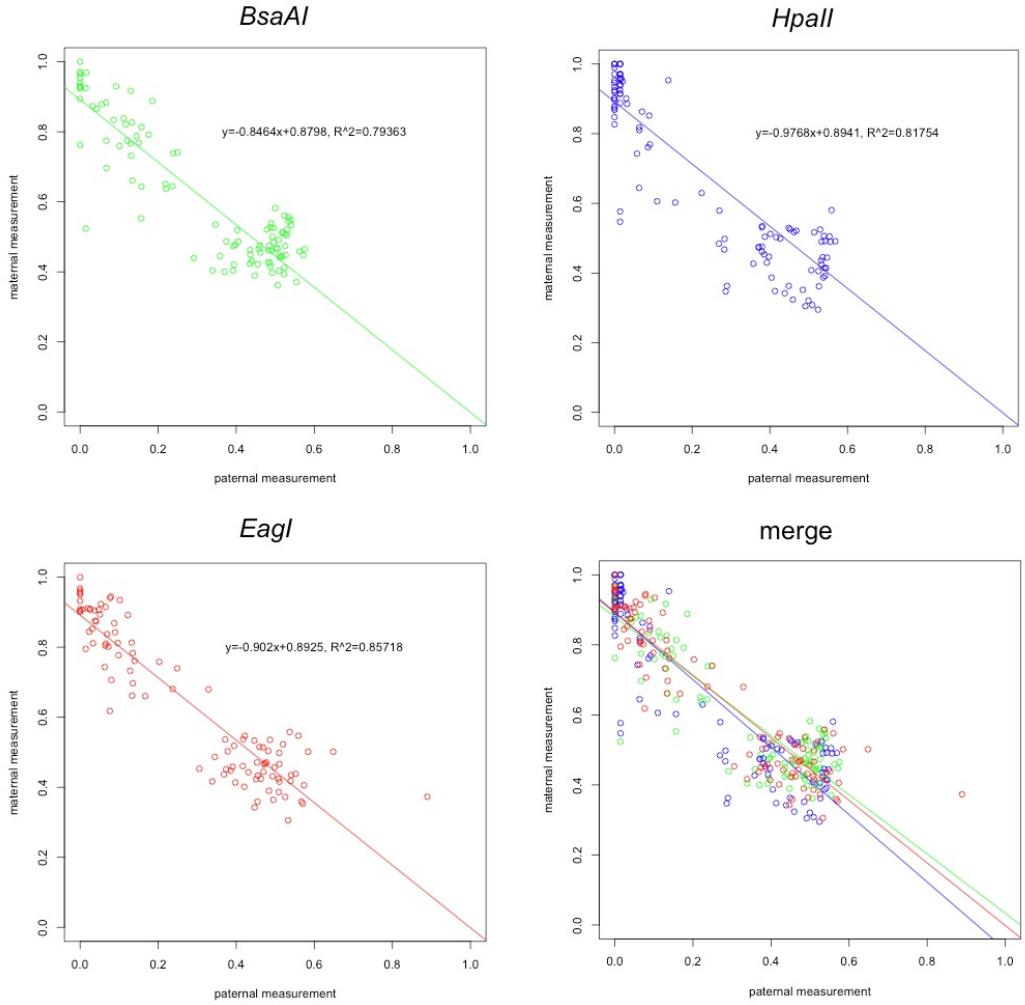


Figure S3-1: Correlation of maternal and paternal allelic methylation measurements at the *Actn1* DMR. Depending on the direction of the cross, the percent maternal methylation and the percent paternal methylation measurements are calculated by the ratios of *Sty*I or *Ahd*I restriction fragment densities. The direct measurements of maternal methylation are plotted against the direct measurements of paternal methylation for each individual methylation-sensitive endonuclease. Fitted line equations and R^2 values are shown in the graph interior.

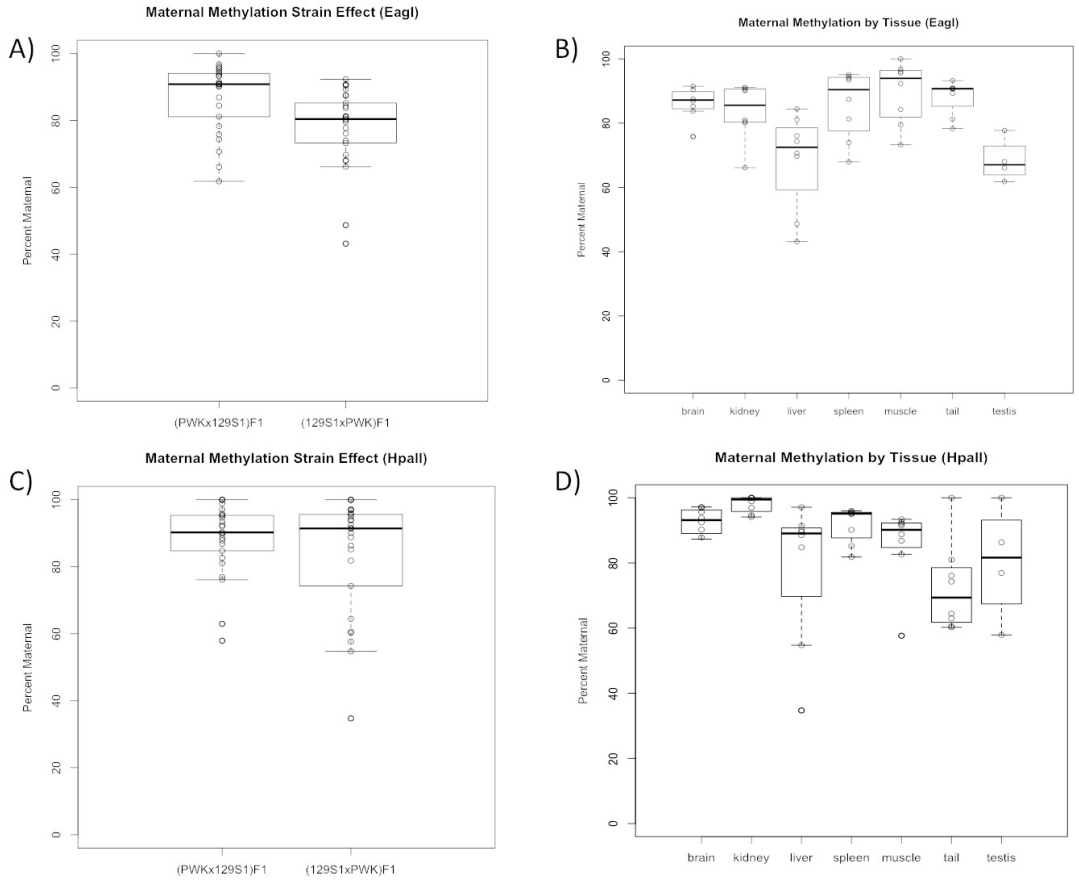


Figure S3-2: Percent maternal methylation of *Actn1* DMR based on *Eag*I and *Hpa*II MS-RFLP. Box and whisker plots showing the lower quartile, median, and upper quartile of percent maternal methylation by cross and by tissue type determined by *Hpa*II or *Eag*I MS-RFLP.

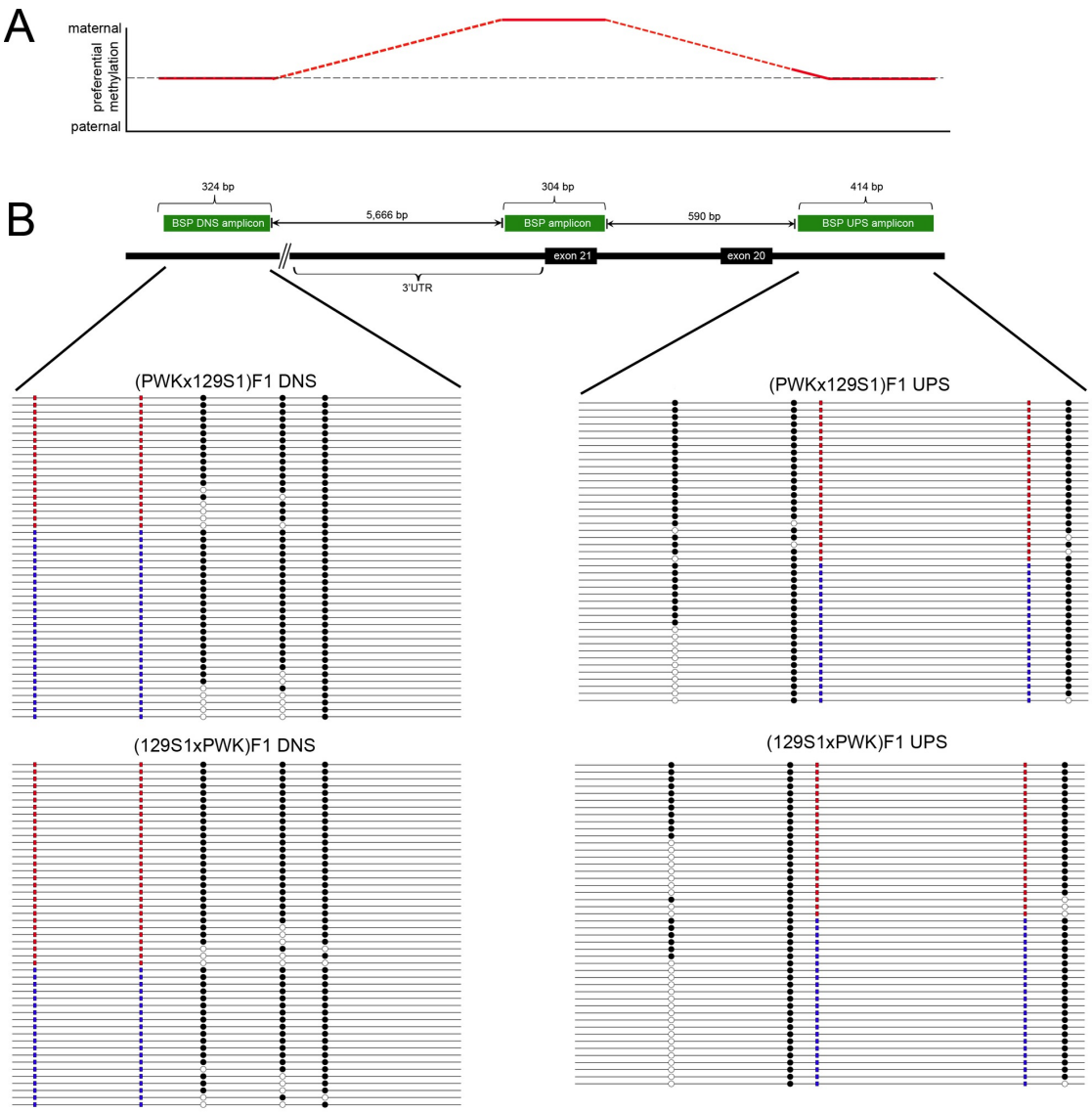
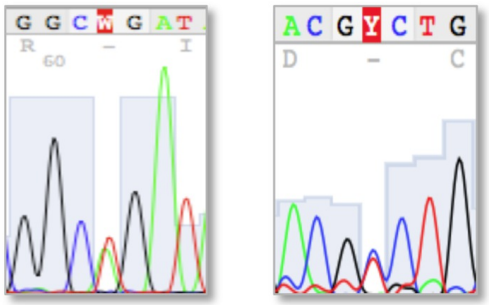


Figure S3-3: Bisulfite sequencing analysis of two regions flanking the *Actn1* DMR in mouse liver tissues. Panel A shows regions of preferential methylation investigated by bisulfite sequencing. Solid red lines represent sequenced regions, while dotted lines represent gaps in sequenced regions. **Panel B** shows a schematic representation of the positions and sizes of the regions selected for methylation analysis by bisulfite sequencing with respect to the location of the last two exons of *Actn1* (exons 20 and 21, ENSMUSE00000114871 and ENSMUSE00000335764, respectively). Two regions, situated downstream (DNS BSP amplicon) and upstream (UPS BSP amplicon) of the region in which we observed differential methylation (BSP amplicon) (**Figure 3-2**), were selected for bisulfite sequencing analysis and the results are shown below the schematic. Each horizontal line represents a unique clone. Red and blue marks symbolize maternal and paternal alleles, respectively, of strain-specific variants. Open circles represent unmethylated CpGs, while closed circles are methylated CpGs.

A)

	(129S1xPWK)F1	(PWKx129S1)F1
brain	0.48 ± 0.01	0.47 ± 0.06
liver	0.47 ± 0.04	0.49 ± 0.05
kidney	0.48 ± 0.01	0.48 ± 0.02
spleen	0.47 ± 0.04	0.51 ± 0.03
muscle	0.46 ± 0.06	0.47 ± 0.03
tail	0.49 ± 0.04	0.49 ± 0.01
testes	0.47 ± 0.00	0.51 ± 0.08

B)



SNP chr12: 81269902

SNP chr12: 81284503

Figure S3-4: Allelic expression analyses of *Actn1* in diverse mouse tissues shows biallelic expression. A) Results of SNuPE analyses of *Actn1* RNA of adult tissues of 2 females and 2 males of each cross, expressed as average proportion of 129S1 allele ± S.D. B) Examples of *Actn1* cDNA sequence analysis at two polymorphisms.

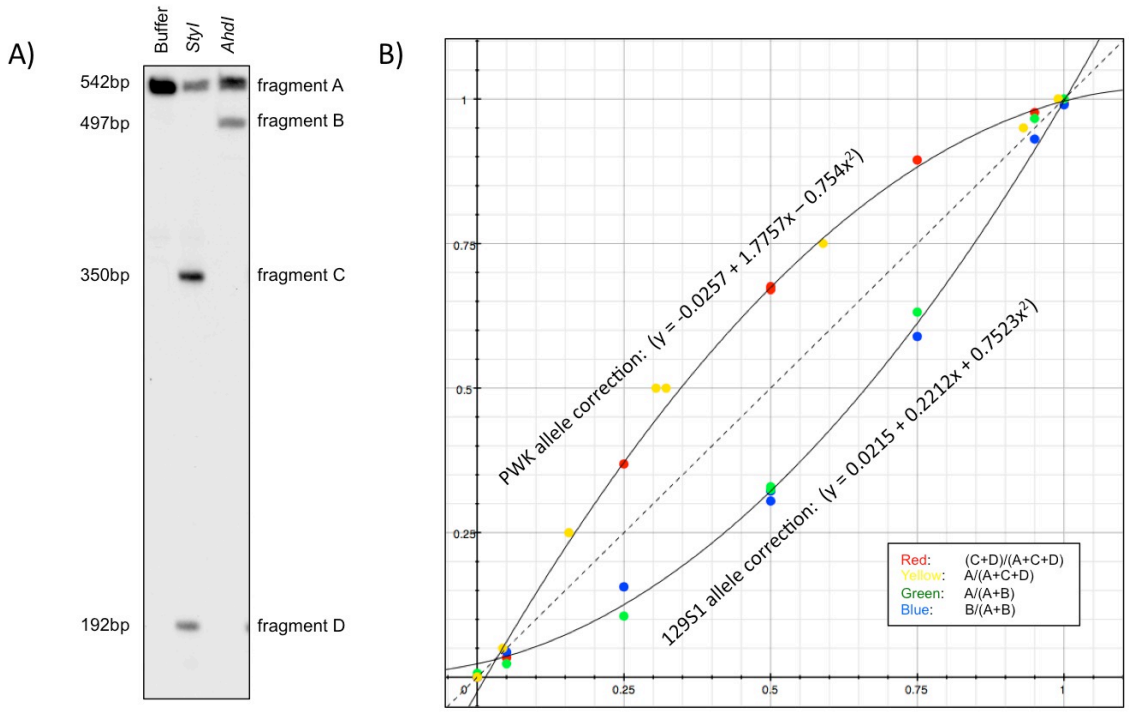


Figure S3-5: Actn1 DMR analysis by RFLP. Panel A shows a sample gel displaying DNA fragments resulting from RFLP analysis of the *Actn1* DMR. The undigested amplicon is arbitrarily named fragment A (542 bp). *Sty*I digestion of this amplicon yields fragments C (350 bp) and D (192 bp). *Ahd*I digestion yields fragment B (497 bp). A smaller, 45 bp fragment is generated from the *Ahd*I digestion but migrates with free αP32-dCTP and, therefore, was not included in the data analysis. **Panel B** shows a plot of artificially created PWK/129S1 allelic ratios for the analysis of MS-RFLP data of *Actn1* DMR. The X- and Y-axes are the fraction of expected and observed methylated parental alleles, respectively. Also shown are the polynomial interpolation equations used to normalize the observed allelic ratios.

(PWKx129S1)F1 *StyI*

	brain	**kidney**	**liver**	**muscle**	**spleen**	**tail**
kidney	1.000					
liver	0.275	0.541				
muscle	1.000	1.000	0.105			
spleen	1.000	0.642	0.025	1.000		
tail	2.40E-06	5.50E-06	2.30E-04	9.50E-07	3.10E-07	
testis	0.002	0.005	0.116	0.001	2.70E-04	0.524

(129S1xPWK)F1 *AhdI*

	brain	**kidney**	**liver**	**muscle**	**spleen**	**tail**
kidney	1.000					
liver	0.276	1.000				
muscle	0.993	1.000	1.000			
spleen	1.000	1.000	0.952	1.000		
tail	0.004	0.037	0.890	0.382	0.017	
testis	0.952	1.000	1.000	1.000	1.000	0.993

(PWKx129S1)F1 *AhdI*

	brain	kidney	liver	muscle	spleen	tail
kidney	0.396					
liver	0.212	1.000				
muscle	1.000	0.179	0.092			
spleen	1.000	0.598	0.396	1.000		
tail	0.000	0.015	0.031	0.000	0.000	
testes	0.941	1.000	1.000	0.595	1.000	0.042

(129S1xPWK)F1 *StyI*

	brain	kidney	liver	muscle	spleen	tail
kidney	1.000					
liver	0.078	1.000				
muscle	1.000	1.000	0.173			
spleen	1.000	1.000	0.855	1.000		
tail	0.007	0.198	1.000	0.019	0.093	
testes	1.000	1.000	1.000	1.000	1.000	0.855

Shown are p-values determined by pairwise T-Tests between tissues ($\alpha < 0.05$)

Table S3-1: Pair wise t-tests of percent maternal methylation at the *Actn1* DMR between tissues. Shown are the *p*-values ($\alpha<0.05$)

name	primer (5'->3')	organism
RFLP-F	ggTTAgAggTCgCTCTCgCCATAC	mouse
RFLP-Rev	TAAggTAggATgTgCTgACgCTgA	mouse
BSP-F	ggTggTTgAgTTgAAAATAATg	mouse
BSP-Rev	TATTTAAAATACCATATACAAATAATA	mouse
BSP-F	gAgTAggAgTAgggTggg	rat
BSP-Rev	AATACCTAACCTTACTAACAAAC	rat
BSP-F	ggTggAggTTgggAgTTg AAA	human
BSP-Rev	ACCTTCCTAACCACTCTCTCCT	human
Actn1-F2	ATCCCATggCTggAgAATCg	mouse
Actn1-PR2	AATggTggTgAgCAgCTgC	mouse
Actn1-RT2	ggTTCTCCACTTCATTgATgg	mouse
Actn1-SnuF	gCAgAgTTTgCCCgAATCATg	mouse
Actn1-SnuR	gCAgTACTCggCTTggTCAg	mouse
Actn1-SnuRT	CCATTCTTgCgATgCAgTAC	mouse
Actn1-SNUPE	CTTATgTCCCgAgAgACggC	mouse
Actn1-F3	CAAAgATTgACCAgCTggAg	mouse
Actn1-NMSM	CAACgACCCCCAgAAgAAg	mouse
Actn1-NM20	gCAACgACCCCCAgggAg	mouse
Actn1-18SM	ACTTTgACCggAAgAAgACg	mouse
AK037382-F	CTCACTTCTTTAAgAggAgAAg	mouse
AK037382-R	TggAAATTCTAACAACAggCTC	mouse
Zfp36l1-F	ACCTTggACAACTCAAgACg	mouse
Zfp36l1-R	TTgTgATTTggCACTTAAggC	mouse
Zfp36l1-RT	CTAAgTTgCTTCTgTAAACgg	mouse
Dcaf5-F	CAATggAgCCTTCATggTgC	mouse
Dcaf5-R	AgATTCCgAgTCAgTgTAgC	mouse
Dcaf5-RT	gCAgTgAggCTgAAgATTCC	mouse
BSP-DNS-F	TAgAAgATgAgTTAgTAAATTT	mouse
BSP-DNS-Rev	CCTAAAAAAAAACACACATTA	mouse
BSP-UPS-F	TgTTTATggTTggAATAgg	mouse
BSP-UPS-Rev	CTTATCTAACACACTTATCATAT	mouse

Table S3-2: List of primers used in the MS-RFLP (RFLP-), Bisulfite-PCR (BSP-), RT-PCR and sequencing or SNuPE (Snu-) analyses.

Tissue	Individuals	Assay	Primers	SNPs analyzed	Expression
Brain, liver, kidney, spleen, tail, muscle & testes	(PWKx129S1)F1 & (129S1xPWK)F1 ♀& ♂ (n=8)	SNuPE	Actn1SnuRT, Actn1SnuF, Actn1SnuR, Actn1SNUPE	chr12: 81269902	Biallelic
Brain, liver, kidney, spleen, tail, muscle & testes	(PWKx129S1)F1 & (129S1xPWK)F1 ♀& ♂ (n=8)	RT-PCR + sequencing	Actn1SnuRT, Actn1SnuF, Actn1SnuR	chr12: 81269902, 81269896 & 81269456	Biallelic
Brain	(PWKx129S1)F1 & (129S1xPWK)F1 ♀& ♂ (n=4)	RT-PCR + sequencing	Actn1RT2, Actn1F2, AcntPR2	chr12: 81284503 & 81274013	Biallelic
E 9.5 embryos and placentas	(PWKx129/S1)F1 & (129S1xPWK)F1 ♀& ♂ (n=4)	RT-PCR + sequencing	Actn1SnuRT, Actn1F3, Actn1SnuR	chr12: 81269902, 81269896 & 81269456	Biallelic
Brain	(PWKx129S1)F1 & (129S1xPWK)F1 ♀& ♂ (n=4)	RT-PCR + sequencing	Actn1SnuRT, Actn1-18SM, Actn1NM20, Actn1NMSM, Actn1SnuF, Actn1SnuR	chr12: 81269902, 81269896 & 81269456	All 3 isoforms biallelic
E 9.5 embryos and placentas	(PWKx129S1)F1 & (129S1xPWK)F1 ♀& ♂ (n=4)	RT-PCR + sequencing	Actn1SnuRT, Actn1F3, Actn1-18SM, Actn1NM20, Actn1NMSM, Actn1SnuR	chr12: 81269902, 81269896 & 81269456	All 3 isoforms biallelic

Table S3-3: Summary of *Actn1* allelic expression analyses performed (see **Table S3-2** for primer sequences)

CHAPTER IV: GENETIC ARCHITECTURE OF SKEWED X INACTIVATION CHOICE IN THE LABORATORY MOUSE[3]

BACKGROUND AND INTRODUCTION

Mammals have a female XX (homogametic) and male XY (heterogametic) allosomal complement. It is postulated that the X and Y chromosomes originated from a common autosomal ancestor that over time have diverged both in gene content and structure [117, 118]. So much so, that the Y chromosome is one-tenth the physical size of the X chromosome and contains 20-fold fewer genes. Without a means to compensate for the unequal number of X chromosomes, males and females would have drastically different transcription levels for X-linked genes. Mammalian females undergo a dosage compensation mechanism called X chromosome inactivation (XCI) that restricts expression to one parental X chromosome per cell. This balances X-linked gene expression between

[3] The following chapter describes work done in collaboration with Dr. Alan B. Lenarcic, John P. Didion, Dr. Jeremy R. Wang, Dr. Jeremy B. Searle, Dr. Leonard McMillan, Dr. William Valdar and Dr. Fernando Pardo-Manuel de Villena. The aim of these experiments was to fine map *Xce,* investigate the haplotype diversity of our new *Xce* candidate interval in species and subspecies of *Mus* and *Mus musculus*, and characterize the refined *Xce* candidate interval. I conducted the mouse breeding (with the exception of the crosses reported previously [115, 116]), sample preparation, and the design and implementation of the pyrosequencing assays. I significantly contributed to the manuscript writing and the figures were of my design (with the exception of **Figure 4-2**). These results and a portion of the introduction were published in PLoS Genetics (Calaway *et al.* 2013). **Figure 4-1** is from an unpublished study currently under review (Crowley *et al.* submitted). I was involved in the mouse dissections and sample preparation, RNAseq library preparation, and I provided input into the X chromosome inactivation portion of the manuscript including Supporting Figure design. This work was done in collaboration with James J Crowley, Vasyl Zhabotynsky, Wei Sun, Shunping Huang, Isa Kemal Pakatci, Yunjung Kim, Jeremy R Wang, Andrew P Morgan, David L Aylor, Zaining Yun, Timothy A Bell, Ryan J Buus, Mark E Calaway, John P Didion, Terry J Gooch, Stephanie D Hansen, Nashiya N Robinson, Ginger D Shaw, Jason S Spence, Corey R Quackenbush, Cordelia J Barrick, Yuying Xie, Dr. William Valdar, Alan B Lenarcic, Wei Wang, Catherine E Welsh, Chen[3]-Ping Fu, Zhaojun Zhang, James Holt, Zhishan Guo, David W Threadgill, Lisa M Tarantino, Darla R Miller, Fei Zou, Leonard McMillan, Patrick F Sullivan, Fernando Pardo-Manuel de Villena.

males and females (**Figure 4-1A**). To compensate for the hemizygous state of the X chromosome, it was postulated by Ohno in 1967 that the X chromosome evolved to upregulate expression 2-fold to match the expression levels of the autosomes [118]. **Figure 4-1B** demonstrates that X-linked genes are indeed expressed at twice the level of autosomal genes.

X chromosome inactivation (XCI): a paradigm of genetic-epigenetic regulation

For simplicity, the XCI process may be divided into five discrete steps or stages: counting/sensing, choice, initiation, spreading and maintenance. The latter three stages

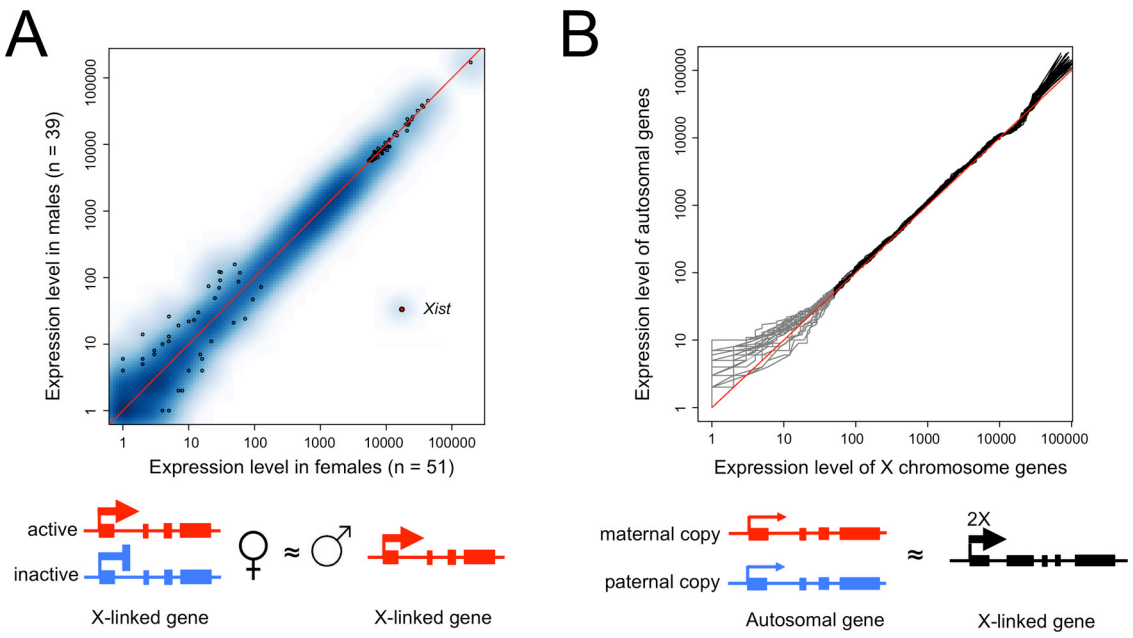


Figure 4-1: Global analysis of X inactivation and dosage compensation. Panel A shows mean expression values for each gene on the X chromosome for males (n=39) versus females (n=51). The 1:1 linear relationship indicates that, as expected, inactivation of one X in females creates roughly equivalent expression levels between the sexes. **Panel B**: For each of 90 animals, a distribution of gene expression levels was generated for autosomal and X chromosome genes separately. These distributions were then plotted against each other as shown, with one line per animal. The result was a roughly 1:1 relationship in the levels of expression from the autosomes and X chromosome. These results support Ohno's hypothesis that X-linked genes are expressed at roughly two times the level of autosomal genes [118].

have been the subject of intense study over the past two and a half decades [9], while the mechanisms and molecular players of the first and second stages of XCI (counting/sensing and choice) remain elusive to this day. This is largely due to technical and biological challenges associated with the initial stages of XCI process.

Sensing/counting and choice occur early during development within a small timeframe (in mouse, random choice is initiated between embryonic days (E) 5.5-6.3 [119, 120]). This poses technical challenges for isolating embryos: 1) without contamination with extraembryonic or maternal tissue, 2) at the correct developmental time point, and 3) with enough tissue for molecular study. Embryonic stem cells (ES) and induced pluripotent stem cell lines (iPS) have been used to address some of these problems. However, care must be taken to ensure that cultured ES and iPS cells do not have gross karyotypic abnormalities [121].

The XCI process is regulated by a locus called the *X-inactivation center (Xic)* that resides near the center of the X chromosome (100.3 Mb, NCBI37/mm9) and is required for the initiation and spreading of XCI in *cis* [122]. The inactivation of one X chromosome is preceded by upregulation of a long non-coding RNA called the *X-inactive specific transcript* (*Xist*) [123, 124]. *Xist* coats the X chromosome from which it was transcribed and elicits a wave of epigenetic reprogramming including DNA methylation [37] and histone post-translational modifications [9, 10, 85]. It is these epigenetic changes and not *Xist* that ultimately maintain the inactive X chromosome state [125]. The active and inactive X chromosomes are morphologically distinguishable with the inactive X physically condensed (facultative heterochromatinized) [85], and devoid of any evidence of transcription demonstrated by a lack of RNA polymerase II association [126].

Before *Xist* expression, two critical steps must occur. First, the cell must 'sense' its sex and 'count' the number of X chromosomes. And secondly, if female, it must 'choose' to inactivate all but one parental X chromosome. There have been a few models proposed

over the years to explain the sensing and counting mechanism that include blocking factors that originate from the autosomes or X chromosomes [28], two-factor blocking [127], the sensing and counting mechanism [128], the stochastic model [129], and the feedback loop model [130]. The counting mechanism requires communication between the autosomes and X chromosomes demonstrated by the relative number of active X chromosomes in human and rabbit triploid and tetraploid cells [131-133]. The apparent interconnection between the sensing and counting mechanisms alludes to a common regulatory pathway.

XCI choice

The choice of which parental X chromosome undergoes XCI occurs within the developmental timeframe of E5.5-E6.3 [119, 120]. In mouse, the maternal X chromosome remains active during the preimplantation stages [134, 135], while the paternal X chromosome is preferentially inactivated until after ~E3.5 [119]. At which time, epiblast cells within the embryo proper reactivate the paternal X chromosome and random choice is poised to occur [136-139]. The paternal X chromosome remains preferentially inactivated in the extraembryonic tissues [119]. By an unknown mechanism, each cell randomly chooses to inactivate one of the two parental X-chromosomes and then commits to that choice by initiating a cascade of transcriptional and epigenetic regulation that modifies both chromosomes to distinguish the future inactive X from the active X [31, 37, 140-143]. The initial choice each epiblast cell makes is preserved and transmitted mitotically to all its daughter cells [144]. As a result, each female is a unique mosaic of somatic cells that express either the maternally or paternally derived X chromosome. The degree of mosaicism (overall ratio and spatial distribution of cells) is determined by the initial number of cells that undergo independent choice, by the developmental fate of each epiblast cell and its multiplication rate.

Genetics of XCI choice: *The X chromosome controlling element (Xce)*

The role of genetics in XCI choice was initially discovered by skewed XCI ratios in female hybrids between certain stocks derived from classical inbred mouse strains. These female hybrids, on average, preferentially inactivated one X chromosome over the other in a strain dependent manner [145, 146]. The effect was later mapped to a single location on the X chromosome and given the name *X-chromosome controlling element* (*Xce*) for its role in XCI choice [147]. Since its initial discovery, four functional alleles of *Xce* have been characterized in *Mus* inbred strains*, (*Xce^a*,* Xce^b*,* Xce^c and Xce^d) and are distinguished by their relative resistance or susceptibility to inactivation [145, 148-153]. The four *Xce* alleles form an allelic series of XCI skewing, the magnitude and direction of which depends on the *Xce* genotype of the female. Furthermore, XCI skewing is only observed in *Xce* heterozygotes while female homozygotes display no preference towards inactivating either parental X chromosome [154]. The order of *Xce* allele strength is $Xce^a < Xce^b < Xce^c < Xce^d$ (**Figure 4-2A**). In other words, in female heterozygotes the X chromosome carrying the stronger *Xce* allele has a higher probability of remaining active and thus, these females will have a larger number of cells with that X chromosome active (**Figure 4-2B**). From a genetic standpoint, alleles at *Xce* are overdominant and therefore *Xce* acts in *cis*.

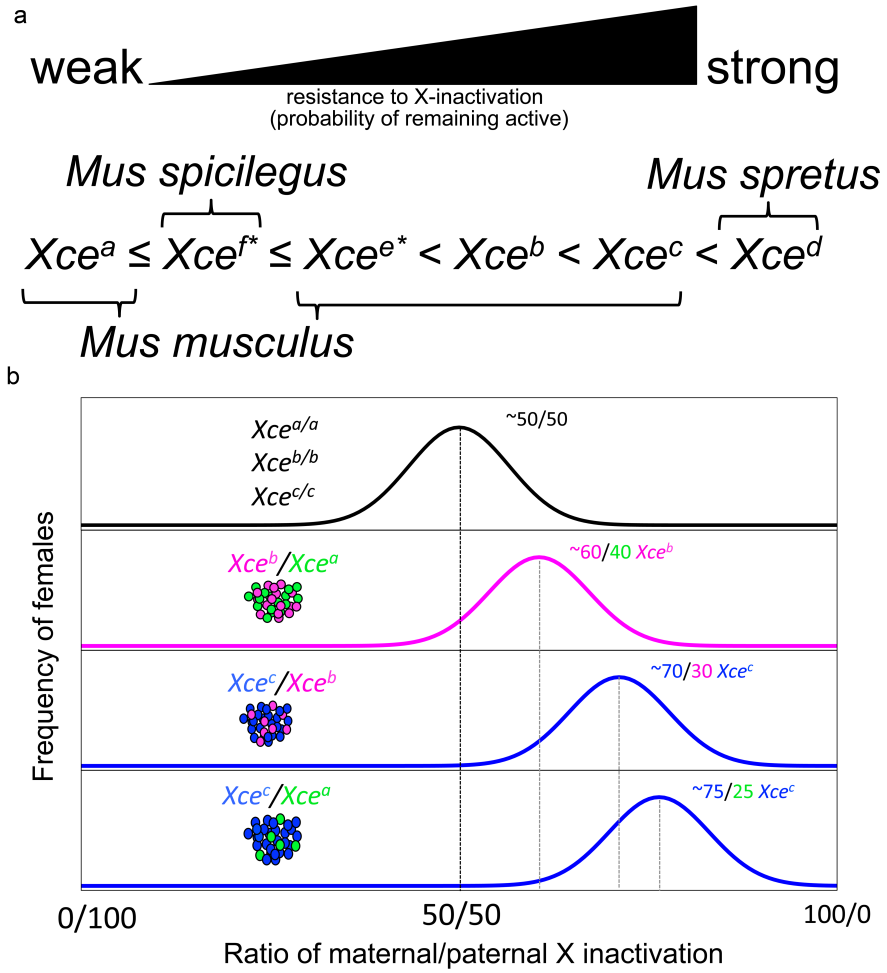


Figure 4-2. The *Xce* allelic series. Panel (A) shows the order of *Xce* allele strength. Panel (B) shows hypothetical distribution and mean XCI ratio skewing in female populations that are either homozygous or heterozygous for *Xce* alleles.

Xce has been mapped within a 1.85 Mb candidate interval that overlaps with the current definition of the *X inactivation center* (*Xic*) which includes three long non-coding RNAs *Xist, Tsix* and *Xite* that play major roles in murine XCI [155]. It has been postulated that the *Xce* allelic series can be explained by genetic variation within these long non-coding RNAs, specifically *Xite* [156]. An alternative hypothesis is that XCI choice is controlled by X-linked and autosomal dosage factors [157-159] and thus *Xce* would serve as a binding site for a *trans*-acting factor(s) that influences *Tsix* or *Xist* expression [158-161]. Nonetheless, the identity of *Xce* remains unknown. This is in part due to the technical challenges of

measuring XCI choice and to the relatively high level of stochastic variation in XCI in isogenic female populations, which together make it difficult to infer with certainty the *Xce* allele present in an individual female (**Figure 4-2B**). Mapping *Xce* is further complicated by the comparatively low recombination rate of the X chromosome and the fact that only females are informative for the phenotype.

Parent-of-origin effects, autosomal modifiers and secondary skewing

Although *Xce* is the major locus controlling XCI choice, previous studies have demonstrated that parent-of-origin and autosomal factors significantly influence XCI choice [148, 149, 162-164]. A large mapping experiment identified suggestive loci on five autosomes but none reached genome–wide significance [155]. The parent-of-origin effect was first described by Forrester and Ansell in 1985 as a difference in XCI skewing depending on whether the Xce^{c} allele was maternally or paternally inherited in $Xce^{c/b}$ heterozygotes. The evidence available at the time, however, could not discriminate among *Xce*, another X-linked locus or autosomal loci. A more recent study provided additional evidence of a parent-of-origin effect and postulated that its cause could be *Xce* itself or epigenetic differences of one or more X-linked loci [163]. The same study showed an increased variance in XCI skewing in F2 females heterozygous for the same combination of *Xce* alleles as F1 hybrids, indicating the existence of autosomal factors that influence XCI choice [163]. A more recent study used mouse lines with recombinant X chromosomes derived from two genetically divergent mouse inbred strains (129S1/SvImJ and CAST/EiJ) to show that multiple regions along the X chromosome influence XCI choice, but was unable to map any of them, including *Xce* [165].

Lastly, there are well-documented cases of secondary XCI skewing that influence the XCI patterns observed in adults [166-168]. Secondary skewing occurs when an X linked mutation impacts cell survival or proliferation.

Challenges of mapping *Xce*

Technical issues associated with measuring XCI choice further complicate the identification *Xce*. A well-established surrogate for XCI choice is X-linked allele-specific gene expression. Nonetheless, gene expression in a female mouse can be influenced by many factors in addition to XCI choice itself. And thus, it is important to carefully choose X-linked genes that most accurately reflect the true ratio of XCI while minimizing the presence of misleading factors such as differential expression due to *cis*-acting regulatory variants, tissue-specific skewing, or XCI escape. As a general rule, estimation of XCI skewing improves with the number of X-linked genes used.

In this study, we developed an approach that overcomes major challenges of mapping *Xce.* Our approach is based on association mapping of XCI skewing phenotypes in classical inbred strains that have recently been genotyped at very high density [1] or had their genome sequenced (whole genome sequence, WGS) [3]. Our analysis was restricted to the previously defined candidate interval [155] and generated a new candidate interval of much smaller size. By generating multiple F1 hybrid females between inbred strains we accurately determined the mean and the variance in XCI ratio within genetically identical mice. We also generated reciprocal crosses to determine the parent-of-origin effects. Lastly, we performed these analyses in multiple tissues and thus determined whether tissue choice had an effect on the estimation of skewing of XCI. In order to analyze the X-linked expression phenotype data we developed a hierarchical Bayesian model and inference procedure that allows to us to estimate both the mean and the variability of XCI within an individual female or female population. We extended our phenotyping to wild-derived inbred strains with different haplotypes of known subspecific origin [1], and used these data to reconstruct the evolutionary history of the *Xce* locus itself.

RESULTS

Association mapping based on public data narrows the *Xce* candidate interval to 194 kb

In our initial approach to reduce the candidate interval we first identified a subset of inbred mouse strains that had both a known *Xce* allele and high-density genotype [1] or sequence data [2] available. Over the past four decades, several inbred mouse strains have been phenotyped for XCI skewing and these strains include representatives of each one of the four known *Xce* alleles (**Figure 4-3**). At the *Xce* candidate interval defined by Chadwick and coworkers (2006), referred hereafter as the Chadwick interval, these strains have haplotypes derived from two different *Mus* species, *Mus spretus* and *Mus musculus*, and two subspecies of the latter, *M. m. castaneus* and *M. m. domesticus* [1]. Two strains, CAST/EiJ and SPRET/EiJ, cannot be used to refine the candidate interval using single locus association mapping techniques because they are singletons for both an *Xce* allele and the specific or subspecific origin (**Figure 4-3**). The remaining 25 strains are almost evenly distributed between Xce^{a} and Xce^{b} carriers and all have a *M. m. domesticus* haplotype in the candidate interval [1]. Furthermore, all of them are classical inbred strains descended from a small pool of founders and thus it is reasonable to assume that they share by descent the same causative genetic variant for their differences in *Xce* alleles. Eleven of these strains (or a closely related sister strain) have been genotyped at high density and eight have been sequenced [1, 2]. Importantly, both alleles are represented among genotyped and sequenced strains (Xce^{a}, seven genotyped and five sequenced strains and Xce^{b}, four genotyped and three sequenced strains, **Figure 4-3**).

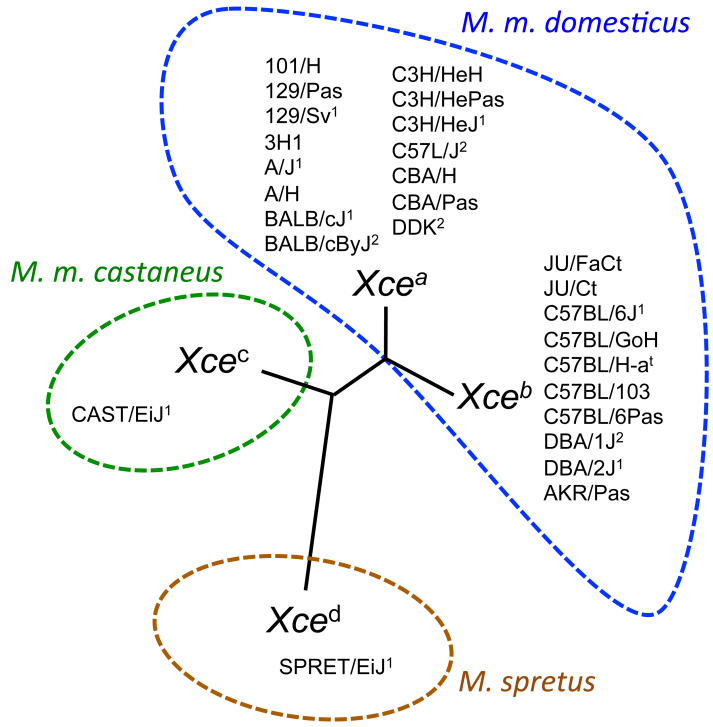


Figure 4-3. Inbred mouse strains with known *Xce* phenotype and their phylogenetic relationship. Shown is a phylogenetic tree that reflects the sequence divergence within the Chadwick candidate interval for inbred mouse strains with known *Xce* alleles. Inbred strains with a number one superscript have both MDA and Sanger sequencing information available, while mouse strains with a number two superscript have only MDA genotype data available. Inbred strains with no number are assumed to have identical genotypes to a closely related strain that has been genotyped. Blue and green shading denotes the subspecific origin of the Chadwick interval for each strain (*M. m. domesticus* and *M. m. castaneus,* respectively).

For every SNP and indel present within the Chadwick interval, we determined the pattern of allelic similarities and differences among the subset of inbred strains with known *Xce* alleles (Strain Distribution Pattern, SDP: **Figure S4-1**) [67, 169]. SDPs were then classified into three categories based on consistency between phenotype and genotype: 1) fully consistent with the *Xce* phenotype (black tick marks), 2) inconsistent with the *Xce* phenotype (red tick marks), or 3) partially consistent (gray tick marks) (**Figure 4-4 and Table S4-1**). We focused our association analysis within the Chadwick interval, which is based on genetic mapping in populations segregating for the Xce^a, Xce^b, and Xce^c alleles.

Analysis of Mouse Diversity Array (MDA, [62]) genotypes and sequence data shows an enrichment of consistent SDPs (eight MDA SNPs, 120 Sanger SNPs and indels) at an 194 kb interval spanning from rs29082048 to Sanger Mouse Genomes Project (SMGP) SNP position at 100,119,750 bp (**Table S4-1**). This interval does not contain any inconsistent SNPs. In addition, there are 23 SNPs with consistent SDPs randomly distributed throughout the distal portion of the Chadwick candidate interval (**Figure 4-4**). These SNPs do not cluster and this region is punctuated with inconsistent SNPs.

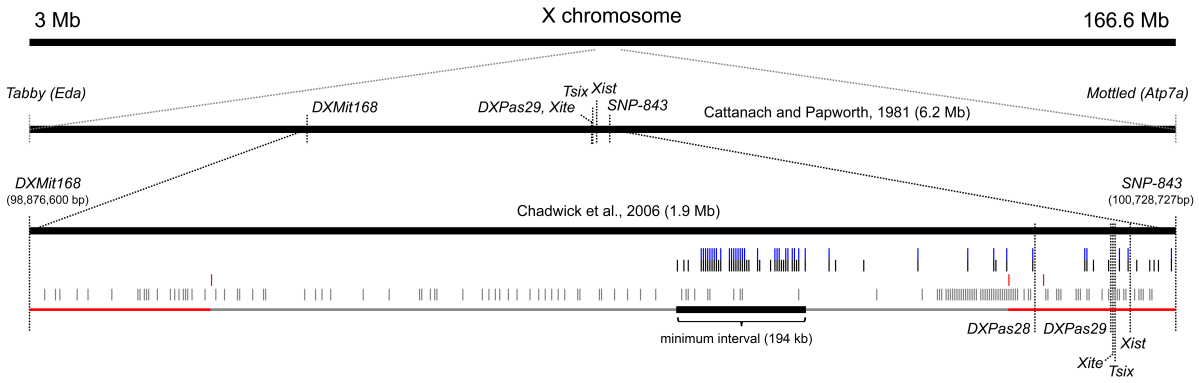


Figure 4-4: The *Xce* candidate interval based on historical data. Shown is a physical map that shows the locations of the previous *Xce* candidate intervals [155, 170]. Below the historical candidate intervals are the results of the SDP analyses using inbred strains selected from Panel A (See Methods). Tick marks represent SDPs classified as consistent (black), inconsistent (red), and partially consistent (gray). SNPs that retain consistent SDPs after inclusion of ALS/LtJ, LEWES/EiJ, PERA/EiJ, SJL/J, TIRANO/EiJ, WSB/EiJ, and ZALENDE/EiJ in the analysis are shown as blue tick marks above consistent SDPs. Our new maximum candidate interval is shown in gray below the tick marks. The minimum candidate interval is shown in black, while regions excluded are shown in red.

We conclude that the minimum *Xce* candidate interval is located approximately 558 kb proximal to *Xist* (note that the maximum *Xce* candidate interval based on this analysis spans from inconsistent SMGP-SNP at position 99,091,507 bp to inconsistent SMGP-indel at 100,460,107 bp). Within this candidate interval all phenotyped strains with the Xce^a allele share the same haplotype and all strains with the Xce^b allele share a different haplotype based on MDA genotypes.

XCI skewing in experimental F1 hybrids derived from inbred strains within unknown *Xce*.

Our ability to reduce further the *Xce* candidate interval depended on the number of inbred strains with known *Xce* allele and high-density genotype data available. Ideally we would like to phenotype inbred strains that have Xce^{a} and Xce^{b} recombinant haplotypes in the candidate interval. Furthermore, we would like to characterize the *Xce* alleles of additional *M. m. domesticus* strains with haplotypes that are not associated with known *Xce* allele carriers. These strains will provide additional information about *Xce* functional diversity within *M. m. domesticus* and depending on their *Xce* phenotype, may further refine the *Xce* candidate interval. We selected three strains with $Xce^{a/b}$ recombinant haplotypes ALS/LtJ, SJL/J and WLA/Pas because of their availability and their ability to refine further the new candidate interval. Based on phylogenetic analysis of the new candidate interval (See Methods), we selected six wild-derived inbred strains, PERA/EiJ, TIRANO/EiJ, ZALENDE/EiJ, LEWES/EiJ, and WSB/EiJ to represent each of the major haplotypes present in *M. m. domesticus* (with the exception of *b3* which has only been observed in wild mice). We selected PWK/PhJ to characterize the *Xce* allele in a third *M. musculus* subspecies, *M. m. musculus*. We selected WSB/EiJ and PWK/PhJ because they are wild-derived strains of *M. m. domesticus* and *M. m. musculus* origin, they have available whole genome sequence [2] and they are founder strains in mouse genetic resources such as the CC [81] and DO [64]. Finally, we selected PANCEVO/EiJ to characterize the *Xce* allele present in a third species of mouse, *Mus spicilegus.* A summary of the justification for selecting each mouse strain and the information it provided towards mapping *Xce* is provided in Table S4-2.

To determine which *Xce* allele is present in each strain, we generated genetically defined F1 female hybrids by crossing the unknown strain to inbred strains with well-characterized *Xce* alleles: Xce^{a}, A/J and 129S1/SvImJ; Xce^{b}, C57BL/6J; and Xce^{c},

CAST/EiJ. To estimate the presence, direction and extent of XCI skewing in each F1 hybrid female, we developed highly quantitative pyrosequencing assays and measured allele-specific X-linked gene expression (see **Methods**). On average, for each strain with an unknown *Xce* allele, we tested allele-specific expression in 69 F1 females (ranging from 40 to 120 females per strain, **Table S4-3**).

To analyze and integrate the X-linked expression data set, we developed a hierarchical Bayesian model and inference procedure. The method is described briefly in the Methods section, and full description will be reported elsewhere [171]. Briefly, our model parameterizes gene-tissue bias and precision, parent-of-origin effects, and genetic background effects (strain) to account for gross sources of uncertainty and error associated with our XCI phenotyping method. This allows us to combine the different gene measurements and tissues from individual females and establish a mean XCI ratio (see **Materials and Methods**) for a given cross.

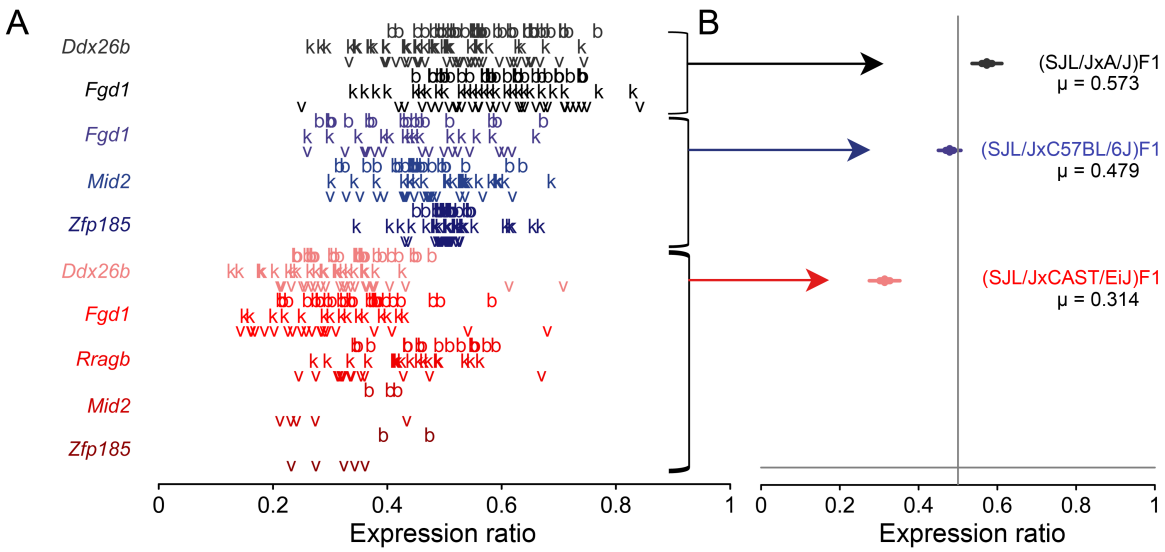
For each F1 cross, we tested whether the two parental strains carry the same *Xce* allele. **Figure 4-5** shows the gene expression data (**panel A**) and posterior mean and confidence intervals inferred from it (**Panel B**) for the SJL/J F1 crosses performed. The posteriors in **Panel B** estimate the mean inactivation proportion associated with each cross. They show where and how posterior probability for the underlying cross mean is concentrated on the scale of 0 (representing full maternal inactivation) to 1 (representing full paternal inactivation), with 0.5 indicating a cross average of about 50% paternal and maternal X-inactivation. By choosing regions of 95% posterior coverage, we see that the data allows us to measure mean X inactivation proportions accurately within 7.7% (+/−5%), placing for instance, the (SJL/JxCAST/EiJ)F1 firmly to the left of 50%, around 33.6% of cells with an active SJL/J X chromosome. As a rule, when a distribution shows a strong bias, in other words, when most of the posterior is concentrated on one side of 0.5 boundary, we use this as evidence to conclude that the two strains involved the cross have functionally

different *Xce* alleles. To quantify this bias, we used the tail posterior probability (*i.e*., the amount of posterior probability that lies on the side of 0.5 line, **Figure 4-5C**. These tail probabilities are like *p*-values and their small values strongly support the presence of skewed XCI.

Using this approach, we conclude that seven inbred strains, ALS/LtJ, SJL/J, LEWES/EiJ, PERA/EiJ, TIRANO/EiJ, WSB/EiJ and ZALENDE/EiJ carry an Xce^{b} allele (**Figure 4-5** and **Figure S4-3**). The *M. m. musculus* strain, PWK/PhJ has a new allele, named herein Xce^{e}. Within the allelic series, the strength of this new allele falls between Xce^{a} and Xce^{b} (**Figure 4-2A**). Finally, PANCEVO/EiJ has an allele that is similar in strength to Xce^{a} (**Figure S4-3**). The results for the WLA/Pas strain are inconclusive and will be discussed later.

Incorporation of the ALS/LtJ and SJL/J strains to our association mapping further reduced the proximal boundary of the new *Xce* candidate interval by 9.6 kb. Furthermore, by including ALS/LtJ, SJL/J, LEWES/EiJ, PERA/EiJ, TIRANO/EiJ, WSB/EiJ and ZALENDE/EiJ into our SDP analysis, we reduced the number of SNPs with consistent SDPs within the *Xce* interval to 69 and further reduced the proximal boundary by 8.2 kb (**Figure 4-4B**, blue tick marks and **Table S3-4**). The minimum refined *Xce* candidate interval is bounded by SMGP-SNPs at positions 99,943,259 bp and 100,119,750 bp.

Outside of the refined candidate interval but within the Chadwick interval only 14 SNPs (WGS and MDA data) have consistent SDPs (**Table S4-4**). These SNPs (highlighted blue in **Figure 4-4**) do not cluster and are interspersed with SNPs with inconsistent SDPs. Lastly, only three SNPs on the entire X chromosome (rs29079362, rs13483921 and rs29081860) outside of the Chadwick interval have SDPs consistent with the *Xce* alleles.



C

Strain	Crossed to strains	# females phenotyped	Exlcuded *Xce* alleles	*Xce* allele (accepted or not rejected)
ALS/LtJ	A/J (*Xce*^{a}), C57BL/6J (*Xce*^{b}), CAST/EiJ (*Xce*^{c})	120	*Xce*^{a} (.08), *Xce*^{c} (<.001)	*Xce*^{b} (.0009)*
SJL/J	A/J (*Xce*^{a}), C57BL/6J (*Xce*^{b}), CAST/EiJ (*Xce*^{c})	81	*Xce*^{a} (.001), *Xce*^{c} (<.001)	*Xce*^{b} (.142)
WLA/Pas	129S1/SvImJ (*Xce*^{a}), SJL/J (*Xce*^{b})	67	*Xce*^{a} (.009)	*Xce*^{b} (.45)
LEWES/EiJ	129S1/SvImJ (*Xce*^{a}), SJL/J (*Xce*^{b})	45	*Xce*^{a} (<.001)	*Xce*^{b} (.68)
WSB/EiJ	129S1/SvImJ (*Xce*^{a}), A/J (*Xce*^{a}), C57BL/6J (*Xce*^{b}), CAST/EiJ (*Xce*^{c})	40	*Xce*^{a} (.02), *Xce*^{c} (.004)	*Xce*^{b} (.54)
PWK/PhJ	129S1/SvImJ (*Xce*^{a}), A/J (*Xce*^{a}), C57BL/6J (*Xce*^{a}), CAST/EiJ (*Xce*^{c}), WSB (*Xce*^{b})	66	*Xce*^{a} (.008), *Xce*^{b} (.004), *Xce*^{c} (.03)	*New*
TIRANO/EiJ	DDK/Pas (*Xce*^{a}), C57BL/6J (*Xce*^{b})	81	*Xce*^{a} (<.001)	*Xce*^{b} (.510)
ZALENDE/EiJ	DDK/Pas (*Xce*^{a}), C57BL/6J (*Xce*^{b})	82	*Xce*^{a} (<.001)	*Xce*^{b} (.08)
PERA/EiJ	C57BL/6J (*Xce*^{b})	42		*Xce*^{b}(.10)
PANCEVO/EiJ	DDK/Pas (*Xce*^{a}), C57BL/6J (*Xce*^{b})	70	*Xce*^{b} (<.001)	*Xce*^{b} (.08)

Figure 4-5: Allelic imbalance in selected female F1 hybrids. Panel A is a plot of the allele-specific expression data from F1 hybrids, where each colored letter represents an individual gene measurement from brain ("b"), kidney ("k"), and liver ("v") from an individual female. Panel B is a plot of the posterior mean and 95% credibility intervals for XCI fraction inferred for each genetic cross, based on our statistical model. Throughout, the x-axis reports the fraction of X-linked allele-specific expression from the strain with the unknown *Xce* allele. The color of each letter (on the right) and each corresponding posterior (on the left) denote the known *Xce* allele to which it is paired: black *Xce*^{a}; blue *Xce*^{b} and red *Xce*^{c}. Panel C shows the inbred strains phenotyped for *Xce*, the strains each were crossed to, the total number of F1 females tested and the *Xce* alleles excluded and included based on posterior tail probabilities.

Analysis of the *Xce* Candidate Interval Reveals a Set of Segmental Duplications Associated with Each Functional *Xce* Allele

After phenotyping of the additional strains, the minimum candidate interval spans 176 kb and its size and relative position with respect to the *Xic* does not change in the latest mouse genome assembly (GrCm38/mm10). The final interval contains five protein coding genes, six pseudogenes, and three novel rRNAs. The G+C content is elevated compared to the X chromosome average (44% *versus* 39%, respectively [172]). Repeat masker

[173] identified 50 LINEs and 60 SINEs as well as 194 other DNA features such as LTRs and regions of low complexity. However, the most dramatic feature is the presence of a set of tandem duplications and inversion (**Figure 4-6A**). The NCBI37/mm9 (and the GrCm38/mm10) reference assembly contains four tandem duplications and one inversion herein referred as segmental duplication (SD) 1 (99,909,337–99,942,773 bp), SD2 (99,940,942–99,961,388 bp), SD3 (99,959,575–100,013,166 bp), SD4 (100,013,346–100,035,061 bp), and inversion (I) 5 (100,040,370–100,084,982 bp) (**Figure 4-6A**). The average size of the duplications is 35 kb, the C+G content is 45%, and they typically span three genes, nine LINEs and 13 SINEs. The phylogenetic tree reveals that two pairs of duplications (SD1 and SD2 and SD3 and I5) are relatively recent events while duplication 4 is the oldest (**Figure 4-6B**). The topological arrangement of these SDs cannot be explained simply by a set of tandem duplications. In particular, the phylogenetic origin, location and orientation of SD3, SD4 and I5 requires both an inversion and a deletion after the duplication event of their common ancestor (**Figure 4-6B**).

Because genotypes in segmental duplications are notoriously unreliable [1, 62], we investigated whether probes designed to track the duplications in the newly released MegaMUGA array (to be reported elsewhere) support our haplotype assignment and mapping conclusions. The MegaMUGA array was designed on Illumina's (San Diego, CA) Infinium BeadChips platform that consistently produces high signal-to-noise ratio compared to conventional hybridization based arrays as demonstrated by previous studies [174, 175].

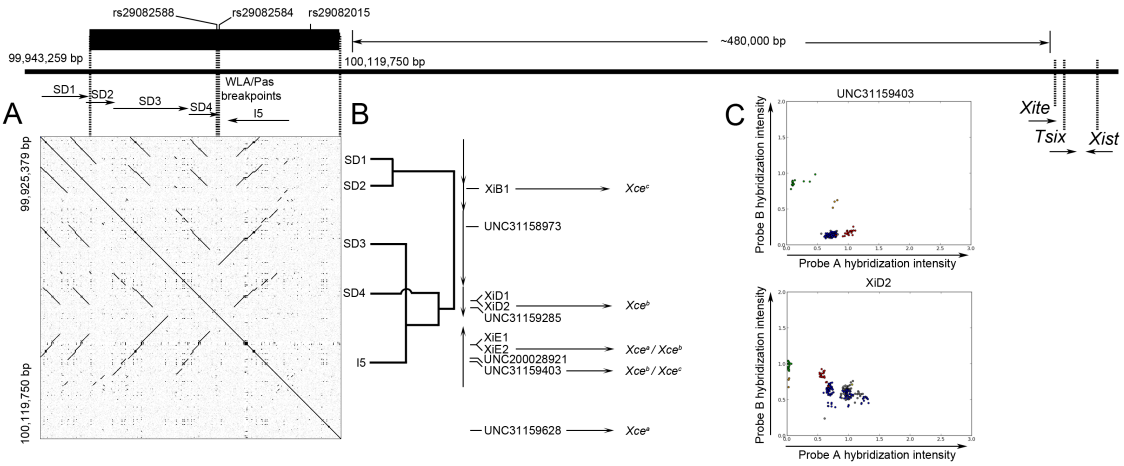


Figure 4-6. Sequence analysis of the candidate interval. In panel A, the candidate interval is show as a thick black bar. Below the candidate interval is a dotplot generated from pairwise sequence concordance in the mm9 genome assembly. Diagonal lines slanting down from left to right are duplications, while diagonal lines slanting up from left to right are inversions. Above the dotplot are arrows that show the four duplications (SD1-4) and inversion (I5) identified. Panel B is a phylogenetic tree that depicts the relationship between the duplications. The phylogenetic tree was generated using the CLUSTALW2 alignment software [4]. Also shown are the ten MegaMUGA markers used for the PCA analysis and their positions in relation to the segmental duplications. Shown in panel C are probe hybridization plots for two of these markers, UNC31159403 and XiD2 (all plots are provided in **Figure S4-2**). The axes represent hybridization intensities for probes tracking alternative alleles at each marker. The colors correspond to the different functional *Xce* alleles: gray Xce^a; blue Xce^b; red Xce^e; green Xce^c; yellow Xce^d. Note that these plots do not agree with the expectations for standard biallelic variants. Typically biallelic variant plots show three distinct clusters representing homozygous A, homozygous B, or heterozygous A/B.

These probes (**Figure 4-6B, C** and **Table S4-5**) consist of standard SNPs and probes with off target variants (VINOs) [67, 176] in addition to probes designed specifically to target the five duplications within the *Xce* candidate interval. Haplotype inference based on probe hybridization has been used successfully in other mouse populations such as the CC [81, 176]. We found a striking consistency between the haplotypes defined by nominal genotypes and the haplotypes based on principal component analysis (PCA) of probe intensities in the segmental duplications. In fact, MegaMUGA probe intensities perfectly partition all mouse inbred strains according to their experimentally defined *Xce* alleles.

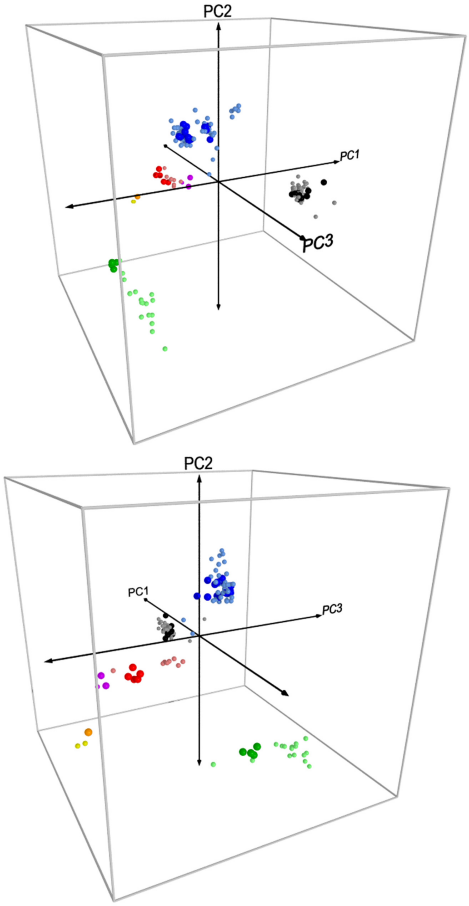


Figure 4-7. Principal component analysis of *Xce* MegaMUGA probes. This figure shows a three-dimensional PCA plot based on hybridization intensity of ten MegaMUGA probes (Figure 4-6 and Table S4-10) within the refined *Xce* candidate interval. Mouse strains with known *Xce* alleles are shown as large spheres, while predicted mouse strains and wild-mice are shown as smaller spheres. Mouse samples are shaded according to *Xce* allele or *Xce* haplotype: Known Xce^a allele, black; predicted Xce^a allele, gray; known Xce^b allele, blue; predicted Xce^b allele, light blue; known Xce^c allele, green; predicted Xce^c allele, light green; known Xce^d allele, orange; predicted Xce^d allele, yellow; known Xce^d allele, orange; predicted Xce^d allele, yellow; known Xce^e allele, red; predicted Xce^e allele, pink; known Xce^f allele, magenta.

This is true not only for Xce^a and Xce^b carriers, but also for known Xce^c, Xce^d, Xce^e, and Xce^f carriers (**Figure 4-8**). We extended this approach to analyze 110 genotyped samples with unknown *Xce* alleles (**Figure 4-7** and **Table S4-10**). Samples with *M. m. domesticus* haplotypes in the candidate interval are partitioned into two groups corresponding to known carriers of Xce^a and Xce^b alleles, matching perfectly the results obtained by standard phylogenetic analysis. In addition, we found that wild-derived inbred strains as well as wild-caught mice with *M. spretus*, *M. spicilegus*, *M. m. castaneus* and *M. m. musculus* haplotypes cluster with the appropriate known carriers of an Xce^d, Xce^f, Xce^c, and Xce^b, respectively. We note that the probes used in the PCA do not share sequence similarity and they do not track homologous regions within the duplications and inversion. Finally, no single probe (nor pair of probes) is able to partition all samples according to *Xce* haplotype or functional allele. There are, however, certain probes that contribute to

the partitioning of the *Xce* alleles more than others (highlighted in **Figure 4-6B**). These results indicate that no single probe can explain the *Xce* allelic series and that each probe does not track a different *Xce* allele.

Phylogenetic analysis of the *Xce* candidate interval

To investigate the evolutionary history of the *Xce* locus, we generated phylogenetic trees based on genotype or sequence data (depending on availability) within the final minimum *Xce* candidate interval for 99 classical inbred strains, 66 wild-derived inbred strains and 124 wild-caught mice (**Figure 4-8** and **Table S4-6**). This tree partitions these samples among five taxa, *M. spicilegus*, *M. spretus*, *M. m. castaneus*, *M. m. musculus* and *M. m. domesticus* that are consistent with previous studies [1, 2].
The *Xce* phenotype has been determined for at least one strain from each one of these taxa (**Table S4-6**). We found that each taxon (species or major subspecies) has a different functional *Xce* allele and there is no evidence of shared of alleles among taxa (**Figure 4-8**). Skewed XCI is present in all crosses between wild-derived strains belonging to different taxa. In contrast, skewing is not present in crosses involving strains from the same taxon. Within the *M. m. domesticus* subspecies we identified five haplotypes (*a*, *b1*, *b2*, *b3* and *b4*). The *a* haplotype is associated with Xce^{a} while two haplotypes, *b1* and *b2* are associated with Xce^{b}. The *b3* haplotype can be explained as recombination between a proximal *b2* and distal *b1* haplotype. The *b3* haplotype has been observed in either a small mouse population on the Farallon islands off the coast of San Francisco, CA, and in one wild-caught mouse from Barcelona, Spain. The *b4* haplotype appears to be a recombination between the *a* and *b1* haplotypes and is found only in the WLA/Pas strain that carries an ambiguous *Xce* allele.

Interestingly, there is an unequal distribution in the number and origin of *M. m. domesticus* stocks that carry each haplotype. For example, classical inbred strains are almost evenly divided among the *a* haplotype (*n* = 52) and the *b1* haplotype (*n* = 47) (**Figure**

4-8). One classical inbred strain, CE/J carries the *b2* haplotype. CE/J has been reported to be an outlier among classical inbred strain because it has the smallest fraction of haplotype sharing genome wide with strains with WGS available [63].

In contrast, wild-derived and wild-caught *M. m. domesticus* mice exclusively carry the *b1*, *b2*, *b3* and *b4* haplotypes (**Figure 4-8**). Note that we have determined experimentally the *Xce* allele for a wild derived representative of these two haplotypes. WSB/EiJ, PERA/EiJ, TIRANO/EiJ and ZALENDE/EiJ carry the *b1* haplotype and LEWES/EiJ carries the *b2* haplotype. All five wild-derived strains (WSB/EiJ, PERA/EiJ, TIRANO/EiJ, ZALENDE/EiJ and LEWES/EiJ) carry the Xce^{b} allele.

We conclude that in natural populations *M. m. domesticus* mice predominantly (or exclusively) carry the Xce^{b} allele. We further conclude that given its absence among 121 wild mice and wild-derived strains the *a* haplotype associated with the Xce^{a} allele is likely a derived allele that arose during the domestication of fancy mice. Another possibility is that Xce^{a} represents a rare allele in the wild (See **Discussion**, **Figure 4-8** and **Figure S4-4**).

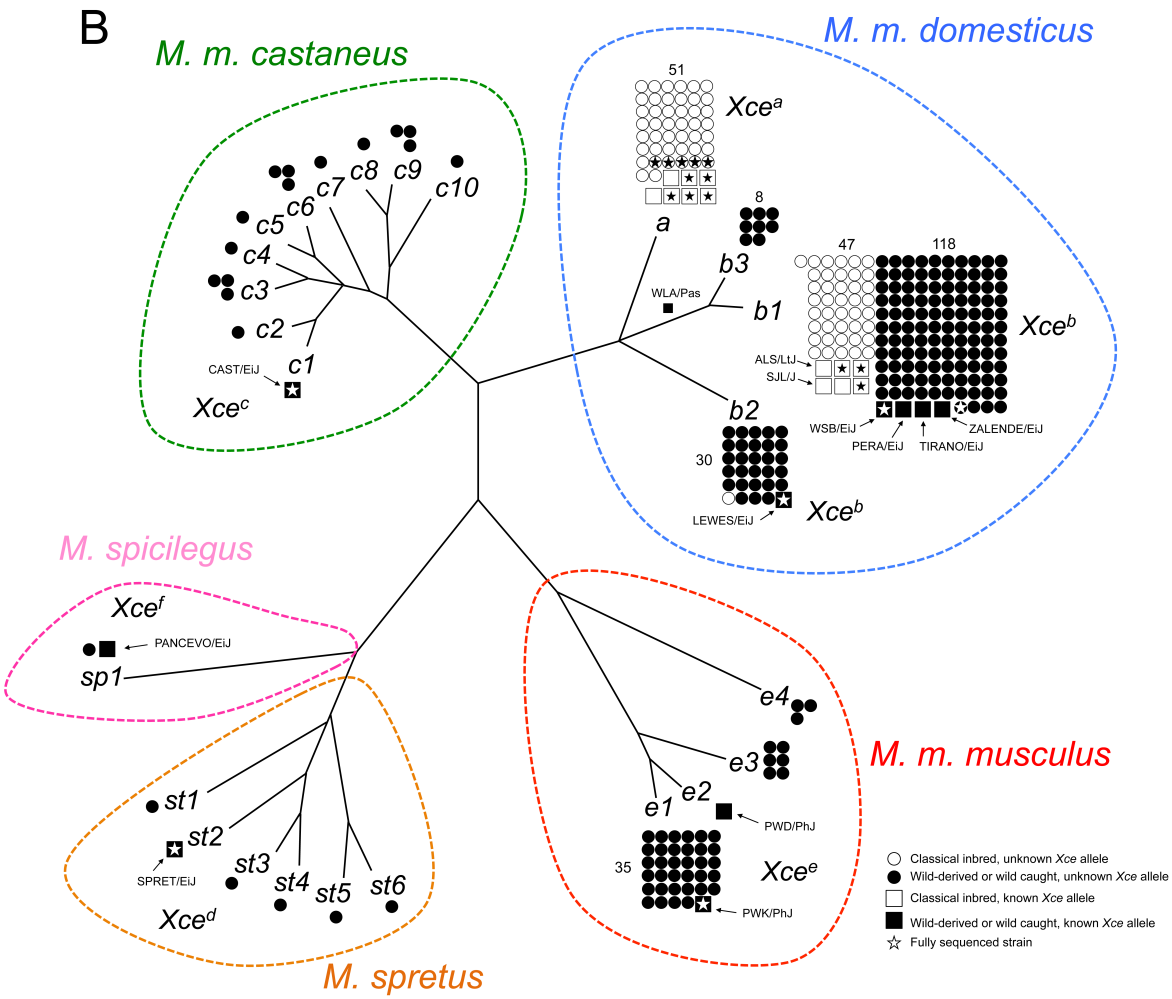


Figure 4-8. Natural history of *Xce*. This Figure shows a phylogenetic tree based on 18 MDA SNP probes within the new *Xce* candidate interval. The topography of the tree accurately reflects the genetic relationship between the *Xce* alleles, however because of the limited number of SNP used to generate the tree and the ascertainment bias of the SNPs present on the MDA [1, 2], the tree is misleading with respect to the true genetic distance between *Xce* haplotypes (see **Figure S4-4** for a more accurate representation of branch lengths). Open circles represent classical inbred strains with unknown *Xce* alleles; filled circles represent wild-derived or wild-caught mice with unknown *Xce* alleles; open squares represent classical inbred strains phenotyped for *Xce*; filled squares represent wild-derived strains with known *Xce* alleles. Strains with whole genome sequence data are shown with a star. We color coded the specific or subspecific origin of the candidate interval for the four major branches of the tree: red, *M. m. musculus*; blue, *M. m. domesticus*; green, *M. m. castaneus*, orange, *M. spretus*, pink, *Mus spicilegus* [3].

Maternal Inheritance of the Strong *Xce* Allele Magnifies XCI Skewing

Previous studies have shown that the parent-of-origin of the *Xce* allele can influence the skewing of XCI [148, 149, 163, 164]. To investigate this effect in our data set, we examined the XCI skewing in reciprocal F1 female hybrids (**Table S4-3**) and tested whether the effect of the parent-of-origin on X inactivation ratio was statistically significant. In order to increase the statistical power to detect parent-of-origin effects we aggregated crosses with the same combination of *Xce* alleles, doing so under the assumption that the parent-of-origin effects are substantially greater than putative effects of genetic background [163].

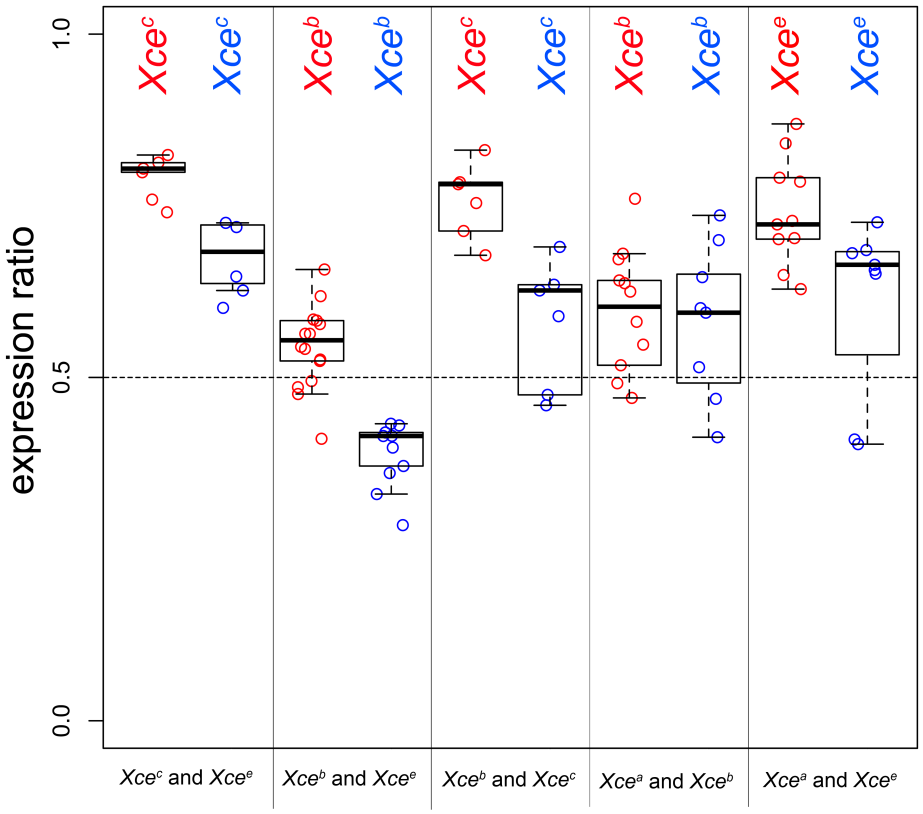


Figure 4-9. Maternal inheritance magnifies XCI skewing. Shown is allele-specific expression from reciprocal F1 *Xce* heterozygotes. The X-axis is partitioned according to *Xce* allele pairs. The Y-axis is the ratio of allele-specific expression from the X chromosome harboring the stronger *Xce* allele. Ratios were determined using either RNAseq or pyrosequencing.

We found that the parent-of-origin effect was highly significant overall (p = 0.0023) and was consistent in its direction, magnifying XCI skewing in the F1 female hybrids

inheriting the stronger *Xce* allele from their mothers (**Figure 4-9**). The magnitude of its effect varied between 18% (the X-inactivation proportion in (CAST/EiJxWSB/EiJ)F1 females minus that in (WSB/EiJxCAST/EiJ)F1 females) and 2% (WSB/EiJxA/J)F1 females minus (A/JxWSB/EiJ)F1 females), averaging 9% among all crosses where reciprocals were tested. We note that the parent-of-origin effect is observed independent of whether XCI measurement is based on pyrosequencing or RNAseq data. We found less support for the parent-of-origin effect on X inactivation skewing in reciprocal F1 females generated by crosses between the WSB/EiJ strain (Xce^b) and Xce^a allele carriers (**Table S4-3**). Retrospective analysis of reported parent-of-origin effects is fully consistent with our hypothesis that maternal origin of a strong *Xce* allele magnifies the skewing (data not shown).

DISCUSSION

Recent advances in mouse genetic resources [1, 2] provide an opportunity to resolve unanswered biological questions. Our method for association mapping integrates historical phenotyping data with these new genetic resources enabling us to reduce rapidly existing candidate intervals to a size amenable to mechanistic studies. Our method is comparable to approaches used to identify candidate genes within candidate intervals reported previously [177-179]. The method guides subsequent experiments by identifying additional mouse strains that could reduce the candidate interval through informative historical recombinations. Moreover, our comparative analysis of different subspecies of mouse provides unique insight into the evolutionary history of the locus that is key to explaining its allelic series [1].

The validity of our approach relies on the fulfillment of several assumptions. These include the requirement that the locus under study explains a large fraction of the genetic variance and its action to be largely independent of other loci; that the causative mutation(s) for each functional allele has arisen once during evolutionary history; and that the genetic markers used in the analysis reflect the true haplotype diversity in the entire candidate interval.

In our mapping of the *Xce* locus, fulfillment of the first assumption of a large genetic effect relies on 40 years of evidence that support the existence of a single major locus on the X chromosome near *Xic* that influence XCI choice [145, 149-152, 155, 170, 180, 181]. Note that these studies arrive at the same conclusion regardless of the combination of *Xce* alleles (Xce^a, Xce^b and Xce^c) used in each particular study. Although parent-of-origin and autosomal effects have been reported, the consensus is that their contribution to XCI skewing variation is small compared with that of *Xce* [149, 155, 163]. The need to fulfill the second assumption, that each allele arose once, guided the decision to restrict our initial association mapping analysis to classical inbred strains only, since the probability of multiple

recurring mutations are extremely low based on their history [1, 2, 63]. Lastly, fulfilling the third assumption, we have previously shown that the marker density in MDA is sufficient to accurately reflect the underlying haplotype diversity genome wide and in particular in regions with lower levels of recombination such as the X chromosome [1, 2, 63].

We have shown that this approach was effective at rapidly reducing the *Xce* candidate interval 10-fold and that it may prove useful to map other genetic traits of interest provided that they meet the above listed criteria. In fact, *Xce* is a particularly difficult test case because of complexity of the XCI process and the reduced recombination rate on the X chromosome.

We tailored our experimental design to anticipate the challenges of phenotyping mouse strains with unknown *Xce* alleles. First, the functional allele in a strain with an unknown *Xce* allele can be determined only by generating heterozygous females with known *Xce* alleles and then determining the ratio of XCI in the heterozygous progeny. The precision in identifying the unknown allele increases with the number of different alleles to which it is paired in the experimental F1 hybrids. We, therefore, crossed each strain with an unknown *Xce* allele to at least two strains with known and different *Xce* alleles.

To estimate mean XCI skewing accurately, we phenotyped multiple females per cross. Moreover, for most females, we measured XCI skewing in at least three different tissues that roughly represent the three germ layers, brain (ectoderm), liver (endoderm) and kidney (mesoderm). Our results confirm previous reports that mean XCI skewing is similar between different tissues [154, 162, 182, 183]. We do, however, observe differences in the variance of XCI skewing between different tissues (brain ±6% kidney ±7.5%, and liver ±8.2%). From a practical standpoint, whole brain had the smallest variance and thus would require fewer animals to accurately determine mean XCI skewing.

It is appropriate to use gene expression to measure the proportion of cells using the maternal *versus* paternal X chromosomes. However, expression at single genes can be

misleading because of measurement bias or allelic imbalance independent of XCI choice such as *cis*-acting regulatory variants or XCI escape. To mitigate these potential issues, we measured multiple X-linked genes using pyrosequencing and/or RNAseq. By combining multiple gene measurements, we can better estimate the mean XCI skewing. Both technologies simultaneously measure maternal and paternal expression, reducing the concern of parent-specific measurement bias.

Despite our thoroughness, we could not conclusively assign an *Xce* allele to the WLA/Pas strain, although we can exclude both Xce^{c} and Xce^{d}. A possible reason for this is that in all crosses involving WLA/Pas the $Xce^{WLA/Pas}$ allele was inherited through the paternal germline and in the absence of reciprocal crosses the parent-of-origin can potentially complicate *Xce* allele calling. A second, and more interesting explanation is that WLA/Pas has a *b4* haplotype that appears to be *a/b1* recombinant whose breakpoints fall within the SD4 in the candidate interval (see below and **Figure 4-6A**).

Although only a small number of readily available mouse strains carry *M. m. castaneus* or *M. m. musculus* haplotypes, a previous study measured XCI skewing in reciprocal F1 hybrids between PWD/PhJ and AKR/J [184]. This study reported that PWD/PhJ has an *Xce* allele that is weaker than Xce^{b}. This result matches our conclusion that PWK/PhJ, a closely related wild-derived inbred strain [1], carries the Xce^{e} allele. Furthermore, we conclude that *M. m. musculus* do not carry the Xce^{c} allele as reported in a congenic mouse line believed to be of *M. m. musculus* origin within the *Xce* candidate interval [183].

Our conclusion that the structural variants in the duplications within the candidate interval are likely to be responsible for the different *Xce* alleles provides simple and satisfactory answers to questions such as the presence of the allelic series, the overdominant nature and mechanism of action of *Xce*, and the evolutionary origin of the interspecific differences for XCI choice. Copy number variation within a region with complex

segmental duplications and inversions can explain the large number (six alleles described so far in *Mus*) and different strength of the alleles at *Xce*. For example, the different strength of *Xce* alleles can be attributed to the number of copies of a binding site for a *trans*-factor that is critical for the initiation of XCI [157-161].

One of the conclusions of our study is that each one of the five taxa (species or major subspecies) analyzed for XCI choice in *Mus* has a different functional allele and that there is no evidence of shared alleles between them. The rate of mutation for CNV at segmental duplicated regions fits well with the observed functional diversity at *Xce*. Given that unequal recombination is thought to be the primary process generating CNVs, it is noteworthy that two of the haplotypes reported here (*b3* and *b4*) involve crossing over within the duplications. In fact, we observe an apparently correct heterozygous call at SNP rs29082017 in two males with the *b3* haplotype. Given that males cannot be true heterozygotes for X linked markers, the result strongly suggests that an unequal crossing over has generated a new haplotype with paralogous variation. Resequencing the candidate interval in these strains should provide important information on the relationship between CNVs and functional *Xce* alleles.

It is striking that each species and subspecies examined thus far has a different functional allele. Furthermore, in the six wild-derived *M. m. domesticus* mouse strains phenotyped in this study, we do not find the occurrence of multiple functional alleles. We conclude that in *M. m. domesticus*, Xce^{b} is the prevalent allele and other functional alleles are either rare or absent. The broad geographic origin of the wild-derived strains analyzed here strongly support this conclusion (**Table S4-6**). The only apparent exception to this rule is the presence of two functional alleles in classical inbred strains, Xce^{a} and Xce^{b}. That said, it is likely that Xce^{b} is the ancestral allele within the *domesticus* subspecies and Xce^{a} is a new, derived allele that originated early during the domestication of fancy mice. However, the phylogenetic tree shown in **Figure 4-8** reveals deep branching

between *Xce*^{a} and *Xce*^{b} haplotypes that at first glance suggests that both are old alleles. Upon further investigation, there is evidence that the deep branching observed in **Figure 4-8** may be an artifact generated by genotyping and alignment problems in regions with segmental duplications (*i.e.*, the apparent SNP are paralogous variants rather that allelic ones). **Figure S4-4** provides evidence in favor of this later scenario as the deep branching disappears immediately proximal (**Figure S4-4A**) and distal (**Figure S4-4C**) to the duplicated regions. Furthermore, there is a dramatic increase in the density of heterozygous calls in the WGS data for inbred strains that overlaps the region of segmental duplications (**Figure S4-4D**).

The phylogenetic analysis also provides an explanation for the apparent differences in the genetics of XCI choice between mouse and humans. Mouse geneticists were able to find evidence of genetic control of XCI because they used mice derived from multiple taxa and because *Xce*^{a} and *Xce*^{b} are equally represented among classical laboratory inbred strains. In fact, were we to have studied only wild-derived or wild mice of *M. m. domesticus* origin, we would very likely have concluded that XCI choice is not under the control of a X chromosome linked locus. We speculate that this is probably the situation in humans too, but note that this conclusion would be due to a lack of functional variation at the *Xce* locus and not proof of the absence of a locus controlling XCI choice.

We conclude that *Xce* is the major determinant of primary XCI choice and maps 500 kb proximal to key components of the murine *Xic* (*Xist*, *Tsix* and *Xite*). This conclusion is compatible with a previous study that used the association between *Xce* alleles and microsatellite markers to refine the distal end of the *Xce* candidate interval [179]. However, the exclusion of *Xist* was dependent on a single classical inbred strain (JU/Ct) with a recombinant haplotype, which is now extinct. Our results are also compatible with the general conclusions reached by Thorvaldsen and coworkers (2012). Nonetheless a direct comparison of both studies is difficult. Thorvaldsen and colleagues (2012) used only two

functional alleles, *Xce*^{a} and *Xce*^{c} from highly divergent mouse strains to map roughly X-linked regions influencing XCI choice. They found that all their crosses, regardless of heterozygosity within the Chadwick interval, there is some degree of skewing in favor of the 129S1/SvlmJ and CAST/EiJ recombinant chromosome X. This led to the conclusion that multiple X-linked loci influence XCI choice. Although we provide strong evidence that the *Xce* allelic series is due to structural variation in the *Xce* candidate interval, we cannot exclude that a selected few SNPs within the Chadwick interval may also contribute to XCI choice. There are 14 SNPs distal to the *Xce* interval reported here with consistent SDPs in *M. m. domesticus* after the incorporation of the four strains with *M. m. domesticus* phenotyped. None of these SNPs individually can explain the allelic series and no simple combination of them within a single gene can be directly tied to the phenotype. On the other hand our reciprocal crosses between ALS/LtJ and C57BL/6J agree with Thorvaldsen's hypothesis that additional loci may have an effect in XCI choice as we find that the parent-of-origin effect is present despite homozygosity at the *Xce* locus (**Figure S4-3**). Both studies strongly predict the presence of an additional X-linked locus (or loci) controlling the parent-of-origin effect.

The genetic analysis of the *Xce* locus presented in this study sets the stage for the molecular characterization of *Xce*. However, the most direct experiments will require access to the cells and biological material of the critical window at which XCI choice is made either by *in vivo* or *ex vivo* using ES cell lines.

MATERIALS AND METHODS

Mouse breeding and tissue isolation

Mice from nine inbred strains (129S1/SvlmJ, A/J, ALS/LtJ, C57BL/6J, CAST/EiJ, LEWES/EiJ, PWK/EiJ, SJL/J, and WSB/EiJ,) were originally obtained from the Jackson

Laboratory (Bar Harbor, ME). Mice of the WLA/Pas strain were generously provided by Xavier Montagutelli from the Pasteur Institute (Paris, FR). Mice were bred at UNC-Chapel Hill for multiple generations and interbred to generate F1 hybrids. Litters of F1 mouse pups were sacrificed within 24 hours after birth. We harvested whole brain, whole liver, right kidney, tail and a forepaw (for sexing, [185]). Tissues were infused with RNAlater (Qiagen) and frozen at −80°C to preserve RNA integrity until extraction. Whole brain was isolated from mouse pups derived from crosses (DDKxC57BL/6J)F1 X PANCEVO/EiJ, (C57BL/6J X DDK)F1 X TIRANO/Ei and (C57BL/6J X DDK)F1 X ZALENDE/Ei [115] and (C57BL/6J X PERA)F1 X C57BL/6J [116]. These mouse crosses were generated for previous studies and reported elsewhere. All mice were treated according to the recommendations of the Institutional Animal Care and Use Committee (IACUC) of the University of North Carolina at Chapel Hill.

Genotypes

Mouse genotypes were acquired from recent studies that employed next-generation sequencing [2, 3] and high-density genotyping array technology [1, 62]. **Tables S4-1, S4-4,** and **S4-6** provide a list of all mice (inbred and wild-caught) and the origin of the genotype information. As an initial filtering step, heterozygous and low-confidence genotyping calls were removed from the data set. Heterozygosity within the *Xce* candidate interval was determined in F2 mouse pups using microsatellite marker *DXMit16* (~99.3 Mb) [186]. Genomic DNA was amplified according to previously reported conditions with the exception of a fluorescent label covalently bound to one *DXMit16* primer (6-FAM-5′-CTgCAATgCCTgCTgTTTTA-3′). 0.5 µl of amplified products were resuspended in 9.0 µl of HIDI formamide (Life Technologies) and 0.5 µl of LIZ1200 sizing ladder (Life Technologies). Samples were run on the ABI 3730xl DNA analyzer using long-run fragment analysis conditions. Traces were analyzed with ABI PeakScanner software.

Association mapping

At each diallelic variant within the Chadwick interval, we represented the C57BL/6J (or C57BL/6JN) allele as zero and all other strains with the same genotype as zero. Strains with the alternative allele are represented with the number one. We then generated strain distribution patterns for each variant as a series of ones and zeros for the strains in the following order: 129S1/SvlmJ, A/J, BALB/cByJ, C3H/HeJ, CBA/J, DDK/Pas, C57L/J, DBA/1J, DBA/2J, and AKR/J (**Table S4-1**). We classified an SDP as completely consistent when all *Xce*^{a} allele carriers are ones (share the same allele) and all *Xce*^{b} allele carriers are zeros (share the same allele as C57BL/6J) (**Tables S4-1** and **S4-4**). We defined an inconsistent SDP when one or more *Xce*^{a} strain(s) are zeros and one or more *Xce*^{b} strain(s) are ones (*i.e.,* A/J, 129S1/SvlmJ, BALB/cByJ, C3H/HeJ, CBA/J, AKR/J opposite to DDK, C57BL/6J, DBA/1J, DBA/2J) (**Tables S4-1** and **S4-4**). Lastly, we defined a diallelic variant as partially consistent when one or more *Xce*^{a} strain(s) are zeros or one or more *Xce*^{b} strain(s) are ones (**Tables S4-1** and **S4-4**).

Measuring allelic imbalance in F1 female hybrids

mRNA was extracted from tissues of F1 mice using an automated bead-based capture technology (Maxwell 16 LEV Total RNA Kits, Promega). Purified mRNA was checked for quality and quantity using a Nanodrop spectrophotometer (Thermo Scientific). For each sample, mRNA was retrotranscribed (SuperScript III, Life Technologies) to produce cDNA. We designed primers (**Table S4-7**) to capture expression SNPs (**Table S4-8**) within X-linked genes to serve as surrogates for maternal and paternal XCI status. In individual reactions, we amplified 1 µl of cDNA in a final volume of 30 µl for 35 cycles (See **Table S4-7** for PCR cycling conditions). One primer for each assay was biotinylated in order to immobilize and purify the amplified products using streptavidin beads (GE Healthcare) according to the manufacturer's protocol (Qiagen). We used Pyrosequencing technology to measure the proportion of maternal and paternal X-linked gene expression

simultaneously. Pyrosequencing quantitatively measures, in real-time, the release of pyrophosphate as a result of nucleotide incorporation during the polymerase chain reaction [187]. Purified, single-stranded amplicons were primed for pyrosequencing using gene-specific primers (**Table S4-7**) and pyrosequenced using the PyroMark Q96 MD instrument (Qiagen) and PyroMark Gold Q96 Reagents (Qiagen) according to manufacturer's protocols. Allelic proportions were determined by the quantitative analysis option of the PyroMark Q96 MD Software. Raw results are show in **Table S4-9**.

RNAseq analysis

RNAseq data used in this study is reported elsewhere (Crowley *et al.* 2013, unpublished). Briefly, we generated cDNA libraries (Illumina (San Diego, CA) TruSeq RNA Sample Preparation Kit v2) from whole brain mRNA of female reciprocal F1 hybrids between CAST/EiJ, PWK/PhJ, and WSB/EiJ. Using the Illumina HiSeq 2000 instrument, we sequenced 100 bp paired end reads (2×100). For each F1 hybrid, we mapped 100 bp paired-end RNAseq reads to pseudogenomes of each parent (CAST/EiJ, PWK/PhJ and WSB/EiJ) using TopHat. Pseudogenomes are approximations of CAST/EiJ, PWK/PhJ and WSB/EiJ strain genomes constructed by incorporating all known SNPs and indels into the C57BL/6 genome (mm9) [188]. We allowed two mismatches total per 100 bp read. For each read, we annotated the number of maternal and paternal alleles (using SNPs and indels). XCI ratios were determined by counting the number of maternal reads versus the number of paternal reads. To measure XCI ratios, we selected 10 X-linked genes that are distributed across the X chromosome (*Wdr13*, *Atp6ap2*,*Usp9x*, *Cask*, *Cd99l2*, *Idh3g*, *Dlg3*, *Zcchc18*, *Tsc22d3*, *Iqsec2*). For each gene, we selected two informative SNPs between PWK, CAST, and WSB so that at least five of the ten genes were informative for a given F1 hybrid. For each informative SNP, we counted allele-specific reads to determine XCI ratios. Results are summarized in **Table S4-9**.

Statistical model for cross-specific X-inactivation ratios

Pyrosequencing and RNAseq provided estimates of the X-inactivation ratios obtaining for particular genes in specific tissues in particular individuals. In order to infer X-inactivation ratios pertaining to individual mice and to the crosses that generated them, we developed a hierarchical Bayesian model linking the observed experimental measurements to a structured set of higher order parameters. These parameters reflected not only the stochastic relationships between measurements, individuals and crosses, but also between different sources of experimental variation. Let Yij be the measured X-inactivation proportion from pyrosequencing or RNAseq in the jth gene-tissue combination of the ith mouse, and let g be the F1 cross to which mouse i belongs, where for instance, crosses (129S1/SvImJxPWK/PhJ)F1 and (PWK/PhJx129S1/SvImJ)F1 are distinct. We first model a latent variable Pi representing the X-inactivation proportion inherent to the individual mouse i as if arising from a beta distribution

$$Pi \sim \text{Beta}\alpha g\mu g+1, \alpha g1-\mu g+1,$$

with cross-specific mean governed by μg and cross-specific variance proportional to $\alpha g-1$. This individual-specific parameter Pi then forms the basis of a further beta distribution, which models tissue-gene specific measurements Yij as if generated by

$$Yij \sim \text{Beta}(Sje\eta jgRjPi+1, Sje\eta jg1-Rj1-Pi+1),$$

where Rj and $Sj-1$ are the bias and variance introduced by tissue-gene combination j, and where ηjg allows for cross-specific variance in X-inactivation. All higher order parameters are themselves modeled in loosely-specified grouped hierarchies based realistic but vague priors (as in, *eg,* [189]). This hierarchical structure allows information and uncertainty to propagate within and between parameters, and results in improved estimation through shrinkage (see, *eg*, [190]). We obtain posterior distributions for all parameters, including those representing unobserved data, using Markov Chain Monte Carlo (MCMC). Marginal

posterior probability densities are computed for μg parameters for crosses between mice with unknown *Xce* alleles using information from mice with known alleles. The μg posterior density that includes the most support for $\mu g=0.5$ is taken as the most plausible candidate for having *Xce* allele shared by the unknown strain. In general, posteriors for μg concentrated near 0.5 are more consistent with there being a shared allele between maternal and paternal pairs, whereas posterior densities shifted from 0.5 suggest that the *Xce* is different.

Significance test of parent-of-origin effects

The statistical significance of parent-of-origin effects was determined by permutation. We first estimated the difference in specimen-level X-inactivation, Pi, between genetically matched individuals of reciprocal parentage and unequal *Xce* alleles, and used this estimate as our test statistic. We then repeated this estimation under 10000 shuffles of the parent-of-origin labels in order to generate a null distribution of the test statistic, and thereby estimate a *p*-value for the parent-of-origin effect in the real data.

Principal Component Analysis (PCA)

For each sample, we constructed a vector of Illumina probe intensities of MegaMUGA markers within the refined *Xce* candidate interval (**Table S4-10**). We then performed principal component analysis on these vectors and report the projection of each sample onto the first three principal components.

Phylogenetic analysis.

For each inbred strain and wild-caught mouse, we assigned the subspecific origin of the Chadwick and new *Xce* candidate interval based on diagnostic alleles from SNP and VINO calls [1, 67]. We then built DNA distance, maximum likelihood, and DNA parsimony phylogenetic trees (PHYLIP (Phylogeny Inference Package) [191]) based on all variation within the candidate interval. No major differences were observed between analysis types,

so we chose maximum likelihood with 100 bootstraps to represent the phylogenetic relationship between mice in **Figure 4-8**.

SUPPORTING MATERIAL

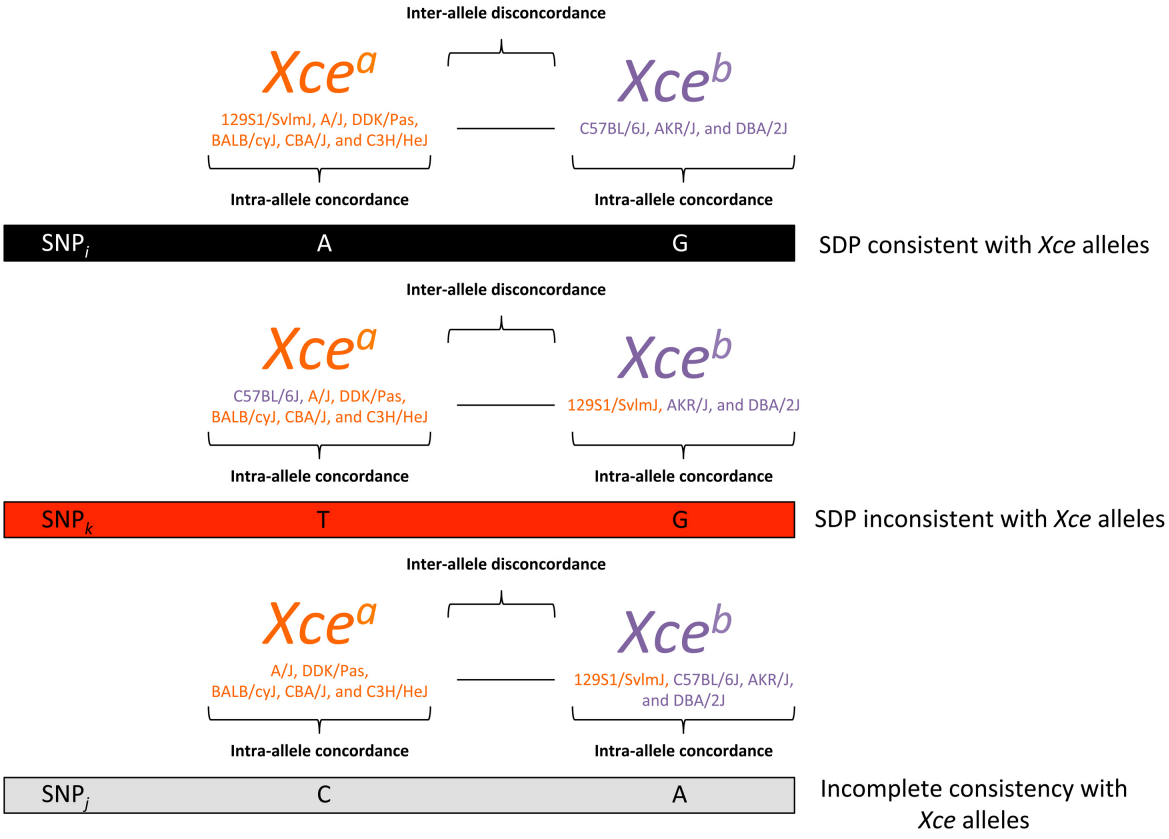


Figure S4-1: Strain Distribution Patterns (SDP). This Figure depicts how the patterns of strain genotypes were classified as consistent, inconsistent or incompletely consistent with the *Xce* phenotypes. SNPs or indels that partition the strains according to their *Xce*^{a} and *Xce*^{b} phenotype were classified as "consistent" and represented as a black (or blue) tick mark in **Figure 4-4**. SNPs or indels that are shared by both *Xce*^{a} and *Xce*^{b} strains were classified as an SDP that is "inconsistent" with the *Xce* phenotypes and represented as a red tick mark in **Figure 4-4**. Lastly, A SNP or indel that is partially consistent but not inconsistent with the *Xce* phenotypes was classified as "partially consistent" and represented with a gray tick mark in **Figure 4-4**.

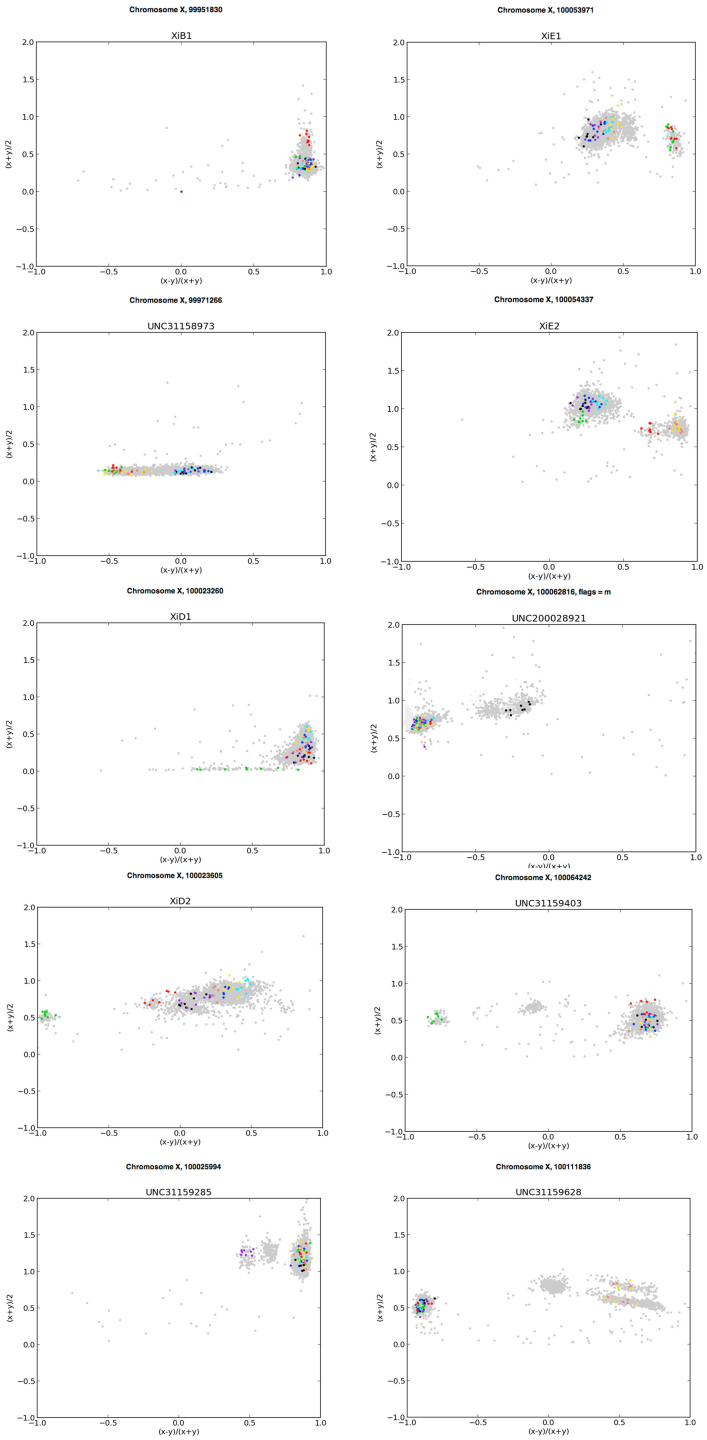


Figure S4-2: MegaMUGA probe plots. Each of the ten panels is a hybridization plot of an individual MegaMUGA probe targeting the *Xce* candidate interval. As described in **Figure 4-6C**, the axes represent hybridization intensities for probes tracking alternative alleles at each marker. The colors correspond to eight biological replicates of the eight founder inbred strains of the CC. Yellow A/J; black C57BL/6J; pink 129S1/SvImJ; blue NOD/ShiLtJ; light blue NZO/HiLtJ; green CAST/EiJ; red PWK/PhJ, and purple WSB/EiJ. Samples in gray represent 300 control DNAs.

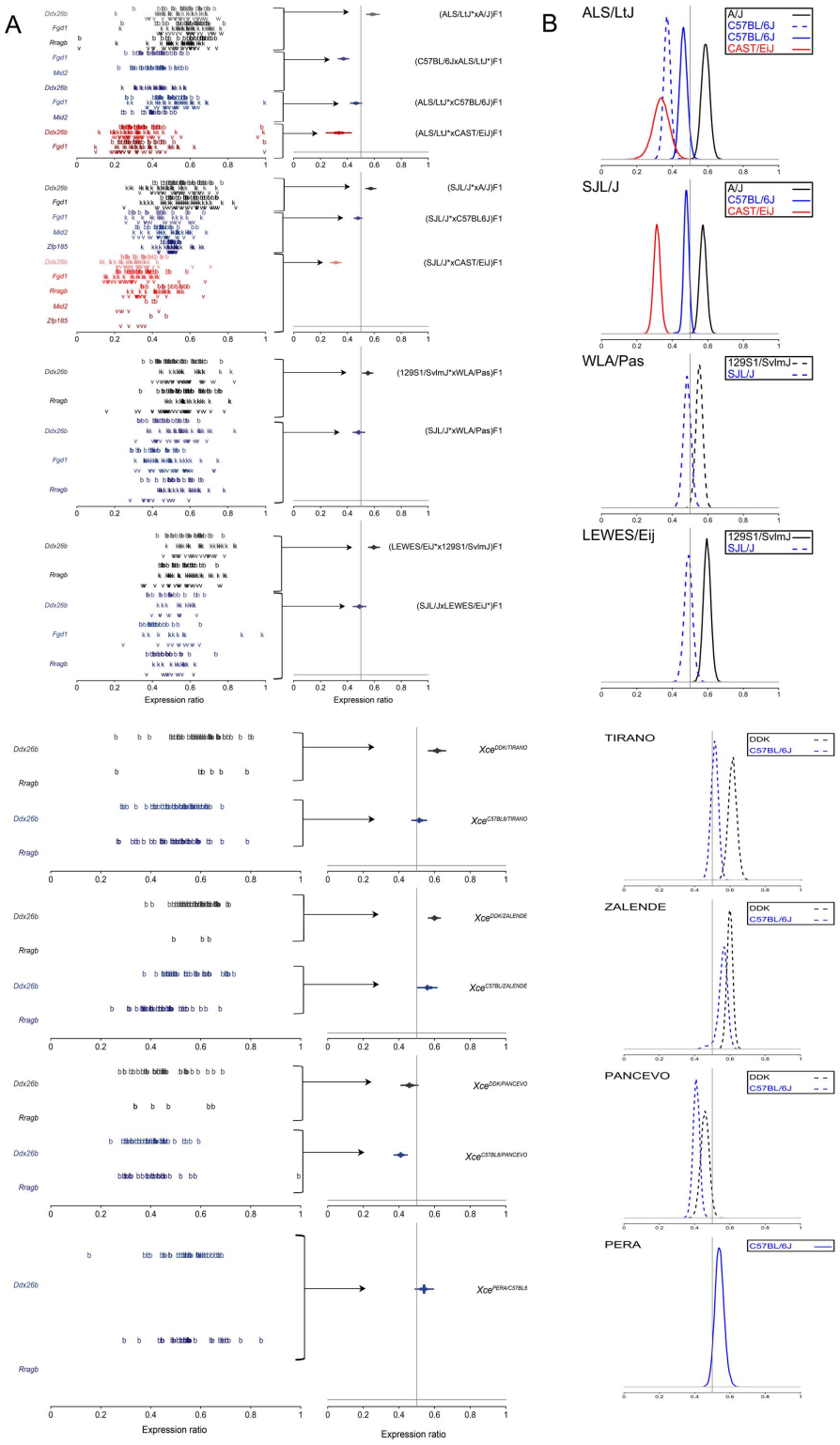


Figure S4-3: Allelic imbalance in additional strains characterized. Shown in **Panel A** are scatter plots and posterior mean and 95% credibility intervals for additional strains phenotyped in this study. Shown in **Panel B** are the posterior distributions of the phenotyping data in **Panel A**.

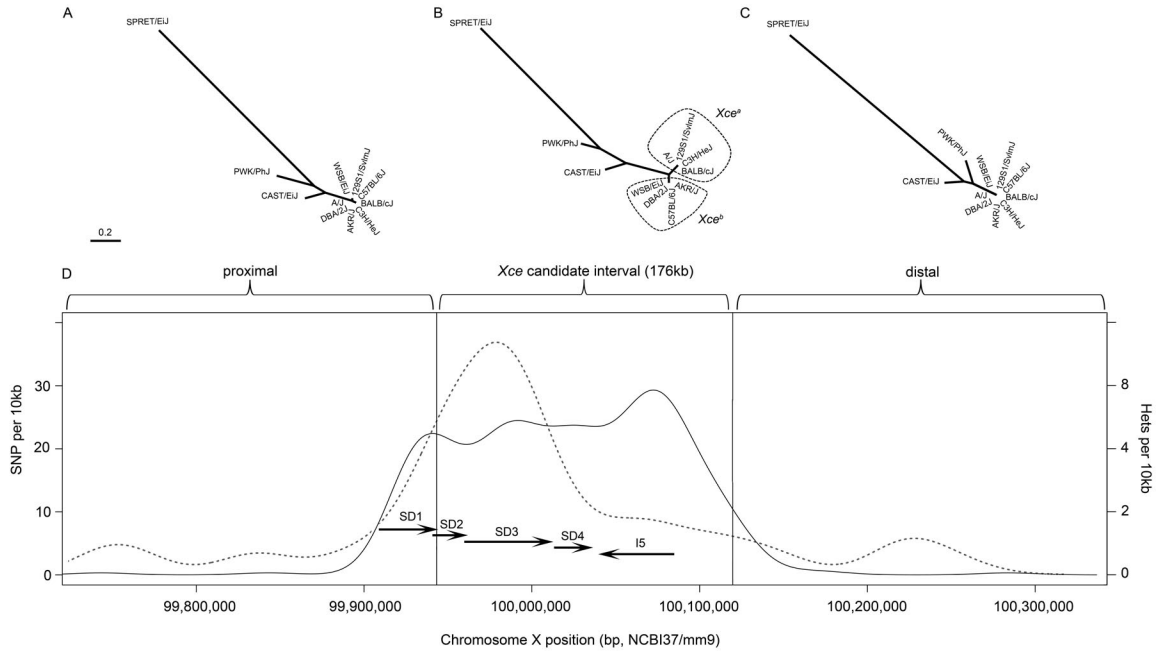


Figure S4-4: Phylogenetic analysis of the *Xce* and flanking intervals using whole genome sequence data. Shown are DNA distance trees based on whole genome sequence data [2, 3] within the corresponding intervals. **Panel D** shows the SNP density (solid line) and heterozygosity (dashed lined) within the candidate (**Panel B**) and flanking intervals (**Panels A and C**).

http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1003853#s5

Table S4-1: Genotype data in the Chadwick interval for strains with previously known *Xce* allele. This table summarizes consistent, inconsistent and partially consistent SDPs for inbred mouse strains with previously known *Xce* alleles. The data includes MDA and Sanger sequencing data.

Strain	F1s phenotyped	# data points	Justification for phenotyping
ALS/LtJ	120	606	*Xce a/b* recombinant haplotype, classical inbred, availability
SJL/J	81	581	*Xce a/b* recombinant haplotype, classical inbred, availability
WSB/EiJ	52	179	Wild derived, *M. m. domesticus*, *b1* haplotype, availability, WGS (Keane et al. 2012), mouse genetic resources (Collaborative Cross and Diversity Outbred)
TIRANO/EiJ*	81	123	Wild derived, *M. m. domesticus*, availability
ZALENDE/EiJ*	82	119	Wild derived, *M. m. domesticus*, availability
PERA/EiJ*	42	63	Wild derived, *M. m. domesticus*, availability
WLA/Pas	67	404	Wild derived, *M. m. domesticus*, availability, *b4* haplotype
LEWES/EiJ	45	270	Wild derived, *M. m. domesticus*, availability, *b2* haplotype
PWK/PhJ	54	285	Wild derived, *M. m. musculus*, availability, WGS (Keane et al. 2012), mouse genetic resources (Collaborative Cross and Diversity Outbred)
PANCEVO/EiJ*	65	136	Wild derived, *M. pancevo*, availablility
Totals	689	2766	

Table S4-2: Justification of selected inbred strains. This table lists the justification for selecting each strain and summarizes the number of F1 females phenotyped for each inbred strain with an unknown *Xce* allele.

dam	sire	# of females	Xce dam	Xce sire	method	reciprocals?	posterior-Mean	posterior-Median	left CI	rightCI	comments
PWK/PhJ	129S1/SvImJ	5	*e*	*a*	pyrosequencing	yes	NA	NA	*	*	
129S1/SvImJ	PWK/PhJ	4	*a*	*e*	pyrosequencing		NA	NA	*	*	
129S1/SvImJ	WSB/EiJ	4	*a*	*b*	pyrosequencing	yes	NA	NA	*	*	
WSB/EiJ	129S1/SvImJ	5	*b*	*a*	pyrosequencing		NA	NA	*	*	
A/J	PWK/PhJ	2	*a*	*e*	pyrosequencing	yes	0.327	0.325	0.235	0.425	*Pooled with PWK-129
PWK/PhJ	A/J	7	*e*	*a*	pyrosequencing		0.725	0.730	0.601	0.839	*Pooled with 129-PWK
A/J	WSB/EiJ	7	*a*	*b*	pyrosequencing	yes	0.403	0.403	0.324	0.479	*Pooled with 129-WSB
WSB/EiJ	A/J	4	*b*	*a*	pyrosequencing		0.630	0.634	0.457	0.813	*Pooled with WSB-129
WSB/EiJ	C57BL6	6	*b*	*b*	pyrosequencing	yes	0.474	0.473	0.372	0.574	
C57BL/6J	WSB/EiJ	2	*b*	*b*	pyrosequencing		0.498	0.495	0.246	0.738	
C57BL/6J	PWK/PhJ	15	*b*	*e*	pyrosequencing	yes	0.543	0.544	0.497	0.593	
PWK/PhJ	C57BL/6J	10	*e*	*b*	pyrosequencing		0.609	0.609	0.543	0.674	
ALS/LtJ	A/J	30	*b*	*a*	pyrosequencing	no	0.587	0.587	0.541	0.634	
ALS/LtJ	C57BL/6J	30	*b*	*b*	pyrosequencing	yes	0.463	0.463	0.423	0.506	
C57BL/6J	ALS/LtJ	29	*b*	*b*	pyrosequencing		0.628	0.629	0.589	0.669	
ALS/LtJ	CAST/EiJ	31	*b*	*c*	pyrosequencing	no	0.336	0.337	0.242	0.429	
SJL/J	WLA/Pas	27	*b*	?	pyrosequencing	no	0.516	0.516	0.472	0.559	
SJL/J	LEWES/EiJ	19	*b*	*b*	pyrosequencing	no	0.509	0.509	0.462	0.556	
SJL/J	CAST/EiJ	24	*b*	*c*	pyrosequencing	no	0.314	0.314	0.275	0.352	
SJL/J	A/J	34	*b*	*a*	pyrosequencing	no	0.573	0.573	0.536	0.611	
LEWES/EiJ	129S1/SvImJ	26	*a*	?	pyrosequencing	no	0.596	0.596	0.556	0.637	
129S1/SvImJ	WLA/Pas	40	*a*	?	pyrosequencing	no	0.448	0.448	0.41	0.488	
SJL/J	C57BL/6J	23	*b*	*b*	pyrosequencing	no	0.479	0.479	0.449	0.507	
PWK/PhJ	WSB/EiJ	6	*e*	*b*	RNAseq	yes	0.514	0.514	0.371	0.665	
WSB/EiJ	PWK/PhJ	6	*b*	*e*	RNAseq		0.588	0.592	0.402	0.757	
PWK/PhJ	CAST/EiJ	5	*e*	*c*	RNAseq	yes	0.340	0.336	0.229	0.457	
CAST/EiJ	PWK/PhJ	6	*c*	*e*	RNAseq		0.785	0.787	0.714	0.853	
WSB/EiJ	CAST/EiJ	6	*b*	*c*	RNAseq	yes	0.421	0.419	0.296	0.549	
CAST/EiJ	WSB/EiJ	6	*c*	*b*	RNAseq		0.753	0.755	0.668	0.835	
C57BL/6J	TIRANO/EiJ	43	*b*	*b*	pyrosequencing	no	0.486	0.486	0.445	0.527	
DDK/Pas	TIRANO/EiJ	38	*a*	*b*	pyrosequencing	no	0.385	0.385	0.338	0.434	
PERA/EiJ	C57BL/6J	42	*b*	*b*	pyrosequencing	no	0.541	0.540	0.492	0.594	
C57BL/6J	ZALENDE/EiJ	42	b	*b*	pyrosequencing	no	0.441	0.437	0.386	0.498	
DDK/Pas	ZALENDE/EiJ	40	*a*	*b*	pyrosequencing	no	0.401	0.400	0.368	0.435	
C57BL/6J	PANCEVO/EiJ	35	*b*	*f*	pyrosequencing	no	0.590	0.590	0.554	0.628	
DDK/Pas	PANCEVO/EiJ	30	*a*	*f*	pyrosequencing	no	0.541	0.541	0.494	0.588	
		Average	19.1	females							

Table S4-3: Summary of crosses. This table summarizes all strains and crosses phenotyped in this study, their corresponding *Xce* alleles, and the molecular method used to measure allele-specific expression. In addition, listed are the posterior mean, median and confidence intervals determined by the Bayesian hierarchical model.

http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1003853#s5

Table S4-4: Genotype data in the Chadwick interval for strains with known *Xce* allele. This table summarizes consistent, inconsistent and partially consistent SDPs for inbred mouse strains with previously known *Xce* alleles combined with mouse strains phenotyped in this study.

Probe name	Position (mm9 Build 37)	PCV1	Rank1	PCV2	Rank2	PCV3	Rank3
UNC200028921.x	100062816	0.075305	12	-0.13603	12	-0.04516	15
UNC200028921.y	100062816	0.018192	15	0.081997	14	-0.18142	7
UNC31158973.x	99971266	0.015932	16	-0.026549	16	-0.018492	18
UNC31158973.y	99971266	-0.003868	18	0.025469	17	0.01759	19
UNC31159285.x	100025994	0.064732	14	0.224796	8	-0.248758	3
UNC31159285.y	100025994	-0.002396	19	-0.01903	18	0.1027	11
UNC31159403.x	100064242	-0.101471	9	-0.148717	11	0.205282	6
UNC31159403.y	100064242	0.237501	6	0.364256	3	-0.098832	12
UNC31159628.x	100111836	-0.585142	1	0.286473	4	-0.15838	8
UNC31159628.y	100111836	0.430314	2	-0.276628	5	-0.045955	14
XiB1.x	99951830	-0.069933	13	-0.114092	13	0.78509	1
XiB1.y	99951830	0.000564	20	0.005926	20	0.033563	17
XiD1.x	100023260	-0.244671	5	-0.207727	9	-0.12135	9
XiD1.y	100023260	-0.00722	17	-0.006115	19	-0.01492	20
XiD2.x	100023605	-0.420769	3	-0.446799	1	-0.107118	10
XiD2.y	100023605	0.146286	8	0.241725	7	0.034679	16
XiE1.x	100053971	-0.077803	10	0.202988	10	0.085225	13
XiE1.y	100053971	-0.173648	7	-0.261203	6	-0.212964	5
XiE2.x	100054337	-0.07654	11	0.048647	15	-0.220128	4
XiE2.y	100054337	0.299917	4	-0.427828	2	-0.258211	2

Table S4-5: MegaMUGA probe information. Summarized in this table are the ten MegaMUGA probes used in the principal component analysis. Shown are the probe names, sequences and ranking according to how much each probe contributes to each principal component.

http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1003853#s5

Table S4-6: List of all mouse samples. This table lists each mouse samples used in this study (total of 327). We annotated haplotypes based on its association with mouse strains with known *Xce* alleles; we assigned the subspecific origin of the *Xce* candidate interval, whether the mouse is a classical inbred [1], wild-derived inbred, or wild-caught; and we assigned each classical strain to a subclass [1] and each wild-derived or wild-caught to a geographic origin. For each mouse sample, we list the haplotype based on 18 MDA genotypes, the name of the haplotype (*i.e., a, b1, b2,* etc), The letter “V” stands for variable intensity oligonucleotide (VINO) [67], the letter “H” stands for heterozygous, and the letter “N” stands for no call.

Name	Sequence	anneal temperature used	
Ddx26b-F	5'-biotin-ggCCTCCATACTACTTAATAACCAAg-3'	56°C	
Ddx26b-Rev	5'-ggTAggCCACATgCAgAATg-3'		
Ddx26b-Pyro	5'-TCTTCCCCACTgATgCTAgAATT-3'	N/A	
Fdg1-F	5'-biotin-CTCCAACCTCAACATgCCTCg-3'	62°C	
Fdg1-Rev	5'-TTCAggAgggTggAATTgATggC-3'		
Fdg1-Pyro	5'-gTCCTggCCTgCAgTTCgAg-3'	N/A	
Rragb1-F	5'-CTgTATAAggCATggTCCAgCATTg-3'	56°C	
Rragb1-Rev	5'-biotin-ATCggTgggCATCTCgCTgTTC-3'		
Rragb1-Pyro	5'- CTgATgAAgTTCTTCTgTTTgA-3'	N/A	
Mid2-F	5'-CTggACCACgAgAATgAgAAgg-3'	60°C	
Mid2-Rev	5'-biotin-gCAgTATTCACCTCAACTTgCTg-3'		
Mid2-Pyro	5'-TCgTCACCgAgACCATCAgg-3'	N/A	
Zfp185-F	5'-CCCTgAgCACTCCAgATTCTTg-3'	60°C	
Zfp185-R	5'-biotin-CAgCATgTTAgTACAgTCCTCgg-3'		
Zfp185-Pyro	5'-AgATCTCAgCATCCTAgAgCC-3'	N/A	
Xist-F	5'-biotin-TggAgTCTgTTTTgTgCTCCTgCC-3'	58°C	Thorvaldsen et al., 2012 [169]
Xist-Rev	5'-CCTTgCTgggTTCAggAAAgCgTC-3'		Thorvaldsen et al., 2012 [169]
Xist-pyro	5'-ATAggCTgCTggCAgTCCTTgA-3'	N/A	

Cycling conditions	
95x2min	
95x15sec	35X
annealx20sec	
72x30sec	
72x10min	

Table S4-7: Primers and conditions for pyrosequencing assays. Primer sequences and annealing temperatures for primer pairs are shown. For amplification prior to pyrosequencing, a universal PCR protocol was used but the annealing temperature was tailored specifically to each primer pair.

	Ddx26b	*Fgd1*	*Rragb1*	*Zfp185*	*Mid2*	*Xist*
A/J	C	A	A	T	T	G
129S1/SvlmJ	T	A	A	A	T	G
C57BL/6J	T	A	A	A	C	G
CAST/EiJ	C	A	G	A	C	A
PWK/PhJ	C	A	G	A	C	A
WSB/EiJ	C	A	G	A	C	G
LEWES/EiJ	C	A	G	N/D	N/D	N/D
WLA/Pas	C	A	G	N/D	N/D	N/D
ALS/LtJ	T	G	G	N/D	T	G
SJL/J	T	G	A	T	N/D	N/D
TIRANO/EiJ	C	N/D	G	N/D	C	G
ZALENDE/EiJ	C	N/D	G	N/D	C	G
PERA/EiJ	C	N/D	G	N/D	C	G
PANCEVO/EiJ	C	N/D	G	N/D	C	A
DDK/Pas	T	N/D	G	N/D	C	G

Table S4-8: Pyrosequencing expression assay allele information. The table shows the mouse strains phenotyped and their genotype for each pyrosequencing assay used. Strains without genotype information are labeled "N/D."

http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1003853#s5

Table S4-9: Pyrosequencing and RNAseq raw data. This matrix shows the fraction of maternal expression generated from pyrosequencing and RNAseq of mouse pups. Each row represents an individual mouse and each column represents a gene measurement. NA is used to show missing data.

http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1003853#s5

Table S4-10: PCA results. Shown are the first three principal components used to generate **Figure 4-7** for each mouse sample.

CHAPTER V: SUMMARY AND FUTURE DIRECTIONS[4]

SECTION ONE

Listed below are general conclusions drawn from section one:

- *cis*-acting sequence variation is a major contributor to differential CpG methylation in mouse, with an estimated 13,000 differentially methylated CpGs genome-wide.
- Although further mechanistic studies are needed, differential CpG methylation appears to be functionally relevant by its association with nearby differential gene expression.
- *cis*-acting sequence variation influence parent-of-origin DMRs and may alter allelic imbalance.

Sequence variation and parent-of-origin DMRs

The results show that CpG methylation at the *Actn1* DMR is dependent on both parent-of-origin and strain (**Figures 3-1C and 3-2**). Specifically, I demonstrated that the maternal PWK allele is more consistently methylated than the maternal 129S1 allele using three independent molecular assays: MSNP (**Chapter II**), MS-RFLP, and sodium bisulfite

[4] The following chapter summarizes and draws conclusions regarding the genetic regulation of epigenetic processes and discusses ongoing work. I would like to thank Dr. Jack Griffith's laboratory, specifically Brian Bower for providing assistance in the setup of the pulse field gel electrophoresis system. I would like to thank Dr. Fernando Pardo-Manuel de Villena for his guidance with Southern blots. The MSNP optimization was done in collaboration with John Didion, a fellow graduate student in the Pardo-Manuel de Villena laboratory. As with **Chapter IV**, **Figure 5-1 and 5-7** are from an unpublished study currently under review (Crowley *et al.* 2013*).*

sequencing. Sodium bisulfite sequencing reveals the true nature of the DMR methylation. A greater proportion of PWK alleles are methylated when inherited through the maternal germline than when 129S1 is maternal (**Figure 3-2**). Furthermore, the methylated PWK clones isolated are consistently methylated across all CpGs sequenced. On the other hand, methylation is sporadic across all CpGs sequenced for the maternal 129S1 allele (**Figure 3-2**).

Within a 1.5 kb region that encompasses the *Actn1* DMR, there are 21 annotated SNPs and 3 annotated indels that distinguish PWK from 129S1 [2, 3]. In fact, one variant is a CpG to TpG transition in PWK (rs32640412). Given the requirement that strain-specific DMRs in F1 hybrids are caused by a nearby *cis* acting variant(s) (genetic or epigenetic) it stands to reason that parent-of-origin DMRs are subject to the same genetic influences. It is possible to further test this hypothesis using inbred strains with different sequence variation within the *Actn1* DMR. These strains could be used to generate F1 hybrids to compare the consistency of DMR methylation to that of 129S1 and PWK. The sequence similarities and differences (correlated with differential methylation) between strains may shed light on the causative variant(s) and perhaps divulge possible molecular mechanisms. For instance, *Dnmt3a,* a *de novo* methyltransferase in mouse, is influenced by the periodicity of CpGs [192]. The sequence variation in PWK does create differences in the CpG periodicity and may explain the strain-specific methylation differences.

129S1xCAST, PWKxC57BL6, and PWKxA/J reciprocal F1 mice were tested using MS-RFLP and showed the same maternal DMR (data not show). Rough estimates using densitometry of strain-specific RFLP bands reveals that each cross shows a similar PWK maternal effect (data not shown). Furthermore, 129S1, A/J and C57BL/6J are IBD within that region [2, 3]. However, future CpG methylation profiling is required for an accurate comparison between strains, including additional MS-RFLP and bisulfite sequencing.

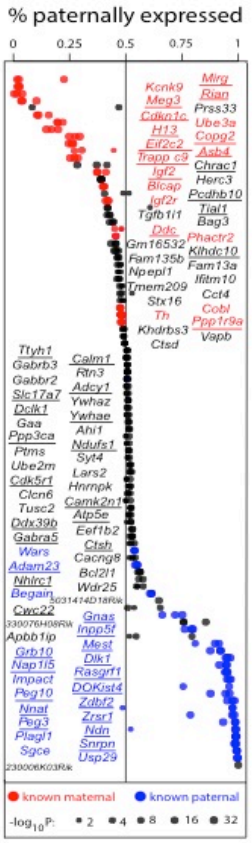


Figure 5-1. Genes imprinted in mouse brain. Plotted is the paternal expression ratio for 98 genes declared to be imprinted. Each dot corresponds to a reciprocal cross (*i.e.,* CASTxPWK versus PWKxCAST) and dot size is proportional to the parent-of-origin *p*-value significance level. Genes known from the literature to be maternally expressed are shown in red, those known to be paternally expressed in blue, and novel imprinted genes in black (N = 56 novel genes). Genes with a strain by parent-of-origin effect are indicated by underline (N = 47 genes).

After extensive investigation, the functional role of the *Actn1* DMR is unknown, and therefore offering any possible phenotypic consequences of the strain by parent-of-origin effects would be purely speculative. However, it is possible to extend this observation to other parent-of-origin DMRs, for instance imprinting control regions (ICR). Sequence variation within these ICRs may alter the highly-conserved alleleic imbalance required for proper development. Although the phenotypic consequences of severe alterations to known imprinted regions is well established [193], little is known about how subtle changes influence phenotype [57, 194, 195]. There is convincing evidence that imprinted regions are not all-or-nothing as once originally thought (**Figure 5-1**, Crowley *et al.* under review). Crowley and colleagues showed, using reciprocal F1 hybrids between three divergent mouse inbred strains PWK/PhJ, CAST/EiJ, and WSB/EiJ, that imprinting expression is

surprising incomplete (**Figure 5-1**). In fact, they report 47 genes that are dependent on both strain and parent-of-origin (**Figure 5-1**, underlined). There are annotated sequence variants within the ICRs that control these imprinted genes [2, 3]. Subtle changes in methylation may alter the function of key imprinting regulators such as CCCTC-Binding factor (CTCF; a *trans*-factor involved in imprinting regulation). It has been shown that CTCF occupancy is affected by differential methylation [196]. The genetic diversity within classical and wild-derived inbred strains could be used to target causative genetic variation for mechanistic study [1, 169]. It must be noted, however, that the RNAseq library was generated from mRNA isolated from the right hemisphere of the brain. It is possible that the incompletely imprinted genes are caused by differential expression between different cell types within the brain.

In the early 20th century, Conrad Waddington put forth the idea of canalization, or phenotypic robustness [197]. He postulated that the evolution of development incorporated a buffering system to shield large phenotypic changes caused by small environmental or genotypic changes. It is therefore possible that despite the apparent widespread differentially methylated CpGs, their effects on developmental phenotypes may be minimal. However, it is clear that understanding the genetic regulation of cytosine methylation is far from being fully realized and will require both discovery and mechanistic studies to fully appreciate the role epigenetic variation plays within a population. The genome-wide methylation survey (**Chapter II**) and the characterization of the *Actn1* DMR (**Chapter III**) significantly contribute to the emerging field of the genetic regulation of cytosine methylation in mammals.

Ongoing work: An optimized MSNP protocol to investigate additional variables affecting DNA methylation in mouse

After identifying the shortcomings of the first MSNP genome-wide survey, we used a similar but computationally optimized approach and expanded the scope of the original

experiment by introducing environmental and age-related variables. We generated reciprocal crosses between two genetically divergent mouse inbred strains, NOD/LtJ and PWK/PhJ, in order to maximize the number of SNPs that distinguish the two parental haplotypes and to track parent-of-origin DNA methylation. Two cohorts of mice were fed either a diet depleted of key methyl-donors (choline and folic acid) or a control diet (**Figure 5-2**). Both cohorts were sampled at 15 weeks of age to study acute environmental effects on DNA methylation. Mice that were fed a methyl-deficient diet were then switched to the control diet for an additional 10 weeks to study the stability of environmentally induced changes in allele-specific DNA methylation. Our preliminary analysis confirms our original observation that genetic variation in *cis* drives the methylation state of nearby CpGs.

We observed global changes in DNA methylation in the two experimental conditions (**Figure 5-3**). These results also suggest a strain by parent-of-origin by diet effect. We also observe local effects such as parent-of-origin, strain, and diet (**Figure 5-4**).

Given the different strains used, this study may be used to expand the catalog of strain-specific DMRs (**Chapter II**) and perhaps parent-of-origin DMRs. Also, the optimized protocol increases the number of methylation informative probes by 37% and reduces the ambiguity seen by using two different fragmentation enzymes. In summary, this dataset shows promise for untangling the effects diet has on the methylome both at a global and local level.

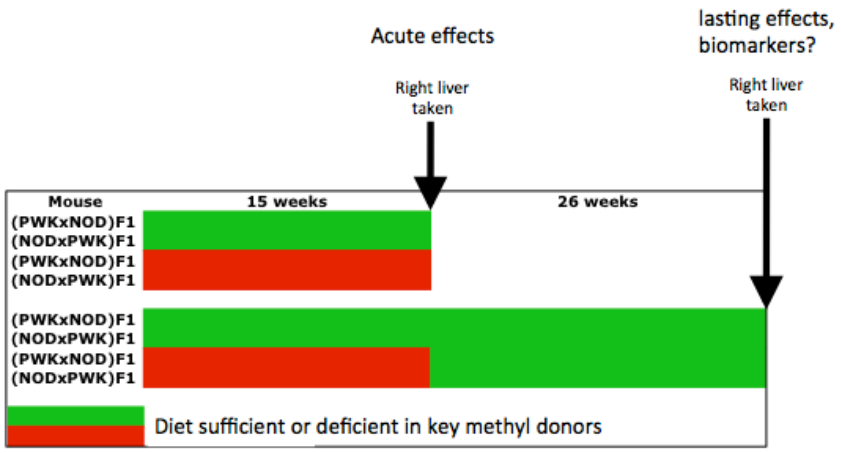


Figure 5-2. Experimental Design. Shown are the crosses, diet and time of sacrifice for the 48 mice in the study. There were six biological replicates for each experimental condition.

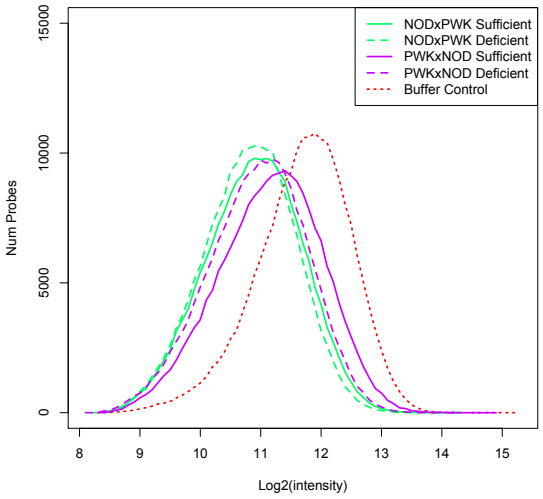
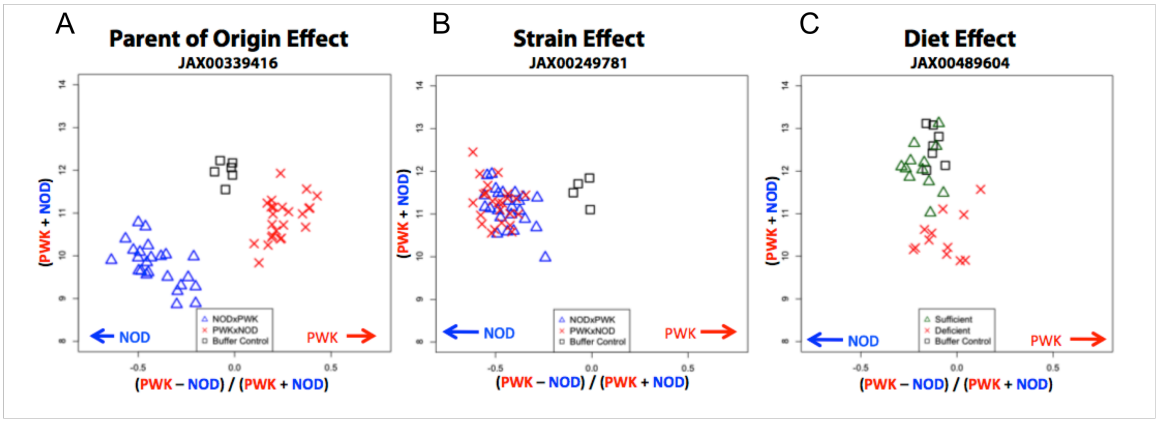


Figure 5-3. Global differences in DNA methylation at 26-weeks of age. Shown are the distributions of probe hybridization intensity for each of the four experimental conditions.



Actn1 Maternal DMR | Gene desert on proximal chr. 1 | *Fmnl2* intron *1*

Figure 5-4. Local DNA methylation effects. Shown are hybridization plots for three different loci showing parent-of-origin, strain, and diet effects. In panels A and B, triangles represent (NODxPWK)F1 hybrids, and red X's represent (PWKxNOD)F1 hybrids. In panel C, green triangles represent mice fed a choline sufficient diet, while red X's are those fed a choline deficient diet.

RNAseq libraries were created at the Jackson Laboratory from a subset of the 48 mouse livers used for MSNP analysis shown above. Among other analyses, this data set could be used to study the phenotypic effects of differential DNA methylation. Interesting comparisons would be gene expression changes between:

- diet
- age (15 and 26 weeks)
- allele-specific
- parent-of-origin
- combinations of all four

If informative SNPs tagging DNA methylation reside near these differentially expressed genes, it would be interesting to see if there is a correlation between changes in gene expression and changes in DNA methylation. Also, are environmental exposures that

change the methylome acute, or do they persist over time? And does this correlate with changes in gene expression with age?

SECTION TWO

Shown below are general conclusions drawn from section two:

- *Xce* is a distinct locus and not merely the result of genetic variation within *Xist*, *Tsix*, or *Xite*.
- The *Xce* candidate interval contains a segmentally duplicated region and CNV between allele carriers may explain the *Xce* allelic series
- The genetic architecture of *Xce* is rich in *Mus,* with six functional alleles identified to date.
- Maternal inheritance of the strong *Xce* allele magnifies XCI skewing 9%.
- Our results support the hypothesis that *Xce* is a *trans*-factor binding site that acts upstream of *Xist*.
- Taken together, genetic variation impacts both CpG methylation and XCI in mouse.

The *Xce* candidate interval

After 40 years of research, the *Xce* candidate interval stood at 1.85 Mb and included major players in the XCI process, *Xist, Tsix,* and *Xite*. The failure to substantially reduce the candidate interval size was primarily due to the high stochastic variability and large recombinant population needed to fine map. However, our approach overcame these challenges by integrated new high-density genotype data [1] and whole genome sequencing data [2, 3] with historical *Xce* phenotyping data. By doing so, were able to utilize association mapping and select strains with historical recombinations and significantly reduce the *Xce* candidate interval 10-fold to 176kb. Furthermore, we were able to exclude *Xist, Tsix* and *Xite* as candidates for *Xce* (**Chapter IV, Figure 4-4**).

Importantly, the new candidate interval is of much smaller size and thus is more amenable for mechanistic studies. For instance, the interval is small enough to perform homologous recombination experiments. These may include sequence exchanges in ES cells to recapitulate the XCI skewing in *Xce* heterozygotes. 129S1/SvlmJ ES cell lines are homozygous for Xce^{a}. By exchanging the candidate interval with a different functional allele, say CAST/EiJ (Xce^{c}), and select for heterozygous ES tranformants, the skewing of XCI choice in the differentiated ES population should mirror skewing seen in 129S1/SvlmJxCAST/EiJ F1 female hybrids. Alterations to the copy number and linear arrangement of the segmentally duplicated region could be used to gain insight into the molecular mechanism itself. For instance, is it the copy number, linear arrangement or both that distinguishes the different functional alleles? Are there specific duplications that distinguish the functional alleles, or can some be removed without influencing the function? Answers to these questions may divulge whether the molecular mechanism is involves binding site duplications, orientation, or topology differences that alters *Xce's* proximity to the *Xic*.

The smaller interval size may also be amenable to biochemical studies. These include, but are not limited to: Crosslinking proteins to DNA at the critical window of XCI choice and then pulling down the *Xce* candidate sequence; characterizing chromatin structure and post-translational modifications during XCI choice; and characterizing DNA methylation changes before and after XCI choice. However, the success of these experiments depends on determining the *Xce* sequence for each functional *Xce* allele.

Xce allelic series

Xce was initially discovered in 1965 [146] through the observation that female mice heterozygous for Cattanach's translocation ((*T(1;X)Ct*), [198]) segregated for 'high' and 'low' levels of white coat color variegation. Unbeknownst to Cattanach and colleagues, the mouse strains used to derive the translocation line carried two functional alleles (CBA, Xce^{a};

PCT, *Xce*^{b}). Thus the two 'states' of coat color variegation were caused by either XCI skewing in females heterozygous for *Xce*$^{a/b}$ or random XCI in females homozygous for *Xce*$^{a/a}$ or *Xce*$^{b/b}$. Since that time, four functional *Xce* alleles have been identified in *Mus* (*Xce*$^{a,b,c,\ and\ d}$) and a total of 16 inbred strains phenotyped (not including sister strains). Our study brings that number to 26, nearly doubling the strains phenotyped. Furthermore, we have identified two new functional alleles, one in a strain of *M. m. musculus* descent (PWK/PhJ, *Xce*^{e}), and one derived from a mouse of *M. spicilegus* descent (PANCEVO/EiJ (*Xce*^{f}). These findings reveal that genetics play a significant role in XCI in the laboratory mouse. The *Xce* allelic series creates a wide-range of XCI skewing from as little as 45:55 (*Xce*$^{a/e}$*)* to nearly 0:100 (*Xce*$^{a/d}$). *Xce* skewing is present in both classical inbreds as well as intersubspecific crosses. Although the degree of mean XCI skewing is minimal in commonly used classical inbred strains (mean XCI skew in *Xce*$^{a/b}$ heterozygotes is 40/60), nearly 50% of all F1 crosses between classical inbred strains will have mean XCI skewing (**Figure 4-8**). Furthermore, we predict that all intersubspecific crosses will result in *Xce* heterozygosity and mean XCI skewing. These results demonstrate that *Xce* has a significant effect on X-linked gene expression and should be incorporated into future X-linked gene expression models.

The evolution of the *Xce* allelic series

Taken together, our results significantly expand our understanding of the genetic architecture of XCI choice. The number of functional *Xce* alleles is consistent with the idea that the *Xce* allelic series is the result of CNV of a *trans*-factor binding site. Unlike a SNP that may or may not result in a new functional allele, each change in the copy number of the *Xce* region results in a new functional allele. Interestingly, we find that each species or subspecies has its own functional allele and there is no sharing between them. This observation could be due to the limited number of strains phenotyped in each branch. However, the results from the *M. m. domesticus* and *M. m. musculus* branch suggest otherwise. Within the *M. m. domesticus* branch, we find that *Xce*^{b} is common to both

classical inbred strains as well the sole allele in wild-derived strains (**Figure 4-8**). Yet, *Xce*^{a} is specific to classical inbred strains only. These results indicate that either *Xce*^{a} is a rare allele in the wild or a new, derived allele-specific to classical inbred strains. If the latter is the case, then *Xce*^{b} is the ancestral allele.

Furthermore, we have phenotyped two additional strains of *M. m. musculus* origin, SKIVE/EiJ and JF1/Ms. The *Xce*^{SKIVE} and *Xce*^{JF1} alleles were tested against known *Xce*^{a} and *Xce*^{b} carriers, and yet produced inconclusive results (data not shown). Although, we cannot assign a specific allele to the two strains, we can conclude that SKIVE/EiJ and JF1/Ms do not harbor the *Xce*^{c} or *Xce*^{d} alleles. Unfortunately, we did not have access to reciprocal crosses involving SKIVE/EiJ and JF1/Ms and believe the parent-of-origin effect is the cause of our uncertainty. However, the XCI skewing recorded is completely consistent with a maternal parent-of-origin effect involving the *Xce*^{e} allele (**Figure 4-9**, data not shown).

If the lack of functional diversity observed is not merely a sampling issue, then perhaps there is a fitness component to XCI skewing. Eutherian mammals undergo random XCI, unlike the rest of theria, such as metatheria and monotremes [199]. From a fitness perspective, random XCI balances the use of each parental X chromosome in half for a given female. It is possible that within a panmictic population, skewing of XCI would bias X chromosome expression towards one X chromosome over another. Mutations in linkage disequilibrium with *Xce* would either be preferentially activated or inactivated depending on the strength of the *Xce* allele. It would stand to reason that deleterious or advantageous alleles linked to *Xce* might be selected against or for, depending on the strength of the *Xce* allele. Therefore the apparent lack of functional diversity may be a result of a selective sweep towards one functional allele over another in a randomly mating population. Dr. Alan Lenarcic and Dr. Will Valdar have attempted to simulate this effect, with limited success, using a computer model of a population segregating for two different functional *Xce* alleles alleles. Interestingly, the simulations yielded either insignificant or no selection whatsoever,

and no simulation completely eliminated an allele from the population. These results may be due to incomplete modeling, or simply that selection does not occur at *Xce*. However, this is an attractive hypothesis that deserves additional investigation.

How our results fit the current model of XCI

It was proposed by Mary Lyon in 1971 that *Xce* may serve as a binding site for a jointly synthesized autosomal *trans*-factor involved in the early stages of XCI sensing and choice [200]. Alternatively, she proposed a model where the initial factor originates from the X chromosomes and actives an autosomal factor, which in turn inactivates one X chromosome [28]. Critical to these models is the stochiometry of X to autosomal factors [131-133]. Since that time, a few models have been proposed that incorporate recent findings in XCI research [130]. Common to all of these models is communication between autosomes and X chromosomes, and that typically involves X-linked *trans*-factor binding site or X-linked signal that induces an autosomal response. Keeping with these proposed models, I put forth an expanded hypothesis that incorporates our *Xce* findings. I propose that the different functional *Xce* alleles are distinguished by expansions and/or contractions of a series of tandem segmental duplications. Specifically, the variation in the segmentally duplicated region changes the probability of either a *trans*-factor binding *Xce* that promotes *Xist* expression (leading to XCI in *cis*) or promotes *Tsix* (leading to an active X) (**Figure 5-5A**). An alternative hypothesis is that *Xce* is a non-coding RNA that either promotes *Xist* or *Tsix* expression in *cis*. Duplications of this non-coding RNA changes the probability of that chromosome remaining active (**Figure 5-5B**). The simplest model to explain the *Xce* allelic series and the rapid evolution of the locus is that variation in the number of segmental duplications result in the change of copy number of the *trans*-factor binding site. Therefore, the 'weaker' *Xce* alleles (*i.e., Xce*^{a}) would have more copies and would increase the probability of binding the sensing/counting *trans*-factor and initiate the XCI process (**Figure 5-5B**, blue chromosome). Likewise, the 'stronger' *Xce* alleles would have fewer copies and

thus have a lower probability of binding the *trans*-factor and undergoing XCI (**Figure 5-5B**, red chromosome). To test this hypothesis the *Xce* candidate interval will have to be assembled properly for at least two different functional alleles (see **Ongoing work**).

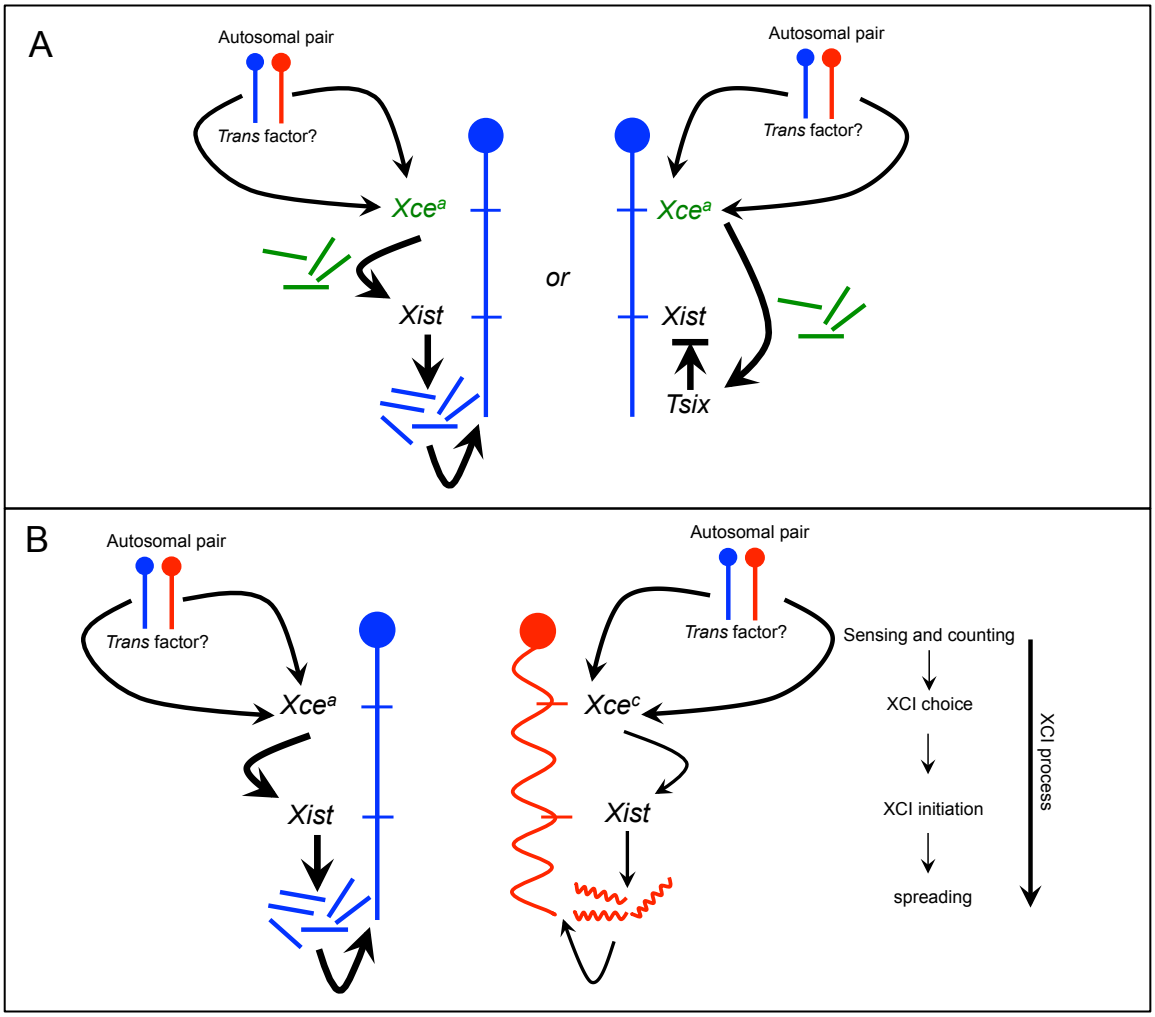


FIGURE 5-5. How *Xce* fits into the current XCI choice mechanism. Shown in **Panel A** are two possible mechanisms of *Xce*. First, an autosomal *trans*-factor binds *Xce* that in turn induces either *Xist* or *Tsix*. *Xce* may act as an enhancer or insulator, or may produce a non-coding RNA that acts in *cis*. In **Panel B** two X chromosomes with different functional *Xce* alleles (*Xce*^{a} *and Xce*^{c}*)* are represented with blue and red lines, respectively. Shown to the left is a timeline of the XCI process beginning with the unknown sensing and counting mechanism and ending with the spreading and maintenance stages. The different functional alleles are distinguished by their probabilities of inducing *Xist* expression. Shown are *Xce*^{a} *and Xce*^{c}. The *Xce*^{a} allele has a greater chance of binding the *trans*-factor and thus has a higher chance of inactivation when in heterozygosity with *Xce*^{c}.

Modeling XCI choice

The overarching goal of our XCI model is to precisely estimate the XCI pattern in an individual female and then use that information to understand stochastic, genetic and epigenetic factors influencing XCI choice in female populations. Here, I provide a summary of our XCI choice model; relay what it tells us about different factors influencing choice; and discuss its impact on future XCI research. A more detailed explanation can be found elsewhere [171].

Unlike a typical regulatory locus that controls gene expression of one or a few gene products nearby, *Xce* is unique because it indirectly influences the expression of hundreds of genes across the entire X chromosome. Currently, the effects of *Xce* can be observed only after the fact; once the XCI process has occurred. We measure XCI choice by estimating the proportion of cells that have an inactive maternal versus paternal X chromosome. To do so, we use allele-specific (in this case, parental specific) X-linked gene expression as a surrogate. We then measure X-linked gene expression in a tissue with the expectation that it truly reflects the original maternal/paternal XCI progenitor cell stoichiometry caused by primary XCI choice.

Our model [171] attempts to parameterize sources of XCI choice variability and uncertainty (**Figure 5-6**). Three factors mainly contribute to mean XCI skewing (although we have yet to determine if and to what extent they may also influence variance): *Xce*, parent-of-origin, and secondary skewing. And, there are two factors that mainly contribute to XCI variance within a population: chance and autosomal modifiers (**Figure 5-6**).

The variance of XCI within a female population depends on the number of cells that undergo independent choice. An analogy that best describes the two possible outcomes of XCI is flipping a coin, which represents the probability facing each cell that undergoes independent choice. In *Xce* homozygotes, the coin has an equal probability of landing on heads or tails. The number of flips allowed will buffer the mean from deviating too far from

50/50, which represents the number of cells that undergo independent choice. The variance, therefore, will be greater in females with fewer cells that undergo independent choice (orange, **Figure 5-6**). In *Xce* heterozygotes, the coin is weighted and thus one side has a higher probability of landing face up, leading to mean XCI skewing (dark green, **Figure 5-6**). The weight differential between two sides of the coin (*Xce* allele strength) determines the number of females needed to show significant deviation from 50/50. For instance, $Xce^{c/a}$ heterozygotes (~75/25) deviate far from 50/50, while $Xce^{a/e}$ heterozygotes are close to 50/50 (45:55). For this reason, it would take far more females to show mean skewing in $Xce^{a/e}$ heterozygotes than $Xce^{c/a}$ heterozygotes. Therefore, the certainty of *Xce*-induced XCI skewing is dependent on the number of females phenotyped and the number of cells that undergo independent choice (which is related to the tissue type used).

It has been shown that the variation in XCI is also impacted by autosomal modifiers [163]. Mapping these modifiers is of great interest to the XCI community, because they could be the dosage factors that communicate with the X chromosome during the early stages of sensing and counting. Currently, we are using the CC and DO populations to map both the autosomal modifiers and parent-of-origin effect (see **Ongoing work**). The combination of our model and these two mouse resources provide an opportunity to finely map these factors.

In the past, X –linked traits were considered low-hanging fruit: easily identified by their severe phenotypes in males [201], XO females [202], or XX females with 100% skewing in favor of the deleterious allele [203]. But more recently, mapping complex traits involving the X chromosome have fallen by the wayside because of XCI skewing and variability within female populations. Our study facilitates future mapping studies by estimating and predicting the contribution of each factor influencing XCI (**Figure 5-6**). Furthermore, our model allows for characterization of additional factors influencing XCI choice such as secondary skewing caused by X-linked mutations affecting cell survival or

proliferation (**Figure 5-6**). Taken together, this model and our large experimental dataset create a novel resource for future X-linked association mapping studies.

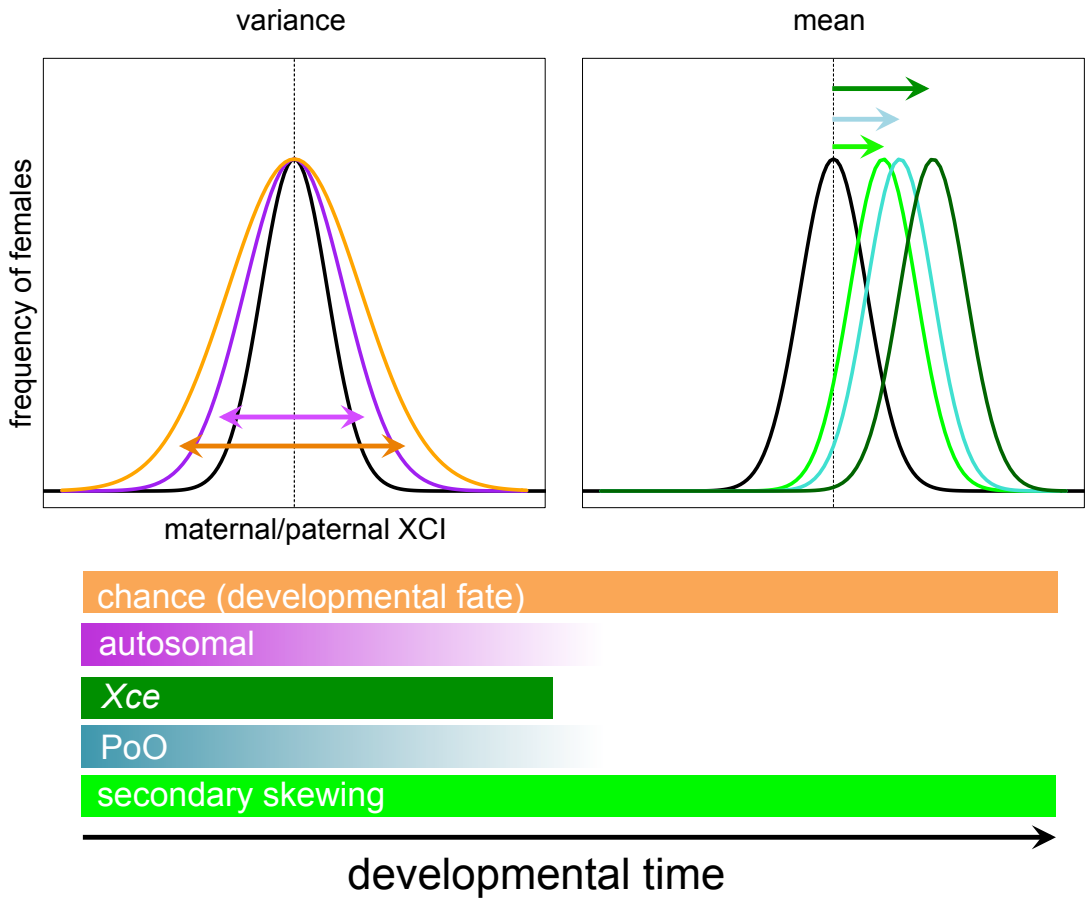


Figure 5-6. Factors influencing XCI choice. Shown are five different factors influencing XCI choice. The distributions show the XCI patterns seen in hypothetical female populations. Chance and autosomal factors have been shown to influence the variability of XCI choice within a population [163, 204]. Mean XCI choice is influenced by *Xce*, PoO and secondary skewing. The amount of skewing shown is not set in stone and will change depending on the genotype of *Xce* and X-linked variation (secondary skewing).

From a developmental perspective, our study may provide a platform for characterizing mammalian cellular lineages during key developmental stages [205, 206]. For example, the variation in XCI patterns between tissues suggests that the pool of cells that gives rise to each tissue type varies. Our model gives a sense of the number of cells that

undergo independent choice and ultimately contribute to each tissue (**Chapter IV, Discussion**).

Finally, given the predictive power of our phylogenetic analysis of the *Xce* candidate interval, this information could be used to generate crosses that lack mean XCI skewing, or intentionally select strains with XCI skewing. The *Xce* allelic series provides a natural way to dial up or down the X chromosome expression from one strain or another. This provides an opportunity to study X-linked variation by crossing strains harboring different *Xce* functional alleles to create naturally skewed female mosaics.

Phenotypic consequences of skewed XCI

RNAseq is an especially powerful tool for determining XCI skewing because of the large number of eSNP measurements across the X chromosome. Where as a single gene may mislead due to tissue or *cis*-driven differential expression, RNAseq provides comprehensive allelic-imbalance information anywhere informative eSNPs are expressed. In addition to increased accuracy for estimating the ratio of maternal to paternal XCI, RNAseq also reveals the significant and chromosome-wide phenotypic consequences of *Xce* functional heterozygosity. Take for example **Figure 5-7.** The female in **Panel A** displays chromosome-wide allelic imbalance in favor of the CAST X chromosome (76% CAST). This is interpreted as 76% of the cells are expressing the CAST X chromosome over the WSB X chromosome. (CASTxWSB)F1 0113 has an X expression profile similar to a CAST inbred female instead of a true hybrid between CAST and WSB. The female hybrid shown in **Panel C** is a prime example of the stochastic and parent-of-origin factors influencing choice. Given the genotype at *Xce* ($Xce^{b/c}$) the expectation is a mean XCI skewing of ~70% in favor of the CAST X chromosome. Although both females are genetically identical (with the exception of mitochondria), the parent-of-origin of the X chromosomes is different. This has the effect of reducing XCI skewing if the strong allele is inherited through the paternal germline (**Chapter IV, Figure 4-9**). Lastly, *Xce*

heterozygosity may have subtle changes in X chromosome skewing as seen in **Panel B**. (PWKxWSB)F1 0084 has XCI skewing in favor of the WSB X chromosome (~53%).

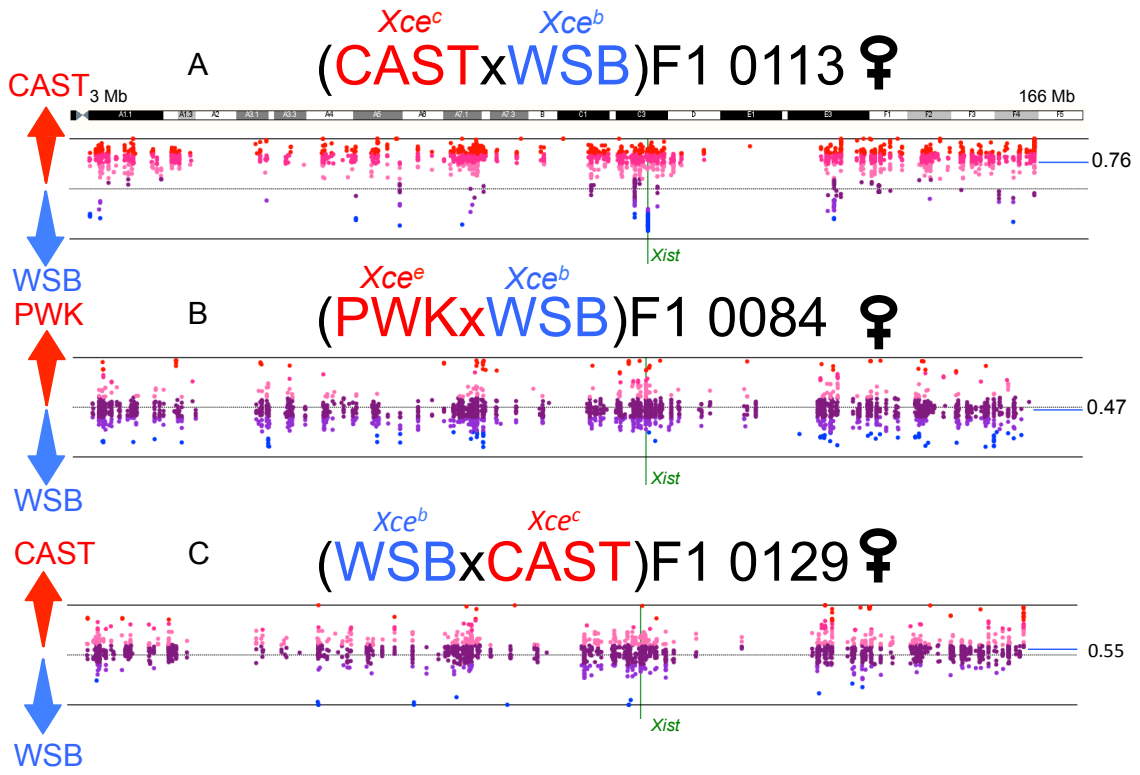


Figure 5-7. X chromosome-wide allelic imbalance as a result of XCI skewing. Shown in **Panel A** is allele-specific gene expression across the entire X chromosome for a single female F1 hybrid (CASTxWBS)F1 0113. Each dot represents an informative eSNP that tracks allelic ratio of expression between the CAST and WSB X chromosomes. The y axis is the number of CAST reads divided by total number of reads. The color of the dot signifies the proportion of RNAseq reads that comes from each parental X chromosome. Red, CAST (PWK in **Panel B**); Blue, WSB and genes expressed close to 50/50 are a hybrid of the two colors, purple. **Panel B** is a (PWKxWSB)F1 hybrid female. **Panel C** is a (WSBxCAST)F1 hybrid female.

Ongoing work: sequence characterization of the *Xce* candidate interval

The *Xce* candidate interval must be assembled correctly for future mechanistic studies. Previously, we attempted to use publically available next-generation sequencing of mouse strains with known *Xce* alleles to determine CNV within the candidate interval ([2, 3], data not shown), but were unsuccessful. The main reasons are: 1) segmentally duplicated

regions are difficult to assemble *de novo*; and 2) CNVs are difficult to interpret with short next-generation sequencing reads available [2, 3]. Listed below is a step-by-step tailored approach to overcome the challenges that accompany segmentally duplicated regions:

1) Characterize bacterial artificial chromosomes (BACs) that span the *Xce* candidate interval from mouse libraries that were derived from inbred strains with different functional *Xce* alleles.

    a. I cultured 18 BAC clones from five different mouse libraries (C3H/HeJ, C57BL/6J, C57BL/6NJ, NOD/LtJ and MsM/Ms) that represent 3 different functional *Xce* alleles (*Xce*^{a}, *Xce*^{b}, *Xce*^{e}) **Figure 5-8**.

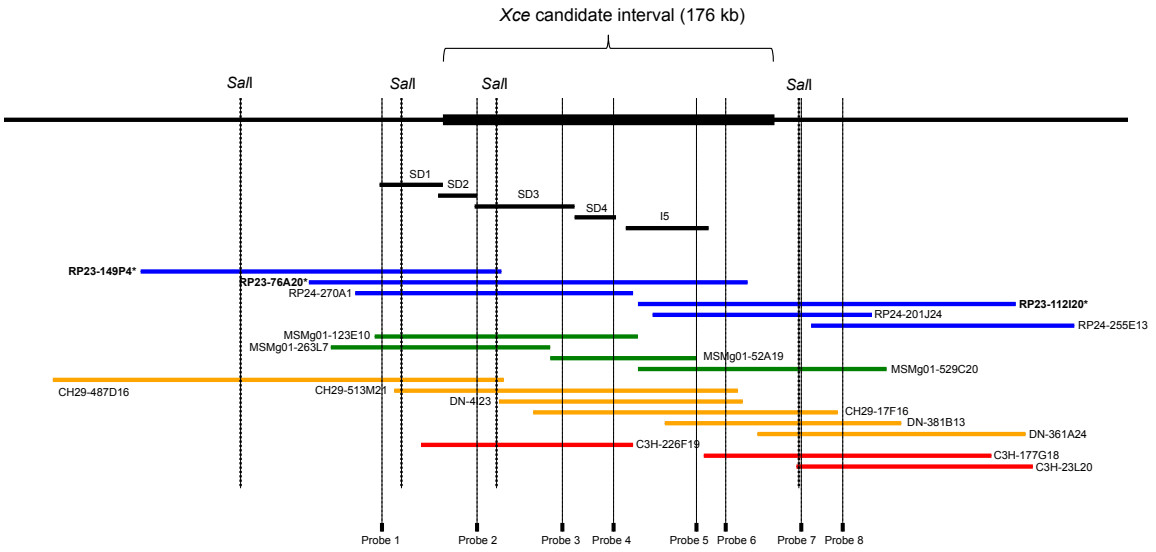


Figure 5-8. Map of BACs that span the candidate interval. This figure displays the segmental duplications labeled SD1-I5. Shown underneath the duplications are BACs that span the *Xce* candidate interval. BACs derived from mouse strains C57BL/6J and C57BL/6NJ are shown in blue. BACs in bold with an asterisk are fully sequenced [172], while the remaining are mapped based on end-sequencing. In orange are BAC clones derived from two different NOD/LtJ libraries; in red, from C3H/HeJ; and in green, from MSM/Ms. Also shown are the restriction sites for *Sal*I and Southern blot probes.

2) Assemble the C57BL/6J reference by determining BAC sizes using Pulse Field Gel Electrophoresis and by using restriction fragment analysis.

    a. Shown in **Table 5-1** are the BAC sizing results for the 18 BAC clones.

PFGE Sizing Results

	Clone name	Library	Derived from	*Xce* allele	Start (mm9)	Stop (mm9)	End-sequencing length (kb)	Estimate size (PFGE) (kb)	difference (kb)		
C57BL/6J (Xce^b)	RP23-149P4	RPCI-23	C57BL/6J	b	99,781,893	99,974,664	192.8	196.2	3.4		Xce^b
	RP23-76A20	RPCI-23	C57BL/6J	b	99,871,536	100,105,736	234.2	97.0	-137.2	*	
	RP23-112I20	RPCI-23	C57BL/6J	b	100,047,468	100,249,207	201.7	198.7	-3.0		
	RP24-270A1	RPCI-24	C57BL/6NJ	b	99,896,469	100,044,114	147.6	151.1	3.5		
	RP24-201J24	RPCI-24	C57BL/6NJ	b	100,055,682	100,172,447	116.8	157.4	40.6	*	
	RP24-255E13	RPCI-24	C57BL/6NJ	b	100,139,454	100,279,500	140.0	145.6	5.5		
NOD/LtJ (Xce^b)	DN-4I23	DIL-NOD	NOD/LtJ	b	99,972,803	100,103,402	130.6	162.9	32.3	*	
	DN-381B13	DIL-NOD	NOD/LtJ	b	100,061,148	100,188,157	127.0	131.7	4.7		
	CH29-487D16	CHORI-29	NOD/LtJ	b	99,735,210	99,975,240	240.0	250.3	10.2	*	
	CH29-513M21	CHORI-29	NOD/LtJ	b	99,916,632	100,101,122	184.5	224.6	40.1	*	
	CH29-17F16	CHORI-29	NOD/LtJ	b	99,991,978	100,153,987	162.0	199.7	37.6	*	
C3H/HeJ (*Xce^a*)	C3H-226F19	C3H iBAC	C3H/HeJ	a	99,931,247	100,044,075	112.8	154.6	41.8	*	Xce^a
	C3H-177G18	C3H iBAC	C3H/HeJ	a	100,082,631	100,235,772	153.1	165.0	11.8		
	C3H-23L20	C3H iBAC	C3H/HeJ	a	100,131,701	100,257,848	126.1	168.5	42.3	*	
MSM/Ms (Xce^e)	MsMg01-263L7	MsMg01	MSM/Ms	e	99,883,258	100,000,816	117.6	143.5	25.9	*	Xce^e
	MsMg01-123E10	MsMg01	MSM/Ms	e	99,907,218	100,047,448	140.2	95.5	-44.7	*	
	MsMg01-52A19	MsMg01	MSM/Ms	e	100,000,811	100,077,933	77.1	99.8	22.7	*	
	MsMg01-529C20	MsMg01	MSM/Ms	e	100,047,443	100,179,746	132.3	160.8	28.5	*	

Table 5-1. Pulse field gel electrophoresis BAC sizing result. This table shows PFGE results for each BAC examined. Clones with a PFGE size that does not agree with its end-sequencing size have an asterisk.

b. Restriction mapping using *Sal*I and *Sac*II revealed size discrepancies between predicted fragment sizes and observed fragment sizes (**Table 5-2, Figure 5-8**).

		Predicted (kb)						Actual (kb)				Southern	
Clone name	Library	total length	frag 1	frag 2	frag 3	frag 4	PFGE length	frag 1	frag 2	frag 3	frag 4	predicted (kb)	actual(kb)
RP23-149P4	C57BL/6J	203.8	54	86	50	3	196.2	85	52	-	-		
RP23-76A20	C57BL/6J	245.2	50	50	134	-	97.0	96	-	-	-		
RP23-112I20	C57BL/6J	212.7	85	116	-	-	198.7	-	-	-	-		
RP24-270A1	C57BL/6NJ	158.6	25	50	73	-	151.1	32	58	83	-	50	57
RP24-201J24	C57BL/6NJ	127.8	77	40	-	-	157.4	119	47	-	-		
RP24-255E13	C57BL/6NJ	151.0	140	-	-	-	145.6	156	-	-	-		
MsMg01-263L7	MSM	128.6	38	50	29	-	143.5	64	43	-	-	50	57
MsMg01-123E10	MSM	151.2	14	50	76	-	95.5	98	-	-	-		
MsMg01-52A19	MSM	88.1	77	-	-	-	99.8	119	-	-	-	N/A	77
MsMg01-529C20	MSM	143.3	85	47	-	-	160.8	111	90	70	55		
CH29-487D16	NOD/LtJ	251.0	100	86	50	-	250.3	112	93	57	-	50	46
CH29-513M21	NOD/LtJ	195.5	130	50	-	-	224.6	162	47	-	-		
DN-4I23	NOD	141.6	131	-	-	-	162.9	167	-	-	-		
CH29-17F16	NOD/LtJ	173.0	141	21	-	-	199.7	183	17	-	-		
DN-381B13	NOD	138.0	72	55	-	-	131.7	75	61	-	-		
C3H-226F19	C3H	123.8	40	73	-	-	154.6	105	29	23	-		
C3H-177G18	C3H	164.1	50	103	-	-	165.0	47	112	-	-		
C3H-23L20	C3H	137.1	137.1	-	-	-	168.5	114	13	12	-		

Table 5-2: *Sal*I digestion results and Southern blot analysis using probe two. Clones in green shown Southern results that agree with the predicted sequence and size of the restriction fragment. Clones in yellow either did not contain probe sequence or had a size discrepancy.

c. Southern blot analysis confirmed that restriction fragments of digested BAC DNA containing the probe sequence are present and are of the correct size.

3) Identify BACs of interest for next-generation sequencing using PacBio long-read technology.

a. As controls, RP23-149P4, RP23-76A20 (rederived), and RP23-112I20 will be sequenced. In addition, MsM-263L7, MsM-123E10, MsM-52A19, MsM-

529C20, C3H-226F19, C29-513M21, and DN-4I23 will be sequenced because they span the segmentally duplicated region and show multiple lines of evidence that their end-sequence map length is inaccurate and their internal sequence cannot be represented by the reference sequence.

4) Assemble the *Xce* candidate interval correctly for each functional *Xce* allele carrier. This includes determining the proper copy number and linear arrangement of the duplications.
    a. The success of this step requires that there is adequate read depth and sufficient overlap between BACs to assemble the *Xce* region *de novo* for each allele carrier. Several assembly problems are anticipated even with the long sequencing technology such as incorporating sequencing errors during multiple reads of the same circularized template DNA [207]. The long-read technology of PacBio has a trade-off: The length of the read is correlated to error rate. Therefore, the longer the read, the higher the error rate. A solution to this problem would be to combine both short read (Illumina) sequencing and BacBio technology.
5) Annotate CNVs, SNP and indels for each strain.
    a. A read threshold will need to be established for *de novo* SNP, CNVs and indels. Previous studies using PacBio technology will guide this process [207].
6) Identify candidate *Xce* loci (locus).
    a. The candidates must fulfill two requirements: 1) strains with the same *Xce* allele should share the same sequence and 2) that sequence must be different from unlike *Xce* allele carriers.

7) Expand sequence analysis to mouse strains without genomic libraries but with known *Xce* alleles. This is an important step because of the limited number of mouse genomic libraries available.
    a. Pyrosequencing assays will be designed that target paralogous variation between duplications identified by PacBio. These assays will be sensitive enough to detect two-fold increase in copy number between inbred strains.
8) Examine the sequence of *Xce.* The sequence itself (*e.g.,* secondary structure or binding motifs) might provide insight into the molecular mechanism of XCI choice and direct future experiments.

Essentially, this approach will mirror our association mapping in **Chapter IV**.

Ongoing work: Mapping parent-of-origin and autosomal modifiers

Previous studies have shown that parent-of-origin influences XCI skewing [148, 149, 163]. We have further characterized this phenomenon by demonstrating that maternal inheritance of the strong *Xce* allele magnifies *Xce* skewing by ~9% (**Chapter IV**). It is unclear if the parent-of-origin locus(i) is *Xce* or somewhere else on the X chromosome., The CC mouse resource population and perhaps sibling pairs of DO mice may be used to map the parent-of-origin effect [64, 81]. Briefly, the CC and DO are genetic mosaics of eight inbred founders [64, 81] that have four functional *Xce* alleles segregating between them (**Figure 4-8**).

Reciprocal CC-RIX female mice (F1's between finished CC lines) offer a quick approach to roughly map the parent-of-origin effect. For example, the haplotype reconstruction of the X chromosome between CC-RIX lines associated with the presence or absence of the parent-of-origin effect can be used to quickly include or exclude regions of identity by descent (IBD) (**Figure 5-9**). The presence of the parent-of-origin effect excludes regions of IBD between CC-RIX X chromosomes, while regions of IBD must be included in the absence of the parent-of-origin effect (**Figure 5-9**). The analysis could be tailored to

include the possibility of multiple loci contributing to the parent-of-origin effect by allowing effects less than the average maternal effect of 9%.

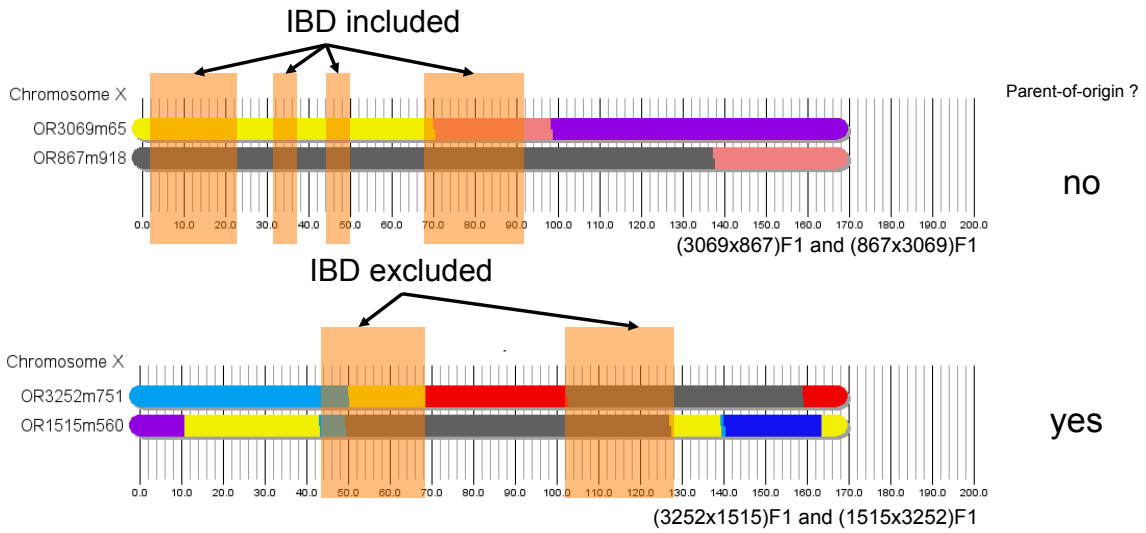


Figure 5-9. Mapping the parent-of-origin effect. Shown are the X chromosome haplotype reconstruction of two CC-RIX mice, (3069x867)F1 and (3252x1515)F1. Each color denotes a CC founder strain, A/J-yellow, C57BL6-gray, 129S1-pink, NOD-dark blue, NZO-light blue, CAST-green, PWK-red, and WSB-purple. Regions of IBD (orange areas) are included or excluded in reciprocal CC-RIX depending on the presence or absence of the parent-of-origin effect.

Mapping autosomal modifiers affecting XCI choice is not as straightforward as mapping the parent-of-origin effect. A recent study attempted to map these modifiers, but none reach genome-wide significance despite nearly 1000 mice used in the study [163]. The Diversity Outbred population should address this problem given the recombination landscape and large number of segregating variants compared to the F2 mice used in the study above. Nevertheless, given the apparent multi-locus results shown by Chadwick *et al.* 2005, at least 1000 DO mice will need to be phenotyped for a robust QTL analysis.

To date, I have extracted RNA from 590 female DO mice, 456 of which have good MegaMUGA genotypes for haplotype reconstruction. Of those 456 mice, 398 have at least one informative pyrosequencing assay to measure XCI ratios (**Figure 5-10**). I have

measured XCI ratios in 247 of these mice for a total of 296 pyrosequencing measurements. Additional DO mice from various studies at UNC and the Jackson Laboratory are available and pyrosequencing analysis is underway to phenotype these animals.

The identification of parent-of-origin and autosomal modifiers may be a critical leap forward in XCI research if these modifiers are indeed the key players in the sensing and counting mechanism.

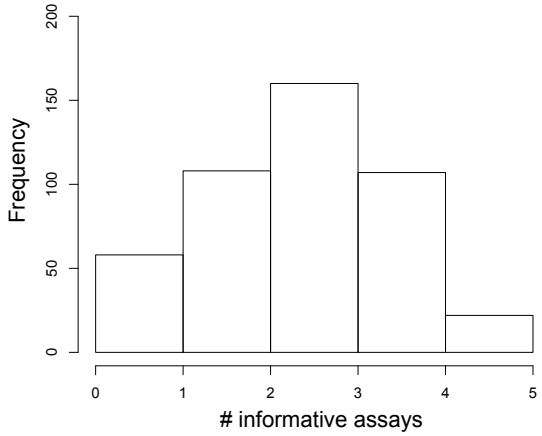


Figure 5-10: Distribution of females with informative pyrosequencing assays. Of the 456 DO females with genotype data, 58 do not have an informative eSNP capture by the pyrosequencing assays, 108 have one, 161 have two, 107 have three, and 22 females have four.

The genetics of epigenetics

Taken together, my work provides a genetic perspective on epigenetic processes in mouse and demonstrates that genetic variability within populations is a major contributor to epigenetic variance. Furthermore, this work reinforces the idea that underlying DNA sequence and epigenetic mechanisms are intimately linked and influence one another. Ultimately, epigenetic variability may have small or large phenotypic consequences demonstrated by the spectrum of XCI skewing in *Xce* heterozygotes. Finally, this dissertation provides significant insight into the genetic regulation of DNA methylation and XCI in mouse and builds a solid foundation for future mechanistic work.

MATERIALS AND METHODS

BAC culture and purification

A total of 20 BACs were ordered from the Children's Hospital Oakland Research Institute (RP23-149P4, RP23-76A20, RP23-112I20, RP24-270A1, RP24-201J24, RP24-255E13, CH29-513M21, CH29-487D16, CH29-17F16), Riken (MsMg01-123E10, MsMg01-263L7, MsMg01-52A19, MsMg01-529C20), The Sanger Institute (C3H-226F19, C3H-177G18, C3H-23L20), and the Center for Applied Genomics (DN-4I23, DN-117I4, DN-361A24, DN-381B13). LB stabs were used to strike LB agar plates containing 5% glucose and 12.5 µg/ml chloramphenicol. A single colony was plucked and cultured overnight in 100 ml of 2X YT media (Sigma) with 12.5 µg/ml chloramphenicol. Bacterial cultures were pelleted and BAC DNA extracted using BACMAX DNA purification kit (Epicentre). DNA quality and quantity was checked with a nanodrop spectrophotometer (Thermo Scientific).

PFGE and alkaline gel transfer

Approximately 30 µg of purified BAC DNA was suspended in 1% low melt agarose (Calbiochem) and transferred to CHEF Mapper plug molds (BioRad). Solidified plugs were cut in thirds and either placed in buffer (10mM Tris), digested with *Sal*I, or with *Sac*II according to manufacturer's protocol (New England Biolabs). Plugs were loaded into 0.9% low melt agarose gels and electrophoresed for 22 hours (6 v/cm, switch time 15"-35", 120°) using a CHEF-DR III pulse field electrophoresis system (BioRad). Gels were transferred to positively charged membranes (Zeta-Probe GT, BioRad) overnight using alkaline transfer.

Southern blotting

Membranes were prehybridized overnight using Ultrahyb (Ambion). 100 ng of purified DNA probes (see T**able S5-1**) targeting the duplicated region were radiolabeled using random hexamer primers and P_{32} dCTP (Invitrogen) and hybridized to the membrane

overnight. Membranes were washed in 2X SSC, 1% SDS for 30 minutes, followed by 0.2X SSC, 1% SDS for 30 minutes. Membranes were exposed to X ray film for visualization.

Name	Sequence	Tm	5'position
Probe1-F	GCCTTGGTTTCAGTATACAG	50.7	99910236
Probe1-Rev	AACTGTATTGTGTTTTCATTGC	50.4	99910890
Probe2-F	TTCTCCTTACAGGAGTGAAGG	53.5	99961227
Probe2-Rev	ACAACCGCCTGATCCATA	53.5	99962064
Probe3-F	CTCCAAAGCAAGACGGACATGG	58.3	100007398
Probe3-Rev	CAGGTGTTCGTGCAAGAGATGG	58.3	100008182
Probe4-F	CATGAAACAAGCATCACTCTG	52.7	100034409
Probe4-Rev	CAGAAATGATACAGCCACTAAGG	53.3	100034970
Probe5-F	TGCGATAGTGGGCTATGG	54.2	100093728
Probe5-Rev	AAATGCTGAAACTGCTAGAACG	53.7	100094575
Probe6-F	CACACAGACAGTCCTAGTCTAG	53.8	100133760
Prove6-Rev	GGTGTTCGATGAACCTGG	53.2	100134259
Probe7-F	GGTTTCCACGCATGTTATCC	54.1	100156526
Probe7-Rev	GATTGTTGGAGACATGGCTC	53.6	100157352
Probe8-F	GTAGTCCTTGCAGTTATGAAGG	53.0	100078161
Probe8-Rev	AGGGTATGGGGTACTTTTGG	54.1	100078764

Table S5.1: Primers used to generate probes for Southern blots.

REFERENCES

1. Yang, H., et al., *Subspecific origin and haplotype diversity in the laboratory mouse.* Nat Genet, 2011. **43**(7): p. 648-55.

2. Keane, T.M., et al., *Mouse genomic variation and its effect on phenotypes and gene regulation.* Nature, 2011. **477**(7364): p. 289-94.

3. Yalcin, B., et al., *Sequence-based characterization of structural variation in the mouse genome.* Nature, 2011. **477**(7364): p. 326-9.

4. Larkin, M.A., et al., *Clustal W and Clustal X version 2.0.* Bioinformatics, 2007. **23**(21): p. 2947-8.

5. Grewal, S.I. and S.C. Elgin, *Transcription and RNA interference in the formation of heterochromatin.* Nature, 2007. **447**(7143): p. 399-406.

6. Berger, S.L., *The complex language of chromatin regulation during transcription.* Nature, 2007. **447**(7143): p. 407-12.

7. Reik, W., *Stability and flexibility of epigenetic gene regulation in mammalian development.* Nature, 2007. **447**(7143): p. 425-32.

8. Feinberg, A.P., *Phenotypic plasticity and the epigenetics of human disease.* Nature, 2007. **447**(7143): p. 433-40.

9. Gendrel, A.V. and E. Heard, *Fifty years of X-inactivation research.* Development, 2011. **138**(23): p. 5049-55.

10. Augui, S., E.P. Nora, and E. Heard, *Regulation of X-chromosome inactivation by the X-inactivation centre.* Nat Rev Genet, 2011. **12**(6): p. 429-42.

11. Eckhardt, F., et al., *DNA methylation profiling of human chromosomes 6, 20 and 22.* Nat Genet, 2006. **38**(12): p. 1378-85.

12. Deaton, A.M., et al., *Cell type-specific DNA methylation at intragenic CpG islands in the immune system.* Genome Res, 2011. **21**(7): p. 1074-86.

13. Baccarelli, A. and V. Bollati, *Epigenetics and environmental chemicals.* Curr Opin Pediatr, 2009. **21**(2): p. 243-51.

14. Maegawa, S., et al., *Widespread and tissue specific age-related DNA methylation changes in mice.* Genome Res, 2010. **20**(3): p. 332-40.

15. Fraga, M.F., et al., *Epigenetic differences arise during the lifetime of monozygotic twins.* Proc Natl Acad Sci U S A, 2005. **102**(30): p. 10604-9.

16. Feinberg, A.P., *Epigenetics at the epicenter of modern medicine.* JAMA, 2008. **299**(11): p. 1345-50.

17. Fresard, L., et al., *Epigenetics and phenotypic variability: some interesting insights from birds.* Genet Sel Evol, 2013. **45**: p. 16.

18. Manavalan, L.P., et al., *RNAi-mediated disruption of squalene synthase improves drought tolerance and yield in rice.* J Exp Bot, 2012. **63**(1): p. 163-75.

19. Resendiz, M., et al., *Epigenetic medicine and fetal alcohol spectrum disorders.* Epigenomics, 2013. **5**(1): p. 73-86.

20. Beard, C., E. Li, and R. Jaenisch, *Loss of methylation activates Xist in somatic but not in embryonic cells.* Genes Dev, 1995. **9**(19): p. 2325-34.

21. Li, E., T.H. Bestor, and R. Jaenisch, *Targeted mutation of the DNA methyltransferase gene results in embryonic lethality.* Cell, 1992. **69**(6): p. 915-26.

22. Okano, M., et al., *DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development.* Cell, 1999. **99**(3): p. 247-57.

23. Young, T.J. and A.L. Kirchmaier, *Cell cycle regulation of silent chromatin formation.* Biochim Biophys Acta, 2012. **1819**(3-4): p. 303-12.

24. Peng, J.C. and G.H. Karpen, *Epigenetic regulation of heterochromatic DNA stability.* Curr Opin Genet Dev, 2008. **18**(2): p. 204-11.

25. Slotkin, R.K. and R. Martienssen, *Transposable elements and the epigenetic regulation of the genome.* Nat Rev Genet, 2007. **8**(4): p. 272-85.

26. Fedoroff, N.V., *Presidential address. Transposable elements, epigenetics, and genome evolution.* Science, 2012. **338**(6108): p. 758-67.

27. Schilling, E., C. El Chartouni, and M. Rehli, *Allele-specific DNA methylation in mouse strains is mainly determined by cis-acting sequences.* Genome Res, 2009. **19**(11): p. 2028-35.

28. Lyon, M.F., *X-chromosome inactivation and developmental patterns in mammals.* Biol Rev Camb Philos Soc, 1972. **47**(1): p. 1-35.

29. Aylor, D.L., et al., *Genetic analysis of complex traits in the emerging Collaborative Cross.* Genome Res, 2011. **21**(8): p. 1213-22.

30. Holliday, R. and J.E. Pugh, *DNA modification mechanisms and gene activity during development.* Science, 1975. **187**(4173): p. 226-32.

31. Riggs, A.D., *X inactivation, differentiation, and DNA methylation.* Cytogenet Cell Genet, 1975. **14**(1): p. 9-25.

32. Cedar, H. and Y. Bergman, *Linking DNA methylation and histone modification: patterns and paradigms.* Nat Rev Genet, 2009. **10**(5): p. 295-304.

33. Lazarovici, A., et al., *Probing DNA shape and methylation state on a genomic scale with DNase I.* Proc Natl Acad Sci U S A, 2013. **110**(16): p. 6376-81.

34. Ratel, D., et al., *N6-methyladenine: the other methylated base of DNA.* Bioessays, 2006. **28**(3): p. 309-15.

35. Yan, J., J.R. Zierath, and R. Barres, *Evidence for non-CpG methylation in mammals.* Exp Cell Res, 2011. **317**(18): p. 2555-61.

36. Reik, W., et al., *Genomic imprinting determines methylation of parental alleles in transgenic mice.* Nature, 1987. **328**(6127): p. 248-51.

37. Sharp, A.J., et al., *DNA methylation profiles of human active and inactive X chromosomes.* Genome Res, 2011. **21**(10): p. 1592-600.

38. Yoder, J.A., C.P. Walsh, and T.H. Bestor, *Cytosine methylation and the ecology of intragenomic parasites.* Trends Genet, 1997. **13**(8): p. 335-40.

39. Baylin, S.B. and J.E. Ohm, *Epigenetic gene silencing in cancer - a mechanism for early oncogenic pathway addiction?* Nat Rev Cancer, 2006. **6**(2): p. 107-16.

40. Rizwana, R. and P.J. Hahn, *CpG methylation reduces genomic instability.* J Cell Sci, 1999. **112 (Pt 24)**: p. 4513-9.

41. Irizarry, R.A., et al., *The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores.* Nat Genet, 2009. **41**(2): p. 178-86.

42. Hansen, K.D., et al., *Increased methylation variation in epigenetic domains across cancer types.* Nat Genet, 2011. **43**(8): p. 768-75.

43. Feinberg, A.P. and B. Tycko, *The history of cancer epigenetics.* Nat Rev Cancer, 2004. **4**(2): p. 143-53.

44. Walsh, C.P. and T.H. Bestor, *Cytosine methylation and mammalian development.* Genes Dev, 1999. **13**(1): p. 26-34.

45. Borgel, J., et al., *Targets and dynamics of promoter DNA methylation during early mouse development.* Nat Genet, 2010. **42**(12): p. 1093-100.

46. Christensen, B.C., et al., *Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context.* PLoS Genet, 2009. **5**(8): p. e1000602.

47. Rauch, T.A., et al., *A human B cell methylome at 100-base pair resolution.* Proc Natl Acad Sci U S A, 2009. **106**(3): p. 671-8.

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

49. Ball, M.P., et al., *Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells.* Nat Biotechnol, 2009. **27**(4): p. 361-8.

50. Tribioli, C., et al., *Methylation and sequence analysis around EagI sites: identification of 28 new CpG islands in XQ24-XQ28.* Nucleic Acids Res, 1992. **20**(4): p. 727-33.

51. Hellman, A. and A. Chess, *Gene body-specific methylation on the active X chromosome.* Science, 2007. **315**(5815): p. 1141-3.

52. Li, E., C. Beard, and R. Jaenisch, *Role for DNA methylation in genomic imprinting.* Nature, 1993. **366**(6453): p. 362-5.

53. Kerkel, K., et al., *Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation.* Nat Genet, 2008. **40**(7): p. 904-8.

54. Czyz, W., et al., *Genetic, environmental and stochastic factors in monozygotic twin discordance with a focus on epigenetic differences.* BMC Med, 2012. **10**: p. 93.

55. Kaminsky, Z.A., et al., *DNA methylation profiles in monozygotic and dizygotic twins.* Nat Genet, 2009. **41**(2): p. 240-5.

56. Wong, C.C., et al., *A longitudinal study of epigenetic variation in twins.* Epigenetics, 2010. **5**(6): p. 516-26.

57. Heijmans, B.T., et al., *Heritable rather than age-related environmental and stochastic factors dominate variation in DNA methylation of the human IGF2/H19 locus.* Hum Mol Genet, 2007. **16**(5): p. 547-54.

58. de La Casa-Esperon, E. and C. Sapienza. *Epigenetic variation: amount, causes and consequences*. Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics 2006; Available from: http://onlinelibrary.wiley.com/book/10.1002/047001153X.

59. Jablonka, E.L., Marion J; Zeligowski, A, *Evolution in Four Dimensions: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life (Life and Mind: Philosophical Issues in Biology and Psychology)*. 2006: A Bradford Book

60. Jablonka, E.a.L., Marion J, *Epigenetic Inheritance and Evolution: The Lamarckian Dimension*. 1995: Oxford University Press, USA.

61. Yuan, E., et al., *A single nucleotide polymorphism chip-based method for combined genetic and epigenetic profiling: validation in decitabine therapy and tumor/normal comparisons.* Cancer Res, 2006. **66**(7): p. 3443-51.

62. Yang, H., et al., *A customized and versatile high-density genotyping array for the mouse.* Nat Methods, 2009. **6**(9): p. 663-6.

63. Wang, J.R., et al., *Imputation of single-nucleotide polymorphisms in inbred mice using local phylogeny.* Genetics, 2012. **190**(2): p. 449-58.

64. Svenson, K.L., et al., *High-resolution genetic mapping using the mouse diversity outbred population.* Genetics, 2012. **190**(2): p. 437-47.

65. *The genome architecture of the collaborative cross mouse genetic reference population.* Genetics, 2012. **190**(2): p. 389-401.

66. Gregg, C., et al., *Sex-specific parent-of-origin allelic expression in the mouse brain.* Science, 2010. **329**(5992): p. 682-5.

67. Didion, J.P., et al., *Discovery of novel variants in genotyping arrays improves genotype retention and reduces ascertainment bias.* BMC Genomics, 2012. **13**(1): p. 34.

68. Weber, M., et al., *Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells.* Nat Genet, 2005. **37**(8): p. 853-62.

69. Wutz, A., et al., *Non-imprinted Igf2r expression decreases growth and rescues the Tme mutation in mice.* Development, 2001. **128**(10): p. 1881-7.

70. Stoger, R., et al., *Maternal-specific methylation of the imprinted mouse Igf2r locus identifies the expressed locus as carrying the imprinting signal.* Cell, 1993. **73**(1): p. 61-71.

71. Seidl, C.I., S.H. Stricker, and D.P. Barlow, *The imprinted Air ncRNA is an atypical RNAPII transcript that evades splicing and escapes nuclear export.* EMBO J, 2006. **25**(15): p. 3565-75.

72. Schulz, R., et al., *Transcript- and tissue-specific imprinting of a tumour suppressor gene.* Hum Mol Genet, 2009. **18**(1): p. 118-27.

73. Kobayashi, H., et al., *Identification of the mouse paternally expressed imprinted gene Zdbf2 on chromosome 1 and its imprinted human homolog ZDBF2 on chromosome 2.* Genomics, 2009. **93**(5): p. 461-72.

74. Kelsey, G., et al., *Identification of imprinted loci by methylation-sensitive representational difference analysis: application to mouse distal chromosome 2.* Genomics, 1999. **62**(2): p. 129-38.

75. Guillemot, F., et al., *Genomic imprinting of Mash2, a mouse gene required for trophoblast development.* Nat Genet, 1995. **9**(3): p. 235-42.

76. Gregg, C., et al., *High-resolution analysis of parent-of-origin allelic expression in the mouse brain.* Science, 2010. **329**(5992): p. 643-8.

77. Bartolomei, M.S., S. Zemel, and S.M. Tilghman, *Parental imprinting of the mouse H19 gene.* Nature, 1991. **351**(6322): p. 153-5.

78. Bell, A.C. and G. Felsenfeld, *Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene.* Nature, 2000. **405**(6785): p. 482-5.

79. Fujita, N., et al., *Methylation-mediated transcriptional silencing in euchromatin by methyl-CpG binding protein MBD1 isoforms.* Mol Cell Biol, 1999. **19**(9): p. 6415-26.

80. Hendrich, B. and A. Bird, *Identification and characterization of a family of mammalian methyl-CpG binding proteins.* Mol Cell Biol, 1998. **18**(11): p. 6538-47.

81. Collaborative-Cross-Consortium, *The genome architecture of the Collaborative Cross mouse genetic reference population.* Genetics, 2012. **190**(2): p. 389-401.

82. Zvetkova, I., et al., *Global hypomethylation of the genome in XX embryonic stem cells.* Nat Genet, 2005. **37**(11): p. 1274-9.

83. Nohara, K., et al., *Global DNA methylation in the mouse liver is affected by methyl deficiency and arsenic in a sex-dependent manner.* Arch Toxicol, 2011. **85**(6): p. 653-61.

84. Durcova-Hills, G., et al., *Influence of sex chromosome constitution on the genomic imprinting of germ cells.* Proc Natl Acad Sci U S A, 2006. **103**(30): p. 11184-8.

85. Wutz, A., *Gene silencing in X-chromosome inactivation: advances in understanding facultative heterochromatin formation.* Nat Rev Genet, 2011. **12**(8): p. 542-53.

86. Kraus, P., et al., *Mouse strain specific gene expression differences for illumina microarray expression profiling in embryos.* BMC Res Notes, 2012. **5**: p. 232.

87. Deaton, A.M. and A. Bird, *CpG islands and the regulation of transcription.* Genes Dev, 2011. **25**(10): p. 1010-22.

88. Ramirez-Carrozzi, V.R., et al., *A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling.* Cell, 2009. **138**(1): p. 114-28.

89. Cui, X., et al., *Improved statistical tests for differential gene expression by shrinking variance components estimates.* Biostatistics, 2005. **6**(1): p. 59-75.

90. Doi, A., et al., *Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts.* Nat Genet, 2009. **41**(12): p. 1350-3.

91. de La Casa-Esperon, E., et al., *X chromosome effect on maternal recombination and meiotic drive in the mouse.* Genetics, 2002. **161**(4): p. 1651-9.

92. Latham, K.E., E. De la Casa, and R.M. Schultz, *Analysis of mRNA expression during preimplantation development.* Methods Mol Biol, 2000. **136**: p. 315-31.

93. Bartolomei, M.S. and A.C. Ferguson-Smith, *Mammalian genomic imprinting.* Cold Spring Harb Perspect Biol, 2011. **3**(7).

94. Fowden, A.L., et al., *Imprinted genes and the epigenetic regulation of placental phenotype.* Prog Biophys Mol Biol, 2011. **106**(1): p. 281-8.

95. Frost, J.M. and G.E. Moore, *The importance of imprinting in the human placenta.* PLoS Genet, 2010. **6**(7): p. e1001015.

96. Kremerskothen, J., et al., *Brain-specific splicing of alpha-actinin 1 (ACTN1) mRNA.* Biochem Biophys Res Commun, 2002. **295**(3): p. 678-81.

97. Benstead, K. and J.V. Moore, *Quantitative histological changes in murine tail skin following photodynamic therapy.* Br J Cancer, 1989. **59**(4): p. 503-9.

98. Feil, R., et al., *Developmental control of allelic methylation in the imprinted mouse Igf2 and H19 genes.* Development, 1994. **120**(10): p. 2933-43.

99. McMinn, J., et al., *Imprinting of PEG1/MEST isoform 2 in human placenta.* Placenta, 2006. **27**(2-3): p. 119-26.

100. Weber, M., et al., *Extensive tissue-specific variation of allelic methylation in the Igf2 gene during mouse fetal development: relation to expression and imprinting.* Mech Dev, 2001. **101**(1-2): p. 133-41.

101. Woodfine, K., J.E. Huddleston, and A. Murrell, *Quantitative analysis of DNA methylation at all human imprinted regions reveals preservation of epigenetic stability in adult somatic tissue.* Epigenetics Chromatin, 2011. **4**(1): p. 1.

102. Choufani, S., et al., *A novel approach identifies new differentially methylated regions (DMRs) associated with imprinted genes.* Genome Res, 2011. **21**(3): p. 465-76.

103. Gertz, J., et al., *Analysis of DNA methylation in a three-generation family reveals widespread genetic influence on epigenetic regulation.* PLoS Genet, 2011. **7**(8): p. e1002228.

104. Schalkwyk, L.C., et al., *Allelic skewing of DNA methylation is widespread across the genome.* Am J Hum Genet, 2010. **86**(2): p. 196-212.

105. Xie, W., et al., *Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome.* Cell, 2012. **148**(4): p. 816-31.

106. Dockery, L., et al., *Differential methylation persists at the mouse Rasgrf1 DMR in tissues displaying monoallelic and biallelic expression.* Epigenetics, 2009. **4**(4): p. 241-7.

107. Das, R., D.D. Hampton, and R.L. Jirtle, *Imprinting evolution and human health.* Mamm Genome, 2009. **20**(9-10): p. 563-72.

108. Riesewijk, A.M., et al., *Maternal-specific methylation of the human IGF2R gene is not accompanied by allele-specific transcription.* Genomics, 1996. **31**(2): p. 158-66.

109. Smrzka, O.W., et al., *Conservation of a maternal-specific methylation signal at the human IGF2R locus.* Hum Mol Genet, 1995. **4**(10): p. 1945-52.

110. Weidman, J.R., et al., *Imprinting of opossum Igf2r in the absence of differential methylation and air.* Epigenetics, 2006. **1**(1): p. 49-54.

111. de la Casa-Esperon, E. and C. Sapienza, *Natural selection and the evolution of genome imprinting.* Annu Rev Genet, 2003. **37**: p. 349-70.

112. Pardo-Manuel de Villena, F., E. de la Casa-Esperon, and C. Sapienza, *Natural selection and the function of genome imprinting: beyond the silenced minority.* Trends Genet, 2000. **16**(12): p. 573-9.

113. Nomura, S., et al., *Clonal analysis of isolated single fundic and pyloric gland of stomach using X-linked polymorphism.* Biochem Biophys Res Commun, 1996. **226**(2): p. 385-90.

114. Abramoff, M.D., P.J. Magelhaes, and S.J. Ram, *Image Processing with ImageJ.* Biophotonics International, 2004. **11**(7): p. 36-42.

115. Kim K, T.S., Howard B, Bell T, Doherty H, Ideraabdullah F, Detwiler D, Pardo-Manuel de Villena F, *Meiotic drive at the Om locus in wild-derived inbred mouse strain.* Biological Journal of the Linnean Society, 2005. **84**: p. 487-492.

116. Bell, T.A., et al., *The paternal gene of the DDK syndrome maps to the Schlafen gene cluster on mouse chromosome 11.* Genetics, 2006. **172**(1): p. 411-23.

117. Charlesworth, D., B. Charlesworth, and G. Marais, *Steps in the evolution of heteromorphic sex chromosomes.* Heredity, 2005. **95**(2): p. 118-28.

118. Ohno, S., *Sex Chromosomes and Sex-linked genes*, ed. T.M. A. Labhart, L. T. Samuels, J. Zander. 1967, Berlin: Springer-Verlag.

119. Mak, W., et al., *Reactivation of the paternal X chromosome in early mouse embryos.* Science, 2004. **303**(5658): p. 666-9.

120. Takagi, N., O. Sugawara, and M. Sasaki, *Regional and temporal changes in the pattern of X-chromosome replication during the early post-implantation development of the female mouse.* Chromosoma, 1982. **85**(2): p. 275-86.

121. Ross, A.L., et al., *Genomic instability in cultured stem cells: associated risks and underlying mechanisms.* Regen Med, 2011. **6**(5): p. 653-62.

122. Rastan, S., *Non-random X-chromosome inactivation in mouse X-autosome translocation embryos--location of the inactivation centre.* J Embryol Exp Morphol, 1983. **78**: p. 1-22.

123. Penny, G.D., et al., *Requirement for Xist in X chromosome inactivation.* Nature, 1996. **379**(6561): p. 131-7.

124. Brown, C.J., et al., *A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome.* Nature, 1991. **349**(6304): p. 38-44.

125. Brown, C.J. and H.F. Willard, *The human X-inactivation centre is not required for maintenance of X-chromosome inactivation.* Nature, 1994. **368**(6467): p. 154-6.

126. Huynh, K.D. and J.T. Lee, *Inheritance of a pre-inactivated paternal X chromosome in early mouse embryos.* Nature, 2003. **426**(6968): p. 857-62.

127. Gartler, S.M. and A.D. Riggs, *Mammalian X-chromosome inactivation.* Annu Rev Genet, 1983. **17**: p. 155-90.

128. Augui, S., et al., *Sensing X chromosome pairs before X inactivation via a novel X-pairing region of the Xic.* Science, 2007. **318**(5856): p. 1632-6.

129. Monkhorst, K., et al., *X inactivation counting and choice is a stochastic process: evidence for involvement of an X-linked activator.* Cell, 2008. **132**(3): p. 410-21.

130. Starmer, J. and T. Magnuson, *A new model for random X chromosome inactivation.* Development, 2009. **136**(1): p. 1-10.

131. Bomsel-Helmreich, *Fate of heteroploid embryos.* Advanced Biosciences, 1971. **6**: p. 381-403.

132. Carr, D.H., *Chromosome studies in selected spontaneous abortions. Polyploidy in man.* J Med Genet, 1971. **8**(2): p. 164-74.

133. Jerome H, C.L., Bomsel-Helmreich O, *Enzymatic activities of triploid and diploid rabbit embyro cells in vitro.* Proc VIe Int Cong Reprod et Insem Artific, 1968: p. 143.

134. Shin, J., et al., *Maternal Rnf12/RLIM is required for imprinted X-chromosome inactivation in mice.* Nature, 2010. **467**(7318): p. 977-81.

135. Tada, T., et al., *Imprint switching for non-random X-chromosome inactivation during mouse oocyte growth.* Development, 2000. **127**(14): p. 3101-5.

136. Takagi, N., *Differentiation of X chromosomes in early female mouse embryos.* Exp Cell Res, 1974. **86**(1): p. 127-35.

137. Monk, M. and H. Kathuria, *Dosage compensation for an X-linked gene in pre-implantation mouse embryos.* Nature, 1977. **270**(5638): p. 599-601.

138. Monk, M., *Biochemical studies on X-chromosome activity in preimplantation mouse embryos.* Basic Life Sci, 1978. **12**: p. 239-46.

139. Epstein, C.J., et al., *Both X chromosomes function before visible X-chromosome inactivation in female mouse embryos.* Nature, 1978. **274**(5670): p. 500-3.

140. Jeppesen, P. and B.M. Turner, *The inactive X chromosome in female mammals is distinguished by a lack of histone H4 acetylation, a cytogenetic marker for gene expression.* Cell, 1993. **74**(2): p. 281-9.

141. Heard, E., et al., *Methylation of histone H3 at Lys-9 is an early mark on the X chromosome during X inactivation.* Cell, 2001. **107**(6): p. 727-38.

142. Gilbert, S.L. and P.A. Sharp, *Promoter-specific hypoacetylation of X-inactivated genes.* Proc Natl Acad Sci U S A, 1999. **96**(24): p. 13825-30.

143. Brockdorff, N., et al., *Conservation of position and exclusive expression of mouse Xist from the inactive X chromosome.* Nature, 1991. **351**(6324): p. 329-31.

144. Krietsch, W.K., et al., *The expression of X-linked phosphoglycerate kinase in the early mouse embryo.* Differentiation, 1982. **23**(2): p. 141-4.

145. Cattanach, B.M. and J.H. Isaacson, *Controlling elements in the mouse X chromosome.* Genetics, 1967. **57**(2): p. 331-46.

146. Cattanach, B.M. and J.H. Isaacson, *Genetic control over the inactivation of autosomal genes attached to the X-chromosome.* Z Vererbungsl, 1965. **96**(4): p. 313-23.

147. Cattanach, B.M., *Control of chromosome inactivation.* Annu Rev Genet, 1975. **9**: p. 1-18.

148. Fowlis, D.J., J.D. Ansell, and H.S. Micklem, *Further evidence for the importance of parental source of the Xce allele in X chromosome inactivation.* Genet Res, 1991. **58**(1): p. 63-5.

149. Forrester, L.M. and J.D. Ansell, *Parental influences on X chromosome expression.* Genet Res, 1985. **45**(1): p. 95-100.

150. Cattanach, B.M.R.C., *Identification of the Mus castaneus Xce allele.* Mouse Genome, 1994(92): p. 114.

151. Cattanach, B.M. and C.E. Williams, *Evidence of non-random X chromosome activity in the mouse.* Genet Res, 1972. **19**(3): p. 229-40.

152. Cattanach, B.M., C.E. Pollard, and J.N. Perez, *Controlling elements in the mouse X-chromosome. I. Interaction with the X-linked genes.* Genet Res, 1969. **14**(3): p. 223-35.

153. Cattanach, B.M., *Identification of the Mus spretus Xce allele.* Mouse Genome, 1991. **89**: p. 565-566.

154. Krietsch, W.K., et al., *Expression of X-linked phosphoglycerate kinase in early mouse embryos homozygous at the Xce locus.* Differentiation, 1986. **31**(1): p. 50-4.

155. Chadwick, L.H., et al., *Genetic control of X chromosome inactivation in mice: definition of the Xce candidate interval.* Genetics, 2006. **173**(4): p. 2103-10.

156. Ogawa, Y. and J.T. Lee, *Xite, X-inactivation intergenic transcription elements that regulate the probability of choice.* Mol Cell, 2003. **11**(3): p. 731-43.

157. Russell, L.B., *Mammalian X-chromosome action: inactivation limited in spread and region of origin.* Science, 1963. **140**(3570): p. 976-8.

158. Monkhorst, K., et al., *The probability to initiate X chromosome inactivation is determined by the X to autosomal ratio and X chromosome specific allelic properties.* PLoS One, 2009. **4**(5): p. e5616.

159. Brown, S.W. and H.S. Chandra, *Inactivation system of the mammalian X chromosome.* Proc Natl Acad Sci U S A, 1973. **70**(1): p. 195-9.

160. Percec, I., et al., *An N-ethyl-N-nitrosourea mutagenesis screen for epigenetic mutations in the mouse.* Genetics, 2003. **164**(4): p. 1481-94.

161. Percec, I., et al., *Autosomal dominant mutations affecting X inactivation choice in the mouse.* Science, 2002. **296**(5570): p. 1136-9.

162. Plenge, R.M., et al., *Expression-based assay of an X-linked gene to examine effects of the X-controlling element (Xce) locus.* Mamm Genome, 2000. **11**(5): p. 405-8.

163. Chadwick, L.H. and H.F. Willard, *Genetic and parent-of-origin influences on X chromosome choice in Xce heterozygous mice.* Mamm Genome, 2005. **16**(9): p. 691-9.

164. Bittner, R.E., et al., *Dystrophin expression in heterozygous mdx/+ mice indicates imprinting of X chromosome inactivation by parent-of-origin-, tissue-, strain- and position-dependent factors.* Anat Embryol (Berl), 1997. **195**(2): p. 175-82.

165. Thorvaldsen, J.L., et al., *Nonrandom x chromosome inactivation is influenced by multiple regions on the murine x chromosome.* Genetics, 2012. **192**(3): p. 1095-107.

166. Vickers, M.A., et al., *Assessment of mechanism of acquired skewed X inactivation by analysis of twins.* Blood, 2001. **97**(5): p. 1274-81.

167. Puck, J.M., R.L. Nussbaum, and M.E. Conley, *Carrier detection in X-linked severe combined immunodeficiency based on patterns of X chromosome inactivation.* J Clin Invest, 1987. **79**(5): p. 1395-400.

168. Kristiansen, M., et al., *Twin study of genetic and aging effects on X chromosome inactivation.* Eur J Hum Genet, 2005. **13**(5): p. 599-606.

169. Ideraabdullah, F.Y., et al., *Genetic and haplotype diversity among wild-derived mouse inbred strains.* Genome Res, 2004. **14**(10A): p. 1880-7.

170. Cattanach, B.M. and D. Papworth, *Controlling elements in the mouse. V. Linkage tests with X-linked genes.* Genet Res, 1981. **38**(1): p. 57-70.

171. Lenarcic, A.B., et al., *Restricted-Contour Bayesian Inference of X-inactivation Ratios From Allele-Specific Expression.* Annals of Applied Statistics, 2013.

172. Waterston, R.H., et al., *Initial sequencing and comparative analysis of the mouse genome.* Nature, 2002. **420**(6915): p. 520-62.

173. Smit, A., R. Hubley, and P. Green. *RepeatMasker Open-3.0*. 2006; Available from: http:www.repeatmasker.org.

174. Peiffer, D.A., et al., *High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping.* Genome Res, 2006. **16**(9): p. 1136-48.

175. Assie, G., et al., *SNP arrays in heterogeneous tissue: highly accurate collection of both germline and somatic genetic information from unpaired single tumor samples.* Am J Hum Genet, 2008. **82**(4): p. 903-15.

176. Fu, C.-P., et al., *inferring ancestry in admixed populations using microarray probe intensities.* Proceedings of the ACM Conference on Bioinformatics, 2012: p. 105-112.

177. Kaelin, C. and G. Barsh, *Tabby pattern genetics - a whole new breed of cat.* Pigment Cell Melanoma Res, 2010. **23**(4): p. 514-6.

178. Eizirik, E., et al., *Defining and mapping mammalian coat pattern genes: multiple genomic regions implicated in domestic cat stripes and spots.* Genetics, 2010. **184**(1): p. 267-75.

179. Simmler, M.C., et al., *Mapping the murine Xce locus with (CA)n repeats.* Mamm Genome, 1993. **4**(9): p. 523-30.

180. Cattanach, B.M.R.C., *Identification of the Mus spretus Xce allele.* Mouse Genome, 1991. **89**: p. 565-566.

181. Cattanach, B.M., et al., *Genetic and molecular evidence of an X-chromosome deletion spanning the tabby (Ta) and testicular feminization (Tfm) loci in the mouse.* Cytogenet Cell Genet, 1991. **56**(3-4): p. 137-43.

182. Nesbitt, M., *X chromosome inactivation mosaicism in the mouse.* Developmental Biology, 1971. **26**: p. 252-263.

183. Johnston, P.G. and B.M. Cattanach, *Controlling elements in the mouse. IV. Evidence of non-random X-inactivation.* Genet Res, 1981. **37**(2): p. 151-60.

184. Wang, X., et al., *Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain.* PLoS One, 2008. **3**(12): p. e3839.

185. Clapcote, S.J. and J.C. Roder, *Simplex PCR assay for sex determination in mice.* Biotechniques, 2005. **38**(5): p. 702, 704, 706.

186. Taylor, B.A. and P.C. Reifsnyder, *Typing recombinant inbred mouse strains for microsatellite markers.* Mamm Genome, 1993. **4**(5): p. 239-42.

187. Ronaghi, M., M. Uhlen, and P. Nyren, *A sequencing method based on real-time pyrophosphate.* Science, 1998. **281**(5375): p. 363, 365.

188. Church, D.M., et al., *Lineage-specific biology revealed by a finished genome assembly of the mouse.* PLoS Biol, 2009. **7**(5): p. e1000112.

189. Congdon, P., *Applied Bayesian hierarchical methods*. Vol. xiii. 2010, London: Chapman and Hall. 590.

190. Parmigiani, G. and L. Inoue, *Decision Theory: Prinicples and Approaches*. 2009: Wiley. 372.

191. Felsenstein, *PHYLIP-Phylogeny Inference Package (Version 3.2).* Cladistics, 1989. **5**: p. 164-166.

192. Jia, D., et al., *Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation.* Nature, 2007. **449**(7159): p. 248-51.

193. Williamson, C.M., Blake A, Thomas S, Beechey CV, Hancock J, Cattanach BM, and Peters J *World Wide Web Site - Mouse Imprinting Data and References.* 2013.

194. Murrell, A., et al., *An association between variants in the IGF2 gene and Beckwith-Wiedemann syndrome: interaction between genotype and epigenotype.* Hum Mol Genet, 2004. **13**(2): p. 247-55.

195. Flanagan, J.M., et al., *Intra- and interindividual epigenetic variation in human germ cells.* Am J Hum Genet, 2006. **79**(1): p. 67-84.

196. Wang, H., et al., *Widespread plasticity in CTCF occupancy linked to DNA methylation.* Genome Res, 2012. **22**(9): p. 1680-8.

197. Waddington, C., *Canalization of development and the inheritance of acquired characters.* Nature, 1942. **150**(3811): p. 563-565.

198. Cattanach, B.M., *A chemically-induced variegated-type position effect in the mouse.* Z Vererbungsl, 1961. **92**: p. 165-82.

199. Al Nadaf, S., et al., *Activity map of the tammar X chromosome shows that marsupial X inactivation is incomplete and escape is stochastic.* Genome Biol, 2010. **11**(12): p. R122.

200. Lyon, M.F., *Possible mechanisms of X chromosome inactivation.* Nat New Biol, 1971. **232**(34): p. 229-32.

201. Martin, J.P. and J. Bell, *A Pedigree of Mental Defect Showing Sex-Linkage.* J Neurol Psychiatry, 1943. **6**(3-4): p. 154-7.

202. Turner, H.H., *A Syndrome of Infantilism, Congenital Webbed Neck, and Cubitus Valgus.* Endocrinology, 1938. **23**: p. 566-574.

203. Redonnet-Vernhet, I., et al., *Uneven X inactivation in a female monozygotic twin pair with Fabry disease and discordant expression of a novel mutation in the alpha-galactosidase A gene.* J Med Genet, 1996. **33**(8): p. 682-8.

204. Calaway, J.D., et al., *Genetic architecture of skewed x inactivation in the laboratory mouse.* PLoS Genet, 2013. **9**(10): p. e1003853.

205. Rossant, J. and P.P. Tam, *Blastocyst lineage formation, early embryonic asymmetries and axis patterning in the mouse.* Development, 2009. **136**(5): p. 701-13.

206. Zernicka-Goetz, M., S.A. Morris, and A.W. Bruce, *Making a firm decision: multifaceted regulation of cell fate in the early mouse embryo.* Nat Rev Genet, 2009. **10**(7): p. 467-77.

207. Powers, J.G., et al., *Efficient and accurate whole genome assembly and methylome profiling of E. coli.* BMC Genomics, 2013. **14**(1): p. 675.