AN INVESTIGATION OF THE GENETIC AND EPIGENETIC FACTORS UNDERLYING ESCAPE FROM X-CHROMOSOME INACTIVATION

by

Bradley Balaton

B.Sc., The University of Saskatchewan, 2014

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Medical Genetics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

May 2021

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

An investigation of the genetic and epigenetic factors underlying escape from X-chromosome inactivation

submitted by Bradley Balaton in partial fulfillment of the requirements for

the degree of Doctor of Philosophy

in Medical Genetics

Examining Committee:

Carolyn Brown, Professor, Medical Genetics, UBC
Supervisor

Louis Lefebvre, Associate Professor, Medical Genetics, UBC
Supervisory Committee Member

Elizabeth Rideout, Assistant Professor, Cellular & Physiological Sciences, UBC
University Examiner

Pamela Hoodless, Professor, Medical Genetics, UBC
University Examiner

Additional Supervisory Committee Members:

Wendy Robinson, Professor, Medical Genetics, UBC
Supervisory Committee Member

Wyeth Wasserman, Professor, Medical Genetics, UBC
Supervisory Committee Member

Abstract

X-chromosome inactivation (XCI) is the process by which one of the X chromosomes in XX females is silenced to express similar levels of X-linked genes with XY males. This silencing is incomplete as some genes escape from XCI and other genes vary their XCI status across populations, tissues or samples. Here I derive consensus XCI status calls in humans, extend XCI status calls across species, and determine the relationship between XCI status and various epigenetic marks.

I aggregated XCI status calls from multiple studies, deriving XCI status calls for 639 human genes. I found 12% of genes escaping from XCI, 8% variably escaping XCI, and 7% discordant across studies. To make XCI status calls across species I obtained DNA methylation data for 12 species, allowing us to generate an average of 387 XCI calls per species. Overall, 12% of genes escaped XCI, with mouse an outlier with only 5%. Of the genes with predictions across at least four species, 74.8% of them were entirely consistent and only 6% had more than one inconsistent species. Many genes were seen to have primate-specific escape from XCI, while only one gene had an artiodactyla-specific XCI status. The consensus XCI status calls were compared to DNA methylation and commonly analyzed histone marks. I found the expected trend where repressive marks were enriched at genes subject to XCI and activating marks were enriched at genes escaping XCI; however, the histone marks had a large overlap between levels seen at genes subject to XCI and genes escaping from XCI. Only DNAme could accurately predict an individual gene’s XCI status. I combined the marks and found that we could make XCI status calls with 75% accuracy for genes escaping from XCI and 90% accuracy for genes subject to

XCI. The marks with the greatest contribution to this predictor were DNAme, H3K27me3 and H3K4me3.

The results of these projects further our understanding of which genes escape from XCI, which may be important for analysis of sexual dimorphism and further provide us a means to examine how silencing may be regulated in humans and across mammals.

Lay Summary

X-chromosome inactivation is the inactivation of one of the two X chromosomes in XX females so that they have the same dosage of X-linked genes as XY males. Not all genes on that X are inactivated however, approximately 12% of genes escape X-chromosome inactivation and 15% vary whether they inactivate between populations, tissues, or individuals. I assembled data from multiple studies, increasing confidence in which genes escape or variably escape from X-chromosome inactivation in humans. Using DNA methylation, which is strongly correlated with this inactivation, I determined which genes are escaping X-chromosome inactivation across 12 mammalian species and observed multiple features associated with X-chromosome inactivation across species. I determined how other expression regulating modifications are related to escape from X-chromosome inactivation and can be used to predict whether novel genes are escaping or subject to X-chromosome inactivation. Understanding these escape genes is important for understanding sex-differences and X-linked diseases.

Preface

Much of the content in chapter one has featured in reviews that I have helped write (Balaton et al. 2018; Balaton and Brown 2016; Navarro-Cobos, Balaton, and Brown 2020). Chapters 2, 3 and 4 are based on work that has been published previously or are currently submitted for publication in the following publication: Chapter 2 is Balaton, B., Cotton, AM., and Brown, CJ. (2015) Derivation of consensus inactivation status for X-linked genes from genome-wide studies. Biol. Sex. Differ. DOI: 10.1186/s13293-015-0053-7. Original concept and tabulations were done by AM Cotton, but the rest of the analysis and writing the manuscript were done by me. Chapter 3 is Balaton, BP., Fornes, O., Wasserman, WW., and Brown, CJ. (2021) Cross-species examination of X-chromosome inactivation highlights domains of escape from silencing. Epigenetics and Chromatin. DOI: 10.1101/2020.12.04.412197. I wrote the manuscript and performed all analyses but the CTCF predictions. All authors contributed conceptually. Chapter 4 is Balaton, BP., and Brown CJ. (2021) Contribution of genetic and epigenetic changes to escape from X-chromosome inactivation. bioRxiv, DOI: 10.1101/2021.03.03.433635. All analysis and the writing of the manuscript were done by me. All authors contributed conceptually.

Access to the raw human genomic data used in chapters 3 and 4 required ethics board approval. This was approved by the UBC Clinical Research Ethics Board. The certificate number for access to the CEMT data used in both chapters is: H17-01363. The certificate number for access to the TCGA data used in chapter 4 is: H19-02018.

Table of Contents

Abstract .... **iii**

Lay Summary .... **v**

Preface .... **vi**

Table of Contents .... **vii**

List of Tables .... **xii**

List of Figures .... **xiii**

List of Abbreviations .... **xiv**

Acknowledgements .... **xvii**

Chapter 1: Introduction **1**

1.1 Thesis overview 1

1.2 X-chromosome inactivation 2

1.3 Escape from XCI 4

1.4 The interaction of epigenetics and XCI status 8

1.4.1 DNAme 8

1.4.2 Heterochromatic histone marks 10

1.4.3 Euchromatic histone marks 11

1.4.4 Other marks 11

1.5 Determining XCI status experimentally 11

1.5.1 Mouse-human hybrids 13

1.5.2 Xi/Xa expression 13

1.5.3 Sex-specific Expression 15

1.5.4 Sex-specific DNAme 15

1.5.5 Other methods 16

1.6 Theoretical determinants of XCI status 17

1.6.1 Waystations 17

1.6.2 Escape elements 18

1.6.3 Boundaries 19

1.7 Thesis objectives 19

Chapter 2: Derivation of consensus inactivation status for X-linked genes from genome-wide studies 21

2.1 Introduction 21

2.2 Methods 24

2.2.1 Categorization of X-linked genes 24

2.2.2 Expression analysis 26

2.2.3 Domain analysis 26

2.3 Results and discussion 27

2.3.1 Creation of a consensus XCI status 27

2.3.2 Discordancies between studies 30

2.3.3 Domains of escape and boundaries 34

2.3.4 Comparison to additional studies examining XCI 34

2.3.5 XCI status of genes with Y chromosome homology 37

2.3.6 Our consensus XCI status calls and sex differences in expression 38

2.4 Conclusions 38

Chapter 3: Cross-species examination of X-chromosome inactivation highlights domains of escape from silencing 40

3.1 Introduction... 40

3.2 Methods... 42

3.2.1 Xi/Xa expression-based XCI status calls.. 42

3.2.2 DNAme based XCI status calls.. 43

3.2.3 Clustering.. 44

3.2.4 Conservation analysis 45

3.2.5 Statistical tests.. 45

3.2.6 Domain analysis... 45

3.2.7 ATAC-seq analysis 46

3.2.8 CTCF predictions... 46

3.3 Results... 46

3.3.1 XCI status calls from allelic expression... 46

3.3.2 Establishing thresholds for calling XCI status from DNAme 49

3.3.3 XCI status calls from DNAme.. 51

3.3.4 Conservation of XCI status calls across species.................................. 54

3.3.5 Role for alternative promoter usage in escape from XCI 56

3.3.6 Domains of escape from XCI across species.. 58

3.3.7 Correlation of features with XCI status across species......................... 61

3.4 Discussion.. 64

3.5 Conclusions.. 71

Chapter 4: Contribution of genetic and epigenetic changes to escape from X-chromosome inactivation 73

4.1 Introduction 73

4.2 Methods 74

4.2.1 Previous XCI status calls 74

4.2.2 Histone ChIP-seq analysis 74

4.2.3 Expression analysis 75

4.2.4 DNAme analysis 75

4.2.5 Histone-based XCI status predictions 77

4.2.6 Statistical comparisons 78

4.3 Results 78

4.3.1 Histone marks differ with sex and XCI status 79

4.3.2 Using Xi/Xa expression ratio to identify XCI status 84

4.3.3 Using DNA methylation to identify XCI status 88

4.3.4 An epigenetic model to predict XCI status across samples 93

4.3.5 Independent regulation of variable escape across a region 100

4.4 Discussions 102

4.5 Conclusions 106

Chapter 5: Discussion 108

5.1 Thesis Findings 108

5.2 Limitations 111

5.3 Outstanding questions 113

5.4 Future directions 113

5.5 Conclusion 116

References..118

Appendices..137

Appendix A Supplementary materials for Chapter 2.. 137

A.1 Supplementary Figures 137

A.2 Supplementary Tables... 140

Appendix B Supplementary materials for Chapter 3.. 141

B.1 Supplementary Figures 141

B.2 Supplementary Tables... 151

Appendix C Supplementary materials for Chapter 4.. 158

C.1 Supplementary Figures 158

C.2 Supplementary Tables... 170

List of Tables

Table 1.1 Epigenetic marks and their relation to genes which are subject to or escaping from XCI. 9

Table 2.1 Sample sizes of previous studies. 21

Table 2.2 Most studies show a trend with what they are calling discordantly. 30

Table 3.1 Mean DNAme for genes subject to XCI per dataset. The mean DNAme of CpG islands at genes found subject to XCI was calculated per dataset. 50

Table 4.1 Significance of differences in epigenetic marks between samples with opposite XCI statuses at genes found variably escaping XCI by Xi/Xa expression. 86

Table 4.2 The percentage of variably escaping genes found by DNAme that have significant differences in epigenetic marks (BH corrected p-value<0.01). 90

List of Figures

Figure 1.1 Sex differences in expression for genes in the PARs or that escape from XCI. 7

Figure 1.2 Approaches to identifying genes that escape from inactivation. 12

Figure 1.3 Theoretical determinants of XCI status. 18

Figure 2.1 The majority of X-linked protein coding genes have an XCI status call. 27

Figure 2.2 Consensus XCI status calls. 29

Figure 2.3 Comparison of discordancies. 31

Figure 2.4 Domains of XCI and the enrichment of discordant and mostly variable escape genes at boundaries. 36

Figure 3.1 Using Xi/Xa expression ratio to establish thresholds of DNAme for XCI status calls. 47

Figure 3.2 The number and type of XCI status calls per species. 52

Figure 3.3 Concordant and discordant escape genes across species. 55

Figure 3.4 Featured genes compared across species. 58

Figure 3.5 DNAme across the variably escaping gene UBA1 60

Figure 3.6 Enrichment of elements which may be related to XCI status. 62

Figure 4.1 The Xi has more heterochromatic and less euchromatic marks than the Xa. 81

Figure 4.2 Epigenetic marks do not change consistently with XCI status for variably escaping genes. 85

Figure 4.3 DNAme varies at genes variably escaping from XCI. 89

Figure 4.4 XCI status predictions with an epigenetic model expands the number of genes examineable. 95

Figure 4.5 XCI status calls are independent between neighboring variably escaping genes. 101

List of Abbreviations

450k	Illumina Infinium Human Methylation450 BeadChip array
AI	allelic imbalance
ATAC-seq	Assay for Transposase Accessible Chromatin with sequencing
BAC	Bacterial Artificial Chromosome
CEMT	Center for Epigenome Mapping Technologies
ChIP-seq	Chromatin Immuno-precipitation with sequencing
CpG	cytosine followed by guanine
CREST	Core Research for Evolutional Science and Technology
CTAG	cancer-testes antigen gene
DNAme	DNA methylation at CpG dinucleotides
GTEx	Genotype-Tissue Expression project
IHEC	International Human Epigenome Consortium
lncRNA	long non-coding RNA
mCH	non-CpG methylation
PAR	pseudoautosomal region
RRBS	reduced representation bisulfite sequencing
scRNA-seq	single cell RNA sequencing
TAD	topologically associated domain
TSS	transcription start site
UCSC	University of California Santa Cruz Genome Browser
WGBS	whole genome bisulfite sequencing
X	X chromosome

Xa	active X chromosome
XCI	X-chromosome inactivation
Xi	inactive X chromosome
Y	Y chromosome

Gene Names

ARSD	Arylsulfatase D
BCOR	BCL6 Corepressor
CDK16	Cyclin Dependent Kinase 16
CDKL5	Cyclin Dependent Kinase Like 5
CTCF	CCCTC-binding Factor
CXorf38	Chromosome X Open Reading Frame 38
DDX3X	Dead-box Helicase 3 X-linked
EZH2	Enhancer of Zeste 2 Polycomb Repressive Complex 2 Subunit
EIF1AX	Eukaryotic Transcription Initiation Factor 1A X-linked
EIF2S3	Eukaryotic Transcription Initiation Factor 2 Subunit Gamma
GEMIN8	Gem Nuclear Organelle Associated Protein 8
HBG2	Hemoglobin Subunit Gamma 2
HPRT	Hypoxanthine Phosphoribosyltransferase 1
KDM5C	Lysine Demethylase 5C
KDM6A	Lysine Demethylase 6A
LOC389906	Zinc Finger Protein 839 Pseudogene
macroH2A	Histone macroH2A

MED14	Mediator Complex Subunit 14
PNPLA4	Patatin Like Phospholipase Domain Containing 4
PRKX	Protein Kinase X-Linked
RPS4X	Ribosomal Protein S4 X-Linked
SMC1A	Structural Maintenance of Chromosomes 1A
SMCHD1	Structural Maintenance of Chromosomes Flexible Hinge Domain Containing 1
STS	Steroid Sulfatase
TIMP1	TIMP Metallopeptidase Inhibitor 1
TRIM6	Tripartite Motif Containing 6
TSIX	XIST Antisense RNA
XIST	X Inactive Specific Transcript
YY1	Yin Yang 1
ZSCAN9	Zinc Finger and SCAN Domain Containing 9

Histone Marks

H3K4me1	mono-methylation of lysine 4 on histone H3
H3K4me2	di-methylation of lysine 4 on histone H3
H3K4me3	tri-methylation of lysine 4 on histone H3
H3K9me3	tri-methylation of lysine 9 on histone H3
H3K27ac	acetylation of lysine 27 on histone H3
H3K27me3	tri-methylation of lysine 27 on histone H3
H3K36me3	tri-methylation of lysine 36 on histone H3
H4K20me3	tri-methylation of lysine 20 on histone H4

Acknowledgements

I would like to thank Dr. Carolyn Brown for the years of help and guidance on this project. Thanks to all the members of her lab that I have overlapped with, for showing me techniques and giving countless amounts of advice. I would like to especially thank Dr. Samantha Peeters and Dr. Thomas Dixon-McDougall who overlapped as my fellow graduate students for the longest time and Dr. Allison Matthews from whom I inherited the project.

Thank you to my committee members Dr. Louis Lefebvre, Dr. Wendy Robinson and Dr. Wyeth Wasserman for their guidance, within the committee and outside of it. I would also like to thank the many members of the Wasserman lab who have attended our DNA elements meetings over the years, but especially my co-authors Dr. Julie Chen and Dr. Oriol Fornes.

Many thanks to the other members of the Molecular Genetics Group at UBC, but especially to those in the bioinformatics office: Aaron Bogutz, Dr. Julien Richard Albert, and Dr. Benjamin Martin. Without them the computational aspects of these projects would have been much more difficult or may not have happened at all. I would also like to thank Amanda Ha from the Lefebvre lab for being my closest friend in the wing.

Lastly, I would like to thank my family for supporting me through this, and especially for giving me a better place to work from during the 2020 pandemic.

Chapter 1: Introduction

1.1 Thesis overview

X-chromosome inactivation (XCI) is the epigenetic inactivation of one of the two X chromosomes in XX female mammals to have dosage compensation with males who have only one X chromosome (X). Some genes escape this inactivation and are expressed from both Xs in females. These genes that escape from XCI have implications for male-female differences in gene expression and disease susceptibility. XCI is a well-studied epigenetic phenomenon however the mechanisms of how and why genes escape from XCI are not fully understood. In addition to genes which constitutively escape from XCI, there are also genes which variably escape from XCI, which escape in some populations, tissues or individuals while being subject to XCI in others. These variably escaping genes may have implications for inter-individual differences in females, and also provide an opportunity to study genes escaping and subject to XCI in the same genomic context.

With the onset of the genomics era there has been an influx of datasets hosted publicly online for researchers to analyze. Due to difficulties in analyzing X data, with males having one active X (Xa) while females have an Xa and an inactive X (Xi), many investigators do not include the X in their analyses. The work in this thesis brings together a variety of sources and types of genomic and epigenomic data in order to determine which genes are escaping, variably escaping or subject to XCI, in humans and other mammals, and to determine how histone marks and DNA methylation at CpG dinucleotides (DNAme) interact with the genomic environment around genes with each XCI status, in females and males.

1.2 X-chromosome inactivation

In eutherian mammals, females generally have 2 Xs while males have an X and a Y chromosome (Y). While the X and Y are both descended from the same pair of ancestral autosomes, the Y has slowly diverged and lost many of the genes shared with the X (reviewed in (Posynick and Brown 2019)). Because XY males only have one copy remaining of most X-linked genes, XX females have evolved a form of dosage compensation known as XCI to epigenetically inactivate one of their two X chromosomes (Lyon 1961, 1962).

XCI is initiated by expression of the long non-coding RNA (lncRNA) *XIST*, which then recruits various heterochromatin factors to the soon to be Xi (reviewed in (M. Almeida, Bowness, and Brockdorff 2020; Dixon-McDougall and Brown 2015)). In mice there are two rounds of XCI. Early in development mice have imprinted XCI and inactivate their paternal X (Mak et al. 2004; Moreira de Mello et al. 2010; Okamoto et al. 2004). This is followed by reactivation of the paternal X everywhere but in placental trophoblasts, and then a second round of XCI, but with a random choice of which X to inactivate. In humans, inactivation occurs later in development but there is only one round of XCI and the choice of which X to inactivate is random (Okamoto et al. 2011). Random inactivation leads to females being a mosaic for which X is inactive, with a random sample of cells having roughly equal expression from both alleles, even though each of the cells is expressing only from one allele.

Some females have one X chromosome allele more commonly as the Xi than the other. This is referred to as skewed XCI. For the purposes of this thesis, I am interested in samples with >90%

skewing, so that we see <10% expression from one allele at genes subject to XCI. This level of skewing can occur naturally, being more common in blood (Vacca et al. 2016) and cancer (Larson et al. 2017). Cells that have become monoclonal during cell culture and those with a deleterious allele on one X have also been seen to have skewed XCI and used for studying XCI (Berletch et al. 2015; Carrel and Willard 2005). In mice, there are strains available with knock-outs in *Xist*, which are used to study XCI as they will be completely skewed to have the *Xist* knock-out on the Xa (Berletch et al. 2015); similar results are obtained by knocking out *Xist*'s agonist *Tsix*, with the deleted allele now always being the Xi (Luikenhuis, Wutz, and Jaenisch 2001). These strains are often outbred to a distant mouse strain so that the F1 generation has many polymorphisms differentiating the Xa from the Xi.

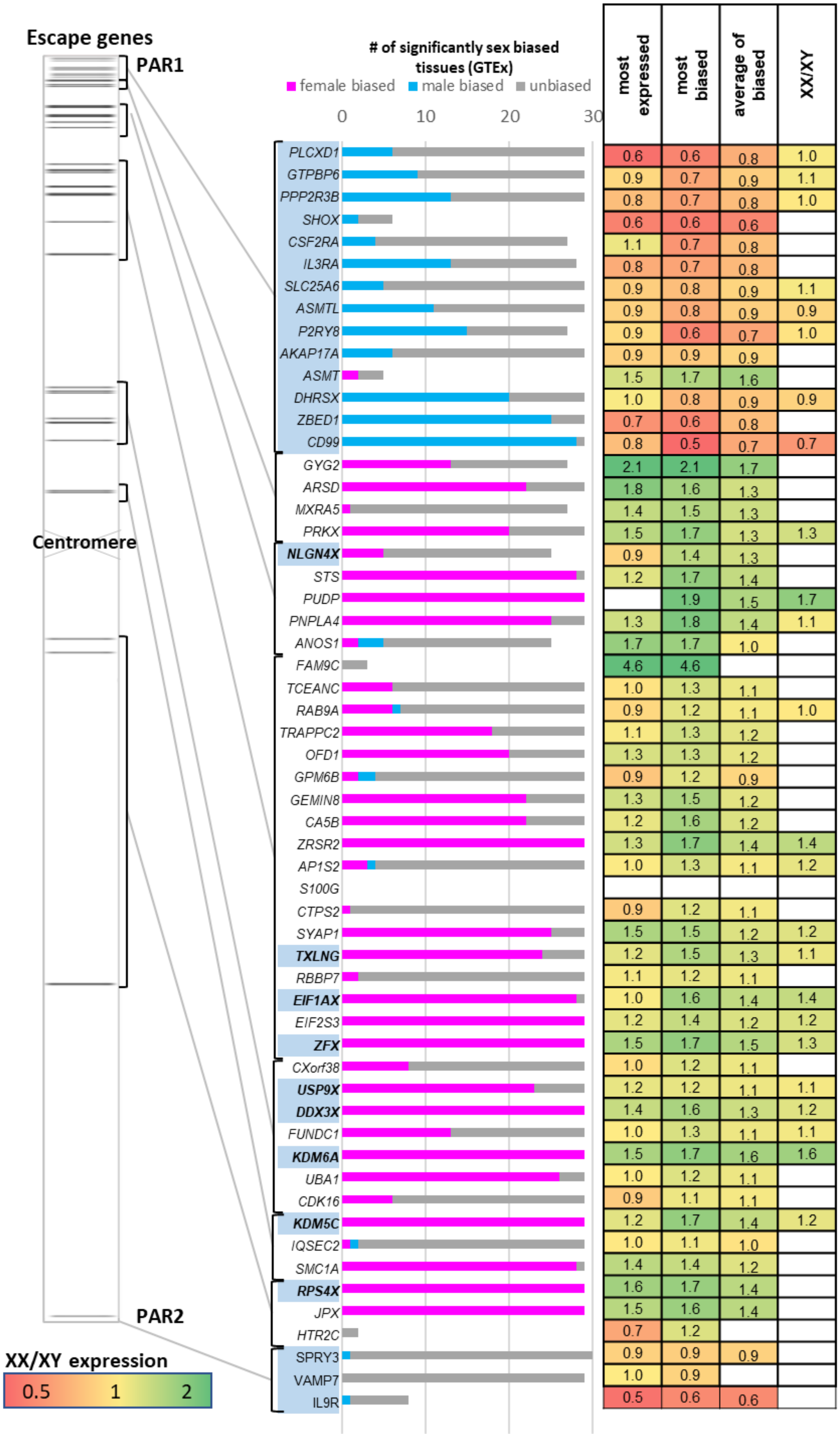
It is important to study XCI as there are many genes on the X chromosome linked to a variety of diseases that will be impacted by the silencing of one allele. For example, intellectual disability is linked to 141 X genes (Neri et al. 2018). X-linked diseases are seen more commonly in males, where one mutated allele will have an effect in males while females will still have half of their cells expressing the healthy allele. However, skewed XCI can expose the female to mutations on the more common Xa and some mutations are lethal at the cellular level and will cause skewed XCI (Mitterbauer et al. 1999; Naumova et al. 1998). There are also cases where the mosaic nature of XCI is uniquely relevant, having a negative phenotype when two neighboring cell populations are expressing opposite alleles on their Xa, but no phenotype if all the cells are expressing either or both alleles (Twigg et al. 2013).

1.3 Escape from XCI

Not all genes on the Xi are subject to XCI; in humans between 8% (Cotton et al. 2013) and 15% (Carrel and Willard 2005) of genes escape XCI with expression level from the Xi at least 10% of that from the Xa. A study in mice with a less stringent threshold reported even less escape, with only 3-7% of genes escaping from XCI (Berletch et al. 2015). Additionally, some genes vary in their XCI status between different tissues, populations or individuals and are called variably escaping from XCI. The number of variably escaping genes varies widely between studies, with one study finding up to 32% of genes variably escaping from XCI (13% variable in all populations and tissues, 9% tissue-specific and 10% population-specific) (Cotton et al. 2013). Another study found <1% of genes variably escaping overall and across tissues but found 29% of genes were variably escaping within 1-2 tissues while having a consistent XCI status in the remainder of the 27 tissues analyzed (Cotton et al. 2015). These differences may be due to the first study using expression to determine XCI status while the second used DNAme. In mice, tissue-specific escape genes have also been seen, with those escaping only in one tissue often having a tissue-specific function (Berletch et al. 2015). Chromosome-wide studies calling XCI status of genes have not been done previously for other eutherian mammals, so we do not know how well conserved escape from XCI is, or if human or mouse are outliers in this regard.

There is a pseudoautosomal region (PAR) at each end of the X and Y, each of which has retained homology and ability to recombine. In humans PAR1, located on the short arm of the X, is 2.7 MB and contains 24 genes, while PAR2, on the long arm of the X, is only 0.33 MB and contains only five genes (Flaquer et al. 2008). The genes in PAR1 are thought to all escape from XCI (Carrel and Willard 2005) as they are also on the Y and therefore do not need dosage compensation between males and females. In contrast, some of the genes in PAR2 are silent on both the Xi and the Y (De Bonis et al. 2006; Ciccodicola et al. 2000) to obtain dosage compensation. As the X has differentiated from the Y through a series of inversion events, there are strata which have differentiated for various lengths of time (reviewed in (Posynick and Brown 2019)). The younger strata which diverged from the Y more recently, are enriched for genes escaping XCI (Carrel and Willard 2005). Additionally, X genes which have retained a functional Y homolog are also enriched for escape from XCI (Bellott et al. 2014). Many of these X-Y homologs are conserved across species and are hypothesized to be more dosage-sensitive so that the loss of the allele on the Y and Xi would be detrimental to survival (Bellott et al. 2014). Having a single X is less severe in mice than humans (Lyon 1962), likely due to the reduced size of the PAR and fewer escape genes in mice compared to humans (Deng et al. 2014).

Dosage compensation is not 100% effective between XX females and XY males; genes which escape XCI in either PAR tend to have male-biased expression while genes which escape XCI outside of the PAR tend to have female-biased expression (Figure 1.1) (Navarro-Cobos, Balaton, and Brown 2020; Tukiainen et al. 2017). While these expression biases could be from hormones and their downstream effects, expression comparisons between sex aneuploidies show increased expression of genes which escape XCI with increased X count and increased expression of PAR genes with increased X or Y count, although the effects were not linear so other compensation is present (Raznahan et al. 2018). Additionally, those genes with X-Y homology had more significant increases in expression with both increasing X and Y copy number. To differentiate the effects of X and Y dosage from hormonal or other sex-related effects, the four core genotypes model in mice compares XX males and XY females with the usual XX females and



	most expressed	most biased	average of biased	XX/XY
PLCXD1	0.6	0.6	0.8	1.0
GTPBP6	0.9	0.7	0.9	1.1
PPP2R3B	0.8	0.7	0.8	1.0
SHOX	0.6	0.6	0.6	
CSF2RA	1.1	0.7	0.8	
IL3RA	0.8	0.7	0.8	
SLC25A6	0.9	0.8	0.9	1.1
ASMTL	0.9	0.8	0.9	0.9
P2RY8	0.9	0.6	0.7	1.0
AKAP17A	0.9	0.9	0.9	
ASMT	1.5	1.7	1.6	
DHRSX	1.0	0.8	0.9	0.9
ZBED1	0.7	0.6	0.8	
CD99	0.8	0.5	0.7	0.7
GYG2	2.1	2.1	1.7	
ARSD	1.8	1.6	1.3	
MXRA5	1.4	1.5	1.3	
PRKX	1.5	1.7	1.3	1.3
NLGN4X	0.9	1.4	1.3	
STS	1.2	1.7	1.4	
PUDP		1.9	1.5	1.7
PNPLA4	1.3	1.8	1.4	1.1
ANOS1	1.7	1.7	1.0	
FAM9C	4.6	4.6		
TCEANC	1.0	1.3	1.1	
RAB9A	0.9	1.2	1.1	1.0
TRAPPC2	1.1	1.3	1.2	
OFD1	1.3	1.3	1.2	
GPM6B	0.9	1.2	0.9	
GEMIN8	1.3	1.5	1.2	
CA5B	1.2	1.6	1.2	
ZRSR2	1.3	1.7	1.4	1.4
AP1S2	1.0	1.3	1.1	1.2
S100G				
CTPS2	0.9	1.2	1.1	
SYAP1	1.5	1.5	1.2	1.2
TXLNG	1.2	1.5	1.3	1.1
RBBP7	1.1	1.2	1.1	
EIF1AX	1.0	1.6	1.4	1.4
EIF2S3	1.2	1.4	1.2	1.2
ZFX	1.5	1.7	1.5	1.3
CXorf38	1.0	1.2	1.1	
USP9X	1.2	1.2	1.1	1.1
DDX3X	1.4	1.6	1.3	1.2
FUNDC1	1.0	1.3	1.1	1.1
KDM6A	1.5	1.7	1.6	1.6
UBA1	1.0	1.2	1.1	
CDK16	0.9	1.1	1.1	
KDM5C	1.2	1.7	1.4	1.2
IQSEC2	1.0	1.1	1.0	
SMC1A	1.4	1.4	1.2	
RPS4X	1.6	1.7	1.4	
JPX	1.5	1.6	1.4	
HTR2C	0.7	1.2		
SPRY3	0.9	0.9	0.9	
VAMP7	1.0	0.9		
IL9R	0.5	0.6	0.6	

Figure 1.1 Sex differences in expression for genes in the PARs or that escape from XCI.

The genes shown are ones which have been shown to escape from XCI in multiple studies. At the left, a schematic shows the location of the PARs and genes escaping XCI. The number of tissues in the Genotype-Tissue Expression project data (GTEx) (out of 29) with sex biased expression per gene are shown (center left). The first three columns are the XX/XY expression ratio per gene for the tissues in GTEx with the most expression, the most biased expression, or averaged for the significantly sex-biased tissues. The final column shows matched XX/XY ratios for lymphoblastoid cell lines from (Raznahan et al. 2018) (center right). Genes with Y homology are shown in blue, with those outside the PARs in bold font. Modified from a figure made by me that was featured in (Navarro-Cobos, Balaton, and Brown 2020).

XY males (reviewed in (Arnold and Chen 2009)). Phenotypes observed in XX males and XX females but not in XY males and XY females should be due to the presence or absence of the Xi and Y, and many are presumed to result from dosage differences in PAR genes or escape genes.

The phenotype of sex chromosome aneuploidies reinforces the importance of knowing the XCI status of genes, as genes that escape from XCI have mis-regulated expression in patients with sex aneuploidies. In XXY, some of these mis-regulated escape genes have been correlated with negative phenotypes (Zitzmann et al. 2015), and others have also been linked to the XXY phenotype (reviewed in (Navarro-Cobos, Balaton, and Brown 2020). In addition, escape from XCI can affect disease susceptibility in XX females. Some X-linked tumor suppressors, such as *KDM6A* escape XCI and therefore require two mutations in females to lose their tumor suppressor capabilities while only needing one mutation in males (Van Der Meulen et al. 2015) This is known as the EXiTS hypothesis (Escape from X inactivation Tumor Suppressors) (Dunford et al. 2017). Genes that escape from XCI can also have different phenotypes between males and females, as females can have a heterozygous phenotype while males will only ever be hemizygous for a mutation or healthy. For example the escape gene *DDX3X* has different severity and disease mechanism between males and females (Snijders Blok et al. 2015). Another example is autoimmune disorders, which have an increased incidence in XX and XXY individuals compared to X and XY, with many of the immune related genes escaping XCI only in immune cells or only in affected individuals (Souyris et al. 2018; Syrett and Anguera 2019). The XCI status of genes is also important to know for genome and epigenome wide association studies as this will influence the effect size of heterozygous alleles (B. Chen, Craiu, and Sun 2018; Xu and Hao 2018) and is also important for genetic selection in agriculture where one male can sire a large portion of the population (Couldrey et al. 2017).

1.4 The interaction of epigenetics and XCI status

As genes which escape XCI are actively expressed on the Xi and genes subject to XCI are silent on the Xi, we would expect different epigenetic marks associated with these genes either as a cause or consequence of silencing (Table 1.1). These marks would also be different between the Xi and Xa, at least for genes subject to XCI.

1.4.1 DNAme

The most studied epigenetic mark as it pertains to XCI status is DNAme. Most studies examining DNAme look at CpG islands that are enriched for CpGs and have their DNAme correlate with Xi expression of nearby promoters (Cotton et al. 2015). DNAme at promoters is associated with gene silencing. The interaction of DNAme and XCI status can most easily be seen when Xa

Mark	Subject Genes	Escape Genes	Promoter?	Gene Body?	References
DNAme					
Promoter DNAme	Hemi-methylation	Hypomethylation	Yes	No	(Cotton et al. 2015; Sharp et al. 2011)
Gene body DNAme	Lower methylation	Higher methylation	No	Yes	(Cotton et al. 2015; Sharp et al. 2011)
Promoter mCH	Lower methylation	Higher methylation	Yes	No	(Lister et al. 2013; Schultz et al. 2015)
Gene body mCH	Lower methylation	Higher methylation	No	Yes	(Lister et al. 2013; Schultz et al. 2015)
Heterochromatic histone marks					
H3K9me3	Enriched				(Cotton et al. 2014; Goto and Kimura 2009)
H4K20me3	Enriched				(Goto and Kimura 2009)
H3K27me3	Enriched	Depleted	Yes	Yes	(Cotton et al. 2014; Goto and Kimura 2009; Kelsey et al. 2015; Marks et al. 2015; Yang et al. 2010)
MacroH2A	Enriched	Depleted			(Changolkar et al. 2010)
Euchromatic histone marks					
H3K4me2 and 3	Depleted	Enriched	Yes	Yes	(Goto and Kimura 2009; Sadreyev et al. 2013)
H3K9ac		Enriched	Yes	Yes	(Goto and Kimura 2009)
H3K27ac	Depleted	Enriched			(Cotton et al. 2013; Kelsey et al. 2015)
H3K9me1		Enriched	No	Yes	(Goto and Kimura 2009)
Other					
XIST	Enriched	Depleted			(Engreitz et al. 2013; Murakami et al. 2009; Simon et al. 2013)
RNA Pol II	Depleted	Enriched			(Berletch et al. 2015; Goto and Kimura 2009; Murakami et al. 2009)
ATAC-seq		Enriched	Yes		(Qu et al. 2015)

Table 1.1 Epigenetic marks and their relation to genes which are subject to or escaping from XCI.

Modified from a table in (Balaton and Brown 2016)

DNAme is low; at these promoters, genes escaping XCI have low DNAme (approximately 10%) while those subject to XCI have moderate DNAme (approximately 40%) (Cotton et al. 2015).

At gene bodies, DNAme is associated with expression. Genes escaping XCI have approximately 75% DNAme while those subject to XCI have approximately 60%. Non-CpG DNAme (mCH) is rarer, with analyses showing <15% mCH in frontal cortex (Lister et al. 2013). As with DNAme, mCH is also correlated with silencing at promoters and expression in gene bodies (Schultz et al. 2015). mCH can be used to separate genes which are escaping from those subject to XCI (Keown et al. 2017; Lister et al. 2013; Schultz et al. 2015), but mCH based calls are quite tissue-specific and find many genes as having tissue-specific escape from XCI (Schultz et al. 2015) where DNAme does not (Cotton et al. 2015).

1.4.2 Heterochromatic histone marks

Heterochromatic marks such as H3K9me3, H4K20me3, H3K27me3 and macroH2A are associated with gene silencing and are enriched on the Xi at genes subject to XCI (Table 1.1) (Changolkar et al. 2010; Cotton et al. 2014; Goto and Kimura 2009; Kelsey et al. 2015; Marks et al. 2015; Yang et al. 2010). H3K27me3 and macro H2A have also been seen to be depleted at genes escaping from XCI. EZH2, the gene that catalyzes H3K27me2 to H3K27me3, is also enriched at genes subject to XCI and depleted at genes escaping from XCI (Cotton et al. 2014). Heterochromatic marks tend to be recruited in broad domains across the Xi, and there are many studies detailing their recruitment and spread (reviewed in (Dixon-McDougall and Brown 2015)).

1.4.3 Euchromatic histone marks

Euchromatic marks such as H3K4me2/3, H3K9ac, H3K27ac and H3K9me1 are associated with active transcription and are enriched on the Xi at genes escaping from XCI (Table 1.1) (Goto and Kimura 2009; Kelsey et al. 2015; Sadreyev et al. 2013). H3K4me2/3 and H3K27ac are additionally depleted at genes subject to XCI. For euchromatic marks, it is hard to tell whether they cause expression from the Xi or are there because of expression from the Xi. It is likely that there is a positive reinforcement cycle where euchromatic marks recruit transcriptional machinery, and transcription recruits euchromatic marks.

1.4.4 Other marks

Other factors vary between genes that are escaping and those subject to XCI. The lncRNA XIST initiates XCI and interacts with genes that are subject to XCI and their promoters (Simon et al. 2013) while being depleted at genes that escape from XCI (Engreitz et al. 2013; Murakami et al. 2009; Simon et al. 2013). RNA polymerase II, which is responsible for mRNA transcription, is found on the Xi at genes which escape XCI but not at genes subject to XCI (Berletch et al. 2015; Goto and Kimura 2009; Kucera et al. 2011). Assay for Transposase Accessible Chromatin with sequencing (ATAC-seq) is an assay for open chromatin, which shows peaks at the promoters of genes (Buenrostro et al. 2013) and twofold enrichment in females at genes that escape XCI as compared to males (Qu et al. 2015).

1.5 Determining XCI status experimentally

There are multiple ways to determine which genes are escaping from XCI, all of which rely on somehow differentiating the Xi from the Xa (Figure 1.2, reviewed in

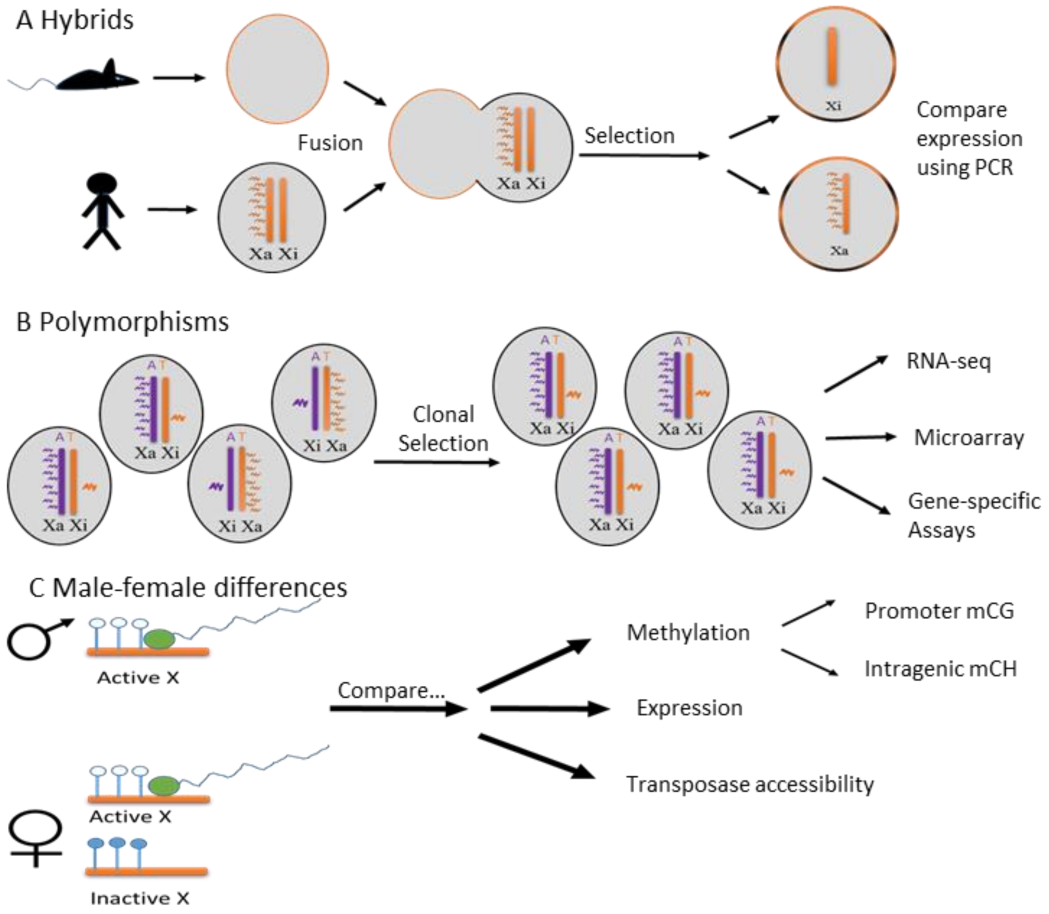


Figure 1.2 Approaches to identifying genes that escape from inactivation.

(A) Human/mouse hybrids. Hybrid cells are made by fusing human and mouse somatic cells. Cells containing a human active X (Xa) or inactive X (Xi) are then selected and X-linked expression compared between these cell lines to determine which genes are escaping (expressed in both sets of hybrids) or subject (only expressed in the Xa hybrids) to XCI. (B) Polymorphisms. Normal females are a mosaic for which X is the Xa or Xi. Clonal selection generates a population of cells having the same Xa. Quantifying the proportion of expression from the alleles on the Xa and Xi is used to determine if the gene containing the polymorphism is escaping or subject to XCI. (C) Male-female differences. Male cells contain only an Xa while female cells contain both an Xa and Xi. Examining male-female differences allows you to determine what effect having an Xi has on X-linked methylation, expression and transposase accessibility. Genes that escape XCI have methylation levels similar to males while having higher expression and transposase accessibility than males; genes that are subject to XCI are the opposite. Modified from a figure made by me and featured in (Balaton and Brown 2016).

(Navarro-Cobos, Balaton, and Brown 2020)). Many accomplish this by using male data as a stand-in for the Xa and subtracting male from female data (Xa+Xi or $\frac{Xa+Xi}{2}$, depending on the data) to approximate the Xi value. Other studies use allelic differences between the Xa and Xi or have physical separation of the Xa and Xi alleles.

1.5.1 Mouse-human hybrids

One of the older, but often used ways to differentiate the Xa from the Xi is using mouse-human hybrid cells. This method involves the fusion of a mouse cell line with a mutation in the X-linked gene *Hprt* with human cells with a functional HPRT and then using various chemicals to select for cells retaining the human Xa (with functional HPRT) or the human Xi (without functional HPRT) (Figure 1.2A) (C. J. Brown and Willard 1989). Expression can then be examined and compared between cells which have only the human Xi or Xa, with genes that are expressed from both (with the Xi having at least 10% as much expression as the Xa) being called as escaping from XCI, while those that are only expressed from the Xa being called as subject to XCI (Carrel and Willard 2005). These mouse-human hybrid cells can also be used to look at other epigenetic marks and how they differ between the Xi and Xa at regions where human specific probes can be made (Goto and Kimura 2009). These hybrids show defects in XIST localization to the Xi, so there may be epigenetic mis-regulation or missing factors in these cell lines (Clemson et al. 1998).

1.5.2 Xi/Xa expression

The best way to make XCI status calls is to use genetic polymorphisms to differentiate expression from the Xi and Xa within the same sample (Figure 1.2B) (Berletch et al. 2015;

Carrel and Willard 2005). The advantage here is that the alleles are in the exact same cellular context, and do not need to be normalized across samples. Two large limitations of using Xi/Xa expression are that you need an expressed heterozygous SNP and need Xi choice to be completely skewed. Heterozygous SNPs cannot be found in all genes, limiting the genes which can have XCI status calls with this method. Additionally, not all samples will be heterozygous at even the most common SNPs, this means that you will need more samples to get the same effective sample size. Sample choice is limited by the need for complete skewing of Xi choice. In humans, samples that have been used include those with X-linked diseases and those that have been clonally cultured (Carrel and Willard 2005).

In mice, there are strains engineered to have an *Xist* knockout on one allele, so that allele is always on the Xa (Berletch et al. 2015). These strains are then bred to a distantly related wild-derived strain to get an F1 cross with maximal number of heterozygous SNPs. A limitation here is that there are only a few mouse strains used so population-specific XCI may not be seen as readily.

Single cell RNA sequencing (scRNA-seq) can be used to examine Xi/Xa expression without the need for samples to have skewed Xi choice (Tukiainen et al. 2017; Wainer Katsir and Linial 2019) (reviewed in (Keniry and Blewitt 2018)). With scRNA-seq, you can also see variation in the XCI status of genes depending on which allele is on the Xi, and even randomly between cells with the same Xi allele in the same sample (Hagen et al. 2020; Tukiainen et al. 2017). One study has attributed this heterogeneity in XCI status to differences in cell cycle and XIST expression level between cells (Garieri et al. 2018). scRNA-seq does suffer from a lack of sequencing depth per cell, limiting the ability to call XCI status of lowly expressed genes. One method to avoid this problem commonly used by scRNA-seq studies is to select for reads at the 3’ or 5’ of genes; this works well for quantifying gene expression but is less useful when you require a SNP to differentiate reads from the Xa and Xi, as you are then limited to SNPs at the 3’ or 5’ end of the gene. Combining cells with the same Xi alleles together can increase the overall sequencing depth for allelic analysis, however this eliminates the scRNA-seq advantage of being able to see heterogeneity between cells with the same allele.

1.5.3 Sex-specific expression

As mentioned in section 1.3, genes that escape XCI tend to have female-biased expression (Figure 1.1) and this has been used as a proxy for XCI status (Tukiainen et al. 2017). While sex-biased expression was enriched at genes escaping XCI when compared to genes variably escaping or subject to XCI, only 74% of genes escaping from XCI were ever seen to have sex-biased expression, and many of these were not biased in the majority of tissues (Tukiainen et al. 2017). Female-male expression differences are more complicated than just the addition of an Xi in females. Feedback regulation and hormonal effects can diminish or increase any expression differences caused by the addition of an Xi, in a difficult to predict fashion (Navarro-Cobos, Balaton, and Brown 2020). This can give some idea of which genes may be escaping from XCI in a new species or tissue, but other methods should be used to make confident XCI status calls.

1.5.4 Sex-specific DNAme

As mentioned in section 1.4.1, genes with CpG islands at their promoter and low male DNAme are expected to have DNAme levels reflective of their XCI status (Cotton et al. 2015). Genes escaping from XCI will have low DNAme on both alleles, while those subject to XCI will have low DNAme on the Xa and high DNAme on the Xi, averaging as moderate DNAme overall. Male DNAme can be used as a stand in for the Xa, so that you do not call genes as subject to XCI if both the Xa (males) and Xi ($male + 2 * (female - male)$) have high DNAme (Figure 1.2C). Gene body DNAme, while different between genes escaping and subject to XCI, is more subtle and not as conclusive. Gene body mCH is enriched at genes escaping from XCI in females and has been used to calls genes as escaping from XCI (Keown et al. 2017; Lister et al. 2013; Schultz et al. 2015). Allelic DNAme gives further confidence to these calls, with mCH on the Xi being undetectable except at the gene body of genes found escaping from XCI (Keown et al. 2017).

1.5.5 Other methods

There are also other methods to determine XCI status of genes. These include RNA-FISH, ATAC-seq, and predictive models. RNA-FISH allows you to visualize the expression of genes from the Xa and Xi, so that one locus can be seen for genes which are subject to XCI and two loci are seen for genes escaping from XCI (Al Nadaf et al. 2012). Probes against XIST can be included to further support that one of the expression loci is from the Xi. The weakness to RNA-FISH is that only well-expressed genes can be visualized with it. It is also a low throughput method.

ATAC-seq shows the location of accessible chromatin, with females having approximately twice as much signal as males at genes which escape from XCI (Qu et al. 2015). Predictive models have been used to classify genes which escape XCI from those subject to XCI (De Andrade E

Sousa et al. 2019; Z. Wang et al. 2006). A human model using only genomic repeat elements achieved accuracy over 80% (Z. Wang et al. 2006) while a mouse model trained on a combination of genomic and epigenomic data had an accuracy of 78% (De Andrade E Sousa et al. 2019).

1.6 Theoretical determinants of XCI status

There are three main categories of elements theorized to control XCI status: waystations, escape elements and boundaries (Figure 1.3, reviewed in (Balaton et al. 2018)). All three categories have evidence to support them and may have complementary or conflicting roles in determining which genes are escaping or subject to XCI.

1.6.1 Waystations

Waystations are elements which help spread the silencing of XCI. *Xist* transgenes are less capable of silencing autosomes than they are the X (Loda et al. 2017), and in X:autosome translocations, silencing does not spread as well as on the X (Cotton et al. 2014). These and similar studies suggest that there is some characteristic of the X that allows XCI to spread more efficiently along it. The main waystation element that has been proposed is LINE repeats as they are enriched on the X, and cells have machinery to silence these repeats which may be co-opted for XCI (Lyon 1998). LINE1 elements were additionally seen enriched on the region of the X containing the *XIST* gene while being depleted at regions containing genes which escape from XCI (Bailey et al. 2000; Loda et al. 2017; Z. Wang et al. 2006). Additionally in X:A translocations, autosomal genes which had efficient XCI were enriched for having pre-existing heterochromatic marks, such as H3K27me3, RING1B and EZH2

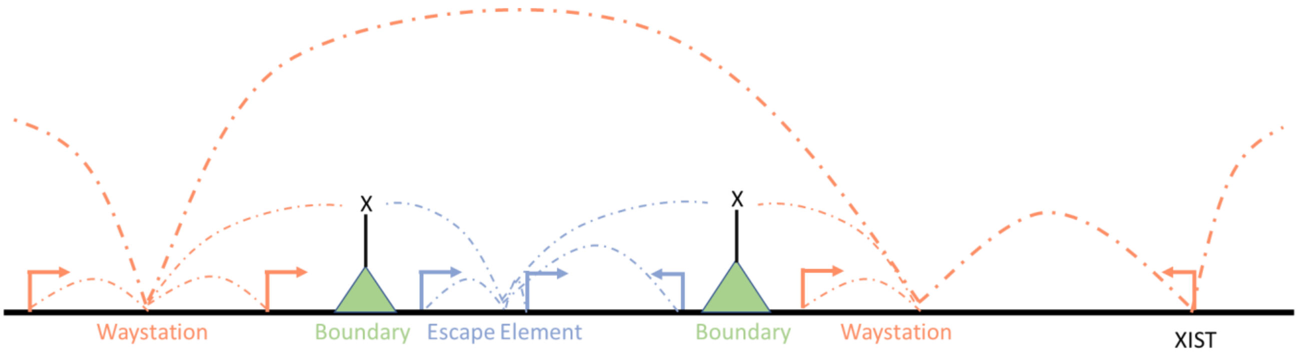


Figure 1.3 Theoretical determinants of XCI status.

Waystations spread XCI initiated from the *XIST* gene. Escape elements allow nearby to genes to escape from XCI. Boundary elements block the spread of escaping or silencing across them.

(Cotton et al. 2014; Loda et al. 2017). So XCI may be more capable of spreading to and from locations with pre-existing heterochromatin.

1.6.2 Escape elements

Escape elements allow nearby genes to escape from XCI and protect them from XCI. The strongest evidence for escape elements is that some Bacterial Artificial Chromosomes (BACs) containing genes which escape from XCI are still able to escape from XCI when placed at other locations, where the local genes are subject to XCI (Horvath, Li, and Carrel 2013; Peeters et al. 2018). These elements were conserved and capable of acting across species when human escape genes were integrated onto the mouse X (Peeters et al. 2018). The identity of these escape elements remain inconclusive, however the transcription factors CTCF and YY1 have been seen enriched near genes which escape from XCI (C. Y. Chen et al. 2016; Loda et al. 2017) and ALU repeat elements have also been seen enriched at genes escaping from XCI (Cotton et al. 2014; Z. Wang et al. 2006).

1.6.3 Boundaries

Boundary elements theoretically block the spread of heterochromatin or euchromatin on the X and delineate regions which are escaping from regions subject to XCI. Deletion of the edge of an inserted escape BAC allowed escape from XCI to spread beyond the edge of the BAC, affecting the genes neighboring the integration site (Horvath, Li, and Carrel 2013). CTCF has been found enriched at boundaries between domains of genes escaping and those subject to XCI and is a likely candidate boundary element (Filippova et al. 2005). Genes escaping from XCI cluster within topologically associating domains (TADs), with a tendency of genes within a TAD to have the same XCI status (Marks et al. 2015). TAD boundaries may therefore be enriched for boundary elements, and we do see enrichment of CTCF at TAD boundaries (Dixon et al. 2012).

1.7 Thesis objectives

The goal of this thesis was to classify genes by their XCI status and identify genetic and epigenetic differences which may control which genes are escaping and which are subject to XCI. Chapter 2 is a meta-analysis of existing studies making XCI status calls in humans, giving a combined XCI status call with improved confidence for future studies. These combined XCI status calls were used for the remaining chapters as a baseline for comparison of new XCI status predictions and for comparison with epigenetic marks.

In Chapter 3, I used Xi/Xa expression and previous XCI status calls in human and mouse to determine DNAme thresholds separating genes which escape from those which are subject to XCI across species. These thresholds were then applied to high-throughput DNAme datasets across 12 different mammalian species to determine the XCI status of genes. I observed conservation of XCI status across species, and multiple features which had previously been associated with genes escaping or subject to XCI were also seen associated across species. These XCI status calls will be useful to labs using these other species as a model and those labs studying these species for agricultural purposes.

Lastly, in Chapter 4 I examined how epigenetic marks differ across genes with differing XCI status (as determined in Chapter 2), showing that histone marks do not have the same Xi to Xa pattern across all genes escaping from or subject to XCI. Additionally, genes which variably escape from XCI had different epigenetic marks in samples escaping vs subject to XCI, but these were not consistent across genes. Overall, the work in this thesis advances our understanding of which genes have each XCI status, in humans and across mammals, and of the genetic and epigenetic features which may control escape from XCI.

Chapter 2: Derivation of consensus inactivation status for X-linked genes from genome-wide studies

2.1 Introduction

The goal of this study was to integrate the results from studies that have done large-scale analyses of which genes escape from, are subject to, or variably escape from XCI and to come up with a catalog of consensus XCI status calls. The first of the three main studies to be integrated used two methods (Carrel and Willard 2005). Human-mouse hybrid cell lines with an active or inactive human X chromosome allowed the direct examination of which genes are expressed from the Xi. Comparison of the expression of each gene from the Xi cell lines to the expression from the Xa cell lines led to a call of escape from XCI when there was 10% or more relative Xi expression. These results will be referred to as the Carrel hybrid study. The Carrel hybrid study used nine Xi hybrid cell lines and made XCI status calls for 465 genes (Table 2.1). Genes which escaped in only 0, 1, or 2 cell lines were called as being subject to XCI, and genes which escaped in 7, 8, or 9 cell lines were called as escaping from XCI. Genes which escaped XCI in 3 to 6

Study	Carrel hybrid	Carrel SNP	Cotton AI	Cotton DNAme
XCI status calls	465	84	429	406
Number of samples	9	40	99	1875
Average number of informative samples	-	12	25	-

Table 2.1 Sample sizes of previous studies.

The number of samples used and XCI status calls made per study for the Carrel hybrid, Carrel SNP, Cotton AI and Cotton DNAme studies. The average number of informative samples was also included for the Carrel SNP and Cotton AI studies as only samples which were heterozygous at a SNP could be used for these studies.

hybrid cell lines were called as variably escaping from XCI. The same publication examined the allelic ratio of X-linked expressed SNPs in fibroblast cell lines which were skewed completely for which X was inactivated, such that in a population of cells, the same allele was always on the Xa and biallelic expression would reflect escape from XCI. These results will be referred to as the Carrel SNP study. The Carrel SNP study examined a panel of 40 cell lines and made XCI status calls for 84 genes, with an average of 12 informative cell lines per gene (Table 2.1). Genes which had less than 23 % of their cell lines escaping from XCI were called as subject to XCI while genes with over 78 % of their cell lines escaping XCI were called as escaping from XCI. Genes with between 23 and 78 % of their cell lines escaping from XCI were called as variably escaping from XCI.

The second study looked at the expression of X-linked SNPs using microarray data to include assessment of intronic polymorphisms (Cotton et al. 2013). The allelic imbalance (AI) between the allele on the Xa and the allele on the Xi for genes which already had strong evidence for being subject to XCI was used to assess how much skewing of XCI was present in each cell line, and this was then used to calculate how much of the AI was due to mosaicism and how much was due to escape from XCI. This will be referred to as the Cotton AI study. The Cotton AI study used 99 cell lines and made XCI status calls for 419 genes with an average of 25 informative samples per gene. The same thresholds were used for the AI study as the SNP study (Table 2.1).

The third study used CpG island methylation data from the Illumina Infinium Human Methylation450 BeadChip platform (450k) (Cotton et al. 2015). It compared the female and male

DNAme levels at CpG islands at the promoters of genes known to be subject to XCI and those known to escape from XCI to develop a classifier which could predict the XCI status of other genes. This classifier was then used on genes with unknown or less evident XCI status to make new XCI status calls. This will be referred to as the Cotton DNAme study. The Cotton DNAme study examined 1875 female samples and 1053 male samples, giving XCI status calls for 409 genes (and multiple transcription start sites for most genes) (Table 2.1). XCI status calls were given individually by tissue, and the overall XCI status call was a list of calls which were obtained in at least one tissue. An uncallable designation was used when less than 50 % of samples in that tissue had a methylation level and male-female difference within two standard deviations of the subject or escape training genes in that tissue (50 genes were left in an uncallable category because they were uncallable in over half of the tissues examined). Genes were called as subject to or escaping from XCI in a tissue if all samples that were given an XCI status call gave the same call. Genes were called as variably escaping from XCI if they had at least one sample giving each XCI status call (subject and escape). Variable escape from XCI was rare in this study with a maximum of one third of all tissues showing variable escape for any given gene.

Additional approaches to determine XCI status, which have examined fewer genes, include DNAme analysis at non-CpG sites (Lister et al. 2013), SNP expression analysis in single cells (Carrel and Willard 1999), RNA-FISH to detect expression from both X chromosomes (Hacisuleyman et al. 2014), analysis of protein polymorphisms in clonal cells by size (Davidson, Nitowsky, and Childs 1963) or by enzyme activity (Migeon et al. 1981), microarray analysis of cellular expression with varying numbers of X chromosomes (Sudbrak et al. 2001), microarray analysis of expression differences between males and females (Craig et al. 2004), and allelic expression analysis of RNA-seq data from clonal cells (Rozowsky et al. 2011).

Each of the three studies integrated in this analysis have examined over 400 different genes, and combined there is data for 639 genes. Generally, multiple studies agree, and only 47 genes show substantial discordancies between studies, which we discuss. There is an enrichment of discordancies and calls of mostly variable escape from XCI at putative XCI boundaries. Seventy percent of protein-coding messenger RNA (mRNA) genes have an XCI status call with the hypermethylated cancer-testes antigen gene family accounting for 42 % of the remaining uncalled mRNA genes. However, fewer of the non-protein-coding genes have a defined XCI status.

2.2 Methods

2.2.1 Categorization of X-linked genes

A full list of genes on the X chromosome was downloaded from University of California, Santa Cruz (UCSC)'s HG19.knownGene table browser (Karolchik et al. 2004). The table was condensed manually from having an entry for each transcription start site to having an entry for each gene. XCI calls from the studies were added to the table, matching alternate gene names from the National Center for Biotechnology Information (NCBI) (G. R. Brown et al. 2015) along with using the in silico PCR tool in UCSC (Hinrichs et al. 2006) with published primers (Carrel and Willard 2005).

Genes were placed into eight categories for an overall XCI status call. If all of a gene's calls from different studies were the same, then the gene was placed in a category for all subjects, all escapes or all variable escapes. If the majority of studies (2 out of 3 or 3 out of 4) gave the same call, then the gene was placed in the mostly subject, mostly escape or mostly variable escape categories. Genes that had one-call subject or one-call escape and a variable escape call which leaned towards the same call (variable escape in a study, with less than 34 % or greater than 65 % of samples escaping XCI) were also placed in the mostly subject and mostly escape categories. The Cotton DNAme study gave some calls that were escape + variable escape or subject + variable escape; for my categorization, these genes were considered to be whichever call was given in the most tissues, this was usually subject or escape. Genes that had no calls in any of the studies were designated as the no call category, while genes that did not fit any of these other categories were placed in the discordant category. Discordant genes had either an even split of different calls or had one of each call (subject, escape, and variable escape from XCI).

Genes were sorted by their transcript type (mRNA, micro RNA (miRNA), ncRNA, snRNA, transfer ribonucleic acid (tRNA)) as determined by UCSC's HG19.kgXref table (Karolchik et al. 2004) and if still unknown, a search of NCBI. A list of cancer-testis antigen genes was taken from CTdatabase (L. G. Almeida et al. 2009).

To determine the source of discordancies, genes with three or four calls and only one study giving a different call from the other studies were examined. The study which gave the discordant call was noted, along with the call it gave and the call agreed upon by the other studies.

2.2.2 Expression analysis

Expression data for the lymphoblast cell line GM12878 was downloaded from GEO dataset GSE30400 (Rozowsky et al. 2011), and expression data for the fibroblast cell line IMR90 was downloaded from GEO dataset GSM981249 (Yue et al. 2014). This data was annotated using Seqmonk (Babraham Bioinformatics) using our condensed X chromosome gene list. A Tukey test was performed to determine if expression levels in lymphoblasts differed amongst the various categories using the multcomp package in R (Hothorn, Bretz, and Westfall 2008; R core Team 2014). This was repeated for the calls given by each individual study.

2.2.3 Domain analysis

Domains were annotated by labeling any genes between escape genes, without crossing a subject gene, as being in an escape domain and labeling any genes between subject genes without crossing an escape gene as being in a subject domain. Genes between a subject and escape gene, with no other subject or escape genes in between, were classified as boundaries; boundaries can start inside of the gene body of a gene which is subject to or escaping from XCI, as a gene's XCI status is likely determined by its promoter. Enrichment was determined using a chi-square test (chisq.test from the MASS package in R (R core Team 2014; Venables, WN. Ripley 2002)). Standardized residuals were extracted from the chi-square test and used to determine enrichment of certain categories (Sharpe 2015), followed by a chi-square test comparing the enrichment of variable, mostly variable and discordant genes in boundaries, individually against genes with no call. Genes with no call were shown to be a good control (p value >0.95) by a chi-square comparison between genes with no call and genes with a call, in boundaries compared to the outside of boundaries.

2.3 Results and discussion

2.3.1 Creation of a consensus XCI status

Gencode currently lists 1144 genes on the human X chromosome (Harrow et al. 2006, 2012). Between the four datasets examined, 639 (54 %) of these genes have an XCI status call (Figure 2.1A). There is a roughly equal distribution of genes that have been examined in one, two, or

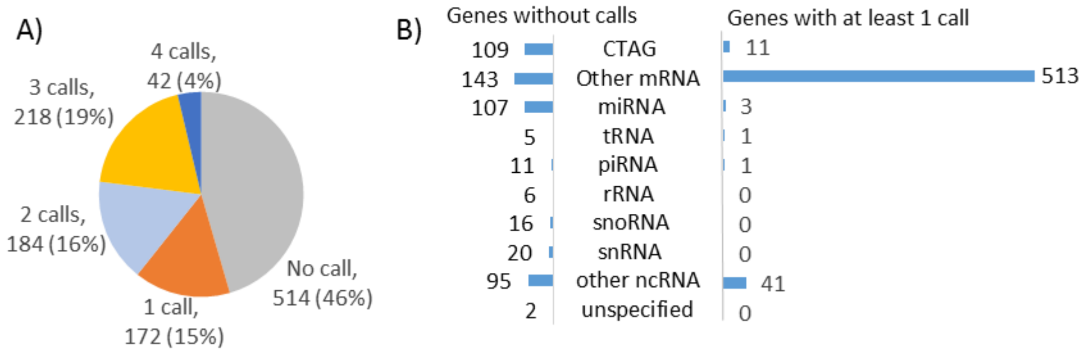


Figure 2.1 The majority of X-linked protein coding genes have an XCI status call.

(A) The number of datasets contributing an XCI status call per gene. The number of calls is the number of studies which gave an XCI status call of subject, escape or variable escape from XCI. Genes with no call were not mentioned in any of the studies but were included in Gencode for HG19 [38,39]. (B) The distribution of RNA transcript types for genes with and without an XCI status call. Transcript type was taken from Gencode or an NCBI search [30]. CTAG are cancer testes antigen genes which are protein coding genes expressed exclusively in cancer and in testes and hypermethylated in other tissues making XCI status calls very difficult. Other mRNA are mRNA genes that are not members of the CTAG family.

three of these studies; however, very few genes have an XCI status call in all four studies because the Carrel SNP study has a small sample size of 84 (Figure 2.1A). Comparing the distribution of transcript types between genes with XCI status calls and those without, protein-coding genes are much more likely to have a call whereas genes for non-coding RNA such as miRNA and tRNA are more likely to not have an XCI status call (Figure 2.1B). A large proportion of the protein-coding genes without a call can be explained by them belonging to the Cancer-Testis Antigen Gene (CTAG) family (Figure 2.1B). CTAG genes are hypermethylated and silenced on both Xs in healthy female cells and are normally only expressed in cancer cells or in the testes of males (L. G. Almeida et al. 2009). Other genes lacking calls have very low expression (RPKM values less than 0.1) in the fibroblasts and lymphoblasts examined in the hybrid, SNP, and AI studies (102 out of 143 non-CTAG genes without a call (Supplemental Table S2.1)), and all genes without calls either are not present on or filtered out from the DNAme microarray used for assessment in the DNAme study (reasons for filtering include hypermethylation in male samples and mapping to repetitive elements or to the autosomes (Cotton et al. 2015)) or were found to have methylation levels in an uncallable region between that found for known subject and escape genes. There were only 24 genes that lacked expression and were called by the DNAme study but could not be called by the expression studies. Enrichment of calls for protein-coding genes likely reflects the more recent identification of lncRNA genes. The smaller RNA types are too small or too tissue-specific to have their XCI status determined in these studies; furthermore, high homology to another gene might prevent assessment of XCI status and the X is enriched for large inverted repeats (Warburton et al. 2004).

Genes were divided into eight categories based on what XCI status the studies called the gene and how often the studies agreed (Figure 2.2A). Seventy-three percent of genes were given an overall call of subject or mostly subject, roughly agreeing with the percent found to be subject in each individual study (Figure 2.2B). The percent of escape and mostly escape genes (12 %) was also similar to the percent of escape genes found by each individual study. The variable escape and mostly variable escape categories (8 %) agreed with the Carrel studies; however, the Cotton studies have large differences in the amount of genes they call variable escape. This difference in the number of variable escape calls contributed to a fair amount of the discordancies between studies. Seven percent of genes on the X were discordant between studies and no consensus call

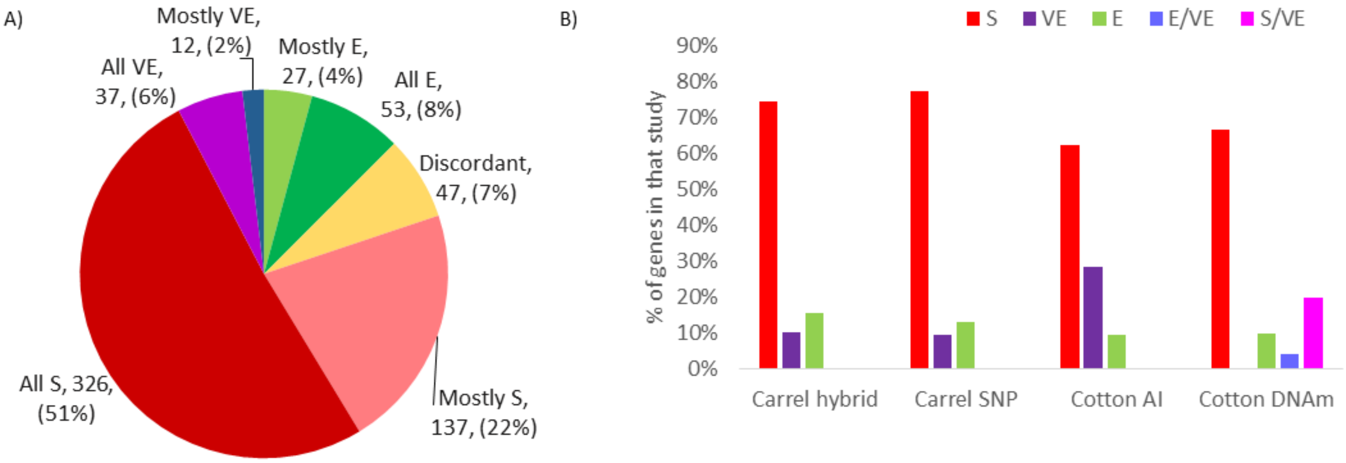


Figure 2.2 Consensus XCI status calls.

(A) Distribution of our consensus XCI status calls. E is escape from XCI, S is subject to XCI and VE is variably escaping from XCI in some individuals or tissues. The mostly E, S or VE categories are genes which have two out of three or three out of four XCI status calls agree on a call of E, S or VE and the last study disagree. The all E, S or VE categories had at least one XCI status call for E, S or VE and had no XCI status calls disagree. Discordant calls had either an even split of different XCI status calls or had one of each call. Genes with no call were left out of this graph. (B) The distribution of XCI status calls given by each individual study. See above for a description of E, S and VE. E/VE and S/VE are calls from the Cotton DNAm study where most tissues were given a call of escape or subject but some tissues were given a call of variable escape. For the sample sizes of each study see Table 2.1.

could be assigned, while another 28 % had a single discordancy (categorized into one of the mostly escape, mostly subject, or mostly variable escape categories) (Figure 2.2A).

2.3.2 Discordancies between studies

To understand the nature of the discordancies between studies, we tabulated the frequency with which studies disagreed and the difference from the consensus call (Table 2.2). The Cotton AI study was the most discordant study with 11 % of its calls disagreeing with two or three other studies and a tendency to call genes as variable escape when other studies called that gene escape or subject (Figure 2.3A). This tendency to call variable escape could be due to the extra calculations involved to correct for using cells which were only partially skewed. Another contributing factor could be that the AI study, in addition to the exonic SNPs used in the SNP study, also used intronic SNPs which are spliced out and degraded and would be present in lower levels which may affect the XCI status calls drawn from them. The AI study also used more

		Discordant Study			
Discordant call	**Consensus call**	**Carrel hybrid**	**Carrel SNP**	**Cotton AI**	**Cotton DNAme**
E	**VE**	0	1	0	2
	S	7	1	1	2
VE	**E**	1	1	17	0
	S	9	0	26	0
S	**E**	0	1	3	0
	VE	0	1	1	3

Table 2.2 Most studies show a trend with what they are calling discordantly.

E is escape from XCI, S is subject to XCI and VE is variable escape from XCI. Discordant call is which XCI status call is being given by the discordant study while consensus call is the XCI status call agreed upon by two or more other studies.

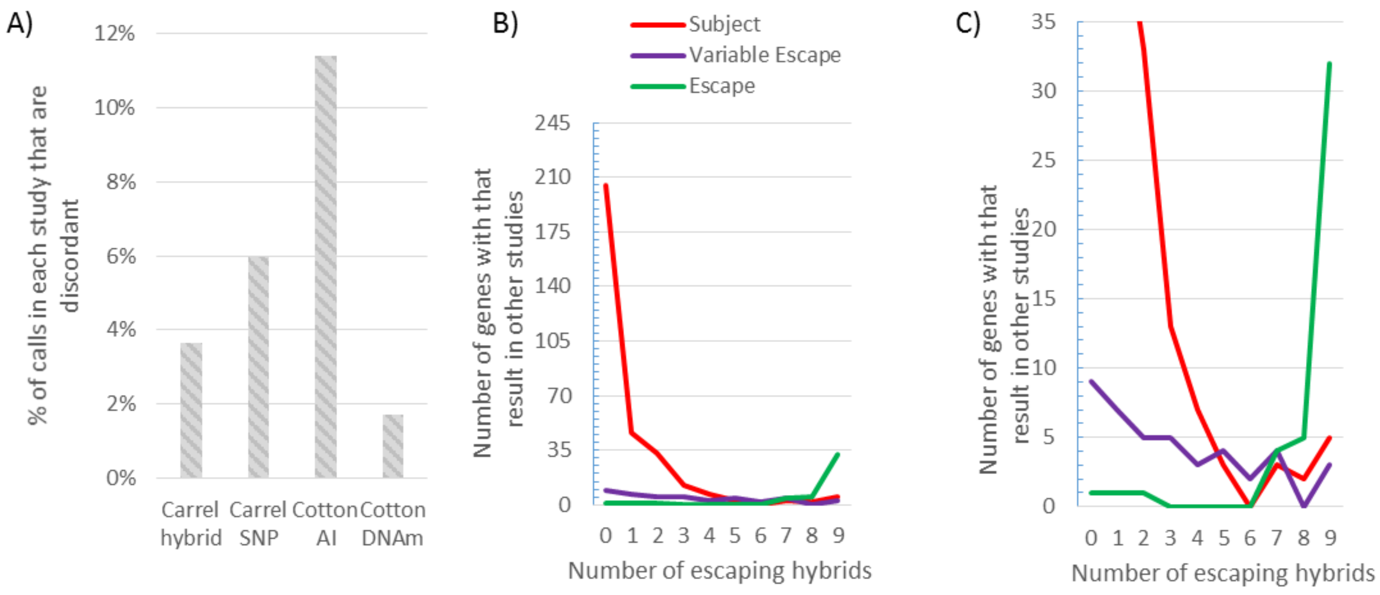


Figure 2.3 Comparison of discordancies.

(A) The level of discordancies in each study. A gene is counted as discordant in a study if that study gives a call and at least two other studies agree on a different call. For the sample sizes of each study see Table 2.1. (B) Comparison of the Carrel hybrid calls to calls from other studies. The number of escaping hybrids is, for each gene, in how many mouse-human hybrid cell lines (out of 9) did that gene escape XCI. The Y axis is how many genes one or more other studies agreed were subject to, escaping from or variably escaping from XCI. (C) A magnified version of B to better show escape and variable escape from XCI.

samples than the other expression studies (an average of 25 informative samples per gene compared to 12 in the SNP study and 9 in hybrids) which would increase the chance of finding variable escape genes. The Cotton DNAme study was the most concordant study with only 2 % of its calls disagreeing with 2 or three other studies; however, it also had an uncallable category for genes which had methylation levels or male-female methylation differences between the thresholds set by training sets of known subject and escape genes (the threshold was set at two standard deviations away from the training set mean). Cotton did not give these genes a call and they were not considered in this analysis. The discordancies in the Cotton DNAme study were mostly due to it not finding any genes with a high level of variable escape from XCI (Figure S1.1). The hybrid study discordancies arose from genes called escape or variable escape when other studies gave a subject call.

Tissue-specific differences in XCI status are an important possible source of discordancies between studies. The Carrel hybrid and SNP studies were both done in a single tissue type, fibroblasts. The Cotton AI study used both lymphoblasts and fibroblasts and found that 10 % of genes showed evidence of tissue-specific escape from XCI; these genes would not appear to be variably escaping in the Carrel studies. However, the Cotton DNAme study looked at 27 tissue types (including fibroblasts and whole blood (which includes lymphoblasts)) and found high concordance between tissues and very few tissue-specific differences in escape from XCI. Therefore, a more likely source of differences between studies could be from differences acquired in cell culture. The Cotton DNAme study was the only study to use primary cells; the Carrel studies and Cotton AI study used cultured cells. Previous studies have shown differences in XCI between primary cells and cultured cells from the same organism (Berletch et al. 2015; Nino-Soto et al. 2005) and between individuals at different ages (Bennett-Baker, Wilkowski, and Burke 2003). Genes with discordancies between studies or calls of variable escape in individual studies may be the genes most prone to epigenetic changes in culture. In the mostly subject and mostly escape categories, 90% of the genes have variable escape as the discordant call and 82% of the discordant genes have at least one variable escape call (Supplemental Table S2.1). This difference between the studies could also be due to differences between the methylation status and XCI status of some of the more variable genes; however, most genes which are found variable by other studies are not given an XCI status call by the Cotton DNAme study (Figure S1).

The mouse-human hybrid cells may be the most different from primary cells. In hybrid cells, XIST fails to properly localize to the Xi (Clemson et al., 1998). This may reflect a loss of some heterochromatin marks on the Xi, leaving X inactivation to be maintained by fewer marks, including DNAme (Gartler, SM., Dyer KA., Marshall Graves JA. 1985). X-inactivated genes in hybrids are more vulnerable to reactivation by 5-azacytidine, a methylation inhibitor (Mohandas, Sparkes, and Shapiro 1981), and approximately 1 in 105 hybrid cells will spontaneously reactivate the HPRT gene which is normally subject to inactivation (Marshall Graves and Young 1982). Reactivation could explain the genes being called escape or variable escape in the Carrel hybrid study while being called subject in other studies. When compared with consensus calls from other studies, genes found to escape in three or four hybrid cell lines in the Carrel hybrid study (which were thus classified as variable escape in that study) are more often called subject to XCI than variably escaping from XCI (Figure 2.3B, Supplemental Table S2.2). Reactivation of subject genes appears to occur for a small percentage of genes in hybrid cell lines.

Most of these studies have used expression to monitor XCI status. We therefore examined whether expression level has an effect on a gene’s XCI status call (Figure S2). None of the categories had significantly different expression levels ($p > 0.05$) nor were there significant differences in expression levels for the calls in each individual study (not shown).

2.3.3 Domains of escape and boundaries

It has been hypothesized that there are domains on the Xi with coordinately regulated XCI caused by nearby XCI way stations spreading XCI or escape elements promoting euchromatin with boundaries separating the two (N. Li and Carrel 2008; Miller and Willard 1998; Pinter et al. 2012). We used our categories to locate these domains and examined the domain enrichment of discordancies and variably escaping genes (Figure 2.4). Fully variable escape genes were most often found in subject domains at a frequency similar to the overall distribution of genes (Figure 2.4B). Genes which mostly variable escape were most often in escape domains and boundary regions suggesting variation in escape genes. Discordant genes were equally abundant in subject domains and boundary regions, despite the substantially smaller size of the boundary regions. Boundaries between domains may provide clues to the mechanisms controlling XCI. Fully variable escape genes were not enriched in boundaries (*p* value >0.95) whereas mostly variable escape and discordant genes each had an approximately threefold enrichment (from 2 to 6 % of genes for mostly variable escape (*p* value $<5*10^{-4}$) and from 7 to 20% for discordant genes (*p* value $< 4*10^{-7}$)) (Figure 2.4C). We hypothesize that these genes may be variable due to either natural variability in the position of a boundary or from instability of boundaries due to cell culture. These discordant and variable genes are spread throughout the different boundaries; 42% of boundaries have discordant or variable genes in them and 45% of all the discordant genes and 60% of all the mostly variable escape genes are in boundaries.

2.3.4 Comparison to additional studies examining XCI

We compared our XCI status calls to those found by various studies examining the XCI status of single genes or regions and generally found agreement (Supplemental Table S2.1). A chi-square

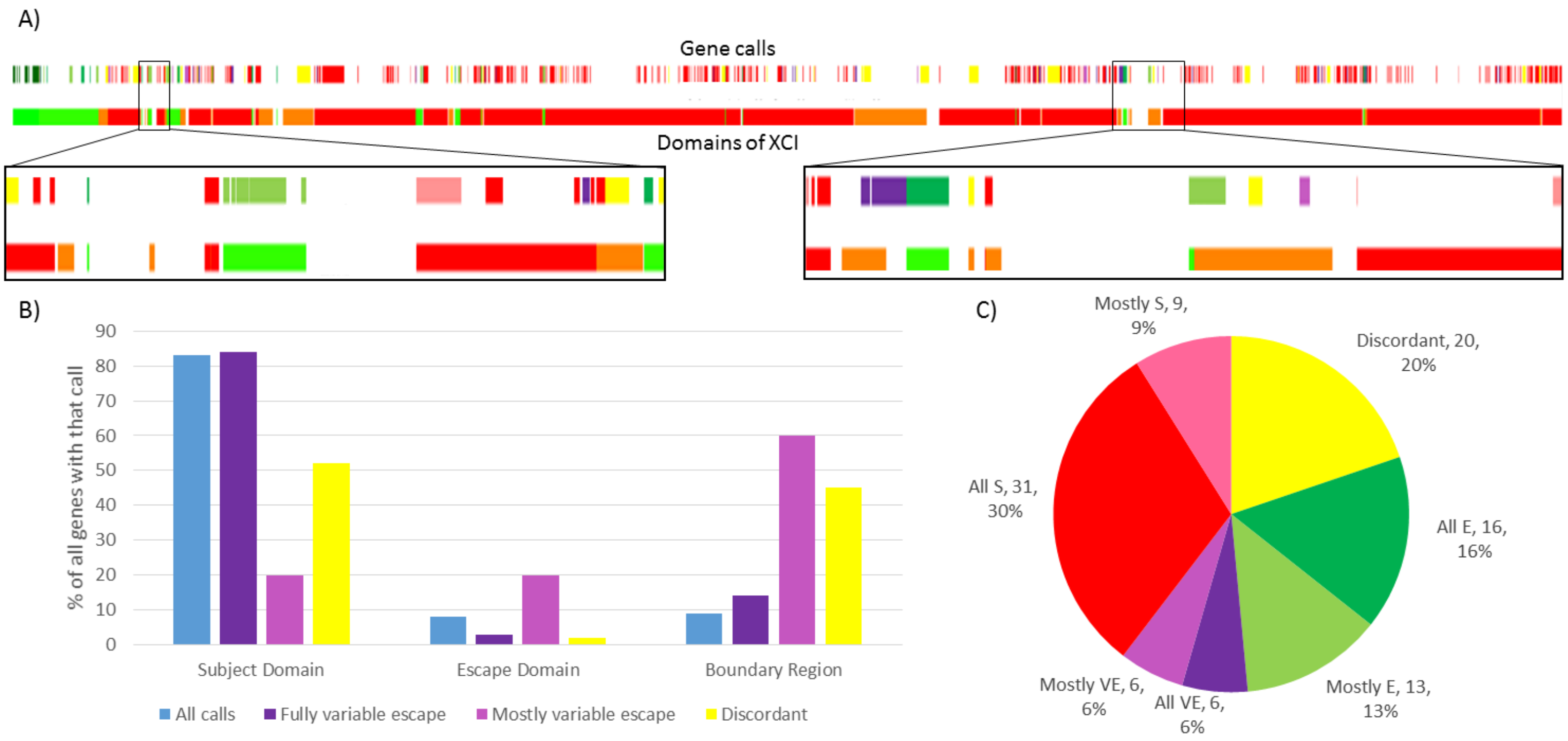

A)
Gene calls
Domains of XCI
B)
% of all genes with that call
0
10
20
30
40
50
60
70
80
90
Subject Domain
Escape Domain
Boundary Region
All calls
Fully variable escape
Mostly variable escape
Discordant
C)
Mostly S, 9, 9%
Discordant, 20, 20%
All S, 31, 30%
All E, 16, 16%
Mostly VE, 6, 6%
All VE, 6, 6%
Mostly E, 13, 13%

Figure 2.4 Domains of XCI and the enrichment of discordant and mostly variable escape genes at boundaries. (A) Our consensus gene calls and the domains of XCI along the X chromosome. The top row is the XCI status calls for all genes with a call on the X while the second row is the domains of XCI called from the consensus calls (see methods). For the XCI status calls, the colors are defined in Figure 2.4C. For the domains of XCI: red is subject, green is escape, orange is boundaries and white space is between domains. A magnification of two regions is shown below, demonstrating how genes line up with domains. Domains are defined by the first and last gene in the domain, even if they start or end inside of other genes which do not share the same domain call. See additional files 1 and 2 in the original publication for the BED files used to generate the UCSC browser track upon which this graph is based. (B) Distribution of genes into XCI status domains. The graph shows what percent of genes with each call are in each domain type. Percent is determined by dividing the number of genes with that XCI status call in that domain type by the total number of genes with that XCI status call. The all calls category includes all genes on the X chromosome, including genes with no calls. (C) Distribution of genes at boundaries. This figure includes the subject and escape genes which define the edges of the boundaries.

standardized residual analysis between the results of other studies and our analysis shows that our study was strongly enriched for calls of fully escape and mostly escape calls when other studies called a gene as escaping from XCI. Our analysis was also strongly enriched for calls of fully subject and enriched for calls of mostly subject and fully variable escape when other studies called a gene subject to XCI. When other studies disagreed with each other, our study tended to call genes discordant.

Another method of examining XCI, using non-CpG methylation (mCH), was recently reported (Lister et al. 2013) and was also compared to our results. Genes called escape by mCH were enriched for the mostly variable escape category while being strongly enriched for the escape and mostly escape categories and depleted for the subject category. Genes called subject by mCH were almost entirely in our subject and mostly subject categories. Another study used mCH to examine XCI across multiple tissue types and found tissue-specific differences (Schultz et al. 2015). Our consensus results were most concordant for genes that escaped XCI across multiple tissues. Together, these comparisons to various calls associated with XCI have shown that the XCI calls presented in our analysis are robust and are relevant to further studies.

2.3.5 XCI status of genes with Y chromosome homology

The X and Y chromosomes were once a homologous pair of chromosomes, and XCI is hypothesized to provide dosage compensation as the Y homologs have decayed. The number of genes escaping XCI is higher on the evolutionarily more recent regions of the X chromosome (Ross et al. 2005), so we compared our consensus calls to which genes have been identified as having Y homologs or Y pseudogenes (Wilson Sayres and Makova 2013). X-linked genes with Y homologs are enriched for genes that escape and mostly escape from XCI (Figure S2.3A). X-linked genes with pseudogenes on the Y are not particularly enriched in any XCI category, although they have less genes with no call (Figure S2.3B). Genes with Y homologs might be anticipated to escape from XCI as having a functioning Y homolog would negate the need for dosage compensation. In addition, these genes could also have been too dosage-sensitive for the stepwise process of upregulation and becoming subject to XCI (Lahn and Page 1999), reviewed in (Veitia et al. 2015). The XCI pattern for genes with Y pseudogenes may be more random, as these genes have had time to evolve XCI. Being enriched for genes with calls may be an artifact due to pseudogenes and XCI calls both being enriched for genes that are better known and well annotated.

2.3.6 Our consensus XCI status calls and sex differences in expression

Genes that escape from XCI tend to not be expressed to the level that is observed from the active X chromosome. A threshold of 10% has been used, and at this level expression from females would only be minimally higher than males; however, expression up to approximately 95% of the Xa has been demonstrated (Carrel and Willard 2005), which would result in sex-biased expression. Recent genome-wide comparisons of expression across multiple tissues (GTEx (Melé et al. 2015)) tested for sex-based expression, and the results correlate well with our consensus calls. Genes with a female expression bias were strongly enriched (*p* value $<10^{-15}$) for the escape and mostly escape from XCI categories. This makes sense as genes which escape have two transcriptionally active copies of a gene in females while only having one in males. Genes with a male expression bias are enriched for being in the PAR1 (*p* value $<10^{-15}$) supporting the theory that there is a minor spread of inactivation into the PAR so that the Y chromosomal copy of the gene has more expression than the Xi copy (Johnston et al. 2008).

2.4 Conclusions

We have compiled a list of XCI status calls from three large studies that used different methodologies. We generated a stringent list in which multiple studies were entirely concordant for subject, escape, or variable categories. We extend those calls with a “mostly” category, allowing single discrepancies. Together, these classifications can be applied to 50 % of genes on the X, including 80 % of all non-CTAG protein-coding genes. Having a reference list of XCI statuses will prove valuable in the future as more research begins to consider sex differences and the effect of having an inactivated X chromosome. This table can be used by researchers to consider the sex effects of their genes of interest or for comparison to larger scale -omics studies such as the GTEx analysis project (Melé et al. 2015). The table can also be informative for the impact of rearrangements, aneuploidies, or copy number variants on the Xi. This XCI status call list will also be valuable for labs such as ours studying X chromosome inactivation. Having a confident XCI status call is needed when attempting to determine patterns across genes with similar XCI statuses or when looking for boundaries between domains with differences in XCI.

Chapter 3: Cross-species examination of X-chromosome inactivation highlights domains of escape from silencing

3.1 Introduction

Human and mouse differ in both the initiation and completeness of XCI (Carrel and Brown 2017; Okamoto et al. 2011). In contrast to human, mouse has imprinted XCI early in development, which is maintained in extraembryonic (trophoblast and primitive endoderm) tissues (Mak et al. 2004; Moreira de Mello et al. 2010; Okamoto et al. 2004). In placenta, rat (Wake, Takagi, and Sasaki 1976) and vole (Shevchenko et al. 2011) also have imprinted XCI while horse/donkey hybrids (X. Wang et al. 2012) and pig (Zou et al. 2019) have random XCI. The story is unclear in cow, where both random (Z. Chen et al. 2016) and imprinted (Xue et al. 2002) XCI have been reported. At the blastocyst stage, human as well as rabbit express XIST from both alleles, while mouse has exclusively paternal Xist expression (Okamoto et al. 2011). Cow has been observed to upregulate XIST at a similar stage to human and rabbit (Yu et al. 2020). Human and rabbit also showed later inactivation timing than mouse (Okamoto et al. 2011). See (Shevchenko et al. 2019) for a review of XCI across species.

Not all genes are subject to XCI, and here again, there is a substantial difference between human and mouse. Escape from XCI is generally defined as having an inactive X expression of at least 10% of active X expression (Carrel and Willard 2005). Around 12% of X chromosome genes are escaping XCI in human (Balaton, Cotton, and Brown 2015), while in mouse the proportion of genes escaping from XCI is only 3-7% (Berletch et al. 2015). In human, an additional 15% of genes variably escape from XCI, differing in their XCI status between different tissues, populations, individuals or studies (Balaton, Cotton, and Brown 2015; Tukiainen et al. 2017). Large-scale studies have not been reported in species outside of human and mouse, and the studies in mouse generally report only on the genes escaping from XCI. The variation between species highlights the importance of studying XCI across a range of species; particularly as the most common model organism, mouse, appears quite different from human.

Knowing the XCI status of genes is important, as genes that escape from XCI often have sex-biased expression, being higher in males if a gametolog is also present on the Y, and higher in females if not (Tukiainen et al. 2017). Furthermore, having two active copies of a gene has been argued to protect females from cancers as both copies will need to be mutated in order to have loss of function (Dunford et al. 2017). In individual species, knowing which genes escape from XCI will be useful for mapping the effect of X-linked genes to various traits, and understanding XCI within a species is important for genomic selection strategies in breeding for agriculture (Couldrey et al. 2017). Additionally, the knowledge of which genes escape from XCI across species can further our understanding of the underlying mechanism allowing some genes to escape XCI and give insight into the evolutionary development of XCI.

Here, we first examined allelic expression and DNAme in human and mouse to establish robust thresholds of DNAme as an indicator of XCI. We then used DNAme data across two separate groups, one of nine different mammalian species, and one of five different primate species, to examine conservation of XCI escape status across species. Finally, we performed analyses testing elements previously seen enriched at genes with various XCI statuses (repetitive elements, CTCF and ATAC-seq peaks) for enrichment with our XCI status calls across species.

3.2 Methods

3.2.1 Xi/Xa expression-based XCI status calls

Human whole genome seq and RNA-seq data was obtained for 11 samples, from the Center for Epigenome Mapping Technologies. This data is from cancer samples, and because cancer has a clonal origin, we anticipated they would show skewing of XCI. Eight of the samples had skewed Xi choice, as could be seen by the majority of genes having an Xi/Xa ratio below 0.1. These samples were from brain, blood, breast and thyroid, however neither of the brain samples had fully skewed Xi choice and could be used in this analysis. Mouse RNA-seq data was obtained from two studies using crosses between two distantly related mouse strains, one of which used an *Xist* knockout to skew Xi selection (Berletch et al. 2015) and another which used fluorescent markers expressed on each X chromosome to separate cells by Xi choice (Wu et al. 2014). These mouse datasets have previously been used to find genes escaping XCI, but most mouse studies do not call genes which are subject to XCI, so they were reanalyzed here.

The different species were processed differently due to different starting file types. The human data was pre-aligned, starting as DNA VCF files and RNA bam files. The DNA VCF files were indexed and then filtered to only heterozygous SNPs in exons using the bcftools view tool (H. Li, 2011). A BCF file was made for the expression data using samtools mpileup with the -t DP,AD options, followed by bcftools filter to filter for depth 30 or higher (H. Li et al. 2009). The RNA BCF file was then indexed and then bcftools call used to find indels and bcftools view used to filter for quality 30+ calls. In mouse, the data was available as fastq files and were aligned using the MEA pipeline (Richard Albert et al. 2018). The resulting unnormalized big wig files were then quantified at known polymorphisms to determine the number of reads on the Xi and Xa. The levels of each allele in the RNA were then extracted using R and compared at all the heterozygous sites found in the DNA analysis (R core Team 2014). The ratio between alleles was used for graphing and the error rate determined using a binomial model with an α of 0.05 (Berletch et al. 2015). Genes were assigned XCI status calls per SNP, with a ratio of 0.1 being used as a threshold between genes escaping and subject to XCI and not giving an XCI status for genes who cross this threshold with their error rates.

SNPs were mapped to splice variants which include the SNP and the closest TSS of these was used to connect DNAme and Xi/Xa expression for Figures 3.1, and supplemental Figures S3.1 and S3.2.

3.2.2 DNAme based XCI status calls

GEO was searched for all WGBS, RRBS or 450k array data that was in eutherian mammals other than mouse and human. Human data was downloaded from the International Human Epigenomics Consortium (IHEC) (Bujold et al. 2016), while a single mouse dataset with a high number of samples was downloaded (Duncan et al. 2018). Data was downloaded for Homo sapiens (human), Pan troglodytes (chimp), Pan paniscus (bonobo), Gorilla gorilla and Gorilla beringei (gorilla), Pongo pygmaeus and Pongo abelii (orangutan), Mus musculus (mouse), Bos Taurus (cow), Ovis aries (sheep), Capra aegagrus hircus (goat), Sus scrofa (pig), Equus ferus caballus (horse) and Canis familiaris (dog). When processed bigwig files were available, they were chosen over processing from raw data. Relevant genomes were downloaded from UCSC (Supplemental Table S3.1) and raw reads were aligned to them using BISMARK (Krueger and Andrews 2011). BISMARK methylation extractor was used to get bedGraph files and then UCSC tools bedGraphToBigWig tool used to make bigwig files. Gene and CpG island maps were downloaded from UCSC, and the UCSC tools bigWigAverageOverBed tool was used to quantify the mean methylation level across CpG islands. R was then used to annotate CpG islands within 2kb of a gene's TSS as belonging to that gene and XCI status calls were made, with islands with a mean DNAme below 10% being called as escaping XCI and islands with between 15 and 60% DNAme being called as subject to XCI. Islands for which over half of males had 15% DNAme or higher were discarded as having male hypermethylation and being uninformative. The mean DNAme across each sex was also calculated and compared per CpG island. The lack of TSSs mapped within each species precluded robust examination of non-CpG island promoter regions, as we were unsure of the exact location of the TSS.

For datasets generated on the human 450k DNAme array, data was downloaded and filtered for promoter associated probes. The mean DNAme of probes sharing an annotated CpG island were matched to their annotated genes and this was used for making XCI status calls as above.

3.2.3 Clustering

XCI calls per species were transformed into numeric values, with escape as 0, variable escape as 0.5 and subject to XCI as 1. The daisy function from the cluster package in R was used to compute distance and then hclust with the gower metric and complete method were used to perform the clustering. The phylogenetic tree was generated using the online interactive Tree of Life tool (Letunic and Bork 2007).

3.2.4 Conservation analysis

R was used to collect and match all the XCI status calls across species. Genes were matched based on their name, controlling only for capitalization changes across species. Genes with XCI status calls in four or more species were included in further analysis. Datasets analyzed were split into two different groups: all mammals (human, chimp, mouse, cow, pig, sheep, and goat WGBS data, with horse RRBS and dog 450k array data) and primates (human, chimp, bonobo, gorilla and orangutan 450k array data). The two separate groups allowed us to examine conservation of genes without our analyses being biased toward primate specific calls.

3.2.5 Statistical tests

Statistical tests comparing enrichment of CpG island statistics and various repeat classes between genes subject to or escaping from XCI were done using R. We used a t-test with the Benjamini Hochberg method for multiple testing correction (Benjamini and Hochberg 1995).

3.2.6 Domain analysis

Domains were identified based on the conservation calls above and examined using the UCSC browser to compare the arrangement of genes. TAD boundaries were taken from (Dixon et al. 2012) and were annotated to genes if they were between it and the next gene or were within the gene body. Additionally, to confirm that UBA1 TSSs were within the same TAD, we used a larger set of TADs in the 3D genome browser (Y. Wang et al. 2018).

3.2.7 ATAC-seq analysis

ATAC-seq data was downloaded, see Supplemental Table S3.1 for data sources. If bigwig files were available they were used, but if not we downloaded raw data and aligned it using HISAT2 (Kim et al. 2019). The bamcoverage tool from the deepTools package (Ramírez et al. 2016) was used to generate bigwig files (normalized using RPKM) and bigWigAverageOverBed from UCSC utilities was used to determine the mean coverage in 250bp up and downstream of each TSS. Each TSS was matched to the closest CpG island within 2kb and any XCI status call from that island used for the TSS.

3.2.8 CTCF predictions

The CTCF binding predictions used here were made by Oriol Fornes of the Wasserman lab for the study this chapter is based off (Balaton et al. 2021). For the purpose of quantifying CTCF binding signal per TSS, we counted the number of bins with an over 50% predicted probability of being a CTCF-bound region within 4kb of each TSS. For analysis of the TCEANC to GEMIN8 region, we counted the number of bins with over 50% probability of CTCF within each region.

3.3 Results

3.3.1 XCI status calls from allelic expression

To obtain DNAme thresholds separating genes escaping XCI from genes subject to XCI, we first needed to establish which genes were escaping versus subject to XCI using allelic expression data. Allelic expression data requires skewed Xi choice and thus was only available for two species: human and mouse (Figure 3.1, Supplemental Figures S3.1, S3.2). Expression-based XCI status calls were determined using a binomial model as previously described (Berletch et al. 2015), with genes having an Xi/Xa expression ratio significantly over 0.1 being called as escaping XCI and those with Xi/Xa significantly under 0.1 being called as subject to XCI. For

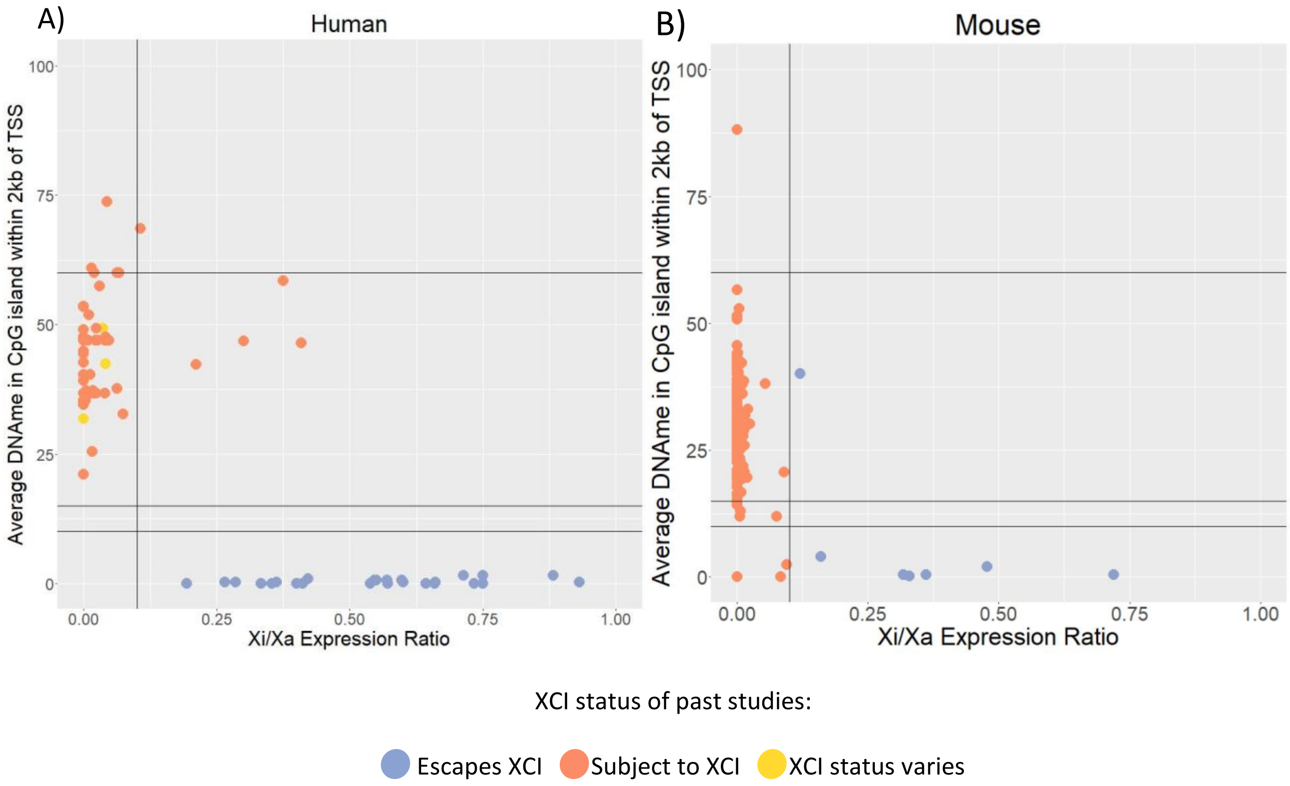


Figure 3.1 Using Xi/Xa expression ratio to establish thresholds of DNAme for XCI status calls. Two species are featured: (A) human (B) and mouse. Each point is a SNP with Xi/Xa expression data, matched to the closest CpG island within 2kb of the closest TSS (accounting for which splice variants would include the SNP) in order to have matched DNAme values. Lines are drawn at 0.1 Xi/Xa expression and at 10, 15 and 60% DNAme as they were used as thresholds to call XCI escape status subsequently. Points are colored based on their XCI status calls. For human, previously published XCI status calls were used (Balaton et al., 2015), while in mouse, which did not have studies calling genes as subject to XCI, they were colored based on their Xi/Xa expression-based XCI status calls featured here. Genes in the pseudoautosomal region, which matches to the Y chromosome, were filtered out. CEMT30, a leukemia cancer sample was used for A, while the Keown et al. data was used for B.

human, we obtained data for eight skewed samples from cancer-related samples and we identified 44 genes escaping XCI, 262 genes subject to XCI and 21 genes variably escaping from XCI in them (Supplemental Table S3.2). We called genes as variably escaping if they had at least 33% of informative samples with each XCI status. The majority of these XCI status calls agreed with previous studies, with discordance for only 53 genes, (17% of genes with an XCI status call in both), 39 of which were reported to variably escape from XCI here or previously in chapter 2 (Balaton, Cotton, and Brown 2015). We attribute the low number of genes variably escaping in our current study to the limited number of samples available and the frequency of informative, heterozygous SNPs per sample, resulting in a mean of 3.5 informative samples per gene. With more samples, we would expect to observe more variably escaping genes.

In mouse we classified 16 genes as escaping XCI, 662 genes subject to XCI and 10 genes variably escaping from XCI (Supplemental Table S3.3). We used three different mouse expression datasets (Berletch et al. 2015; Keown et al. 2017; Wu et al. 2014) and results were 97%, 87% and 90% concordant when datasets were compared with each other. Most of the discordance in our results arise from identifying more genes variably escaping in the Wu dataset than the other two datasets. Additionally, our use of a threshold of 0.1 rather than 0 to call escape from XCI and the inclusion of a variable escape category resulted in more discordant calls relative to those assigned by Berletch. Figure 3.1 shows a clear DNAme difference between genes with an Xi/Xa expression ratio under this 0.1 threshold and genes with an Xi/Xa expression ratio over the threshold.

3.3.2 Establishing thresholds for calling XCI status from DNAme

DNAme data has also been used to call XCI status (Cotton et al. 2015), and is now available from a number of species where expression in individuals with skewed Xi choice is not available. Our search of GEO (Barrett et al. 2013) for DNAme data across eutherian species found datasets with females for 12 different species: human, chimp, bonobo, gorilla, orangutan, mouse, cow, sheep, pig, horse, goat and dog (Supplemental Table S3.1). Most of the datasets used whole genome bisulfite sequencing (WGBS), while horse was limited to a reduced representation bisulfite sequencing (RRBS) dataset and many of the primates and dog were processed on the 450k array, with probes that did not map well to the species in question being filtered out by the source publications. Plots of male versus female DNAme at promoter CpG islands on the X chromosome showed similar trends across species (Supplemental Figure S3.3) with a cluster of sites with less than 10% methylation in both, the bulk of sites showing higher female and low male methylation, and the cluster that is over 70% methylated in both sexes being under-represented on the array data. There are some differences in the amount of male hemi-methylated islands and the female DNAme average across species, which could be due to differences across species or due to the different tissues and methods of assessing DNAme used.

DNAme levels for human and mouse were compared to Xi/Xa expression in order to establish thresholds of DNAme for calling escape from XCI (Figure 3.1, Supplemental Figures S3.1, S3.2). There was good correlation between XCI status calls made using Xi/Xa expression and DNAme with a 10% DNAme threshold. An uncallable zone between 10-15% DNAme was added to lower the chance of miscalling genes, as most discordancies between Xi/Xa expression-based calls and DNAme-based calls had DNAme levels in this range. DNAme at genes subject to

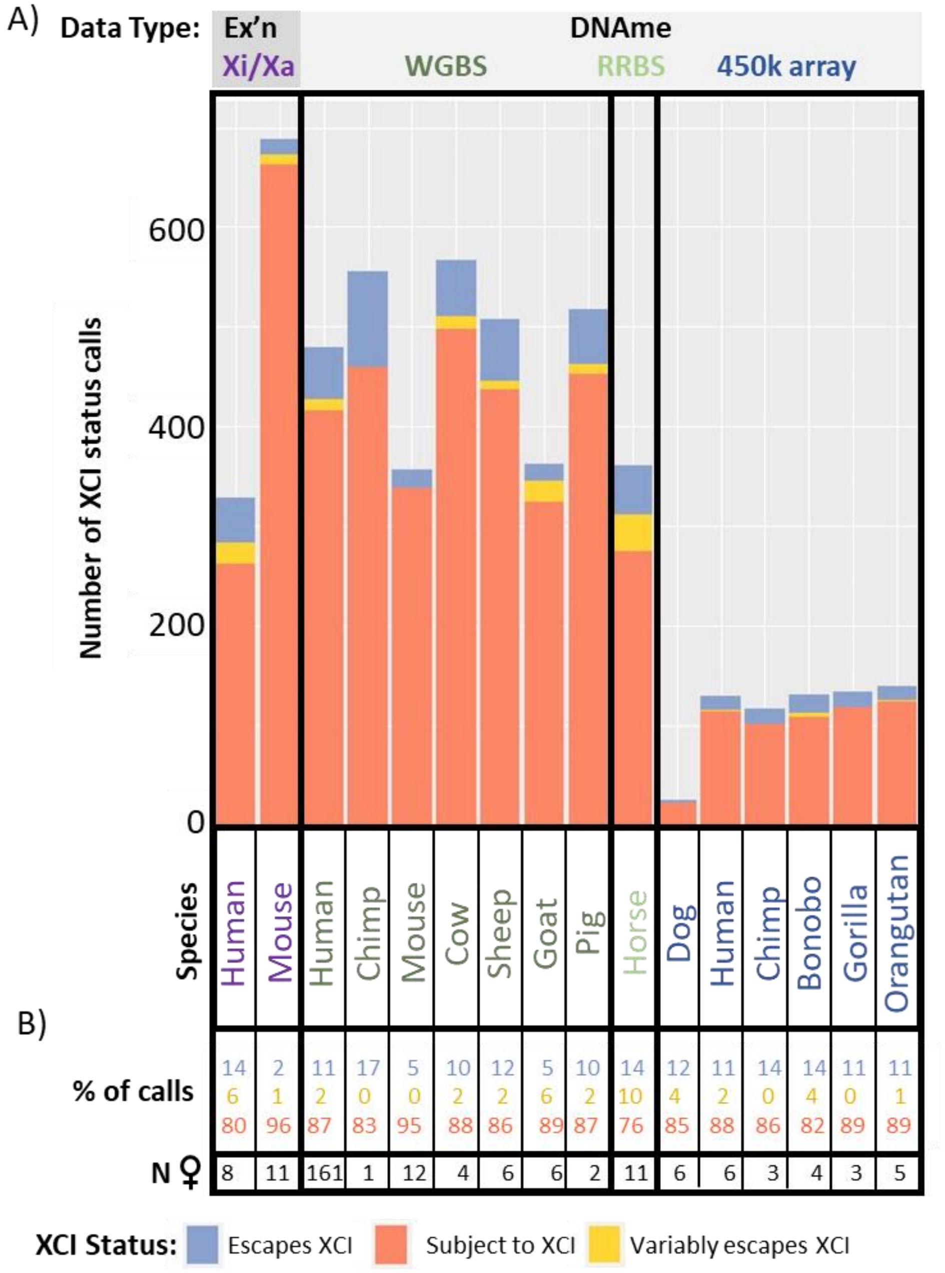
XCI was lower than expected if the Xi was completely hypermethylated, with an average DNAme of 38% and 27% in human and mouse, respectively (Table 3.1). This shows that the DNAme on the Xi is not complete at these CpG islands. Looking at autosomal imprinted genes, the expected 50% DNAme ratio was found, demonstrating that lower methylation is not a problem inherent with this analysis or datasets, rather it reflects the DNAme levels of the Xi (Supplemental Figure S3.4).

Species	Data Type	Average DNAme
Human	WGBS	38%
	450k array	41%
Chimp	WGBS	35%
	450k array	41%
Bonobo	450k array	38%
Gorilla	450k array	39%
Orangutan	450k array	39%
Mouse	WGBS	27%
Cow	WGBS	37%
Sheep	WGBS	31%
Goat	WGBS	33%
Pig	WGBS	38%
Horse	RRBS	37%
Dog	450k array	39%

Table 3.1 Mean DNAme for genes subject to XCI per dataset. The mean DNAme of CpG islands at genes found subject to XCI was calculated per dataset.

3.3.3 XCI status calls from DNAme

Applying our DNAme thresholds across species to make XCI status calls generated between 26 and 567 XCI status calls per species, with a median of 342 calls per species (Supplemental Tables S3.2, S3.3). Most species had 80-90% of genes identified as subject to XCI by DNAme (Figure 3.2), while mouse had 95% of genes subject to XCI and horse only had 76% of genes subject to XCI. The decreased number of genes subject to XCI in horse may be due to the data being generated using RRBS, which provides sparser data and, unlike 450k array data, the sparse CpGs assessed are not the same across samples. In other species the average DNAme at genes subject to XCI ranged from 31% in sheep to 41% in the chimp 450k array data. The 450k array data tended to have higher DNAme than WGBS data, with values between 38 and 41%. Comparison between human and chimp WGBS and 450k array data at the same genes showed that the WGBS and 450k array data differ in DNAme levels, with R^2 values of 0.04 in chimp and 0.59 in human (Supplemental Figure S3.5). Differences may be due to having more CpG sites averaged in the WGBS data. Of the genes that had XCI status calls from both DNAme determining methods, 98% of human genes had the same XCI status calls when analyzed by WGBS or 450k array, as did 92% of chimp genes. The largest impact of using the 450k array instead of WGBS was at genes escaping from XCI, which occasionally crossed the threshold to being called subject to XCI, particularly in chimp, likely due to the low sample size in WGBS (only one sample). Many genes were not assigned a call in one of the datasets as they were hypermethylated. XCI status calls made using our DNAme thresholds were generally consistent however, so we did not discard the 450k array datasets.



Species	Human	Mouse	Human	Chimp	Mouse	Cow	Sheep	Goat	Pig	Horse	Dog	Human	Chimp	Bonobo	Gorilla	Orangutan
% of calls	14 6 80	2 1 96	11 2 87	17 0 83	5 0 95	10 2 88	12 2 86	5 6 89	10 2 87	14 10 76	12 4 85	11 2 88	14 0 86	14 4 82	11 0 89	11 1 89
N ♀	8	11	161	1	12	4	6	6	2	11	6	6	3	4	3	5

Figure 3.2 The number and type of XCI status calls per species.

(A) The number of XCI status calls per dataset and (B) the percentage of calls with each XCI status per dataset are shown. Datasets (columns) were sorted by technique used to generate the data. Species names are colored by the type of data used to generate XCI status calls.

Horse had elevated numbers of variably escaping genes (10%), which was close to that seen previously in human (in chapter 2), while other species (including human) only had 0-5% of genes found variably escaping from XCI. The variation in proportion of variable escape genes seen here could be due to low sample size (in everything except human WGBS), or from our methods of calling variable escape genes being more stringent than in previous studies. We required at least 33% of informative samples to have each XCI status before calling a gene as variably escaping from XCI, similar to the initial survey of human XCI status by Carrel and Willard (Carrel and Willard 2005). Reducing this requirement to only 10% of samples increased the number of variably escaping genes found in human to 63 - almost a quarter of informative genes. These include 37 new genes called which did not have enough informative samples to be called as escaping or subject to XCI with our initial thresholds, as well as 15 genes which changed from an initial call of escaping XCI (12 genes) or subject to XCI (three genes). Although this lower threshold called more genes, we used our 33% threshold of variable escape calls for subsequent studies as we wished to focus on genes that we were confident changed their XCI status between species, rather than differing levels of variable escape from XCI.

Overall, we saw that calls of XCI status using DNAme agreed well with those made using allelic expression and provided an opportunity to examine XCI across multiple species. While WGBS resulted in the most XCI status calls, 450k array DNAme-based calls were generally concordant. These studies showed an average of 11% of genes escaping from XCI across 12 different species, with mouse and goat being an outlier with 95% of genes subject to XCI.

3.3.4 Conservation of XCI status calls across species

XCI status calls per gene were compared across species, focusing on genes that were informative in at least four species. We observed 267 genes being completely conserved across all informative species, with only eight of these genes escaping from XCI and the rest being subject to XCI. Of the eight conserved XCI escapees, two (*DDX3X* and *KDM6A*) have Y homologues across eutherian mammals (Bellott et al. 2014), five have Y pseudogenes in human (*ARSD, STS, PNPLA4, EIF2S3 and MED14*) (Wilson Sayres and Makova 2013), and one has no known Y homology (*CTPS2*) (Figure 3.3A). To avoid biasing the analysis with the more conserved primates, the species were grouped into two groups: primates with 450k array data, and other datasets (including the human and chimp WGBS data). A clear difference in conservation of status was seen between these two groups, with 97% of genes having completely conserved XCI status across primates, while only 75% of genes had conserved XCI status across all mammals (Supplemental Table S3.2). Of the genes which were usually subject to XCI (>75% of informative species subject to XCI), 79% of these had all informative species subject to XCI. Genes that usually escaped from XCI were less concordant, with only 61% of these genes having entirely conserved XCI status across all informative species. A similar trend was seen in the all primates group.

There were 16 genes that varied frequently (2+ species escaping XCI and 2+ species subject to XCI) in the all mammals group and none that varied greatly across primates, again showing the higher similarity in XCI status across closely related species (Figure 3.3). Of these 16 genes, four showed primate-specific escape from XCI (*RPS4X, CDK16, EIF1AX and GEMIN8*) and one showed artiodactyla-specific (cow, sheep, goat, pig) XCI (*KDM5C*). The pattern of conservation

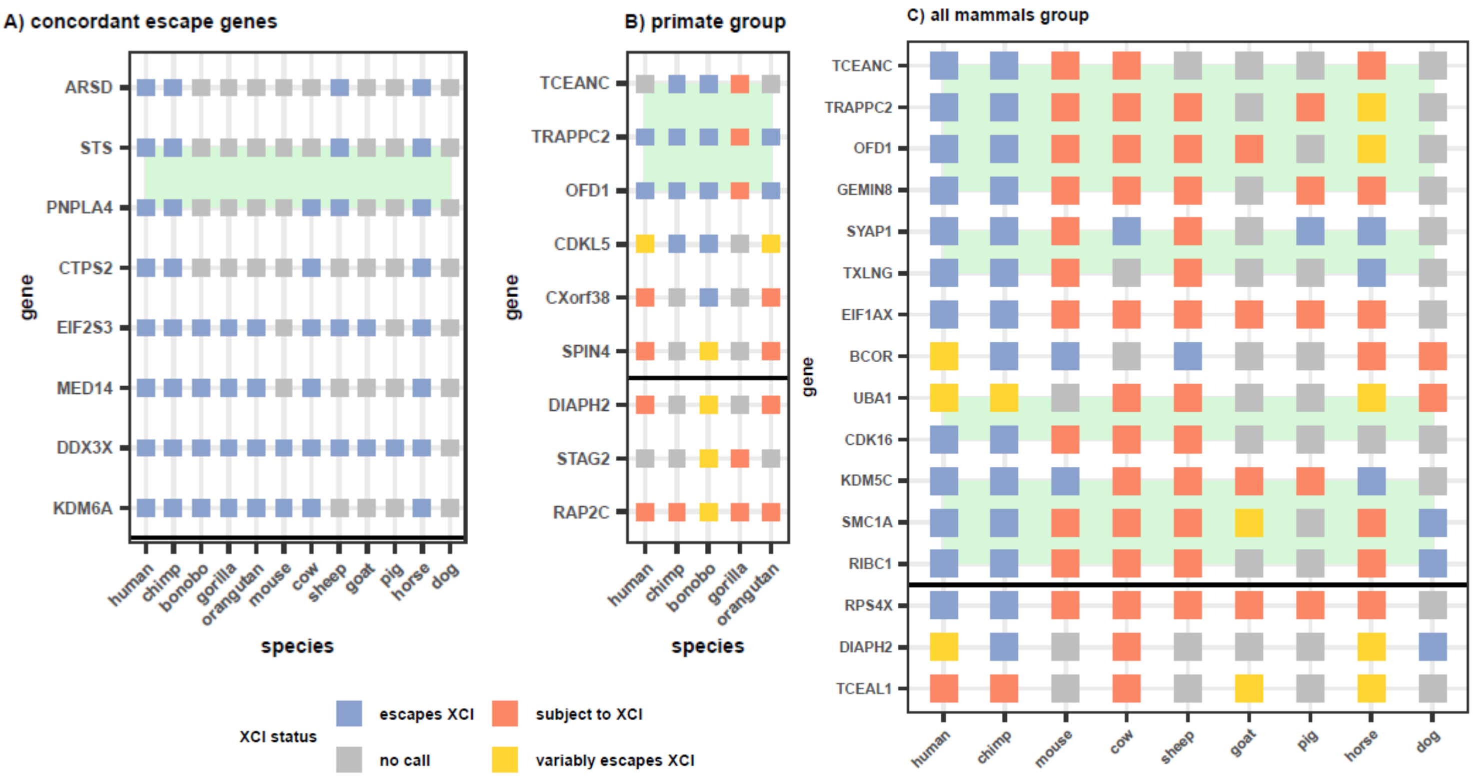


Figure 3.3 Concordant and discordant escape genes across species.

(A) Eight genes escape XCI in all informative species, while 259 genes were subject to XCI in all informative species (not shown). Discordant genes in two different groups of species were examined, (B) only primates and (C) all mammals (limited to only 2 primate species). The intersection of a gene and species is colored based on that gene's XCI status call in that species. Genes that did not have an XCI status call in a species are colored grey. Only escape genes informative in at least 4+ species were selected for A. Genes were selected for B if they had at least one discordant primate species while genes in C required two XCI statuses with two or more species. To match best across species within groups, 450k array data was prioritized in B and WGBS data was prioritized in C. Genes are organized based on their position on the human X chromosome with a horizontal black line denoting the centromere. Green boxes highlight domains of adjacent genes with similar changes to XCI statuses across species.

of the other genes variably escaping across species did not match any phylogenetic patterns. The primate-specific escape genes *RPS4X* and *EIF1AX* have been shown to have primate-specific retention of their Y homolog while *KDM5C*, the gene that is subject to XCI only in artiodactyla has lost its Y homolog in bulls, while retaining it in mouse and primates (Bellott et al. 2014). We show the WGBS data surrounding the CpG island at the transcription start site (TSS) of the ubiquitous escape gene *KDM6A*, the artiodactyla-specific subject gene *KDM5C* and the primate-specific escape gene *RPS4X* (Figure 3.4).

CDKL5 was the only gene seen to have more than one discordant species in primates (Figure 3.3B), being subject to XCI in the human WGBS data, variable in orangutan and the human 450k array data and escaping in chimp and bonobo. In gorilla, *CDKL5* appeared subject to XCI, but half of the data was in the uncallable region between 10 and 15% DNAme so it was not called as subject to XCI. Other genes had only one species of primates discordant from the rest, usually gorilla or bonobo.

3.3.5 Role for alternative promoter usage in escape from XCI

UBA1 was particularly interesting as it has been shown previously in human to have two different TSSs with differing XCI statuses (Goto and Kimura 2009). This pattern of multiple TSSs with differing XCI status was seen also in chimp and horse (although data is sparse in horse) (Figure 3.5). In cow, the upstream TSS and CpG island are not annotated, but the region homologous to the human upstream TSS showed a DNAme pattern consistent with a promoter subject to XCI, and in pig the CpG islands are annotated but the gene is not. Similarly, in mouse both TSSs (which are annotated but lack CpG island definition) had female-specific DNAme. Mouse has been shown to have fewer CpG islands than human, with CpG island loss from the ancestral genome being four times as high in mouse as human (Jiang et al. 2007). The island is

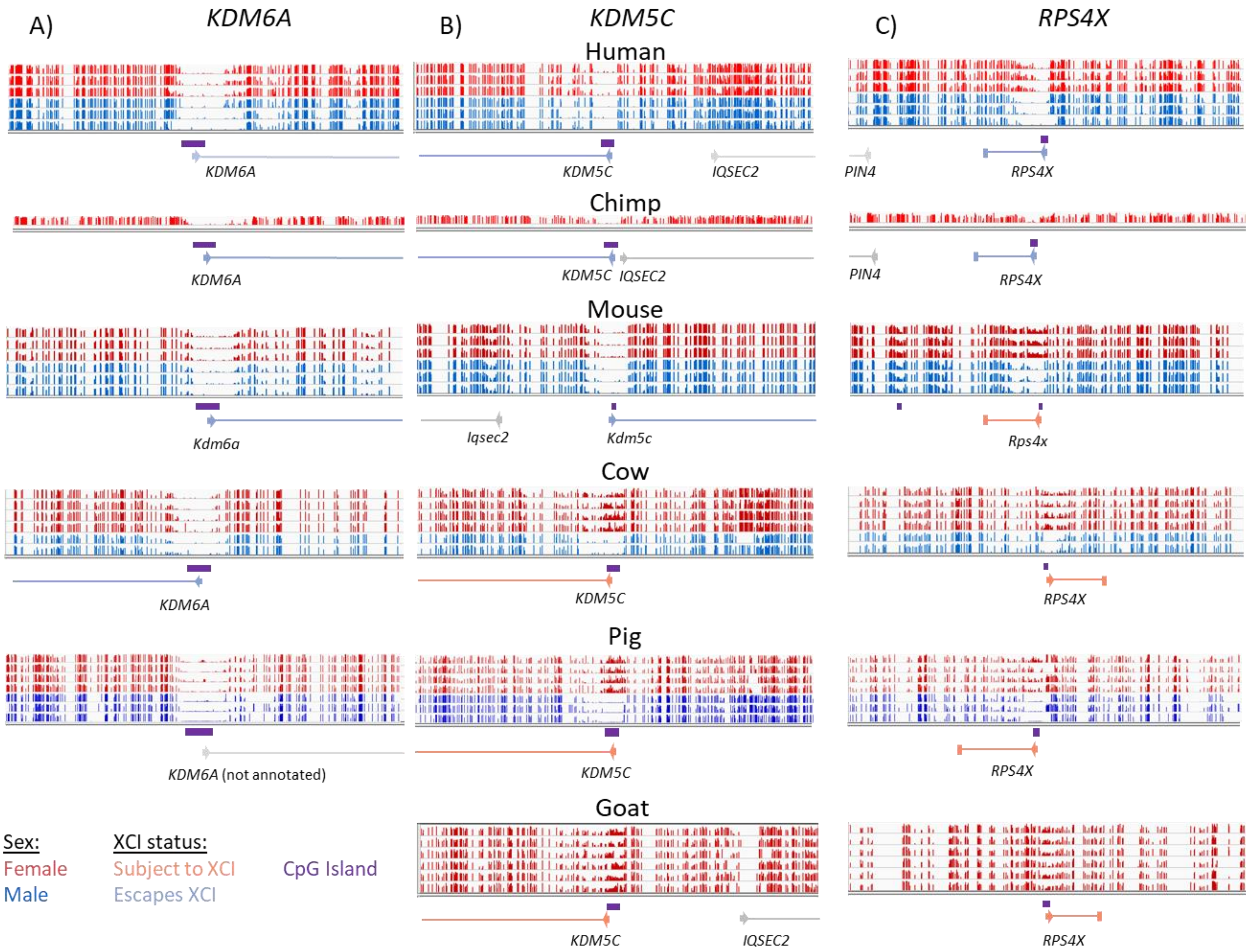

A)
KDM6A
B)
KDM5C
C)
RPS4X
Human
KDM6A
KDM5C
IQSEC2
PIN4
RPS4X
Chimp
KDM6A
KDM5C
IQSEC2
PIN4
RPS4X
Mouse
Kdm6a
Iqsec2
Kdm5c
Rps4x
Cow
KDM6A
KDM5C
RPS4X
Pig
KDM6A (not annotated)
KDM5C
RPS4X
Goat
KDM5C
IQSEC2
RPS4X
Sex:
Female
Male
XCI status:
Subject to XCI
Escapes XCI
CpG Island

Figure 3.4 Featured genes compared across species.

Male and female DNAme values are graphed by gene and dataset. (A) *KDM6A* is featured as it is concordantly escaping across species. (B) *KDM5C* is featured because it is known to escape XCI across species but is here shown to be subject to XCI in artiodactyla (cow, sheep, pig and goat). (C) *RPS4X* is featured because it is a well known primate-specific escape gene. Male methylation is shown in blue and female in red. Annotated CpG islands are shown under the methylation data in purple. Genes are shown as arrows colored at the TSS pointing in the direction of transcription, colored by their XCI status. All of the methylation data shown is from WGBS. Pig did not have KDM6A annotated, but predictions from other species show it located at this CpG island. Goat did not have a CpG island or lowly methylated region at the annotated KDM6A.

still large enough to see low methylation on the Xa so the cutoff for minimum island size may be too high in some species. Overall, the alternative TSSs are conserved across species; however, the XCI status of the downstream TSS changes from escaping from XCI in human, chimp and horse to being subject to XCI in mouse and cow. In humans, both TSSs were always found within the same TAD and sub-TAD. Examining TSS usage in the other genes featured in Figure 3.3C, we were able to map the TSS and CpG islands using either the University of California Santa Cruz Genome Browser (UCSC) (Kent et al. 2002) for that species or using the UCSC liftover tool across species, suggesting that the change in XCI status across species was not due to differences in TSS usage between species.

3.3.6 Domains of escape from XCI across species

Looking at the position of genes escaping XCI along the human X chromosome, we saw that most genes escaping XCI clustered into domains on the short arm of the X chromosome, similar

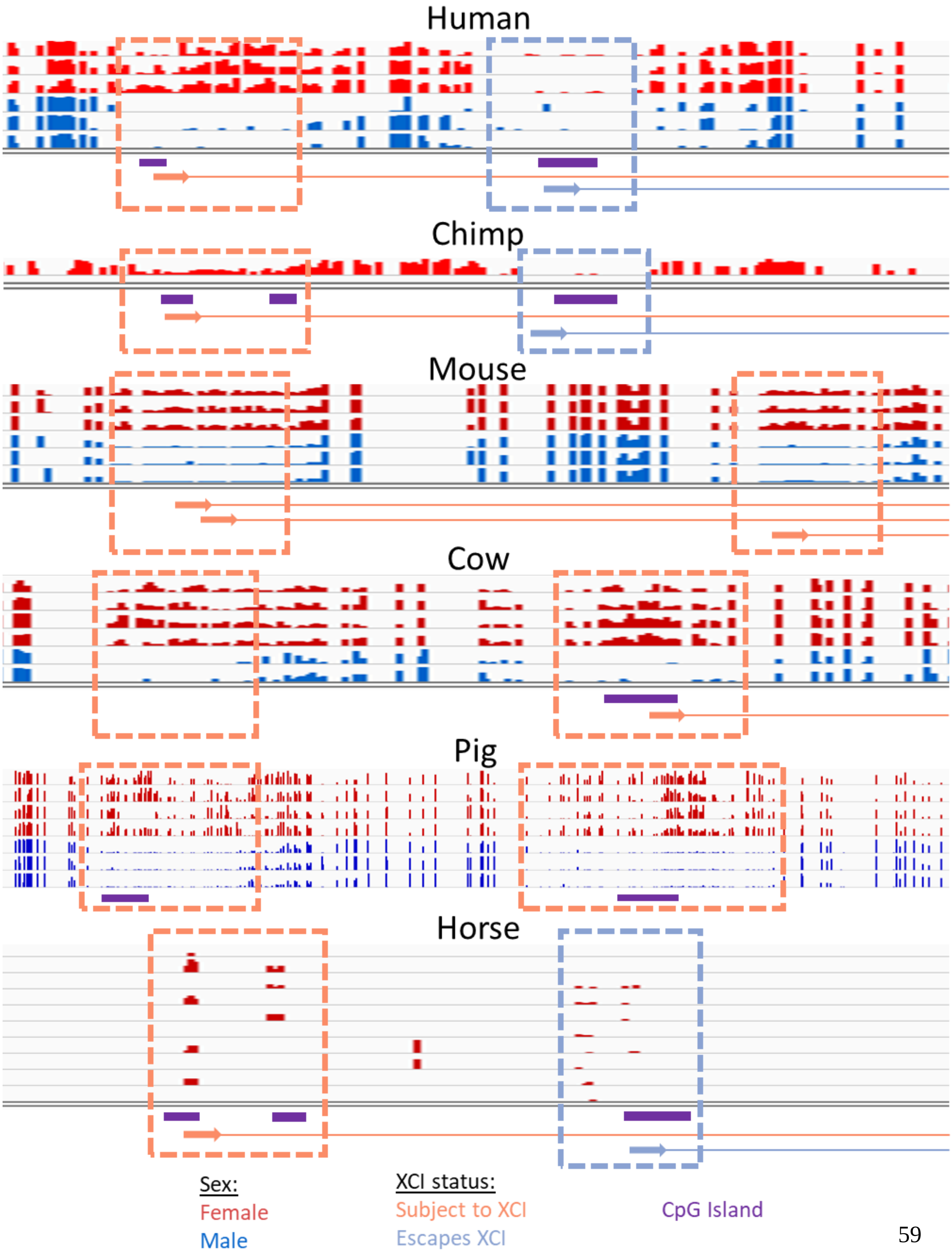

Human
Chimp
Mouse
Cow
Pig
Horse
Sex:
Female
Male
XCI status:
Subject to XCI
Escapes XCI
CpG Island

Figure 3.5 DNAme across the variably escaping gene UBA1.

UBA1 is featured as it has multiple different TSSs with CpG islands that have different XCI statuses. Male methylation is shown in blue and female in red. Annotated CpG islands are shown under the methylation data in purple. Genes are shown with arrows at the TSS pointing in the direction of transcription, colored by the TSSs XCI status. All the methylation data shown, except for horse is from WGBS. Horse used RRBS data, which is why the data is so sparse.

to what has been described previously (Carrel and Willard 2005). Ten of the 23 transitions between clusters of genes escaping or variably escaping from XCI and genes subject to XCI fell near TAD boundaries in human (Dixon et al. 2012), again similar to what has been seen previously (Marks et al. 2015). These clusters of genes escaping from XCI often matched across species. Genes discordant in more than one species were also often clustered, while the genes discordant in only one species were generally scattered by themselves. Some of the genes within discordant clusters were not featured in Figure 3.3 as they were missing data in some species. Only two of the strongly discordant genes featured in Figure 3.3 are located on the long arm of the X chromosome and they did not form a cluster.

We investigated these domains of changing XCI status further by examining whether the discordant species had altered the chromosomal arrangement of these genes. For the primate-specific region of genes escaping XCI spanning the genes *TCEANC* to *GEMIN8*, most species had the same gene order, orientation and flanking genes as observed for human (Supplemental Figure S3.6), although some small changes were observed in gorilla, mouse, cow and sheep. In human and mouse, the two species with Hi-C data, there is a TAD spanning from *EGFL6* (which neighbors *TCEANC*) to *GEMIN8*, which may coordinate the regulation of this region, although if regulated as a domain, *EGFL6* would be expected to also escape XCI in primates. There was no data here giving an XCI status for *EGFL6*, but a previous study had seen it as subject to XCI in human (Cotton et al. 2013). Gorilla was the only primate that did not demonstrate escape from XCI across this domain, with only the gene GEMIN8 escaping XCI. A small insertion was present in gorilla, but it was outside of the TAD which casts doubt about whether it could be the cause of this discordance from the other primates. None of the structural differences in this region were conserved across species with concordant XCI status; thus, we found no detectable genomic correlate underpinning the change in XCI status. Similar results were found for the other discordant regions.

3.3.7 Correlation of features with XCI status across species

These genes that transition their inactivation status across species provided a dataset to interrogate for factors underlying establishment of silencing or escape from silencing. We considered various factors pertaining to CpG islands in addition to enrichment of various classes of DNA repeats. No differences were seen in CpG island size, nor CpG and GC content between species with discordant XCI status at specific genes. Differences in islands between all genes escaping from versus subject to XCI per species were seen in some species, but no characteristic was seen to be significant after multiple testing correction or in more than one species. Different classes of repeats were tested for correlation with genes escaping from versus subject to XCI in human, chimp, mouse, cow, sheep, pig and horse (Supplemental Table S3.4). There were significantly more LINE repeats within 15kb upstream of genes subject to XCI than for genes escaping from XCI in chimp, mouse, sheep and horse (Figure 3.6A, Table S3.4, t-test, corrected p- values<0.01). Other repeat classes found enriched across multiple species include

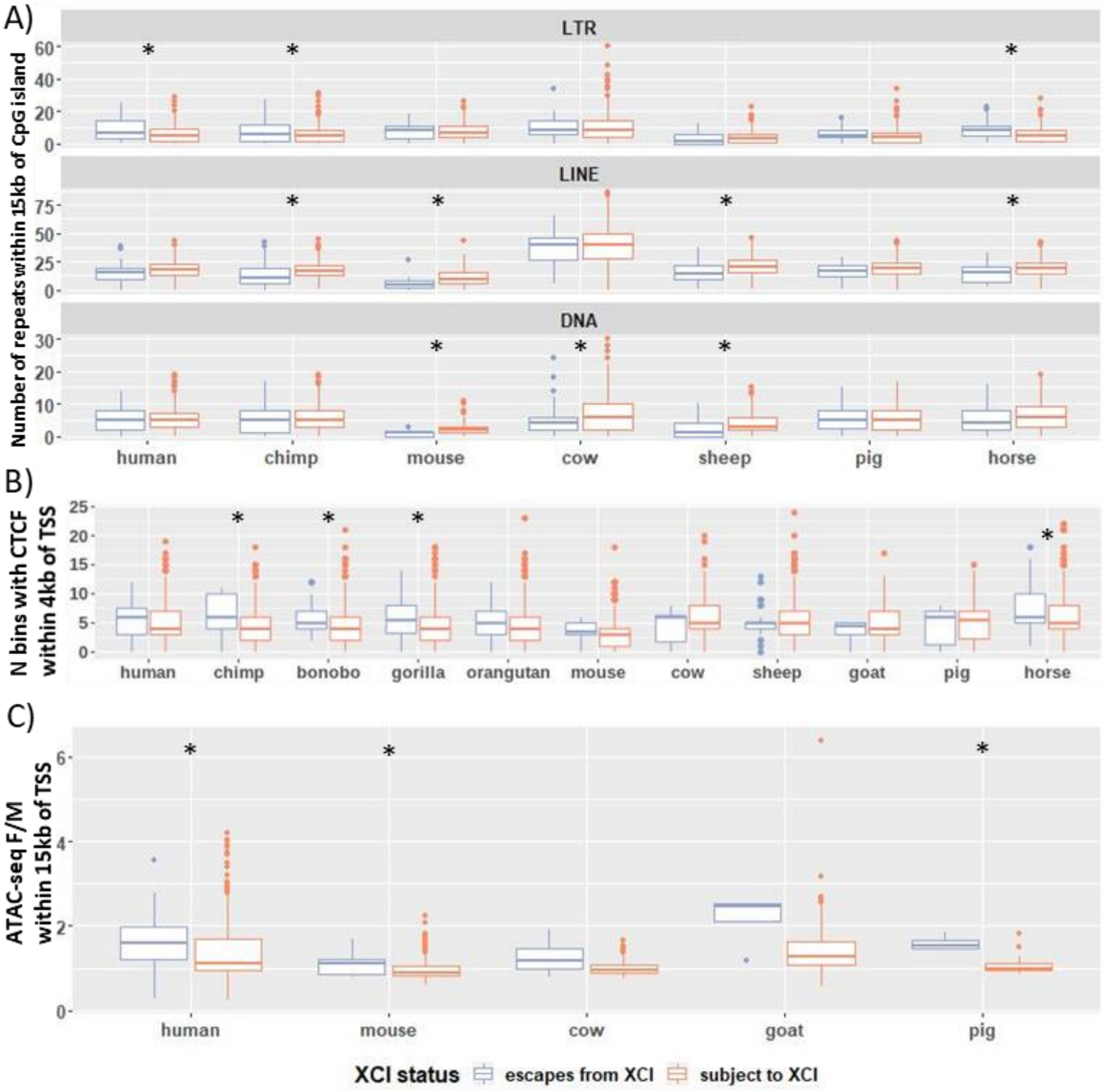


Figure 3.6 Enrichment of elements which may be related to XCI status.

(A) The number of repetitive elements of each class within 15kb of each CpG island, sorted by XCI status. See Supplemental Figure S3.7 for the repeat classes not shown here. (B) CTCF binding in overlapping 200 bp bins was predicted using a DanQ model (Quang and Xie 2016). The Y axis shows the number of bins with >50% predicted probability of having CTCF binding within 4kb of each TSS. (C) Female/male ATAC-seq signal averaged across samples within 250bp of each TSS. F/M is female over male. Species with a * have significant differences between TSSs escaping XCI and those subject to XCI (t-test, adjusted p-value<0.01). P-values are listed in supplemental table S3.4, along with the number of CpG islands or TSSs per XCI status used for each species in the analysis.

LTR, DNA and snRNA repeats, which were enriched at genes escaping XCI in 3 species (Supplemental Figure S3.7). SINE repeats, which have previously been seen enriched at genes escaping from XCI (Cotton et al. 2014), were only found significant in horse, which unexpectedly had more SINE repeats near genes subject to XCI than at genes escaping from XCI. Human still had more SINE repeats near genes escaping XCI than subject to XCI on average, but this difference failed to reach significance in this study.

We compared CTCF binding signal between genes found escaping vs subject to XCI across species. For this, we predicted the probability of CTCF binding across species by using a DanQ model (Quang and Xie 2016) trained on human CTCF ChIP data from ENCODE (Davis et al. 2018) and validated on mouse. There were significant differences in the amount of CTCF binding signal within 4kb of TSSs escaping vs subject to XCI in chimp, bonobo, gorilla, and horse but not in human, gorilla, mouse, cow, sheep, goat or pig (Figure 3.6B, Supplemental Table S3.4, t-test, corrected p- values<0.01). All the species with significant differences had more CTCF binding signal near genes escaping XCI. We also examined whether there were regions in the TCEANC to GEMIN8 cluster of discordant genes which correlated with a change in XCI status across species but did not find any differences consistent across species (Supplemental Table S3.5).

Comparing ATAC-seq signal 250bp up and downstream of TSSs across species revealed significant differences in the mean female/male ratio across genes that were escaping vs subject to XCI in human, mouse and pig but not in cow or goat (Figure 3.6C, Supplemental Figure S3.4, t-test, corrected p- values<0.01). ATAC-seq signal had a higher female/male ratio in genes escaping XCI than genes subject to XCI, as seen previously in human (Qu et al. 2015), and the same trend existed in species where the differences failed to reach significance. In the species with significant differences in ATAC-seq signal with XCI status, we did not see all tissues showing significant differences (Supplemental Figure S3.8). The differences were significant in the only tissue examined in human, two of the three examined in pig, and one out of ten examined in mouse.

Across all species examined, mouse genes appeared uniquely well-silenced. We clustered all species based on their XCI status calls (Supplemental Figure S3.9). The bovids (cow, sheep and goat) as a group clustered together, although mouse clusters with them for an unknown reason. Dog has very sparse data which may explain it clustering as an outlier, but we are unsure of the reason why pig clustered with dog instead of with the more closely related bovids. We observed clear separation of the primates from most other species due to the large number of primate-specific escape genes.

3.4 Discussion

Escape from XCI is an important contributor to sex differences in expression and has even been argued to underlie a male predisposition to cancer (Dunford et al. 2017; Tukiainen et al. 2017). In addition, genes subject to XCI can also have unique effects on phenotype, with some mutations having phenotypic effects only when separate cell populations are expressing two different alleles (CenterWall and Benirschke 1975; Twigg et al. 2013). Mutations that are deleterious at the cellular level or affect the region controlling choice of Xi can lead to skewed

Xi choice, leaving the individual vulnerable to recessive mutations on the opposite X chromosome (Mitterbauer et al. 1999; Naumova et al. 1998). Knowing the XCI status of genes is also important for estimating the effect of an X-linked allele in genome- or epigenome-wide association studies (B. Chen, Craiu, and Sun 2018; Xu and Hao 2018) and is important for genetic selection of X-linked genes in agriculture (Couldrey et al. 2017).

To validate our use of DNAme to call XCI status, we compared expression-based calls with DNAme in human and mouse. The human Xi/Xa expression-based calls had 83% agreement with previous calls, with the discrepancies largely in genes variably escaping from XCI (Balaton, Cotton, and Brown 2015). As cancer samples were used to allow Xi/Xa analysis, some epigenetic dysregulation may have occurred (Larson et al. 2017). We took human DNAme data from IHEC which included multiple consortia, one of which was mostly cancer samples while the other two were not. DNAme based XCI status calls were quite similar between the consortia with only one gene being called as escaping in one consortium and subject to XCI in another (Table S6). Our study was further limited by the need for heterozygous polymorphisms, thus with only 8 samples, any mis-regulation may not have been noticeable, or led to false or missed calls of variable escape from XCI. Our human DNAme calls were 94% (WGBS) and 91% (450k array) concordant with previous XCI calls, and the two datasets analyzed here gave calls that were 97% concordant with each other. Of the few XCI status calls that were inconsistent with previous studies, 80% were in genes called as variably escaping from XCI, and are likely due to differences in the population or tissues sampled. While our mouse Xi/Xa expression-based calls had a median 90% concordance across datasets, we only identified 60-86% of previously identified mouse escape genes, likely due to differences in thresholds between studies. There were no discordancies between our mouse DNAme calls and previous mouse studies; however, the genes discordant between our Xi/Xa expression calls and previous mouse studies were not informative in our DNAme calls due to lack of CpG islands. Comparing our mouse DNAme calls to a previous study by Keown et al., which examined DNAme on the X chromosome in mouse brain, revealed no discordancies in genes called as escaping XCI, but there were differences in which genes were informative (Keown et al. 2017).

In this study we have made an average of 342 XCI status calls per species, for 12 different species. The proportion of genes subject to XCI differs, with most species having 80-90% of genes subject to XCI. The only species with more genes subject to XCI is mouse at 95%, and the only species with fewer was horse at 76%. Additionally, horse had elevated numbers of genes variably escaping from XCI (10), while other species only had 0-5% of genes variably escaping from XCI. A meta-analysis in human found 8% of genes variably escaping from XCI and a further 7% varying between studies (Balaton, Cotton, and Brown 2015), while our current study identified 6% variable escape in human by expression and only 2% by DNAme. Our analysis is consistent with a previous study using DNAme to make XCI status calls that did not see many genes consistently variably escaping from XCI (Cotton et al. 2015). Of the genes previously predicted to variably escape from XCI (Balaton, Cotton, and Brown 2015), 69% had no data in this study due to lack of a CpG island and another 10% were hypermethylated in males or females and therefore XCI status could not be determined.

Our DNAme analysis found that human genes subject to XCI have promoter CpG DNAme between 38% (in WGBS) and 41% (in 450k array analysis) which agrees with a previous analysis using the 450k DNAme array which showed genes subject to XCI having an average DNAme around 40% (Cotton et al. 2015) (Table 3.1). Mouse had a lower 27% DNAme average for genes subject to XCI; other mouse studies have not examined genes which are subject to XCI. Other species had DNAme averages in a range between human and mouse, but most were closer to human than mouse. Our DNAme thresholds to call genes as escaping from or subject to XCI were consistent across human and mouse WGBS but, as our data was from different studies using different techniques on different tissues in different species there may be variation unaccounted for with our thresholds. However, WGBS and 450k array-based XCI status calls were consistent in both human and chimp and, with a few notable exceptions, genes had concordant XCI status calls across species. Past studies of XCI status calls using DNAme in human did not see many differences in DNAme-based XCI status across tissues (Cotton et al. 2015), so different tissues analyzed may not cause many discordancies. Having male DNAme as a control and an upper threshold for calling genes as subject to XCI should reduce the chance of calling a gene as subject to XCI if it is instead silenced on both copies of the X in a tissue-specific manner. For the primate and dog samples which used the human 450k DNAme array, only probes which mapped consistently between the species were kept by the source publications (Epiphanio et al. 2019; Hernando-Herraez et al. 2013), and so these species may be enriched for genes with a conserved XCI status. Utilizing datasets from different studies confounds the species differences with other experimental differences including sample size as well as inclusion of male samples. The lack of male samples in some species prohibited us from filtering out genes that are methylated on the Xa and therefore would never be seen to escape XCI by DNAme.

Many of the genes escaping from XCI have previously been seen grouped in domains (Marks et al. 2015), and here we see these domains conserved across species. Furthermore, we see that many of the genes that change XCI status across species are clustered into domains and many of these domains coincide with TADs in human. These domains suggest escape from XCI may be regulated at a domain level; however, we also see some genes being regulated individually and even separate TSSs for the same gene can have opposite XCI statuses. Individual escape genes are often discordant in a few species. Coincidence of changes in XCI status with loss of Y homology emphasizes the importance of dosage for determining genes whose escape from XCI is vital to survival. Generally, the TSS is seen to be conserved, even when a gene changes XCI status. Previous studies have suggested that CTCF and YY1 may be enriched near genes escaping from XCI (Berletch et al. 2015; C. Y. Chen et al. 2016; Giorgetti et al. 2016). CTCF has also been seen enriched at boundaries between domains of genes with opposite XCI statuses (Filippova et al. 2005). Repeat elements (SINE for genes escaping XCI and LINEs for genes subject to XCI) have also been seen enriched in 100kb windows around TSSs as well as windows 15kb upstream (Cotton et al. 2014; Z. Wang et al. 2006).

Our XCI status calls across species also allow us to check conservation of elements that may control XCI. A region escaping XCI in human was still able to escape from XCI when inserted at a mouse region which is normally subject to XCI, showing that the mechanisms controlling escape from XCI are conserved and functional across species (Peeters et al. 2018). We suspect that any elements found to be important in human or mouse research will be conserved across species with the same XCI status; having a variety of mammalian species with XCI status calls gives us a platform to test this hypothesis.

We compared DNA repeats and CpG island characteristics with XCI status within and across species and found none varied significantly across species per discordant gene, few varied between XCI statuses within a species and none varied between XCI statuses in all species. Previous studies have examined enrichment of repetitive elements across differently sized regions ranging from 15kb to 100kb. The enrichment closer to the promoter may reflect gene-specific control whereas enrichment across a broader range suggests regulation at the level of domains. These studies have seen enrichment of LINE and LTR MLT1K repeats at genes subject to XCI and SINE and MER33 repeats at genes escaping from XCI (Cotton et al. 2014; Z. Wang et al. 2006). Here, with a window of 15kb, we replicated the enrichment for LINE repeats, with SINE repeats failing to reach significance, and LTR and DNA repeats (which MLT1K and MER33 belong to) showing the opposite trend of previous studies. However, no element was consistently found across all species. We also predicted CTCF binding and observed that some species have more CTCF binding signal around genes escaping XCI than genes subject to XCI as has been seen previously (Berletch et al. 2015; C. Y. Chen et al. 2016; Giorgetti et al. 2016). ATAC-seq signal, which has previously been seen enriched at genes escaping XCI, was also seen enriched here, but again, only in some species (Qu et al. 2015). A deeper bioinformatic analysis comparing our XCI status calls to features which differ across species with differing XCI status but are conserved in species with conserved XCI status might identify important regulatory features which control the XCI status of nearby genes or control XCI in general.

These XCI status calls may be improved in the future through new techniques such as single-cell RNA-seq (scRNA-seq) which can make expression-based XCI status calls without the need for samples with skewed Xi choice. Cells can be analyzed individually, or their Xi choice can be identified and then all of the cells with the same Xi can be pooled. scRNA-seq has also identified variable escape at the cellular level within a tissue (Tukiainen et al. 2017), with most genes XCI status per cell based on which X was the Xi in that cell and one gene (*TIMP1*) seen to vary randomly with no observed difference in Xi choice between cells with different XCI status. Current scRNA-seq datasets have a limitation of low read depth per cell, which limits the ability to examine lowly expressed genes (Moreira de Mello et al. 2010). Methods to enrich for the 3' end of genes, such as the Chromium Next GEM Single Cell pipeline, are useful for quantifying expression per gene but further limits the number of polymorphisms available for study. As sequencing becomes cheaper and scRNA-seq technology continues to develop, scRNA-seq may become the new gold standard for making XCI status calls.

Non-CpG DNAme may allow us to use DNAme to examine XCI status in genes without CpG islands, as this mark is seen enriched in the gene body of transcribed genes (Lister et al. 2013). Brain and pluripotent cells have the most abundant non-CpG DNAme, with other tissues having less than 1% non-CpG DNAme (He and Ecker 2015). A study across multiple tissues in human found 18% of genes (109 of 612) had female-specific non-CpG DNAme in at least one tissue, but of these 66% (72 genes) were only significant in one tissue (usually brain) (Schultz et al. 2015). Another study, in brain only, found 20% of genes escaping from XCI (Lister et al. 2013). These numbers are higher than other reports of escape, likely due to many of these genes variably escaping from XCI and only escaping from XCI in brain.

Improved gene and genome annotations in some of the less well-studied species would enhance our XCI status calls across species. Many of the species examined here had gene annotations generated bioinformatically using CESAR (Sharma, Elghafari, and Hiller 2016) mapping of human genes instead of being annotated with mRNA from that species. This may not have captured the correct TSS, and if transcription was no longer close to the same CpG island these XCI status calls would be invalid. With better annotations in the future, these datasets could be reprocessed to provide more up-to-date XCI status calls with improved confidence.

As mouse has considerably fewer genes escaping from XCI than other species, there may be a better species to use as a model for research related to which genes escape from XCI. Unfortunately, none of the species other than mouse examined here are small or make affordable model systems. Rabbit, for which there was no DNAme data available, has been shown to be more similar to human than mouse in aspects of XCI and may be a good species for further examination (Okamoto et al. 2011).

3.5 Conclusions

Our study created reference XCI status calls for 12 species, so that labs working with diverse mammalian species will have improved understanding of how their genes of interest are expressed in their species of interest. We have confirmed that mouse has substantially fewer genes escaping from XCI than human and shown that other mammals are more similar to human in this regard. Additionally, we showed conservation of XCI status across the majority of X-linked genes and highlighted some genes of interest which are discordant across species. Interestingly, many of these discordant genes occur in domains of similarly regulated genes. In the future, we hope to use these XCI status calls to identify elements which are controlling escape from XCI and which are conserved across species, and these discordant genes are ideal candidate regions to investigate.

Chapter 4: Contribution of epigenetic changes to escape from X-chromosome inactivation

4.1 Introduction

Relatively little is known about how a gene can be expressed from the midst of heterochromatin. The factors determining XCI status remain unresolved with evidence from previous chapters suggesting regional control, but there are also lone genes that escape XCI while flanked by genes subject to XCI (Balaton, Cotton, and Brown 2015) and even genes with two TSSs with opposite XCI status (Balaton et al. 2021; Goto and Kimura 2009). Beyond the direct examination of allelic expression, the modifications to DNA and chromatin that accompany XCI can be used as surrogates to determine if a gene is inactivated. For these features it is unclear if the mark enables or reflects XCI status. As noted earlier, promoter DNAme at CpG islands is an epigenetic mark which is strongly predictive of a gene's XCI status and has been used to differentiate genes which escape XCI from those subject to XCI without the need for heterozygous SNPs or skewed Xi choice (Cotton et al. 2015). Other epigenetic marks such as histone marks have been reported to be correlated with a gene's XCI status. Active marks such as H3K4me2/3, H3K9ac, H3K27ac, H3K9me1, RNA polymerase II and transposase accessibility are enriched at genes escaping from XCI, while inactive marks such as H3K9me3, H4K20me3, H3K27me3 and macroH2A are enriched at genes subject to XCI (Balaton and Brown 2016; Kucera et al. 2011; Qu et al. 2015), reviewed in chapter 1 and (Balaton and Brown 2016). A predictive model using many epigenetic and genetic features in mice was able to predict a gene's XCI status accurately 78% of the time (De Andrade E Sousa et al. 2019) and in humans a model obtained over 80% accuracy using only genomic repeats (Z. Wang et al. 2006).

There are now numerous consortia generating genome-wide data for an assortment of genetic and epigenetic marks. To expand our knowledge of how XCI status is regulated, we herein study the relationship between multiple epigenetic marks and XCI status, with XCI status being determined with previous XCI status calls as well as by expression, DNAme and other epigenetic marks. Additionally, we investigate the co-regulation of XCI status for genes within a variably escaping domain.

4.2 Methods

4.2.1 Previous XCI status calls

We used XCI meta-status calls from chapter 2 (Balaton, Cotton, and Brown 2015) for all comparisons with past XCI statuses and to train our models. Genes which escape and mostly escaped were combined together due to the small size of these categories, with genes in the PAR1 being left out or having their own separate category depending on the analysis. Genes which were mostly subject to XCI were combined with genes subject to XCI for comparisons between studies but were left out when training models. Genes which were annotated as variably escaping, mostly variably escaping and discordant across studies were combined together as variably escaping genes for comparisons.

4.2.2 Histone ChIP-seq analysis

Histone Chromatin Immuno-precipitation (ChIP-seq) bigwig files were downloaded from the IHEC data portal (Bujold et al. 2016) and quantified with bigWigAverageOverBed (Kent et al. 2010) for a region 500bp upstream of TSSs as annotated by Gencode (Harrow et al. 2012). We normalized the data across samples by multiplying samples to have the same total depth (including all chromosomes). The metagene plots for Figure 4.5 were generated using Deeptools computeMatrix and plotProfile (Ramírez et al. 2016).

4.2.3 Expression analysis

Xi/Xa expression based XCI status calls per sample were generated for chapter 3 and reused here (Balaton et al. 2021). We used a different threshold to find variably escaping genes here, requiring at least two samples with each XCI status. This narrowed down the number of variably escaping genes and increased the chance that those found would have enough samples to reach significance. The overall expression level of genes was calculated using bigwig files downloaded from the IHEC data portal (Bujold et al. 2016) and quantified as RPKM using VisRseq (Younesy et al. 2015).

4.2.4 DNAme analysis

WGBS bigwig files were downloaded from the IHEC data portal (Bujold et al. 2016) and quantified with bigWigAverageOverBed (Kent et al. 2010) for a region 500bp upstream of TSSs as annotated by Gencode (Harrow et al. 2012). DNAme thresholds established in (Balaton et al. 2021) were used to determine which genes were escaping XCI and which were subject to XCI. These thresholds are: DNAme<10% escapes XCI, 15%<DNAme<60% subject to XCI, and DNAme>60% hypermethylated. A threshold of DNAme<15% in males was used to filter out TSSs which were methylated on the Xa and therefore not informative for this analysis. To see the differences between adjacent CpGs, we converted bigWig files to bedGraphs and for each island we used R to find the mean absolute value difference between each adjacent CpG.

DNAme per read was calculated by downloading WGBS bam files and using a script to count the number of unmethylated and methylated CG dinucleotides per read within CpG islands within 2kb of TSSs. For allelic DNAme, we did similar but only examined reads that overlapped heterozygous SNPs identified in our Xi/Xa analysis and had to reconstruct the read from the CIGAR string in the bam file in order to determine the allele of origin. We analyzed allelic DNAme for SNPs within 2kb of TSSs and noted which were found in CpG islands.

We used bins for every 10% increase in mean DNAme and chose bins for each individual gene per sample separately. All of the reads per bin, across all genes and female samples were used for Figure 4.3C. For allelic DNAme we first filtered out polymorphisms where the alleles were CT or GA as bisulfite conversion makes it impossible to differentiate these. We then lumped all reads with a C or T allele, and all reads with a G or A allele together and filtered out polymorphisms without at least five of each allele type in a sample. The mean DNAme per read per allele type was then calculated and this was used to make XCI status calls per polymorphism in each sample with enough reads. The thresholds were 0.25 and 0.75 for our XCI status calls with polymorphisms with both alleles below 0.25 being called as escaping from XCI and both alleles higher than 0.75 being called as hypermethylated. Polymorphisms with one allele above 0.75 and the other allele below 0.25 were called as subject to XCI. The DNAme per read per polymorphism was binned as above, but instead of using the mean DNAme across all reads, we determined the mean DNAme per allele and used the mean of that; this was done so that we get the mean between the Xi and Xa if there are more reads for one than the other. Additionally, we determined which allele was lower for each polymorphism and graphed the low allele separately from the high allele, per bin.

4.2.5 Histone-based XCI status predictions

A simple histone predictor was made by using genes with known XCI status as published in chapter 2 (Balaton, Cotton, and Brown 2015), and defining XCI status for genes within two standard deviations of the mean for each XCI status, similar to a model used in (Cotton et al. 2015). Because the mean of genes subject to XCI and the mean of genes escaping XCI were often within two standard deviations of each, the average of these two means was often used as a threshold instead.

For our random forest models, we wanted to include both male and female data, and breast did not have any male data so we used the kmeans function in R to cluster all of our samples based on autosomal levels of all seven epigenetic marks used herein. With three clusters we had multiple male and female samples in each cluster. As input for our models, we used individual female data per sample and matched it with the mean values per gene across males in the same cluster.

Random forest models were trained using the R package caret (Kuhn 2008) with the trainControl method cv and the train method rf. We trained the model on genes known to escape or be subject to XCI (Balaton, Cotton, and Brown 2015). The training metric was ROC, tunelength was 5 and ntree was 1500. Three genes escaping and subject to XCI were left out of the training set and used to check accuracy of overall calls. We trained twenty models per sample, with each model being trained on a random sample of 75% of genes escaping XCI and twice as many genes subject to XCI, with each iteration of the model using 75% of the number of input escaping genes. Accuracy per model was tested on the remaining genes with known XCI status. Genes were considered as escaping or subject to XCI if 15+ of 20 models predicted them as escaping or subject to XCI respectively. Separate categories were made for genes where only 12-14 of the models agreed on the gene's XCI status, being annotated as leaning subject or leaning escape. Overall calls were made across samples with genes with 66% or more of samples agreeing on a gene's XCI status being called as subject to or escaping from XCI, genes with at least 33% or more of all samples having each XCI status being called as variably escaping from XCI, and genes which required the leaning categories to reach 66% of samples having a status being annotated with a similar leaning status.

4.2.6 Statistical comparisons

All statistical comparisons were done in R. The majority were t-tests with a Benjamini-Hochberg (BH) multiple testing correction (Benjamini and Hochberg 1995) with results deemed significant if they had an adjusted p-value<0.01. The one test with a different threshold was for comparing genes variably escaping XCI as determined by Xi/Xa expression. This test used a 0.05 threshold and had no multi testing correction due to a low sample size, with most genes only having 2 or 3 of each category. If we had a larger sample size, a more stringent test would be preferred. We also used a chi-square test to determine enrichment of significant histone differences between tissues and TSSs, with p-value of 0.01.

4.3 Results

To understand the interplay of various epigenetic marks and XCI status we sought a dataset with both a broad range of epigenetic marks available and matched expression data to determine XCI status. We thus turned to data from the International Human Epigenome Consortium (IHEC), which has standardized ChIP-seq datasets for the core histone marks H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K27me3 and H3K36me3 along with WGBS to examine DNAme. We specifically used data from the Center for Epigenome Mapping Technologies (CEMT) as we could get all the core marks for each sample, along with raw data for whole genome seq, RNA-seq and WGBS for some samples (Supplemental Table S4.1). As these samples were derived from cancer they have a high frequency of skewed XCI, allowing us to use allelic expression to determine XCI status (Balaton et al. 2021). We additionally examined data from another group within IHEC, this one from Core Research for Evolutional Science and Technology (CREST). The advantage to the CREST data was that the samples were not derived from cancer thus allowing us to determine whether any trends that we observe in the CEMT data are due to the samples being cancer; however, the CREST samples had less sequencing depth, fewer females (only nine), and could only be examined for DNAme and histone marks.

4.3.1 Histone marks differ with sex and XCI status

We compared published XCI status calls from a previous meta-analysis of various studies (our consensus XCI status calls from chapter 2, hereby referred to as meta-status) (Balaton, Cotton, and Brown 2015) and sex to the levels of histone marks within 500bp upstream of a gene's TSS except for the mark H3K36me3 which is associated with gene bodies and so was examined at exons (Barski et al. 2007). For the purposes of this and all future analyses, genes in the PAR were not included with genes escaping from XCI as they may be epigenetically distinct, especially when comparisons with males are included.

Comparing males and females, the median level per TSS was significantly different (p-value < 0.01) for most marks at both genes escaping and subject to XCI in both datasets (Supplemental Table S4.2). Fewer marks showed significant differences when comparing between genes escaping XCI and those subject to XCI within each sex, especially in the CEMT data. The euchromatic marks H3K4me3 and H3K27ac were significantly different between genes subject to vs escaping from XCI in both CEMT and CREST females, but the heterochromatic marks H3K9me3 and H3K27me3 were only significantly different with the CREST dataset. Comparing XCI statuses within males gave the fewest significantly different marks, with H3K4me1, H3K9me3 and H3K27ac being significant in CEMT and only H3K4me1 being significant in CREST.

To visualize the differences between the Xi and Xa, we calculated the Xi to Xa fold change for each mark by taking log2 of the female-male difference (the contribution from the Xi) and dividing it by the male value (the contribution from the Xa) (Figure 4.1A, Supplemental Figure S4.1, Supplemental Table S4.2). This allowed us to see that heterochromatic marks are generally higher on the Xi than Xa, especially for genes subject to XCI. H3K27me3 has a higher Xi:Xa fold change than H3K9me3 in both XCI statuses and both datasets. Both marks are highest at genes subject to XCI in the CEMT dataset, while in the CREST dataset the differences between the median gene escaping XCI and the median gene subject to XCI are small. For euchromatic marks, the Xi:Xa trend is close to 1:1 at genes escaping XCI, while lower for genes subject to

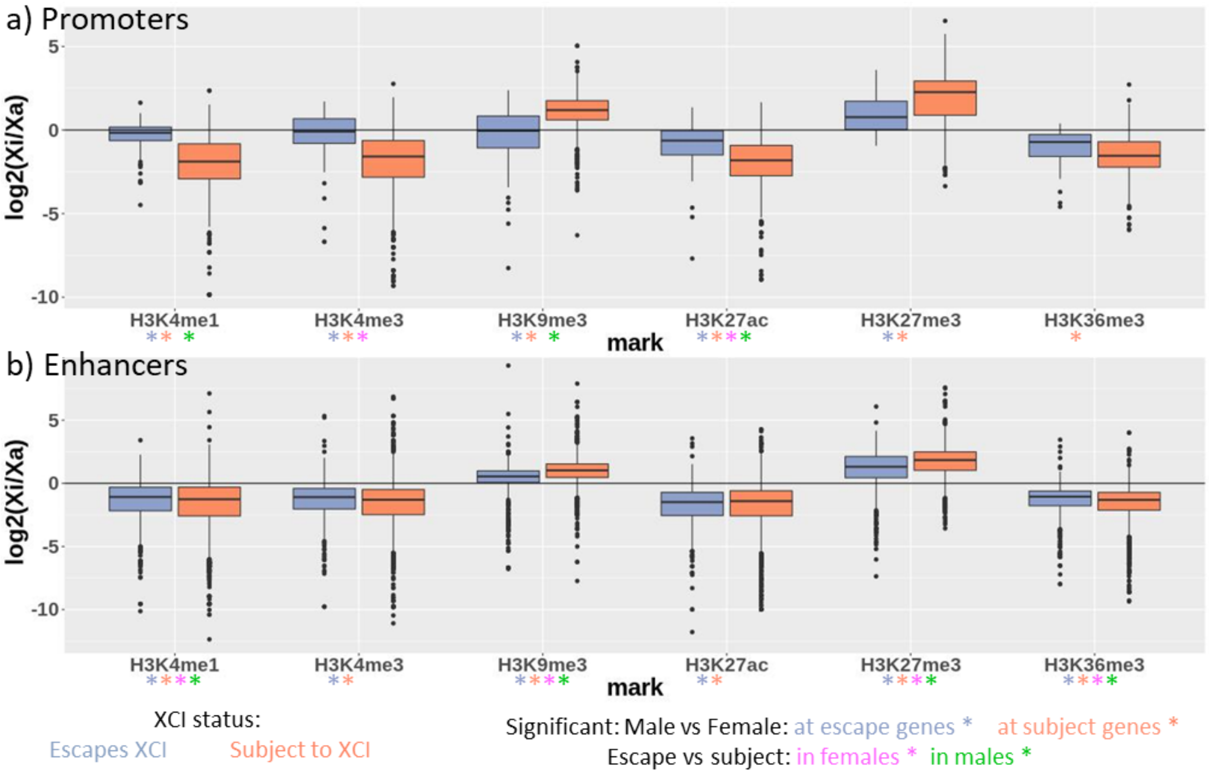


Figure 4.1 The Xi has more heterochromatic and less euchromatic marks than the Xa.

Log_2(Xi/Xa ratio) for the histone marks examined here, split by XCI status. Data from CEMT is shown. Significance for the various t-tests featured in Table S2 are shown by the differently colored star (adjusted p-values <0.01).

XCI. H3K36me3 is fairly equivalent between the two XCI statuses however, with the Xi being around half of the Xa. For genes subject to XCI, the healthy CREST samples had less of an Xi to Xa difference than the CEMT cancer samples, while at genes escaping from XCI the differences were weaker, and more variable between the datasets.

Examining the number of genes per XCI status which were significantly different between sexes in the CEMT data (t-test, adjusted p- values<0.01), we found H3K27me3 significant in over 85%

of the genes per XCI status category except for genes in the PAR, which have 2 copies in both sexes (Supplemental Table S4.3). H3K9me3 was significant for over 75% of genes that are subject to or variably escape from XCI, with a decreasing percentage of significant genes in each of the no call, escape and PAR groups. H3K4me3 was most often significant in genes escaping from XCI, but it was only significant in 56% of escape genes. H3K4me1, H3K27ac and H3K36me3 were never significant in more than 50% of the genes with any XCI status. With the CREST dataset, none of the histone marks significantly differed between sexes for more than 1% of genes. We attributed this to CREST having lower depth to their data than CEMT did, and therefore having more samples per gene with zero reads in both sexes. Additionally, CREST had fewer samples to power our statistical tests. To test whether CEMT or CREST was the outlier, we downloaded H3K27me3 data from ENCODE and found a similar trend to the CEMT data, with over 70% of genes in the escaping, subject to XCI and variably escaping categories being significantly different between sexes. We analyzed chromosome 7 as an example autosome and saw a much lower percentage of genes with significant male-female differences significant for H3K9me3 and H3K27me3 than even genes escaping from XCI, so even genes which escape from XCI have a significant increase of these heterochromatic marks from males.

In addition to our promoter based analysis, we also compared histone marks at enhancers annotated to genes on the X (Fishilevich et al. 2017) and found that all marks differed significantly between males and females, in both XCI statuses (Supplemental Table S4.2, t-test, adjusted p-values<0.01). Our enhancer annotation included 2695 enhancers within genes and 11565 intergenic enhancers. Of the genic enhancers, 170 are annotated to genes escaping XCI and 1169 annotated to genes subject to XCI. For intergenic enhancers, 866 are annotated to genes escaping XCI and 3908 are annotated to genes subject to XCI. When only considering genic enhancers, H3K27ac was no longer significant in either XCI status, while at intergenic enhancers H3K4me3 was not significant at genes escaping XCI. Comparing enhancers that map to genes escaping XCI vs those subject to XCI, we see that H3K4me1, H3K9me3, H3K27me3 and H3K36me3 are significantly different in both females and males when considering all enhancers. In females many of these marks significantly differ between XCI status when examining only genic enhancers, but not when examining intergenic enhancers (H3K4me3, H3K9me3, H3K27ac) with H3K36me3 being significant only in intergenic enhancers and not genic enhancers. Males have the opposite trend, where many marks (H3K4me1, H3K9me3, H3K27me3 and H3K36me3) differ significantly with XCI status at intergenic enhancers but not at genic enhancers, with H3K4me3 showing the opposite trend. Looking at the Xi:Xa fold change at enhancers (Figure 4.1B), all of the heterochromatic marks were higher on the Xi than the Xa, while euchromatic marks were higher on the Xa than the Xi. This did not differ greatly between enhancers which were annotated to interact with genes escaping XCI vs subject to XCI.

We also compared DNAme and expression across XCI status and sex and had similar results between datasets (Supplemental Table S4.2). DNAme was significantly different between males and females at genes subject to XCI and in females between genes subject to XCI and those escaping from XCI. Expression was not found significantly different in any of the comparisons here.

4.3.2 Using Xi/Xa expression ratio to identify XCI status

To further analyze the interaction between XCI status and epigenetic marks, we compared these marks to XCI status calls made within the same sample. To determine sample-specific XCI status, we used our Xi/Xa expression based XCI status calls in a subset of the CEMT samples from chapter 3 (Balaton et al. 2021). Across the eight skewed samples, 30 genes escaped XCI, 202 genes were subject to XCI and 8 genes variably escaped from XCI (requiring at least two samples with each XCI status to be called variable) (Figure 4.2A, Supplemental Table S4.4)). The genes found to be escaping XCI by this analysis were previously described to escape from XCI or be located in the PAR1, with two genes (*AX746622* and *LOC389906*) having no prior XCI status call (Balaton, Cotton, and Brown 2015). Genes called subject to XCI here are less consistent, with 158 of 187 genes called subject to XCI herein being called subject to XCI previously; only five genes were completely discordant, being found subject to XCI here and escaping from XCI previously. Two of the genes found to variably escape herein were previously called as variably escaping from XCI, while five were escaping XCI and one subject to XCI.

Using these XCI status calls did not substantively alter the histone mark representations found using meta-status calls (Supplemental Table S4.2). The histone marks which were found significant in all three comparisons (CEMT marks vs meta-status, CREST marks vs meta-status, CEMT marks vs CEMT Xi/Xa calls) are H3K4me3 and H3K9me3 (in all but the male XCI comparison), H3K27me3 (between sexes for both XCI statuses), H3K27ac (between sexes at genes subject to XCI and between XCI statuses in females), and H3K4me1 and H3K36me3 (only between sexes at genes subject to XCI). DNAme was again different between males and

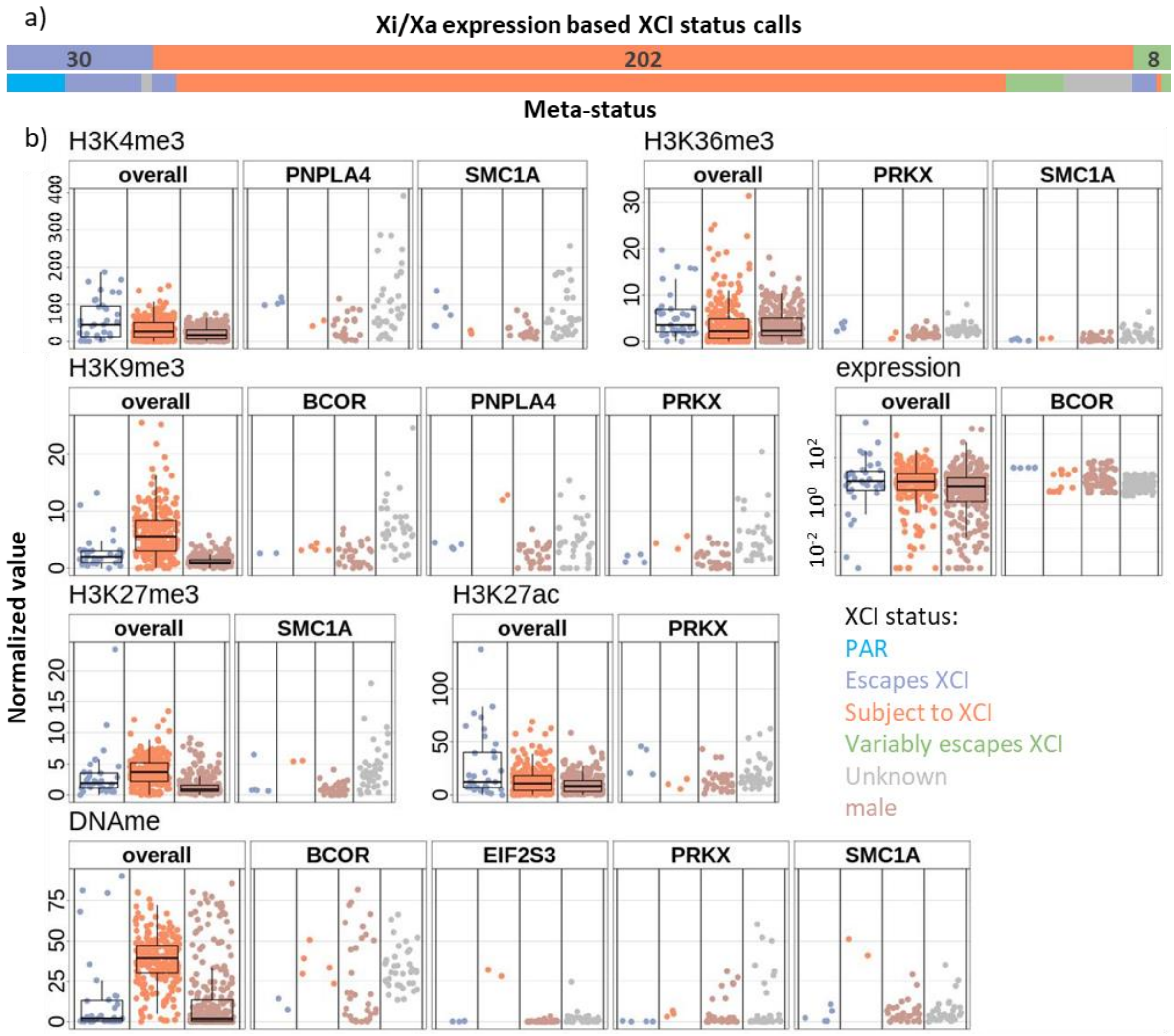


Figure 4.2 Epigenetic marks do not change consistently with XCI status for variably escaping genes.

(A) The number of genes with each XCI status call across all samples, determined by Xi/Xa expression with their distribution by meta-status underneath. (B) The interaction of histone marks and Xi/Xa expression determined XCI status. On the left for each mark is a comparison to the overall XCI status across samples per gene and on the right are shown the variably escaping genes which had significant differences in the histone mark between samples that were subject to and those escaping from XCI. A p-value of 0.05 was used for significance. Unknown XCI status is for samples which were uninformative in the Xi/Xa expression analysis. Expression is on a $\log_{10}$ scale while the others are on a linear scale.

females at genes subject to XCI and different at genes escaping from vs subject to XCI in females. Expression levels were not different between sexes or XCI statuses. An additional benefit of having both histone marks and XCI status on individual samples is the ability to examine how histone marks correlate with XCI status at variably escaping genes.

Genes that variably escape from XCI provide a unique opportunity to study differences between genes escaping vs subject to XCI in the same genomic context. All of the marks available except for H3K4me1 were significantly different (p-value <0.05) between samples escaping XCI vs those subject to XCI in at least one of these eight variably escaping genes, but never for the majority of genes (Figure 4.2B, Table 4.1). As might be anticipated, when active marks were

gene	H3K4me1	H3K4me3	H3K9me3	H3K27ac	H3K27me3	H3K36me3	DNAme	expression	nE	nS
BCOR	0.076	0.42	**0.019**	0.43	0.065	0.56	**0.011**	**0.0097**	**2**	**5**
CXorf38	0.071	0.26	0.82	0.52	0.18	0.24	0.10	0.18	**3**	**2**
EIF2S3	0.054	0.097	0.80	0.12	0.33	0.54	**0.040**	0.070	**4**	**2**
MED14	0.66	0.98	0.32	0.39	0.28	0.45	0.42	0.86	**2**	**3**
PNPLA4	0.84	**0.029**	**0.0076**	0.069	0.070	0.42	0.15	0.74	**4**	**2**
PRKX	0.73	0.070	**0.029**	**0.047**	0.27	**0.021**	**0.048**	0.053	**5**	**3**
SMC1A	0.11	**0.043**	0.24	0.61	**0.036**	**0.048**	**0.046**	0.15	**6**	**2**
TIMP1	0.59	0.40	0.55	0.054	0.78	0.18	0.21	0.87	**3**	**4**

Table 4.1 Significance of differences in epigenetic marks between samples with opposite XCI statuses at genes found variably escaping XCI by Xi/Xa expression.

We also tested whether expression differed between samples with opposite XCI status. Presented here are the p-values of t-tests. Those with p-values less than 0.05 are in bold. nE and nS are the number of samples escaping or subject to XCI for each gene.

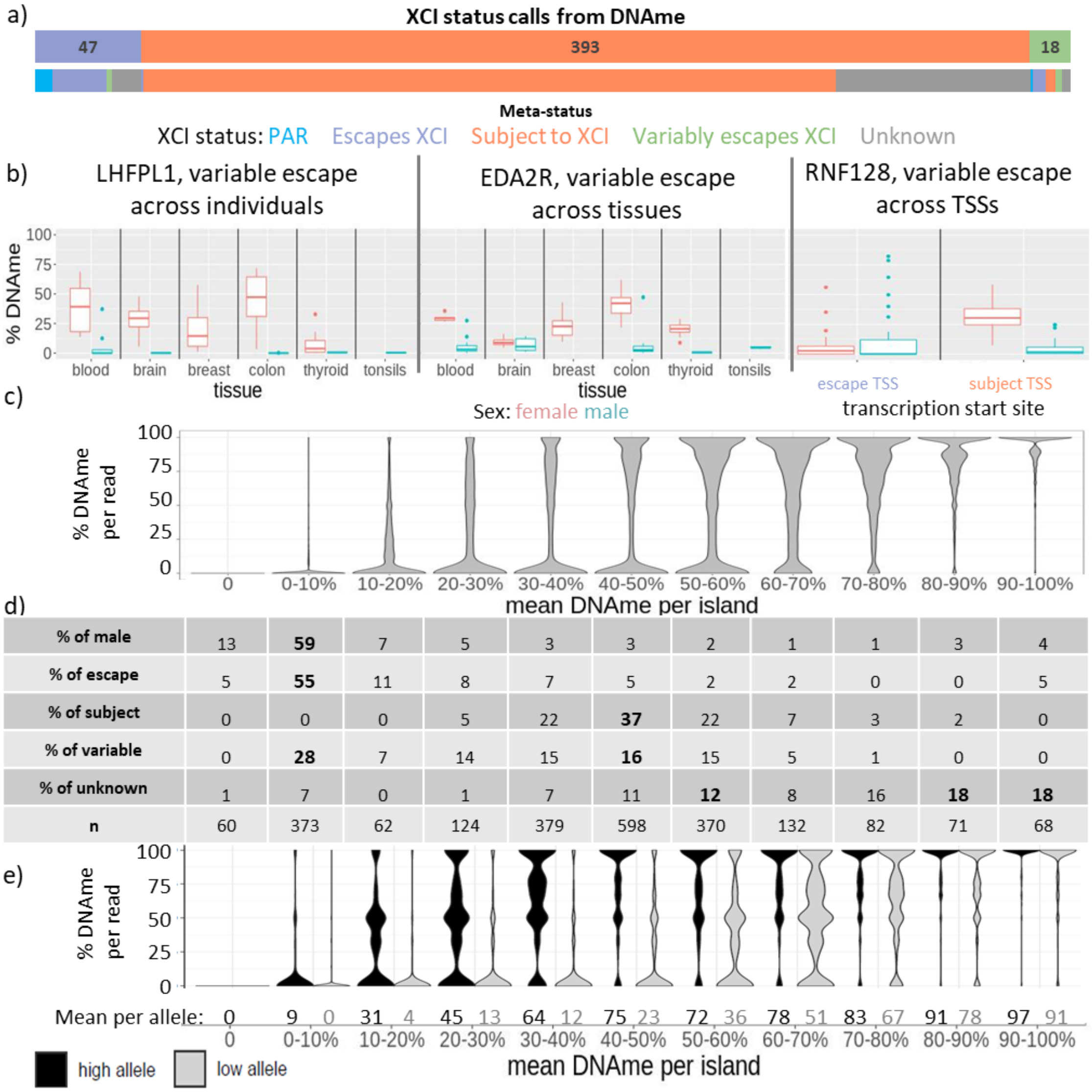
significantly different, they tended to be higher in samples escaping XCI, while inactive marks were lower in samples escaping XCI (Supplemental Table S4.5). The exception to this is H3K36me3 in gene bodies. For one gene, the samples subject to XCI had higher H3K36me3, while in another gene, it was the samples escaping XCI that were higher.

We found that four out of the eight variably escaping genes had significant differences in DNAme (p value<0.05). The samples subject to XCI in *PRKX* had significantly higher DNAme, but were not above the DNAme thresholds for XCI status calls that we established previously [23]. All of the other genes with significant DNAme differences showed a clear switch from an 'escape' DNAme pattern to one matching genes subject to XCI. *TIMP1*, one of the four that was not significant, has low CpG density and high male DNAme so was not expected to differ with XCI status. For the other three, they had few informative samples and we lacked the power to detect differences, they may have false XCI status calls or there may be more complicated epigenetic processes involved. Interestingly, the two genes found to be variably escaping by Xi/Xa expression and meta-status (*MED14* and *TIMP1*) did not show DNAme differences while many of the genes without meta-status calls of variable escape had significant DNAme differences, supporting that these genes are truly variable across these samples. Three of the variably escaping genes did not show significant differences at any of the examined marks; increasing the sample size may give us the power to see more consistent differences across variably escaping genes as some of these genes only had 2 informative samples per XCI status.

We considered a series of other potential contributors to variability in escape from XCI including differences in sequence, expression, or tissue. We wanted to examine whether the Xi allele or the genotype of nearby polymorphisms had any correlation with XCI status, but with an average of only six samples informative per gene, we did not have sufficient power. Two genes showed significant expression differences between samples that escaped XCI vs those subject to XCI (Supplemental Figure S4.3, p-value<0.05). In *BCOR*, samples escaping XCI had higher expression across all exons, while in *EIF2S3* some exons were higher in samples subject to XCI while other exons were higher in samples escaping XCI. XCI status and expression per exon may be linked by different TSSs having different XCI status or possibly different tissues having different XCI status and dominant splicing variants. To test whether variable escape may be tissue-specific, XCI status per sample was compared with tissue of origin; only one of the eight genes showed tissue-specificity, *EIF2S3*, which was the gene with significant differences in exon expression per XCI status. However, with only eight samples across three tissue types and being limited by heterozygous polymorphisms, there are likely other tissue-specific variable escape genes that were not identified here, as many genes did not have multiple informative samples per tissue.

4.3.3 Using DNA methylation to identify XCI status

To increase our sample size, we used promoter DNAme levels to determine XCI status across all genes within the larger 45 sample CEMT dataset. Only TSSs with high CpG density and low male methylation were considered informative, but within this group we found 47 genes escaping XCI, 393 subject to XCI and 18 variably escaping across samples (Figure 4.3A,



% of male	13	**59**	7	5	3	3	2	1	1	3	4
% of escape	5	**55**	11	8	7	5	2	2	0	0	5
% of subject	0	0	0	5	22	**37**	22	7	3	2	0
% of variable	0	**28**	7	14	15	**16**	15	5	1	0	0
% of unknown	1	7	0	1	7	11	**12**	8	16	**18**	**18**
n	60	373	62	124	379	598	370	132	82	71	68

Figure 4.3 DNAme varies at genes variably escaping from XCI.

(A) The number of genes with each XCI status call by DNAme, with their distribution by meta-status underneath. (B) From left to right: An example of a gene that variably escapes XCI across individuals (and within multiple tissues), a gene that variably escapes from XCI between tissues, and a gene that variably escapes from XCI between TSSs. (C) The percent DNAme per read for genes on the X, binned together by their mean DNAme across the CpG island. Only reads overlapping the CpG island were included here. (D) The distribution of genes with each XCI status across the bins of mean DNAme per island. Bins with a peak number of genes in them are bolded. (E) Allelic DNAme, shown as the percent DNAme per read by allele. The mean DNAme across all reads per allele in each bin is shown underneath.

Supplemental Table S4.4). Our DNAme based calls had strong concordance with meta-status; there were no genes called as escaping XCI here that were previously called as subject to XCI, while only one of the genes called as subject to XCI here was previously called as escaping XCI. We included genes in the variably escaping from XCI category if at least one of their TSSs had 33% or more of its samples escaping XCI and another 33% or more samples subject to XCI. Additionally, one gene had opposing XCI statuses at separate TSSs and 36 had opposite XCI statuses across tissues (Figure 4.3B). An additional 67 genes were found variably escaping in at least one tissue but were not identified as variably escaping from XCI in the larger dataset. Only

VEtype	n VE transcripts	n VE genes	H3K4me1	H3K4me3	H3K9me3	H3K27ac	H3K27me3	H3K36me3	DNAme	expression
Across dataset	22	17	6%	17%	18%	17%	17%	16%	**100%**	0%
Across tissues	70	40	12%	**28%**	7%	16%	14%	21%	**51%**	**39%**
Between TSSs	2	1	0%	100%	0%	0%	100%	0%	**100%**	0%
Within blood	74	51	2%	8%	2%	12%	4%	8%	**63%**	3%
Within brain	10	9	13%	0%	0%	13%	0%	11%	**63%**	0%
Within breast	29	24	0%	0%	4%	0%	0%	4%	18%	0%
Within colon	8	4	**25%**	**25%**	0%	0%	0%	20%	**100%**	0%
Within thyroid	45	27	16%	0	7%	10%	0%	24%	**97%**	0%

Table 4.2 The percentage of variably escaping genes found by DNAme that have significant differences in epigenetic marks (BH corrected p-value<0.01).

Different categories of variable escape are included on the left. The number of variably escaping (VE) transcripts found per category and the number of unique genes is also included. Categories with 25% or more variably escaping genes found significant are bolded, excluding variable escape between TSSs which only had 1 gene available.

one of the genes found variably escaping from XCI in the Xi/Xa expression-based calls was also found variably escaping here. In addition, 96% of genes escaping and 87% of genes subject to XCI identified by Xi/Xa expression in these samples had concordant status in our DNAme based calls, with most of the discrepancies between calls being due to genes being called as variably escaping in only one of the datasets.

Comparing epigenetic marks to DNAme based XCI status calls, H3K4me1 and H3K36me3 were not significantly different between genes with opposite XCI status calls but the rest of the marks (H3K4me3, H3K9me3, H3K27me3 and H3K27ac) were very significant (Supplemental Table S4.6). We again compared epigenetic marks at variably escaping genes to see if they differed between samples in which the gene escaped XCI vs those in which it was subject to XCI. We categorized variable escape genes as those variably escaping across the dataset, across TSSs, across tissues or within specific tissues. For variable escape from XCI between individuals across the dataset, every mark examined was found to be significant (p-value<0.01) in at least one gene; however across all categories of variable escape from XCI, only DNAme, expression and H3K4me3 were significant in more than 25% of genes in any type of variable escape category (Table 4.2) . The direction of histone marks changes was less consistent than for Xi/Xa expression based XCI status calls, with the majority of genes still having higher active marks in genes escaping XCI and higher inactive marks in genes subject to XCI, but with many genes showing the opposite results (Supplemental Figure S4.4).

We have previously seen that the average DNAme at genes subject to XCI was 38%, less than expected if the Xi were completely methylated and that some genes subject to XCI had DNAme as low as 15% (Balaton et al. 2021). To further understand this low DNAme, we examined the DNAme per WGBS read at CpG islands for the six female samples for which we had WGBS aligned reads, along with one male as a control. We sorted each gene per sample into bins based on their mean DNAme, with separate bins for every 10% DNAme increase (Figure 4.3C). Over 60% of genes in males or with a meta-status of escape from XCI had less than 10% mean DNAme, while over 80% of genes with a meta-status of subject to XCI had a mean DNAme between 30% and 60% (Figure 4.3D). Variably escaping genes were found distributed in the range where genes escaping and subject to XCI were found; however, genes with intermediate 20-30% DNAme had more variably escaping genes than genes with a consistent XCI status. Genes with no known XCI status tended to have high DNAme, with over half of them having 70% DNAme or higher. The genes in the 30-40% DNAme and 40-50% DNAme bins had a surprisingly low amount of reads with high DNAme (24% and 35% of reads over 75% DNAme, respectively) so it appears to be that the majority of cells are partially methylated and not that some cells are methylated while others are not (Supplemental Table S4.7). Intermediate DNAme (33-66%) is found most frequently in the 20-30% and 70-80% bins.

To further differentiate DNAme from the Xa and Xi, we examined DNAme per read overlapping heterozygous SNPs within 2kb of TSSs. In addition to the usual limitations of mapping allelic reads, we had to exclude C T and G A polymorphisms as the bisulfite conversion step in WGBS converts unmethylated C to T and on the opposite strand this appears as a G to A conversion. Separating these genes into the same bins of 10% mean DNAme as earlier (Figure 4.3E), we can see that the intermediately methylated reads tend to be on the high allele (the presumed Xi) for bins with less than 40% DNAme and are on the low allele for bins with greater than 50%

DNAme. We further used this allelic DNAme to call XCI status, using thresholds at 25% and 75% DNAme per allele, with genes having both alleles below 25% being called as escaping from XCI and those having one allele below 25% and one above 75% being called as subject to XCI. We find that these calls had good accuracy for SNPs within CpG islands, all 28 of the loci found escaping here and 50/51 of the loci found subject to XCI here agree with previous XCI status calls. For SNPs outside of CpG islands this was not as predictive, only 91/182 loci found escaping XCI and 235/274 loci found subject to XCI here were concordant with their meta-status.

To explain the enrichment of intermediately methylated reads, we examined the DNAme per CpG across some of these islands where we can see that the DNAme level is not consistent (Supplemental Figures S4.5, S4.6) and that these intermediately methylated WGBS reads and CpG island DNAme averages likely capture this phenomena. We can also see this variability in DNAme in males although it is rarer. This problem seems exacerbated in the cancer samples examined here as compared to healthy samples from CREST. Examining the average DNAme difference between adjacent CpG sites we saw an average difference of 24% in cancer and 13% in healthy samples.

4.3.4 An epigenetic model to predict XCI status across samples

As DNAme has been shown repeatedly to be a strong predictor of XCI status, we wanted to test whether the other epigenetic marks examined could predict XCI. Our first model was using simple thresholds to separate genes that have low values for a mark vs those with high values; however, there was still a large overlap between the regions occupied by the opposing XCI statuses causing accuracy to be poor (Supplemental Table S4.8). The most accurate mark here was H3K27me3 with an accuracy of 68%, ignoring genes called as variably escaping XCI. These histone thresholds overcalled genes as variably escaping XCI, with gene-body H3K36me3 being the worst and calling 87% of genes as variably escaping XCI across samples. We moved to a random forest predictor model to predict the XCI status of genes using each individual mark per female sample along with matched male data. Using this predictor, we could predict escape from XCI with accuracies ranging from 42% with H3K9me3 to 69% with H3K4me3 and for genes subject to XCI with accuracies ranging from 85% with gene-body H3K36me3 to 99% with H3K27ac. In contrast, a similar model using CpG island DNAme data obtained a much better accuracy of 87% for predicting genes as escaping XCI and 99% for predicting genes as subject to XCI, showing the higher predictive ability of DNAme.

To get XCI status calls from histone mark data with an increased accuracy, we combined data from all of the histone marks and DNAme data from CEMT and trained a new random forest model (Kuhn 2008). This model was trained on XCI meta-statuses (Balaton, Cotton, and Brown 2015) and was able to accurately predict genes escaping vs subject to XCI, with a median accuracy for genes outside the training set of 75% for genes escaping from XCI and 90% for genes subject to XCI (Figure 4.4A). We trained the model 20 separate times per sample and were confident in a prediction if 75%+ of the models agreed. Across all samples, the model called 46 genes as escaping XCI, 780 genes as subject to XCI and seven genes as variably escaping from XCI (Figure 4.4B, Supplemental Table S4.4). While none of the genes predicted to escape XCI here have a meta-status of subject to XCI, 11 of the genes predicted to be subject to XCI have a meta-status of escaping XCI and an additional six genes are located in the PAR1 and are

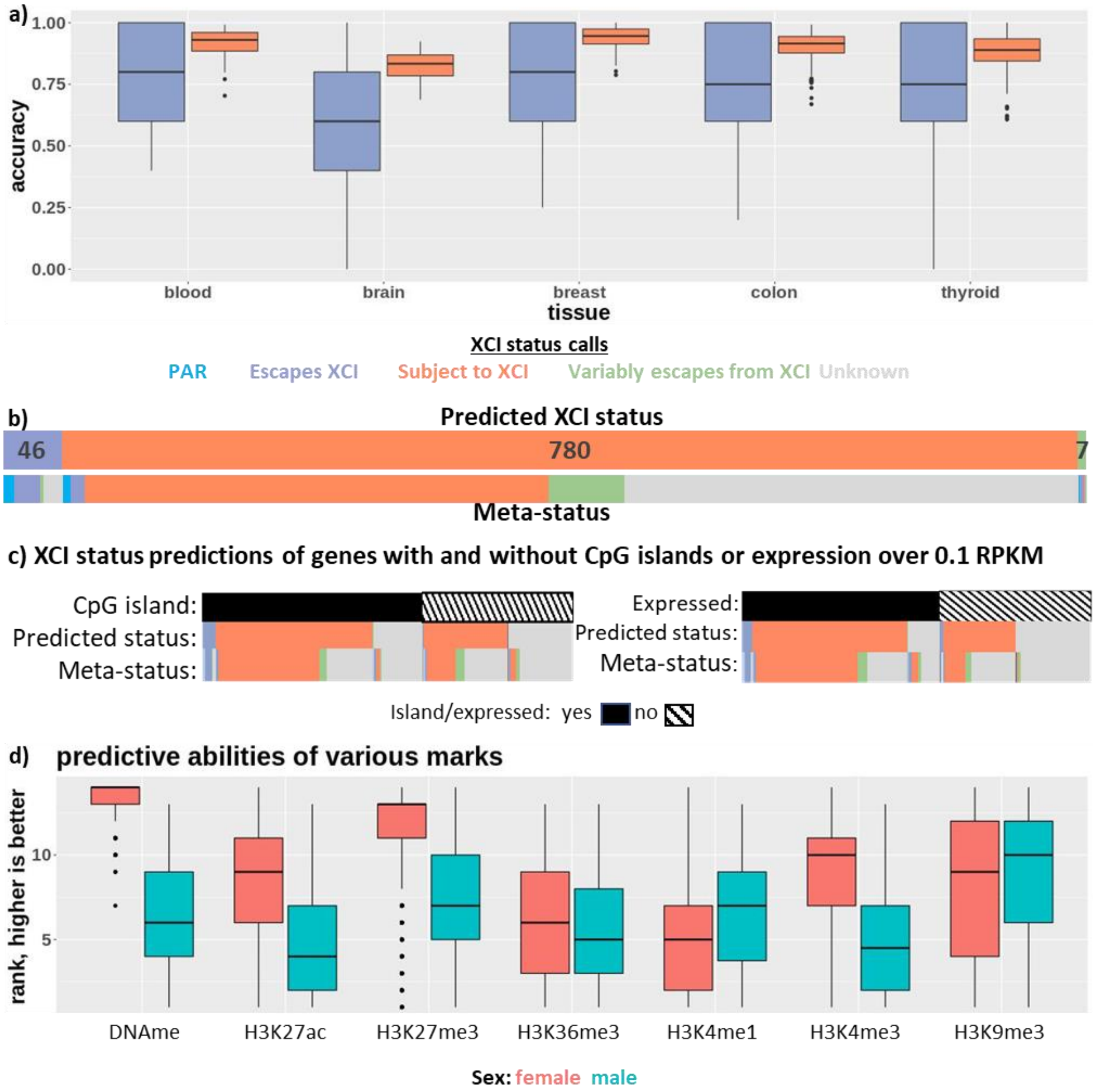


Figure 4.4 XCI status predictions with an epigenetic model expands the number of genes examinable.

(A) Accuracy of our epigenetic predictor using DNAme and all six histone marks. Each point is one of the 20 models per sample. This accuracy is tested on genes outside of the training set. (B) The number of genes with each XCI status as predicted by our model, with their distribution by meta-status underneath. (C) As b, but further split by the presence of a CpG island or by an expression threshold of 0.1 RPKM. (D) The predictive ability of each mark. Each mark was ranked per model on how important it was to the model, with the most important mark being ranked 14 and the least important being ranked 1. We used the marks within each female sample paired with the mean mark in similar male samples for the predictor, so both the female and male marks are featured here.

expected to escape XCI (Balaton, Cotton, and Brown 2015). Comparing to our Xi/Xa expression based XCI status calls, 23 genes escape XCI in both sets while only two were called as escaping XCI by Xi/Xa expression and subject to XCI in this model and three genes had the opposite calls. Of the eight genes found variably escaping by Xi/Xa expression, three of them (*CXorf38*, *PRKX* and *SMC1A*) had their predicted XCI status across samples perfectly match that found by Xi/Xa expression. There are no genes which were called opposite calls across samples by our DNAme based calls and this model, however some of the genes found to variably escape differed between the two.

This epigenetic based model can predict XCI status across all genes on the X chromosome, without being limited by CpG density and male values as DNAme is; however, transcripts without high promoter CpG density are almost twice as likely to have inconsistent XCI status calls within the same sample while genes that are lowly expressed (median RPKM across samples <0.1) are over three times more likely to have an inconsistent XCI status call (Figure 4.4c, Supplemental Table S4.9). We predicted an XCI status for over 300 genes that did not have previously annotated XCI statuses, however ~200 of these had low expression and so may actually be silent on both the active and inactive X chromosome making an XCI status call moot. DNAme was the most important input for the models, with H3K27me3 being the next most important (Figure 4.4D).

A separate model was trained and used within each sample, however the models are capable of being used across samples within the same tissue with reduced accuracy and even across tissues

(Supplemental Figure S4.7). Some tissues trained models which were accurate across all tissues while other tissues had accurate predictions made in them, from models trained on any tissue (Supplemental Figure S4.8). Brain was the best tissue at training models, and breast was the best tissue for predicting accurately with other tissue's models. Models in some tissues tended to overcall genes as subject to XCI while others overcalled genes as escaping from XCI.

In addition to the seven genes which our epigenetic predictor called as variably escaping XCI across samples, we found 48 genes with tissue-specific escape from XCI, and one gene with separate TSSs with opposite XCI status. We compared our epigenetic marks across samples, tissues and TSSs with opposite XCI statuses (Supplemental Table S4.10). We found that very few marks had significant (t-test, adjusted p-value <0.01) differences between samples found escaping vs subject to XCI in our set of genes found variably escaping across samples. DNAme was the exception to this with four of seven genes having significant DNAme differences. For the genes found variably escaping across tissues, all the marks had multiple genes significantly different between tissues subject to XCI vs tissues escaping from XCI, but many of the genes that didn't variably escape also had significant differences across tissues. Tissue-specific variable escape genes had significant enrichment (chi-square test, adjusted p-value<0.01) for genes with tissue-specific H3K27me3, H3K4me3, DNAme and expression over genes that did not variably escape from XCI. There was only one gene found to variably escape between TSSs so no statistical tests were possible, however there were differences between TSSs for H3K27ac, H3K4me1 and DNAme and at gene-body H3K9me3 for the different exons used.

To further examine variable escape across samples, we decided to lower our threshold for what percentage of samples need each XCI status in order to be called as variably escaping from XCI. At our threshold requiring 33% of samples to have each XCI status in order to be called as variably escaping from XCI, we found 7 of 1155 genes to be variably escaping. Lowering this threshold to 25% found 35 variably escaping genes, at 10% we found 304 genes and at 5% we found 476 genes. This shows that there is some level of variability in XCI status of 41% of genes, but few genes are highly variable across samples. As the threshold for calling genes as variably escaping decreased, the percentage of these genes with significant DNAme differences between samples with opposite XCI statuses decreased down to 20% and the percentage of genes with H3K27me3 differences rose to 27% (Supplemental Table S4.11). So H3K27me3 differences may be driving these differences in XCI status calls for genes that are less likely to variably escape from XCI. The high number of genes variably escaping XCI with a low variable escape threshold could also be because this data is from cancer, with a limited number of samples being epigenetically mis-regulated at each gene.

To validate our conclusions from this model on healthy samples, we trained our overall epigenetic predictor on a non-cancer dataset from CREST with all of the same epigenetic marks. The CREST dataset contains nine samples with which we were able to obtain all the required epigenetic data for our predictor. With this data, we predicted 84 genes escaping from XCI, 791 subject to XCI, six variably escaping across samples, ten across tissues and six across TSSs. These calls are similar to those in the CEMT data, with 94% of genes with calls from both datasets agreeing **(**Supplemental Table S4.12). The genes variably escaping from XCI in the CEMT dataset tended to be escaping XCI in CREST while genes variably escaping in CREST tended to be subject to XCI in the CEMT dataset. The number of variably escaping genes is similar between datasets, even with the difference in the number of samples. The number of tissue-specific genes is much reduced in CREST however, likely due to having only two tissues rather than five. CREST tissue-specific genes had significant differences in H3K27me3, DNAme and expression between tissues, all three of which were also significant in CEMT samples. CREST had enough genes variably escape across TSSs to see that H3K4me3, H3K27me3 and DNAme were significantly different between TSSs escaping and TSSs subject to XCI in females (Supplemental Table S4.13). Males had significant differences in H3K4me3, H3K27ac, H3K27me3, H3K36me3 and DNAme between TSSs escaping vs subject to XCI in females, which suggests that these TSS also differ significantly on the Xa. These TSSs may be predisposed to have different XCI statuses based on their epigenetic landscape prior to XCI or the Xa differences may be misleading the predictor causing it to predict different XCI statuses.

To test whether the variably escaping genes found by our CEMT predictor intrinsically differ from genes with a consistent XCI status, we trained a model to differentiate genes which are escaping XCI consistently from genes which are escaping XCI in the sample but variably escape from XCI across all samples. To have enough variably escaping genes for this model, we used the 10% variable escape threshold. The model differentiating escape from variable escape genes overcalled variable escape (median accuracy of 75%) but all other accuracy metrics were over 85%, while a model differentiating genes subject to XCI from genes variably escaping from XCI was much worse, overcalling genes as subject to XCI with a median accuracy across samples of 15% for genes called as variably escaping from XCI (Supplemental Figure S4.9). This could suggest that many of these variably escaping genes are just genes subject to XCI that have been miscalled, or it could be that they are epigenetically closer to genes subject to XCI. These genes were depleted for meta-status calls of subject to XCI, however there were still more with a meta-status call of subject to XCI than any other XCI status. These variably escaping genes were predicted to be subject to XCI in more samples than they were predicted to escape from XCI (median of 18 samples subject to XCI and seven escaping from XCI) (Supplemental Figure S4.10). The models differentiating genes escaping XCI from those variably escaping XCI tended to rely on DNAme, H3K27ac and H3K4me3, in both the female samples and their matched male controls while those differentiating genes subject to XCI from those variably escaping XCI tended to rely equally on all marks except H3K36me3 and H3K9me3 (Supplemental Figure S4.11).

4.3.5 Independent regulation of variable escape across a region

To understand the scale at which variably escaping genes are regulated, we examined XCI status calls per sample across a region that is enriched in genes variably escaping from XCI according to their meta-status (Figure 4.5A) (Balaton, Cotton, and Brown 2015). We found that many of the genes in this region which are annotated as variably escaping from XCI had low levels of variable escape with few samples differing from the most common XCI status. The genes that vary their XCI status across samples change their XCI status independent from their neighboring genes, suggesting that regulation of variably escaping genes happens at the single gene level and not at the level of TADs or subTADs. Additionally, we saw genes which had differing TSSs with different XCI statuses and genes that are bidirectional from the same promoter with opposite XCI status showing that the scale could be narrowed even further. All of the genes in this region that showed variable escape here, except for *IRAK1*, had significant differences for some

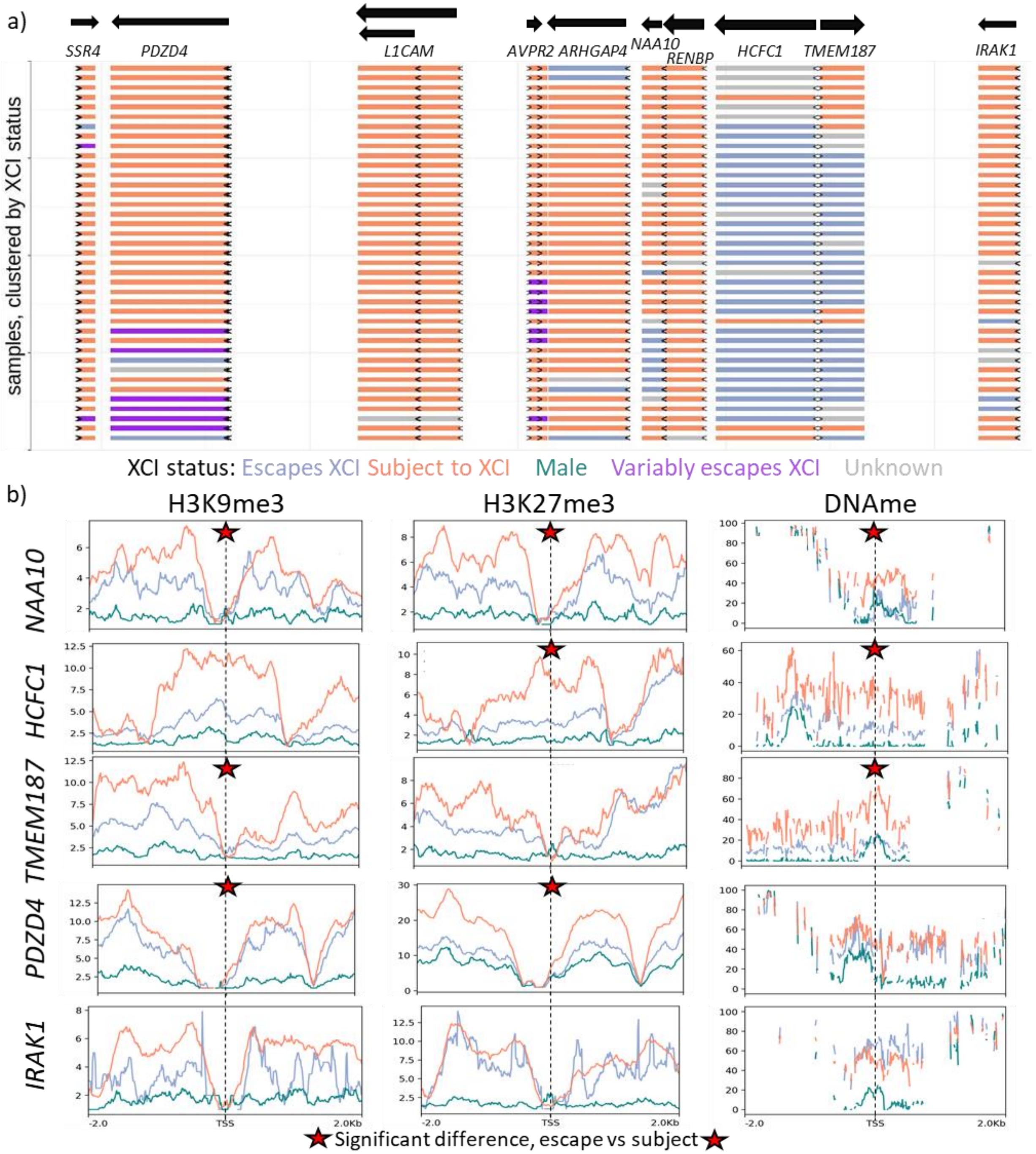


Figure 4.5 XCI status calls are independent between neighboring variably escaping genes.

(A) A map of a variably escaping region, with genes colored by their XCI status as predicted by our random forest model using all epigenetic marks available. The samples from CEMT were clustered based on their XCI status calls within the region. Arrows indicate where each TSS is located, and they point in the direction of transcription. (B)

Metagene plots for the epigenetic marks which were most commonly significantly different between samples subject to XCI vs those escaping from XCI at the above variably escaping genes. Genes were chosen to show every combination of which mark is significant per gene, that we saw in this region. Marks that were significant at a gene are marked with a star.

combination of marks including H3K9me3, H3K27me3 and DNAme between genes escaping vs subject to XCI (p-val<0.05, Figure 4.5B, Supplemental Figure S4.12). Euchromatic marks were not significantly different as often but were different for a few of the genes.

4.4 Discussions

XCI is a classic paradigm for epigenetic regulation, yet why some genes are resistant to silencing (or the maintenance of silencing) and escape XCI remains unresolved. Here we have examined the epigenetic differences between genes escaping and those subject to XCI. Epigenetic marks tended to be more different between males and females than between genes escaping vs subject to XCI. Genes escaping XCI tended to have similar epigenetic marks between the Xi and Xa, except for H3K27me3 which is higher on the Xi. The increased Xi H3K27me3 at genes escaping XCI may be why escape genes can have as low as 10% expression from the Xi compared to the Xa; the other marks being Xa-like allows some expression to continue. Genes subject to XCI tend to have higher heterochromatic marks on the Xi and lower euchromatic marks, which supports these genes not being expressed on the Xi. Other studies have seen a general enrichment of heterochromatic marks at genes subject to XCI and euchromatic marks enriched at genes escaping from XCI (reviewed in (Balaton and Brown 2016)).

Across all our epigenetic analyses, DNAme stood out as being the most reflective of a gene's XCI status. The euchromatic mark H3K4me3 was the histone mark that was most significant for differentiating genes escaping vs subject to XCI, while the heterochromatic mark H3K27me3 had the largest Xi:Xa difference and was the most predictive histone mark for our epigenetic predictor. A previous study, which used a random forest model to predict XCI statuses and silencing timing in mice, found that DNAme often ranked below many of their histone marks, including H3K27ac, H3K4me1 and H3K27me3 (De Andrade E Sousa et al. 2019). In addition to the possible species differences, their model may not rely on DNAme as much due to them including numerous genomic features and transcription factor binding annotations, with distance to *Xist* and gene density being their top two features. We chose not to incorporate more features as we wanted to find XCI differences across samples, and therefore wanted to only include features that were sample-specific. Our model however does not account for the interaction genomic features may have with the epigenetic marks examined here. If we had genetic and epigenetic data for a large number of samples, that would have been ideal for determining the genetic and epigenetic determinants of XCI.

In this study we used multiple different methods to predict the XCI status of genes and examined how different epigenetic marks changed across genes with differing XCI statuses. We found similar distributions of genes escaping, variably escaping or subject to XCI across our DNAme analyses as our previous Xi/Xa expression analysis (Balaton et al. 2021), while our epigenetic predictor predicted twice as many genes as subject to XCI with similar levels of genes escaping and variably escaping from XCI. A large proportion of the additional genes found subject to XCI by our epigenetic predictor may in fact be inactive on both the Xa and Xi, as 68% of them had a median expression across samples under 0.1 RPKM. The threshold at which to call genes as 'variable' in XCI status is arbitrary. We used a threshold requiring 33% of samples to have each XCI status to call variable escape from XCI in our DNAme and epigenetic predictors as used previously (Balaton et al. 2021; Carrel and Willard 2005), with the greater number of samples with each XCI status improving the power of our statistical tests comparing epigenetic marks across samples with opposite XCI statuses. Decreasing the threshold increased the number of genes variably escaping from XCI and the number of epigenetic marks that were significant in at least one gene, but decreased the percentage of genes significant for DNAme which was the only mark ever significant for over 50% of genes in a dataset.

We observed that variable escape from XCI was regulated at the level of single genes, with adjacent genes varying their XCI status independently. In contrast, a study in mice found clusters of genes which variably escape across their three cell lines, with adjacent genes often having the same XCI status across lines (Marks et al. 2015). They also found that these clusters colocalize with TADs, with one line having the majority of a TAD escaping XCI and another line having only part of it escaping. An interesting candidate regulator of regional control is SMCHD1; regions enriched with variably escaping genes were upregulated when *SMCHD1* was disrupted, while genes which constitutively escaped from XCI were not affected (C. Y. Wang et al. 2019). This was found in mice with *Smchd1* knocked-out but not in human patients with heterozygous *SMCHD1* mutations. Another study found variants with low expression of *SMCHD1*, *ZSCAN9* and *HBG2/TRIM6* associated with hypomethylation of X-linked CpG islands, with affected islands enriched near genes that variably escape from XCI (Luijk et al. 2018). Combined with our evidence, we suggest that some genes may be regulated as clusters while others are regulated individually. It could also be that XCI status is mostly regulated at the domain level, but the domain featured in Figure 4.5 and other variably escaping domains are at a threshold where individual genes can have either XCI status based on local factors.

One drawback to this study is that many of our results relied on cancer datasets which may have differences from healthy tissues and epigenetic instability. DNA methyltransferases and histone modifying enzymes are commonly mutated in cancer, and 5-10% of CpG islands which should be unmethylated become methylated (reviewed in (Dawson and Kouzarides 2012)). We would expect the changes from epigenetic instability to differ between cancers and cause more genes to variably escape from XCI, however we saw a similar number of genes variably escaping from XCI in the CEMT cancer dataset as in the healthy CREST dataset. This may however be why we see a vast increase in the number of genes variably escaping from XCI when we lower our threshold to require only 5% of samples to differ in their XCI status. Despite these problems, we used the CEMT dataset because it had a standardized set of epigenetic marks across many samples and the clonality of cancer allowed us to examine expression and DNAme allelically. We found that other datasets, even within IHEC, did not always have all the marks from the same samples, were lacking females or sex labels or had mislabeled sex.

We had some discordant calls between the various methods employed here. Genes could be falsely called as subject to XCI in the Xi/Xa expression-based analysis if the alternate SNP allele no longer mapped to the same region or if heterozygosity was miscalled. DNAme has been seen misregulated in many cancers (Dawson and Kouzarides 2012). The cancer cells could have mutations mosaic between the parts sampled for different analyses. Our epigenetic predictor did not obtain 100% accuracy on its training data so we expect some of the calls made with it to be false, while the training data could also have false calls further hurting its ability to make accurate XCI status calls.

4.5 Conclusions

Our study has shown that most of the epigenetic marks assayed (H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K27me3, H3K36me3, and DNAme) had male-female differences, while fewer marks were significantly different at genes with opposite XCI statuses. To account for dosage differences, we calculated the contributions of the Xa and Xi to these sex-biased modifications. Genes subject to XCI had higher heterochromatic marks on the Xi and lower euchromatic marks on the Xi while genes escaping XCI tended to have equal levels of marks on the Xa and Xi, except for H3K27me3 which was high on the Xi. Genes which escape from XCI are not expressed at 100% of the level of the Xa, which supports this conclusion. No mark other than DNAme was very accurate at predicting XCI status; however, combining all the epigenetic marks together allowed us to call XCI status for genes without CpG islands, where DNAme alone is unable to establish a call. Most marks were significantly different between samples escaping vs subject to XCI at variably escaping genes, but which marks were significant was not consistent between genes and no mark was significant across all the genes. This may be due to variably escaping genes having multiple ways in which they are regulated. DNAme intermediate to what is expected for genes escaping vs subject to XCI is enriched at variably escaping genes and is mostly due to inconsistent DNAme on the Xi. Neighboring variably escaping genes were seen to regulate their XCI status independently from each other, suggesting local regulatory elements. Overall, we see that escape from XCI is influenced by local regulatory elements as well as chromatin modifications which can be independent of each other. Understanding how genes escape from XCI will further our understanding of epigenetics in general and may allow us to control which genes are escaping from XCI and rescue X-linked mutations in females.

Chapter 5: Discussion

5.1 Thesis Findings

Through the various analyses featured in this thesis, I have discovered many differences and similarities between genes with the same XCI status, and between genes with opposite XCI statuses. The human consensus XCI status calls made in chapter 2 were ideal for testing the later methods of determining a species or sample-specific XCI status and were used for my comparisons of epigenetic marks. Figure 3.1A and Supplemental Figure 3.1 show the good concordance between our consensus XCI status calls, our Xi/Xa expression-based calls and our DNAme based calls for most of the samples analyzed. Figures 4.2A, 4.3A and 4.4B-C compare my later XCI status calls to my consensus calls. Xi/Xa expression remains the standard for XCI status calls, with table 2.2 showing that the Carrel SNP study (Carrel and Willard 2005) has the least discordancies with my consensus status calls; however the limitation of requiring skewed XCI status and heterozygous SNPs limits its potential for comparing XCI status across all genes and all samples. DNAme enables us to examine XCI without those requirements, which allowed us to examine XCI across species and all our human samples. Expanding our XCI status predictor with other epigenetic marks allowed us to make XCI status calls for genes without CpG islands, however these became less accurate without CpG islands and for genes with low expression.

This study expanded upon what has been seen for genes variably escaping from XCI. Previous investigators have seen variable escape across individuals, tissues, populations, TSSs and cells (Cotton et al. 2013; Goto and Kimura 2009; Hagen et al. 2020; Tukiainen et al. 2017) . Here we show increased evidence of variable escape across individuals and tissues, while also observing discordancy across studies (which may be due to variable escape between the samples, populations or tissues used) and discordancy across species (which may have similar mechanisms to variable escape across populations). We found 16 genes highly discordant across species, with 5 of these having an obvious phylogenetic pattern, and some notably coinciding with loss of the gene’s Y homolog. The XCI status calls in chapter 3 did not find many genes variably escaping within species, with the DNAme based method finding less variably escaping genes than the Xi/Xa expression-based method.

The Cotton DNAme study (Cotton et al. 2015) examined in chapter 2 also did not call many genes as variably escaping except within a few tissues, so this may be due to DNAme not associating with XCI status completely. In chapter 4 we find a smaller proportion of genes variably escaping across samples by DNAme and by the epigenetic predictor (which is strongly affected by DNAme), while still seeing many genes variably escaping within specific tissues. We did, however, see DNAme as the most commonly significant epigenetic mark differing between samples escaping and subject to XCI by our Xi/Xa expression based XCI status calls; DNAme may be differing in 50% of variably escaping genes, although with only 8 variably escaping genes and only 8 informative samples it is difficult to tell whether that trend would continue across a larger sampling of genes. Other epigenetic marks were also seen to vary with a genes XCI status, however there was no mark that was significant across all the variably escaping genes identified by Xi/Xa expression (Table 4.1, Figure 4.2). Similarly when looking at which epigenetic marks differed at genes in a variably escaping region, all but one gene had significant differences between samples escaping vs subject to XCI in two of the three heterochromatic marks (H3K9me3, H3K27me3 or DNAme) (Figure 4.5, Supplemental Figure S4.12).

The gene *UBA1* was previously reported to have TSS-specific escape from XCI in humans (Goto and Kimura 2009), and here we see this trend continue in chimp and horse, but not in mouse, cow or pig. In chapter 4 we see other genes with TSS specific XCI status and also that some epigenetic marks differ between the TSS which escapes XCI and the one which is subject to XCI; many of these marks also differed in males however so this is likely on the Xa and may not be linked to XCI. Having epigenetic differences on the Xa prior to XCI may help pre-determine the TSSs XCI status, however, as it has been shown that genes with an enrichment for heterochromatic marks prior to XCI are more likely to be subject to XCI (Cotton et al. 2014; Kelsey et al. 2015; Loda et al. 2017).

An ongoing question has been whether XCI status is regulated at the level of domains. It was previously shown that genes escaping from XCI tend to be clustered in domains (Marks et al. 2015). Our results in chapter 2 agree with this, but also add that genes which variably escape from XCI are found clustered between the domains of genes escaping from XCI and domains of genes subject to XCI. Results in chapter 3 expand this by finding that the domains of genes escaping from XCI are concordant across species and also that genes which are discordant across species cluster with other genes with a similar pattern of discordancy across the same species. In Chapter 4 we examined how variably escaping genes are regulated within a domain of variably escaping genes, and we found that the genes vary their XCI status independently of their neighbours. Combined with our earlier data we propose that these variably escaping domains have some feature which predisposes the genes within them to be able to escape or be subject to XCI, and that there is some sample and gene specific genetic or epigenetic effect which decides which genes will be escaping XCI and which will be subject to XCI in any given sample.

5.2 Limitations

There are a few limitations with the studies featured in this thesis. In chapter 2, my consensus XCI status calls are limited to one per gene instead of being TSS specific. This erased all the genes with TSS-specific XCI statuses seen in chapters 3 and 4. This needed to be done however as three of the four studies used to generate the consensus calls were expression-based and featured shared exons where it was not possible to determine which TSS was dominant.

Another limitation was that cancer cells were used in chapters 3 and 4, and these may have genetic and epigenetic aberrations not found in healthy cells (Larson et al. 2017). We compared XCI status calls between cancer and non-cancer samples and found few discordancies (Table S4.12). The cancer cells had more variation in DNAme between adjacent CpGs within the same CpG island than healthy cells did (Figure S4.6).

I saw some discordancies between XCI status calls made using Xi/Xa expression and DNAme within the same samples (Figures 3.1, S3.1, S3.2). In human these were all cancer based due to the need for skewed Xi choice, but in mouse these were taken from healthy mice with *Xist* knocked out on one allele to create skewed Xi choice. There were considerably less discordancies in mouse, however. The human discordancies were also found mostly in three samples, which were removed from further analysis. Later analyses featuring human Xi/Xa expression based XCI status calls were limited due to the low sample size caused by selecting for only skewed samples and further restricting to only samples with a heterozygous polymorphism within the gene of interest.

Chapter 3 had further limitations in that it used different tissues and different methods to determine DNAme across the species examined. Some of the genes found discordant across species could instead be variably escaping between the tissues used between discordant species. From these results we cannot tell if the high number of genes variably escaping in horse is something specific to the species or if it is due to the use of RRBS data to determine XCI status. Similarly, the low number of variably escaping genes found in many species could be due to a single tissue being analyzed or due to a low number of samples. We do find lower than expected levels of variable escape in human here, even though our analysis included multiple tissues and a high number of samples. In the 450k array DNAme based study featured in chapter 2 (Cotton et al. 2015) they found less variable escape from XCI than the other studies, however, no genes consistently variably escaped across all tissues. In the same study only two genes had tissue specific escape from XCI, and many genes only variably escaped within one or two tissues but had a consistent XCI status in the remainder of the 37 tissues examined. This suggests that DNAme may not always follow a gene's XCI status across samples for variably escaping genes. Figure 4.3B shows examples of genes with DNAme varying at a variably escaping gene so DNAme is still capable of reflecting some change at variably escaping genes. An additional limitation for chapter 3 is that I used 450k array data for most of the XCI status calls in primates. I showed that XCI status calls made by the 450k array and WGBS were fairly concordant, but using the human 450k array to make calls in other species is limited to only using probes which map well between species, which may select for genes with conserved XCI status. This causes some doubt in my finding that 97% of genes had a completely conserved XCI status across primates, this number may be lower once less conserved CpG islands are included.

5.3 Outstanding questions

The elements which regulate variable escape from XCI remain unknown. In chapter 2, I showed that variably escaping genes cluster into domains between genes which are subject to XCI and those escaping XCI and in chapter 4 I expanded on this to show that within the largest of these domains, neighbouring genes vary independently of one another. I also found tissue-specific escape from XCI which must be driven by an epigenetic change or tissue-specific transcription factor binding. I showed that many epigenetic marks follow a change in XCI status between samples, but these marks are not always consistent and so there may be multiple variants of what causes the change. Currently we do not know whether the change in histone marks causes the change in XCI status, or whether the change in histone marks simply reflects the change in XCI status.

5.4 Future directions

Many of the analyses featured herein could be expanded in the future. As these analyses were all based off datasets found online, and as more data is generated and uploaded online, the power of these analyses can be improved. This is most pronounced with my cross-species analysis, as the number of mammalian species with DNAme available was the reason why only 12 species were analyzed. As the cost of sequencing drops, more studies will generate whole-genome datasets such as WGBS, which allows us to make XCI status calls at more genes, and in a less biased way, than using the human 450k array. Scientists and funding sources are also becoming increasingly interested in including both sexes in science (Clayton 2016) so in the future it may be easier to find datasets with properly labeled female and male samples. This work could also be expanded by generating our own WGBS datasets. The benefits to this are that we would then have a more standardized approach, with all species being examined in the same tissues, and with a similar number of samples. The problem here being the high cost and forcing the choice between having a larger sample size and being able to see variable escape from XCI or having less samples per species and having more species analyzed.

Another analysis that would be a useful expansion of chapter 3 would be to further examine genetic differences between species near genes with discordant XCI status across species. The problem here is that you would need many species analyzed in order to have statistical power to identify which evolutionary changes were significantly associated with a change in XCI status. For genes with an order specific XCI status, such as the many primate-specific escape genes or the one artiodactyl-specific subject gene, it will be difficult to determine which genetic differences cause the change in XCI status and which are coincidentally limited to that specific order through evolution. Also, for the genes which have independently evolved a change in XCI status, we do not know whether this would have been separate mutations or a common site which is frequently mutated across evolution.

To find genes which vary their XCI status in humans, it would be useful to have XCI status calls in the same samples as whole-genome sequencing. We had whole-genome sequencing for our Xi/Xa expression-based calls in chapter 3, but with only 8 samples we did not have the power to search for polymorphisms. With enough whole-genome sequencing data, we would expect to find any genetic polymorphisms with a major effect on XCI status. Whole-genome data would also allow us to see polymorphisms with a smaller effect size and see if there is enrichment of polymorphisms at certain regulatory regions or at repetitive elements. We have seen that variably escaping genes change their XCI status independently of neighboring genes, and so we would expect these causative polymorphisms to be found nearby in 2D or 3D space.

Whole-genome sequence of samples with skewed Xi choice would also be useful for allelic ChIP-seq, allowing us to see where each mark is specifically bound to the Xi and Xa. Our current estimates of Xi and Xa binding for histone marks rely on subtracting male data from female data, but there are other factors which may have a sex-specific effect on epigenetic marks or even differ between females. Allelic expression would allow us to see epigenetic marks as they differ between the Xa and Xi in the same cellular environment. We were able to do this for DNAme, but the CEMT samples for which we had whole-genome data did not have deep enough ChIP-seq for us to be comfortable examining ChIP-seq allelically.

Another direction that this research leads, is to use a different model system to study escape from XCI, as mouse seems to be an outlier for the number of genes escaping from XCI. Mice and rats also have imprinted XCI, whereas most other eutherian mammals do not (as summarized in section 3.1). As I am unfamiliar with many other mammalian model systems I am unable to propose a new model with certainty, but rabbits are small and do not have imprinted XCI (Okamoto et al. 2011), so they may be a good model system. A recent publication has generated

WGBS for 24 female rabbit samples, which could be used to generate XCI status calls (Shao et al. 2020).

Many groups are attempting to reactivate the inactive copy for genes subject to XCI, with the hopes of reactivating the healthy Xi copy of a gene with a deficient copy on the Xa (Halmai et al. 2020; Leko et al. 2018; Przanowski et al. 2019). I think the most promising method for medical use is to use targeted epigenome modifications to selectively activate the one disease-associated gene instead of trying to reactivate the whole Xi (Halmai et al. 2020). Chapter 4 shows that not all genes with the same XCI status will have the same epigenetic marks. Marks commonly found to significantly differ with XCI status such as DNAme, H3K27me3 and H3K4me3 may be promising targets, with DNAme already having studies targeting it (Halmai et al. 2020). My work also suggests that each gene target may need a different set of epigenetic marks targeted in order for reactivation to occur.

5.5 Conclusion

Overall, the studies included in this thesis increase the number and confidence of XCI status calls in humans and other mammalian species. I also found epigenetic effects which are implicated in the regulation of XCI and which genes escape from it. Many studies, including my own, have already used the consensus XCI status calls derived in chapter 2 and I hope that many studies will benefit from the cross-species XCI status calls presented in chapter 3.

I have shown the benefit of using DNAme to call XCI status by using it to generate XCI status calls across species and also expanded our XCI status calls using an epigenetic predictor with the core histone marks featured in IHEC. My discovery of domains of variably escaping genes and of genes which are discordant across species invites other studies to determine how these domains are regulated. I showed here that the variably escaping genes within these domains appear to be independently regulated. The variably escaping region featured in chapter 4 contains genes with XCI status changing within a limited genomic space, which may help narrow down the regions responsible for regulating differences in XCI status. The genes found with TSS specific XCI statuses identified here are also a promising target for this future research.

The conservation of various XCI implicated features across species detailed in chapter 3 helps support conclusions made by others about how XCI is regulated, and our epigenetic analysis in chapter 4 furthers our understanding of how XCI is regulated by showing that H3K27me3 was higher on the Xi than the Xa even at genes escaping from XCI. These results build on our knowledge of which genes escape from XCI and how they are regulated, will help future studies determine how escape from XCI affects phenotypes and disease, and may aid efforts to inactivate or reactivate genes epigenetically as a genetic medicine.

References

Almeida, Luiz Gonzaga et al. 2009. “CTdatabase: A Knowledge-Base of High-Throughput and Curated Data on Cancer-Testis Antigens.” *Nucleic Acids Research* 37(SUPPL. 1).

Almeida, Mafalda, Joseph S. Bowness, and Neil Brockdorff. 2020. “The Many Faces of Polycomb Regulation by RNA.” *Current Opinion in Genetics and Development* 61: 53–61. /pmc/articles/PMC7653676/?report=abstract (December 16, 2020).

De Andrade E Sousa, Lisa Barros et al. 2019. “Kinetics of Xist-Induced Gene Silencing Can Be Predicted from Combinations of Epigenetic and Genomic Features.” *Genome Research* 29(7): 1087–99. https://pubmed.ncbi.nlm.nih.gov/31175153/ (December 18, 2020).

Arnold, Arthur P., and Xuqi Chen. 2009. “What Does the ‘Four Core Genotypes’ Mouse Model Tell Us about Sex Differences in the Brain and Other Tissues?” *Frontiers in Neuroendocrinology* 30(1): 1–9. /pmc/articles/PMC3282561/?report=abstract (December 17, 2020).

Bailey, Jeffrey A., Laura Carrel, Aravinda Chakravarti, and Evan E. Eichler. 2000. “Molecular Evidence for a Relationship between LINE-1 Elements and X Chromosome Inactivation: The Lyon Repeat Hypothesis.” *Proceedings of the National Academy of Sciences of the United States of America* 97(12): 6634–39. https://pubmed.ncbi.nlm.nih.gov/10841562/ (February 2, 2021).

Balaton, Bradley P., and Carolyn J. Brown. 2016. “Escape Artists of the X Chromosome.” *Trends in Genetics* 32(6): 348–59.

Balaton, Bradley P., Allison M. Cotton, and Carolyn J. Brown. 2015. “Derivation of Consensus Inactivation Status for X-Linked Genes from Genome-Wide Studies.” *Biology of Sex Differences* 6(1). https://pubmed.ncbi.nlm.nih.gov/26719789/ (December 15, 2020).

Balaton, Bradley P., Thomas Dixon-McDougall, Samantha B. Peeters, and Carolyn J. Brown. 2018. “The EXceptional Nature of the X Chromosome.” *Human molecular genetics* 27(R2): R242–49. https://pubmed.ncbi.nlm.nih.gov/29701779/ (December 18, 2020).

Balaton, Bradley P, and Carolyn J Brown. 2021. “Contribution of Epigenetic Changes to Escape from X-Chromosome Inactivation.” *bioRxiv*: 2021.03.03.433635. https://doi.org/10.1101/2021.03.03.433635 (March 4, 2021).

Balaton, Bradley P, Oriol Fornes, Wyeth W Wasserman, and Carolyn J Brown. 2021. “Cross-Species Examination of X-Chromosome Inactivation Highlights Domains of Escape from Silencing.” *Epigenetics and Chromatin*: 2020.12.04.412197. https://doi.org/10.1101/2020.12.04.412197 (December 16, 2020).

Barrett, Tanya et al. 2013. “NCBI GEO: Archive for Functional Genomics Data Sets - Update.” *Nucleic Acids Research* 41(D1).

Barski, Artem et al. 2007. “High-Resolution Profiling of Histone Methylations in the Human Genome.” *Cell* 129(4): 823–37. https://pubmed.ncbi.nlm.nih.gov/17512414/ (February 2, 2021).

Bellott, Daniel W. et al. 2014. “Mammalian y Chromosomes Retain Widely Expressed Dosage-Sensitive Regulators.” *Nature* 508(7497): 494–99.

Benjamini, Yoav, and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society: Series B (Methodological)* 57(1): 289–300.

Bennett-Baker, Pamela E., Jodi Wilkowski, and David T. Burke. 2003. “Age-Associated Activation of Epigenetically Repressed Genes in the Mouse.” *Genetics* 165(4): 2055–62. /pmc/articles/PMC1462878/?report=abstract (December 15, 2020).

Berletch, Joel B. et al. 2015. “Escape from X Inactivation Varies in Mouse Tissues.” *PLoS Genetics* 11(3).

De Bonis, M. L. et al. 2006. “Maintenance of X- and Y-Inactivation of the Pseudoautosomal (PAR2) Gene SPRY3 Is Independent from DNA Methylation and Associated to Multiple Layers of Epigenetic Modifications.” *Human Molecular Genetics* 15(7): 1123–32.

Brown, Carolyn J., and Huntington F. Willard. 1989. “Noninactivation of a Selectable Human X-Linked Gene That Complements a Murine Temperature-Sensitive Cell Cycle Defect.” *American Journal of Human Genetics* 45(4): 592–98. /pmc/articles/PMC1683489/?report=abstract (December 18, 2020).

Brown, Garth R. et al. 2015. “Gene: A Gene-Centered Information Resource at NCBI.” *Nucleic Acids Research* 43(D1): D36–42.

Buenrostro, Jason D. et al. 2013. “Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position.” *Nature Methods* 10(12): 1213–18.

Bujold, David et al. 2016. “The International Human Epigenome Consortium Data Portal.” *Cell Systems* 3(5): 496-499.e2.

Carrel, Laura, and Carolyn J. Brown. 2017. “When the Lyon(Ized Chromosome) Roars: Ongoing Expression from an Inactive X Chromosome.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 372(1733).

Carrel, Laura, and Huntington F. Willard. 1999. “Heterogeneous Gene Expression from the Inactive X Chromosome: An X-Linked Gene That Escapes X Inactivation in Some Human Cell Lines but Is Inactivated in Others.” *Proceedings of the National Academy of Sciences of the United States of America* 96(13): 7364–69.

———. 2005. “X-Inactivation Profile Reveals Extensive Variability in X-Linked Gene Expression in Females.” *Nature* 434(7031): 400–404.

CenterWall, WR;, and K Benirschke. 1975. “An Animal Model for the XXY Klinefelter’s Syndrome in Man: Tortoiseshell and Calico Male Cats.” *American journal of veterinary research* 9: 1275–80.

Changolkar, Lakshmi N. et al. 2010. “Genome-Wide Distribution of MacroH2A1 Histone Variants in Mouse Liver Chromatin.” *Molecular and Cellular Biology* 30(23): 5473–83. https://pubmed.ncbi.nlm.nih.gov/20937776/ (December 17, 2020).

Chen, Bo, Radu V Craiu, and Lei Sun. 2018. “Bayesian Model Averaging for the X-Chromosome Inactivation Dilemma in Genetic Association Study.” *Biostatistics* 21(2): 319–35. https://academic.oup.com/biostatistics/advance-article/doi/10.1093/biostatistics/kxy049/5105904 (December 16, 2020).

Chen, Chih Yu et al. 2016. “YY1 Binding Association with Sex-Biased Transcription Revealed through X-Linked Transcript Levels and Allelic Binding Analyses.” *Scientific Reports* 6(1): 1–14. www.nature.com/scientificreports (December 16, 2020).

Chen, Zhiyuan et al. 2016. “Global Assessment of Imprinted Gene Expression in the Bovine Conceptus by next Generation Sequencing.” *Epigenetics* 11(7): 501–16.

Ciccodicola, Alfredo et al. 2000. “Differentially Regulated and Evolved Genes in the Fully Sequenced Xq/Yq Pseudoautosomal Region.” *Human Molecular Genetics* 9(3): 395–401. https://pubmed.ncbi.nlm.nih.gov/10655549/ (December 17, 2020).

Clayton, Janine Austin. 2016. “Studying Both Sexes: A Guiding Principle for Biomedicine.” *FASEB Journal* 30(2): 519–24. /pmc/articles/PMC4714546/?report=abstract (February 1, 2021).

Clemson, Christine Moulton, Jennifer C. Chow, Carolyn J. Brown, and Jeanne Bentley Lawrence. 1998. “Stabilization and Localization of Xist RNA Are Controlled by Separate Mechanisms and Are Not Sufficient for X Inactivation.” *Journal of Cell Biology* 142(1): 13–23. https://pubmed.ncbi.nlm.nih.gov/9660859/ (December 18, 2020).

Cotton, Allison M. et al. 2013. “Analysis of Expressed SNPs Identifies Variable Extents of Expression from the Human Inactive X Chromosome.” *Genome Biology* 14(11). https://pubmed.ncbi.nlm.nih.gov/24176135/ (December 15, 2020).

———. 2014. “Spread of X-Chromosome Inactivation into Autosomal Sequences: Role for DNA Elements, Chromatin Features and Chromosomal Domains.” *Human Molecular Genetics* 23(5): 1211–23. https://pubmed.ncbi.nlm.nih.gov/24158853/ (December 16, 2020).

———. 2015. “Landscape of DNA Methylation on the X Chromosome Reflects CpG Density, Functional Chromatin State and X-Chromosome Inactivation.” *Human Molecular Genetics* 24(6): 1528–39.

Couldrey, C. et al. 2017. “Bovine Mammary Gland X Chromosome Inactivation.” *Journal of Dairy Science* 100(7): 5491–5500.

Craig, Ian W. et al. 2004. “Application of Microarrays to the Analysis of the Inactivation Status of Human X-Linked Genes Expressed in Lymphocytes.” *European Journal of Human Genetics* 12(8): 639–46.

Davidson, Ronald G., Harold M. Nitowsky, and Barton Childs. 1963. “DEMONSTRATION OF TWO POPULATIONS OF CELLS IN THE HUMAN FEMALE HETEROZYGOUS FOR GLUCOSE-6-PHOSPHATE DEHYDROGENASE VARIANTS.” *Proceedings of the National Academy of Sciences of the United States of America* 50(3): 481–85.

https://pubmed.ncbi.nlm.nih.gov/14067093/ (December 15, 2020).

Davis, Carrie A. et al. 2018. “The Encyclopedia of DNA Elements (ENCODE): Data Portal Update.” *Nucleic Acids Research* 46(D1): D794–801. https://pubmed.ncbi.nlm.nih.gov/29126249/ (December 16, 2020).

Dawson, Mark A., and Tony Kouzarides. 2012. “Cancer Epigenetics: From Mechanism to Therapy.” *Cell* 150(1): 12–27. https://pubmed.ncbi.nlm.nih.gov/22770212/ (February 2, 2021).

Deng, Xinxian, Joel B. Berletch, Di K. Nguyen, and Christine M. Disteche. 2014. “X Chromosome Regulation: Diverse Patterns in Development, Tissues and Disease.” *Nature Reviews Genetics* 15(6): 367–78.

Dixon-McDougall, Thomas, and Carolyn Brown. 2015. “The Making of a Barr Body: The Mosaic of Factors That EXIST on the Mammalian Inactive X Chromosome1.” *Biochemistry and Cell Biology* 94(1): 56–70. https://pubmed.ncbi.nlm.nih.gov/26283003/ (December 16, 2020).

Dixon, Jesse R. et al. 2012. “Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions.” *Nature* 485(7398): 376–80.

Duncan, Christopher G. et al. 2018. “Dosage Compensation and DNA Methylation Landscape of the X Chromosome in Mouse Liver.” *Scientific Reports* 8(1).

Dunford, Andrew et al. 2017. “Tumor-Suppressor Genes That Escape from X-Inactivation Contribute to Cancer Sex Bias.” *Nature Genetics* 49(1): 10–16. https://doi.org/10.1038/ng.3726.

Engreitz, Jesse M. et al. 2013. “The Xist LncRNA Exploits Three-Dimensional Genome Architecture to Spread across the X Chromosome.” *Science* 341(6147).

https://pubmed.ncbi.nlm.nih.gov/23828888/ (December 17, 2020).

Epiphanio, Tatiane Moreno Ferrarias et al. 2019. "Global DNA Methylation of Peripheral Blood Leukocytes from Dogs Bearing Multicentric Non-Hodgkin Lymphomas and Healthy Dogs: A Comparative Study." *PLoS ONE* 14(3): e0211898. https://doi.org/10.1371/journal.pone.0211898 (December 16, 2020).

Filippova, Galina N. et al. 2005. "Boundaries between Chromosomal Domains of X Inactivation and Escape Bind CTCF and Lack CpG Methylation during Early Development." *Developmental Cell* 8(1): 31–42.

Fishilevich, Simon et al. 2017. "GeneHancer: Genome-Wide Integration of Enhancers and Target Genes in GeneCards." *Database : the journal of biological databases and curation* 2017. https://pubmed.ncbi.nlm.nih.gov/28605766/ (February 2, 2021).

Flaquer, Antonia, Gudrun A. Rappold, Thomas F. Wienker, and Christine Fischer. 2008. "The Human Pseudoautosomal Regions: A Review for Genetic Epidemiologists." *European Journal of Human Genetics* 16(7): 771–79. www.nature.com/ejhg (December 17, 2020).

Garieri, Marco et al. 2018. "Extensive Cellular Heterogeneity of X Inactivation Revealed by Single-Cell Allele-Specific Expression in Human Fibroblasts." *Proceedings of the National Academy of Sciences of the United States of America* 115(51): 13015–20. /pmc/articles/PMC6304968/?report=abstract (December 18, 2020).

Gartler, SM., Dyer KA., Marshall Graves JA., Rocchi M. 1985. "A Two Step Model for Mammalian X-Chromosome Inactivation." *Prog Clin Biol Res* 198: 96–102.

Giorgetti, Luca et al. 2016. "Structural Organization of the Inactive X Chromosome in the Mouse." *Nature* 535(7613): 575–79.

Goto, Yuji, and Hiroshi Kimura. 2009. "Inactive X Chromosome-Specific Histone H3

Modifications and CpG Hypomethylation Flank a Chromatin Boundary between an X-Inactivated and an Escape Gene." *Nucleic Acids Research* 37(22): 7416–28.

Hacisuleyman, Ezgi et al. 2014. "Topological Organization of Multichromosomal Regions by the Long Intergenic Noncoding RNA Firre." *Nature Structural and Molecular Biology* 21(2): 198–206.

Hagen, Sven Hendrik et al. 2020. "Heterogeneous Escape from X Chromosome Inactivation Results in Sex Differences in Type I IFN Responses at the Single Human PDC Level." *Cell reports* 33(10): 108485. http://www.ncbi.nlm.nih.gov/pubmed/33296655 (December 18, 2020).

Halmai, Julian A.N.M. et al. 2020. "Artificial Escape from XCI by DNA Methylation Editing of the CDKL5 Gene." *Nucleic Acids Research* 48(5): 2372–87. https://academic.oup.com/nar/article/48/5/2372/5700548 (March 8, 2021).

Harrow, Jennifer et al. 2006. "GENCODE: Producing a Reference Annotation for ENCODE." *Genome biology* 7 Suppl 1.

———. 2012. "GENCODE: The Reference Human Genome Annotation for the ENCODE Project." *Genome Research* 22(9): 1760–74.

He, Yupeng, and Joseph R. Ecker. 2015. "Non-CG Methylation in the Human Genome." *Annual Review of Genomics and Human Genetics* 16: 55–77.

Hernando-Herraez, Irene et al. 2013. "Dynamics of DNA Methylation in Recent Human and Great Ape Evolution." *PLoS Genetics* 9(9).

Hinrichs, Angela S. et al. 2006. "The UCSC Genome Browser Database: Update 2006." *Nucleic acids research* 34(Database issue).

Horvath, Lindsay M., Nan Li, and Laura Carrel. 2013. "Deletion of an X-Inactivation Boundary

Disrupts Adjacent Gene Silencing." *PLoS Genetics* 9(11). /pmc/articles/PMC3836711/?report=abstract (December 18, 2020).

Hothorn, Torsten, Frank Bretz, and Peter Westfall. 2008. "Simultaneous Inference in General Parametric Models." *Biometrical Journal* 50(3): 346–63.

Jiang, Cizhong et al. 2007. "Features and Trend of Loss of Promoter-Associated CpG Islands in the Human and Mouse Genomes." *Molecular Biology and Evolution* 24(9): 1991–2000. https://pubmed.ncbi.nlm.nih.gov/17591602/ (December 16, 2020).

Johnston, Colette M. et al. 2008. "Large-Scale Population Study of Human Cell Lines Indicates That Dosage Compensation Is Virtually Complete." *PLoS Genetics* 4(1): 0088–0098.

Karolchik, Donna et al. 2004. "The UCSC Table Browser Data Retrieval Tool." *Nucleic Acids Research* 32(DATABASE ISS.).

Kelsey, Angela D. et al. 2015. "Impact of Flanking Chromosomal Sequences on Localization and Silencing by the Human Non-Coding RNA XIST." *Genome Biology* 16(1). https://pubmed.ncbi.nlm.nih.gov/26429547/ (December 17, 2020).

Keniry, Andrew, and Marnie E. Blewitt. 2018. "Studying X Chromosome Inactivation in the Single-Cell Genomic Era." *Biochemical Society Transactions* 46(3). https://pubmed.ncbi.nlm.nih.gov/29678955/ (December 18, 2020).

Kent, W. J. et al. 2002. "The Human Genome Browser at UCSC." *Genome Research* 12(6): 996–1006.

———. 2010. "BigWig and BigBed: Enabling Browsing of Large Distributed Datasets." *Bioinformatics* 26(17): 2204–7. /pmc/articles/PMC2922891/?report=abstract (February 2, 2021).

Keown, Christopher L. et al. 2017. "Allele-Specific Non-CG DNA Methylation Marks Domains of Active Chromatin in Female Mouse Brain." *Proceedings of the National Academy of Sciences of the United States of America* 114(14): E2882–90.

Kim, Daehwan et al. 2019. "Graph-Based Genome Alignment and Genotyping with HISAT2 and HISAT-Genotype." *Nature Biotechnology* 37(8): 907–15.

Krueger, Felix, and Simon R. Andrews. 2011. "Bismark: A Flexible Aligner and Methylation Caller for Bisulfite-Seq Applications." *Bioinformatics* 27(11): 1571–72. https://pubmed.ncbi.nlm.nih.gov/21493656/ (December 16, 2020).

Kucera, Katerina S. et al. 2011. "Allele-Specific Distribution of RNA Polymerase II On Female X Chromosomes." *Human Molecular Genetics* 20(20): 3964–73. /pmc/articles/PMC3177651/?report=abstract (December 17, 2020).

Kuhn, Max. 2008. "Caret Package." *Journal of Statistical Software* 28(5).

Lahn, Bruce T., and David C. Page. 1999. "Four Evolutionary Strata on the Human X Chromosome." *Science* 286(5441): 964–67.

Larson, Nicholas B. et al. 2017. "An Integrative Approach to Assess X-Chromosome Inactivation Using Allele-Specific Expression with Applications to Epithelial Ovarian Cancer." *Genetic Epidemiology* 41(8): 898–914. http://doi.wiley.com/10.1002/gepi.22091 (December 16, 2020).

Leko, Vid et al. 2018. "Pooled ShRNA Screen for Reactivation of MeCP2 on the Inactive X Chromosome." *Journal of Visualized Experiments* 2018(133). https://pubmed.ncbi.nlm.nih.gov/29553562/ (March 8, 2021).

Letunic, Ivica, and Peer Bork. 2007. "Interactive Tree Of Life (ITOL): An Online Tool for Phylogenetic Tree Display and Annotation." *Bioinformatics* 23(1): 127–28. https://pubmed.ncbi.nlm.nih.gov/17050570/ (December 16, 2020).

Li, Heng et al. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics* 25(16): 2078–79. /pmc/articles/PMC2723002/?report=abstract (December 16, 2020).

Li, Nan, and Laura Carrel. 2008. “Escape from X Chromosome Inactivation Is an Intrinsic Property of the Jarid1c Locus.” *Proceedings of the National Academy of Sciences of the United States of America* 105(44): 17055–60.

Lister, Ryan et al. 2013. “Global Epigenomic Reconfiguration during Mammalian Brain Development.” *Science* 341(6146).

Loda, Agnese et al. 2017. “Genetic and Epigenetic Features Direct Differential Efficiency of Xist-Mediated Silencing at X-Chromosomal and Autosomal Locations.” *Nature Communications* 8(1). https://pubmed.ncbi.nlm.nih.gov/28947736/ (December 18, 2020).

Luijk, René et al. 2018. “Autosomal Genetic Variation Is Associated with DNA Methylation in Regions Variably Escaping X-Chromosome Inactivation.” *Nature Communications* 9(1). https://pubmed.ncbi.nlm.nih.gov/30218040/ (February 2, 2021).

Luikenhuis, Sandra, Anton Wutz, and Rudolf Jaenisch. 2001. “Antisense Transcription through TheXist Locus Mediates Tsix Function in Embryonic Stem Cells.” *Molecular and Cellular Biology* 21(24): 8512–20. /pmc/articles/PMC100014/?report=abstract (January 11, 2021).

Lyon, Mary F. 1961. “Gene Action in the X-Chromosome of the Mouse (Mus Musculus L.).” *Nature* 190(4773): 372–73.

———. 1962. “Sex Chromatin and Gene Action in the Mammalian X-Chromosome.” *American Journal of Human Genetics* 14: 135–48.

———. 1998. “X-Chromosome Inactivation: A Repeat Hypothesis.” *Cytogenetics and Cell Genetics* 80(1–4): 133–37. https://pubmed.ncbi.nlm.nih.gov/9678347/ (December 18, 2020).

Mak, Winifred et al. 2004. “Reactivation of the Paternal X Chromosome in Early Mouse Embryos.” *Science* 303(5658): 666–69.

Marks, Hendrik et al. 2015. “Dynamics of Gene Silencing during X Inactivation Using Allele-Specific RNA-Seq.” *Genome Biology* 16(1).

Marshall Graves, Jennifer A., and Graham J. Young. 1982. “X-Chromosome Activity in Heterokaryons and Hybrids between Mouse Fibroblasts and Teratocarcinoma Stem Cells.” *Experimental Cell Research* 141(1): 87–97.

Melé, Marta et al. 2015. “The Human Transcriptome across Tissues and Individuals.” *Science* 348(6235): 660–65.

Van Der Meulen, Joni et al. 2015. “The H3K27me3 Demethylase UTX Is a Gender-Specific Tumor Suppressor in T-Cell Acute Lymphoblastic Leukemia.” *Blood* 125(1): 13–21.

Migeon, B. R. et al. 1981. “Adrenoleukodystrophy: Evidence for X Linkage, Inactivation, and Selection Favoring the Mutant Allele in Heterozygous Cells.” *Proceedings of the National Academy of Sciences of the United States of America* 78(8 I): 5066–70.

Miller, Andrew P., and Huntington F. Willard. 1998. “Chromosomal Basis of X Chromosome Inactivation: Identification of a Multigene Domain in Xp11.21-P11.22 That Escapes X Inactivation.” *Proceedings of the National Academy of Sciences of the United States of America* 95(15): 8709–14.

Mitterbauer, Gerlinde et al. 1999. “Clonality Analysis Using X-Chromosome Inactivation at the Human Androgen Receptor Gene (HUMARA): Evaluation of Large Cohorts of Patients with Chronic Myeloproliferative Diseases, Secondary Neutrophilia, and Reactive Thrombocytosis.” *American Journal of Clinical Pathology* 112(1): 93–100.

Mohandas, T., R. S. Sparkes, and L. J. Shapiro. 1981. “Reactivation of an Inactive Human X

Chromosome: Evidence for X Inactivation by DNA Methylation." *Science* 211(4480): 393–96.

Moreira de Mello, Joana Carvalho et al. 2010. "Random X Inactivation and Extensive Mosaicism in Human Placenta Revealed by Analysis of Allele-Specific Gene Expression along the X Chromosome." *PLoS ONE* 5(6).

Murakami, K. et al. 2009. "Identification of the Chromatin Regions Coated by Non-Coding Xist RNA." *Cytogenetic and Genome Research* 125(1): 19–25. https://pubmed.ncbi.nlm.nih.gov/19617692/ (December 17, 2020).

Al Nadaf, Shafagh et al. 2012. "A Cross-Species Comparison of Escape from X Inactivation in Eutheria: Implications for Evolution of X Chromosome Inactivation." *Chromosoma* 121(1): 71–78. http://link.springer.com/10.1007/s00412-011-0343-8 (December 18, 2020).

Naumova, Anna K. et al. 1998. "Genetic Mapping of X-Linked Loci Involved in Skewing of X Chromosome Inactivation in the Human." *European Journal of Human Genetics* 6(6): 552–62.

Navarro-Cobos, Maria Jose, Bradley P. Balaton, and Carolyn J. Brown. 2020. "Genes That Escape from X-Chromosome Inactivation: Potential Contributors to Klinefelter Syndrome." *American Journal of Medical Genetics, Part C: Seminars in Medical Genetics* 184(2): 226–38. /pmc/articles/PMC7384012/?report=abstract (December 16, 2020).

Neri, Giovanni, Charles E. Schwartz, Herbert A. Lubs, and Roger E. Stevenson. 2018. "X-Linked Intellectual Disability Update 2017." *American Journal of Medical Genetics, Part A* 176(6): 1375–88. /pmc/articles/PMC6049830/?report=abstract (December 17, 2020).

Nino-Soto, M. I. et al. 2005. "Differences in the Pattern of X-Linked Gene Expression between Fetal Bovine Muscle and Fibroblast Cultures Derived from the Same Muscle Biopsies."

Cytogenetic and Genome Research 111(1): 57–64.

Okamoto, Ikuhiro et al. 2004. "Epigenetic Dynamics of Imprinted X Inactivation during Early Mouse Development." *Science* 303(5658): 644–49.

———. 2011. "Eutherian Mammals Use Diverse Strategies to Initiate X-Chromosome Inactivation during Development." *Nature* 472(7343): 370–74.

Peeters, Samantha B., Andrea J. Korecki, Elizabeth M. Simpson, and Carolyn J. Brown. 2018. "Human Cis-Acting Elements Regulating Escape from X-Chromosome Inactivation Function in Mouse." *Human Molecular Genetics* 27(7): 1252–62. https://academic.oup.com/hmg/article-abstract/27/7/1252/4833562 (December 16, 2020).

Pinter, Stefan F. et al. 2012. "Spreading of X Chromosome Inactivation via a Hierarchy of Defined Polycomb Stations." *Genome Research* 22(10): 1864–76.

Posynick, Bronwyn J., and Carolyn J. Brown. 2019. "Escape From X-Chromosome Inactivation: An Evolutionary Perspective." *Frontiers in Cell and Developmental Biology* 7. /pmc/articles/PMC6817483/?report=abstract (December 16, 2020).

Przanowski, Piotr, Zeming Zheng, Urszula Wasko, and Sanchita Bhatnagar. 2019. "A Non-Random Mouse Model for Pharmacological Reactivation of Mecp2 on the Inactive X Chromosome." *Journal of Visualized Experiments* 2019(147). https://pubmed.ncbi.nlm.nih.gov/31180354/ (March 8, 2021).

Qu, Kun et al. 2015. "Individuality and Variation of Personal Regulomes in Primary Human T Cells." *Cell Systems* 1(1): 51–61.

Quang, Daniel, and Xiaohui Xie. 2016. "DanQ: A Hybrid Convolutional and Recurrent Deep Neural Network for Quantifying the Function of DNA Sequences." *Nucleic Acids Research* 44(11). https://pubmed.ncbi.nlm.nih.gov/27084946/ (December 16, 2020).

R core Team. 2014. “R: A Language and Environment for Statistical Computing.” http://www.r-project.org/.

Ramírez, Fidel et al. 2016. “DeepTools2: A next Generation Web Server for Deep-Sequencing Data Analysis.” *Nucleic acids research* 44(W1): W160–65. https://pubmed.ncbi.nlm.nih.gov/27079975/ (December 16, 2020).

Raznahan, Armin et al. 2018. “Sex-Chromosome Dosage Effects on Gene Expression in Humans.” *Proceedings of the National Academy of Sciences of the United States of America* 115(28): 7398–7403. /pmc/articles/PMC6048519/?report=abstract (December 17, 2020).

Richard Albert, Julien et al. 2018. “Development and Application of an Integrated Allele-Specific Pipeline for Methylomic and Epigenomic Analysis (MEA).” *BMC Genomics* 19(1).

Ross, Mark T. et al. 2005. “The DNA Sequence of the Human X Chromosome.” *Nature* 434(7031): 325–37.

Rozowsky, Joel et al. 2011. “AlleleSeq: Analysis of Allele-Specific Expression and Binding in a Network Framework.” *Molecular Systems Biology* 7.

Sadreyev, Ruslan I., Eda Yildirim, Stefan F. Pinter, and Jeannie T. Lee. 2013. “Bimodal Quantitative Relationships between Histone Modifications for X-Linked and Autosomal Loci.” *Proceedings of the National Academy of Sciences of the United States of America* 110(17): 6949–54. https://pubmed.ncbi.nlm.nih.gov/23564346/ (December 17, 2020).

Schultz, Matthew D. et al. 2015. “Human Body Epigenome Maps Reveal Noncanonical DNA Methylation Variation.” *Nature* 523(7559): 212–16.

Shao, Jiahao et al. 2020. “Genome-Wide DNA Methylation Changes of Perirenal Adipose Tissue in Rabbits Fed a High-Fat Diet.” *Animals* 10(12): 2213. https://www.mdpi.com/2076-2615/10/12/2213 (February 4, 2021).

Sharma, Virag, Anas Elghafari, and Michael Hiller. 2016. “Coding Exon-Structure Aware Realigner (CESAR) Utilizes Genome Alignments for Accurate Comparative Gene Annotation.” *Nucleic Acids Research* 44(11). https://pubmed.ncbi.nlm.nih.gov/27016733/ (December 16, 2020).

Sharp, Andrew J. et al. 2011. “DNA Methylation Profiles of Human Active and Inactive X Chromosomes.” *Genome Research* 21(10): 1592–1600. https://pubmed.ncbi.nlm.nih.gov/21862626/ (December 17, 2020).

Sharpe, D. 2015. “Your Chi-Square Test Is Statistically Significant: Now What?” *PARE* 20.

Shevchenko, Alexander I. et al. 2011. “Variability of Sequence Surrounding the Xist Gene in Rodents Suggests Taxon-Specific Regulation of X Chromosome Inactivation.” *PLoS ONE* 6(8). https://pubmed.ncbi.nlm.nih.gov/21826206/ (December 16, 2020).

Shevchenko, Alexander I., Elena V. Dementyeva, Irina S. Zakharova, and Suren M. Zakian. 2019. “Diverse Developmental Strategies of x Chromosome Dosage Compensation in Eutherian Mammals.” *International Journal of Developmental Biology* 63(3–5): 223–33.

Simon, Matthew D. et al. 2013. “High-Resolution Xist Binding Maps Reveal Two-Step Spreading during X-Chromosome Inactivation.” *Nature* 504(7480): 465–69. https://pubmed.ncbi.nlm.nih.gov/24162848/ (December 17, 2020).

Snijders Blok, Lot et al. 2015. “Mutations in DDX3X Are a Common Cause of Unexplained Intellectual Disability with Gender-Specific Effects on Wnt Signaling.” *American Journal of Human Genetics* 97(2): 343–52.

Souyris, Mélanie et al. 2018. “TLR7 Escapes X Chromosome Inactivation in Immune Cells.” *Science Immunology* 3(19). https://pubmed.ncbi.nlm.nih.gov/29374079/ (December 17, 2020).

Sudbrak, Ralf et al. 2001. "X Chromosome-Specific CDNA Arrays: Identification of Genes That Escape from X-Inactivation and Other Applications." *Human Molecular Genetics* 10(1): 77–83.

Syrett, Camille M., and Montserrat C. Anguera. 2019. "When the Balance Is Broken: X-Linked Gene Dosage from Two X Chromosomes and Female-Biased Autoimmunity." *Journal of Leukocyte Biology* 106(4): 919–32. /pmc/articles/PMC7206452/?report=abstract (December 17, 2020).

Tukiainen, Taru et al. 2017. "Landscape of X Chromosome Inactivation across Human Tissues." *Nature* 550(7675): 244–48.

Twigg, Stephen R.F. et al. 2013. "Cellular Interference in Craniofrontonasal Syndrome: Males Mosaic for Mutations in the x-Linked EFNB1 Gene Are More Severely Affected than True Hemizygotes." *Human Molecular Genetics* 22(8): 1654–62. https://pubmed.ncbi.nlm.nih.gov/23335590/ (December 16, 2020).

Vacca, Marcella, Floriana Della Ragione, Francesco Scalabrì, and Maurizio D'Esposito. 2016. "X Inactivation and Reactivation in X-Linked Diseases." *Seminars in Cell and Developmental Biology* 56: 78–87.

Veitia, Reiner A., Frédéric Veyrunes, Samuel Bottani, and James A. Birchler. 2015. "X Chromosome Inactivation and Active X Upregulation in Therian Mammals: Facts, Questions, and Hypotheses." *Journal of Molecular Cell Biology* 7(1): 2–11.

Venables, WN. Ripley, BD. 2002. *Modern Applied Statistics with S*. 4th ed. New York: Springer.

Wainer Katsir, Kerem, and Michal Linial. 2019. "Human Genes Escaping X-Inactivation Revealed by Single Cell Expression Data." *BMC Genomics* 20(1): 201. https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-5507-6 (December

18, 2020).

Wake, Norio, Nobuo Takagi, and Motomichi Sasaki. 1976. “Non-Random Inactivation of X Chromosome in the Rat Yolk Sac.” *Nature* 262(5569): 580–81.

Wang, Chen Yu et al. 2019. “Role of the Chromosome Architectural Factor SMCHD1 in X-Chromosome Inactivation, Gene Regulation, and Disease in Humans.” *Genetics* 213(2): 685–703. https://pubmed.ncbi.nlm.nih.gov/31420322/ (February 2, 2021).

Wang, Xu, Donald C. Miller, Andrew G. Clark, and Douglas F. Antczak. 2012. “Random X Inactivation in the Mule and Horse Placenta.” *Genome Research* 22(10): 1855–63.

Wang, Yanli et al. 2018. “The 3D Genome Browser: A Web-Based Browser for Visualizing 3D Genome Organization and Long-Range Chromatin Interactions.” *Genome Biology* 19(1): 151. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1519-9 (January 25, 2021).

Wang, Zhong, Huntington F. Willard, Sayan Mukherjee, and Terrence S. Furey. 2006. “Evidence of Influence of Genomic DNA Sequence on Human X Chromosome Inactivation.” *PLoS Computational Biology* 2(9): 0979–88.

Warburton, Peter E. et al. 2004. “Inverted Repeat Structure of the Human Genome: The X-Chromosome Contains a Preponderance of Large, Highly Homologous Inverted Repeated That Contain Testes Genes.” *Genome Research* 14(10 A): 1861–69.

Wilson Sayres, Melissa A., and Kateryna D. Makova. 2013. “Gene Survival and Death on the Human y Chromosome.” *Molecular Biology and Evolution* 30(4): 781–87.

Wu, Hao et al. 2014. “Cellular Resolution Maps of X Chromosome Inactivation: Implications for Neural Development, Function, and Disease.” *Neuron* 81(1): 103–19.

Xu, Wei, and Meiling Hao. 2018. “A Unified Partial Likelihood Approach for X-Chromosome

Association on Time-to-Event Outcomes." *Genetic Epidemiology* 42(1): 80–94. http://doi.wiley.com/10.1002/gepi.22097 (December 16, 2020).

Xue, Fei et al. 2002. "Aberrant Patterns of X Chromosome Inactivation in Bovine Clones." *Nature Genetics* 31(2): 216–20.

Yang, Fan, Tomas Babak, Jay Shendure, and Christine M. Disteche. 2010. "Global Survey of Escape from X Inactivation by RNA-Sequencing in Mouse." *Genome Research* 20(5): 614–22. https://pubmed.ncbi.nlm.nih.gov/20363980/ (December 17, 2020).

Younesy, Hamid et al. 2015. "VisRseq: R-Based Visual Framework for Analysis of Sequencing Data." *BMC Bioinformatics* 16(11): S2. /pmc/articles/PMC4559603/?report=abstract (February 2, 2021).

Yu, Bo, Helena T.A. van Tol, Tom A.E. Stout, and Bernard A.J. Roelen. 2020. "Initiation of X Chromosome Inactivation during Bovine Embryo Development." *Cells* 9(4).

Yue, Feng et al. 2014. "A Comparative Encyclopedia of DNA Elements in the Mouse Genome." *Nature* 515(7527): 355–64.

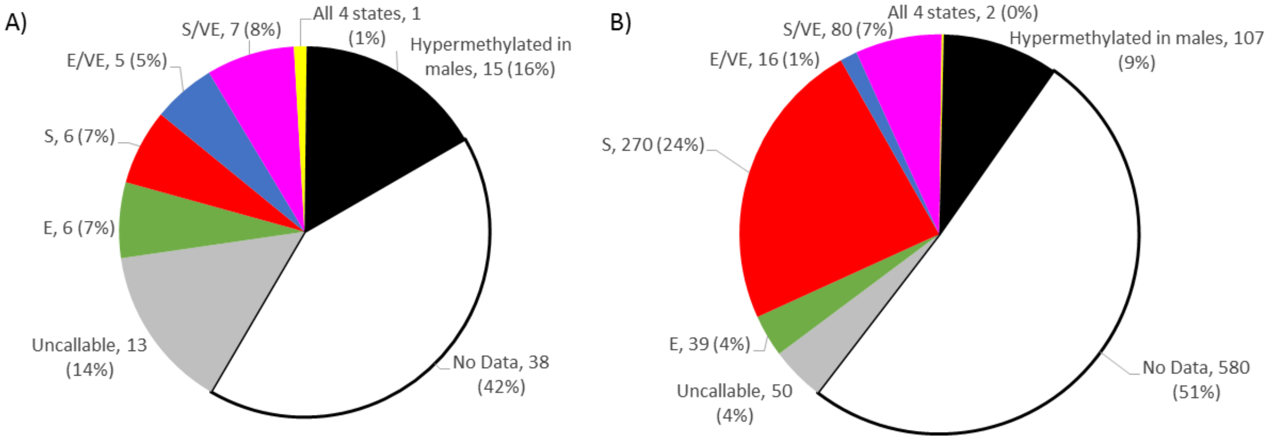
Zitzmann, Michael et al. 2015. "Gene Expression Patterns in Relation to the Clinical Phenotype in Klinefelter Syndrome." *Journal of Clinical Endocrinology and Metabolism* 100(3): E518–23. https://pubmed.ncbi.nlm.nih.gov/25532039/ (December 17, 2020).

Zou, Huiying et al. 2019. "No Imprinted XIST Expression in Pigs: Biallelic XIST Expression in Early Embryos and Random X Inactivation in Placentas." *Cellular and Molecular Life Sciences* 76(22): 4525–38. https://doi.org/10.1007/s00018-019-03123-3 (December 16, 2020).

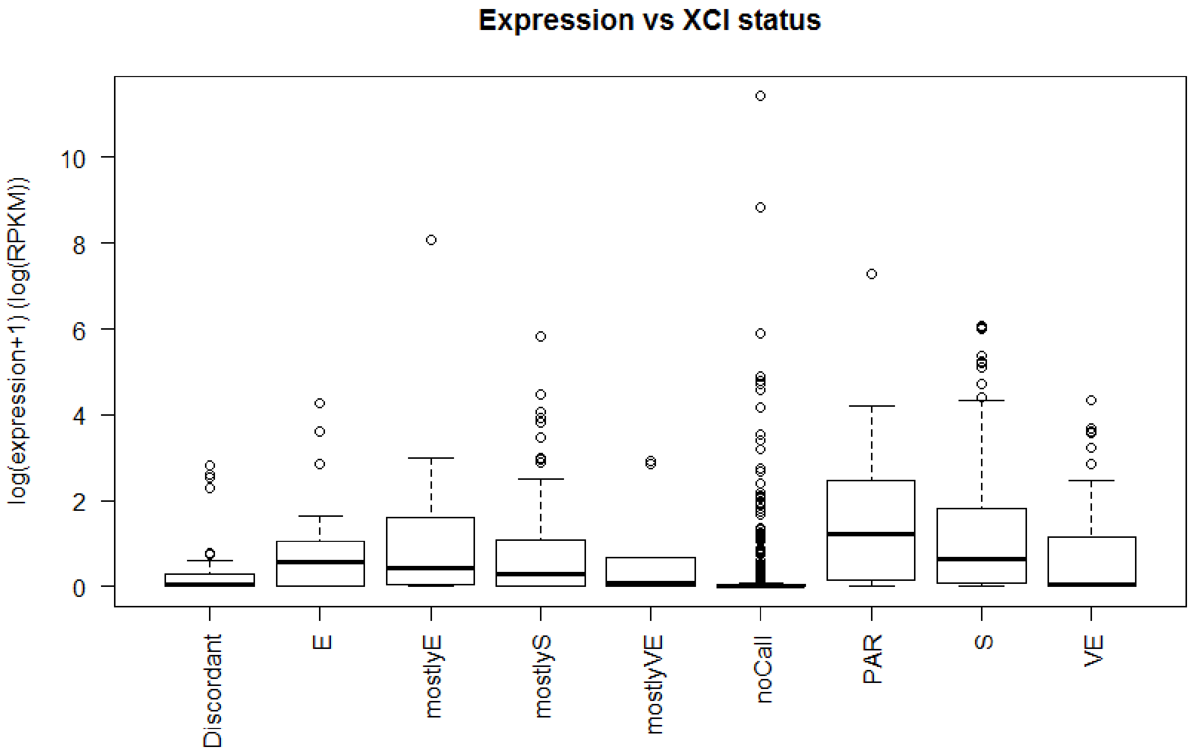
Appendices

Appendix A Supplementary materials for Chapter 2

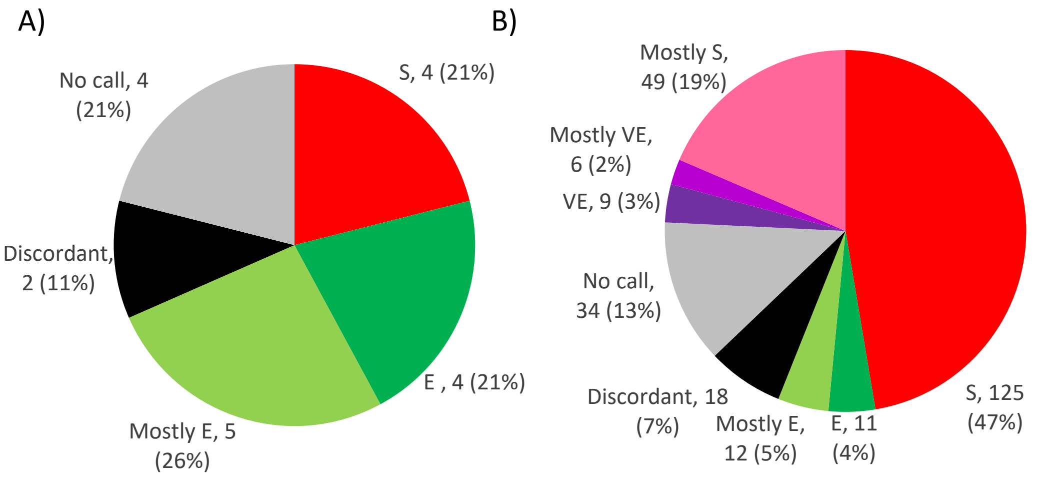
A.1 Supplementary Figures



Supplemental Figure S2.1 Comparing the Cotton DNAme XCI status calls and consensus calls. No data reflects genes which were not called in the DNAme study, primarily due to a lack of CpG islands. Uncallable are genes which had methylation between the subject and escape classifiers and were unable to be confidently called by the DNAme study. S, E and VE are subject, escape and variable escape from XCI. E/VE and S/VE are genes which were fully subject or escape in some tissues while variably escaping in other tissues. All 4 states were genes which had some tissues subject, escaping, variably escaping and uncallable making the gene not fit into any other XCI status category. (A) The Cotton DNAme XCI status calls when the consensus call is variable escape or discordant. N=91. (B) The Cotton DNAme XCI status calls for all genes on the X chromosome for comparison. N=1144.



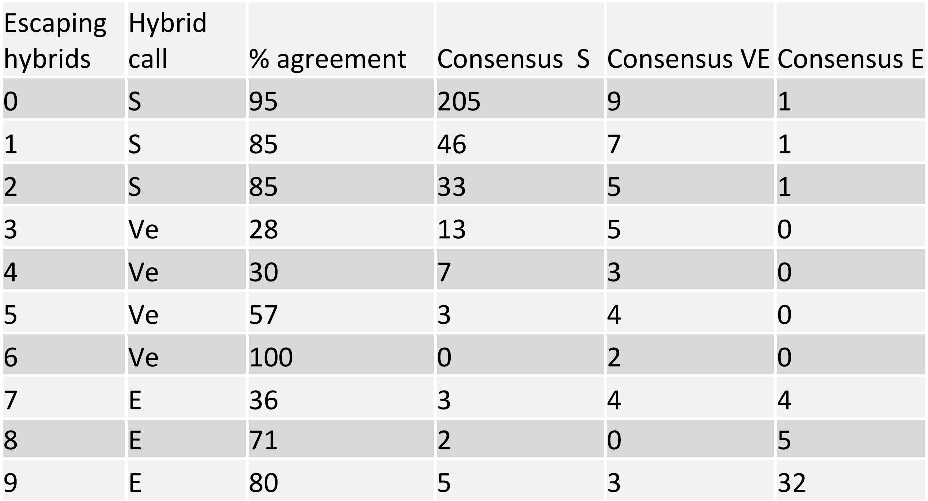
Supplemental Figure S2.2 Expression in GM12878 does not correlate with consensus XCI status call. A box and whisker plot of the log reads per kilobase of transcript per million mapped reads (RPKM) of expression. A value of 1 RPKM was added to each gene in order to include genes with 0 expression in a graph of log10(RPKM). E, VE, S and PAR are escape, variable escape, and subject to XCI and pseudo-autosomal region. The N are: Discordant=44, E=29, mostly E=26, mostly S=129, mostly VE=10, no call=509, PAR=22, S=331, VE=37.



Supplemental Figure S2.3 Consensus XCI status calls of genes with Y homologs or Y pseudogenes. A) XCI status calls of X genes with homologues on the Y chromosome. E is genes which escape from XCI in all studies, mostly E is genes which escape from XCI in the majority of studies, S is genes which are subject to XCI in all studies, discordant is genes which either have an even split of S and E calls or have one of each call (including variable escape), and no call is genes with no XCI status call in any study. N=19. B) XCI status calls of X genes with pseudogenes on the Y chromosome. See above for description of most categories. VE and mostly VE is variable escape from XCI in all studies and variable escape from XCI in the majority of studies. Mostly S is subject to XCI in the majority of studies. N=264.

A.2 Supplementary Tables

Supplemental table S2.1: Our consensus XCI status calls for all genes on the X chromosome. The consensus calls from this study are under the column labeled Balaton consensus calls. The data used for the rest of the analyses in this chapter are also included as columns. The second sheet has descriptions of each column. As the table is too large, please see the version of this chapter published in Biology of Sex Differences (Balaton, Cotton, and Brown 2015).

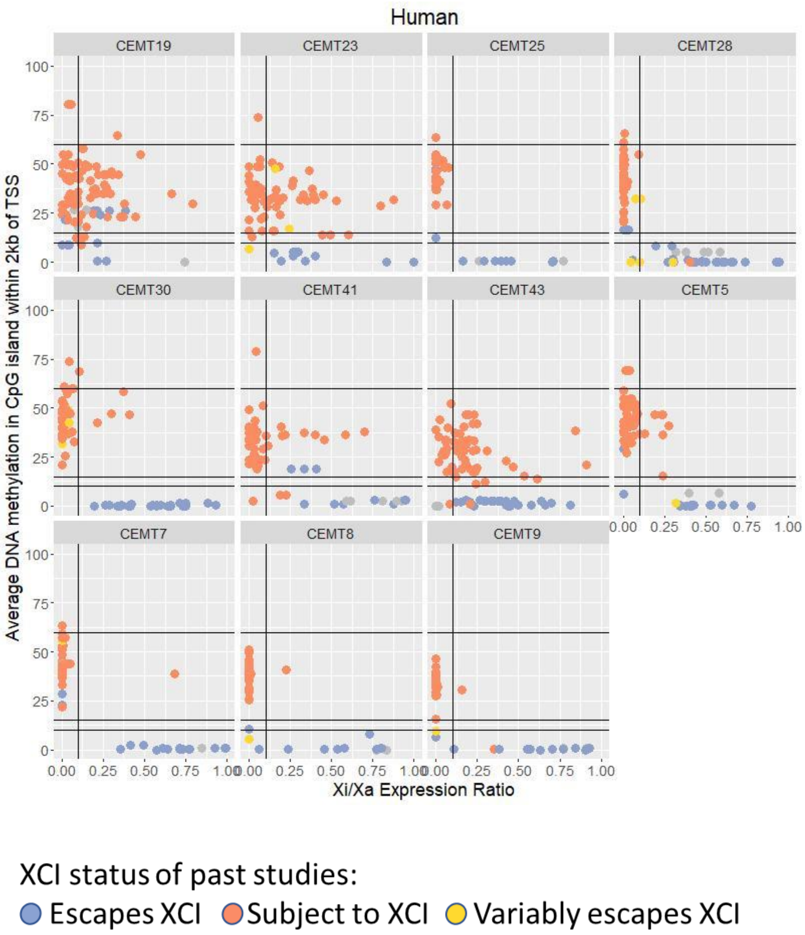


Escaping hybrids	Hybrid call	% agreement	Consensus S	Consensus VE	Consensus E
0	S	95	205	9	1
1	S	85	46	7	1
2	S	85	33	5	1
3	Ve	28	13	5	0
4	Ve	30	7	3	0
5	Ve	57	3	4	0
6	Ve	100	0	2	0
7	E	36	3	4	4
8	E	71	2	0	5
9	E	80	5	3	32

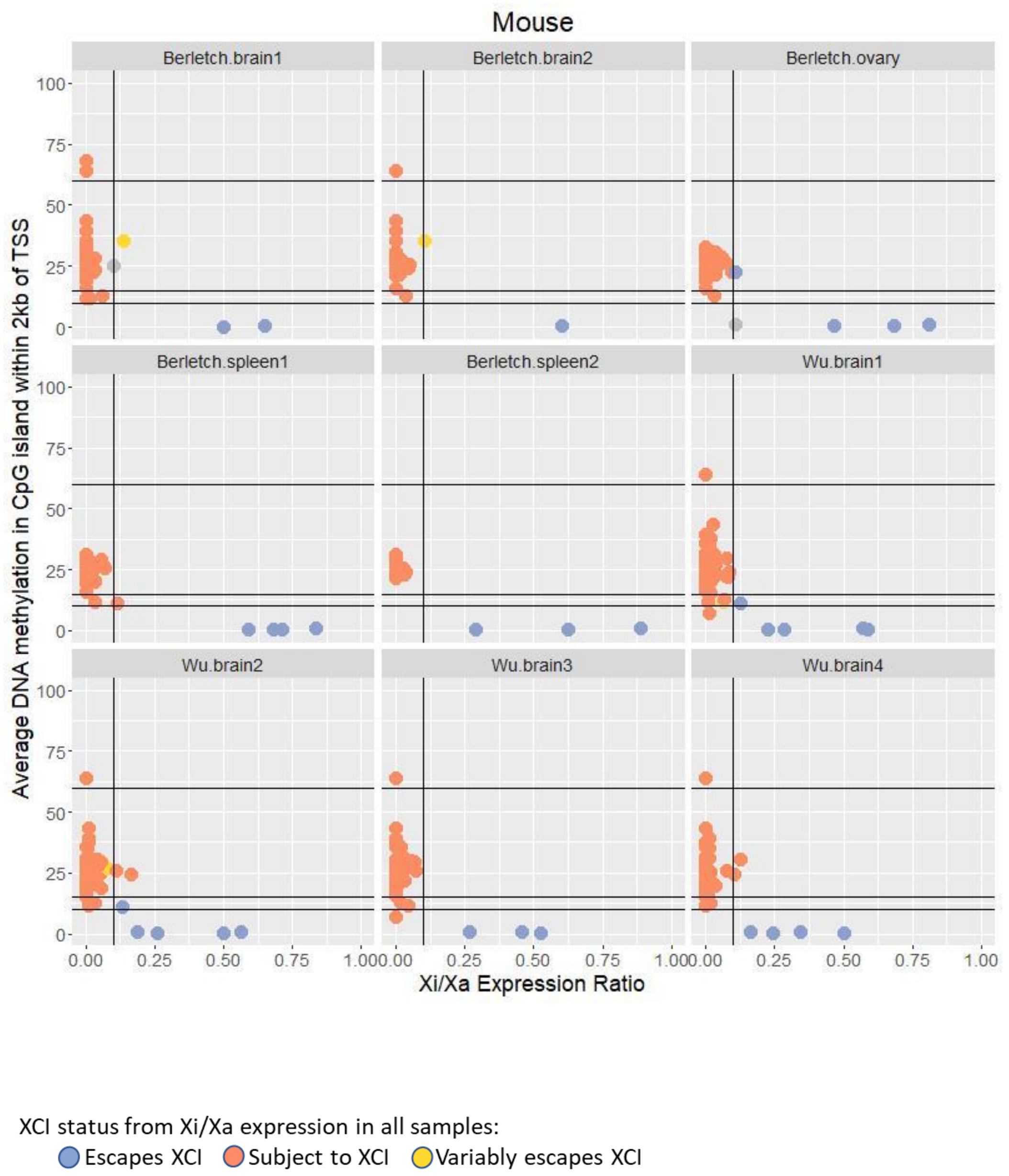
Supplemental table S2.2: The hybrid study tends to call genes variable escape discordantly. The data used to create Figure 2.4. Escaping hybrids is how many human-mouse hybrid cell lines (out of 9) were found to escape from XCI by Carrel, Hybrid call is the XCI status call from the Carrel hybrid study, % agreement is the percent of genes with that number of escaping hybrids whose Carrel hybrid call agrees with one or more other study's call. Consensus S, VE and E are how many genes have other studies agree on a call of subject, variable escape or escape.

Appendix B Supplementary materials for Chapter 3

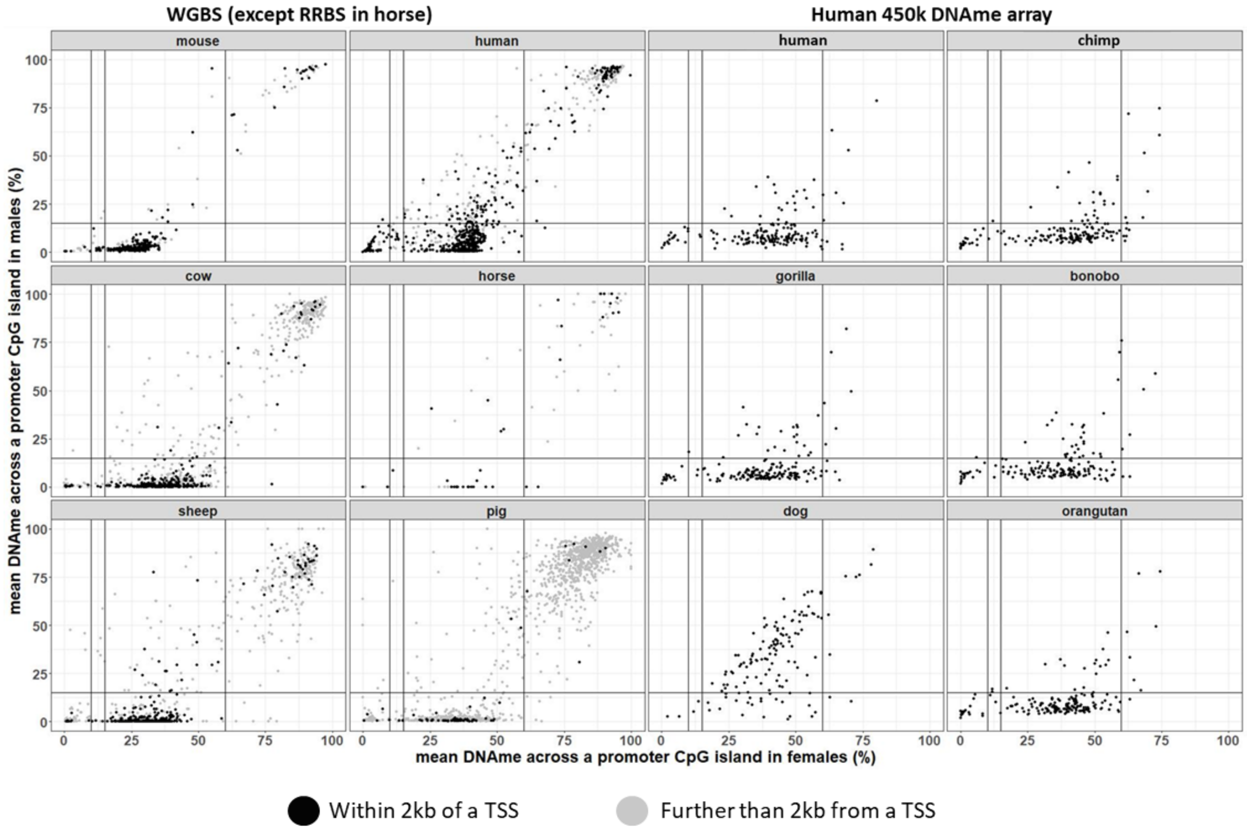
## B.1	Supplementary Figures



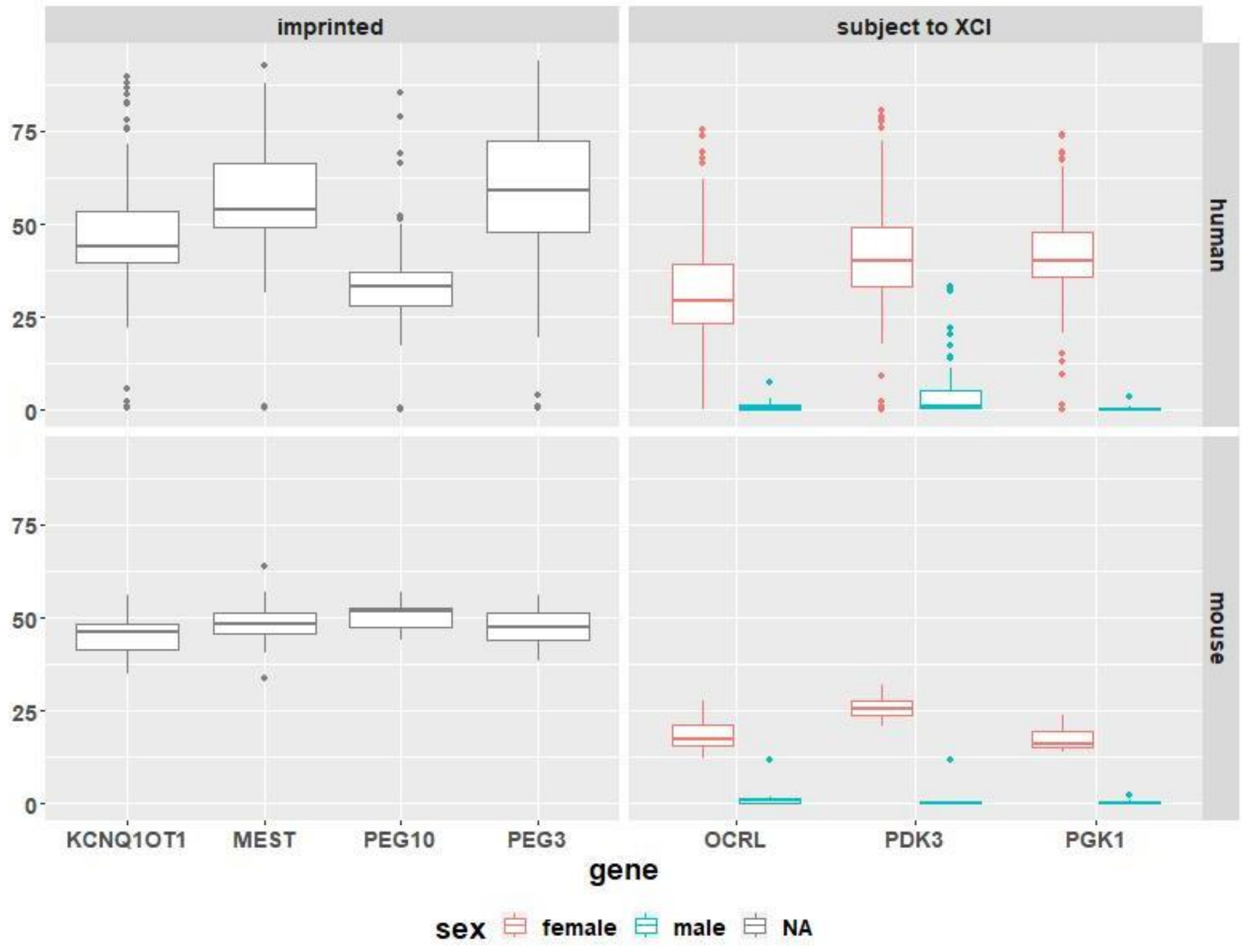
Supplementary Figure 3.1: The Xi/Xa expression ratio vs promoter DNAme level in individual human samples. Each point is a SNP with Xi/Xa expression data, matched to the most likely promoter and any CpG islands within 2kb in order to have matched DNAme values. Lines are drawn at 0.1 Xi/Xa expression and at 10, 15 and 60% DNAme as they were used as thresholds to call XCI escape status later. Points are colored based on their XCI status calls in the previous literature (Balaton, Cotton, and Brown 2015). CEMT30, a leukemia cancer sample, was used for Figure 1. Three samples (CEMT19, CEMT23 and CEMT43) were discarded from downstream analyses, because they did not appear to show skewing of Xi choice, with many genes called as subject to XCI by DNAme and previous studies, with an XiXa expression ratio >>0.1.



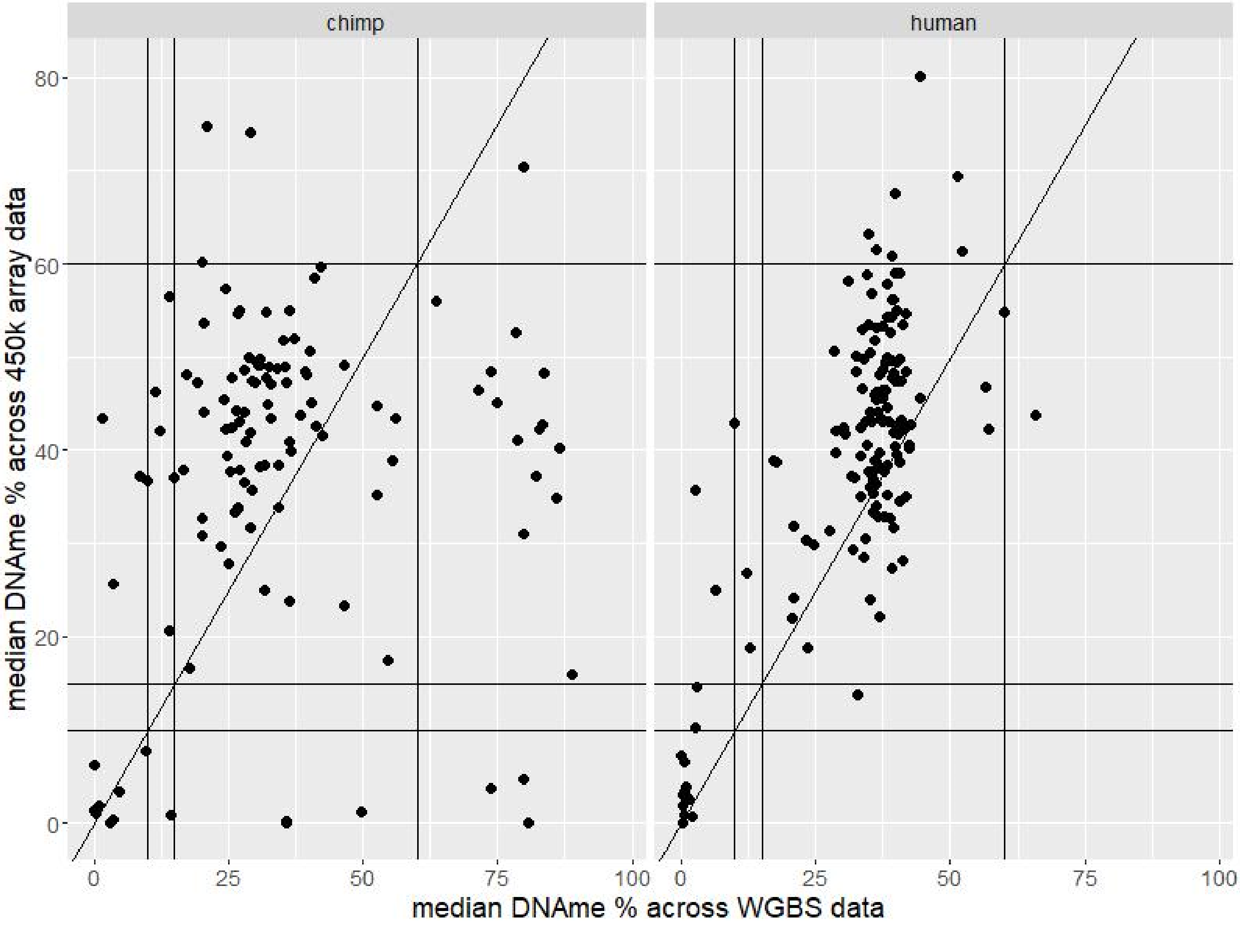
Supplementary Figure 3.2: The Xi/Xa expression ratio vs promoter DNAme level in individual mouse samples. Each point is a SNP with Xi/Xa expression data, matched to the most likely promoter and any CpG islands within 2kb in order to have matched DNAme values. Lines are drawn at 0.1 Xi/Xa expression and at 10, 15 and 60% DNAme as they were used as thresholds to call XCI escape status later. Points are colored based on their XCI status calls made using Xi/Xa expression. Data from 2 different studies are used: one used an Xist knockout to skew Xi choice and the other used differently colored fluorescent proteins expressed from each X chromosome to sort cells based on Xi choice. Data from (Keown et al. 2017) not shown here was used for Figure 3.1.



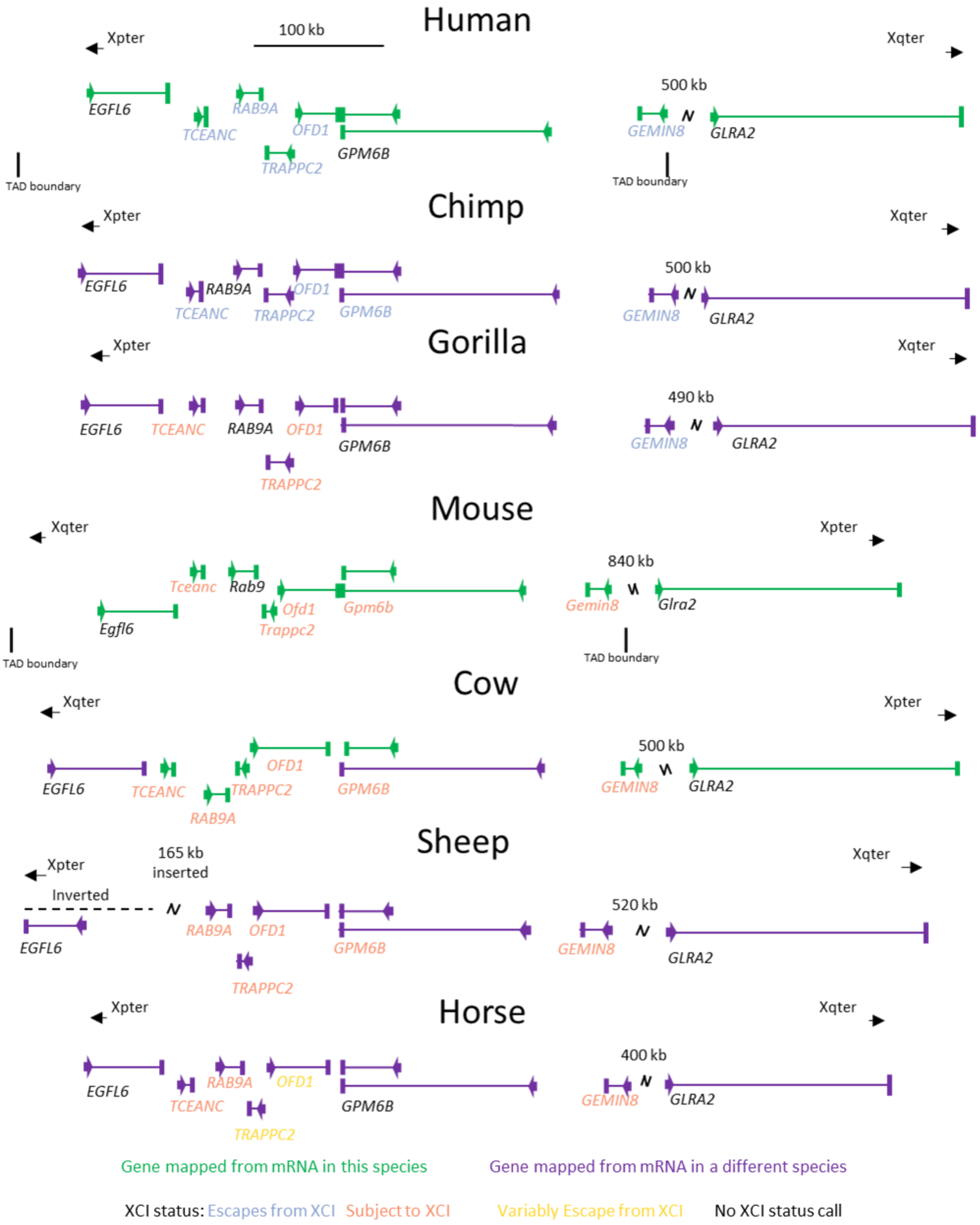
Supplementary Figure 3.3: Male vs female DNAme across species. The DNAme data shown was generated with 3 different methods: WGBS, RRBS and the human 450k DNAme array. Each point is a CpG island. Lines are drawn at female DNAme of 10, 15 and 60 as those thresholds were used to call a gene's XCI status and at male DNAme of 15 as genes with higher than 15% male DNAme were discarded from further analysis. Chimp (WGBS) and goat are not shown due to lack of male samples to compare to. CpG islands are colored based on the distance to their closest TSS.



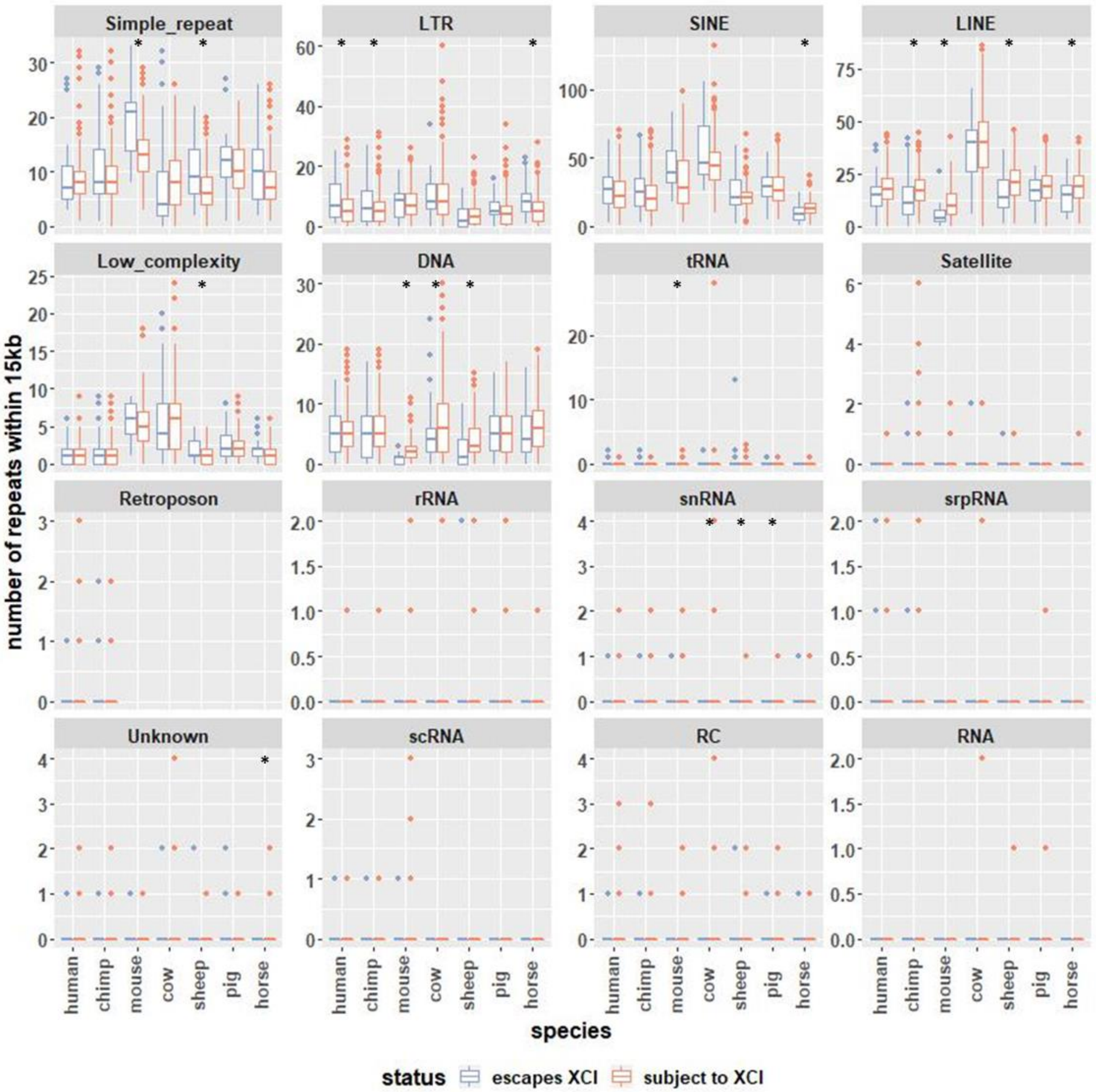
Supplementary Figure 3.4: A comparison of imprinted genes and genes subject to XCI. The average DNAme level at promoter CpG islands are shown for 4 imprinted genes and 4 genes subject to XCI in humans (A) and mouse (B). Genes subject to XCI have males and females separate as females are expected to be hemi-methylated while males are expected to have low methylation.



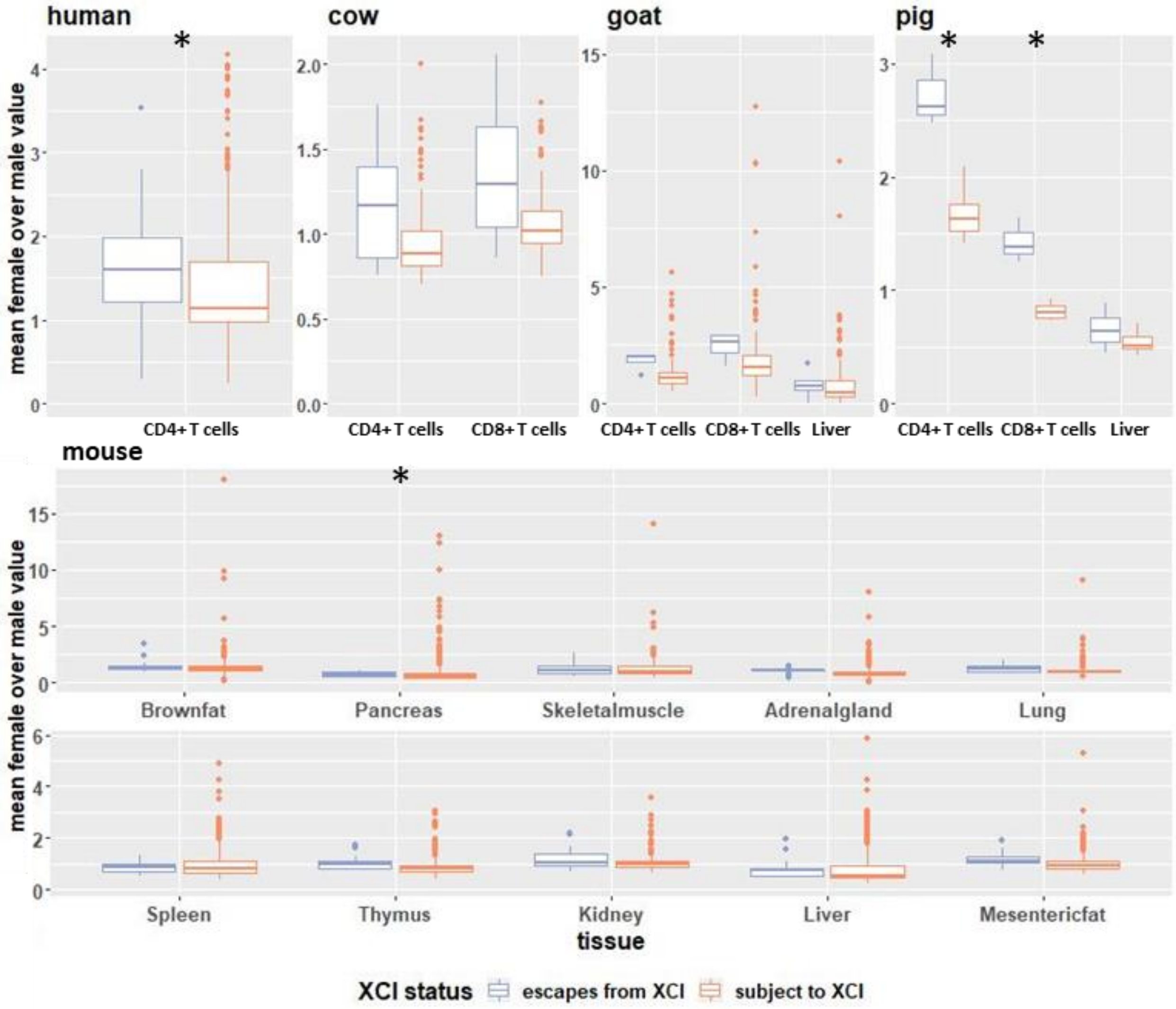
Supplementary Figure 3.5: Comparison of DNAme data generated using WGBS and the 450k array. Human and chimp were the only two species that had data generated using both methods. Lines are drawn at 10,15 and 60% DNAme to show the thresholds used for calling XCI status. Another line was drawn along the diagonal to show where perfect concordance between datasets would be.


Human
Xpter
100 kb
Xqter
EGFL6
TCEANC
RAB9A
OFD1
GPM6B
TRAPPC2
500 kb
GEMIN8
GLRA2
TAD boundary
TAD boundary
Chimp
Xpter
Xqter
EGFL6
RAB9A
OFD1
TCEANC
TRAPPC2
GPM6B
500 kb
GEMIN8
GLRA2
Gorilla
Xpter
Xqter
EGFL6
TCEANC
RAB9A
OFD1
GPM6B
TRAPPC2
490 kb
GEMIN8
GLRA2
Mouse
Xqter
Xpter
Tceanc
Rab9
Ofd1
Trappc2
Gpm6b
Egfl6
840 kb
Gemin8
Glra2
TAD boundary
TAD boundary
Cow
Xqter
Xpter
EGFL6
TCEANC
OFD1
TRAPPC2
RAB9A
GPM6B
500 kb
GEMIN8
GLRA2
Sheep
Xpter
165 kb
inserted
Inverted
Xqter
EGFL6
RAB9A
OFD1
TRAPPC2
GPM6B
520 kb
GEMIN8
GLRA2
Horse
Xpter
Xqter
EGFL6
RAB9A
OFD1
TCEANC
TRAPPC2
GPM6B
400 kb
GEMIN8
GLRA2
Gene mapped from mRNA in this species
Gene mapped from mRNA in a different species
XCI status: Escapes from XCI
Subject to XCI
Variably Escape from XCI
No XCI status call

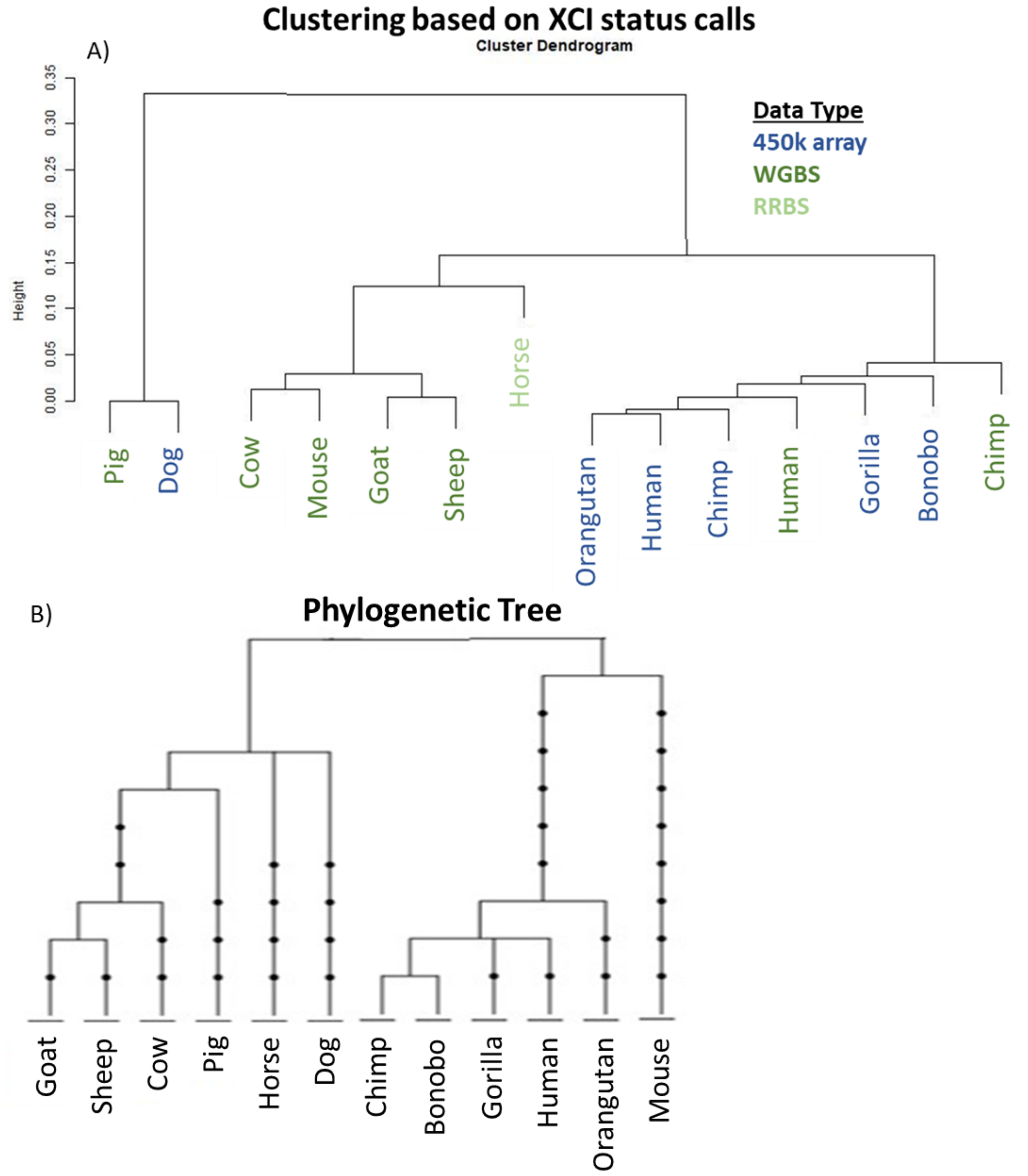
Supplementary Figure 3.6: Cross-species comparison of a primate-specific escape domain. The domain spanning from *TCEANC* to *GEMIN8* and the neighboring gene on each side are shown. Genes names are colored by their XCI status in each species and the gene diagram is colored by whether the gene annotation is from mRNA in that species or from other species. All regions in all species were scaled together, with species aligned at the end of GPM6B. As there is a large gene-free region between *GEMIN8* and *GLRA2* this region has been condensed and the distance between the two genes noted. Dotted lines show the region that is inverted in sheep. Xpter and Xqter show the direction to the short and long arms of the chromosome respectively, note that this region and much of the X chromosome is inverted in mouse and cow (Bujold et al. 2016; Duncan et al. 2018; Vacca et al. 2016). Cow had inconsistencies between bosTau6 (used in our data source and this study) and bosTau9 (the latest cow genome build), with bosTau6 being used here. bosTau9 had duplication or rearrangement of *EGFL6* and *TCEANC*. Gorilla and horse had small pseudo-gene insertions in the region, but these were only around 2kb in size and so were left out.



Supplementary Figure 3.7: Number of repeats within 15kb per TSS. Species with a * have significant differences between genes found escaping XCI and those found subject to XCI at adjusted p-value<0.01.



Supplementary Figure 3.8: Mean female/male ATAC-seq signal across samples within 250bp of TSSs, separated by tissue. Tissues with a * have significant differences between genes found escaping XCI and those found subject to XCI at adjusted p-value<0.01.



Supplementary Figure 3.9: Clustering of species by XCI status calls. Species were clustered by their XCI status calls (A) and compared to a phylogenetic tree showing their evolutionary relations (B). For the clustering, species names are colored by the type of data used to generate the XCI status calls.

B.2 Supplementary Tables

DNAme

Species	n female samples	n male samples	method	tissue	source
human	161	115	WGBS	various	IHEC
human	6	3	450k	peripheral blood	Hernando-Herraez, I., et al.
chimp	1	0	WGBS	bone	Gokhman, D., et al.
chimp	3	2	450k	peripheral blood	Hernando-Herraez, I., et al.
gorilla	4	2	450k	peripheral blood	Hernando-Herraez, I., et al.
bonobo	3	3	450k	peripheral blood	Hernando-Herraez, I., et al.
orangutan	5	1	450k	peripheral blood	Hernando-Herraez, I., et al.
mouse	12	12	WGBS	liver	Grimm, SA., et al.
cow	4	0	WGBS	whole blood and mammary	Zhou, Y., et al.
cow	0	3	WGBS	muscle	Fang, X., and Zhao, Z.
sheep	6	0	WGBS	adipose	Statham, A., and Tellam, R.
sheep	0	2	WGBS	muscle	no publication?
goat	6	0	WGBS	skin	Li, C., et al.
pig	4	0	WGBS	corpus luteum	Zhao, F., et al
pig	0	3	WGBS	liver	Li, Y., et al.
horse	11	1	RRBS	leukocytes	Zabek, T., et al.
dog	6	1	450k	leukocytes	Epiphanio, TMF., et al.

ATAC-seq

Species	**n female samples**	**n male samples**	**method**	**tissue**	**source**
human	**8**	**35**	**ATAC-seq**	**T cells**	**Qu, K., *et al.***
cow	**4**	**4**	**ATAC-seq**	**CD4+ and CD8+ T cells**	**Foissac, S., *et al.***
pig	**6**	**6**	**ATAC-seq**	**CD4+ and CD8+ T cells, liver**	**Foissac, S., *et al.***
mouse	**20**	**20**	**ATAC-seq**	**brown fat, pancreas, skeletal muscle, spleen, thymus, kidney, liver, adrenal, lung, mesenteric fat**	**Liu, C., et al.**

CTCF

Species	**n samples**	**method**	**source**
human	318	ChIP-seq	ENCODE
mouse	37	ChIP-seq	ENCODE

Genome Build

Species	genome build
human	hg38
chimpanzee	PanTro6
gorilla	gorGor5
bonobo	panpan3
orangutan	ponAbe3
mouse	mm9
cow	bosTau6
sheep	oviAri3
goat	CHIR1.0
pig	susScr11
horse	equiCab3

Supplemental Table S3.1: **The sources of data used in this study.**

Supplemental Table S3.2: All XCI status calls made in this study compared to humans. As the table is too large, please see the version of this chapter published in Epigenetics and Chromatin (Balaton et al. 2021).

Supplemental Table S3.3: **Individual XCI status calls per dataset**. Each sheet is a separate dataset analyzed. As the table is too large, please see the version of this chapter published in Epigenetics and Chromatin (Balaton et al. 2021).

Repeats

repeat class	**human**	**chimp**	**mouse**	**cow**	**sheep**	**pig**	**horse**	**n significant species**
Simple_repeat	8.9E-01	3.3E-02	**8.3E-03**	7.2E-01	**1.9E-04**	3.1E-01	2.3E-02	2
LTR	**7.6E-03**	**7.1E-03**	9.6E-01	4.4E-01	1.3E-01	3.3E-02	**2.1E-03**	3
SINE	1.8E-01	6.1E-02	7.4E-02	3.3E-02	1.3E-01	4.9E-01	**7.1E-03**	1
LINE	1.2E-02	**6.1E-04**	**8.0E-03**	1.8E-01	**1.9E-04**	7.0E-02	**5.5E-03**	4
Low_complexity	9.6E-01	1.1E-01	3.7E-01	7.3E-01	**2.4E-03**	3.0E-01	1.7E-01	1
DNA	7.3E-01	6.3E-01	**8.9E-05**	**7.1E-03**	**1.9E-04**	4.9E-01	5.6E-01	3
tRNA	2.1E-01	7.2E-01	**1.5E-10**	4.6E-01	5.5E-01	9.7E-01	1.8E-01	1
Satellite	3.0E-01	7.8E-01	6.4E-02	3.7E-01	4.4E-01	NaN	4.4E-01	0
Retroposon	2.9E-01	4.9E-01	NaN	NA	NA	NA	NA	0
rRNA	5.0E-02	3.3E-02	1.0E-01	1.8E-01	4.1E-01	2.1E-01	4.4E-01	0
snRNA	4.4E-01	7.2E-01	7.2E-01	**2.0E-03**	**1.9E-04**	**3.9E-03**	4.1E-01	3
srpRNA	4.1E-01	4.2E-01	NaN	4.4E-01	NaN	1.8E-01	NaN	0
Unknown	4.2E-01	1.8E-01	3.0E-01	1.2E-01	7.2E-01	6.3E-01	**6.1E-04**	1
scRNA	9.5E-01	9.6E-01	3.4E-01	NaN	NaN	NaN	NaN	0
RC	9.5E-01	9.0E-01	2.1E-01	4.5E-02	8.0E-01	9.6E-01	9.0E-01	0
RNA	NaN	NaN	NaN	4.4E-01	3.0E-01	4.4E-01	NaN	0
n Escape islands	106	101	18	57	61	54	49	
n Subject islands	832	512	339	498	437	452	275	

CTCF

species	**pValue**	**adjusted p-value**		**n Escape TSS**	**n Subject TSS**
human	0.7335188	0.733519		71	551
chimp	0.0002829	**0.001556**		71	324
bonobo	0.0015868	**0.005818**		66	458
gorilla	0.0021464	**0.005903**		54	530
orangutan	0.2318512	0.387336		57	494
mouse	0.6387296	0.702603		14	395
cow	0.1724907	0.37948		12	250
sheep	0.2464863	0.387336		98	2130
goat	0.3671318	0.504806		4	131
pig	0.607208	0.702603		6	62
horse	5.543E-05	**0.00061**		230	1967

ATAC-seq

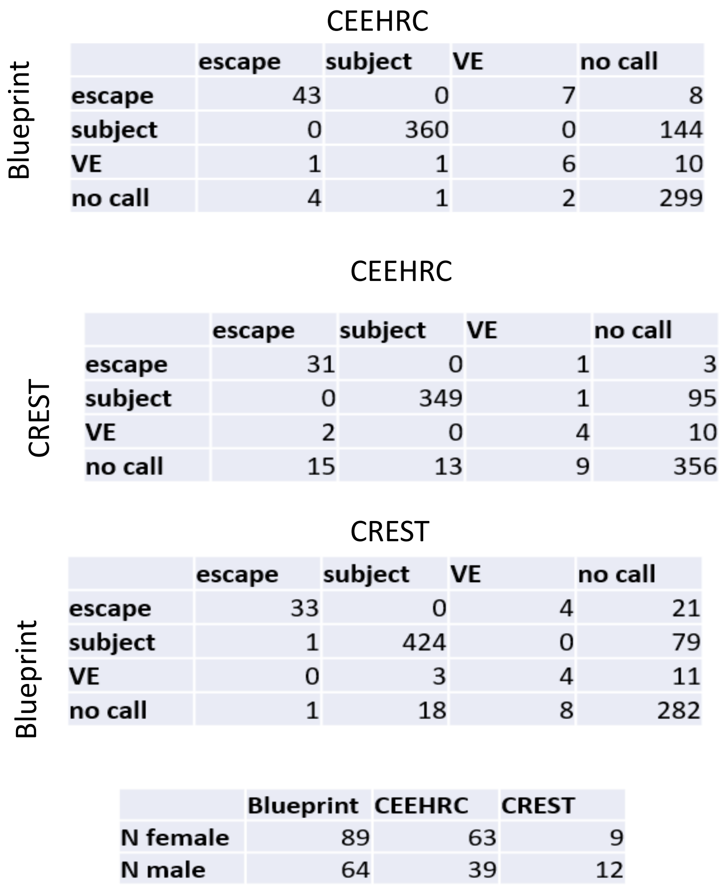
	p-value		**N escape TSS**	**N subject TSS**
Human	**0.000003411**		149	970
Mouse	**0.006876**		25	432
Cow	0.02972		12	174
Goat	0.1069		4	131
Pig	**0.00781**		4	42

Supplemental Table S3.4: Enrichment of repeats, CTCF and ATAC-seq at genes escaping vs subject to XCI. Repeats, CTCF and ATAC-seq are all separate tables. For repeats we tested the number of repeats within 15kb of each CpG island. For CTCF we tested the number of 200bp bins with predicted CTCF binding within 4kb of each TSS. For ATAC-seq we tested the female/male signal within 250bp of each TSS. We also included the number of CpG islands and TSSs per species that were informative for each analysis. Those in bold were found significant (t-test, adjusted p-value<0.01).

		primates					non-primates					
StartGene	StopGene	human	chimp	bonobo	gorilla	orangutan	mouse	cow	sheep	goat	pig	horse
edge	EGFL6	1	2	1	2	2	0	1	0	4	2	2
EGFL6	EGFL6	20	18	19	22	23	5	69	14	2	18	30
EGFL6	TCEANC	12	9	9	6	7	4	370	70	5	23	10
TCEANC	TCEANC	3	5	2	5	4	0	8	NA	2	0	5
TCEANC	RAB9A	2	2	4	3	2	0	0	NA	6	6	4
RAB9A	RAB9A	5	9	7	10	7	0	10	6	0	12	6
RAB9A	TRAPPC2	0	0	0	0	0	0	0	0	0	1	0
TRAPPC2	TRAPPC2	10	10	10	8	10	0	3	3	2	4	15
TRAPPC2	OFD1	0	0	1	1	1	0	0	1	2	2	0
OFD1	OFD1	17	15	18	15	15	18	16	23	22	23	15
OFD1	GPM6B	0	0	0	0	0	0	2	0	0	3	0
GPM6B	GPM6B	13	15	14	11	15	7	18	3	14	25	18
GPM6B	GEMIN8	40	49	45	37	40	51	52	77	57	66	47
GEMIN8	GEMIN8	8	7	5	11	9	7	3	3	0	11	4
GEMIN8	GLRA2	109	108	103	98	100	213	115	122	121	115	102
GLRA2	GLRA2	43	48	47	44	48	67	50	38	36	34	31
GLRA2	edge	3	3	3	3	2	2	0	0	0	0	0
	Mean	16.8	17.6	16.9	16.2	16.8	22.0	42.2	24	16.1	20.3	17.0

Escape XCI | Subject to XCI | variably escapes XCI

Supplemental Table S3.5: The number of predicted CTCF binding sites between genes in a discordant region. A DanQ model was given overlapping 200bp bins of each genome and predicted the likelihood of it containing a CTCF binding site. The number of bins with over 80% chance of having CTCF binding were counted per region. Each region goes from either the start of a gene to its end, or from the end of one gene to the start of the next. Edges were included 5kb from the furthest gene on each side. This discordant region is the one featured in Figure S3.6.



CEEHRC

Blueprint

	escape	subject	VE	no call
escape	43	0	7	8
subject	0	360	0	144
VE	1	1	6	10
no call	4	1	2	299

CEEHRC

CREST

	escape	subject	VE	no call
escape	31	0	1	3
subject	0	349	1	95
VE	2	0	4	10
no call	15	13	9	356

CREST

Blueprint

	escape	subject	VE	no call
escape	33	0	4	21
subject	1	424	0	79
VE	0	3	4	11
no call	1	18	8	282

	Blueprint	CEEHRC	CREST
N female	89	63	9
N male	64	39	12

Table S6: DNAme based XCI status calls compared across IHEC consortia. CREST, Blueprint and CEEHRC were the consortia with the most DNAme data sets when this data was downloaded. The majority of CEEHRC samples are cancer while the CREST samples and the majority of Blueprint samples are not. The 4th table shows the number of male and female samples per consortium. VE is variably escapes from XCI

Appendix C Supplementary materials for Chapter 4

C.1 Supplementary Figures

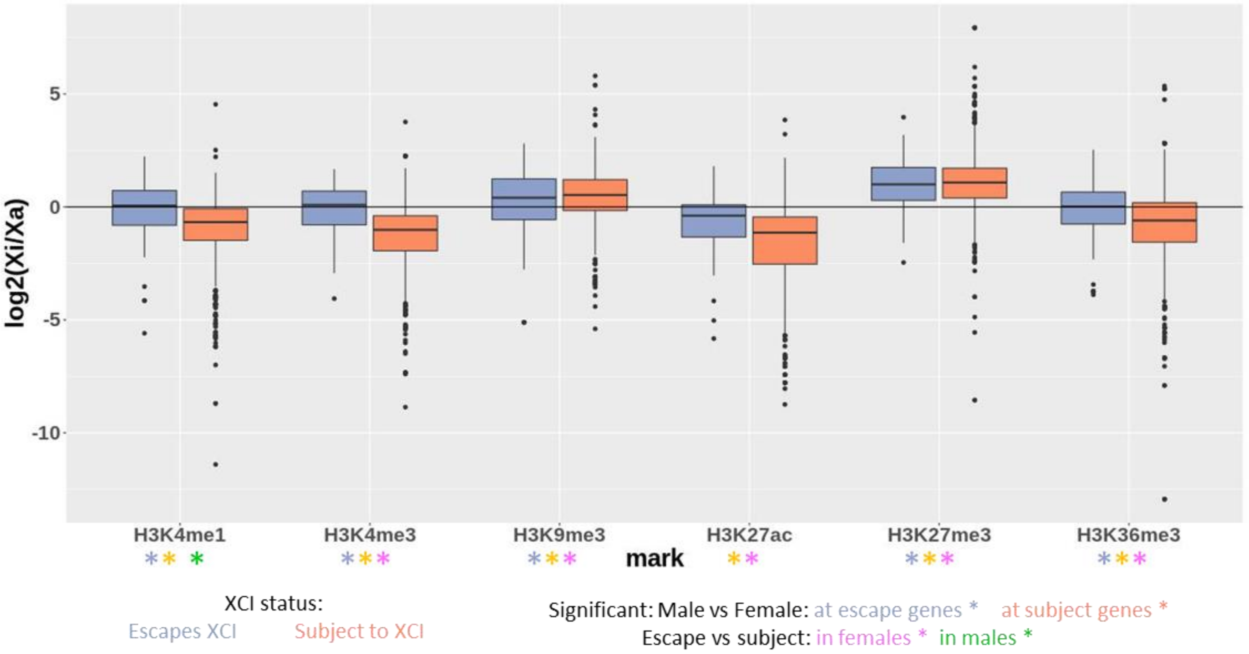
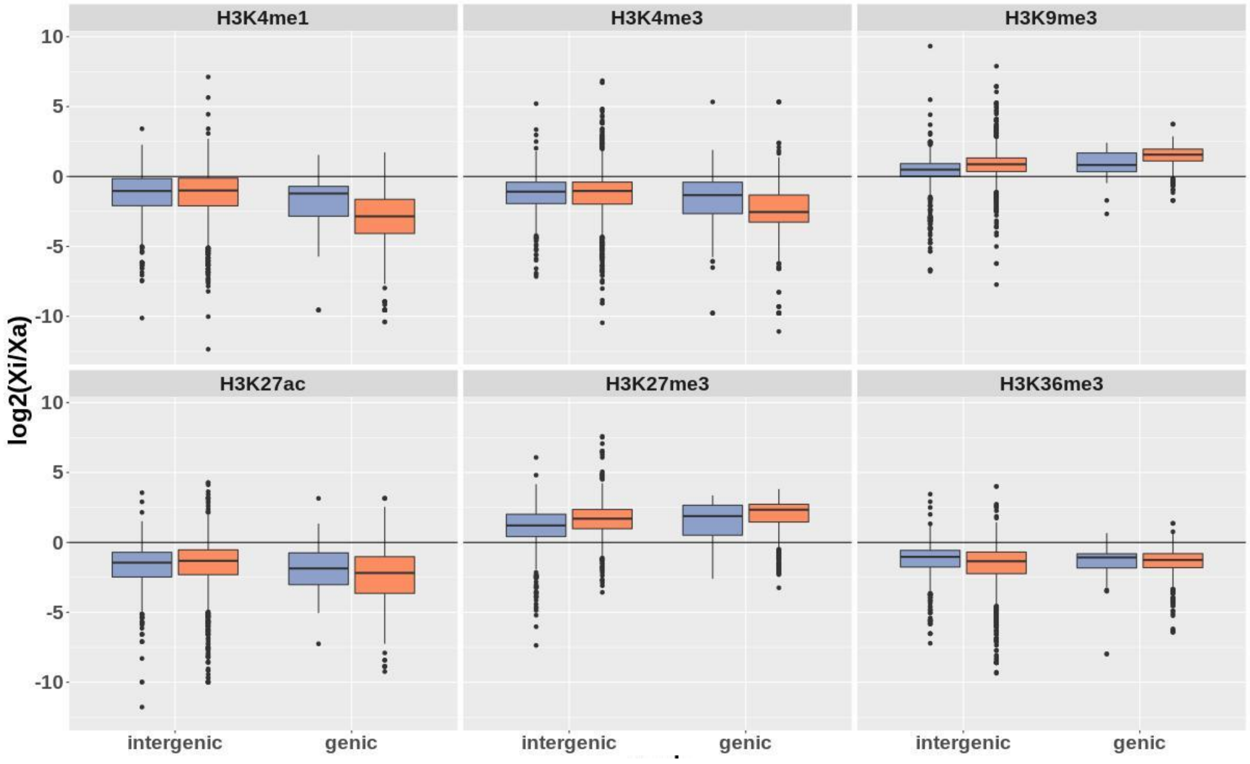


Figure S4.1: log2(Xi/Xa) for epigenetic marks in CREST at promoters. Data from CREST is shown.

Significance for the various t-tests featured in Table S2 are shown by the differently colored star.



XCI Status: Escapes XCI Subject to XCI

Figure S4.2: log2(Xi/Xa) for epigenetic marks in CEMT at enhancers, split by genic vs intergenic location.

Enhancers are split by whether they are located within a gene (genic) or not (intergenic).

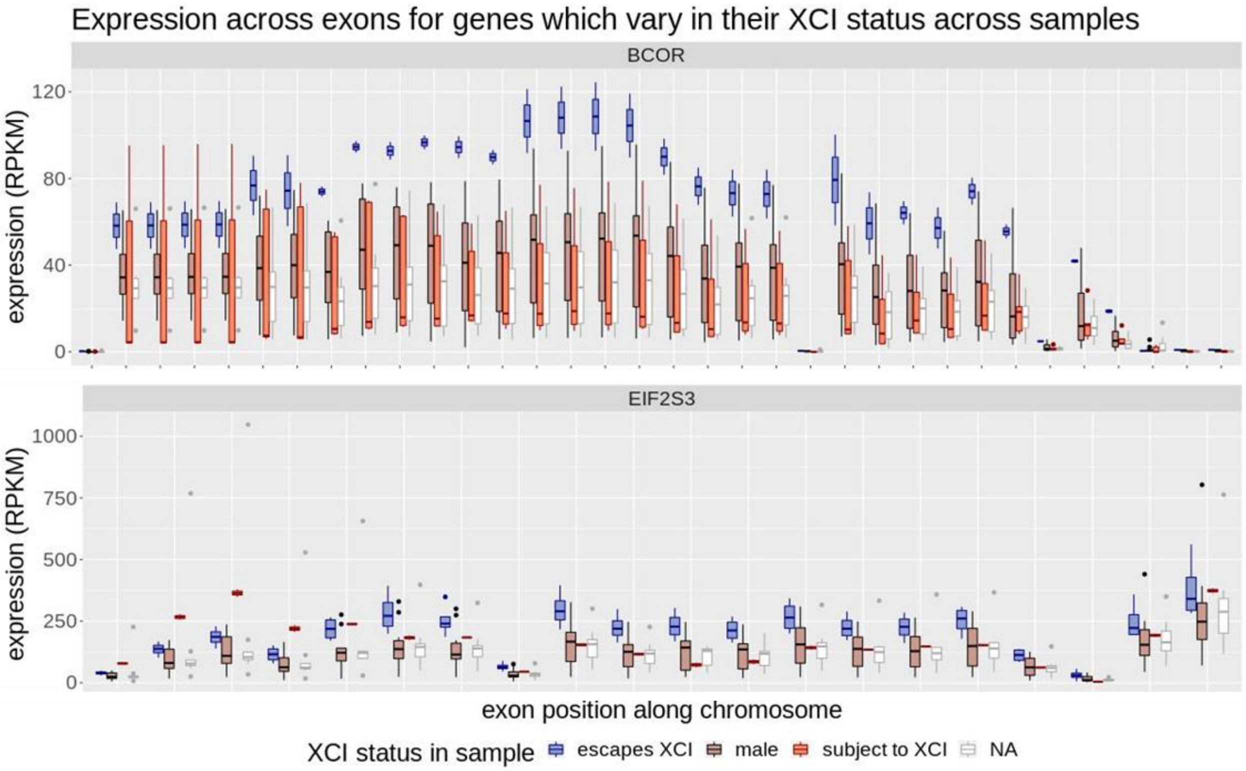


Figure S4.3: Expression across exons for genes with significantly different expression in samples with opposite XCI statuses. XCI status per sample was determine here using Xi/Xa expression.

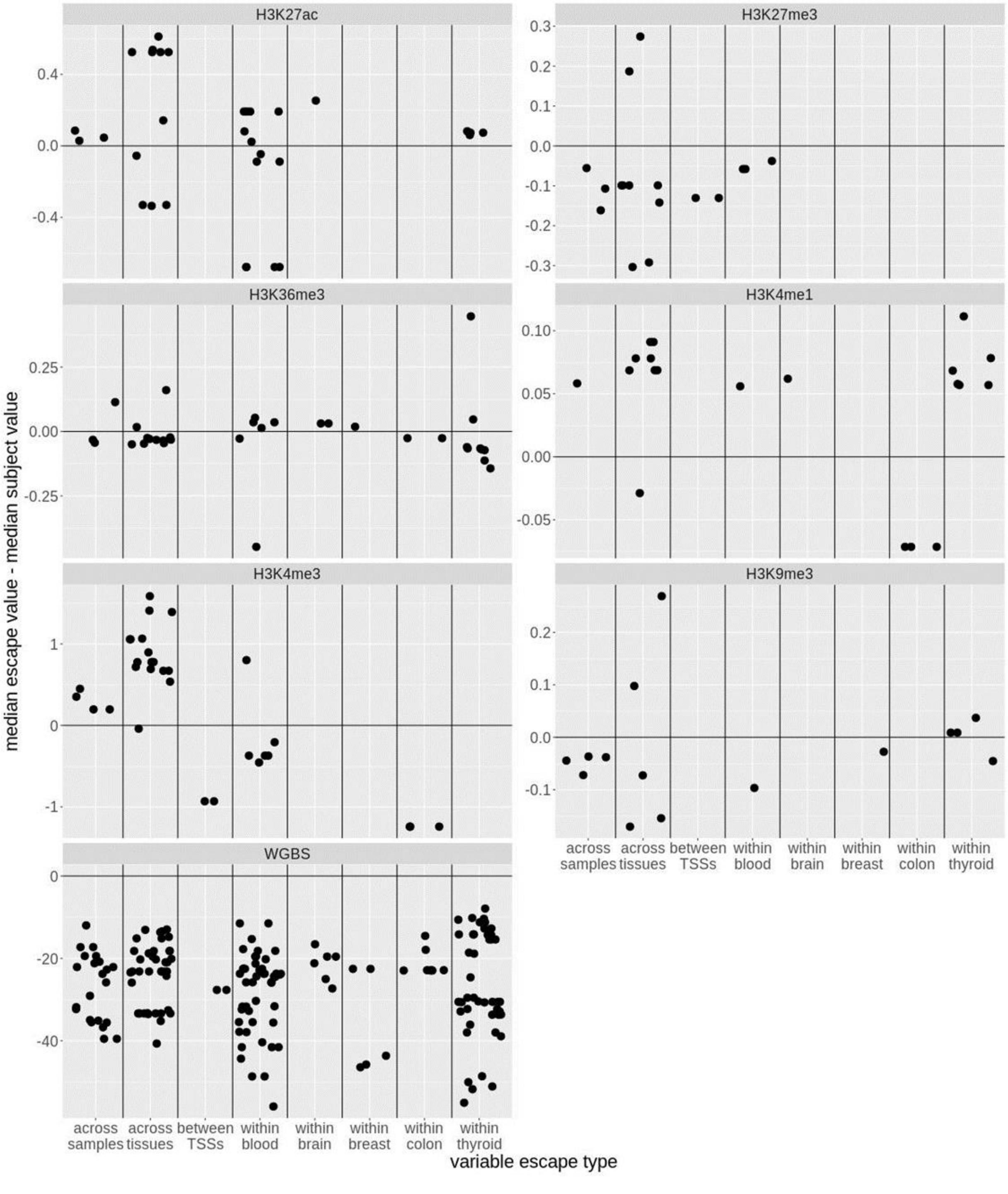


Figure S4.4: Differences in epigenetic marks between samples found escaping vs subject to XCI at variably escaping genes in DNAme. For most of these marks, the region 500bp upstream of the promoter is used, except for H3K36me3 which uses the gene body.

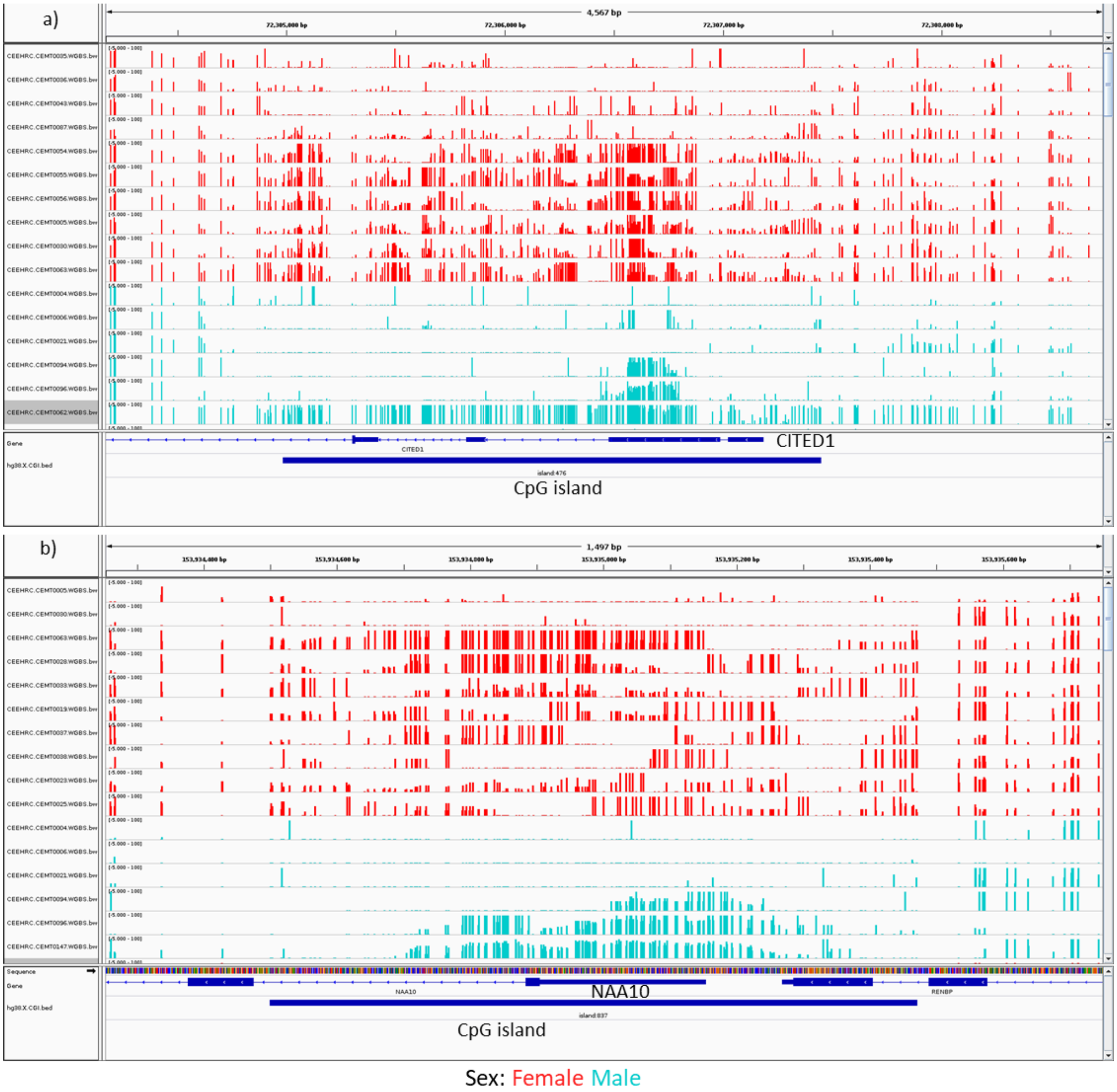


Figure S4.5: IGV view of DNAme bigwig tracks at two variably escaping genes. (A) A view of the CpG island at CITED1. (B) a view of the CpG island at NAA10. A broad representation of samples was sought, some hypomethylated, some hypermethylated and some inconsistent across the CpG island. Broad hypermethylation in males at these genes was rare but is included here as an example of an extreme.

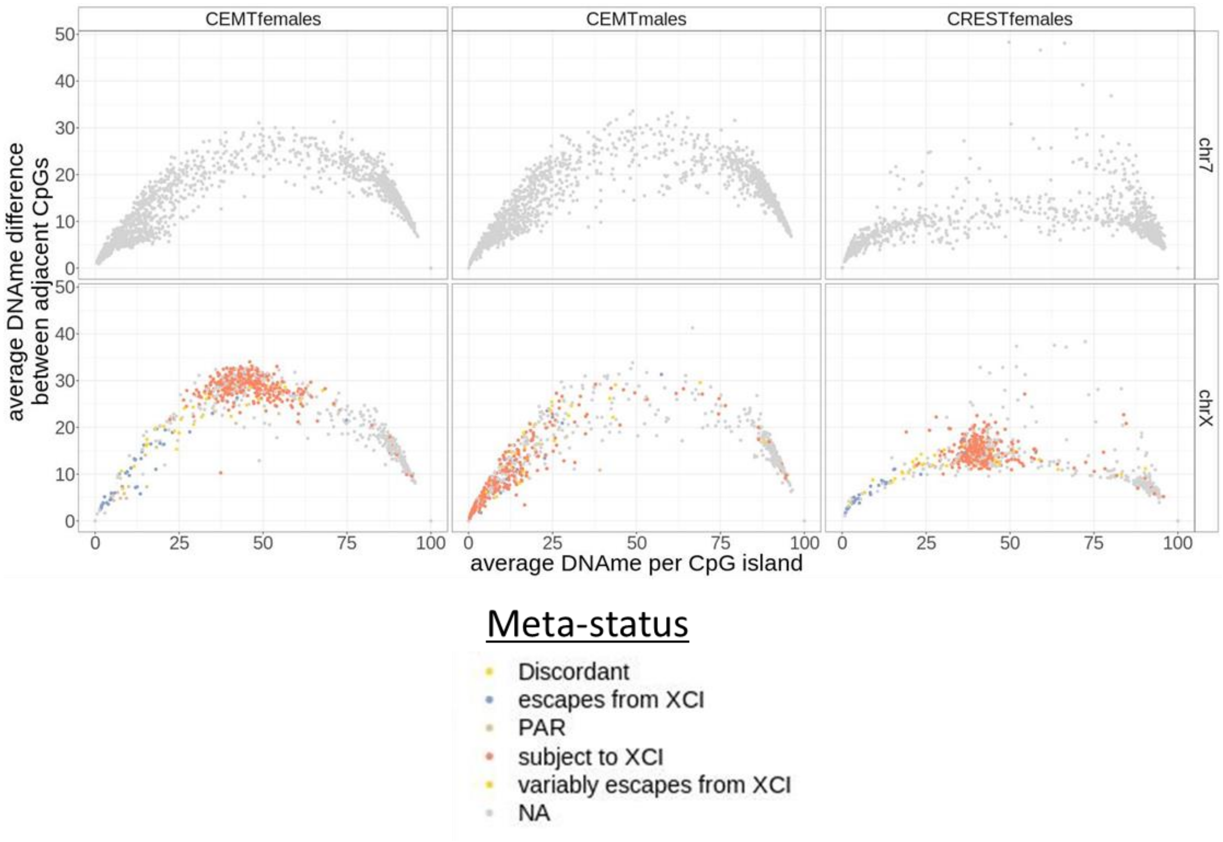


Figure S4.6: average DNAme difference between adjacent CpGs per CpG island. Each point is the average DNAme difference between adjacent CpGs for an individual island, averaged again across samples. Islands are colored by the meta-status of the closest TSS within 2kb. Chr7 was chosen as an autosomal control to show whether the differences are X specific. Males and females from CEMT were used to check for sex specificity and females from CREST were included to check for cancer specificity.

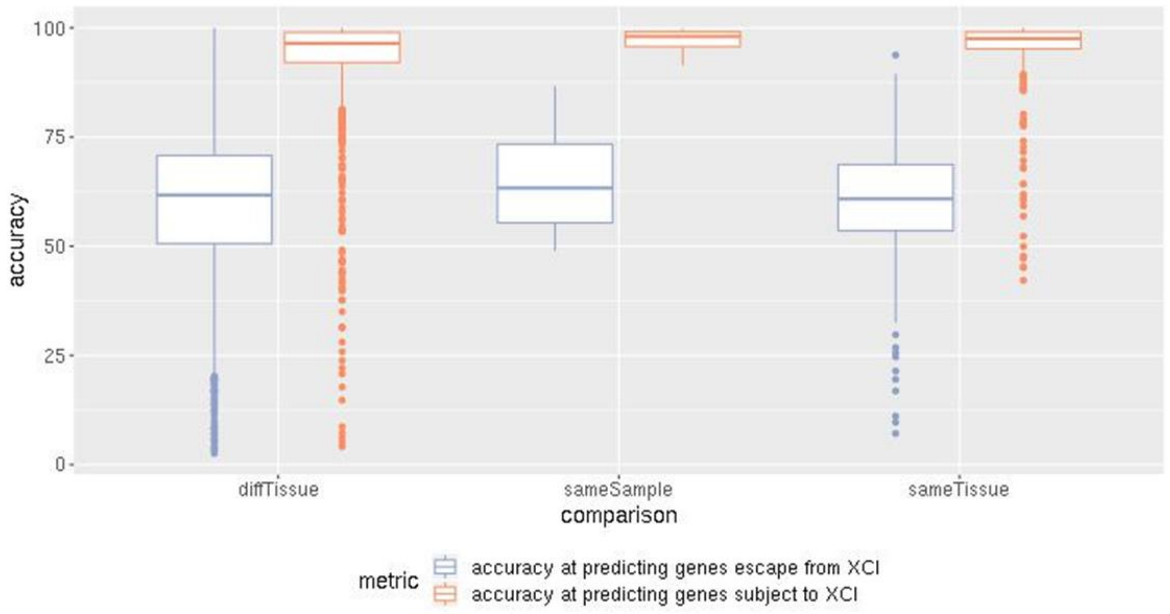


Figure S4.7: Accuracy when models trained in one sample are tested on other models.

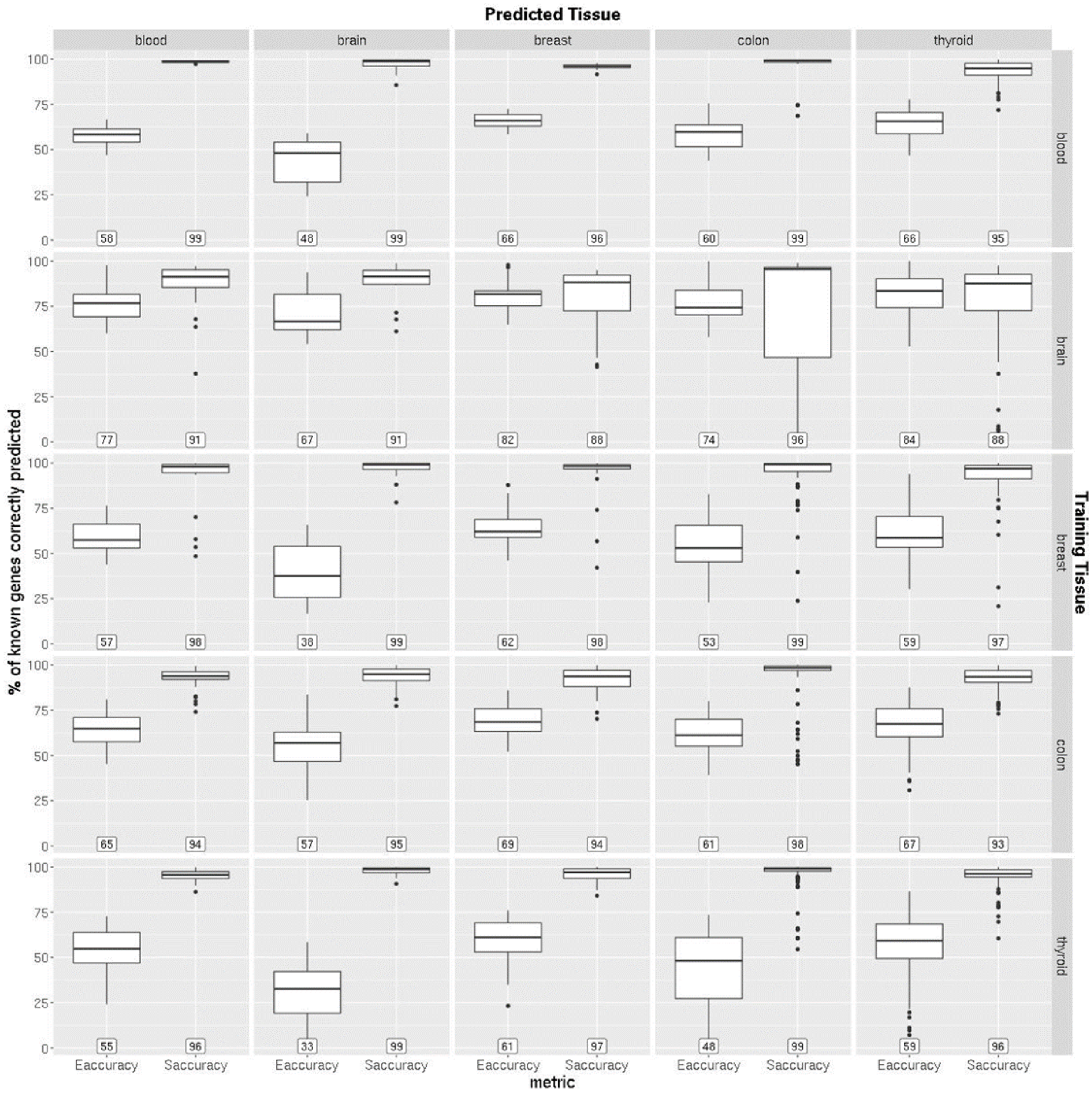


Figure S4.8: Accuracy when models trained in one sample are tested on other models, separated per tissue comparison. The numbers at the bottom of each plot are the median accuracy. Each point is the accuracy at predicted an XCI status when a model from the training tissue on a sample in the predicted tissue. Eaccuracy is accuracy at predicting genes as escape from XCI. Saccuracy is accuracy when predicting genes as subject to XCI.

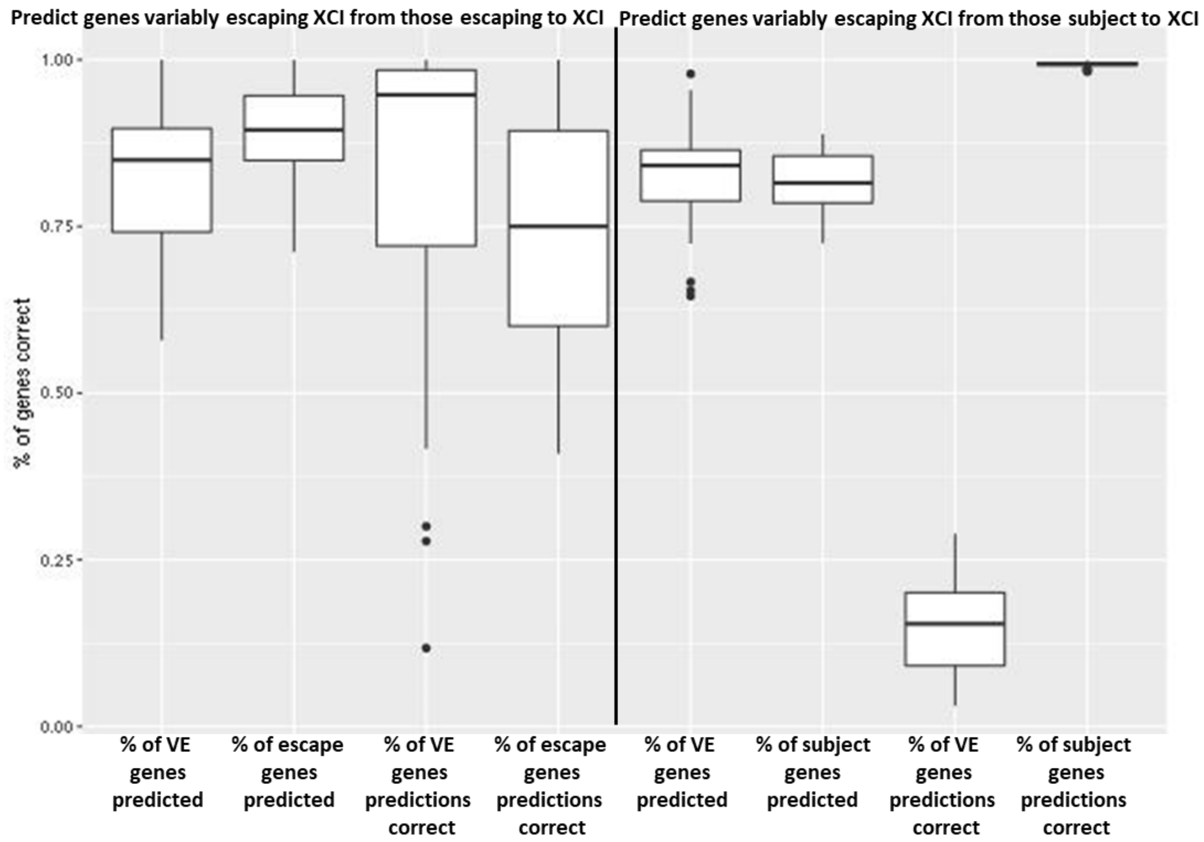


Figure S4.9: Accuracy metrics when predicting which genes variably escape from XCI across samples using data from individual samples. On the left are metrics for when a model is trained on only genes called as escaping XCI in that sample, while the right is metrics when a model is trained on only genes called as subject to XCI. VE is variably escaping from XCI. Subject is subject to XCI.

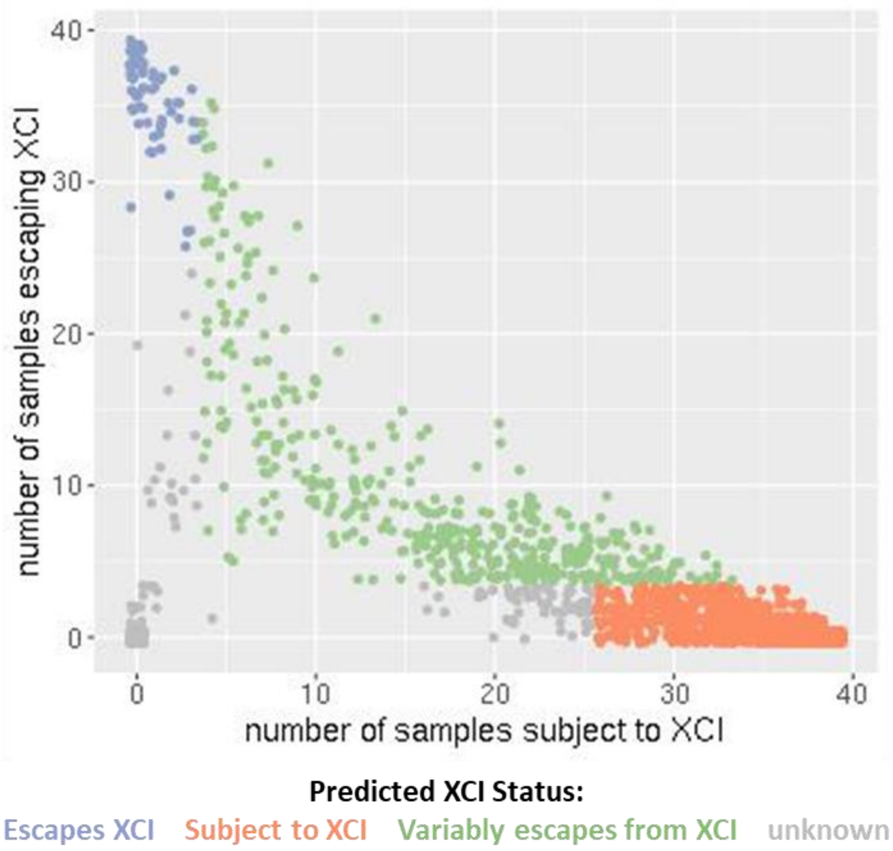


Figure S4.10: The number of samples called as escaping vs subject to XCI per transcript by our epigenetic predictor.

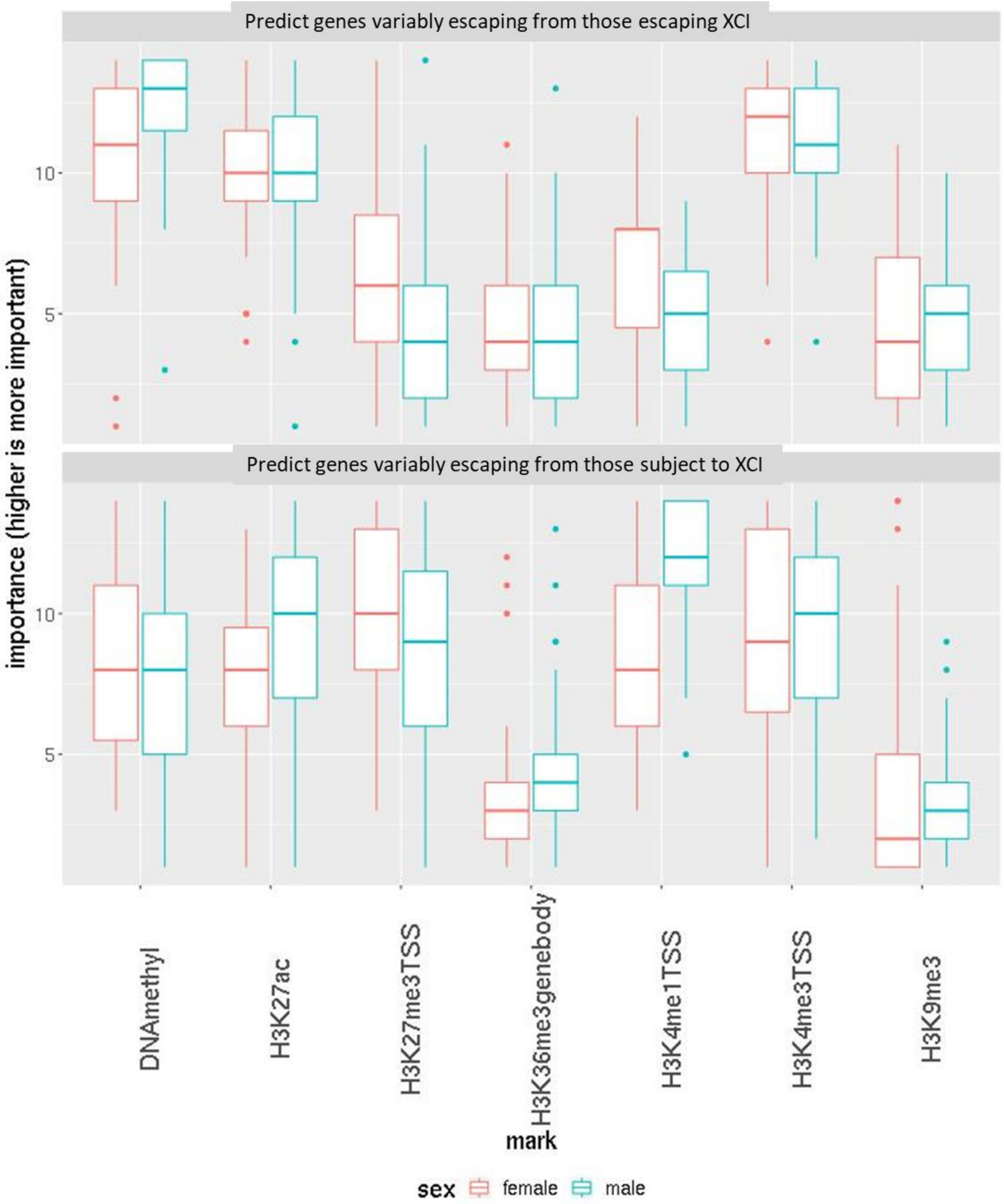


Figure S4.11: Ranked importance of the marks used to predict which genes variably escape across samples. The contributions to each model from each mark were ranked, with rank 14 being the most important and rank one being the least important.

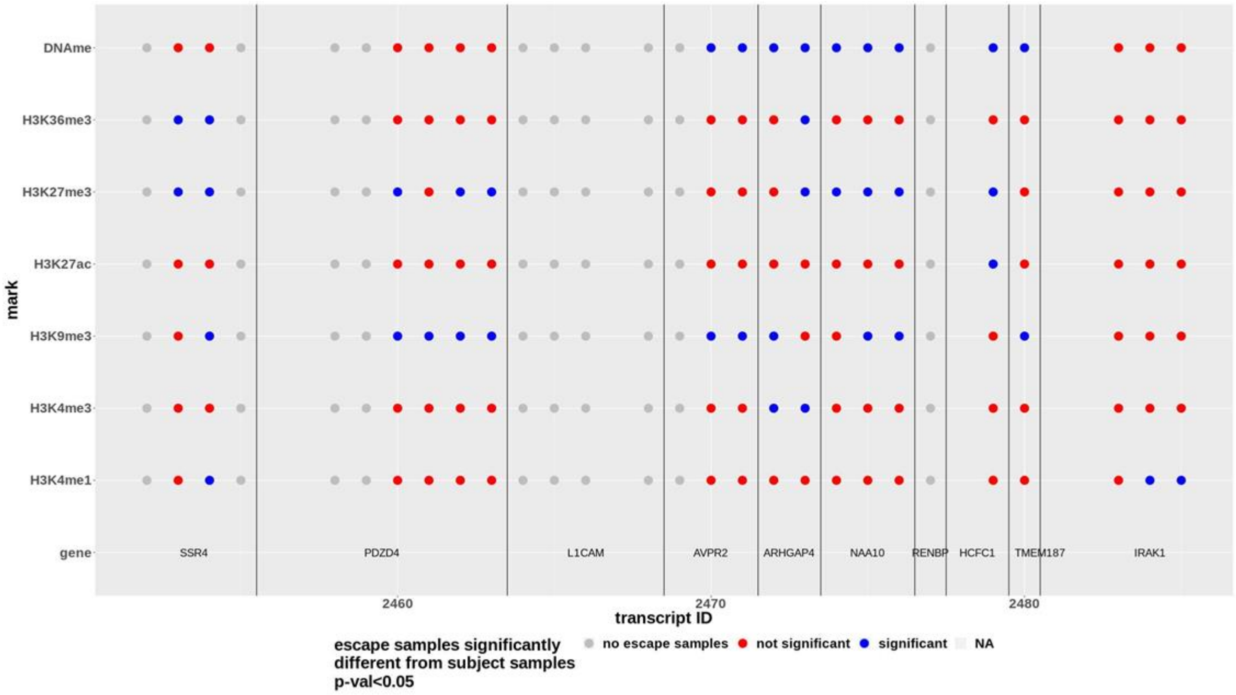


Figure S4.12: Which marks were significantly different between samples predicted as escaping vs subject to XCI in a variably escaping region. Transcript ID is the order that the transcripts are located along the chromosome. There are multiple transcripts per gene but they may be sharing the same TSS and have the same data for all marks but H3K36me3. Vertical lines are drawn denoting which transcripts belong with each gene.

C.2 Supplementary Tables

Table S4.1: List of samples used. See additional file 1**.** For CEMT samples, tissue was manually annotated to combine samples from related areas. Columns D through L are true if the dataset was available for the sample. Patient health status and sample disease are the annotations done by CEMT. CREST samples were only used for the epigenetic predictor and only samples with all datasets available were included here. As the table is too large, please see the version of this chapter available on bioRxiv (Balaton and Brown 2021).

Adjusted p-values

		promoters			enhancers		
		CEMT		CREST	CEMT		
		meta-status	Xi/Xa status	meta-status	meta-status		
		all			all	genic	intergenic
genes escape XCI, Male vs Female	H3K4me1	4.56E-09	2.50E-02	2.89E-20	3.69E-11	1.93E-06	9.41E-08
	H3K4me3	4.54E-07	1.17E-05	7.88E-06	3.26E-04	4.87E-05	1.22E-02
	H3K9me3	5.24E-06	3.00E-03	2.56E-23	2.32E-84	1.46E-28	1.94E-60
	H3K27ac	7.95E-04	1.39E-02	2.54E-02	2.51E-03	6.46E-02	1.02E-03
	H3K27me3	3.78E-07	2.27E-04	5.83E-19	1.30E-187	1.22E-30	3.14E-157
	H3K36me3	1.12E-02	2.40E-01	1.74E-07	3.59E-07	1.49E-04	8.81E-06
	DNAme	9.48E-01	5.60E-01	5.32E-01	9.69E-08	1.17E-08	2.04E-05
	expression	7.37E-01	7.73E-01	6.38E-01			
genes subject to XCI, Male vs Female	H3K4me1	1.99E-09	2.27E-09	6.14E-129	5.88E-74	1.01E-06	1.80E-79
	H3K4me3	4.54E-07	1.69E-05	6.96E-14	7.85E-10	1.65E-08	1.05E-09
	H3K9me3	2.24E-170	1.93E-136	0.00E+00	0.00E+00	0.00E+00	0.00E+00
	H3K27ac	7.29E-05	6.18E-04	1.15E-09	3.11E-14	9.18E-02	8.58E-32
	H3K27me3	5.24E-276	2.28E-255	0.00E+00	0.00E+00	0.00E+00	0.00E+00
	H3K36me3	4.56E-09	2.81E-07	2.38E-89	1.12E-52	1.52E-22	1.77E-34
	DNAme	1.48E-161	3.26E-134	6.10E-162	5.81E-67	9.93E-224	2.53E-06
	expression	2.57E-01	8.31E-01	7.60E-01			
Females, escape genes vs subject genes	H3K4me1	1.12E-02	9.18E-02	8.06E-02	6.04E-05	2.45E-04	3.81E-05
	H3K4me3	6.29E-06	1.21E-04	3.07E-05	5.64E-01	1.12E-06	8.29E-01
	H3K9me3	1.12E-02	3.66E-03	4.45E-03	1.92E-07	4.23E-11	1.06E-01
	H3K27ac	6.05E-06	1.95E-03	1.91E-03	7.62E-02	8.09E-06	3.62E-02
	H3K27me3	9.48E-02	1.89E-06	3.07E-05	3.65E-19	7.24E-05	8.68E-15
	H3K36me3	6.78E-01	3.00E-03	7.60E-01	6.70E-14	1.93E-01	1.29E-14
	DNAme	2.75E-18	7.52E-14	1.48E-19	2.58E-02	1.16E-14	1.45E-07
	expression	5.23E-01	2.58E-01	4.00E-01			
Males, escape genes vs subject genes	H3K4me1	7.29E-05	5.21E-01	6.60E-05	6.69E-05	6.14E-01	4.00E-09
	H3K4me3	9.48E-01	9.54E-01	1.36E-01	2.30E-01	8.23E-03	5.80E-01
	H3K9me3	4.79E-03	2.09E-03	2.30E-01	1.61E-20	3.03E-01	8.17E-20
	H3K27ac	3.63E-03	5.09E-02	2.74E-02	7.86E-02	8.50E-04	4.04E-05
	H3K27me3	2.61E-01	4.54E-01	1.25E-02	2.33E-05	9.00E-01	6.34E-06
	H3K36me3	9.48E-01	3.66E-03	1.26E-01	1.00E-12	3.29E-01	7.07E-13
	DNAme	7.50E-01	8.05E-01	7.60E-01	1.55E-05	1.87E-01	2.24E-02
	expression	5.93E-01	1.90E-01	7.11E-01			

Median Xi/Xa Fold Change		CEMT		CREST	CEMT		
		meta-status	Xi/Xa status	meta-status	meta-status		
		promoters			enhancers		
		all			all	genic	intergenic
females, escape genes	H3K4me1	6.11	5.69	0.46	0.10	0.16	0.09
	H3K4me3	33.87	32.34	2.37	0.02	0.33	0.02
	H3K9me3	2.95	3.07	0.44	0.06	0.06	0.06
	H3K27ac	15.59	7.92	2.09	0.06	0.20	0.05
	H3K27me3	3.81	1.88	0.42	0.12	0.11	0.12
	H3K36me3	3.92	4.24	0.22	0.06	0.05	0.06
	DNAme	3.86	1.45	6.73	65.13	31.86	68.41
	expression	3.73	5.38	0.84			
males, escape genes	H3K4me1	3.21	3.93	0.22	0.08	0.12	0.07
	H3K4me3	15.57	15.50	0.90	0.02	0.27	0.02
	H3K9me3	1.47	1.81	0.17	0.02	0.02	0.02
	H3K27ac	10.40	8.76	0.89	0.05	0.20	0.04
	H3K27me3	0.85	0.84	0.12	0.03	0.02	0.03
	H3K36me3	2.32	2.92	0.12	0.04	0.04	0.05
	DNAme	2.12	0.81	1.48	63.87	11.49	66.63
	expression	2.45	7.15	0.70			
Females, subject genes	H3K4me1	5.16	5.29	0.45	0.10	0.14	0.09
	H3K4me3	20.25	20.68	1.19	0.03	0.27	0.02
	H3K9me3	3.59	3.55	0.47	0.06	0.08	0.06
	H3K27ac	8.07	9.66	0.91	0.06	0.18	0.05
	H3K27me3	5.72	5.55	0.54	0.13	0.13	0.13
	H3K36me3	3.16	3.09	0.23	0.05	0.05	0.05
	DNAme	39.75	38.77	40.67	59.26	45.70	63.74
	expression	4.91	6.95	0.42			
Males, subject genes	H3K4me1	4.18	4.42	0.27	0.08	0.13	0.06
	H3K4me3	16.69	17.27	0.79	0.02	0.24	0.01
	H3K9me3	1.00	1.00	0.18	0.02	0.02	0.02
	H3K27ac	6.94	8.43	0.65	0.04	0.17	0.04
	H3K27me3	0.90	0.78	0.16	0.03	0.02	0.03
	H3K36me3	2.09	1.98	0.14	0.04	0.04	0.04
	DNAme	2.47	1.76	2.43	60.81	12.78	66.00
	expression	4.86	7.33	0.44			

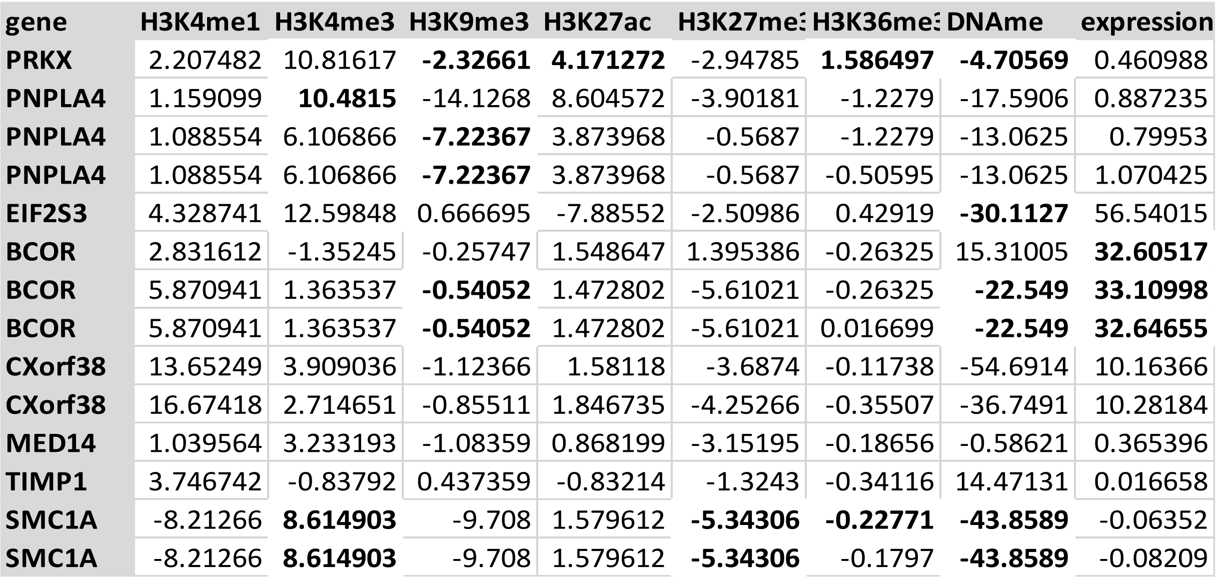
fold change(Xi:Xa) ratio calculated as log2(female-male)/male		CEMT		CREST	CEMT		
		meta-status	Xi/Xa status	meta-status	meta-status		
		promoters			enhancers		
		all			all	genic	intergenic
Xi genes escaping XCI	H3K4me1	-0.14	-1.17	0.11	-2.04	-1.48	-1.85
	H3K4me3	0.23	0.12	0.70	-1.71	-2.20	-1.49
	H3K9me3	0.02	-0.51	0.71	0.74	1.13	0.60
	H3K27ac	-1.00	NA	0.42	-1.83	NA	-1.85
	H3K27me3	1.80	0.30	1.28	1.67	2.46	1.47
	H3K36me3	-0.53	-1.14	-0.37	-1.66	-1.34	-1.70
	DNAme*	5.61	2.08	11.99	66.39	52.22	70.18
	expression	-0.94	NA	-2.34			
Xi genes subject to XCI	H3K4me1	-2.09	-2.34	-0.61	-1.55	-4.59	-1.06
	H3K4me3	-2.23	-2.34	-0.95	-1.08	-2.94	-1.03
	H3K9me3	1.37	1.35	0.62	1.21	1.63	0.99
	H3K27ac	-2.63	-2.78	-1.32	-1.48	-4.08	-1.48
	H3K27me3	2.42	2.61	1.30	2.00	2.48	1.87
	H3K36me3	-0.95	-0.84	-0.70	-1.20	-1.16	-1.23
	DNAme*	77.03	75.78	78.91	57.71	78.62	61.47
	expression	-6.60	NA	NA			

Table S4.2: Comparison of histone marks between sex and XCI status. The first table shows BH adjusted p-values comparing female vs male and escape genes vs those subject to XCI per mark in in CEMT with our meta-status and Xi/Xa expression based XCI status calls, along with CREST data with meta-status calls and CREST data at enhancers with meta-status calls of linked genes. The second table shows the median value per mark with each sex and XCI status and the third shows the Xi/Xa ratio and log2 fold change per mark calculated based off of that median. NA values in the 3rd table mean that the female was lower than male so the Xi/Xa fold change could not be computed. DNAme is grey in the last table as it is calculated differently and on a different scale; here it is showing the estimated level of Xi DNAme.

	H3K4me1	H3K4me3	H3K9me3	H3K27ac	H3K27me3	ENCODE H3K27me3	H3K36me3	DNAme
escapes from XCI	0.45	0.56	0.47	0.08	**0.97**	**0.72**	0.46	0.26
no previous call	0.28	0.15	0.66	0.01	**0.87**	0.49	0.30	0.29
PAR	0	0.00	0.28	0.00	0.04	0.11	0.17	0.04
subject to XCI	0.10	0.07	**0.88**	0.01	**0.90**	**0.83**	0.32	**0.80**
variably escapes from XCI	0.15	0.09	**0.80**	0.01	**0.93**	**0.93**	0.30	0.54
chr7	0.15	0.03	0.05	0.08	0.10	0.17	0.07	0.30

Table S4.3: The ratio of TSSs with significant differences between males and females for various epigenetic marks using CEMT data. The denominator was the total number of informative TSSs for which we had data. For most marks this was measured as 500bp upstream of the promoter, but for H3K36me3 we measured the mark across exons. For H3K36me3 we used unique transcripts instead of unique TSSs. Marks significant in over 70% of informative TSSs are in bold. All of the H3K27me3 data from ENCODE was downloaded and used as a replication dataset. Chromosome 7 (chr7) was included as an example autosome.

Table S4.4: All XCI status calls made here. The first sheet contains a single XCI status call per gene per method. Published calls are from Balaton, *et al*. 2015. Other sheets contain all calls per sample for each method. Each row is one entry into the model, so Xi/Xa is per gene and the others are for unique transcripts. For DNAme, the samples on the far right in shades of grey are males while the samples on the left in color are females. For the epigenetic predictor, separate low confidence categories were made for when transcripts have only 12-14 of the 20 models per sample predicted a certain XCI status. As the table is too large, please see the version of this chapter available on bioRxiv (Balaton and Brown 2021).



gene	H3K4me1	H3K4me3	H3K9me3	H3K27ac	H3K27me3	H3K36me3	DNAme	expression
PRKX	2.207482	10.81617	**-2.32661**	**4.171272**	-2.94785	**1.586497**	**-4.70569**	0.460988
PNPLA4	1.159099	**10.4815**	-14.1268	8.604572	-3.90181	-1.2279	-17.5906	0.887235
PNPLA4	1.088554	6.106866	**-7.22367**	3.873968	-0.5687	-1.2279	-13.0625	0.79953
PNPLA4	1.088554	6.106866	**-7.22367**	3.873968	-0.5687	-0.50595	-13.0625	1.070425
EIF2S3	4.328741	12.59848	0.666695	-7.88552	-2.50986	0.42919	**-30.1127**	56.54015
BCOR	2.831612	-1.35245	-0.25747	1.548647	1.395386	-0.26325	15.31005	**32.60517**
BCOR	5.870941	1.363537	**-0.54052**	1.472802	-5.61021	-0.26325	**-22.549**	**33.10998**
BCOR	5.870941	1.363537	**-0.54052**	1.472802	-5.61021	0.016699	**-22.549**	**32.64655**
CXorf38	13.65249	3.909036	-1.12366	1.58118	-3.6874	-0.11738	-54.6914	10.16366
CXorf38	16.67418	2.714651	-0.85511	1.846735	-4.25266	-0.35507	-36.7491	10.28184
MED14	1.039564	3.233193	-1.08359	0.868199	-3.15195	-0.18656	-0.58621	0.365396
TIMP1	3.746742	-0.83792	0.437359	-0.83214	-1.3243	-0.34116	14.47131	0.016658
SMC1A	-8.21266	**8.614903**	-9.708	1.579612	**-5.34306**	**-0.22771**	**-43.8589**	-0.06352
SMC1A	-8.21266	**8.614903**	-9.708	1.579612	**-5.34306**	-0.1797	**-43.8589**	-0.08209

Table S4.5: Value of differences in epigenetic marks between samples with opposite XCI statuses at genes found variably escaping XCI by Xi/Xa expression. Those found significant in Table 1 are bolded.

H3K4me1	0.093906
H3K4me3	3.34E-11
H3K9me3	7.66E-68
H3K27ac	8.11E-09
H3K27me3	1.03E-29
H3K36me3	0.067493

Table S4.6: adjusted p-values comparing marks in females between genes found subject to XCI vs escaping XCI by DNAme.

bin	**<25%**	33-66%	>75%
0	0.990521	0	0
0-10%	0.967776	0.011134	0.002227
10-20%	0.731334	0.138481	0.035099
20-30%	0.591919	0.172437	0.130438
30-40%	0.521443	0.128352	0.242815
40-50%	0.447618	0.093807	0.359666
50-60%	0.337163	0.11472	0.44407
60-70%	0.213763	0.151308	0.5045
70-80%	0.08201	0.16705	0.602591
80-90%	0.026355	0.094393	0.755547
90-100%	0.0059	0.02496	0.914908

Table S4.7: Distribution summary for DNAme per read. The number is what percent of reads in each bin were below 25%, between 33 and 66% or over 75% DNAme.

Histone Mark	Accuracy for genes escaping XCI	Accuracy for genes subject to XCI
H3K4me1	0.27	0.45
H3K4me3	0.36	0.55
H3K9me3	0.11	0.16
H3K27ac	0.33	0.64
H3K27me3	0.51	0.21
H3K36me3	0.09	0.03

Table S4.8: The accuracy of simple models predicting XCI status from a single histone mark. These accuracies are low because the models overpredicted variable escape from XCI as there is large overlap between the two XCI statuses.

	No CpG island	CpG island	low expression	high expression
escapes XCI	8	95	26	77
subject to XCI	600	1116	549	1167
variably escapes XCI	2	7	1	8
inconsistent prediction	462	346	569	239

Table S4.9: XCI status calls made using a random forest epigenetic predictor, split by presence or absence of a CpG island and expression. The threshold used to split low from high expression is a median of 0.1 RPKM across samples. Inconsistent predictions had over a third of samples with fewer than 15 of the 20 models trained agree on an XCI status.

	Variable escape across individuals	**Variable escape across tissues**		**Variable escape across TSSs**	
Total number of genes	**9**	**65**	**2636**	**1**	**461**
H3K27me3 TSS	0	**51%**	27%	0	12%
H3K27me3 gene-body	0	28%	41%	0	22%
H3K27ac TSS	0	31%	37%	100%	17%
H3K9me3 TSS	0	32%	19%	0	30%
H3K9me3 gene-body	11%	8%	7%	100%	29%
H3K4me3 TSS	0	**40%**	20%	0	22%
H3K4me1 TSS	0	62%	53%	100%	19%
H3K36me3 TSS	0	35%	22%	0	12%
H3K36me3 gene-body	11%	29%	31%	0	38%
DNAme TSS	67%	**58%**	28%	100%	19%
expression	0	**38%**	1%		

Table S4.10: The percent of genes found variably escaping by our epigenetic predictor with significant differences in various epigenetic marks. Genes were counted as significant if BH corrected p-values were less than 0.01 when comparing samples predicted as subject to XCI to samples predicted as escaping from XCI. The total number of genes row shows the total number of genes in each category. The variable escape across tissues and TSSs categories have 2 columns each, the left column being the percent of variably escaping genes with significant differences between tissues/TSSs and the right column being the percent of all genes on the X chromosome with differences between tissues/TSSs. Highlighted in blue are marks which were significantly more likely to have significant differences between tissues/TSSs at genes predicted to variably escape than in all X linked genes.

variable escape threshold	33%	25%	10%	5%
Number of genes variably escaping across samples	9	41	431	740
H3K27me3 TSS	0%	4.90%	23%	27%
H3K27me3 gene-body	0%	0%	3.70%	4.20%
H3K27ac TSS	0%	2%	3.50%	7.80%
H3K9me3 TSS	0%	0%	11%	15%
H3K9me3 gene-body	11%	0%	2.10%	4.30%
H3K4me3 TSS	0%	0%	2.80%	4.20%
H3K4me1 TSS	0%	4.90%	7.70%	10%
H3K36me3 TSS	0%	2.40%	13%	16%
H3K36me3 gene-body	11%	7%	3.50%	6.20%
DNAme	67%	22%	17%	20%
expression	0%	0%	0.90%	0.70%

Table S4.11: The percent of genes found variably escaping by our epigenetic predictor with significant differences in various epigenetic marks across various variable escape thresholds. Variable escape threshold is the number of samples with each XCI status (escaping from XCI and subject to XCI) that were required in order to call a gene as variably escaping from XCI across samples. Genes were counted as significant if BH corrected p-values were less than 0.01 when comparing samples predicted as subject to XCI to samples predicted as escaping from XCI.

	CEMT calls per gene			
		Escapes XCI	Subject to XCI	Variably escapes XCI
CREST calls per gene	Escapes XCI	24	19	6
	Subject to XCI	2	658	1
	Variably escapes XCI	1	12	1

Table S4.12: Comparing XCI status calls made by an epigenetic predictor in the CEMT dataset vs a similar model in the CREST dataset.

	VE across individuals	**VE across tissues**		**VE across TSSs**	
total	**8**	**13**	**2313**	**6**	**1155**
H3K4me1 Female	0	0	0.076%	0	0.087%
H3K4me1 Male	0	<NA>	NA	17%	4.2%
H3K4me3 Female	0	0	0	**50%**	**1%**
H3K4me3 Male	0	<NA>	NA	**67%**	**6%**
H3K9me3 Female	0	0	0	0	0
H3K9me3 Male	0	<NA>	NA	33%	5%
H3K27Ac Female	0	0	0.08%	17%	1%
H3K27Ac Male	0	<NA>	NA	**50%**	**5%**
H3K27me3 Female	0	**15%**	**0.11%**	**17%**	**9%**
H3K27me3 Male	0	<NA>	NA	**33%**	**2%**
H3K36me3 Female	0	0	0	0	2%
H3K36me3 Male	0	<NA>	NA	**50%**	**3%**
WGBS Female	0	**92%**	**2.31%**	**100%**	**6%**
WGBS Male	0	<NA>	NA	**100%**	**1%**
RNA-seq Female	0	**46%**	**1.33%**	0	3%
RNA-seq Male	0	<NA>	NA	17%	3%

Table S4.13: The percent of genes found variably escaping by our epigenetic predictor in the CREST dataset with significant differences in various epigenetic marks. Genes were counted as significant if BH corrected p-values were less than 0.01 when comparing samples predicted as subject to XCI to samples predicted as escaping from XCI. The total number of genes row shows the total number of genes in each category. The variable escape across tissues and TSSs categories have 2 columns each, the left column being the percent of variably escaping genes with significant differences between tissues/TSSs and the right column being the percent of all genes on the X chromosome with differences between tissues/TSSs. Highlighted in blue are marks which were significantly more likely to have significant differences between tissues/TSSs at genes predicted to variably escape than in all X linked genes.