EVOLUTION OF X CHROMOSOME INACTIVATION ESCAPE IN MAMMALS

A Dissertation
Presented to the Faculty of the Graduate School
of Cornell University
In Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by
Andrea Jane Slavney
May 2018

EVOLUTION OF X CHROMOSOME INACTIVATION ESCAPE IN MAMMALS

Andrea Jane Slavney Ph. D.
Cornell University 2018

In eutherian mammals, dosage compensation between XX females and XY males occurs in part through X chromosome inactivation (XCI) of one X homolog per cell in females. XCI was previously assumed to be complete across the X-specific regions of X chromosomes, but recently genes outside the pseudoautosomal regions (PARs) have been observed to exhibit low but significant expression from the inactive X in several species. These genes are termed XCI "escapers". XCI escape introduces gene expression variation between females and males and inflates variation among females. However, it remains unclear whether XCI escape is generally a "bug" in the XCI system – that is, merely a consequence of inefficient XCI – or a "feature" required for normal biological function.

By comparing XCI profiles across mammals, we can gain insight into the evolutionary history of XCI and XCI escape, with the hope that this will improve our understanding of their mechanisms and functions. Here, I describe three analyses motivated by these goals. First, we expanded upon earlier findings to show that human XCI escapers exhibit a greater degree of strong purifying selection, as well as higher and broader gene expression than inactivated genes in both sexes. These results suggest a role for gene expression patterns in determining XCI status after divergence from the Y chromosome.

Next, we generated a novel XCI profile for the domestic dog from single cell RNA-seq from two female F1 crossbreed dogs. Cell-level gene expression data was used to classify

X genes as showing either monoallelic or bialleleic expression in multiple cells. Using this method, we identified 45 putative XCI escapers and 98 putative X-inactivated genes.

Lastly, we performed a comparative analysis of XCI profiles across human, mouse, dog, and opossum 1:1 X orthologs. This analysis revealed that XCI escape is highly lineage-specific. Despite this, XCI escapers in these species overlap highly in their Gene Ontology biological process annotations. XCI escapers also appear to be under greater evolutionary constraint than inactivated genes within most species, though the magnitude varies.

Overall, our findings suggest that the forces driving XCI escape evolution vary extensively across genes within and between species.

BIOGRAPHICAL SKETCH

I was born in Madison, Wisconsin to Linda and Michael Slavney in 1989, and lived in the city of Sun Prairie, Wisconsin between 1989 and 1997. Our family moved to Madison in 1997, a few blocks from my father’s childhood home. I remained there for the next fourteen years, attending Madison West High School from 2003 to 2007 and then the University of Wisconsin – Madison from 2007 to 2011. I graduated from UW as a third-generation alumna, with a B.S. in Genetics and a second major in Anthropology. While at UW, I worked for three years as a laboratory assistant in the lab of Professor David C. Schwartz, primarily collecting data for whole-genome optical restriction mapping. My senior thesis in Molecular Anthropology was supervised by Dr. John Hawks, who first suggested that I begin learning bioinformatics and programming.

After graduating from college in December 2011, I spent six months working as a full-time technician in the Schwartz lab and interviewing for Genetics and Genomics PhD programs. When the time came to choose among them, the Genetics and Development program at Cornell University stood out as the clear winner not only because of its beautiful setting and impressive scientific history, but its collegial faculty and socially-conscious students. My committee members and the members of Andy Clark’s lab exemplify these qualities, and I am proud that several of them have become my mentors and close friends over the past six years.

To my family, and my Cornell family.

ACKNOWLEDGEMENTS

I would first like to thank current and former members of the Clark and Keinan labs for their assistance and support throughout my Ph.D. In particular, I want to acknowledge Diana Chang and Kaixiong Ye for their help with XWAS; Leonardo Arbiza, Aaron Sams, and Aviv Madar for being great research mentors; Alon Keinan for allowing me to work with his group; Xu Wang for taking the time to share his knowledge of X chromosome inactivation and allele specific expression; Elissa Cosgrove for providing her invaluable computational expertise and her work on the dog analysis; Manisha Munasinghe, Florencia Schlamp, Jullien Flynn, and Yasir Ahmed-Braimah for their advice on RNA-seq; and my other office mates Sri Raj, Emily Davenport, Francisco Agosto-Perez, and Ian Caldas for making 101 Biotech such a wonderful place to work.

Additionally, I want to thank our collaborators in the canine XCI profile project: Rory Todhunter and the Cornell Veterinary Biobank staff for generously providing us with sampling materials; Doug Antzcak and Don Miller for their excellent work isolating PBMCs from the blood samples; Peter Schweitzer for performing the scRNA-seq and assisting with the associated software; Aaron Sams for his assistance with the Embark SNP data; Stacy Parks for being so enthusiastic and accommodating in letting us sample her dogs; and the dogs themselves – Brie and Surrie – for being extremely patient with all of the poking and prodding that they endured in the process.

Finally, I thank my committee members, Andy Clark, Chip Aquadro, and Charles Danko, for their guidance and mentorship throughout my time at Cornell. I could not have done this without their vision and encouragement, and I feel extremely privileged to have worked with them.

TABLE OF CONTENTS

Biographical Sketch……………………………………………………………………..v

Dedication………………………………………………………………………......vi

Acknowledgments..……………………………………...……………………...vii

List of Abbreviations……………………………………………………….………x

List of Symbols………………………….…………………………………………..x

Introduction and Literature Review

X chromosome inactivation and escape……………………………………………...1

Extension of the human XCI profile……………….………………………………...3

Importance to the field……………….………………………………….………...4

Research questions……………….…………………………………….…………6

Chapter 1: Gene expression as a potential driver of XCI escape evolution in humans

Introduction………………………...………………………………………………..9

Results……………………………………………………………………………11

1. Composite human XCI profile……………………………………...…11

2. Selection analysis……………………………….…………………………...12

3. Contrasting gene expression patterns of XCI escapers and X-inactivated genes……………………………………………………………………………..13

Conclusions……………………………………………………………………17

Methods………………………………………………………………………......20

Chapter 2: Preliminary canine XCI profile from single-sell RNA-seq

Introduction………………………………………………………………………29

Results……………………………………………………………………………32

1. Genotyping………………………………………..…………………...32

2. Single cell RNA-sequencing……………………..……………….…………32

3. XCI profile…………………………………………………………..…...34

Conclusions…………………………………………………………………………35

Methods……………………………………………………………………………..37

Chapter 3: Conservation of XCI escape across mammals

Introduction…………………………………………………………………………42

Results……………………………………………………………………………44

1. XCI escape conservation……………………………………………………44

2. Functional conservation……………………………………………………46

3. Sequence conservation…………………………………………………...46

Conclusions…………………………………………………………………………48

Methods……………………………………………………………………………..52

Chapter 4: Discussion

Summary……………………………………………………………………………56

Implications……………………………………………………….……….…...57

Future directions……………………………….……………..………………….…58

Tables and Figures…………………………………………………………………...60

List of Supplementary Materials……………………………………………………79

References…………………………………………………………………………….…80

LIST OF ABBREVIATIONS

cDNA – Complementary DNA

CpG site – C (cytosine) G (guanine) DNA dinucleotide site

mCH – Non-CG DNA dinucleotide methylation

PAR – Pseudoautosomal region

PBMC – Peripheral blood mononuclear cell

RNA-seq – RNA sequencing

RT-PCR – Reverse transcription polymerase chain reaction

scRNA-seq – Single cell RNA-seq

SNP – Single nucleotide polymorphism

TSS – Transcription start site

X, chrX – X chromosome

XAR – X-added region

XCI – X chromosome inactivation

XCR – X-conserved region

Y – Y chromosome

LIST OF SYMBOLS

τ - Tissue specificity index, tau (Yanai et al. 2005)

X_a – Active X chromosome

X_i – Inactive X chromosome

INTRODUCTION
AND LITERATURE REVIEW

X chromosome inactivation and escape

In species with chromosomal sex determination, one or more sex chromosomes may encode genes that are essential in both sexes. This introduces the problem that females and males may have intrinsically different stoichiometry between sex chromosome and autosomal gene products, potentially causing deleterious effects. Susumu Ohno and others hypothesized that sex chromosome dosage compensation processes function at the transcription level to equalize gene expression between males and females, and between the sex chromosomes and the ancestral autosomes from which they evolved (Ohno 1967).

A wide variety of species have evolved some form of sex chromosome dosage compensation. In eutherian mammals, dosage compensation between XX females and XY males is achieved through a process known as X chromosome inactivation (XCI), or lyonization (Lyon 1961, 1962). During XCI, one of the two X chromosomes in each female somatic cell is epigenetically silenced early in embryonic development. This process involves extensive chromatin remodeling facilitated by the recruitment of polycomb repressor complexes to the inactivated X (X_i) (de Napoles et al. 2004), which ultimately produces a condensed heterochromatic structure known as the Barr body (Barr and Bertram 1949). As embryonic development proceeds, the epigenetic state of the X_i is transmitted down each cell's lineage, producing a mosaic pattern of X allele expression in the fully developed organism.

XCI was previously assumed to be complete across the X-specific regions of X chromosomes, i.e. the regions that no longer recombine with the Y chromosome. However, genes outside the pseudoautosomal regions (PARs) have more recently been observed to exhibit unique epigenetic profiles (Miller and Willard 1998), and later, low but significant expression in several therian and eutherian species (Carrel and Willard 2005; Wang et al. 2012; Wang et al. 2014; Berletch et al. 2015). These genes, referred to as XCI "escapers", include the long non-coding RNA *XIST*, which is solely transcribed from X_i and facilitates the recruitment of XCI machinery to the rest of the chromosome during lyonization (Brown et al. 1991). XCI escape at another non-PAR human gene was later reported (Carrel and Willard 1999), and the same group went on to identify many more non-PAR human X genes that escaped XCI to some degree (Carrel and Willard 2005).

In this first major study of XCI escape, Carrel and Willard performed semi-quantitative RT-PCR assays on nine human-murine hybrid fibroblast cell lines (Carrel et al. 1999) to assess the degree of expression from the X_i allele relative to the active X (X_a) allele in each line. Because this method was only semi-quantitative, they used an X_i to X_a expression ratio of greater than or equal to 10% as the lower limit at which biallelic expression was considered to be significant. Using this method, they tested over 600 of the approximately 1,100 human X genes annotated at that time and found that over 200 of them show some degree of significant biallelic expression. They further broke down XCI escapers by the proportion of human-murine hybrid cell lines in which they escaped, calling genes escaping in three to six of the nine cell lines "heterogeneous" escapers, and those escaping in seven or more of the nine cell lines "consistent" escapers. Genes escaping in one or two of the nine cell lines were considered to be X inactivated. Using these definitions, they reported 94 consistent XCI escapers, 62 heterogeneous escapers, and 458 X-inactivated genes. This study also confirmed and refined an earlier observation that human XCI escapers are not randomly distributed across the X chromosome (Miller and Willard

1998): most are clustered together in the distal part of the p arm of the chromosome, which diverged from the Y chromosome more recently than the rest of the chromosome (Ross et al. 2005).

Extension of the human XCI profile

The expression level-based definition of XCI escape introduced by Carrel and Willard (2005) has been widely used in other studies of XCI profiles – i.e. which genes escape or are subject to X inactivation in a given species. However, more recent studies have expanded the catalog of human XCI escapers by using a different, but strongly correlated biological feature: DNA methylation. Cotton et al (2015) showed that methylation levels at CpG islands in a gene's transcription start site (TSS) were strongly predictive of its XCI status in the Carrel and Willard 2005 study, with strongly female-biased TSS methylation levels correlating with X-inactivation and lower female bias correlating with XCI escape. Additionally, Schultz et al. (2015) demonstrated that gene body non-CpG methylation (mCH) levels showed the opposite pattern, with escapers showing a higher female bias in mCH levels than X-inactivated genes. Using methylation data to identify escapers allowed these two groups to assay many weakly-expressed genes that could not be tested using allele-specific expression.

Both of these studies also made the innovation of assaying XCI profiles in multiple tissues and cell lines and found that a gene's XCI status can vary across tissues within the same individual. Furthermore, many of the tissues used in the Cotton et al. study included multiple – sometimes hundreds – of individuals. With this data, they were able to show that heterogeneous, or in their terms, "variable" XCI escapers are common, but that most are either primarily X-inactivated or primarily escape across all individuals. Additionally, they found few genes that showed extremely tissue-specific escape. These findings demonstrated that XCI escape is an under-investigated source of gene expression variation across females, as well as between females and males.

Importance to the field

The primary motivation for studying XCI and XCI escape is to gain new insights into the evolution of sex chromosome dosage compensation, which is a quintessential example of large-scale epigenetic regulation, and an essential biological process in many species. Misregulation of sex chromosome dosage compensation can have widespread deleterious effects. For instance, failure to achieve upregulation of the single male X in *Drosophila* results in male lethality (Belote and Lucchesi 1980 a, b), and failure to inactivate an additional maternal X causes embryonic lethality in mouse (Goto and Takagi 1998, 2000). In certain human autoimmune diseases and cancers, reactivation of the X_i can produce global transcriptional effects that are believed to contribute to the disease phenotype (Agrelo and Wutz 2010 a, b).

Despite the high conservation of some form of sex chromosome dosage compensation across a wide range of species, there is extensive variation in the mechanisms by which it is achieved. In *C. elegans,* dosage compensation between XX hermaphrodites and XO males occurs by halving expression from both hermaphrodite X chromosomes (Hsu and Meyer 1994). In *Drosophila,* transcription from the single X in XY males is doubled to be equivalent to XX females (Mukherjee and Beermann 1965). In avians, down regulation of transcription from the Z chromosome in ZZ males appears to occur on a gene-by-gene basis rather than on the chromosome level (Itoh et al. 2007; Mank and Ellegren 2009). In contrast to the random XCI observed in eutherian mammals, marsupials display paternal XCI, meaning that only the paternal

X chromosome is silenced in XX females (Wang et al. 2014). The vastly different nature of these schemes indicates that they have evolved independently across different animal phyla, but it is unclear to what extent they are homologous or convergent within phyla. In eutherian mammals, this question is particularly interesting given the wide variety in the degree of X and Y chromosome divergence and PAR size across species (Raudsepp and Chowdhury 2015), which may result in dosage compensation of different X genes across species.

Another motivation to study XCI and XCI escape is to understand what roles each of these processes play in human health and disease. An excess of Mendelian disorders have been mapped to the X chromosome compared to the rest of the human genome (Hamosh et al. 2005), and it is expected that the X chromosome also plays unique roles in complex genetic diseases. This is supported by the fact that many complex diseases show a pronounced sex bias in incidence, prevalence, etiology, or symptoms (Patsopoulos et al. 2007; Ober et al. 2008), and the possibility that X is enriched for sexually antagonistic alleles (Morrow and Connallon 2013). As mentioned above, XCI escape may contribute to phenotypic differences between the sexes. This idea has driven a great deal of interest in XCI escape among human geneticists, and there is some evidence to support it for at least a few genes. For example, one recent study found that a subset of human XCI escapers show a significant male-specific enrichment for somatic loss-of-function mutations across a variety of cancers (Dunford et al. 2017). The authors suggested that these genes might have roles as tumor suppressors, and that their biallelic expression in females confers a protective effect in the event that one allele acquires a loss-of-function mutation.

XCI and XCI escape may also contribute to phenotypic differences among females. Due to the stochastic nature of random X inactivation in early development, XX individuals have different X allelic expression ratios at the tissue level based on the degree of X_i mosaicism within that tissue. This “XCI skewing” can be extreme, with some individuals showing effectively monoallelic expression within a given tissue (Busque et al. 1996). Variable XCI escape also introduces the potential for allelic expression ratio variation among females based on whether or not a given gene escapes XCI in each individual (Peeters et al. 2014).

Research questions

While research over the past fifty years has illuminated many aspects of XCI and XCI escape, several aspects of both phenomena remain poorly understood. First among these are the details of the physical mechanisms of XCI and XCI escape. Despite the many technological advances that have occurred since their discovery, both XCI and XCI escape are still difficult to observe experimentally because they occur early in development and with extensive intrinsic variation. This is especially true for human studies, which have much more limited access to embryos and a greater extent of atypical epigenetic states in cell lines (Shen et al. 2008; Fan and Tran 2011) than model organisms. The functional significance and consequences of XCI escape also remain uncertain. As mentioned above, it has been suggested that XCI escape could be essential to normal female development. However, it has also been argued that XCI escape is simply the result of inefficient XCI, with no little or functional impact (Migeon 2014). As with the physical mechanisms of XCI and XCI escape, these two models are difficult to evaluate experimentally.

Understanding the evolutionary history of XCI escape might help elucidate its mechanisms as well as its functional significance. The widely accepted model for XCI and XCI escape evolution is that XCI evolved gradually in response to loss of Y chromosome gene content (Ohno 1967; Charlesworth 1991; Skaletsky et al. 2003), and that XCI escape is an intermediate state in this process (Jegalian and Page 1998). At a glance, the relationship between Y gametolog functionality and X gametolog inactivation seems straightforward: XCI rarely occurs in the PARs (Berletch et al. 2011), which contain most of the remaining functional Y genes, and recent lineage-specific Y gametolog loss within the eutherian clade correlates with XCI of the corresponding X gametolog (Jegalian and Page 1998). However, there are many human X genes that do not neatly fit this pattern, such as X-inactivated genes that have retained functional Y gametologs, and XCI escapers located in old evolutionary strata that lost their Y gametologs many millions of years ago (Pandey, Wilson-Sayres, and Azad 2013; Wilson-Sayres and Makova 2013). These apparent discrepancies motivate the central goal of this dissertation: to gain insight into whether XCI escape is simply a "bug" in the XCI system, or if it can also be a functional "feature" maintained by direct or indirect natural selection.

If XCI escape is solely a neutral consequence of inefficient XCI, we would expect that a gene's XCI status would be determined primarily by its evolutionary relationship to the Y chromosome. Mammalian X and Y chromosome divergence is believed to have occurred in a stepwise manner as a consequence of large-scale structural rearrangements on the Y chromosome inhibiting X-Y recombination (Ross et al. 2005). Therefore, if the assumptions of the "bug" model of XCI escape hold for all genes, we would expect that X genes that recently diverged from their Y gametologs would escape XCI, while genes that diverged from their Y gametologs far in the past would be completely X-inactivated.

While it is likely that this "bug" model sufficiently explains some instances of XCI escape, it may also be true for a subset of XCI escapers that XCI escape or some corollary of escape is a biological "feature" that is a target of selection. For instance, XCI escape could result from dosage sensitivity, creating selective pressure to maintain biallelic expression of these genes in females, or to maintain high expression in males and/or females. In these cases, we might see a greater degree of purifying selection on XCI escapers and their *cis* regulatory elements than on inactivated genes. Importantly, X inactivated genes might also be evolving under natural selection, but for different features: Assuming that biallelic expression is the ancestral state for an X chromosome gene, its *cis* regulatory sequences might need to undergo a period of positive selection or relaxed purifying selection in order to develop the ability to interact with the XCI machinery and undergo consistent silencing.

If some genes indeed escape XCI as a consequence of functional constraints, they might show functional characteristics that are distinct from X-inactivated genes and independent of their physical location in the X chromosome. We can identify common functional characteristics of XCI escapers by comparing gene expression patterns, genetic and protein-protein interactions, biological process, cellular component annotations, and other features between XCI escapers and X inactivated genes. Pinpointing these features may provide valuable insight into the physical mechanisms of both XCI escape and XCI in humans and model organisms and clarify the roles of XCI escapers in human development and disease. In this dissertation, I describe my investigation of the relationship between XCI status and gene expression in humans, and comparisons of XCI profiles across four mammalian species, with the goal of elucidating some of these points.

CHAPTER 1

GENE EXPRESSION AS A POTENTIAL DRIVER OF XCI ESCAPE EVOLUTION IN HUMANS

Introduction

The human XCI profile is the most extensive of any organism to date, with over 700 genes having been assayed in at least one major study (Carrel and Willard 2005; Cotton et al. 2015; Schultz et al. 2015). Human XCI escape also holds great interest to the medical genetics community due to its potential role in genetic disease and sex differences. As such, we chose to use publicly-available human XCI data sets to evaluate evidence for the "feature" model of XCI escape evolution, building on previous analyses by other groups that provided important preliminary evidence supporting this model.

Park et al. (2010) were the first to examine this hypothesis by investigating the relative evolutionary rates of human XCI escapers compared to X-inactivated genes, using the Carrel and Willard 2005 XCI profile. By comparing protein coding DNA sequences of human X genes to the reference sequences of their one-to-one orthologs in Rhesus macaque and chimpanzee, they demonstrated that human XCI escapers show significantly lower evolutionary rate than consistently X-inactivated genes, indicating that they have experienced stronger purifying selection on average than X-inactivated genes. This pattern was most pronounced for the nine consistent XCI escapers that have retained a functional Y gametolog, and functionality of the Y gametolog was shown to be a strong predictor of an X gene's XCI status. Later work by Bellott et al. (2014) showed that the set of eighteen X genes with functional Y gametologs across mammals is extremely highly conserved within Mammalia, and shows enrichment for housekeeping functions such as transcription, translation, and protein modification. This finding raised the possibility that the signal of stronger purifying selection observed in human XCI escapers by Park et al. was primarily driven by this small group of ultra-conserved genes rather than being a general feature of XCI escapers.

In addition to their observation of stronger sequence conservation in XCI escapers, Park et al. (2010) showed that human XCI escaper gene expression levels were more conserved across primates than those of X-inactivated genes. Expression level (Drummond et al. 2005; Wall et al. 2005) and to a greater extent, expression breadth (Duret and Mouchiroud 2000; Park and Choi 2010), are known to correlate strongly with nucleotide substitution rates in mammals, with highly, broadly expressed genes evolving more slowly than weakly-expressed or tissue-specific genes. Therefore, there is a strong possibility that gene expression level and/or breadth contribute to the high sequence conservation observed among XCI escapers. Importantly, this is a distinct hypothesis from that of selection on female-biased expression of XCI escapers resulting from XCI escape, which may contribute to sex-specific phenotypes (Trabzuni et al. 2013; Deng et al. 2014).

While previous studies have provided essential insights into the evolution of XCI, none have specifically focused on the relationship between purifying selection and gene expression across all XCI escapers compared to other X-linked genes. In order to investigate the possibility that gene expression patterns contribute to the differential signals of purifying selection between XCI escapers and X-inactivated genes, we used divergence data inferred from publicly available primate genomes, human polymorphism data, and human RNA-seq data to examine the relationship between purifying selection and gene expression within XCI categories (XCI escapers vs. X-inactivated genes), controlling for Y gametolog functionality. We first separated genes with functional Y gametologs from the rest so that we could determine the extent to which the stronger signal of purifying selection in XCI escapers is dependent on the inclusion of these highly conserved genes. Then, we compared gene expression level and breadth across XCI and Y gametolog status combinations in females and males in multiple primary tissues.

Results

1. Composite human XCI profile

In order to maximize the number of genes in our analysis, we assigned consensus XCI statuses to unique protein-coding X genes compiled from three studies that used different methods to determine XCI status: Carrel and Willard et al. (2005), Cotton et al. (2015), and Schultz et al. (2015). A summary of the assays and XCI classifications used in each study is shown in File S1.1, and XCI status calls in for all genes across all three studies are available in File S1.2. Based on the observation by Cotton et al. that most genes either primarily escape XCI or are inactivated across both tissues and individuals, we limited the number of XCI status categories within each data set to two – XCI escapers and X-inactivated genes – either by excluding variable (for Carrel and Willard 2005; Cotton et al. 2015) and tissue-specific (for Schultz et al. 2015) escapers, or by splitting these genes between the XCI escaper and X-inactivated categories based on their behavior in the majority of individuals and/or tissues (details in File S1.2). We considered each data set individually, but also combined data across the three studies using a more inclusive definition of XCI escape. In this combined data set, we classified genes as XCI escapers if they showed significant evidence of escape in any number of tissues or individuals in any study. We classified all other genes as X-inactivated. This inclusive definition left us with a list of 248 X genes that have been shown to escape XCI in at least one individual in one or more tissues in any data set, and 238 that were only X-inactivated in any data set. We report results for the combined data set throughout this chapter.

2. Selection analysis

For each XCI/Y status combination, we calculated two statistics derived from a McDonald-Kreitman (McDonald and Kreitman 1991) framework: *N,* the fraction of sites evolving neutrally, and *S*, the fraction of sites evolving under strong purifying selection (Mackay et al. 2012). Weak purifying selection and population expansion can both produce excesses of rare polymorphisms (Keinan and Clark 2012; Gao and Keinan 2014), which bias the McDonald-Kreitman test. However, these statistics provide some differentiation between weak purifying selection and neutral polymorphism inflation, and all gene categories in this analysis are expected to share the same degree of neutral polymorphism inflation due to their shared history on X. We calculated the XCI escaper-to-X-inactivated ratios of *S* and *N* within each Y gametolog category by pooling polymorphism and divergence data across genes within each category, and resampling gene category membership with replacement 1000 times to account for variability in the extent of selection across genes. In cases where XCI escapers are undergoing more purifying selection than X-inactivated genes, these ratios are expected to be >1 for *S* and <1 for *N*. The three Y gametolog categories we considered were "functional Y" (the X gene has retained a functional, but not identical Y gametolog), "pseudogenized Y" (the X gene has a Y gametolog that is non-functional but detectable by sequence homology), and "lost Y" (no Y gametolog can be detected by sequence homology). There are a total of 15 genes in the functional Y gametolog category, 11 of which escape XCI and only 4 of which undergo X-inactivation, which severely limits the power of any statistical tests used to detect differences between them. However, the direction of the relationship between XCI escapers and X-inactivated genes in this group is still somewhat informative, and in light of their high biological significance (Bellott et al. 2014) we report results for this class throughout this chapter.

Using XCI status assignments from the combined data set, the *S* ratio was significantly greater than one and the *N* ratio was significantly less than one in all but the functional Y gametolog category (Figure 1.1; Table 1.1). To ensure that these results were not dependent on the inclusion of variable and/or tissue-specific XCI escapers, we calculated the same ratios either excluding or including variable and tissue-specific XCI escapers in each individual XCI status data set and in the combined data set. We observed that the *S* and *N* ratios were significantly greater and less than one (respectively) in at least one of the pseudogenized or lost Y gametolog categories in each XCI status data set. The *S* and *N* ratios for each XCI/Y category did not change significantly within each data set when variable and tissue-specific escapers were excluded (File S1.3). We also observed that *S* and *N* ratios were rarely significantly different between the categories of genes with pseudogenized and lost Y gametologs. Therefore, for the remainder of our analyses, we combined the pseudogenized and lost Y gametolog categories within each XCI category into a single “no Y” gametolog category. Combined, the results presented in this section show that the signal of stronger purifying selection in XCI escapers compared to X-inactivated genes 1) is evident across the different methods and definitions used to define XCI escape in each of the previous studies, 2) is not solely driven by XCI escapers with functional Y gametologs, and 3) is generally robust to the exclusion of variable or tissue-specific XCI escapers.

3. Contrasting gene expression patterns of XCI escapers and X-inactivated genes

Next, we sought to identify biological features that might drive constraint on XCI escaper evolution. Additionally, we wanted to evaluate the extent to which the trend of stronger expression level conservation among XCI escapers observed by Park et al. (2010) held when we included variable XCI escapers and genes with XCI status determined by methylation, which might include genes that were too weakly expressed to be assayed by allele-specific expression. To accomplish this, we used gene expression data from primary tissues across many individuals to compare expression level and breadth between XCI escapers and X-inactivated genes as classified in our combined XCI profile. We examined gene expression in both males and females because the gain in overall expression level produced by XCI escape is not necessarily very large. For instance, the Carrel and Willard (2005) data showed an average X_i to X_a expression level ratio among XCI escapers of approximately thirty percent. Furthermore, few human XCI escapers have been found to exhibit statistically significantly female-biased expression (Talebizadeh et al. 2013; Tukiainen et al. 2017). We therefore did not want to take for granted that XCI escapers would show expression differences between the sexes.

For our gene expression analysis, we used the publicly available NIH Genotype Tissue Expression (GTEx) data (v.4) (GTEx Consortium. 2015). This data set included RNA-seq data from multiple individuals of both sexes in twenty-three primary tissues taken from US tissue banks. We used data that had been converted to reads per kilobase per million (RPKM), which normalizes read count to transcript length to reduce bias towards long molecules (Mortazavi et al. 2008). To quantify gene expression breadth, we used the tissue specificity index τ (Yanai et al. 2005), defined in the methods section of this chapter. Highly tissue-specific genes will thus show high values of τ, whereas broadly expressed genes show low τ values.

In the early stages of our gene expression analysis, we observed that many X genes with non-zero expression values do not have any expression values over standard heuristic cutoffs, such as 1 RPKM. Given that our composite XCI profile included some XCI statuses determined based solely on methylation data, we chose to include weak expression values to ensure that these genes were not disregarded. To evaluate the effect of including weakly expressed X genes on our results, we ran each analysis using several different gene expression level cutoffs. First, we tried three commonly used fixed cutoff values: 1, 0.5, and 0.1 RPKM. Next, we adapted the method of Hart et al. (2013) to set expression level cutoffs that took into account the expression profile of each tissue. We set the minimum expression value for each tissue based on its distribution of X gene RPKM values across all individuals and used all expression values within some percentile of this distribution (one, two, and three standard deviations from the mean). We found that the general result of higher mean expression across XCI escapers compared to X-inactivated genes was robust to the choice of minimum expression cutoff (Table 1.2), but more pronounced and significant with lower minima. We show expression results using the three standard deviation (3SD) expression cutoff in the remainder of this document.

Using this definition of significant gene expression, we compared the expression of XCI escapers and X-inactivated genes across 23 broad tissue types in females and males separately using RNA-seq data from the GTEx Consortium (2015). We first obtained a single expression level value for each gene in each tissue by averaging significant expression values across individuals (see Methods) and observed that XCI escapers were generally more highly expressed than X-inactivated genes in individual tissues in both sexes (Figure 1.2.a). We then obtained a global expression value for each gene by taking the average of its expression values across tissues (Figure 1.2.b). To assess the significance of the differences between XCI escapers and X-inactivated genes at a global level, we calculated the average global expression value across genes in each XCI/Y category and calculated the XCI escaper-to-X-inactivated ratio of these values in each Y gametolog category. We permuted gene membership between the XCI groups in each Y gametolog category to obtain a p-value for this ratio. Within the no Y gametolog category, the XCI escaper global expression mean was significantly higher than the X-inactivated global expression mean in both sexes. The XCI escaper-to-X-inactivated global expression mean ratio was 1.97 ($p << 10^{-3}$) in females and 2.06 in males ($p << 10^{-3}$). In the functional Y gametolog category, there was no significant difference between the global expression mean of significantly expressed X-inactivated genes and XCI escapers: the XCI escaper/X-inactivated global expression mean ratio was 0.867 ($p = 0.529$) in females and 1.204 in males ($p = 0.434$). Data for this analysis is available in File S1.4

Comparing the global expression ratios of XCI/Y categories across various minimum expression level cutoffs showed that the XCI escaper-to-X-inactivated global expression ratio was sensitive to the minimum expression cutoff. The purpose of using these filters was to avoid using RPKM values that are indistinguishable from background signal, but this results in a substantial loss of information in some XCI/Y categories (this is readily apparent in Figure 1.4, which is described in the expression breadth section of the Results). Therefore, to examine the impact of these weak expression values on global expression trends without making assumptions about what constitutes significant expression, we performed a non-parametric rank-based comparison of XCI/Y category expression.

For each XCI/Y category, we obtained a single expression distribution by averaging unfiltered RPKM values (including zeros) across individuals for each gene in each tissue, resulting in a distribution of length *n* x 23, where *n* is the number of genes in the XCI/Y category and 23 is the number of tissues. We then performed the Mann-Whitney U test to contrast the XCI escaper and X-inactivated RPKM distributions in each Y gametolog category. In females, the global Mann-Whitney U test p-values were 0.010 for XCI escapers > X-inactivated in the small functional Y gametolog category, and $2.2\text{x}10^{-16}$ for XCI escapers > X-inactivated in the much larger no Y gametolog category. In males, the XCI escapers > X-inactivated p-values were 0.047 and $2.2\text{x}10^{-16}$ for the functional and no Y gametolog categories. This test highlights the fact that X-inactivated genes with functional Y gametologs are a unique group showing extremely weak expression in many tissues but unusually high expression in others, whereas the XCI escapers with functional Y gametologs are more uniformly highly expressed. These results further support our overall conclusion that XCI escapers tend to be more highly expressed than X-inactivated genes.

To compare expression breadth between XCI categories, we calculated the tissue specificity index τ (Yanai et al. 2005; Methods) for each gene and compared its distribution between XCI categories. XCI escapers showed significantly lower values τ of than X-inactivated genes in both sexes and in both Y gametolog categories (Figure 1.3; Table 1.3), indicating that they are more broadly expressed than X-inactivated genes. As an additional expression breadth measurement, we calculated the proportion of tissues for which each gene showed significant expression (Methods) and observed that XCI escapers were significantly expressed in a higher proportion of tissues than X-inactivated genes with the same Y gametolog status in both sexes (Figure 1.4, File S1.6). Consistent with previous work demonstrating that XCI escapers with functional Y gametologs are enriched for highly conserved functions (Bellott et al. 2014), these genes showed the lowest τ values in both sexes across all XCI/Y categories.

Conclusions

In this work, we have collected and analyzed data from three of the most recent and comprehensive studies with information on human XCI escape. The results of our selection analysis are consistent with the hypothesis that XCI escape or some corollary of it is under evolutionary constraint. However, they do not constitute direct evidence for the “feature” model of XCI escape being a general rule. Our gene expression analysis results suggest that expression level and breadth might contribute to this signal of purifying selection, but they do not clarify the nature of the relationship between XCI status and expression. We can broadly characterize this relationship according to two distinct, but not mutually exclusive models.

First, XCI escape itself might be the target of purifying selection. One possibility is that the female expression bias produced by expression of the X_i allele in females is functionally important in normal development. Perhaps the strongest support for this model comes from studies of X aneuploidy syndromes such as Turner syndrome (45, X) and Kleinfelter syndrome (47, XXY). For instance, reduced expression of several human XCI escapers in individuals with Turner syndrome has been implicated in phenotypes associated with this disorder, including ovarian failure, heart defects, and various neurological abnormalities (Ellison et al. 1997; Bione et al. 1998; Zinn and Ross 1998). Another study identified several XCI escapers that were inferred to be highly dosage-sensitive to be promising candidates for various X aneuploidy syndromes (Pessia et al. 2012). However, the rarity of significant female expression biases introduced by XCI escape makes this model a poor fit for the majority of XCI escapers. For instance, in one study, only fifteen ChrX genes surveyed showed significant female expression biases in at least one of eleven human tissues, and only six of these escape XCI (Talebizadeh et al. 2006). In a recent study that also used the GTEx expression data and a more sensitive definition of sex biased expression, sixty XCI escapers exhibited female-biased expression and several others showed male-biased expression in some tissues (Tukiainen et al 2017).

Alternatively, XCI escape may be an indirect consequence, rather than a driver, of gene expression patterns via any of a number of molecular mechanisms, such as local perturbations in chromatin state. For instance, inefficient silencing of strong promoters or other *cis* regulatory elements prior to or during XCI could prevent these genes from undergoing complete silencing (Brown and Greally 2003; Migeon 2014). It is now well-established that XCI escapers co-localize with each other outside of the Barr body in three-dimensional space, producing topologically associated domains that are clearly distinct from the rest of the chromosome in both human and mouse (Lieberman-Aiden et al. 2009; Splinter et al. 2011; Engreitz et al. 2013; Deng et al. 2015; Giorgetti et al. 2016). There is also some evidence in mouse that XCI escapers are more highly expressed than X-inactivated genes early in development (Marks et al. 2015), but it is unknown if this also occurs in humans. However, the observation that human genes escaping XCI in females are more highly expressed than X-inactivated genes in adult tissues from both males and females is consistent with this model, as *cis* regulatory elements would be shared between the sexes. Similarly, Cotton et al. (2015) observed that monozygotic twins are much more likely to show the same XCI status at variable escape genes than is expected by chance.

Importantly, there are several limitations in our study. First, only approximately two thirds of the human X chromosome genes had an XCI status assignment in one or more of the three data sets used in this study. As such, the observations we have made about selection patterns and gene expression must be continuously revaluated as the human XCI profile is extended. Second, we cannot definitively say that the XCI statuses we have assigned to these genes generalize to the broader population and/or other primary tissues. The true extent of variation in XCI profiles is still unknown and will require expanding the number of tissues and individuals in studies of XCI status. In particular, our results highlight the variability and uncertainty in the definition of variable XCI escapers, which were often annotated as variable escapers in one data set and as consistently escaping or X-inactivated in one of the others (Table S2). We suggest that it should be a priority to determine whether variable escape reflects true polymorphism for underlying mechanistic differences among individuals in escape status, or whether it is technical noise that arises from catching the tail of a distribution of allele-specific expression or methylation after studying sufficiently many samples across sufficiently many tissues.

Finally, it is important to note that our result of higher and broader gene expression among XCI escapers compared to X-inactivated genes (but not the purifying selection results; Table 3) is dependent on the inclusion of many weakly-expressed genes that have typically been excluded from XCI and gene expression studies (as described in Results section 3), and the inclusion of the genes with XCI statuses derived from the Schultz et al. 2015 study: if these are excluded from the analysis, the mean expression level and breadth of the X-inactivated genes is not significantly different from the XCI escapers. Schultz et al. determined the criteria for calling a gene X-inactivated or escaping XCI from sex-biased mCH levels based in part on the ability of these criteria to recapitulate the XCI status assignments generated by ASE in Carrel and Willard 2005 for genes assayed in both studies. While we have no reason to distrust the Schultz et al. XCI assignments for this reason, it would be prudent to cross validate them in additional samples and perhaps using other methods. Because is likely that many of these genes could not be assayed by ASE due to either weak gene expression or lack of informative SNPs, these other methods might include other types of epigenetic marks, or chromatin conformation capture.

Overall, we concur with many previous studies that human XCI escapers are promising candidates for investigation as contributing to human genetic disorders that exhibit significant sex differences in incidence, severity, or symptoms, as well as those that show high phenotypic variability among females. In particular, large-scale disease studies are likely to benefit from incorporating information about a gene's XCI status in association analyses, as in Tukiainen et al. (2014), Chang et al. (2014), and Ma et al. (2015).

Methods

1. Combining XCI profiles across studies

We obtained XCI statuses for 758 unique protein-coding human X loci from three data sets: Carrel and Willard (2005), Cotton et al. (2015), and Schultz et al. (2015) (File S1.2). These XCI statuses were determined based on allele-specific expression from the inactivated X (Carrel and Willard 2005), female versus male transcription start site CG dinucleotide (mCG) level (Cotton et al. 2015), and female versus male gene body non-CG methylation (mCH) level (Schultz et al. 2015). Brief descriptions of the assays and categorizations used in each study are shown in File S1.1.

To ascertain the effects of the different methods of determining XCI status and the impact of variable and tissue-specific XCI escapers on the overall patterns of purifying selection across XCI categories, we repeated the analysis of strong purifying selection (described below) in each individual data set. For the two data sets with information about heterogeneous/variable escape, we performed these calculations either excluding variable escapers or splitting them into the consistent escape and consistently X-inactivated categories. For the Schultz et al. data set, we calculated these statistics using the set of genes that escaped in any of eleven tissues as XCI escapers, and using only genes that escaped in six or more tissues as escapers (treating those that escaped in five or fewer tissues as X-inactivated). We observed that excluding variable or tissue-specific escapers did not result in a significant change in the results within the individual data sets, nor did it have a significant effect on the "combined" data set defined by aggregating XCI statuses across all three studies using an inclusive definition of XCI escape: all genes showing escape in any number of tissues and/or individuals in any study were considered XCI escapers, and all others were considered X-inactivated (File S1.2). This definition of XCI is expressly somewhat more inclusive, and meant as an appropriate contrast, being less vulnerable to false negatives in the identification of XCI escape associated to the nuances of the particular methods used in each of the individual studies.

2. Polymorphism data

For our polymorphism analyses, we took advantage of a large, high-depth whole-exome sequence data set from the NHLBI GO Exome Sequencing Project (ESP), which includes a large amount of rare (minor allele frequency<0.5%) single nucleotide variants (SNVs) (Fu et al. 2013). We included only biallelic variants in the European American subsample with an average sequencing depth >20X, and that passed the original (Tennessen et al. 2012) quality control filters. The number of copies *n* available in the ESP data set varies across SNVs, with a range of 2379-6728, a mean of 6558, and a standard error of the mean of 2.51 across non-PAR X sites. To make all sites comparable, we down-sampled derived allele counts at each site to a total of 6056 (90% of the maximum *n* value), according to the method described in Marth et al. (2004), and excluded variants for which data were available for fewer than 6056 copies.

3. Divergence data

To facilitate expression analyses (described below), we considered only the transcript with the greatest total exonic length in the RefSeq database for each unique protein-coding gene, yielding 748 non-overlapping X transcripts. Of the 1,048,661 unique ESP SNVs across the entire human exome, 24,795 were associated with these 748 transcripts, with the majority falling in non-exonic regions. Polymorphic, divergent and conserved monomorphic sites were identified as either synonymous (Syn) or non-synonymous (NonSyn) based on the genetic code. Counts of divergent and polymorphic sites were calculated for each transcript. Divergent sites were those for which the human reference (hg19) allele differed from the ancestral allele, inferred by the reference chimpanzee (panTro4) allele and confirmed by one or both of the reference orangutan (ponAbe2) and macaque (rheMac3) alleles. Sites for which the ancestral state could not be confirmed in this manner were excluded.

4. Quantifying ancient/strong purifying selection via polymorphism and divergence

Transcripts were filtered to remove overlap with annotated segmental duplications, repetitive elements or ampliconic regions, as well as potential CpG dinucleotides (in either the ancestral or derived state) and sites with uncertain ancestral states resulting from poor synteny as defined by the UCSC syntenic net tracks, or ambiguous ancestral state (as defined above). After filtering, we were left with 525 X transcripts that retained at least one exon with usable data, existed in a single-copy, and were inferred to be present on the ancestral mammalian X. These 525 loci were the final set of genes we considered in our analysis. Across these genes, there were 3,581 NonSyn polymorphic sites, 2,279 Syn polymorphic sites, 392 NonSyn divergent sites, 558 Syn divergent sites, 1,431,973 conserved NonSyn sites, and 507,404 conserved Syn sites. All site counts for each gene are available in File S1.7.

The McDonald-Kreitman (MK) test (McDonald and Kreitman 1991) uses the ratios of selected and neutral sites among intra-species polymorphisms compared to inter-species divergences to detect departures from neutral evolution. We used non-synonymous (NonSyn) and synonymous (Syn) sites of protein-coding genes as our selected and neutral site classes, respectively, and calculated three statistics based on the MK test across XCI/Y categories. In these statistics, $\boxed{m_S}$ and $\boxed{m_N}$ are the Syn and NonSyn conserved (monomorphic) site counts, respectively. $\boxed{P_{N\ neut}}$ and $\boxed{P_{N\ weak}}$are the neutral and weakly deleterious fraction of polymorphic sites out of $\boxed{P_N}$ NonSyn polymorphisms overall. These are estimated by partitioning the NonSyn and Syn polymorphic site counts $\boxed{P_N}$ and $\boxed{P_S}$ into arbitrarily defined bins of high and low derived allele frequency, and using the neutral low:high count ratio to calculate the expected fractions of the rarest selected polymorphisms that are weakly deleterious and neutral. Throughout this study, we report values calculated using a derived allele frequency cutoff of >1% to classify SNVs as high frequency, which corresponded to a down-sampled derived allele count of 61. For a more detailed description of the calculation of these statistics, refer to the supplementary material of Mackay et al. (2012). The statistics we calculated for each XCI/Y group were: $\boxed{N = (m_S P_{N\ neut})/(m_N P_S)}$, the fraction of sites that are neutral (called *f* in Mackay et al. 2012); $\boxed{W = (m_S P_{N\ weak})/(m_N P_S)}$, the fraction of sites that are evolving under weak purifying selection (called *b* in Mackay et al. 2012); and $\boxed{S = 1 - (N + W)}$, the fraction of sites evolving under strong purifying selection (called *d* in Mackay et al. 2012). The fraction of sites that are weakly deleterious, *W,* was the only statistic for which the relationships between XCI/Y categories changed across cutoff DAFs (File S1.8). However, because *W* was an order of magnitude smaller than either *N* or *S* in all categories, we concluded that these fluctuations are most likely due to chance, and we did not further analyze this statistic.

5. Y gametolog status data

Y gametolog statuses for 723 unique protein-coding X genes were obtained from Wilson-Sayres and Makova (2013), who identified Y gametologs based on interspecies comparisons and human X/Y homology. The Y gametolog categories described in this study were defined as follows: 594 genes that were inferred to be ancestral in the eutherian lineage (conserved across 4 species among the set including mouse, rat, rabbit, cow, horse, dog, opossum and chicken) were classified as "Functional Y" (*n* = 19), "Pseudogenized Y" (*n* = 265), and "Lost Y" (*n* = 312). Finally, we removed three genes located in the human-specific X-transposed region from consideration, of which two had functional Y gametologs and one had a pseudogenized Y gametolog. For some analyses, based on our initial results, we combined the two categories of "Pseudogenized Y" and "Lost Y" into a single category, "no Y" corresponding to no functional Y gametolog (see Results).

6. X gene copy number data

Human X gene copy classifications were obtained from Mueller et al. (2013). In this data set, genes were annotated as single-copy, X multicopy or ampliconic, where X multicopy genes had at least one *cis* paralog but were not in ampliconic regions, and ampliconic genes were those in or near ampliconic regions. Because ampliconic and multicopy genes are poorly conserved across the great apes, we only considered in this study genes with a single copy on the human X.

7. Gene expression data filtering

To assess differences in human gene expression across XCI/Y categories, we used the publicly available NIH Genotype-Tissue Expression (GTEx) RNA-seq data set (GTEx Consortium 2015), which includes expression data in 51 primary tissue subtypes across 30 broad tissue types from American women (78) and men (138). Many genes in certain XCI/Y categories are very weakly expressed, and retain no usable data if we use typical arbitrary minimum expression values (e.g. 1 RPKM) to remove potentially spurious measurements from the RNA-seq data. However, these weakly expressed genes are not typically statistical outliers in the expression distributions of their gene categories. We therefore elected to use an adaptive method for filtering background expression values in an unbiased manner while including low expression values that were not outliers. For each broad tissue type, we fit a half Gaussian distribution to the higher mode of the distribution of non-zero $\log_2$(RPKM) values across all individuals for all X genes, and mirrored it to capture the expression distribution of actively transcribed genes (Hart et al. 2013). We then used three standard deviations below the means of these distributions as a minimum expression value to consider a gene significantly expressed in each tissue, while excluding the other genes from analysis. The sex- and tissue-specific expression cutoffs defined in this way are available in File S1.9. After individual expression values were filtered according to these minima, the remaining values were averaged across 1000 bootstrap samples of individuals for each gene in each tissue. For the sake of clarity, in this section we refer to each gene’s average expression across individuals in each tissue as its expression value for that tissue.

8. Gene expression level analysis

For GTEx tissues with multiple subtypes, we considered only the subtype with the largest sample size, yielding one expression value for each broad tissue type for each gene. To ensure that all genes had data for equal numbers of tissues in females and males, we excluded sex-specific tissues (cervix, fallopian tube, ovary, uterus, vagina, prostate, and testis) from our expression analyses, leaving 23 broad tissue types in both sexes. The mean expression values across all tissues for each gene are available in File S1.4 and displayed in Figure 1.2.b. For each tissue, we also calculated the mean and standard deviation of global expression values across each XCI/Y category. These data are shown in Figure 1.2.a. and are available in File S1.4.

To assess the significance of global expression ratios between XCI escapers and X-inactivated genes in each Y gametolog category, we performed 1000 permutations of gene membership in pairs of XCI/Y groups and calculated the difference in the mean expression of these genes across all tissues. P-values reflect the proportion of permutations for which the ratio of average expression values between categories, averaged across all tissues, was greater than or equal to the observed. Repeating this analysis using various more and less lenient fixed and adaptive expression minima showed that the adaptive three standard deviation threshold produced results that were closest to those generated using a fixed threshold of 0.1 RPKM (Figure S4), which is the same threshold used by the GTEx Consortium in their own expression analyses (GTEx Consortium 2015) and similar to the threshold value calculated by the analysis of Hart et al. (2013).

To calculate two-sided Mann-Whitney U test statistics and p-values for the unfiltered expression distributions in each Y gametolog category in each sex, we first calculated the average RPKM values for each gene in each tissue across individuals. These values were pooled across genes and tissues in each XCI/Y category, and the test then applied for each pair of these pooled data sets using the wilcox.test function for unpaired data in R.

9. Gene expression breadth analysis

We calculated the tissue-specificity index τ (Yanai et al. 2005) for each gene to quantify the breadth of expression across each category (data available in File S1.5). This statistic is defined for each gene as:

$$\boxed{\tau = \frac{\sum_{i=1}^{N}(1 - x_i)}{N - 1}}$$

where *N* is the number of tissues and x_i is the expression value of this gene in tissue i, normalized by its maximum expression value across all N tissues. High values of this statistic (e.g. >0.6) indicate that the gene's expression is skewed towards a small number of tissues, and low values indicate that it is expressed at similar levels across many tissues. Based on the expression data described in the previous subsection, the genes in each XCI/Y category were resampled with replacement 1000 times, and the mean and standard deviation of the τ values across these resamples were obtained (Table 1.2).

Lastly, for each gene, we calculated the proportion of 23 broad tissue types for which it showed significant expression (greater than the minimum expression value for each tissue, defined as in the previous section) as an additional metric of expression breadth (Figure 1.4). The gene membership of each XCI/Y category was resampled with replacement 1000 times and used to compute the mean and standard deviation (Table 1.3).

CHAPTER 2

PRELIMINARY CANINE XCI PROFILE FROM SINGLE-CELL RNA-SEQ

Introduction

Despite high sequence conservation of X chromosome genes among mammals, XCI profiles differ substantially across the three species for which extensive profiles have been ascertained (Wang et al. 2014; Cotton et al. 2015; Berletch et al. 2016; Supplementary Materials). This suggests that XCI profiles may have evolved in a largely lineage-specific manner within the mammalian clade. Given the history of the therian sex chromosomes, this is likely: These chromosomes appear to have originated from an autosomal pair in the therian most recent common ancestor and show strong sequence homology to the avian chromosomes 1 and 4. Divergence between the therian X and Y chromosome was likely driven by the translocation of the testis-determining gene *SRY* from an ancestral sex chromosome to the proto-Y chromosome, followed by a breakdown of recombination between X and Y as the Y accumulated large-scale structural rearrangements (Charlesworth 1991). This eventually resulted in the pseudogenization or loss of many Y chromosome genes, and XCI of the corresponding X genes is believed to have evolved in a stepwise-manner to compensate for the resulting dosage differences between males and females (Jegalian and Page 1998). This model is supported by the high correlation between the functionality of Y genes and the XCI status of their X gametologs within species (Jegalian and Page 1998; Wilson-Sayres and Makova 2013).

Given that the extent of X and Y divergence differs widely across mammals, (Raudsepp and Chowdhury 2015), it follows that XCI profiles would also differ between species. However, if some selective pressure drives certain genes to escape XCI or be inactivated independent of their Y gametolog status, these genes might share an inactivation status across species despite having different evolutionary histories. Therefore, cross-species comparisons of XCI profiles can improve our understanding of the evolutionary drivers and consequences of XCI and XCI escape within and across mammals.

An analysis of this type would ideally include complete XCI profiles from a wide range of mammalian species, from the most basal to the most derived. While individual genes have been assayed by immunofluorescence (Chaumeil et al. 2011; Al Nadaf et al. 2012), quantitative PCR (Yen et al. 2007), or methylation patterns (Bell et al. 2008) in several species, large-scale XCI profiles are available for only three species at present: human (Carrel and Willard 2005; Cotton et al. 2015; Schultz et al. 2015), mouse (*M. musculus*) (Yang et al. 2010; Berletch et al. 2015), and opossum (*M. domestica)* (Wang et al. 2014). The difficulty in ascertaining such XCI profiles arises from female tissue mosaicism with regards to the active X homolog. Specifically, most tissues contain a mixture of cells expressing the maternal or paternal X homolog, complicating allele-specific expression (ASE) analysis from bulk tissue samples. In light of this, past studies of XCI have typically been confined to artificial hybrid cell systems (Carrel and Willard 2005) or rare tissue samples with highly skewed XCI patterns (Cotton et al. 2013), which is uncommon in karyotypically normal adult tissues (Amos-Landgraf et al. 2006).

The development of high-throughput single-cell RNA-sequencing (scRNA-seq) technology presents a novel opportunity to acquire cell-specific XCI status calls from primary cell populations, as recently implemented in human peripheral blood mononuclear cells (PBMCs) (Tukiainen et al. 2017). Several scRNA-seq platforms are currently available, but all use a general strategy of capturing and lysing individual cells, reverse transcription of their mRNA, cDNA amplification, library construction, and sequencing (Liu and Trapnell 2016). More recently, some systems have incorporated unique molecular identifier (UMI) sequences, which barcode individual RNA molecules during the reverse transcription step. This enables direct transcript quantification, which produces precise ASE information from small amounts of starting material.

However, even with scRNA-seq, the number of genes that can be profiled in an ASE experiment is limited by the degree of heterozygosity of the cell sample donors at loci of interest. In non-human subjects, this limitation can be addressed by using species- or breed-hybrid individuals, which are more likely to carry distinct maternal and paternal alleles for any given locus. For instance, Wang et al. (2012) used horse-donkey hybrids (mules and hinnies) to facilitate ASE analysis to assess imprinting status. The domestic dog (*C. lupus familiaris*) presents an excellent opportunity to apply this strategy to XCI profiling, as intensive breeding over the past two centuries has produced extensive between-breed population structure (Vonholdt et al. 2010) and within-breed homozygosity (Boyko et al. 2010; Shannon et al. 2015). With this in mind, we sampled two individuals of one of the more common breed hybrids, the "goldendoodle" (a golden retriever-poodle hybrid), and performed X chromosome-wide XCI profiling using scRNA-seq. A carnivore XCI profile is particularly useful for studying the evolution of XCI escape because the carnivore X and Y are less diverged than in human and mouse, with a 6.6Mb PAR, and compared to the rodent X, the canine X has experienced fewer large-scale structural rearrangements relative to the human X (Federici et al. 2015). Therefore, comparing the human and canine XCI profile may produce novel insights into the roles of divergence from Y versus gene identity in determining XCI status. To this end, we use this XCI profile in a comparison of XCI profiles a comparative study of XCI profile evolution across mammals in the next chapter of this dissertation.

Results

1. Genotyping

Two adult female F1 goldendoodles, hereafter referred to as Dog1 and Dog2, were genotyped using the Embark Veterinary custom Illumina CanineHD array. The two dogs represented both cross directions: Dog1's father was a golden retriever and its mother was a standard poodle, while Dog2's father was a standard poodle and its mother was a golden retriever (pedigree shown in Figure 2.1). The genotypes of the dogs were concordant with the parental breeds and cross generation specified by their pedigrees. The array includes 6,030 markers on X (chromosome 39). 1,388 and 1,693 of these sites were heterozygous and therefore informative for ASE analysis in Dog1 and Dog2, respectively, giving a total of 2,404 unique informative X sites.

2. Single-cell RNA sequencing

Single-cell RNA sequencing (scRNA-seq) was performed on peripheral blood mononuclear cells (PBMCs) from each dog using the 10X Genomics Chromium platform to generate bar-coded single-cell Illumina RNA sequencing libraries, which include both cell-specific barcodes and UMIs. Barcoded cDNAs were then sequenced with the Illumina Next-Seq 500 instrument. Raw sequencing output was pre-processed to demultiplex reads by cell and transcript, and to remove adapter sequences. Trimmed reads were then mapped to the canine reference genome (canFam3.1), and PCR duplicates were removed. Following these steps, the Dog1 and Dog2 samples produced, respectively: 1,234 and 991 cells, with 213,010,866 and 209,833,555 total reads, and 172,618 and 211,739 mean reads per cell.

The program FASTQC (Andrews et al. 2010) was run on the raw aggregate alignment file for each sample and used to calculate per-base sequence quality, per sequence quality scores, and duplication levels. Based on these results, we removed non-primary alignment reads, which reduced estimated duplication levels to approximately 25 and 29% for Dog1 and Dog2, respectively.

As an additional assessment of the quality of our scRNA-seq data, we identified PBMC subpopulations using the cell clustering workflow developed by Satija et al. (2015), which groups cells by their expression profiles using a K-nearest neighbors strategy and identifies the most highly differentially-expressed genes in each group. As expected, the observed cell clusters derived from gene counts corresponded to known PBMC subtypes, including monocytes, T cells, B cells, and granulocytes. All but one population in the Dog2 sample could be identified based its top ten cluster marker genes. Cell populations in both samples are shown in Figure 2.2, and the workflow steps and outputs can be found in File S2.1.

To confirm that it was possible to effectively differentiate between mono- and biallelic expression at the cell level using our scRNA-seq read data, we used the Integrative Genome Viewer (IGV) software package (Robinson et al. 2011) to examine raw read data at genes with XCI statuses that we are reasonably certain of that also showed evidence of containing heterozygous sites in our data. We show images of raw read data in single cells for two examples genes in Figure 2.3. The first, *Xist* (ENSCAFT00000048497 and ENSCAFT00000045197 in canFam3.1), is expected to show completely monoallelic expression, as it is the non-coding RNA that mediates XCI through exclusive expression from the X_i. The second, *ARSE* (ENSCAFG00000025078), is located well within the canine PAR at ~1.5Mbp and is therefore expected to show expression from both the X_a and X_i. Although the raw reads at these genes show some error, they are consistent with these expectations.

3. XCI profile

To determine the XCI status of canine genes from our scRNA-seq data, we first identified heterozygous sites in each dog from their SNP array genotypes and filtered scRNA-seq data. After identifying these sites, we called each site in each cell as showing either bi- or monoallelic expression within that cell. These calls were then combined across sites within the same gene to get a gene-level expression pattern call in each cell. Genes that showed biallelic expression in two or more cells within a single dog were called as XCI escapers in that dog, while the remaining genes that showed monoallelic expression in at least one cell were called as X-inactivated in that dog. A full description of the methods used to make these calls is available in the Methods section of this chapter, and the workflow steps and outputs are shown in File S2.2.

In total, we obtained an XCI status call for 152 genes across the two dogs (available in File S2.5). Of these, 143 were in the X-specific region of the chromosome, and 9 were in the PAR. Of the non-PAR genes, 21 were observed to escape XCI in both dogs, 9 showed variable XCI escape (i.e. escaped in one dog and were inactivated in the other), 15 escaped XCI in one dog and were uninformative in the other, 24 were X-inactivated in both dogs, and 74 were inactivated in one dog and were uninformative in the other. In total, 45/143 (31%) non-PAR genes tested were observed to escape XCI in at least one dog. The gene-level minor allele frequency for each cell and the consensus XCI status for each informative gene is shown in Figure 2.5.

Lastly, we compared our XCI status calls to previously reported canine XCI statuses for two informative non-PAR X genes: *COL4A5* (ENSCAFG00000018020), which was reported as X-inactivated by Bell et al. 2008, and *CHM* (ENSCAFG00000017406), which was reported as escaping XCI by Chase et al. 2005. Our results showed that *COL4A5* escaped XCI in both dogs and that *CHM* was inactivated in one dog and uninformative in the other. This discordance may be explained by fixed genetic differences between breeds (the breeds used in Bell et al. 2008 and Chase et al. 2005 were Navasota dog and Portuguese water dog). Additionally, it is possible that the different results in our study are a consequence of using a different tissue type than either of the previous studies. Tissue-specific XCI escape has been observed in both human (Cotton et al. 2015) and mouse (Berletch et al. 2015), and certain naïve lymphocytes have been suggested to be particularly prone to partial reactivation of some X-linked immune genes in mouse and human (Wang et al. 2016), causing higher rates of XCI escape than is observed in other cell types. However, a recent study of human XCI using scRNA-seq in whole blood samples found high concordance between their XCI profile and those derived from other tissue types (Tukiainen et al. 2017). Moreover, neither *CHM* nor *COL4A5* are known to have immune functions in either humans or dogs.

Conclusions

In this chapter, we describe a preliminary XCI profile for the domestic dog derived from scRNA-seq of PBMCs from two F1 crossbreed dogs. It consists of 143 non-PAR X chromosome genes, with 45 escaping XCI in one or both dogs. This represents the first ever carnivore XCI profile, and the second XCI profile generated from scRNA-seq. It is comparable in size to the human XCI profile generated from scRNA-seq by Tukiainen et al. (2017), which included a total of 165 genes across 4 samples.

While this analysis represents an important first step in investigating XCI in dogs, its methodology and small sample size limits the extent to which our findings can be generalized. The low coverage inherent to scRNA-seq data precluded analysis of weakly- but potentially significantly- expressed genes, which might be testable by epigenetic marks as in Wang et al. 2014, Cotton et al. 2015, and Shultz et al. 2015. Moreover, sampling bias caused by low mRNA capture efficiency in scRNA-seq severely restricted our ability to detect biallelic expression even in cases where the expected allelic ratio was approximately 50:50. For example, of the 9 PAR genes with XCI status calls, only 3 were called as XCI escapers. Consequently, while we are confident that the genes we have called as XCI escapers genuinely are expressed from both X chromosomes in these samples, there is less evidence to suggest that genes called as X-inactivated are actually consistently inactivated. Further, without genome sequence data from these individuals, we were forced to conservatively identify ASE-informative sites and consequently may have missed some heterozygous sites. Finally, lacking information about the parental genotypes to phase the X chromosomes in our samples, we likely missed some cases of apparent monoallelic expression of the X_i allele caused by sampling bias, which would be consistent with XCI escape. In light of these limitations, it would be prudent to independently validate the XCI statuses identified in this study with a more sensitive method of assessing ASE, such as allele-specific pyrosequencing.

With samples of a single cell type in only two individuals, we did not have the ability to assess the extent of variation in XCI escape across individuals that has been observed in humans (Cotton et al. 2015; Shultz et al. 2016). We suggest that future studies include more individuals, possibly from crosses of more deeply diverged breeds to increase the number of testable genes. Analyzing additional cell and tissues types within individuals would also be informative. In particular, it would be interesting to examine expression of X genes in placental tissues to ascertain whether or not dogs show paternal XCI in these tissues as in marsupials and several placental mammals (Wang et al. 2014).

Of great interest to our group is how similar the dog XCI profile is to other mammalian XCI profiles. Our dog XCI profile shows more XCI escapers (45) than mouse (38) and opossum (24), but less than a third of the number of human XCI escapers (over 200, depending on the exact definition used). Assuming that most of the ~1200 remaining dog X chromosome genes are X inactivated, a substantially lower fraction of X chromosome genes (~4%) escape in dog compared to at least 23% in human (Tukiainen et al. 2017). This is somewhat surprising given that the dog X and Y chromosomes are substantially less diverged than the human X and Y chromosomes, potentially suggesting that dog X genes might have evolved the ability to be X-inactivated more rapidly than human X genes. However, with analysis of more dogs/tissues and using more sensitive methods, more XCI escapers may eventually be identified. We expand on interspecies comparisons of XCI profiles in the next chapter of this dissertation.

Methods

1. PBMC scRNA-seq

We drew a single ~10 milliliter (mL) whole venous blood sample from each of two adult female F1 goldendoodles, i.e. the first-generation offspring of a purebred golden retriever and purebred standard poodle. Breed composition was confirmed by pedigree records and breed mix inference based on SNPs. These two individuals represented both directions of this cross: Dog1 was the daughter of a female standard poodle and male golden retriever, and Dog2 was the daughter of a female golden retriever and a male standard poodle. Having both cross directions allows us to differentiate between allele-specific expression caused by XCI and parent of origin effects. Blood samples were collected in 12 mL EDTA tubes and stored on ice for 2.5 hours before being processed according to the protocol of Viana et al. 2013: Blood samples were placed in a mixture of Ficoll–Hypaque (Sigma Chemical Co., USA, density: 1.119 g/mL) and Ficoll–Hypaque (Sigma Chemical Co., USA, density: 1.077 g/mL) at a 1:3 ratio (Ficoll/blood) in sterile polystyrene conical bottom tubes (Falcon™, Corning ®, USA). Samples were centrifuged at 700 × g for 80 min at 22 °C. The ring of mononuclear cells collected at the Ficoll–Hypaque interface was transferred to another tube with 40 mL of Falcon sterile 1× PBS containing 10% FBS. This tube was centrifuged two times at 400 × g for 10 min at 4 °C. After the supernatant was discarded, the cells were resuspended in 1 mL of cell culture medium RPMI 1640. Cells were counted in a Neubauer hemocytometer chamber to determine the numbers of cells and were diluted a concentration of 1 million cells per mL.

Using the 10x Genomics Chromium platform, these cells were combined with 10x Genomics v1 chemistry 3' Single Cell library reagents for reverse transcription to generate 150bp cDNAs with UMIs and cell-specific barcodes. The cells were then combined and lysed, and cDNAs were amplified using polymerase chain reaction (PCR). Dual-index paired-end sequencing of amplified cDNA was performed on the Illumina NextSeq500 platform according to the protocol recommended by the manufacturer.

2. Processing and filtering scRNA-seq data

The raw RNA-seq read data was processed with the 10x Genomics software CellRanger 2.0.2 to obtain cell-specific expression data. First, the pipeline mkfastq was used to demultiplex the raw Illumina reads, producing FATSQ files for each sample containing all reads across all cells. These were used as input to the cellranger count pipeline, which separated reads according to their cell and molecule barcodes, trimmed barcode and adapter sequences, and used STAR (Dobin et al. 2013) to perform read alignment to the canFam3.1 reference genome. This pipeline also performed filtering to remove PCR duplicates and generated a unique molecular identifier (UMI) count for each gene in each cell.

To identify the blood cell subpopulations represented in each sample, we used the R package Seurat v2.0 (Satija et al., 2015) and the suggested guided clustering workflow (http://satijalab.org/seurat/pbmc3k_tutorial.html). We first removed cells that were outliers in terms of their percent mitochondrial RNA content and total number of detected genes. Next, we identified the top ~2500 most variable genes across each sample to use for cell clustering. Using only these genes, we regressed out unwanted sources of variation and performed dimensionality reduction using principal components analysis on the resulting data. After examining the amount of variance explained by each principal component, we used the top ten principal components to cluster cells using the K nearest neighbors algorithm, and the top ten most differentially expressed genes in each cluster were identified. Cell type markers in these lists were then used to assign a putative blood cell subtype to each cluster, and clusters were visualized using t-Distributed Stochastic Neighbor Embedding (tSNE) (van der Maaten and Hinton 2008). These plots are shown in Figure 2.1. The workflow used to perform this analysis is available as an R Markdown file in the supplementary materials (File S2.1).

3. Genotyping by SNP array

A buccal swab was taken from each dog using the Embark Veterinary sample collection kit. The samples were then genotyped at 214,504 markers across the entire genome on the

Embark Veterinary custom Illumina CanineHD array. The breed mix of the dogs was inferred by Embark based on these SNPs, confirming that both dogs were F1 goldendoodles.

4. Identifying ASE-informative sites

Only heterozygous SNPs are informative for ASE analysis. Based on the SNP array, Dog1 and Dog2 were heterozygous at 1,388 and 1,693 X markers on the Embark SNP array, for a total of 2,404 unique markers. However, only 5 and 7 of these markers were covered by scRNA-seq reads for Dog1 and Dog2, respectively. Therefore, heterozygous sites were also identified directly from the aggregate (i.e. pooled across all cells) scRNA-seq data for each dog. The aggregate BAM files were run through samtools mpileup followed by bcftools call –m to generate a VCF file with, for each cell, base quality and allelic depth data for the reference and alternate allele at all sites covered by scRNA-seq reads in that cell. Because reads containing the reference allele mapped more confidently to the reference genome, the proportion of reads made up by the reference allele at known biallelic sites was typically well over 50% (Figure 2.4). To account for this bias, we considered a site to be heterozygous in a given individual if it 1) had at least 8 reads covering it across all cells; and 2) had a minor allele frequency of at least 15%. The 8 read minimum was based on Tukiainen et al. 2017, and the 15% minor allele frequency threshold was based on the observed minor allele frequencies at the known heterozygous (minimum 20%) and homozygous sites (maximum 14.5%) from the SNP array. We only considered biallelic sites where both alleles were SNPs rather than insertions or deletions. Using these criteria, we identified 701 and 1,351 X chromosome heterozygous sites across 1,010 and 656 cells in Dog1 and Dog2, respectively.

5. Evaluating ASE in single cells

For each ASE-informative site, we counted the number of cells for which the site showed bi- or monoallelic expression. Sites covered by at least 8 reads within a given cell were considered to show biallelic expression within that cell if the minor allele made up at least 10% of reads and were considered to show monoallelic expression within that cell if the minor allele made up less than 10% of reads. For Dog1 and Dog2 respectively, these criteria identified 178 and 228 ASE-informative X chromosome sites showing evidence of biallelic expression in 2 or more cells, and 184 and 502 showing evidence of monoallelic expression in 2 or more cells and biallelic expression in 2 or fewer cells for Dog1 and Dog2, respectively.

For every X gene, the start and stop positions of the transcription range were taken as the union of Ensembl transcript ranges across all transcripts for that gene. The number of ASE-informative sites showing biallelic expression within each gene's transcript range was then calculated for each gene in each cell. If the average minor allele frequency across all ASE-informative sites in a gene was at least 10%, it was called as escaping XCI in that cell. If the average minor allele frequency across all ASE-informative sites in a gene was less than 10%, it was called as X-inactivated in that cell. XCI status calls for all X genes across all cells in both dogs (with usable data) are shown in Figure 2.5.

6. XCI classifications

If the same gene was called as an XCI escaper in 2 or more cells, it was considered a candidate XCI escaper in that individual. If it was not called as escaping XCI in at least 2 cells and was called as X-inactivated in 1 or more cells, it was considered X-inactivated in that individual. If a gene did not meet either of these criteria, it was considered non-informative for XCI status calling in that individual. The consensus XCI status for all X genes is shown in Figure 2.5.

CHAPTER 3

CONSERVATION OF XCI PROFILES ACROSS MAMMALS

Introduction

XCI and other sex chromosome dosage compensation mechanisms are intrinsically tied to the evolutionary histories of the chromosomes they regulate. The therian X and Y chromosomes are estimated to have originated from a pair of autosomes ~310 million years ago (Mya) (Lahn and Page 1999). Sometime between the divergence of monotremes ~166 (Mya) and the divergence of marsupials and placental mammals ~148 Mya (Veyrunes et al. 2008), *Sry* appears to have acquired its dominant sex-determining function (Wallis et al. 2007), accelerating X-Y divergence. After this point, a series of large-scale inversions resulted in at least four distinct evolutionary "strata" on eutherian X chromosomes corresponding to the degree of divergence from Y (Lahn and Page 1999; Pandey, Wilson-Sayres, and Azad 2013), some of which are lineage-specific.

It is unclear when exactly in this history X chromosome dosage compensation originated, and in what form. Marsupials lack *Xist* (Hore et al. 2007) but have their own dosage compensation mechanisms including the XCI-mediating long non-coding RNA *Rsx* (Grant et al. 2012). Because various forms of imprinted XCI are observed in both marsupials and placental mammals, imprinted XCI is assumed to be more ancient than *Xist*-mediated random XCI (Chaumeil et al. 2011; Shevchenko et al. 2013). If this is the case, *Xist*-mediated random XCI likely arose after marsupial-placental divergence at ~148 Mya (Luo et al. 2011) but before the most recent common ancestor of extant placentals ~100 Mya (Wible et al. 2007). Since then, lineage-specific X-Y divergence has resulted in substantial variation in the size and gene content of the mammalian PARs (Raudsepp and Chowdhury 2015).

Where do XCI escapers fit into this history? As mentioned in Chapter 1, several studies have shown that the length of time for which a given X gene has been diverged from its Y gametolog is strongly predictive of its XCI status, with more deeply-diverged genes being more likely to be X-inactivated and more recently-diverged genes being more likely to escape XCI (Jegalian and Page 1998; Wilson-Sayres and Makova 2013). Assuming that biallelic expression is the deep ancestral state for all X chromosome genes, genes escaping XCI have retained their ancestral XCI status. Therefore, if an XCI escaper is sufficiently diverged from its Y homolog to have stopped recombining with it, its status as an XCI escaper could, broadly speaking, either result from inefficient XCI or from some functional constraint related to XCI escape such as high/biallelic expression (e.g. X genes in conserved X-Y gametolog pairs located outside the PARs described in Bellott et al. 2014). We refer to these two possibilities as the "bug" and "feature" models of XCI escape evolution (respectively) in Chapter 1.

With XCI profiles from multiple therian species, we can add another dimension to the "feature" and "bug" models: time. For genes with XCI status calls in three or more species, we can infer their likely ancestral XCI status in the common ancestor of those species, and in some cases, the location of XCI status changes in the phylogeny. When this information is intersected with measurements of selection, we can also identify the specific branches where evolutionary constraint or relaxation of constraint likely occurred. For example, if XCI escape is homologous, i.e. the conserved ancestral state of orthologous genes, it could have persisted either due to ongoing directional selection over millions of years, or simply by chance if selection was relaxed prior to the genes diverging. Genes escaping XCI in multiple mammalian lineages fit this homology model, with deeper divergence between the lineages under consideration more strongly supporting it. Conversely, XCI escape could have evolved independently in diverged lineages either in response to similar selective pressures, or parallel relaxation of selection. Parallelism cases could include non-orthologous genes in different lineages that are experiencing selection on similar functional feature(s) related to XCI escape. They could also include orthologous genes reverting to escaping XCI in multiple mammalian lineages if the ancestral gene was already X-inactivated in the common ancestor of those lineages, or cases where the gene was not yet diverged from its Y gametolog when two lineages split.

Using previously published XCI profiles for human, mouse, and opossum together with the dog XCI profile described in the previous chapter (the number and proportion of genes escaping XCI in each species is shown in Table 3.1)., we performed three analyses to look for evidence of convergent evolution of XCI profiles. First, we evaluated conservation of XCI status across X chromosome orthologs in human, mouse, dog, and opossum to reconstruct the evolutionary history of XCI status changes at these loci. Next, we compared the gene ontology annotations of XCI escapers across species to identify shared biological functions. Lastly, we compared sequence conservation between XCI escapers and X-inactivated genes within each species to evaluate the extent to which the pattern of stronger conservation in human XCI escapers compared to X-inactivated that we describe in Chapter 1 of this dissertation genes generalizes to other lineages.

Results

1. XCI profile conservation

When accounting for the history of X-Y divergence in multiple species, it is important to note that the placental and marsupial X chromosomes differ substantially in their gene content due to a large-scale translocation between the ancestral therian X and an autosome in the placental lineage ~105 Mya. The region comprised by the ancestral X and conserved between placentals and marsupials is referred to as the X conserved region (XCR), and the younger, placental-specific translocated region is referred to as the X added region (XAR). In light of this, we examined conservation of all one-to-one (1:1) X orthologs across human, mouse, dog, and opossum to provide an appropriate basis for comparison to conservation of XCI status.

We found that while many X genes were conserved across multiple species, including 225 1:1:1:1 orthologs (Figure 3.1.a), few genes escaped XCI in more than one species (Figure 3.1.b). Specifically, only one gene, *KDMC5*, escaped XCI in all four species, and only 18 genes escaped XCI in two or more species. Importantly, *XIST* is not among the shared XCI escapers due to its poor sequence homology across mammals. Conservation of X-inactivation was more common, with 10 genes X-inactivated in all four species and 118 more X-inactivated in 2/3 or 3/4 species.

For each of the 144 genes with an XCI status call in three or more species, we inferred its most likely XCI state in the last metatherian common ancestor based on the assumptions that the 1) most parsimonious ancestral state was the one requiring the fewest XCI status changes, and 2) that all genes share the same tree topology as the species whole genomes, as shown in Figure 3.2.a. Following these assumptions, we inferred the most likely locations in the opossum, dog, mouse, human phylogeny where these genes switched XCI status. We found that 129 of these genes appear to have been inactivated in the common ancestor of opossum, dog, mouse, and human. The remaining 15 genes escape XCI in 2/3 or 3/4 species, suggesting that they escaped XCI in the common ancestor and have retained that ancestral state these species for over 300 million years. Of these, 8 became inactivated in the mouse lineage, and 2 became inactivated in the human lineage (Figure 3.2.b). Additionally, 55 genes appear to have been inactivated in the metatherian common ancestor and reverted to escaping XCI within individual lineages, with 12 escaping only in opossum, 16 escaping only in dog, 3 escaping only in mouse, and 24 escaping only in human (Figure 3.2.c).

2. Functional conservation

We obtained PANTHER Gene Ontology Biological Process (BP) annotations for XCI escapers in each species, the 18 genes that escape XCI in two or more species, and the 129 genes that are inactivated in 2/3 or 3/4 species. Across all gene sets in all species, only 12 BP categories were observed: cellular component organization or biogenesis (GO:0071840), cellular process (GO:0009987), localization (GO:0051179), reproduction (GO:0000003), biological regulation (GO:0065007), response to stimulus (GO:0050896), developmental process (GO:0032502), multicellular organismal process (GO:0032501), biological adhesion (GO:0022610), locomotion (GO:0040011), metabolic process (GO:0008152), and immune system process (GO:0002376). All of these except biological adhesion and immune system process appeared in at least one XCI escaper across all four species. The proportion and number of genes in each BP category in each group are shown in Figure 3.3.

3. Sequence conservation

PhyloP (Pollard et al. 2010) is a method that produces a per-base measurement of evolutionary constraint in the context of a given phylogenetic model by looking at the rate of DNA sequence change across the branches of the phylogeny in orthologous sequences. Sites predicted to be under stronger evolutionary constraint have more positive scores, and sites predicted to be under less constraint that have evolved rapidly have more negative scores. The absolute value of a site's phyloP score corresponds to its negative log p-value for a likelihood ratio test under a null hypothesis of neutral evolution.

We obtained phyloP scores from the publicly available phyloP100way track in the UCSC Genome Browser's GRCh38/hg38 database. These scores represent estimates of departures from neutral evolution across 100 vertebrate genomes, including human, mouse, dog, and opossum. For each gene in each species, we pulled all phyloP scores at the human genome positions corresponding to that gene's best local alignment to hg38, including both coding and non-coding sequence. The distributions of phyloP scores across all sites in XCI escapers and X-inactivated genes in each species are shown as violin plots in Figure 3.4.

We observed that XCI escapers in dog and opossum have significantly different distributions of phyloP scores than X-inactivated genes in the same species (Mann Whitney U two-tailed p-values = 2.177e-199 and 2.140e-31, respectively), with XCI escaper and X-inactivated gene mean phyloP scores of 2.626 and 2.007 for dog and 3.508 and 2.857 for opossum. In mouse, the distributions of XCI escaper and X-inactivated genes were not significantly different (Mann Whitney U test two-tailed p-value = 0.472), but the XCI escaper mean phyloP score (2.793) was higher than the X-inactivated mean score (2.707). In human, the distributions of XCI escaper and X-inactivate gene phyloP scores were significantly different (Mann Whitney U two-tailed p-value = 0), with a mean phyloP score of -0.012 in XCI escapers and 0.244 in X-inactivated genes. We also compared the phyloP score distributions of the 18 conserved XCI escapers (escaping XCI in two or more linages) and 129 conserved X-inactivated genes (X-inactivated in 2/3 or 3/4 lineages) and found that the distributions for conserved XCI escapers and X-inactivated genes were significantly different (Mann Whitney U two-tailed p-value = 0.001), with a mean score of 0.525 in XCI escapers and 0.333 in X-inactivated genes.

Conclusions

In this study, we compared XCI statuses across human, mouse, dog, and opossum one-to-one X chromosome orthologs. In addition to suggesting that XCI likely existed and was pervasive in the common ancestor of therian and eutherian mammals, our results showed that XCI escape is poorly conserved across these species and has likely evolved independently in multiple lineages at several genes. This observation is consistent with XCI and XCI escape evolving in a highly lineage-specific manner, reflecting the different evolutionary histories of the X chromosome in each species, such as the addition of the XAR in eutherians and structural rearrangements on the X chromosome in each lineage.

As expected, the proportion of genes escaping XCI mirrors the extent of X-Y divergence in each species, as reflected in the size of the PAR in each species: Dog, which has the largest PAR at 6.6Mb, has the highest proportion (31%) of XCI escapers, followed by human with a PAR size of 2.7Mb and approximately 25% of genes escaping XCI, and then by mouse with a PAR size 0.7 and 3-7% of genes escaping XCI. The PAR boundary in opossum is not currently confidently mapped.

The results of our analysis of the conservation of XCI escape across species provide only a few examples of genes that fit the homology model of XCI escape evolution, in which XCI escape is the conserved ancestral state of orthologous genes. Of the 18 genes that we found to escape in two or more of the four species, four (*CA5B*, *KDMC5*, *FAM122B*, and *HCFC1*) are members of deeply conserved X-Y gene pairs. The X gametologs in these gene pairs are expected to escape XCI, provided that their Y gametologs have not diverged from their ancestral functions, duplicated, or translocated to autosomes (Bellott et al. 2014). A consequence of such divergences is that many of these genes do not have 1:1 orthologs in multiple species and as such were not considered in this analysis.

Most genes escaping XCI in the four species we have considered do so in a single lineage, which could result either from retention of the ancestral state of biallelic expression in that lineage, or from reversion to XCI escape if the gene was X-inactivated in an ancestral species. We observed examples of such genes in all four lineages, but to varying degrees, with mouse showing markedly fewer inferred reversions to XCI escape than human, dog, or opossum. Without knowing more about the mechanisms of XCI and XCI escape, it is unclear whether conservation of ancestral XCI escape in a single lineage or reversion to XCI escape in that lineage is more plausible when both scenarios would require the same number of XCI status changes. It is also important to note that individual gene phylogenies may not match the species phylogeny, which we did not account for.

Despite XCI escape being highly lineage-specific, we found that XCI escapers in these four species overlap highly in their Gene Ontology biological process annotations. Some of these, such as cellular process (the most common annotation among the human XCI escapers, with 50% falling into this category), provide little insight into these genes' specific functions. It is striking that these genes overlap so extensively in their functional annotations given how few of them are orthologous to each other and that in all species XCI escapers represent only a fraction of X chromosome genes. However, these annotations were similarly distributed among the 129 genes that were inactivated in the majority of species. This may indicate that these functional categories are simply common among X chromosome genes rather than being associated with a particular XCI status or degree of XCI status conservation.

Gene Ontology annotations are more complete in human compared to other species. When the human XCI escaper BP annotations are broken down further, the Cellular Process category includes histone and protein modification processes such as acetylation and deacetylation, suggesting that XCI escapers may play important roles in genome-wide regulation of gene and protein expression, in agreement with earlier findings regarding X-linked members of conserved X-Y pairs by Bellott et al. (2014).

Out of all of our results, our analysis of evolutionary conservation in XCI escapers provides the strongest support for the parallelism model of XCI evolution by showing that non-orthologous genes that escape XCI in different species share the common trait of being more conserved at the sequence level as a group than X-inactivated genes in the same species. We note that the phyloP100way scores for human XCI escapers do not follow this trend, in apparent disagreement with our results in Chapter 1. This discrepancy appears to in arise from the exclusion of genes with XCI status calls derived from methylation patterns (Cotton et al. 2015; Schultz et al. 2015) from this analysis, which was done in the interest of using a consistent functional definition of XCI escape across species. It is also worth noting that the phyloP data used in this analysis is expected to capture very ancient evolutionary conservation, in contrast to our analysis in chapter 1, which looked at conservation only among primates. A general conclusion of this analysis is that the magnitude of evolutionary constraint operating on both XCI escapers and X-inactivated genes varies substantially by lineage, and across genes in each XCI class within lineages.

Recently, Naqvi et al. 2018 showed that based on conserved microRNA-gene interactions, sensitivities to gene dosage increase appear to have been conserved from ancestral amniote autosomes between many inactivated X and Z genes in mammals and birds. In contrast, human XCI escapers show significantly less dosage sensitivity conservation by this same metric. The authors of this study suggest that this indicates that XCI escapers tend to be less sensitive to dosage increases than other X genes, with the exception of XCI escapers that are members of deeply conserved X-Y gene pairs. It is not yet clear to what extent this trend generalizes to other mammalian species, but it is consistent with our finding that more recent selection and/or relaxation of selection could be occurring on the distal branches of the mammalian phylogeny, potentially reflecting a higher tolerance for dosage changes among XCI escapers. Further comparisons of lineage-specific evolutionary conservation among XCI escapers and X-inactivated genes in these and other species will be necessary to investigate this possibility.

Overall, we find evidence that XCI escape at individual genes can evolve according to either homology or parallelism scenarios, but that parallelism is likely more common. These results suggest there are likely multiple mechanisms by which a gene's XCI status can evolve in addition to the extreme cases of strong constraint or extensive relaxation of selection. Generation of comprehensive XCI profiles for additional mammalian species will be necessary to further clarify the complicated history of XCI evolution. Additionally, the analyses performed in this study have limited ability to differentiate between genes that escape due to selective pressures and genes for which XCI escape is merely a result of inefficient XCI. We hope to address this possibility in the future by applying methods that more explicitly test for relaxation of purifying selection along individual branches of a phylogeny.

Methods

1. Processing of previously-generated ASE-based XCI profiles

When XCI status is assayed by ASE, different criteria may be used to categorize it as escaping, inactivated, or any number of intermediate states. In this analysis, these classifications must be applied as consistently as possible across the different data sets to ensure that meaningful comparisons are being made. ASE-based XCI profiles are available for human (Carrel and Willard 2005), mouse (Berletch et al. 2015), and opossum (Wang et al. 2014), each of which used slightly different definitions of XCI escape and XCI status categories based on methodology and samples.

Carrel and Willard's definition of an XCI escaper was genes showing an X_i/X_a expression ratio of at least 10% in at least 7 of the 9 XX human-murine hybrid fibrobast cell lines they tested. The remaining genes were considered either "heterogeneous" escapers, or X-inactivated. Berletch et al. obtained their XCI status calls from bulk RNA-seq of brain, ovary, and spleen samples from two female BL6 x Spretus F1s. They defined XCI escapers as genes showing a significant probability (per a binomial test) that the degree of X_i expression was greater than expected for a monoallelicaly expressed gene given the total number reads of each allele observed for that gene. All genes that did not meet this criterion were considered X-inactivated. Wang et al. acquired XCI status calls by performing bulk RNA-seq on brain samples from four female F1 LL1 x LL2 individuals, and defined XCI escapers as genes showing an X_i:X_a expression ratio of at least 10% in at least one individual. All other genes were considered either X-inactivated, except a few which were annotated as candidate escapers based on showing non-zero X_i expression of less than 10% of X_a expression. In the previous chapter, we described calling XCI status in dog based on the average minor allele frequency across heterozygous sites within each gene, with genes showing an average minor allele frequency of 10% or greater in two or more cells in the same individual called as escaping XCI, and those with an average minor allele frequency of less than 10% in one or more cells in a single individual called as X-inactivated. To minimize differences in the functional definition of XCI escape across species, we re-called XCI statuses in all three data set using an X_i:X_a expression ratio of at least 10% in at least 25% of individuals as the cutoff for calling a gene an XCI escaper.

Additionally, to compare XCI statuses across the three species, it was necessary to update the gene names used in each XCI profile. Gene identifiers from each XCI data set were cross-referenced with the Ensembl 89 data base to get up-to-date Ensembl gene IDs, and discard genes without Ensembl gene IDs or update/consolidate those with retired IDs. Following these steps, we were left with 451, 543, 152, and 168 genes with both an XCI status call and an Ensembl gene ID in human, mouse, dog, and opossum, respectively. The data are available in the supplementary materials (File S3.1).

2. Orthology

Ortholog data for human (GRCh38.p10), mouse (GRCm38.p5), dog (canFam3.1) and opossum (monDom5) was downloaded from the Ensembl data base (release 89) (Zerbino et al. 2018). These included the gene IDs, chromosome/scaffold, and homology type for each gene in each species. We were only interested only in one-to-one X chromosome orthologs, meaning that a given X gene had exactly one best match ortholog, also on the X chromosome, in each of the other species it was being compared to. The one exception to this definition was when an X gene in one species was strongly homologous to both members of an X-Y gametolog pair in another species. For instance, the opossum X gene KDM5C (ENSMODG00000009765) has two orthologs in human: KDM5C (ENSG00000126012) on the human X, and KDM5D

(ENSG00000012817) on the human Y. Although the two human genes differ, they originated as a homologous pair and only recently stopped recombining, meaning that neither is a duplicate of the other. Therefore, we can effectively treat the relationship between ENSMODG00000009765 and ENSG00000126012 as one-to-one.

To account for any possible discrepancies in annotations across the data sets for different species, we further filtered the gene data sets to ensure that gene list memberships were reciprocal among compared species, meaning that each one-to-one X ortholog list for a given set of species was identical regardless of which species' data set it was generated from. Final gene lists are available in the supplementary materials (File S3.2).

3. Gene ontology comparisons

We obtained PANTHER Gene Ontology Biological Process (BP) annotations for XCI escapers in each species from http://www.pantherdb.org/ (Version 13.1, accessed May 3rd, 2018) (Mi et al. 2017). These are available in File S3.3. Queries were the Ensembl gene IDs for XCI escapers and X-inactivated genes from each species and the 144 conserved genes, which are available in the File S3.3.

4. Sequence conservation comparisons

The phyloP100way track in the USCS Genome Browser hg38 database was queried using the UCSC Genome Table Browser tool (http://genome.ucsc.edu/cgi-bin/hgTables) (Karolchik et al. 2004). Each query consisted of human genomic intervals corresponding to each mouse, dog, and opossum gene. These intervals were obtained by using the Ensembl BLAT (Kent 2002) web tool (ensembl.org/Multi/Tools/Blast) to pull the DNA sequence for each gene by its Ensembl gene ID and align it to the human (hg38) genome sequence. For each gene, the human interval with the highest E score was taken as the best alignment, and phyloP100 way scores were pulled from the corresponding genomic intervals in the phyloP100way track. The Ensembl gene IDs and corresponding hg38 genomic intervals are provided in File S3.4, and phyloP scores for each interval for each species and conserved genes are provided in File S3.5.

\

CHAPTER 4

DISCUSSION

Summary

In this dissertation, I have described three projects unified by the goal of better understanding the evolutionary drivers and history of XCI escape. In the first, I showed that human XCI escapers tend to show higher and broader gene expression in both males and females in addition to stronger evolutionary constraint since the divergence of the human lineage from other primates. This piqued my interest in the possibility that XCI escape might have important biological consequences in humans and perhaps other mammals, but it was clear that with limited data from other species, we had insufficient information to rigorously evaluate this hypothesis.

Subsequent publication of more comprehensive XCI profiles for opossum (Wang et al. 2014) and mouse (Berletch et al. 2015) introduced the possibility of performing an informative cross-species comparison of XCI profiles. Motivated by this, I next used newly-available scRNA-seq technology to generate an XCI profile for the domestic dog. I then incorporated those data into an analysis of conservation of XCI profiles across human, mouse, dog, and opossum 1:1 X orthologs. This analysis produced two main findings: First, XCI escape is highly lineage-specific, with several XCI escapers appearing to have reverted to biallelic expression after previously evolving to become inactivated. Second, there is extensive heterogeneity in evolutionary conservation within XCI categories in each species.

Implications

This project began as an investigation into whether or not XCI escapers showed distinct patterns of genetic diversity in humans compared to other X chromosome genes, which was motivated by prioritizing genetic association results in human disease studies. In the five years since then, the work that I and others have done on XCI escape has produced results that sometimes appear contradictory. The results described in the previous chapter of this dissertation shed some light on these contradictions by demonstrating that the history of XCI profile evolution has been more complex than existing models generally account for. Taken together, the evidence seems to suggest that XCI escape is neither necessarily a consequence of selection, nor necessarily a neutral feature resulting from inefficient XCI. The majority of genes appear to fall somewhere between these two extremes. Consequently, I believe that we should be cautious in treating XCI classes as functionally cohesive groups.

Our findings support previous predictions that XCI occurred in the common ancestor of marsupials and placentals. Additionally, our observation that XCI escape is deeply conserved for a handful of genes suggests that XCI escape might also have occurred in this common ancestor. It remains unclear whether the mechanism of XCI in this ancestor would resemble the chromosome-wide XCI seen in mammals, the gene-specific silencing seen in the avian Z chromosome (Itoh et al. 2007), or neither. However, the large number of genes that are X-inactivated across multiple species could be interpreted as evidence that XCI was pervasive across the chromosome. Incorporating XCI profiles from additional species might further clarify these points.

Future directions

While this work has provided novel insights into the evolutionary origins of XCI escape, it perhaps most importantly serves to motivate future experiments that more directly interrogate the functional impacts and mechanisms of XCI escape. For example, gene knock down or knock out experiments in female animals or cell lines can help clarify the extent to which the generally small gain in transcript abundance contributed by XCI escape is necessary or sufficient for various phenotypes. Phenotypes associated with human X aneuploidy syndromes, such as fertility and neurological development, may be of particular interest. These approaches combined with scRNA-seq could prove extremely powerful for investigating the impact of abnormal XCI patterns on genome-wide transcription at the cell level.

Another intriguing opportunity for further investigation is afforded by technologies that capture the three-dimensional organization of chromatin inside cells, such as 5C (Dostie et al. 2006; Ferraiuolo et al. 2012) and HiC (Lieberman-Aiden et al. 2009). Combining the topological data provided by these methods with DNA sequence motif detection tools could vastly improve our understanding of how specific regulatory elements affect XCI status of nearby genes. Chromatin conformation capture studies of the X_i have already solidified a role for LINE elements and the CTCF binding sites in XCI (e.g. Berletch et al. 2015; Deng et al. 2015), but modern computational methods could accelerate this process. For instance, A 2006 study (Wang et al.) had limited success in using a machine learning approach to identify DNA sequence motifs that were predictive of XCI status in human, but a similar strategy could perhaps be taken to identify sequence elements associated with specific X_i topological associated domains defined by chromatin conformation capture data.

Lastly, the observation that XCI escape evolves relatively rapidly and varies substantially among species raises the possibility that XCI status could contribute to or result from species-specific biological features. The hypothesis that regulatory evolution underlies much of the phenotypic differences between species is well established and widely supported (e.g. King and Wilson 1975; Brawand et al. 2011; Diehl et al. 2018). Given the important role of several XCI escapers in brain development (reviewed in Berletch et al. 2011) and overall enrichment for brain development genes and brain expression on the mammalian X chromosome (Zechner et al. 2001; Skuse 2005; Nguyen and Disteche 2006), investigating evidence for lineage-specific positive selection on XCI escapers and associated *cis*-regulatory elements might lead to interesting discoveries.

TABLES AND FIGURES

Chapter 1:

Figure 1.1: XCI escaper to X-inactivated gene ratios of fraction of sites under strong purifying selection (*S*) and fraction of neutral sites (*N*) by Y gametolog status.

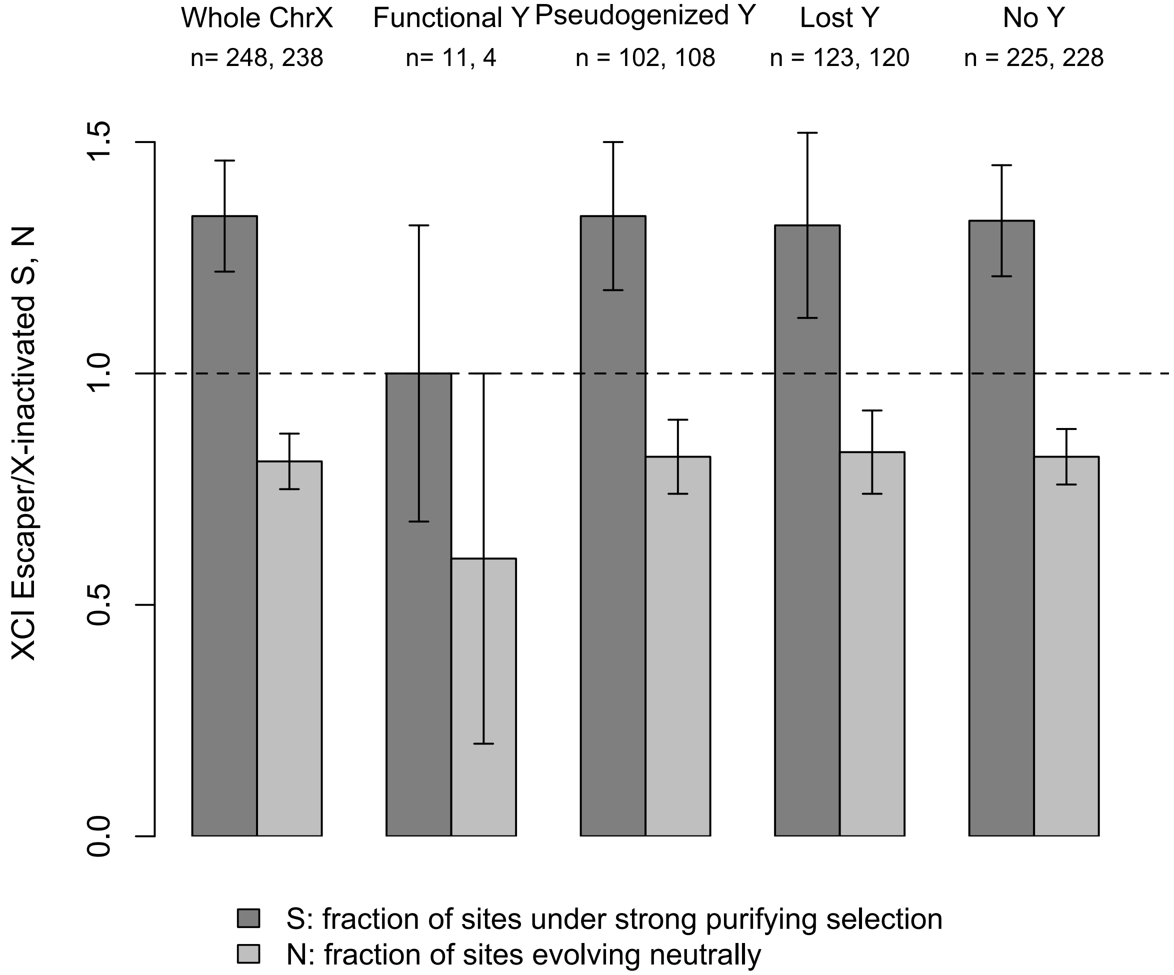


Fig. 1.1: XCI escapers show larger fractions of sites under strong purifying selection (*S*) and smaller fractions of neutral sites (*N*) than X-inactivated genes with the same Y gametolog status. XCI escaper-to-X-inactivated ratios of the values of *N* and *S* are presented, with standard deviations based on 1000 bootstrap samples of genes in each XCI/Y category (error bars denote ±1 standard deviation). The numbers (*n*) of XCI escaper and X-inactivated genes in each Y gametolog category are shown below the category names at the top of the plot. The Whole ChrX category includes genes from all other categories, as well as 18 genes that did not have a Y gametolog status assignment (12 XCI escapers and 6 X-inactivated genes). The No Y category is the union of the Pseudogenized and Lost Y categories.

Figure 1.2:

1.2.a: XCI category average gene expression for genes by XCI/Y status, tissue and sex.

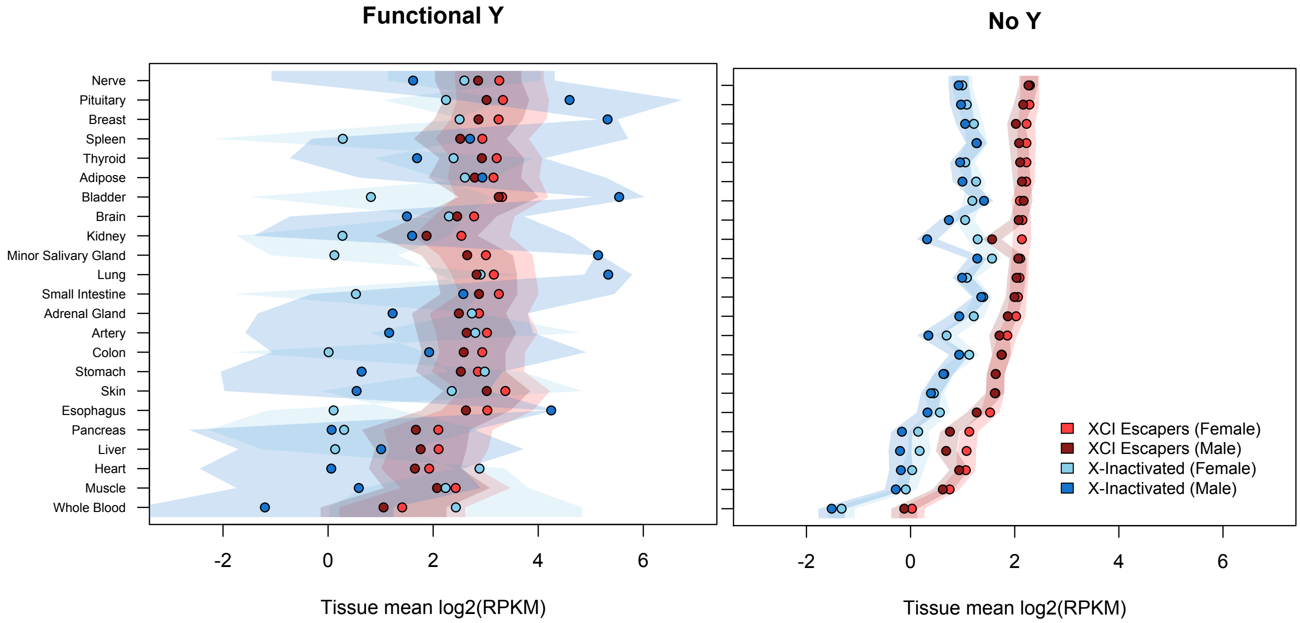


Fig. 1.2.a: XCI category average gene expression by Y gametolog status, tissue and sex. Scatterplots show the XCI/Y category averages of gene expression values for each sex in each GTEx tissue. Shaded areas show the standard deviation of the average across genes. In both plots, individual expression values were filtered by tissue-specific cutoffs (Methods) before being incorporated into gene averages.

1.2.b: Global mean expression for each gene by XCI/Y status and sex.

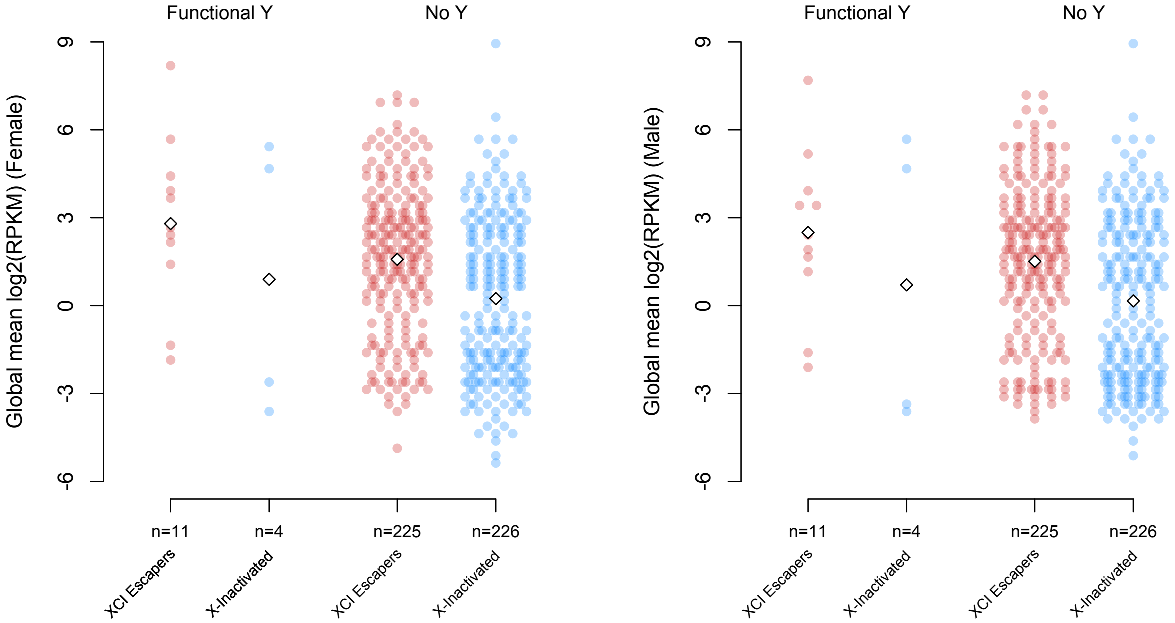


Fig. 1.2.b: Global mean expression for each gene by XCI status, Y gametolog status, and sex. Each point in these one-dimensional scatterplots shows the global mean expression value of a single gene in female (left) and male (right) samples. The white diamonds indicate the mean expression value of the genes in that XCI/Y category. Areas of darker color indicate overlap of two or more points. In both plots, individual expression values were filtered by tissue-specific cutoffs (Methods) before being incorporated into gene averages.

Figure 1.3: Gene tissue specificity index (τ) by XCI/Y status and sex.

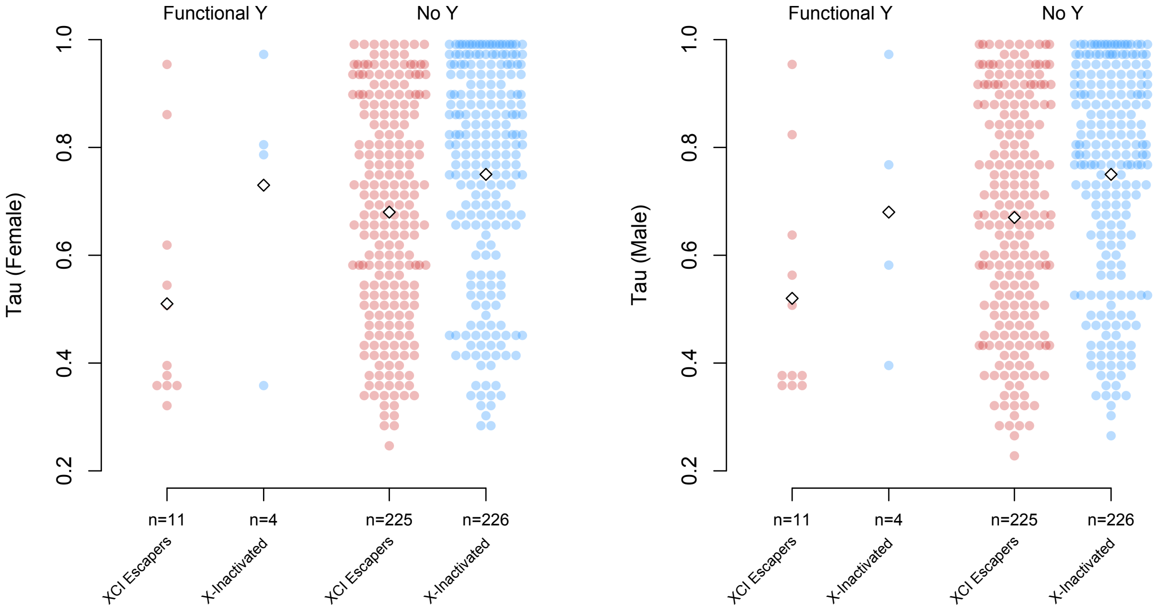


Fig. 1.3: XCI escapers are more broadly expressed than X-inactivated genes with the same Y gametolog status in both sexes: Points show τ statistic values for all genes in each XCI/Y category for female (left) and male (right) samples. White diamonds show the mean of τ across all genes in that XCI/Y category. Areas of darker color indicate overlap of two or more points. Values of τ range from 0 to 1, with higher values corresponding to greater tissue-specificity.

Figure 1.4: Proportion of tissues in which each gene shows significant expression (using 3SD cutoffs) by XCI/Y status and sex.

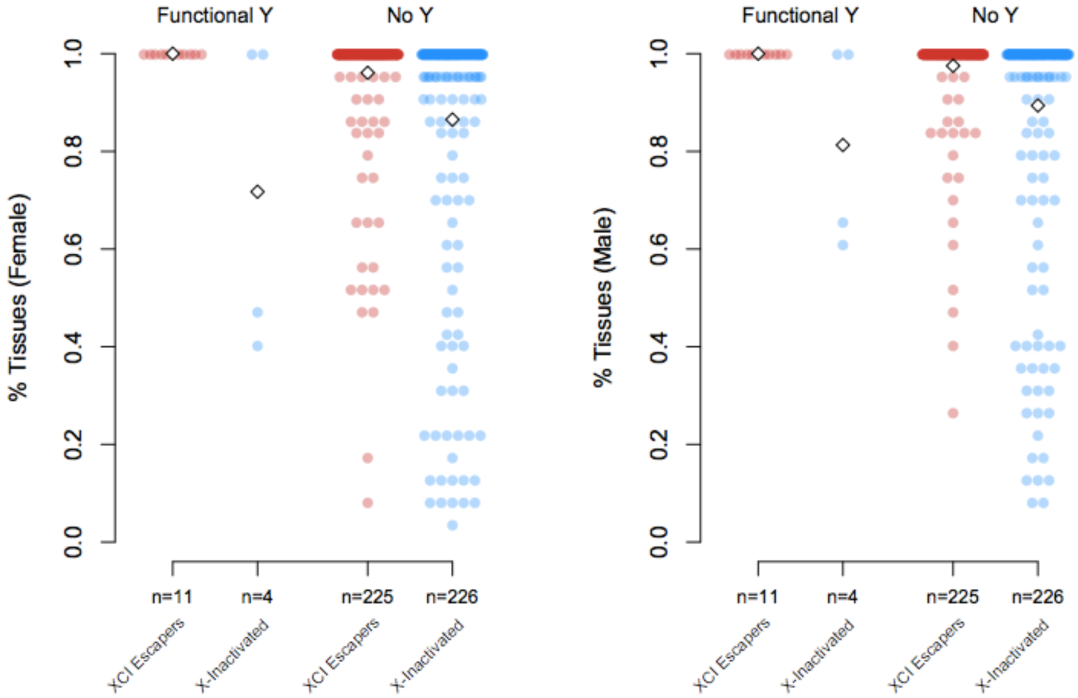
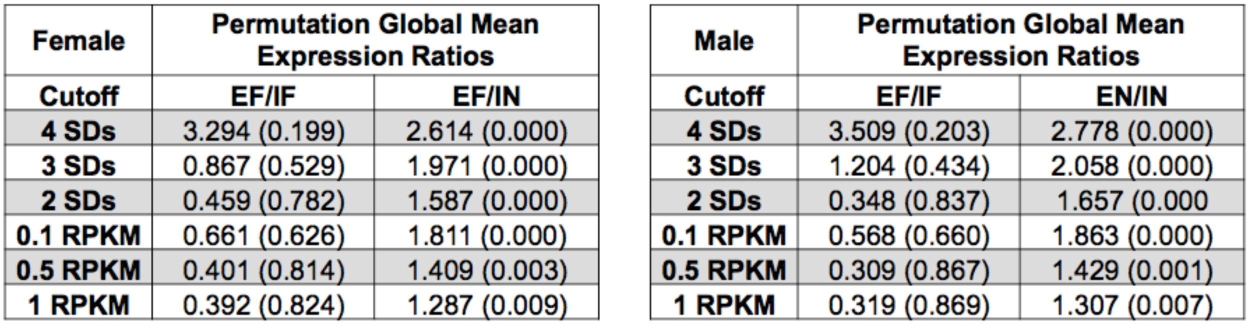


Fig. 1.4: XCI escapers show significant expression in a higher proportion of tissues than X- inactivated genes with the same Y gametolog status. Each point represents the proportion of 23 broad tissue types (excluding sex-specific tissues) for which a gene showed significant expression using the 3SD cutoff described in the text. Gene membership of each XCI/Y category was resampled with replacement 1000 times, and the mean value of these samples for each category is shown as a white diamond.

Table 1.1: Resampling mean and standard deviation of *S* and *N* for XCI/Y categories in the combined XCI data set.

Y Category	XCI category	n Genes	*S*	*N*
All	XCI Escapers	248	0.52+/-0.02	0.46+/-0.02
	X-inactivated	238	0.39+/-0.03	0.57+/-0.03
Functional Y	XCI Escapers	11	0.60+/-0.06	0.37+/-0.05
	X-inactivated	4	0.63+/-0.13	0.32+/-0.14
Pseudogenized Y	XCI Escapers	102	0.52+/-0.03	0.47+/-0.03
	X-inactivated	108	0.39+/-0.04	0.58+/-0.04
Lost Y	XCI Escapers	123	0.50+/-0.04	0.47+/-0.04
	X-inactivated	120	0.38+/-0.04	0.57+/-0.04
No Y (Pseudo.+ Lost Y)	XCI Escapers	225	0.52+/-0.02	0.47+/-0.02
	X-inactivated	228	0.39+/-0.03	0.57+/-0.03

Table 1.2: Global mean expression ratios using various expression minima.



Female	Permutation Global Mean Expression Ratios	
Cutoff	EF/IF	EF/IN
4 SDs	3.294 (0.199)	2.614 (0.000)
3 SDs	0.867 (0.529)	1.971 (0.000)
2 SDs	0.459 (0.782)	1.587 (0.000)
0.1 RPKM	0.661 (0.626)	1.811 (0.000)
0.5 RPKM	0.401 (0.814)	1.409 (0.003)
1 RPKM	0.392 (0.824)	1.287 (0.009)

Male	Permutation Global Mean Expression Ratios	
Cutoff	EF/IF	EN/IN
4 SDs	3.509 (0.203)	2.778 (0.000)
3 SDs	1.204 (0.434)	2.058 (0.000)
2 SDs	0.348 (0.837)	1.657 (0.000
0.1 RPKM	0.568 (0.660)	1.863 (0.000)
0.5 RPKM	0.309 (0.867)	1.429 (0.001)
1 RPKM	0.319 (0.869)	1.307 (0.007)

Table 1.2: XCI/Y category global mean expression level ratios for significantly expressed genes using different minimum expression cutoffs. XCI escaper-to-X-inactivated global mean expression ratios are shown for each Y gametolog category in both sexes. P-values (shown in parentheses) are the proportion of 1000 permutations of gene category membership for which the expression ratio was greater than or equal to the observed ratio. "EF" denotes XCI escapers with functional Y gametologs, "IF" denotes X-inactivated genes with functional Y gametologs, "EN" denotes XCI escapers with no Y gametologs, and "IN" denotes X-inactivated genes with no Y gametolog.

Table 1.3: Resampling mean and standard deviation of the tissue specificity index (τ) for XCI/Y status categories in females and males.

Y Category	XCI category	n Genes	Female τ	Male τ
Functional Y	XCI Escapers	11	0.51+/-0.06	0.52+/-0.06
	X-inactivated	4	0.73+/-0.11	0.68+/0.11
No Y (Pseudo.+ Lost Y)	XCI Escapers	225	0.68+/-0.01	0.67+/-0.01
	X-inactivated	226	0.75+/-0.01	0.75+/-0.01

Chapter 2:

Figure 2.1: Pedigree of Dog1 and Dog2

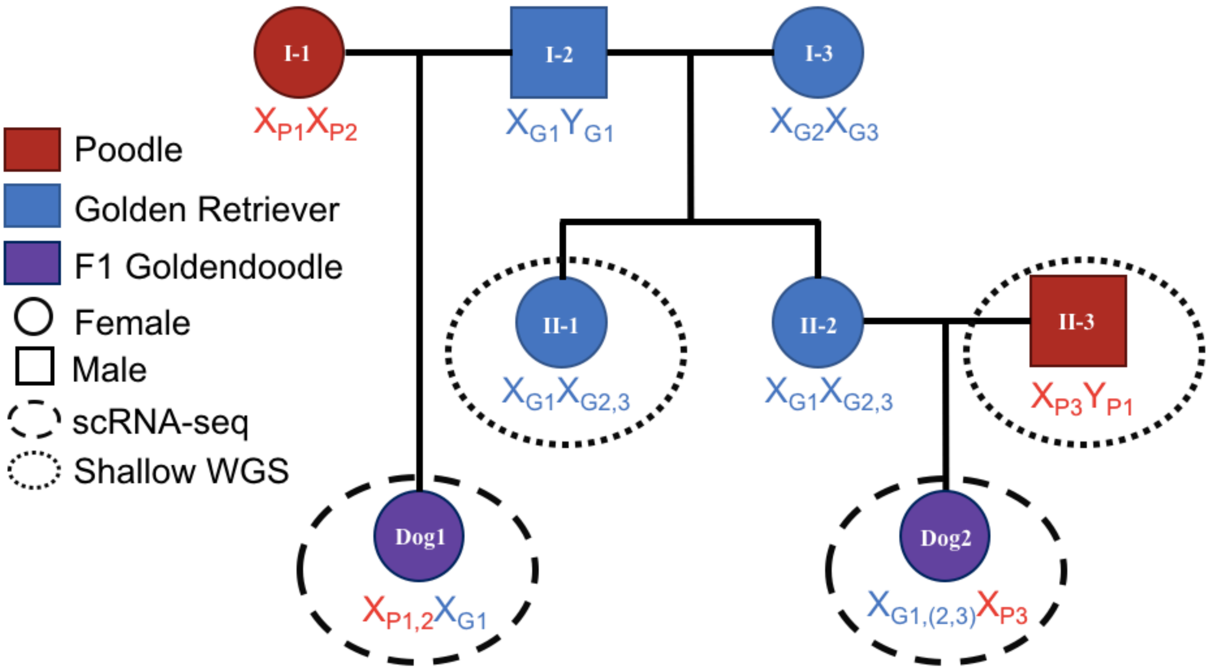


Fig. 2.1: The breed, sex, and the identity of X chromosomes of the parents of Dog1 and the parents and grandparents of Dog2 are shown, as well as for the half-sister of Dog1/aunt of Dog2 (individual II-1). Saliva samples from individuals II-1 and II-3 were subject to shallow whole-genome sequencing to facilitate phasing of the Dog1 and Dog2 scRNA-seq data (analysis ongoing). Individual I-2, the father of Dog1, is deceased and could not be sampled.

Figure 2.2: tSNE plots of PBMC cell subpopulations in scRNA-seq samples.

Fig 2.2.a: Dog1 PBMC populations.

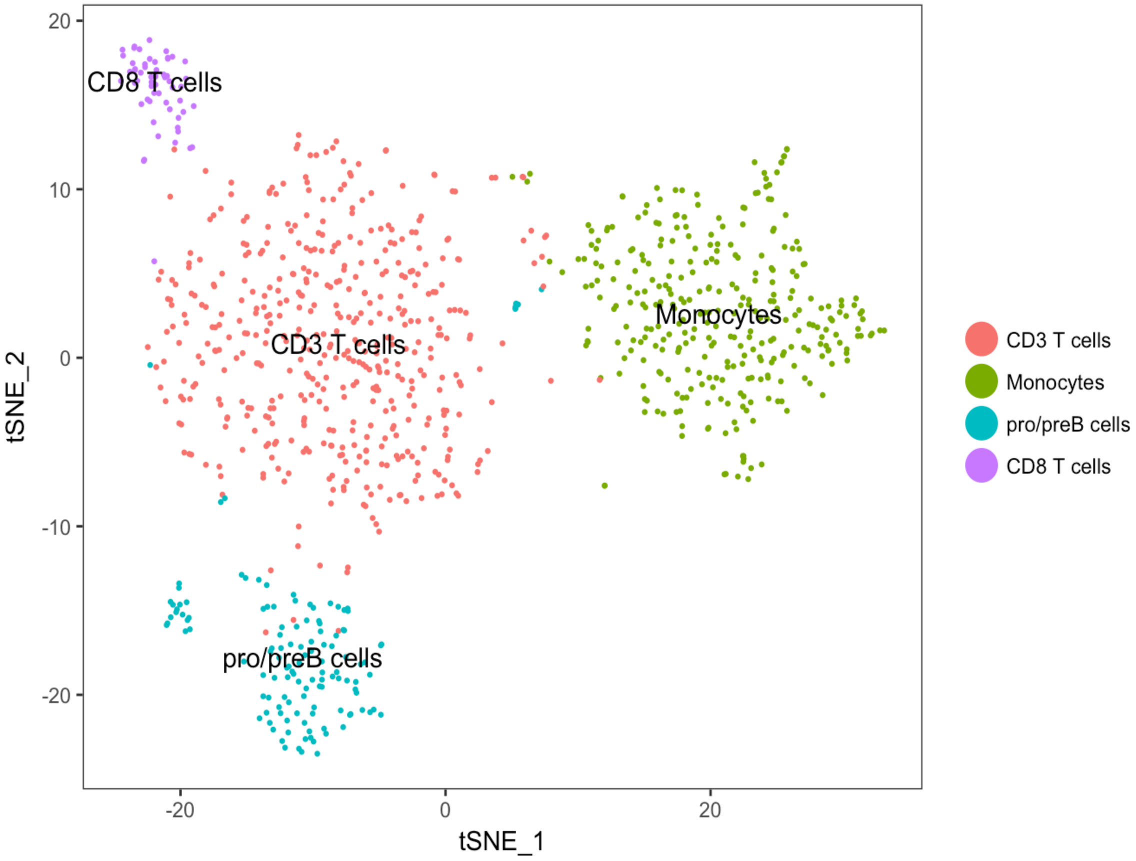


Fig 2.2.b: Dog2 PBMC populations.

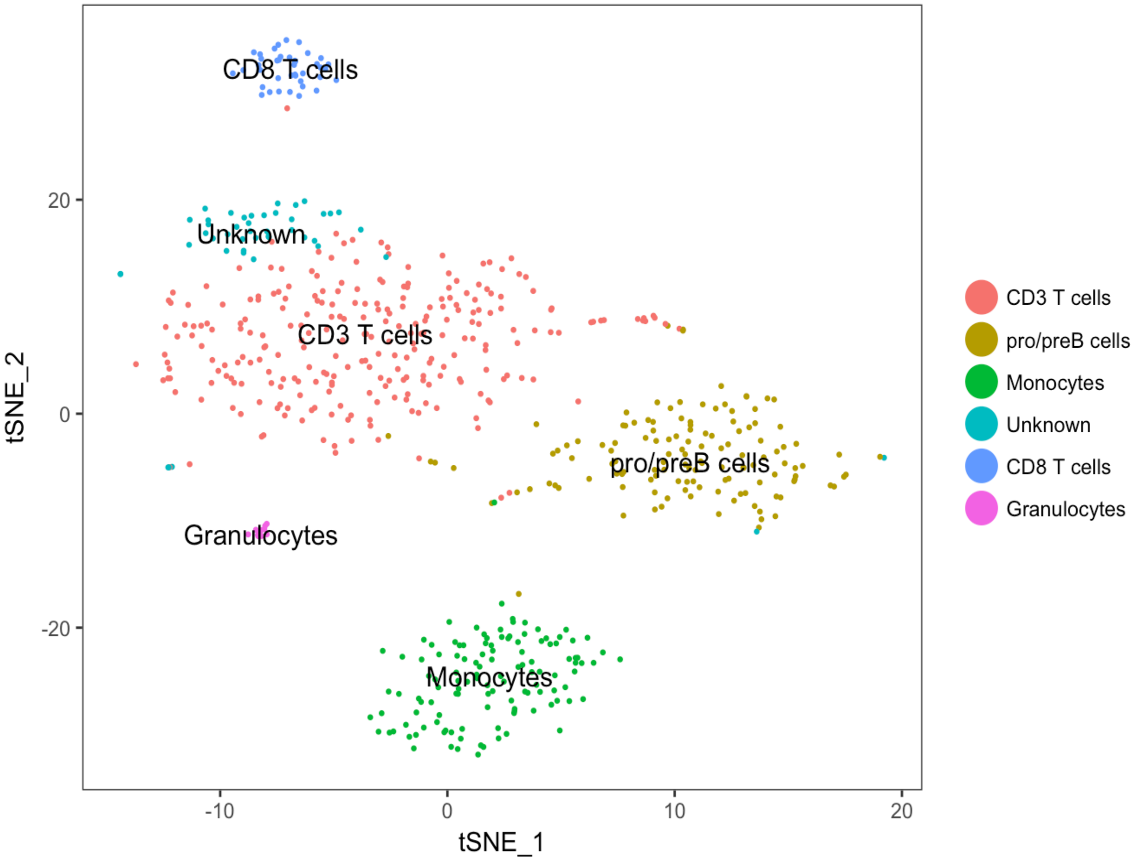


Figure 2.3: Example raw scRNA-seq data for genes with known expression patterns.

Fig. 2.3: These panels show Integrative Genome Viewer (IGV) screenshots of read alignments from individual cells at two genes – one that is known to be expressed solely from X_i, and one in the PAR that is expected to show biallelic expression

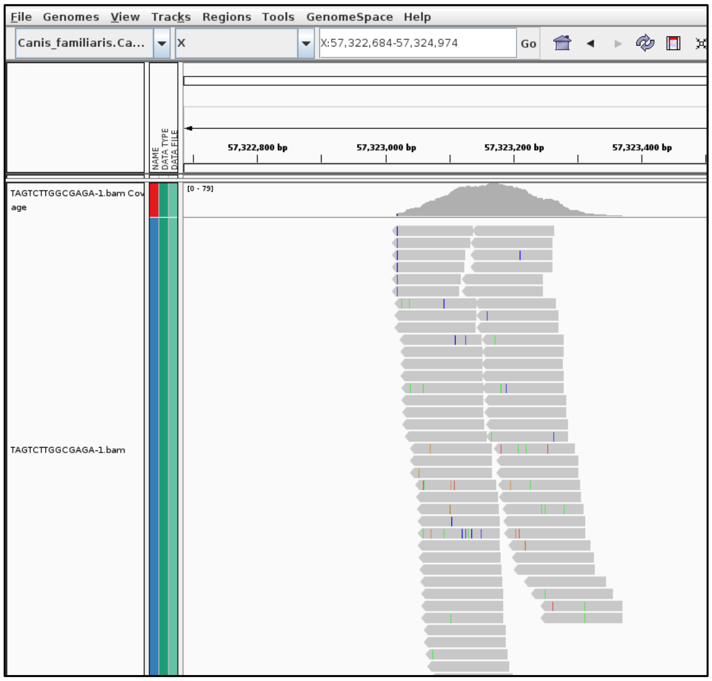


Fig. 2.3.a: *XIST*, the RNA that mediates XCI and is solely expressed from Xi, shows monoallelic expression (with some error).

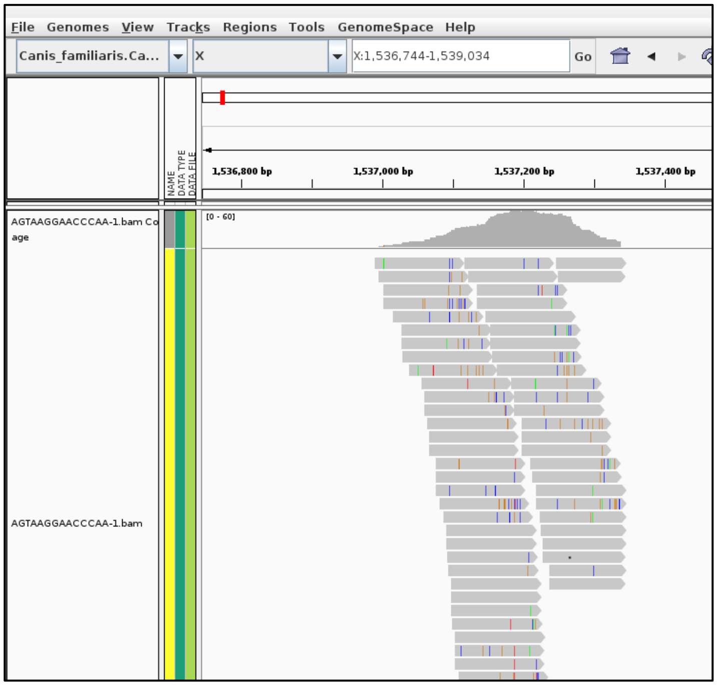


Fig. 2.3.b: ARSE, a PAR gene, shows expression of multiple alleles.

Figure 2.4: Fraction of reads at ASE-informative sites coming from the reference allele across all cells.

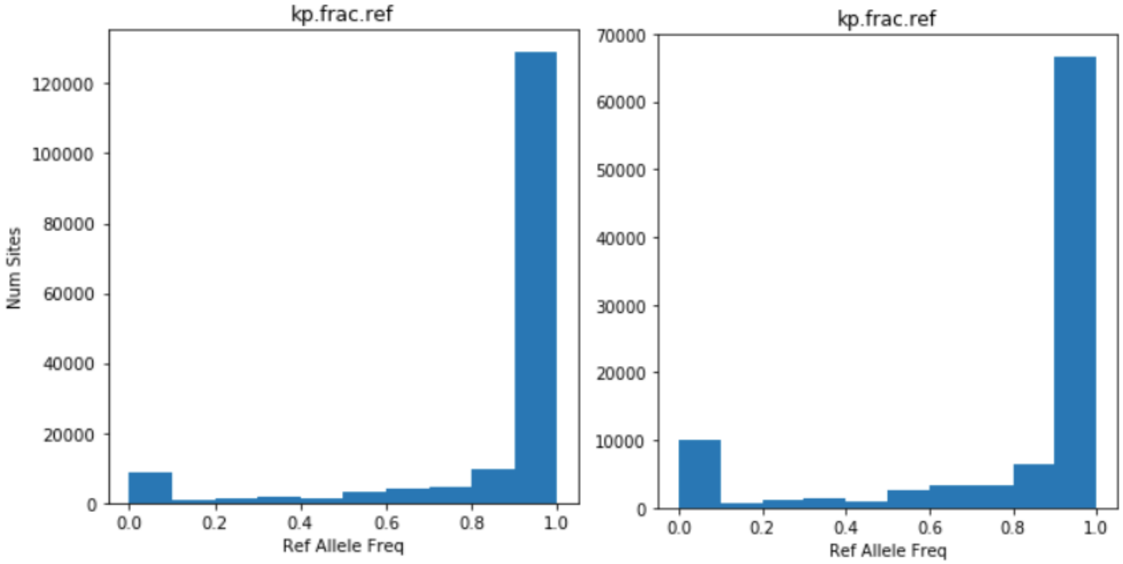


Fig. 2.4.a: Dog1 chrX.

Fig. 2.4.b: Dog2 chrX.

Figure 2.5: XCI profile for dog.

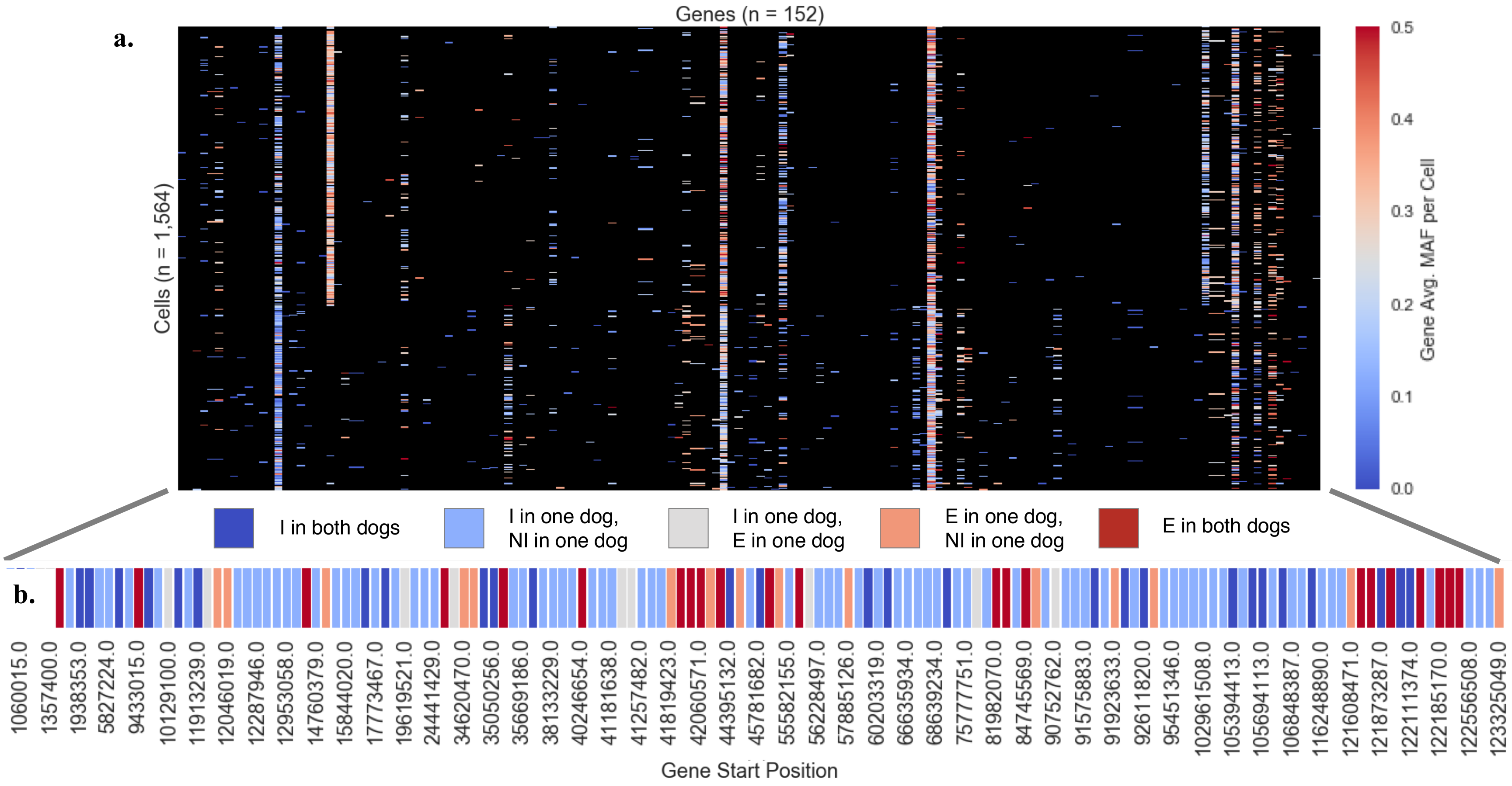


Fig. 2.5.a: Average minor allele frequency (color scale on right) across all ASE-informative sites in 152 chrX genes in 1,564 cells.
Fig. 2.5.b: Consensus XCI call for each of these 152 genes (color key above), with gene start positions on chrX. I = X-inactivated,   E = XCI escaper, NI = non-informative.

Chapter 3:

Figure 3.1: ChrX gene overlap among human, mouse, dog, and opossum.

Fig. 3.1.a: Overlap among 1:1 X orthologs.

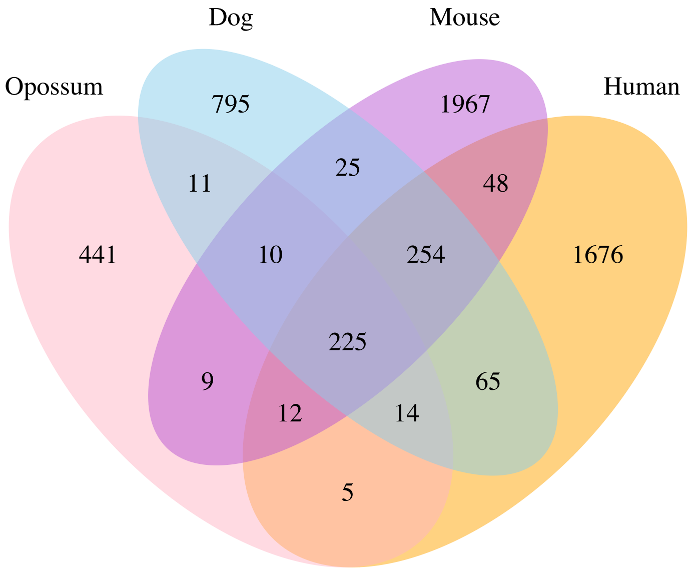


Fig. 3.1.b: Overlap among XCI escapers.

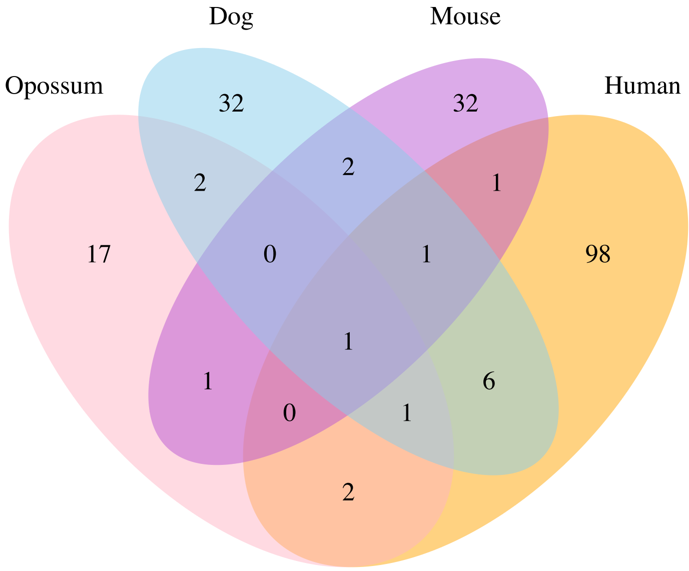


Fig. 3.1.c: Overlap among X-inactivated genes.

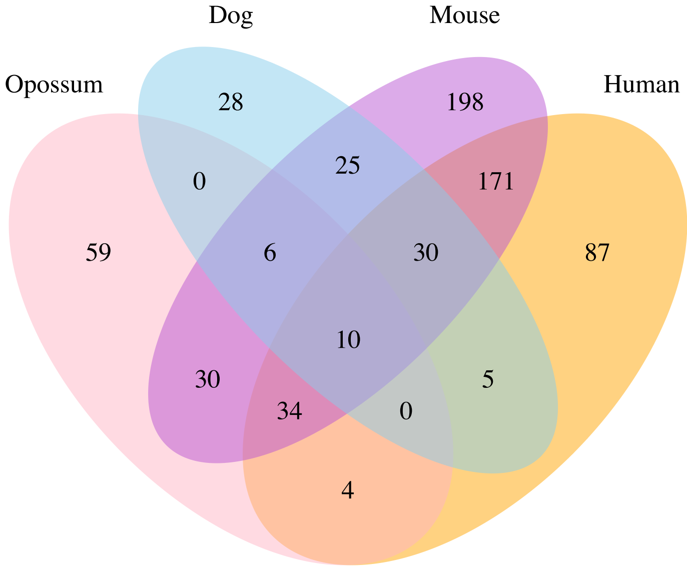


Figure 3.2: Inferred locations of XCI status changes in the opossum, dog, mouse, human phylogeny.

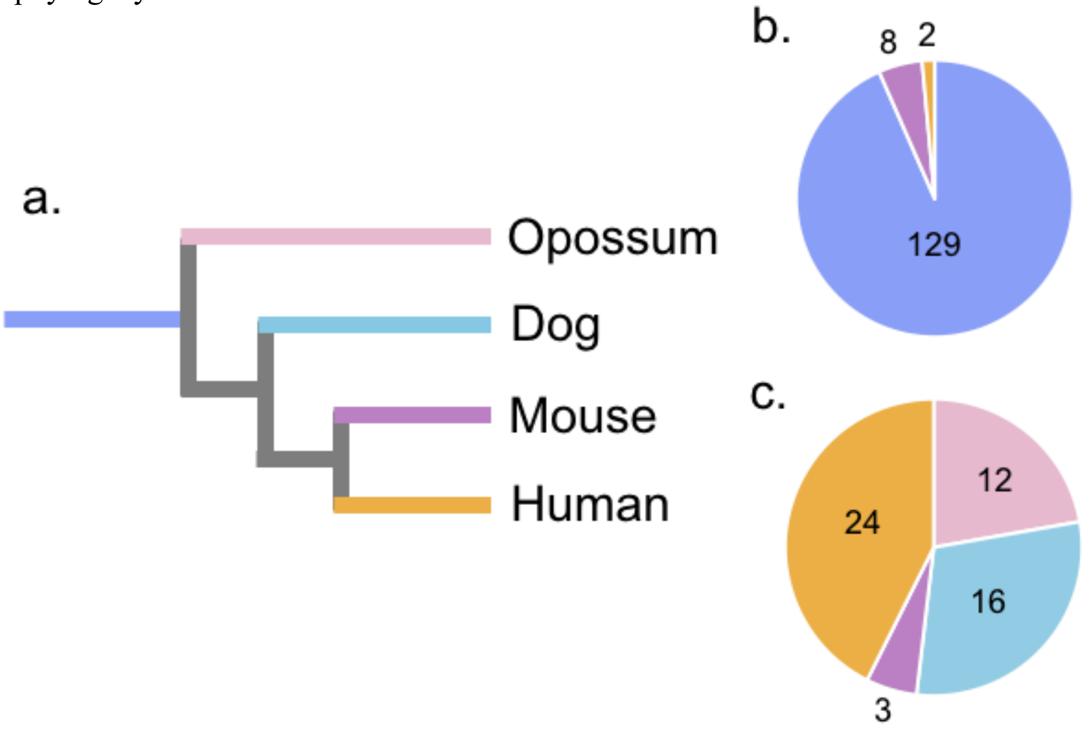


Fig. 3.2.a: Phylogeny for opossum, dog, mouse, and human.

Fig. 3.2.b: Number of genes changing from XCI escapers to X-inactivated genes, colored by the lineage in which the change occurred.

Fig. 3.2.c. Number of genes reverting from X-inactivated to escaping XCI, colored by lineage in which the change occurred.

Figure 3.3: PANTHER Biological Process terms for XCI escapers by species.

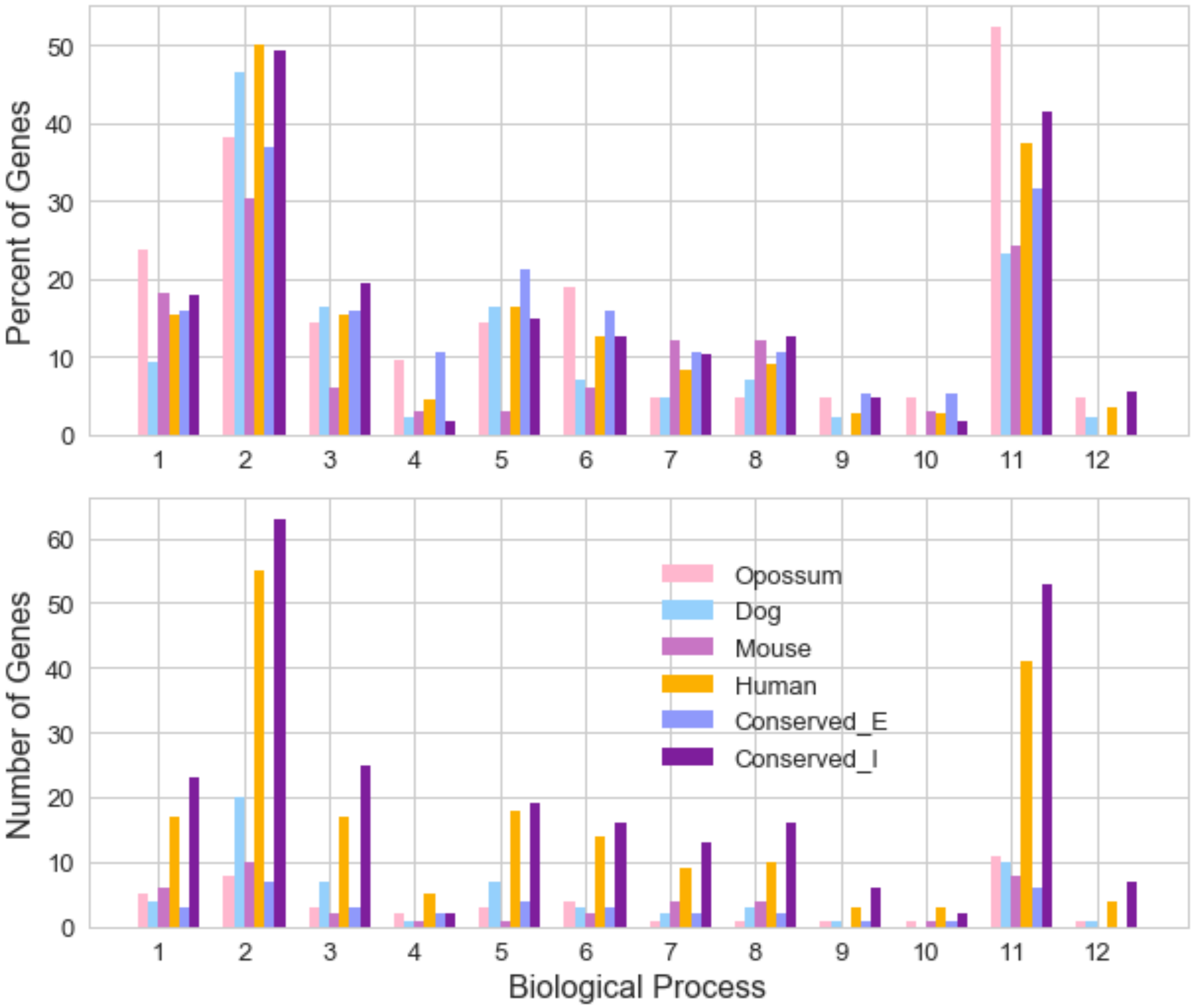


Fig. 3.3: Bars show the percentage (top) and number (bottom) of XCI escapers genes in each species and in the genes that escape XCI in two or more species (Conserved_E) or are X-inactivated in 2/3 or 3/4 species (Conserved_I) that have each of the twelve Gene Ontology Biological Process annotations observed across these species (right).

Biological Processes:

1	Cellular component organization/biogenesis
2	Cellular Process
3	Localization
4	Reproduction
5	Biological Regulation
6	Response to Stimulus
7	Developmental Process
8	Multicellular Organism Process
9	Biological Adhesion
10	Locomotion
11	Metabolic Process
12	Immune System Process

Figure 3.4: 100-way vertebrate phyloP scores by XCI status for each species.

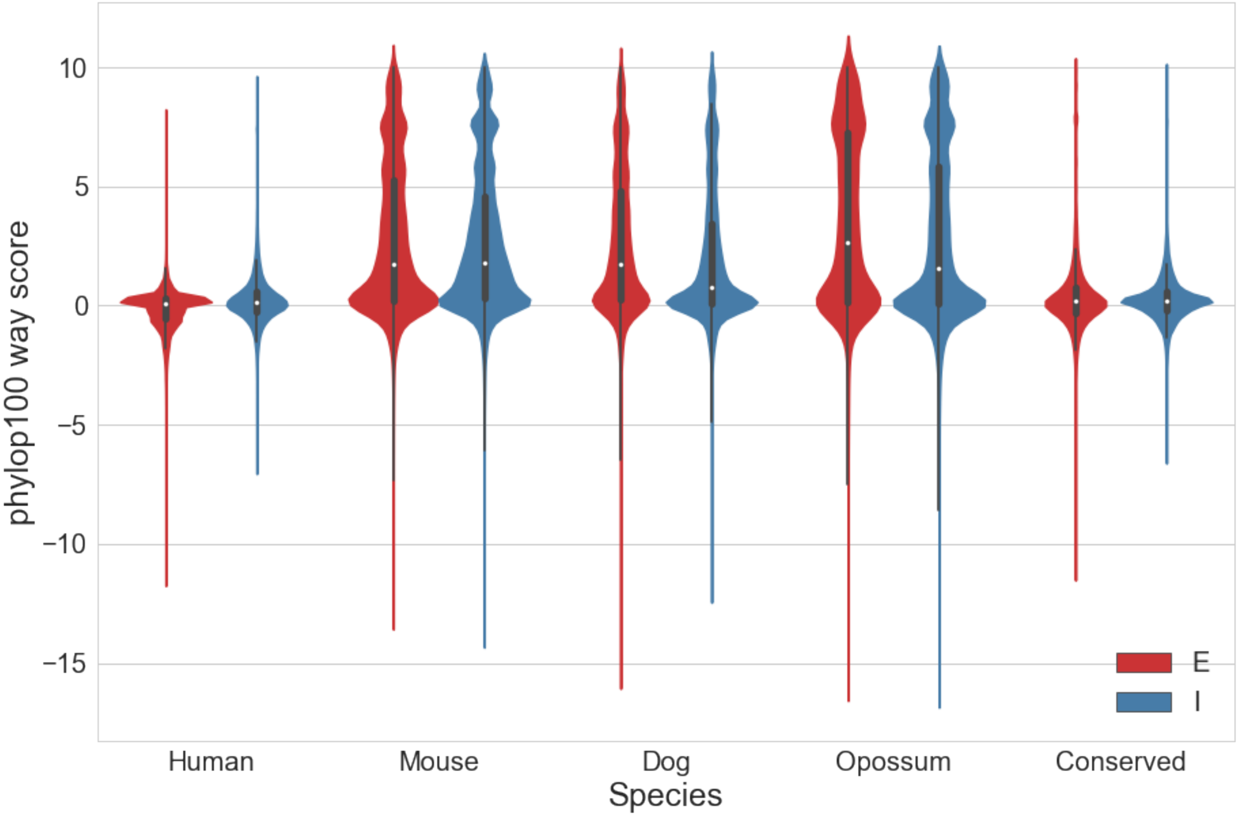


Fig. 3.4: Violin plots of phyloP 100-way (vertebrate) conservation scores for XCI escapers (“E”) and X-inactivated genes (“I”) in each species. The black rectangles inside each violin plot show the interquartile ranges, while the white points show the medians.

Table 3.1: XCI profile summaries for all species.

Species	# Informative X genes	# XCI escapers
Human	612	155/612 (25%)
Mouse	542	38/542 (7%)
Dog	143	45/143 (31%)
Opossum	176	24/176 (13%)

LIST OF SUPPLEMENTARY MATERIALS

Supplementary files for Chapter 1:

File S1.1: Summary of XCI status classification methods in three studies.

File S1.2: XCI status calls in three studies.

File S1.3: *S* and *N* values by XCI/Y status for all XCI status data sets.

File S1.4: Expression level for all genes by XCI/Y status, tissue, and sex.

File S1.5: Tissue-specificity index results for all genes by XCI/Y status and sex.

File S1.6: Percent of tissues with significant expression for all genes by sex.

File S1.7: MK test site counts for all genes.

File S1.8: *W* values by derived allele frequency.

File S1.9: Tissue-specific “3SD” expression level minima.

Supplementary files for Chapter 2:

File S2.1: RMarkdown file and input data for Seurat PBMC cell subtype analysis.

File S2.2: ASE analysis data.

File S2.3: Dog1 and Dog2 genotypes.

File S2.4: XCI status calls for canFam3.1 genes in both dogs.

Supplementary files for Chapter 3:

File S3.1: XCI profiles for all species.

File S3.2: Orthology data for all species.

File S3.3: GO annotation data for XCI escapers in each species and conserved escapers and X- inactivated genes.

File S3.4: hg38 genomic intervals for all species BLAT hits.

File S3.5: PhyloP100way data for all species.

REFERENCES

1. Agrelo, R., and Wutz, A. (2010). X inactivation and disease. Semin Cell Dev Biol 21, 194-200.

2. Agrelo, R., and Wutz, A. (2010). Context of change-X inactivation and disease. Embo Mol Med 2, 6-15.

3. Al Nadaf, S., Deakin, J.E., Gilbert, C., Robinson, T.J., Graves, J.A., and Waters, P.D. (2012). A cross-species comparison of escape from X inactivation in Eutheria: implications for evolution of X chromosome inactivation. Chromosoma 121, 71-78.

4. Barr, M.L., and Bertram, E.G. (1949). A Morphological Distinction between Neurones of the Male and Female, and the Behaviour of the Nucleolar Satellite during Accelerated Nucleoprotein Synthesis. Nature 163, 676-677.

5. Bell, R.J., Lees, G.E., and Murphy, K.E. (2008). X chromosome inactivation patterns in normal and X-linked hereditary nephropathy carrier dogs. Cytogenet Genome Res 122, 37-40.

6. Bellott, D.W., Hughes, J.F., Skaletsky, H., Brown, L.G., Pyntikova, T., Cho, T.J., Koutseva, N., Zaghlul, S., Graves, T., Rock, S., et al. (2014). Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. Nature 508, 494-+.

7. Belote, J.M., and Lucchesi, J.C. (1980). Male-specific lethal mutations of Drosophila melanogaster. Genetics 96, 165-186.

8. Belote, J.M., and Lucchesi, J.C. (1980). Control of X chromosome transcription by the maleless gene in Drosophila. Nature 285, 573-575.

9. Berletch, J.B., Ma, W., Yang, F., Shendure, J., Noble, W.S., Disteche, C.M., and Deng, X. (2015). Escape from X inactivation varies in mouse tissues. PLoS Genet 11, e1005079.

10. Berletch, J.B., Yang, F., Xu, J., Carrel, L., and Disteche, C.M. (2011). Genes that escape from X inactivation. Hum Genet 130, 237-245.

11. Bione, S., Sala, C., Manzini, C., Arrigo, G., Zuffardi, O., Banfi, S., Borsani, G., Jonveaux, P., Philippe, C., Zuccotti, M., et al. (1998). A human homologue of the Drosophila melanogaster diaphanous gene is disrupted in a patient with premature ovarian failure: Evidence for conserved function in oogenesis and implications for human sterility. Am J Hum Genet 62, 533-541.

12. Boyko, A.R., Quignon, P., Li, L., Schoenebeck, J.J., Degenhardt, J.D., Lohmueller, K.E., Zhao, K., Brisbin, A., Parker, H.G., vonHoldt, B.M., et al. (2010). A simple genetic architecture underlies morphological variation in dogs. PLoS Biol 8, e1000451.

13. Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csardi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., et al. (2011). The evolution of gene expression levels in mammalian organs. Nature 478, 343-348.

14. Brown, C.J., and Greally, J.M. (2003). A stain upon the silence: genes escaping X inactivation. Trends Genet 19, 432-438.

15. Busque, L., Mio, R., Mattioli, J., Brais, E., Blais, N., Lalonde, Y., Maragh, M., and Gilliland, D.G. (1996). Nonrandom X-inactivation patterns in normal females: lyonization ratios vary with age. Blood 88, 59-65.

16. Carrel, L., Cottle, A.A., Goglin, K.C., and Willard, H.F. (1999). A first-generation X-inactivation profile of the human X chromosome. Proc Natl Acad Sci U S A 96, 14440-14444.

17. Carrel, L., and Willard, H.F. (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. Proc Natl Acad Sci U S A 96, 7364-7369.

18. Carrell, D.T., Liu, L., Huang, I., and Peterson, C.M. (2005). Comparison of maturation, meiotic competence, and chromosome aneuploidy of oocytes derived from two protocols for in vitro culture of mouse secondary follicles. J Assist Reprod Genet 22, 347-354.

19. Chang, D., Gao, F., Slavney, A., Ma, L., Waldman, Y.Y., Sams, A.J., Billing-Ross, P., Madar, A., Spritz, R., and Keinan, A. (2014). Accounting for eXentricities: Analysis of the X Chromosome in GWAS Reveals X-Linked Genes Implicated in Autoimmune Diseases. Plos One 9.

20. Charlesworth, B. (1991). The Evolution of Sex-Chromosomes. Science 251, 1030-1033.

21. Chase, K., Carrier, D.R., Adler, F.R., Ostrander, E.A., and Lark, K.G. (2005). Interaction between the X chromosome and an autosome regulates size sexual dimorphism in Portuguese Water Dogs. Genome Res 15, 1820-1824.

22. Chaumeil, J., Waters, P.D., Koina, E., Gilbert, C., Robinson, T.J., and Graves, J.A. (2011). Evolution from XIST-independent to XIST-controlled X-chromosome inactivation: epigenetic modifications in distantly related mammals. Plos One 6, e19040.

23. Consortium, G.T. (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348, 648-660.

24. Cotton, A.M., Price, E.M., Jones, M.J., Balaton, B.P., Kobor, M.S., and Brown, C.J. (2015). Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation. Hum Mol Genet 24, 1528-1539.

25. de Napoles, M., Mermoud, J.E., Wakao, R., Tang, Y.A., Endoh, M., Appanah, R., Nesterova, T.B., Silva, J., Otte, A.P., Vidal, M., et al. (2004). Polycomb group proteins Ring1A/B link ubiquitylation of histone H2A to heritable gene silencing and X inactivation. Dev Cell 7, 663-676.

26. Deng, X., Berletch, J.B., Nguyen, D.K., and Disteche, C.M. (2014). X chromosome regulation: diverse patterns in development, tissues and disease. Nat Rev Genet 15, 367-378.

27. Deng, X., Ma, W., Ramani, V., Hill, A., Yang, F., Ay, F., Berletch, J.B., Blau, C.A., Shendure, J., Duan, Z., et al. (2015). Bipartite structure of the inactive mouse X chromosome. Genome Biol 16, 152.

28. Diehl, A.G., and Boyle, A.P. (2018). Conserved and species-specific transcription factor co-binding patterns drive divergent gene regulation in human and mouse. Nucleic Acids Res 46, 1878-1894.

29. Drummond, D.A., Bloom, J.D., Adami, C., Wilke, C.O., and Arnold, F.H. (2005). Why highly expressed proteins evolve slowly. P Natl Acad Sci USA 102, 14338-14343.

30. Dunford, A., Weinstock, D.M., Savova, V., Schumacher, S.E., Cleary, J.P., Yoda, A., Sullivan, T.J., Hess, J.M., Gimelbrant, A.A., Beroukhim, R., et al. (2017). Tumor-suppressor genes that escape from X-inactivation contribute to cancer sex bias. Nat Genet 49, 10-16.

31. Duret, L., and Mouchiroud, D. (2000). Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol Biol Evol 17, 68-74.

32. Ellison, J.W., Wardak, Z., Young, M.F., Robey, P.G., LaigWebster, M., and Chiong, W. (1997). PHOG, a candidate gene for involvement in the short stature of Turner syndrome. Human Molecular Genetics 6, 1341-1347.

33. Engreitz, J.M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C., Kadri, S., Xing, J., Goren, A., Lander, E.S., et al. (2013). The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. Science 341, 1237973.

34. Fan, G.P., and Tran, J. (2011). X chromosome inactivation in human and mouse pluripotent stem cells. Hum Genet 130, 217-222.

35. Federici, F., Mulugeta, E., Schoenmakers, S., Wassenaar, E., Hoogerbrugge, J.W., van der Heijden, G.W., van Cappellen, W.A., Slotman, J.A., van, I.W.F., Laven, J.S., et al. (2015). Incomplete meiotic sex chromosome inactivation in the domestic dog. BMC Genomics 16, 291.

36. Fu, W.Q., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Altshuler, D., Shendure, J., Nickerson, D.A., et al. (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493, 216-220.

37. Gao, F., and Keinan, A. (2014). High burden of private mutations due to explosive human population growth and purifying selection. BMC Genomics 15.

38. Giorgetti, L., Lajoie, B.R., Carter, A.C., Attia, M., Zhan, Y., Xu, J., Chen, C.J., Kaplan, N., Chang, H.Y., Heard, E., et al. (2016). Structural organization of the inactive X chromosome in the mouse. Nature 535, 575-579.

39. Goto, Y., and Takagi, N. (1998). Tetraploid embryos rescue embryonic lethality caused by an additional maternally inherited X chromosome in the mouse. Development 125, 3353-3363.

40. Goto, Y., and Takagi, N. (2000). Maternally inherited X chromosome is not inactivated in mouse blastocysts due to parental imprinting. Chromosome Res 8, 101-109.

41. Grant, J., Mahadevaiah, S.K., Khil, P., Sangrithi, M.N., Royo, H., Duckworth, J., McCarrey, J.R., VandeBerg, J.L., Renfree, M.B., Taylor, W., et al. (2012). Rsx is a metatherian RNA with Xist-like properties in X-chromosome inactivation. Nature 487, 254-258.

42. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., and McKusick, V.A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 33, D514-D517.

43. Hart, T., Komori, H.K., LaMere, S., Podshivalova, K., and Salomon, D.R. (2013). Finding the active genes in deep RNA-seq gene expression studies. BMC Genomics 14, 778.

44. Hore, T.A., Koina, E., Wakefield, M.J., and Graves, J.A.M. (2007). The region homologous to the X-chromosome inactivation centre has been disrupted in marsupial and monotreme mammals. Chromosome Research 15, 147-161.

45. Hsu, D.R., and Meyer, B.J. (1994). The dpy-30 gene encodes an essential component of the Caenorhabditis elegans dosage compensation machinery. Genetics 137, 999-1018.

46. Itoh, Y., Melamed, E., Yang, X., Kampf, K., Wang, S., Yehya, N., Van Nas, A., Replogle, K., Band, M.R., Clayton, D.F., et al. (2007). Dosage compensation is less effective in birds than in mammals. J Biol 6, 2.

47. Jegalian, K., and Page, D.C. (1998). A proposed path by which genes common to mammalian X and Y chromosomes evolve to become X inactivated. Nature 394, 776-780.

48. Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. (2004). The UCSC Table Browser data retrieval tool. Nucleic Acids Res 32, D493-496.

49. Keinan, A., and Clark, A.G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336, 740-743.

50. Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Res 12, 656-664.

51. King, M.C., and Wilson, A.C. (1975). Evolution at two levels in humans and chimpanzees. Science 188, 107-116.

52. Lahn, B.T., and Page, D.C. (1999). Four evolutionary strata on the human X chromosome. Science 286, 964-967.

53. Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326, 289-293.

54. Liu, S., and Trapnell, C. (2016). Single-cell transcriptome sequencing: recent advances and remaining challenges. F1000Res 5.

55. Luo, Z.X., Yuan, C.X., Meng, Q.J., and Ji, Q. (2011). A Jurassic eutherian mammal and divergence of marsupials and placentals. Nature 476, 442-445.

56. Lyon, M.F. (1961). Gene action in the X-chromosome of the mouse (Mus musculus L.). Nature 190, 372-373.

57. Lyon, M.F. (1962). Sex chromatin and gene action in the mammalian X-chromosome. Am J Hum Genet 14, 135-148.

58. Ma, L., Hoffman, G., and Keinan, A. (2015). X-inactivation informs variance-based testing for X-linked association of a quantitative trait. BMC Genomics 16.

59. Mackay, T.F., Richards, S., Stone, E.A., Barbadilla, A., Ayroles, J.F., Zhu, D., Casillas, S., Han, Y., Magwire, M.M., Cridland, J.M., et al. (2012). The Drosophila melanogaster Genetic Reference Panel. Nature 482, 173-178.

60. Mank, J.E., and Ellegren, H. (2009). All dosage compensation is local: Gene-by-gene regulation of sex-biased expression on the chicken Z chromosome. Heredity 102, 312-320.

61. Marks, H., Kerstens, H.H.D., Barakat, T.S., Splinter, E., Dirks, R.A.M., van Mierlo, G., Joshi, O., Wang, S.Y., Babak, T., Albers, C.A., et al. (2015). Dynamics of gene silencing during X inactivation using allele-specific RNA-seq. Genome Biol 16.

62. Marth, G.T., Czabarka, E., Murvai, J., and Sherry, S.T. (2004). The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. Genetics 166, 351-372.

63. McDonald, J.H., and Kreitman, M. (1991). Adaptive Protein Evolution at the Adh Locus in Drosophila. Nature 351, 652-654.

64. Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., and Thomas, P.D. (2017). PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res 45, D183-D189.

65. Migeon, B.R. (2014). Females are mosaics: X inactivation and sex differences in disease. (Oxford: Oxford University Press).

66. Miller, A.P., and Willard, H.F. (1998). Chromosomal basis of X chromosome inactivation: identification of a multigene domain in Xp11.21-p11.22 that escapes X inactivation. Proc Natl Acad Sci U S A 95, 8709-8714.

67. Morrow, E.H., and Connallon, T. (2013). Implications of sex-specific selection for the genetic basis of disease. Evol Appl 6, 1208-1217.

68. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5, 621-628.

69. Mukherjee, A.S., and Beermann, W. (1965). Synthesis of ribonucleic acid by the X-chromosomes of Drosophila melanogaster and the problem of dosage compensation. Nature 207, 785-786.

70. Naqvi, S., Bellott, D.W., Lin, K.S., and Page, D.C. (2018). Conserved microRNA targeting reveals preexisting gene dosage sensitivities that shaped amniote sex chromosome evolution. Genome Res 28, 474-483.

71. Nguyen, D.K., and Disteche, C.M. (2006). High expression of the mammalian X chromosome in brain. Brain Res 1126, 46-49.

72. Ober, C., Loisel, D.A., and Gilad, Y. (2008). Sex-specific genetic architecture of human disease. Nat Rev Genet 9, 911-922.

73. Ohno, S. (1967). Sex chromosomes and sex-linked genes. (Berlin, New York etc.: Springer-Verlag).

74. Pandey, R.S., Wilson Sayres, M.A., and Azad, R.K. (2013). Detecting evolutionary strata on the human x chromosome in the absence of gametologous y-linked sequences. Genome Biol Evol 5, 1863-1871.

75. Park, C., Carrel, L., and Makova, K.D. (2010). Strong Purifying Selection at Genes Escaping X Chromosome Inactivation. Molecular Biology and Evolution 27, 2446-2450.

76. Park, S.G., and Choi, S.S. (2010). Expression breadth and expression abundance behave differently in correlations with evolutionary rates. BMC Evol Biol 10, 241.

77. Patsopoulos, N.A., Tatsioni, A., and Ioannidis, J.P.A. (2007). Claims of sex differences - An empirical assessment in genetic associations. Jama-J Am Med Assoc 298, 880-893.

78. Peeters, S.B., Cotton, A.M., and Brown, C.J. (2014). Variable escape from X-chromosome inactivation: Identifying factors that tip the scales towards expression. Bioessays 36, 746-756.

79. Pessia, E., Makino, T., Bailly-Bechet, M., McLysaght, A., and Marais, G.A.B. (2012). Mammalian X chromosome inactivation evolved as a dosage-compensation mechanism for dosage-sensitive genes on the X chromosome. P Natl Acad Sci USA 109, 5346-5351.

80. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 20, 110-121.

81. Raudsepp, T., and Chowdhary, B.P. (2015). The Eutherian Pseudoautosomal Region. Cytogenet Genome Res 147, 81-94.

82. Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat Biotechnol 29, 24-26.

83. Ross, M.T., Grafham, D.V., Coffey, A.J., Scherer, S., McLay, K., Muzny, D., Platzer, M., Howell, G.R., Burrows, C., Bird, C.P., et al. (2005). The DNA sequence of the human X chromosome. Nature 434, 325-337.

84. Schultz, M.D., He, Y.P., Whitaker, J.W., Hariharan, M., Mukamel, E.A., Leung, D., Rajagopal, N., Nery, J.R., Urich, M.A., Chen, H.M., et al. (2015). Human body epigenome maps reveal noncanonical DNA methylation variation. Nature 523, 212-U189.

85. Shannon, L.M., Boyko, R.H., Castelhano, M., Corey, E., Hayward, J.J., McLean, C., White, M.E., Abi Said, M., Anita, B.A., Bondjengo, N.I., et al. (2015). Genetic structure in village dogs reveals a Central Asian domestication origin. Proc Natl Acad Sci U S A 112, 13639-13644.

86. Shen, Y., Matsuno, Y., Fouse, S.D., Rao, N., Root, S., Xu, R.H., Pellegrini, M., Riggs, A.D., and Fan, G.P. (2008). X-inactivation in female human embryonic stem cells is in a nonrandom pattern and prone to epigenetic alterations. P Natl Acad Sci USA 105, 4709-4714.

87. Shevchenko, A.I., Zakharova, I.S., and Zakian, S.M. (2013). The evolutionary pathway of x chromosome inactivation in mammals. Acta Naturae 5, 40-53.

88. Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P.J., Cordum, H.S., Hillier, L., Brown, L.G., Repping, S., Pyntikova, T., Ali, J., Bieri, T., et al. (2003). The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature 423, 825-U822.

89. Skuse, D.H. (2005). X-linked genes and mental functioning. Hum Mol Genet 14 Spec No 1, R27-32.

90. Splinter, E., de Wit, E., Nora, E.P., Klous, P., van de Werken, H.J., Zhu, Y., Kaaij, L.J., van Ijcken, W., Gribnau, J., Heard, E., et al. (2011). The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. Genes Dev 25, 1371-1383.

91. Talebizadeh, Z., Simon, S.D., and Butler, M.G. (2006). X chromosome gene expression in human tissues: male and female comparisons. Genomics 88, 675-681.

92. Trabzuni, D., Ramasamy, A., Imran, S., Walker, R., Smith, C., Weale, M.E., Hardy, J., Ryten, M., and North American Brain Expression, C. (2013). Widespread sex differences in gene expression and splicing in the adult human brain. Nat Commun 4, 2771.

93. Tukiainen, T., Pirinen, M., Sarin, A.P., Ladenvall, C., Kettunen, J., Lehtimaki, T., Lokki, M.L., Perola, M., Sinisalo, J., Vlachopoulou, E., et al. (2014). Chromosome X-Wide Association Study Identifies Loci for Fasting Insulin and Height and Evidence for Incomplete Dosage Compensation. Plos Genetics 10.

94. Tukiainen, T., Villani, A.C., Yen, A., Rivas, M.A., Marshall, J.L., Satija, R., Aguirre, M., Gauthier, L., Fleharty, M., Kirby, A., et al. (2017). Landscape of X chromosome inactivation across human tissues. Nature 550, 244-248.

95. Van der Maaten L, Hinton G. Visualizing Data using t-SNE. Journal of Machine Learning Research. 2008;9(11)

96. Viana, K.F., Aguiar-Soares, R.D., Roatt, B.M., Resende, L.A., da Silveira-Lemos, D., Correa-Oliveira, R., Martins-Filho, O.A., Moura, S.L., Zanini, M.S., Araujo, M.S., et al. (2013). Analysis using canine peripheral blood for establishing in vitro conditions for monocyte differentiation into macrophages for Leishmania chagasi infection and T-cell subset purification. Vet Parasitol 198, 62-71.

97. Veyrunes, F., Waters, P.D., Miethke, P., Rens, W., McMillan, D., Alsop, A.E., Grutzner, F., Deakin, J.E., Whittington, C.M., Schatzkamer, K., et al. (2008). Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. Genome Res 18, 965-973.

98. Vonholdt, B.M., Pollinger, J.P., Lohmueller, K.E., Han, E., Parker, H.G., Quignon, P., Degenhardt, J.D., Boyko, A.R., Earl, D.A., Auton, A., et al. (2010). Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. Nature 464, 898-902.

99. Wall, D.P., Hirsh, A.E., Fraser, H.B., Kumm, J., Giaever, G., Eisen, M.B., and Feldman, M.W. (2005). Functional genomic analysis of the rates of protein evolution. P Natl Acad Sci USA 102, 5483-5488.

100. Wallis, M.C., Delbridge, M.L., Pask, A.J., Alsop, A.E., Grutzner, F., O'Brien, P.C., Rens, W., Ferguson-Smith, M.A., and Graves, J.A. (2007). Mapping platypus SOX genes; autosomal location of SOX9 excludes it from sex determining role. Cytogenet Genome Res 116, 232-234.

101. Wang, J., Syrett, C.M., Kramer, M.C., Basu, A., Atchison, M.L., and Anguera, M.C. (2016). Unusual maintenance of X chromosome inactivation predisposes female lymphocytes for increased expression from the inactive X. Proc Natl Acad Sci U S A 113, E2029-2038.

102. Wang, X., Douglas, K.C., VandeBerg, J.L., Clark, A.G., and Samollow, P.B. (2014). Chromosome-wide profiling of X-chromosome inactivation and epigenetic states in fetal brain and placenta of the opossum, Monodelphis domestica. Genome Res 24, 70-83.

103. Wang, X., Miller, D.C., Clark, A.G., and Antczak, D.F. (2012). Random X inactivation in the mule and horse placenta. Genome Res 22, 1855-1863.

104. Wang, Z., Willard, H.F., Mukherjee, S., and Furey, T.S. (2006). Evidence of influence of genomic DNA sequence on human X chromosome inactivation. PLoS Comput Biol 2, e113.

105. Wible, J.R., Rougier, G.W., Novacek, M.J., and Asher, R.J. (2007). Cretaceous eutherians and Laurasian origin for placental mammals near the K/T boundary. Nature 447, 1003-1006.

106. Wilson Sayres, M.A., and Makova, K.D. (2013). Gene survival and death on the human Y chromosome. Mol Biol Evol 30, 781-787.

107. Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-Saban, S., Safran, M., Domany, E., et al. (2005). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics 21, 650-659.

108. Yang, F., Babak, T., Shendure, J., and Disteche, C.M. (2010). Global survey of escape from X inactivation by RNA-sequencing in mouse. Genome Res 20, 614-622.

109. Yen, Z.C., Meyer, I.M., Karalic, S., and Brown, C.J. (2007). A cross-species comparison of X-chromosome inactivation in Eutheria. Genomics 90, 453-463.

110. Zechner, U., Wilda, M., Kehrer-Sawatzki, H., Vogel, W., Fundele, R., and Hameister, H. (2001). A high density of X-linked genes for general cognitive ability: a run-away process shaping human evolution? Trends Genet 17, 697-701.

111. Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Giron, C.G., et al. (2018). Ensembl 2018. Nucleic Acids Res 46, D754-D761.

112. Zinn, A.R., and Ross, J.L. (1998). Turner syndrome and haploinsufficiency. Curr Opin Genet Dev 8, 322-327.