The Pennsylvania State University

The Graduate School

X-VIZ WEBTOOL: ANNOTATION AND VISUALIZATION OF GENES IN THE HUMAN X-CHROMOSOME BY THEIR ACTIVATION STATES

A Thesis in

Bioinformatics and Genomics

by

Karine Angel Moussa

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science

December 2021

The thesis of Karine Moussa was reviewed and approved by the following:

Dajiang Liu
Associate Professor of Public Health Sciences
Thesis Co-Advisor

Suzanne Gonzalez
Assistant Professor of Psychiatry and Biobehavioral Health
Thesis Co-Advisor

Laura Carrel
Associate Professor of Biochemistry and Molecular Biology

George Perry
Associate Professor of Anthropology and Biology
Chair, Intercollege Graduate Degree Program in Bioinformatics and Genomics

ABSTRACT

The human X chromosome carries >1000 genes with essential functions that are critical for development and disease. However, the X chromosome is severely under-studied compared to autosomes, in part due to the unique biology of X chromosome inactivation (XCI). XCI is an epigenetic process that silences one X chromosome in each female and balances gene dosage between sexes. Yet, ~12-35% of genes escape XCI in females and exhibit expression in both X chromosomes. Escape from XCI shows inter-individual differences, induces dosage imbalance, and may influence disease. Much progress by us and others has been made to understand XCI. Researchers have developed several methods to classify and predict XCI status, each with their strengths and limitations. Yet, there still lacks a comprehensive and convenient platform for users to explore different analyses results, annotate X-linked genes and associate them with diseases. We present X-Viz, an R Shiny app, to fill this gap. This application is the first genome browser to summarize and synthesize the present knowledge of XCI escape genes through interactive visualization tools.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This research was supported by National Institutes of Health (NIH) R01GM126479, the Lupus Research Alliance, and by CURE funds from the Pennsylvania Department of Health. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Institutes of Health or the Pennsylvania Department of Health.

We would like to thank Renan Sauteraud and Chachrit Khunsriraksakul from the Penn State College of Medicine for their help in data collection and synonymous trait mapping.

Chapter 1

Introduction

The human X chromosome contains many functionally important genes that influence developmental and disease processes. Despite its critical role, the X chromosome is severely understudied compared to autosomal genes. This is partly due to the difference in chromosome numbers between sexes and X chromosome inactivation (XCI), which silences one X chromosome in females to balance gene dosage between sexes. Moreover, around 12-20% of genes are expressed from the inactive X chromosome (Xi) and consistently "escape" inactivation, and an additional 10-15% of genes variably escape XCI in some, but not all, individuals [1]. These complexities make it challenging to analyze X chromosome data in functional genomics or disease association studies [2]. In recent years, researchers have made much progress in understanding XCI states through population-scale datasets [3-5]. There is a growing need to compare results across studies and consider the role of XCI in complex traits, particularly sex-biased traits. We deliver X-Viz, an interactive visualization platform that cross-references escape gene annotations and genome-wide association study (GWAS) results for X-linked genes.

To offer the most comprehensive XCI status annotations, X-Viz integrates XCI states inferred from several methods, including differential gene expression analysis (DGEA) [5, 6], allele-specific expression (ASE) [4], and gene promoter methylation and epigenetic marks [7, 8]. Annotations generated from these different approaches provide complementary information. For example, DGEA identifies genes that show differential

expression between sexes, a characteristic for escape genes, yet does not reveal whether a gene escapes XCI in individual samples. ASE-based analysis fills this gap and can be applied to bulk RNA-seq and single-cell RNA-Seq data to characterize inactive X expression levels across different individuals [9]. Bulk RNA-seq data are broadly available from many individuals, which help characterize inter-individual differences but have limited power for ASE analysis when the sample skewing, or the ratio of maternal and paternal inactive X selection [5], is balanced [3-5]. Single-cell RNA-seq data can complement bulk RNA-seq data in more balanced samples. Analyzing methylation or other repressive chromatin modifications has an advantage over ASE-based methods in samples with balanced skewing, although the correlation between these epigenetic marks and escape status is not perfect. Finally, skewing and XCI status can vary across cells and individuals and may explain some differences in XCI classifications across studies.

X-Viz consolidates the escape states of individual genes across multiple studies for quick meta-study comparisons of escape states. Our application allows users to link XCI escape annotations with GWAS associations to interpret the phenotypic consequence of escape genes and provides an interactive web interface to visualize results. Finally, our application allows users to upload custom XCI annotations for visualization and joint analysis.

Chapter 2

X Chromosome Inactivation and Phenotypic Outcomes

X Chromosome Inactivation

X Chromosome Inactivation (XCI) is the epigenetic process that silences one female X chromosome during early development [10]. The human sex chromosome pair typically comprises two X chromosomes in females and an XY pair in males. XCI in females results in a gene dosage balance between males and females to support homeostasis within females [11]. The inactive X chromosome (Xi) may be the maternally- or paternally-inherited chromosome. This results in random inactivation, or "mosaicism" [12]. Rare instances of non-random inactivation can occur in females where all inactive X chromosomes are inherited by one parental donor, resulting in a "fully skewed" individual [5]. The presence of a second variably expressed X chromosome in females strongly implicates a causal role of X-linked genes in sex-biased traits [1]. XCI research offers valuable insight into eukaryotic gene regulation and the genetic mechanisms that influence sex-biased phenotypes.

Several genetic and epigenetic factors influence XCI regulation. The X inactivation center (XIC) on the X chromosome houses many genes involved in initiating inactivation [13]. The *XIST* gene (X-inactive specific transcript) located in XIC transcribes for long non-coding RNA that coats the chromosome of origin to achieve transcriptional silencing [13]. *XIST* RNA recruits factors that modify chromatin structure and package Xi DNA into heterochromatin, reducing its transcription accessibility [11]. The *XIST*-recruited factors

include polycomb group proteins (PcG) that form complexes to repress gene expression [13], specialized histone variant macroH2A which forms dense macrochromatin bodies on Xi [14], and concomitant trithorax group protein (*ASH21*) and heterogeneous nuclear ribonucleoprotein (*SAF-1*) which are thought to stabilize inactivation by preparing Xi chromatin for protein interactions that suppress expression [15]. During *XIST* coating, active histone marks such as H3K4me2/me3 and H3/H4 acetylation are removed from Xi by preloaded histone deacetylase, such as HDAC3 [16], while PcG proteins PRC1 and PRC2 deposit repressive histone marks H3K27me3 and H2A119b, respectively [17, 18]. Figure 2-1 models the *XIST* coating process and histone recruitment/removal during XCI.

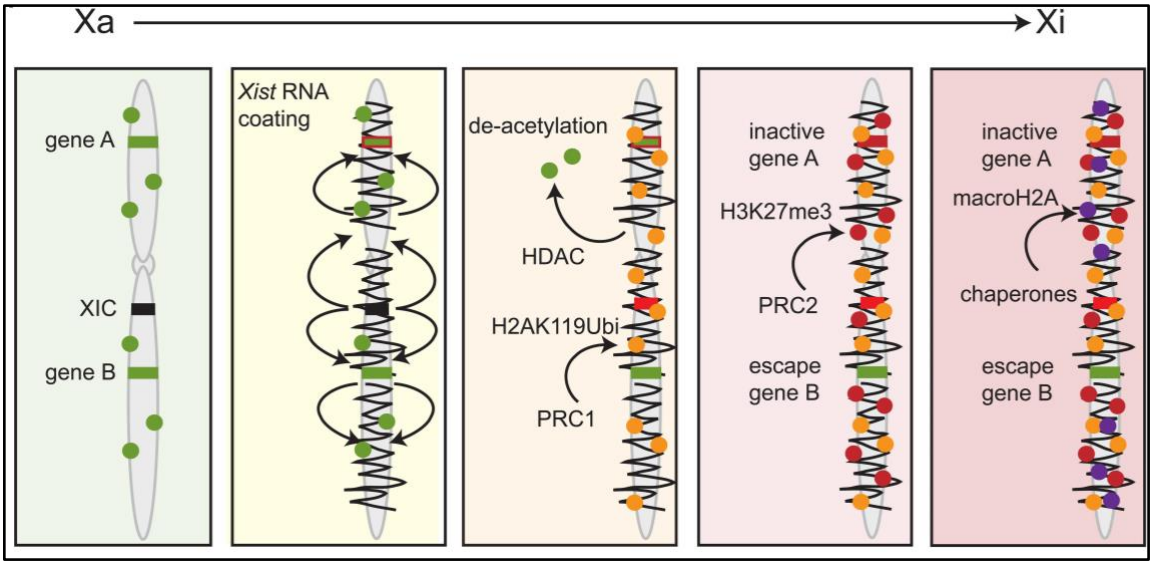


Figure 2-1 *XIST* coating and histone modifications during XCI; Image from Fang et al. [18]; Active histone marks (green dots) are deacetylated by HDAC3, while repressive marks are mediated by PRC1 and PRC2, with specialized macroH2A modifications mediated by *XIST* transcripts. Gene A represents a gene subject to inactivation, and Gene B represents a gene that escapes inactivation.

While *XIST* transcriptionally silences Xi regions, it is not the sole regulating factor of XCI [19]. The 3-dimensional proximity of Xi to nucleolus boundaries, mediated by *FIRRE* (functional intergenic repeating RNA element) RNA [18], supports stable inactivation by stationing Xi heterochromatin at the nuclear membrane [20]. Long interspersed nuclear element-1 (L1) repeat elements are commonly associated with regional inactivation. L1s in the genome can alter genetic function when activated by breaking DNA sequences, altering splicing, and suppressing gene transcription [21]. L1s are more abundant in Xi regions subject to inactivation than in regions that are incompletely silenced [22]. L1s' involvement in XCI may be due to the late onset of L1 expression during silencing [23]; non-expressed L1s are also implicated in XCI by aiding heterochromatin packaging to induce silencing [23]. XCI is a complex biological process that involves several carefully mediated steps and elements to initiate and maintain proper silencing.

Incomplete XCI

X inactivation suppresses a majority of the X chromosome; however, partial activation occurs in a subset of "escape" genes. In humans, 12-20% of genes consistently escape inactivation across all females [1]. Genes that escape inactivation exhibit lower expression levels than their Xa counterparts [24]. An additional 10-15% of genes variably escape inactivation, expressing themselves in some, but not all, tissues or individuals [1]. Much research is required to identify the full set of elements that influence incomplete XCI.

Similar to X inactivation, researchers posit that several epigenetic and genetic factors are associated with XCI state. Regions that maintain inactivation are enriched for repressive histone marks H3K9me3 [25], H4K20me3 [26], and H3K27me3 [25] and are depleted for the active mark H3K4me3 [27], whereas escape gene regions retain active marks H3K4me2 [26], H3K4me3 [27], H3K9ac [26], H3K27ac [3, 28], and H3K9me1 [26] and lack the repressive mark H3K27me3 [27, 29]. DNA methylation levels are more abundant in transcriptional start sites of genes subject to inactivation than in genes which escape inactivation in females [8]. Many genetic factors are also implicated in Xi reactivation. CCCTC-binding factor (CTCF) is an RNA-binding and DNA-binding protein that influences DNA methylation, gene interactions, and chromatin structure [30]. Several studies implicate CTCF in Xi reactivation. CTCF is a known XCI boundary element, delineating regions of opposite activation status [30-33]. Topologically associated domains (TAD) are self-interacting genomic regions delineated by CTCF binding sites and cohesion rings [34]. TAD structures in the X chromosome typically coincide with XCI regions [35],

and locations surrounding CTCF boundaries of TAD structures harbor more variability in escape status across studies [36]. CTCF is also enriched at transcriptional start sites (TSS) of genes that escape inactivation. The presence of CTCF in TSS regions may biomechanically influence escape by serving as an obstacle to DNA methylation and heterochromatic structuring [17]. Another reported escape influencer is a combination of genomic distance to *XIST* and 3D chromatin structures, which affect reactivation timing and reactivation speed along the X chromosome [35, 37]. While each component's level of impact has yet to be fully understood, these discoveries piece together our understanding of Xi reactivation and inform the methods used to predict XCI status.

Sex-biased Phenotypic Outcomes of Incomplete XCI

There is growing evidence that supports the role of inter-individual XCI states in sex-biased traits. X chromosome genes are enriched for several functions implicated in cancer and immune disorders with a strong female bias such as breast cancer (100:1 white female:male, 70:1 black female:male) [38, 39], rheumatoid arthritis (3:1 female:male) [40], and systemic lupus erythematosus (SLE) (9:1 female:male) [41]. Abnormalities in X chromosome count increase the risk of developing certain autoimmune disorders, implicating over-expression of X-linked genes in autoimmune disease pathogenesis. For example, 47,XXX females develop SLE at higher frequencies than 46,XX females [42], and 47,XXY males develop SLE at higher frequencies than 46,XY males [43]. Recent research links SLE risk with aberrant Xi expression in human T cells and B cells [44, 45]. Toll like receptor 7 gene (*TLR7*) encodes for protein products which mediate pathogen

recognition and innate immune response [46]; increased transcription of *TLR7* is implicated in immune dysregulation and can escape XCI in female B lymphocytes, monocytes, and plasmacytoid dendritic cells [47, 48]. Variably escaping genes are another important class of X genes that may contribute to differences in phenotypic outcomes which vary across individuals and populations, such as autoimmune risk [49]. Chromosomal aneuploidies implicate the role of X chromosome expression in autoimmune disorders, while Xi escape and variable escape genes offer insight into female-biased disorders that are linked to XCI irregularities.

Understanding the role of XCI in sex-biased disease regulation is an important factor for targeted drug development. In female human and mice SLE patients, the absence of *XIST* RNA coating and heterochromatic modifications in mature T cells alters the inactive profile of Xi, allowing more genes to escape inactivation [44]. T cells are involved in several immune response functions, including killing infected cells and activating antibody production through B cell differentiation [50]. Aberrant T cell function can lead to over-reactive antibody response and drive autoimmune pathogenesis, as is the case in SLE [51]. Synthetic approaches to maintaining or reinstating the processes of *XIST* coating and histone modifications in Xi within T cells would likely disrupt certain disease pathways that are involved in autoimmune dysregulation. Targeted solutions for sex-based diseases are not necessarily limited to maintaining X inactivation but could also be designed to reactivate commonly inactive genes to mitigate the effects of deleterious mutations. For example, Rett Syndrome is a female-biased neurological disease influenced by a heterozygous mutation in *MECP2* [52], a gene of the X chromosome that encodes for a methyl-CpG-binding protein necessary for nerve cell development and communication

[53]. *MECP2* is commonly inactive on Xi; however, the reactivation of *Mecp2* in mouse models reverses the effects of the neurological deficits caused by Rett Syndrome [54]. Characterizing X inactive and escape genes and establishing their role in disease regulation will substantially inform therapeutic solutions for sexually dimorphic disorders.

Though it has yet to be sufficiently studied, some evidence may support the role of Xi escape genes in guarding females against male-biased diseases. Several loss-of-function mutations that affect male-biased cancer development are also commonly expressed on the inactive X in females [55]; among these are *DDX3X* which regulates cancer proliferation and tumorigenesis [56], *KDM5C* and *KDM6A* which encode for histone demethylases and are associated in tumor progression [57], and *MAGEC3* which encodes for a tumor-specific protein in T lymphocytes [58] and is associated in cancer development caused by paternally inherited mutations [59]. Females also exhibit lower susceptibility to several X-linked neurological disorders [60, 61]. Further analysis on male-biased X-linked disorders is required to identify whether specific X genes escape inactivation to protect females against male-biased diseases. *Chapter 5: Future Steps for XCI research and tools* elaborates on how researchers can investigate this hypothesis.

Chapter 3

Methods Used to Classify XCI Status

The prospective role of XCI reactivation in sex-biased traits highlights the value of identifying commonly escaping or suppressed X genes. Researchers have developed several experimental and computational approaches to both classify and predict escape status. One of the premiere X escape profiles was created by measuring gene expression in human/rodent somatic hybrid cell lines [24], comprised of X chromosomes derived from human skin fibroblasts cell lines that were retained in mutant murine lines [62]. Human-mice somatic cell hybridization allows researchers to isolate X chromosomes based upon whether they are active (Xa) or inactive (Xi) [62]. The researchers assayed nine Xa and Xi hybrid lines to determine whether each gene was solely expressed in Xa lines or in both Xa and Xi lines, indicating escape status. Later methods sampled populations with varying degrees of skew to measure allelic imbalance (AI) and epigenetic marks [3, 7]. As more methods develop, the results of each approach add to the present understanding of XCI reactivation by leveraging different components of the XCI process.

Differential gene expression analysis (DGEA) is a commonly used technique for evaluating gene expression between conditions and can be designed for XCI classification. Comparing X chromosome expression across female and male samples returns genes with more or less expression in one sex. Genes with higher expression in females are potential escape candidates since X expression should theoretically be the same across both sexes if Xi is completely inactive in females. One of the most comprehensive sex-stratified DGEA studies was performed by Oliva et al. 2020 [6] which integrated tissue-specific differential

expression, an important feature for evaluating X-linked diseases. The analysis highlighted how tissue-specific sex-biased expression affects several biological functions, including hormonal, cancer, and immune functions. DGEA is limited in its ability to measure the precise level of escape expression at an individual basis. In addition, DGEA cannot classify escape status at an individual level. While its population-level observations limit DGEA, sex-differential expression is an accessible approach that offers broad insight into the Xi escape landscape.

Allele-specific expression (ASE) methods can determine XCI status from expression data where skew ratio and heterozygous SNPs are available. The skew ratio represents the percentage of samples for which one paternally inherited chromosome is inactivated across the total number of samples [5]. One common method to calculate sample skewing is to measure the allele-specific expression of known silenced genes, which represents the expression of the one inherited chromosome over the total expression of both X chromosomes [4]. ASE classification operates on the basis that genes that escape inactivation in a set of cells exhibit an expression ratio, $\frac{REF}{ALT+REF}$, that differs from the skew ratio of the sample. One clear example of this is in a fully skewed individual, where the skew ratio is 0:100 (either Xmaternal:Xpaternal or Xpaternal:Xmaternal). In this sample, the expression ratio for silenced genes should reflect the skew ratio 0:100, whereas escape genes should differ from 0:100 due to their expression on Xi. Highly skewed samples of >25:75 support statically stronger classifications than those centering around 50:50, which most naturally occurs in females. For this reason, researchers may design cultured cell and hybrid cell experiments to simulate higher skew values [24]. XCI classifications are more

reliable in cultured cell lines than in hybrid cells; however, primary cell cultures harbor unstable epigenetic modifications [63], resulting in false escape calls. ASE is a useful approach to classify XCI states from expression data, provided that samples exhibit reasonable skewing and SNP-level information is available.

ASE methods require a panel of known heterozygous SNPs, which is not always available for each X chromosome gene. DNA methylation and epigenetic marks inform alternative methods that do not rely on skewing or heterozygous allele abundance. During silencing, *XIST* RNA transcripts coat the inactive X and recruit factors that modify 3D chromatin structure and histone marks [37]. These biomarkers are generally strong predictors of XCI status [17, 28]. For example, there are lower levels of DNA methylation (DNAme) in Xa and Xi regions for genes that escape inactivation in females, while there are intermediate or higher levels of DNAme in inactive regions [7]. As mentioned previously, active histone marks (H3K4me2/3, H3K9ac, H3K27ac, H3K9me1) are enriched in genes that escape inactivation, while suppressive marks (H3K9me3, H4K20me3, H3K27me3, macroH2A) are enriched in genes that are subject to XCI. Prediction models are developed by measuring DNAme and histone marker enrichment in females and fitting the model to previously known XCI states or XCI states obtained through ASE calls. Epigenetic prediction models can exceed 75% and 90% accuracy for predicting escape and subject status, respectively [7]. However, epigenetic prediction models are highly sensitive to low read depth, which can result in several false escape calls [7]. Table 3-1compares the utility of DGEA, ASE, and epigenetic prediction models in classifying XCI states. While each method has strengths and weaknesses, the results of these approaches collectively comprise our present knowledge of the XCI profile.

Table 3-1: Summarization of the strengths and limitations of different XCI classification methods.

Method	**Strengths**	**Limitations**
Differential Gene Expression Analysis (DGEA)	A straightforward, uncomplicated approach to identify X-inactive (Xi) escape candidates Requires only X-linked RNA expression data of male and female participants and comfortability using DGEA tools, such as *limma* [64] or *DESEQ* [65]	Requires both female and male expression data Female-biased returns are strong candidates, but are not conclusive escape genes Xi expression ratio information is lost (the ratio of Xi gene expression over total expression across both X chromosomes)
Allele-Specific Expression	Returns Xi expression ratio for each gene Can be applied to bulk or single-cell RNA-seq data	Can only classify X genes with heterozygous single nucleotide polymorphisms (SNPs) Robust calls require a high degree of sample skewing
DNA methylation/Epigenetic prediction models	Does not rely on skewing ratio or heterozygous alleles to make inferences Uses known escape influencers (chromatin modifiers) to inform escape status	Requires both female and male epigenetic data Highly sensitive to low read depth DNA methylation is highly tissue-specific [66]; prediction models derived from one cell type will not perform consistently in other cell types

Chapter 4

X-Viz Application

I developed the X-Viz application to summarize the escape profiles of several key XCI studies and contextualize the results with X-linked epidemiological events. X-Viz allows researchers to observe XCI states and cross-reference XCI states with complex traits and their sex ratios. XCI states and the extent of inter-individual variation can be examined for ten data sets across seven published studies of XCI escape genes [3-5, 7-9, 24]. These studies represent a variety of methods, including DGEA, DNA methylation, and allele-specific ratio. Several sample types are also represented, including lymphoblast cell lines, fibroblast cell lines, hybrid cell lines, and multiple human tissues. To date, no genome browsers or gene ontologies include escape classifications of X genes in their descriptions. The cross-study analysis of XCI states allows for a more complete understanding of X chromosome expression and highlights the biological and technical factors that affect escape classifications.

The X-Viz application was developed with R's shiny library and is hosted on https://liugroupstatgen.shinyapps.io/xci-app-1/. The application source code is available at https://github.com/Karine-Moussa/xci-app-1. *Appendix A* contains descriptions of each data set included in the application.

Explore XCI Classifications

X-Viz provides many useful features to jointly visualize XCI states and their phenotypic relevance. Users may search individual studies to collect detailed results, including a table of XCI states and a graphical display of gene locations on the X chromosome. Users can focus on selected genes within the study and return their escape state, mapped GWAS traits, and sex ratios for the cases based on UK Biobank estimates. All genes include a hyperlink to their corresponding ENSEMBL entry to shed light on surrounding genetic marks, such as CTCF binding sites. Figure 4-1: illustrates the visual and tabulated results for a single study outlined in the application. The ability to consolidate XCI research studies will facilitate hypothesis generation for X-linked genetic studies.

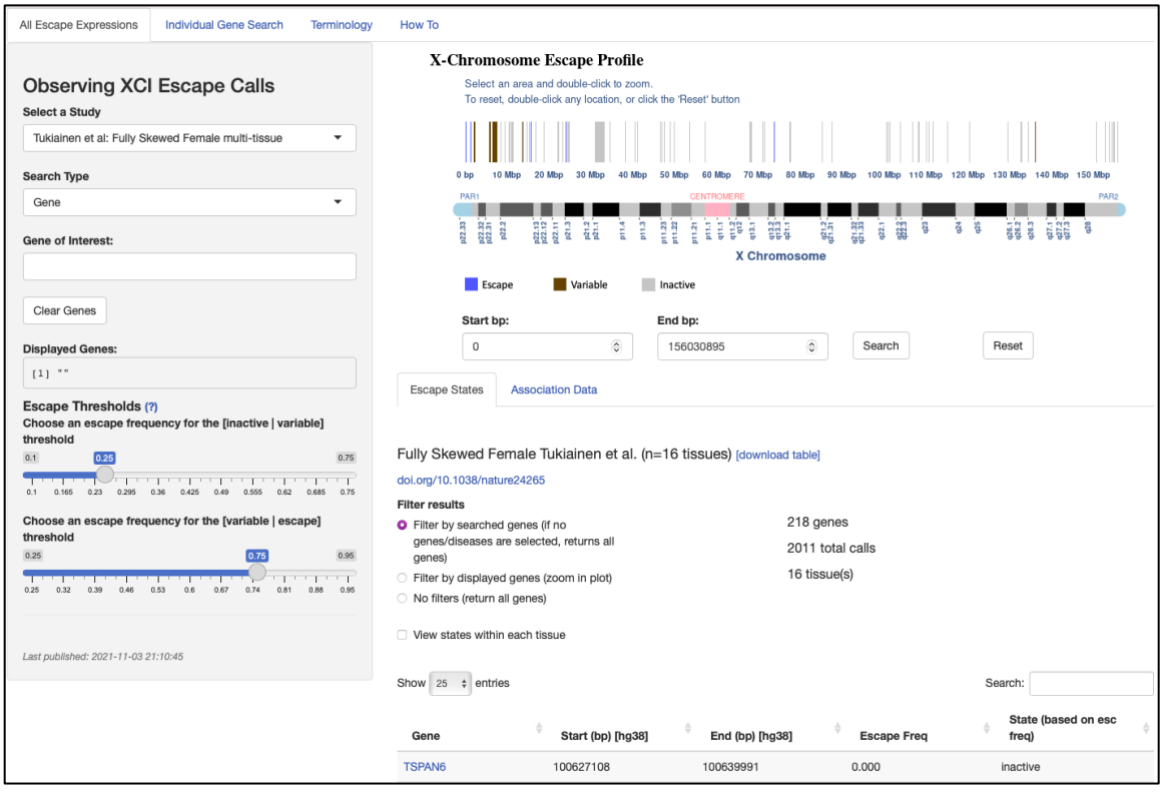


Figure 4-1: X-chromosome graphic and tabulated XCI states for a highly skewed female from the Genotype Tissue Expression (GTEx) project [5].

Cross-Referencing Epidemiological Events

To address the epidemiological implications of incomplete XCI, X-Viz returns information on X-linked genes associated with sex-biased traits and diseases. The multi-gene search engine returns genome-wide association studies (GWAS) hits for mapped traits and UK Biobank sex-ratios for many self-reported traits. Many X-linked genes have been implicated in female-biased autoimmune disorders, including *MECP2* [49], *IRAK1* [49], and *CXORF21* [67]. Users may filter their search for female-biased traits by selecting the "All Female Biased Traits" option in the disease search engine. Figure 4-2 displays the GWAS hits for *CXORF21* in the 2013 Cotton et al. lymphoblast and fibroblast study.

Figure 4-2: Genome-wide association studies (GWAS) hits for chromosome X open reading frame 21 gene (*CXORF21*); 2013 Cotton et al. lymphoblast and fibroblast study.

Characterizing Study-Specific Escape States through Escape Frequencies

X-Viz supports the ability to evaluate a study's escape classifications by adjusting escape frequency thresholds, an attribute of XCI research that is defined per study. The overall escape status – inactive, variable, or escape – for genes in multi-sample or multi-tissue data is typically determined by comparing the sample-specific escape frequency to a pre-specified escape frequency threshold. Escape frequency thresholds are subjectively selected per study and may be modified to account for study design. The included escape threshold sliders facilitate this modularity for studies that have escape frequency information. Adjusting the escape frequency thresholds will update the escape classifications for the selected study. A tutorial and an example of this feature are included in Appendix B of this report and in the "How To" page of the application.

Observing Individual Escape Statuses across Multiple Studies

XCI classifications can vary across studies due to several factors apart from authentic genetic expression. Some variability is attributed to the different cell types [5] or statistical methods [7, 8] used to classify XCI status. Another source of variability is the expression level of variably escaping genes on the inactive X chromosome [1]. Current methods estimate that Xi gene expression is around 33% of its Xa counterpart [1], while other methods assume 10% [3] to 100% [68] relative expression. Escape expression can also vary across tissues [6] and populations [69]. The inter- and intra-variability of XCI escape highlights the importance of distinguishing escape states across studies spanning various methods, tissue types, and populations.

X-Viz allows researchers to observe escape statuses across several studies through the Individual Gene Search engine. The search accepts a list of genes and returns a summary of the escape calls made by the included studies. This summary also includes the XCI statuses of user-uploaded data sets. Figure 4-3 displays the cross-study escape status of Lysine Demethylase 6A (*KDM6A*), a well-known escape gene. The cross-study analysis of escape states will provide a more complete understanding of commonly escaping genes and highlight the biological and technical factors that affect escape classifications.

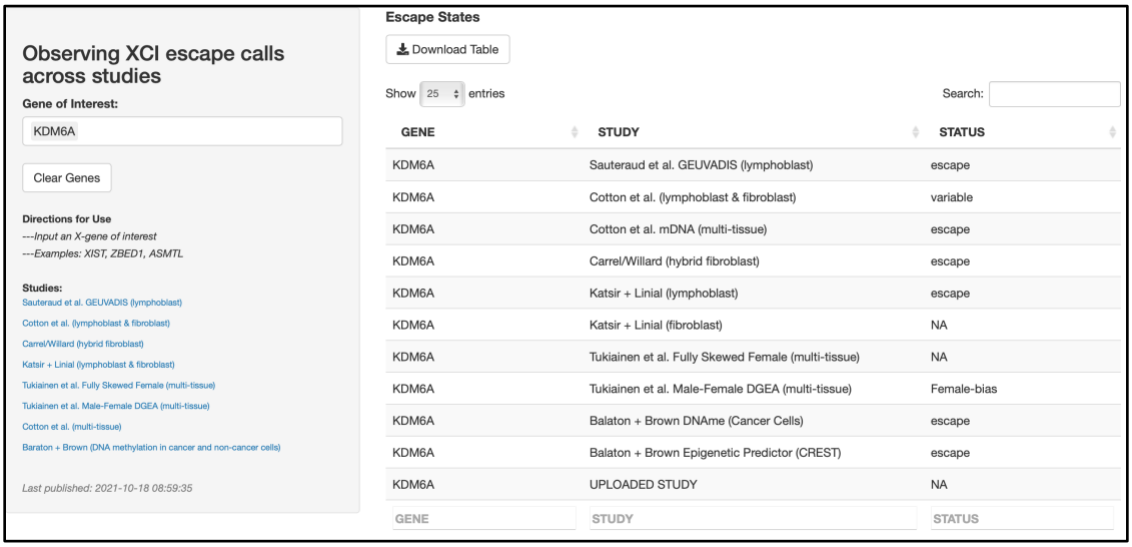


Figure 4-3: Summary of Lysine Demethylase 6A (*KDM6A*) states across XCI research.

Manually Upload Results

Users can visualize research results from their own studies by uploading datasets in comma-delimited format. The minimum information required are gene name, XCI state, and start and end position; users may optionally include tissue or sample descriptions. X-Viz will calculate the sample-specific escape frequencies and cross-reference the disease association database. The gene position and XCI states are also visualized on the scaled X

chromosome to highlight genomic context and nearby genes. The “How To” tab of the application includes detailed information for uploading studies.

Comparison to Existing Tools

X-Viz is designed to complement existing visualization platforms for genetic studies. For example, PheWeb is a popular tool that provides large-scale GWAS and PheWAS result visualizations [70]. However, PheWeb does not include X chromosomal genes, which limits its utility for investigating X-linked traits. Figure 4-4 compares the PheWeb and X-Viz results for the female-biased disease, systemic lupus erythematosus. X-Viz highlights X-linked GWAS traits and their sex biases and is the first utility to additionally annotate XCI escape states to our knowledge. Our application addresses the paucity of X chromosome data in bioinformatics tools and will be helpful in sex-biased trait research.

a

Category	Phenotype	Top variant in locus	P-value	MAF	Nearest Gene(s)
dermatologic	Lupus (localized and systemic)	6:32,631,348 G / A (rs9274253)	6.0e-14	0.091	*HLA-DQB1*
dermatologic	Systemic lupus erythematosus	6:32,080,191 G / C (rs1269852)	6.7e-13	0.13	*TNXB*
dermatologic	Systemic lupus erythematosus	6:25,769,349 T / C (rs1892251)	1.2e-10	0.13	*SLC17A4*

b **GWAS Catalog Search (Disease/Trait)**

Searching *All Assocations v1.02*

Show 25 entries Search:

Date	Mapped Gene	Disease/Trait	Link	UK Bio Desc.	Ratio (f/m)	Bias
9/23/16	CXorf21	systemic lupus erythematosus	www.ncbi.nlm.nih.gov/pubmed/26502338	systemic lupus erythematosis/sle	7.39622641509434	Female
9/23/16	NR0B1 - CXorf21	systemic lupus erythematosus	www.ncbi.nlm.nih.gov/pubmed/26502338	systemic lupus erythematosis/sle	7.39622641509434	Female
10/14/16	IRAK1	systemic lupus erythematosus	www.ncbi.nlm.nih.gov/pubmed/26606652	systemic lupus erythematosis/sle	7.39622641509434	Female
2/9/17	TMEM187 - IRAK1	systemic lupus erythematosus	www.ncbi.nlm.nih.gov/pubmed/26663301	systemic lupus erythematosis/sle	7.39622641509434	Female

Figure 4-4: Comparison of PheWeb and X-Viz returns for the female-biased disease, systemic lupus erythematosus. (a) The PheWeb search is limited to autosomal genes. (b) X-Viz returns GWAS X-linked genes and UK Biobank sex-ratios for the disease of interest.

Scientific Communication

We offer a starting point to consolidate standard XCI terms through the terminology page of the application. XCI research is accelerating as more advanced bioinformatic tools are developed, and the growing implications of X-linked expression in sex-biased disorders are understood. The definitions included in our application are based upon the collection of concepts defined across XCI research studies. The terminology section includes descriptive images and an interactive feature to visualize a key concept in XCI research, Xi genetic expression [3], for the 2021 GEUVADIS study [71]. Figure 4-5 presents a preview of the terminology page. Together, these resources offer a comprehensive educational tool for X chromosome biology and are nicely integrated with features of the browser.

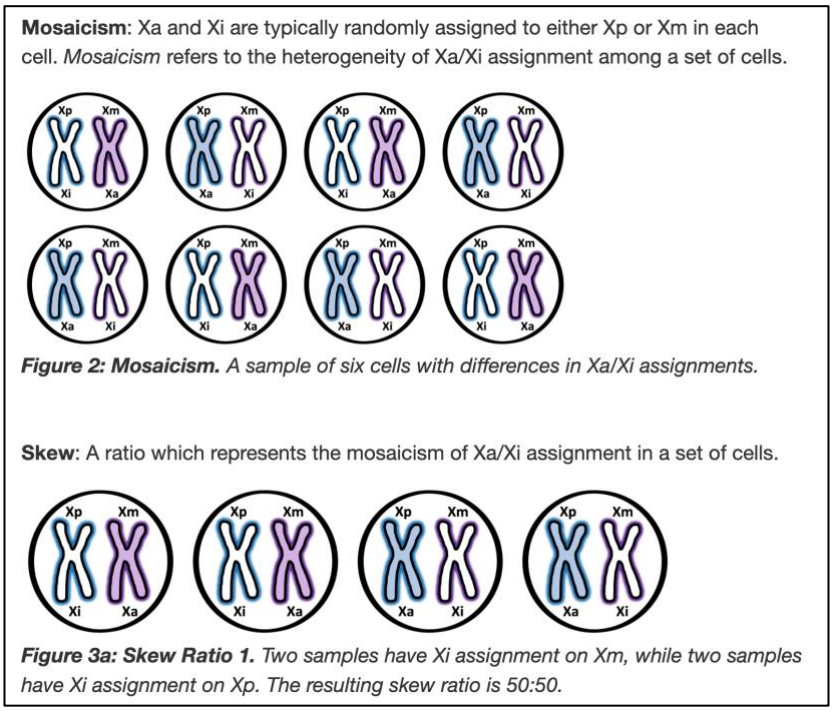


Figure 4-5: An excerpt from the Terminology page of X-Viz; the definitions of mosaicism and skew ratio and the context of the inactive X chromosomes (Xi), active X chromosome (Xa), the paternally inherited chromosomes (Xp) and the maternally inherited chromosomes (Xm).

Summary

We present X-Viz as a useful web-based platform to integrate XCI states with epidemiological findings. XCI escape status and disease associations for X-linked genes can be easily searched, visualized, and summarized. In addition to visualizing data sets and cross-referencing GWAS/UK Biobank data, X-Viz provides a valuable framework for disseminating future research results for X-linked genes. Together with our detailed documentation and tutorials, we anticipate that X-Viz will benefit XCI research and facilitate education on this biological process.

Chapter 5

Future Steps for XCI Research and Tools

XCI Research: Future Steps

The current knowledge of gene regulation and sexually dimorphic traits will advance as more aspects of XCI are better understood. Perhaps one of the most interesting yet understudied aspects of XCI regulation is the existence of genes that do not escape inactivation but show similar sex-biased expression patterns as escape genes. The sex-stratified DGEA study [5] included in the X-Viz application facilitates a simple method to examine this phenomenon. Using the individual gene search engine, a user can search for inactive genes across all or most XCI classification studies that also show female-biased expression in the sex-stratified DGEA study. Applying this method returns the following examples: *ATP6AP1*, *PRDX4*, *ZC4H2*, and *STARD8* which are inactive in all studies; *ACSL4* which is primarily inactive except variable in Sauteraud et al.; *PHKA2* which is primarily inactive except variable in Cotton et al. (methylation model); and *DMD* which is primarily inactive except variable in Balaton and Brown (epigenetic predictor in healthy cells). Table 5-1 summarizes each of these genes' classifications and their biological significance. The higher expression of reportedly inactive genes in females suggests that these genes exhibit higher expression on Xa, which may implicate them as a causal factor in either regulating XCI or dictating reactivation for specific Xi genes. To further examine this hypothesis, researchers could design a DGEA study that captures sex-differentiated expression at a population level as well as the Xa and Xi gene expression at an individual

level. The individual gene expression data is necessary to validate that an X gene's increased female expression is solely due to its expression on Xa. Those genes that do not escape XCI but are upregulated in Xa in females could be targeted in knock-out experiments to monitor their relevance in XCI maintenance. Researchers can utilize experimental approaches such as CRISPR perturbations that suppress specific genomic regions by recruiting transcription-blocking factors through guide RNA [72]. Researchers can thus observe the effects specific Xa genes on X inactivation and reactivated gene expression. Uncovering the role of Xa genes in Xi expression could explain some of the genetic factors that are involved in XCI and Xi reactivation.

Table 5-1: Commonly inactive or variable X genes that are differentially expressed in females

Gene	**DGEA status in Tukiainen et al.**	**XCI status in X-Viz studies**	**Gene Description**
ATPase H+ Transporting Accessory Protein Abbreviation: *ATP6AP1*	Female-biased	Inactive in all studies	ATP6AP encodes proteins that 1transport V-ATPase which is necessary for protein sorting, zymogen activation, and receptor-mediated endocytosis [73]; missense mutations in ATP6AP are associated in immunodeficiency, cognitive impairment, and protein glycosylation [74]
Peroxiredoxin 4 Abbreviation: *PRDX4*	Female-biased	Inactive in all studies	PRDX4 encodes proteins that balance oxidant states in cells and regulate hydrogen peroxide signaling [75]; abnormal protein production is associated in several cancers [76] and cardiovascular risk [75].
Zinc Finger C4H2-Type Containing Abbreviation: *ZC4H2*	Female-biased	Inactive in all studies	ZC4H2 encodes proteins that are part of the zinc finger domain-containing protein family; ZC4H2 is involved in neuronal development and brain/spinal connectivity [77]

STAR related lipid transfer domain containing 8 Abbreviation: *STARD8*	Female-biased	Inactive in all studies	STARD8 encodes proteins that coordinate membrane trafficking and maintain organelle integrity [78]; STARD8 protein products play an integral role in cancer degradation and tumor suppression [79, 80]
Acyl-CoA synthetase long-chain family member 4 Abbreviation: *ACSL4*	Female-biased	Inactive except variable in Sauteraud et al.; lymphoblast	ACLS4 encodes proteins that are part of the long-chain fatty-acid-coenzyme A ligase family and are integral for lipid biosynthesis and fatty acid degradation [81]; ACLS4 protein production abnormalities contribute to cognitive disabilities and aggressive development in breast cancer [81, 82]
Phosphorylase kinase regulatory subunit alpha 2 Abbreviation: *PHKA2*	Female-biased	Inactive except variable in Cotton et al.; DNA methylation; multi-tissue	PHKA2 encodes for the phosphorylase kinase enzyme involved in glycogen breakdown [83]; PHKA2 mutations are strongly associated in X-linked liver glycogen storage disease [83]
Dystrophin Abbreviation: *DMD*	Female-biased	Inactive except variable in Balaton and Brown.; epigenetic predictor; multi-tissue	DMD encodes for a large protein that bridges the inner cytoskeleton and the extracellular matrix [84]; mutations in DMD are associated in X-linked recessive neuromuscular disorders [85]

There is a collective effort to include historically underrepresented populations in data sets within bioinformatics and genomics research. This is motivated by both the statistical utility of including trans-ethnic or sex-stratified data for causal inference, and the pressing need to consider all possible genetic factors for personalized health solutions. Female-biased traits are of particular interest in XCI research due to the sex-specific mechanism of X chromosome inactivation. Several sex-biased disorders are also

disparately represented across populations [39, 49, 86-88]. SLE, for instance, occurs in African-American females nearly three times more frequently than in white females [87]. The studies included in the X-Viz application are comprised primarily of European ancestry, with a small percentage of samples from the Yoruban population [3, 5] and one Asian participant (one fully-skewed individual from Tukiainen et al.) [5]. A critical next step in XCI research is to diversify samples and more adequately represent those minority populations with a statistically higher risk of developing sex-biased disorders. This integration could uncover knowledge on population-specific genetic factors affecting disease development and inform personalized approaches to treating sex-biased disorders.

One fascinating aspect of X reactivation is the overlap between female escape genes and genes implicated in X-linked disorders. As previously noted in *Chapter 2: X Chromosome Inactivation and Phenotypic Outcomes*, certain male-biased cancers have been linked to loss-of-function mutations in *DDX3X*, *KDM5C*, and *KDM6A* genes which also escape or variably escape across several studies. Another example of this occurs in hyper IgM immunodeficiency (HIM), a male-biased disorder mediated by a *CD40LG* mutation in Lymphocytic cells, T and B cells, and fibroblasts [89]. The *CD40LG* gene encodes for proteins expressed in T cells that promote B cell function and stimulate the production of antibodies. Mutations in *CD40LG* contribute to the pathogenesis of HIM, in which hosts are highly susceptible to infections due to a lack of proper immune response [89]. Less severe manifestations occur in females harboring heterozygous mutations of *CD40LG*, likely due to XCI mosaicism; though notably, *CD40LG* is also classified as a variably escaping gene [3, 24, 90]. These examples in male-biased tumor development and immune response prompt the question of whether reactivated genes are a protective feature

informed by deleterious X-linked heterozygous mutations. Enrichment analysis could afford greater insight by observing whether female escape and variable genes are enriched for heterozygous mutations associated in male-biased diseases. X-Viz could facilitate this research through its individual gene search engine, which allows users to input multiple genes of interest. Users can search for genes that have been implicated in male-biased X-linked disorders to return their XCI states in several data sets. X-Viz currently contains a search filter that returns all genes within a study that are associated with female-biased traits. Future upgrades to the application could include a similar search filter for male-biased traits to aid this enrichment analysis. Determining whether certain X reactivation events result from survival-driven mechanisms will give greater insight into the biological purpose of XCI escape.

Expanding the X-Viz application

X-Viz can integrate useful updates to further empower XCI research. X-Viz allows users to manually upload a data set containing already known XCI states to infer broad XCI status based on escape frequency. A helpful upgrade would be to integrate existing R packages that classify XCI status from experimental data. This feature would require users to have data that is compatible with the integrated R package. XCIR, for instance, is a statistical package that classifies escape status based upon RNA sequencing data among heterozygous alleles [91]. The input data is a VCF file comprising counts for heterozygous SNPs. XCIR provides several valuable features, including the ability to specify a statistical model or analyze only highly expressed SNPs. The X-Viz interface could make such

features available to the user (see Figure 5-1 for a conceptual layout). The classified states could then be analyzed in X-Viz similarly to the included studies, allowing for annotations and GWAS cross-analysis. The integration of packages like XCIR in X-Viz would enable scientists to have powerful statistical tools at their fingertips and immediately contextualize their results against epidemiological findings and existing XCI research.

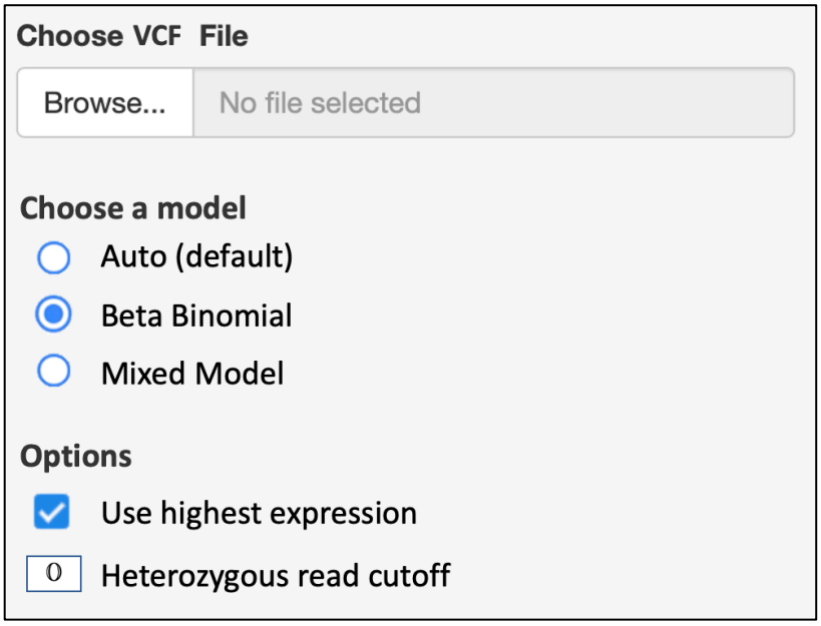


Figure 5-1: Conceptual layout for R statistical package (XCIR) integration in X-Viz

The contextualization of XCI states against known epidemiological events could be enriched through a more extensive search pool and a user-upload feature. The current version of the application includes only GWAS traits in the disease search engine. Future application versions could incorporate other relevant catalogs such as the Online Mendelian Inheritance in Man (OMIM), comprising mendelian traits and disorders [92]. The application could also support manually entering traits or diseases that are not included in the preloaded catalogs. This would allow users to enter a disorder along with its mapped genes and sex bias information from external sources and subsequently have the

information available for each data set to cross-reference. For instance, several studies cite X genes in Rhett Syndrome, a female-biased disease not included in the X-Viz disease/trait search. Manually submitting information on disorders like Rhett Syndrome would improve the disease search pool and expand X-Viz results.

X-Viz could also support technical enhancements to facilitate quicker results and deeper analysis. Table 5-2 below shows a list of technical features that can be added to future application versions.

Table 5-2: Future technical updates to the X-Viz application

Feature	**Description**
Revision History page	A page to give users insight into the changes made to the application
Gene search; text file upload	The individual gene search engine currently allows for multiple inputs; however, it could additionally support a text file upload to handle numerous genes
Ancestry information	As studies incorporate more diverse samples, ancestry information can be included in the tabulated data sets
Disease enriched data sets	The application could include published XCI classifications for data sets that are enriched for sex-biased diseases

Chapter 6

Conclusion

X chromosome inactivation is a remarkable example of the complex biological processes that promote homeostasis and differentiate genetic profiles. Researchers continue to identify various genetic and epigenetic mechanisms that regulate X chromosome silencing and reactivation. Some of the major factors implicated in XCI include the *XIST* transcript, which influences histone and chromatin modifications in the silenced X, and CTCF protein and L1 repeat elements, which delineate and designate XCI regions. Epigenetic factors, such as active histone marks and DNA methylation deposits, seem to significantly impact escape genes' reactivation tendency. XCI classification methods offer novel insight into each biomarker's relative contribution and add incremental information to the full XCI map. As researchers develop experimental and computational approaches to characterize XCI, silencing and Xi reactivation processes will be more fully understood and better inform molecular, genetic, and epidemiological conclusions.

XCI research faces many challenges due to study constrictions and the biological complexities of X inactivation. XCI mosaicism results in heterogeneous Xa expression in females; for ASE methods, mosaic samples have less power to detect escape status. Xi genes vary in their escape expression levels across cells, tissues, and individuals, creating a class of variable escape genes whose functions are not fully understood, nor are they easily designated due to subjective threshold criteria. There is also variability in reactivation timing, which is seemingly influenced by a combination of distance from *XIST* and 3D chromatin structure. Additional complications arise due to results obtained from

commonly used mammal models. Researchers frequently leverage mouse or hybrid models to assess biological phenomenon, however, more recent cross-species analyses find that mouse models are an outlier in XCI studies, specifically when comparing the number of genes subject to inactivation in mice versus other mammal models [31]. XCI research must continually factor for the aforementioned complications by developing computational and experimental approaches that account for intricacies in sample selection, data collection, and the diverse set of XCI factors.

We offer X-Viz as a comprehensive visualization and annotation platform for X chromosome research, a region that is critically underrepresented in genomic studies. X-Viz synthesizes several existing research results to represent a variety of methods and data sets used to classify XCI status [3-5, 7-9, 24]. The scaled X chromosome provides genomic context for commonly escaping regions. X-Viz provides an easy-to-use interface to search the GWAS catalog and the UK Biobank sex biases for traits and disorders mapped to the X chromosome. Sex-biased traits and disorders are of great interest in XCI research due to the possible link to X escape expression. Users can annotate their own data sets and modify escape thresholds to better understand the effect of technical parameters on overall classifications. The application offers additional commentary on XCI concepts and terminology, providing valuable education on XCI research. We expect that X-Viz will be a valuable asset for annotating, contextualizing, and disseminating XCI knowledge and will increase our understanding of X-mediated events.

Appendix A

Data Collection

XCI Studies

We aggregated information across ten different datasets from seven publications. Table A-1 summarizes the data sets included in the application.

Table A-1: Data sets included in the X-Viz application.

Data set	Reference (PUBMED ID)	Methods	Data fields
Sauteraud et al.: Genetic European Variation in Disease (GEUVADIS) lymphoblastoid cell line	34426515	Allele-specific expression	Gene, start, end, escape frequency, state *On Terminology page:* sample name, Xi expression, skew, p-value
Balaton & Brown: Cancer cells, multi-tissue	34187555	DNA methylation mark prediction model	Gene, start, end, state
Balaton & Brown: healthy cells, multi-tissue	34187555	Histone mark prediction model	Gene, start, end, state
Tukiainen et al.: multi-tissue, highly skewed female	29022598	Allele-specific expression	Gene, start end, tissue, escape frequency, state
Tukiainen et al.: multi-tissue, sex specific	29022598	Differential Gene Expression Analysis	Gene, start end, state
Cotton et al.: multi-tissue	25381334	DNA methylation	Gene, DNA methylation position, start, end, tissue, state
Katsir & Linial: lymphoblast	30871455	Single-cell RNA sequencing, Allele-specific expression	Gene, start, end, state
Katsir&Linial: fibroblast	30871455	Single-cell RNA sequencing, Allele-specific expression	Gene, start end, state
Cotton et al.: lymphoblast & fibroblast	24176135	Allele-specific expression	Gene, start end, state
Carrel & Willard: somatic cell hybrids	15772666	Reverse transcription polymerase chain reaction	Gene, start, end, state

Lymphoblast [4]: The Genetic European Variation in Disease (GEUVADIS) [93] dataset contains RNA data for over 460 individuals from the 1000 Genomes project [94]. XCI states were determined in lymphoblast cell lines for 217 genotyped females using the XCIR software package [91]. XCIR leverages allele-specific expression (ASE) methods to estimate sample skewing and determine escape status. Similar to other ASE-based methods [3, 68], XCIR infers whether a given gene escapes XCI in an individual, which is critical

for understanding XCI escape genes with substantial inter-individual differences. The XCIR method outperforms existing ASE-based methods and produces accurately controlled type I error and improved power. The incorporated dataset includes the p-values for testing whether a gene escapes XCI in a given individual (on tab “All Escape Genes”) and the fraction of samples where each gene escapes (%Xi on the tab “Terminology”) in the lymphoblast data set.

Multi-Tissue, Differential Expression Analysis and ASE in one Female with Skewed XCI [5]: The Genotype-Tissue Expression Project [95] contains tissue-specific RNA-seq data for around 1000 individuals. Tukiainen et al. performed differential gene expression analysis on 449 female and male participants from the GTEx v6 release. The analysis spanned 29 adult tissues and identified candidate XCI escape genes that are over-expressed in females relative to male samples. The research group additionally assessed RNA-seq data from one female GTEx donor with highly skewed XCI and made escape calls through biallelic expression analysis. The analysis spanned 16 adult tissues and identified a subset of tissue specific XCI escape genes. Cumulative results for the skewed female sample can be visualized or segregated by tissue.

Multi-Tissue DNA methylation and epigenetic marks [7]: Balaton and Brown examined DNA methylation (DNAme) and histone marks in X promoter regions to predict XCI status. X-Viz includes two data sets from their study: DNAme-based calls in cancer tissues and epigenetic-based calls from healthy tissues. Cancer tissue samples exhibit high XCI skewing and reportedly give more accurate XCI calls when compared to consensus XCI

states or independently assessed ASE calls. DNAme in gene promoter regions was used to predict XCI status in these cancer tissues. This cancer data set was obtained from the Center for Epigenome Mapping Technologies (CEMT) and included six different tissues. Due to the possible epigenetic irregularities in cancer tissues, Balaton and Brown also predicted XCI status in healthy cells using a random forest model that considered H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K27me3, and H3K36me3 levels. These healthy tissue samples were collected from the Core Research for Evolutional Science and Technology (CREST) and included nine female samples derived from hepatocytes and human colon absorptive epithelium.

Multi-Tissue DNA methylation [8]: DNAme patterns in X gene promoter sequences were examined in 1800 female samples spanning twenty-seven tissues. XCI status was predicted by comparing DNAme levels in promoter regions to DNAme thresholds derived from training sets of genes with known XCI patterns.

Fibroblast and Lymphoblast single-cell RNA-seq dataset [9]: Katsir and Linial conducted single-cell RNA-Seq to measure the ASE at heterozygous locations of fibroblasts without haplotype phasing ($n = 104$) and that of clonal lymphoblast cell line GM12878 with full haplotype phasing ($n = 25$). Heterozygous SNPs with sufficient biallelic expression and measurable in at least seven samples were used to determine the XCI escape status.

Fibroblast and Lymphoblast (combined) [3]: Cotton et al. leveraged cDNA hybridization to SNP microarrays to assess allelic expression imbalance from lymphoblast cell lines of the Centre d'Etude du Polymorphisme Humain HapMap population (*n = 30 females*), the Yoruban HapMap population (*n = 31 females*), and fibroblast cell lines from 38 females. A gene was deemed "escape" XCI if more than 78% of individuals exhibited at least 10% Xi expression. A total of 506 classified genes from this study were remapped to the hg19 genome build by Tukiainen et al. [5], which was incorporated in our X-Viz browser.

Somatic cell hybrids (Carrel and Willard, 2005): The first X-inactivation profile was established using nine rodent/human somatic cell hybrids that retain a human Xa or Xi. This approach allows Xa or Xi expression to be independently assessed in each line directly by RT-PCR. Escape status was assigned to genes with Xi expression in more than 75% of the samples. Variable escape status was assigned to genes that escaped in 25% to 75% of the samples, and inactive status was assigned to genes that escaped in less than 25% of the samples.

The NCBI Genome Remapping Service was used to lift over the genomic positions for each study from hg19 to hg38 (https://www.ncbi.nlm.nih.gov/genome/tools/remap).

Disease/Trait Collection

GWAS diseases and traits were collected from version 1.02 of the GWAS catalog [96]. The application queries the MAPPED_GENE column of the GWAS summary table for the gene of interest. The sex ratios of diseases/traits were derived from the sex ratios of personally reported traits in the UK Biobank [97]. Due to the difference in terminology across the UK Biobank and GWAS phenotypes, the Mapping UK Biobank to the Experimental Factor Ontology (EFO) master conversion file was used to connect similar traits between the two sources [98].

Appendix B

Escape Frequency Thresholds: Tutorial and Example

Background

The percent escape (PE) of a gene is the frequency of its escape status among samples. For example, in the gene data shown in Figure B-1 below, *HCFC1* escapes in 3/10 of the samples, resulting in a PE of 30%.

gene	samp_state	sample	start	escape_freq
HCFC1	escape	HG00146	153947557	0.3
HCFC1	inactive	HG00163	153947557	0.3
HCFC1	escape	NA20502	153947557	0.3
HCFC1	inactive	HG00231	153947557	0.3
HCFC1	inactive	HG00235	153947557	0.3
HCFC1	escape	HG00258	153947557	0.3
HCFC1	inactive	NA20756	153947557	0.3
HCFC1	inactive	NA20769	153947557	0.3
HCFC1	inactive	NA20795	153947557	0.3
HCFC1	inactive	NA20819	153947557	0.3

Figure B-1: Example of gene escape frequency calculation from *samp_state* column

The PE of a gene is used to classify its escape status. Genes which escape in nearly all samples are classified as “escape”; genes which escape in some but not all samples are “variable”; genes which escape in few or no samples are “inactive”. **Escape Frequency Thresholds** delineate each classification.

For example, the GEUVADIS lymphoblast study uses 25% and 75% thresholds.

- %escape ≤ 25% → Silenced (S)
- 25% < %escape < 75 Variable escape (VE)
- 75% → Escape (E)

Applying these thresholds to our example data set will classify *HCFC1* as “variable”, as pictured in Figure B-2.

gene	samp_state	sample	start	escape_freq	state
HCFC1	escape	HG00146	153947557	0.3	variable
HCFC1	inactive	HG00163	153947557	0.3	variable
HCFC1	escape	NA20502	153947557	0.3	variable
HCFC1	inactive	HG00231	153947557	0.3	variable

Figure B-2: Example of variable gene status as a result of 25% and 75% escape frequency thresholds

Changing Escape Frequency Thresholds

The escape frequency thresholds of 25% and 75% are arbitrarily selected and can be informed by biological information or study design. The application gives users the freedom to change the escape frequency thresholds to allow for technical flexibility.

The left side panel contains sliders that modify the lower and upper thresholds, as pictured in Figure B-3. These sliders are available only for studies that contain sample-level information of escape status.

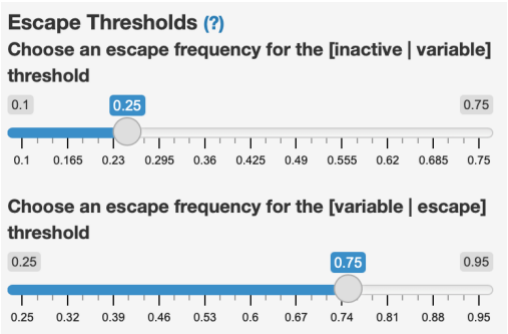
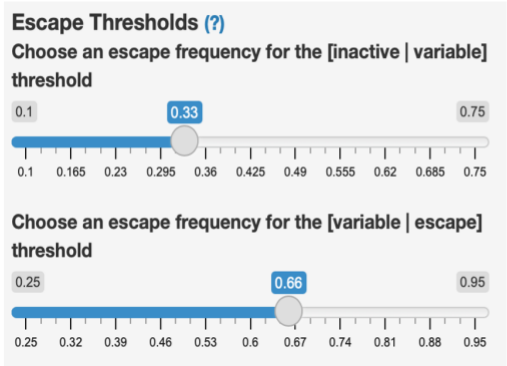


Figure B-3: Escape frequency threshold sliders included in X-Viz application

Changing the escape thresholds will modify the classification sensitivity for the selected study. In our example, modifying the thresholds to 33% and 66% will re-classify the escape status of the *HCFC1* gene (Figure B-4).

a



b

gene	samp_state	sample	start	escape_freq	state
HCFC1	escape	HG00146	153947557	0.3	inactive
HCFC1	inactive	HG00163	153947557	0.3	inactive
HCFC1	escape	NA20502	153947557	0.3	inactive
HCFC1	inactive	HG00231	153947557	0.3	inactive

Figure B-4: Example of modified XCI status as a result of adjusting escape frequency thresholds. a) Escape frequency thresholds adjusted from 25%-75% to 33%-66%. b) Example gene *HFC1* exhibiting inactive status as a result of modified escape frequency thresholds.

Modifying the escape thresholds will also update the gene position graphic. Note, this graphic is only available for studies that include positional data ("start" column). Figure B-5 on page 39 displays modified visualizations for different escape frequency thresholds.

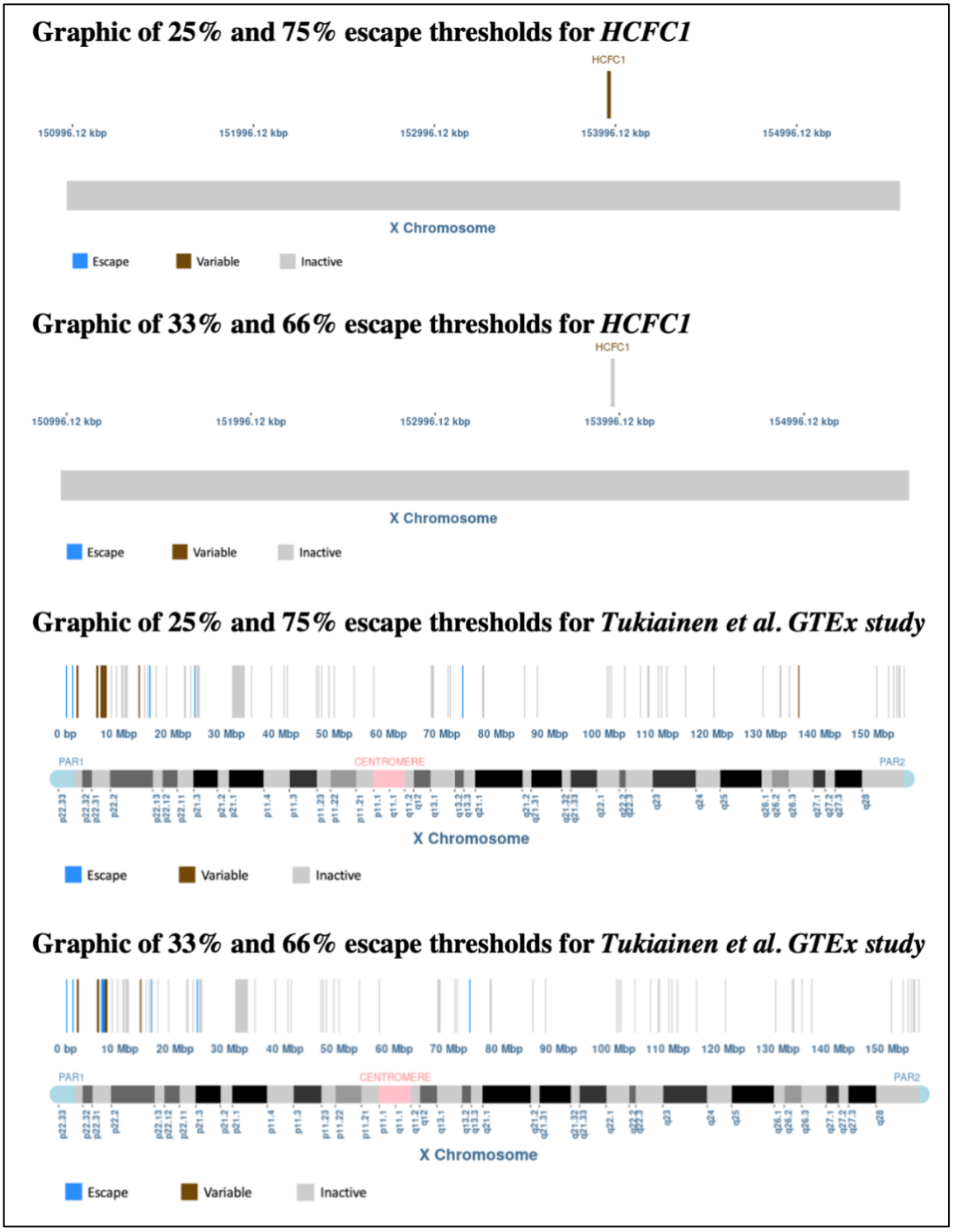


Figure B-5: Examples of modified visualizations as a result of adjusting escape frequency thresholds. The first two images display the adjusted color of *HCFC1* from the original turquoise (variable) to grey (inactive). In the bottom two images, the color profile for the Tukiainen et al. GTEx study is modified; more specifically, fewer variable genes are displayed due to applying stringent escape frequency thresholds.

REFERENCES

1. Carrel, L. and C.J. Brown, *When the Lyon(ized chromosome) roars: ongoing expression from an inactive X chromosome.* Philos Trans R Soc Lond B Biol Sci, 2017. **372**(1733).
2. Khramtsova, E.A., L.K. Davis, and B.E. Stranger, *The role of sex in the genomics of human complex traits.* Nature Reviews Genetics, 2019. **20**(3): p. 173-190.
3. Cotton, A.M., et al., *Analysis of expressed SNPs identifies variable extents of expression from the human inactive X chromosome.* Genome Biol, 2013. **14**(11): p. R122.
4. Sauteraud, R., et al., *Inferring Genes that Escape X Chromosome Inactivation (with XCIR) Reveals Important Contribution of Variable Escape Genes to Sex-biased Diseases.* Genome Research, 2021.
5. Tukiainen, T., et al., *Landscape of X chromosome inactivation across human tissues.* Nature, 2017. **550**(7675): p. 244-248.
6. Oliva, M., et al., *The impact of sex on gene expression across human tissues.* Science, 2020. **369**(6509).
7. Balaton, B.P. and C.J. Brown, *Contribution of genetic and epigenetic changes to escape from X-chromosome inactivation.* Epigenetics & Chromatin, 2021. **14**(1): p. 30.
8. Cotton, A.M., et al., *Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation.* Human Molecular Genetics, 2014. **24**(6): p. 1528-1539.
9. Wainer Katsir, K. and M. Linial, *Human genes escaping X-inactivation revealed by single cell expression data.* BMC Genomics, 2019. **20**(1): p. 201.
10. Lyon, M., *X-chromosome inactivation and human genetic disease.* Acta Paediatrica, 2002. **91**(s439): p. 107-112.
11. Sun, B.K. and H. Tsao, *X-Chromosome Inactivation and Skin Disease.* Journal of Investigative Dermatology, 2008. **128**(12): p. 2753-2759.
12. Migeon, B.R., *Why females are mosaics, x-chromosome inactivation, and sex differences in disease.* Gender Medicine, 2007. **4**(2): p. 97-105.
13. Augui, S., E.P. Nora, and E. Heard, *Regulation of X-chromosome inactivation by the X-inactivation centre.* Nature Reviews Genetics, 2011. **12**(6): p. 429-442.
14. Rasmussen, T.P., et al., *Expression of Xist RNA is sufficient to initiate macrochromatin body formation.* Chromosoma, 2001. **110**(6): p. 411-420.
15. Pullirsch, D., et al., *The Trithorax group protein Ash2l and Saf-A are recruited to the inactive X chromosome at the onset of stable X inactivation.* Development, 2010. **137**(6): p. 935-943.
16. Żylicz, J.J., et al., *The Implication of Early Chromatin Changes in X Chromosome Inactivation.* Cell, 2019. **176**(1): p. 182-197.e23.

17. Loda, A., et al., *Genetic and epigenetic features direct differential efficiency of Xist-mediated silencing at X-chromosomal and autosomal locations.* Nature Communications, 2017. **8**(1): p. 690.
18. Fang, H., C.M. Disteche, and J.B. Berletch, *X Inactivation and Escape: Epigenetic and Structural Features.* Frontiers in Cell and Developmental Biology, 2019. **7**(219).
19. Brown, C.J. and H.F. Willard, *The human X-inactivation centre is not required for maintenance of X-chromosome inactivation.* Nature, 1994. **368**(6467): p. 154-156.
20. Yang, F., et al., *The lncRNA Firre anchors the inactive X chromosome to the nucleolus by binding CTCF and maintains H3K27me3 methylation.* Genome Biol, 2015. **16**: p. 52.
21. Stribinskis, V. and K.S. Ramos, *2.21 - LINE-1*, in *Comprehensive Toxicology (Second Edition)*, C.A. McQueen, Editor. 2010, Elsevier: Oxford. p. 403-426.
22. Bailey, J.A., et al., *Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: The Lyon repeat hypothesis.* Proceedings of the National Academy of Sciences, 2000. **97**(12): p. 6634.
23. Chow, J.C., et al., *LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation.* Cell, 2010. **141**(6): p. 956-69.
24. Carrel, L. and H.F. Willard, *X-inactivation profile reveals extensive variability in X-linked gene expression in females.* Nature, 2005. **434**(7031): p. 400-4.
25. Cotton, A.M., et al., *Spread of X-chromosome inactivation into autosomal sequences: role for DNA elements, chromatin features and chromosomal domains.* Human Molecular Genetics, 2014. **23**(5): p. 1211-1223.
26. Goto, Y. and H. Kimura, *Inactive X chromosome-specific histone H3 modifications and CpG hypomethylation flank a chromatin boundary between an X-inactivated and an escape gene.* Nucleic Acids Research, 2009. **37**(22): p. 7416-7428.
27. Sadreyev, R.I., et al., *Bimodal quantitative relationships between histone modifications for X-linked and autosomal loci.* Proc Natl Acad Sci U S A, 2013. **110**(17): p. 6949-54.
28. Kelsey, A.D., et al., *Impact of flanking chromosomal sequences on localization and silencing by the human non-coding RNA XIST.* Genome Biology, 2015. **16**(1): p. 208.
29. Simon, M.D., et al., *High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation.* Nature, 2013. **504**(7480): p. 465-469.
30. Filippova, G.N., et al., *Boundaries between Chromosomal Domains of X Inactivation and Escape Bind CTCF and Lack CpG Methylation during Early Development.* Developmental Cell, 2005. **8**(1): p. 31-42.
31. Balaton, B.P., et al., *Cross-species examination of X-chromosome inactivation highlights domains of escape from silencing.* Epigenetics & Chromatin, 2021. **14**(1): p. 12.
32. Spencer, R.J., et al., *A Boundary Element Between Tsix and Xist Binds the Chromatin Insulator Ctcf and*

Contributes to Initiation of X-Chromosome Inactivation. Genetics, 2011. **189**(2): p. 441.
33. Berletch, J.B., et al., *Escape from X Inactivation Varies in Mouse Tissues.* PLOS Genetics, 2015. **11**(3): p. e1005079.
34. McArthur, E. and J.A. Capra, *Topologically associating domain boundaries that are stable across diverse cell types are evolutionarily constrained and enriched for heritability.* The American Journal of Human Genetics, 2021. **108**(2): p. 269-283.
35. Marks, H., et al., *Dynamics of gene silencing during X inactivation using allele-specific RNA-seq.* Genome biology, 2015. **16**(1): p. 149-149.
36. Balaton, B.P., A.M. Cotton, and C.J. Brown, *Derivation of consensus inactivation status for X-linked genes from genome-wide studies.* Biology of sex differences, 2015. **6**: p. 35-35.
37. Barros de Andrade E Sousa, L., et al., *Kinetics of Xist-induced gene silencing can be predicted from combinations of epigenetic and genomic features.* Genome research, 2019. **29**(7): p. 1087-1099.
38. Chaligné, R., et al., *The inactive X chromosome is epigenetically unstable and transcriptionally labile in breast cancer.* Genome research, 2015. **25**(4): p. 488-503.
39. Society, A.C., *Cancer Facts & Figures 2021.* Atlanta, Ga; American Cancer Society, 2021.
40. Román-Fernández, I.V., et al., *Assessment of CD40 and CD40L expression in rheumatoid arthritis patients, association with clinical features and DAS28.* Clinical and Experimental Medicine, 2019. **19**(4): p. 427-437.
41. Guerra, S.G., T.J. Vyse, and D.S. Cunninghame Graham, *The genetics of lupus: a functional perspective.* Arthritis research & therapy, 2012. **14**(3): p. 211-211.
42. Liu, K., et al., *X Chromosome Dose and Sex Bias in Autoimmune Diseases: Increased Prevalence of 47,XXX in Systemic Lupus Erythematosus and Sjögren's Syndrome.* Arthritis & Rheumatology, 2016. **68**(5): p. 1290-1300.
43. Bentham, J., et al., *Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus.* Nat Genet, 2015. **47**(12): p. 1457-1464.
44. Syrett, C.M., et al., *Altered X-chromosome inactivation in T cells may promote sex-biased autoimmune diseases.* JCI Insight, 2019. **4**(7).
45. Pyfrom, S., et al., *The dynamic epigenetic regulation of the inactive X chromosome in healthy human B cells is dysregulated in lupus patients.* Proc Natl Acad Sci U S A, 2021. **118**(24).
46. Flo, T.H. and A. Aderem, *Pathogen Recognition by Toll-like Receptors*, in *NeuroImmune Biology*, L. Bertók and D.A. Chow, Editors. 2005, Elsevier. p. 167-182.
47. Souyris, M., et al., *TLR7 escapes X chromosome inactivation in immune cells.* Science Immunology, 2018. **3**(19): p. eaap8855.
48. Deane, J.A., et al., *Control of Toll-like Receptor 7 Expression Is Essential to Restrict Autoimmunity and Dendritic Cell Proliferation.* Immunity, 2007. **27**(5): p. 801-810.

49. Izmirly, P.M., et al., *The Incidence and Prevalence of Systemic Lupus Erythematosus in New York County (Manhattan), New York: The Manhattan Lupus Surveillance Program.* Arthritis Rheumatol, 2017. **69**(10): p. 2006-2017.
50. Cano RLE, L.H., Anaya JM, Shoenfeld Y, Rojas-Villarraga A, et al., *Introduction to T and B lymphocytes*, in *Autoimmunity: From Bench to Bedside*. 2013 Jul 18., El Rosario University Press: Bogota (Colombia).
51. Suarez-Fueyo, A., S.J. Bradley, and G.C. Tsokos, *T cells in Systemic Lupus Erythematosus.* Curr Opin Immunol, 2016. **43**: p. 32-38.
52. Sripathy, S., et al., *Screen for reactivation of MeCP2 on the inactive X chromosome identifies the BMP/TGF-β superfamily as a regulator of XIST expression.* Proceedings of the National Academy of Sciences of the United States of America, 2017. **114**(7): p. 1619-1624.
53. Gonzales, M.L. and J.M. LaSalle, *The Role of MeCP2 in Brain Development and Neurodevelopmental Disorders.* Current Psychiatry Reports, 2010. **12**(2): p. 127-134.
54. Guy, J., et al., *Reversal of neurological defects in a mouse model of Rett syndrome.* Science (New York, N.Y.), 2007. **315**(5815): p. 1143-1147.
55. Dunford, A., et al., *Tumor-suppressor genes that escape from X-inactivation contribute to cancer sex bias.* Nature Genetics, 2017. **49**(1): p. 10-16.
56. Lin, T.C., *DDX3X Multifunctionally Modulates Tumor Progression and Serves as a Prognostic Indicator to Predict Cancer Outcomes.* Int J Mol Sci, 2019. **21**(1).
57. Chang, S., S. Yim, and H. Park, *The cancer driver genes IDH1/2, JARID1C/ KDM5C, and UTX/ KDM6A: crosstalk between histone demethylation and hypoxic reprogramming in cancer metabolism.* Exp Mol Med, 2019. **51**(6): p. 1-17.
58. Lucas, S., E. De Plaen, and T. Boon, *MAGE-B5, MAGE-B6, MAGE-C2, and MAGE-C3: four new members of the MAGE family with tumor-specific expression.* Int J Cancer, 2000. **87**(1): p. 55-60.
59. Eng, K.H., et al., *Paternal lineage early onset hereditary ovarian cancers: A Familial Ovarian Cancer Registry study.* PLoS Genet, 2018. **14**(2): p. e1007194.
60. Snijders Blok, L., et al., *Mutations in DDX3X Are a Common Cause of Unexplained Intellectual Disability with Gender-Specific Effects on Wnt Signaling.* American journal of human genetics, 2015. **97**(2): p. 343-352.
61. Rhee, S.H. and I.D. Waldman, *Etiology of sex differences in the prevalence of ADHD: An examination of inattention and hyperactivity–impulsivity.* American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 2004. **127B**(1): p. 60-64.
62. Brown, C.J. and H.F. Willard, *Noninactivation of a selectable human X-linked gene that complements a murine temperature-sensitive cell cycle defect.* American journal of human genetics, 1989. **45**(4): p. 592-598.
63. Garitaonandia, I., et al., *Increased risk of genetic and epigenetic instability in human embryonic stem cells associated with specific culture conditions.* PLoS One, 2015. **10**(2): p. e0118307.
64. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Res, 2015. **43**(7): p. e47.

65. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.
66. Schultz, M.D., et al., *Human body epigenome maps reveal noncanonical DNA methylation variation.* Nature, 2015. **523**(7559): p. 212-6.
67. Bentham, J., et al., *Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus.* Nature Genetics, 2015. **47**(12): p. 1457-1464.
68. Larson, N.B., et al., *An integrative approach to assess X-chromosome inactivation using allele-specific expression with applications to epithelial ovarian cancer.* Genet Epidemiol, 2017. **41**(8): p. 898-914.
69. Zhang, Y., et al., *Genes that escape X-inactivation in humans have high intraspecific variability in expression, are associated with mental impairment but are not slow evolving.* Mol Biol Evol, 2013. **30**(12): p. 2588-601.
70. Gagliano Taliun, S.A., et al., *Exploring and visualizing large-scale genetic associations by using PheWeb.* Nature Genetics, 2020. **52**(6): p. 550-552.
71. Sauteraud, R., et al., *Inferring genes that escape X-Chromosome inactivation reveals important contribution of variable escape genes to sex-biased diseases.* Genome Res, 2021.
72. Mandegar, M.A., et al., *CRISPR Interference Efficiently Induces Specific and Reversible Gene Silencing in Human iPSCs.* Cell Stem Cell, 2016. **18**(4): p. 541-53.
73. NCBI *ATP6AP1 ATPase H+ transporting accessory protein 1 [Homo sapiens (human)]*. Gene 2021.
74. Jansen, E.J., et al., *ATP6AP1 deficiency causes an immunodeficiency with hepatopathy, cognitive impairment and abnormal protein glycosylation.* Nat Commun, 2016. **7**: p. 11600.
75. Gateva, A., et al., *Increased peroxiredoxin 4 levels in patients with prediabetes compared to normal glucose tolerance subjects.* Clin Endocrinol (Oxf), 2016. **85**(4): p. 551-5.
76. Jia, W., P. Chen, and Y. Cheng, *PRDX4 and Its Roles in Various Cancers.* Technol Cancer Res Treat, 2019. **18**: p. 1533033819864313.
77. May, M., et al., *ZC4H2, an XLID gene, is required for the generation of a specific subset of CNS interneurons.* Hum Mol Genet, 2015. **24**(17): p. 4848-61.
78. Braun, A.C., et al., *The Rho-specific GAP protein DLC3 coordinates endocytic membrane trafficking.* J Cell Sci, 2015. **128**(7): p. 1386-99.
79. Noll, B., et al., *DLC3 suppresses MT1-MMP-dependent matrix degradation by controlling RhoB and actin remodeling at endosomal membranes.* J Cell Sci, 2019. **132**(11).
80. Durkin, M.E., et al., *Deleted in liver cancer 3 (DLC-3), a novel Rho GTPase-activating protein, is downregulated in cancer and inhibits tumor cell growth.* Oncogene, 2007. **26**(31): p. 4580-9.
81. NCBI *ACSL4 acyl-CoA synthetase long chain family member 4 [Homo sapiens (human)]*. 2021.

82. Castillo, A.F., et al., *New inhibitor targeting Acyl-CoA synthetase 4 reduces breast and prostate tumor growth, therapeutic resistance and steroidogenesis.* Cell Mol Life Sci, 2021. **78**(6): p. 2893-2910.
83. Hendrickx, J., et al., *Mutations in the phosphorylase kinase gene PHKA2 are responsible for X-linked liver glycogen storage disease.* Hum Mol Genet, 1995. **4**(1): p. 77-83.
84. NCBI *DMD dystrophin [Homo sapiens (human)]*. Gene, 2021.
85. Onore, M.E., et al., *Linked-Read Whole Genome Sequencing Solves a Double DMD Gene Rearrangement.* Genes (Basel), 2021. **12**(2).
86. Somers, E.C., et al., *Population-based incidence and prevalence of systemic lupus erythematosus: the Michigan Lupus Epidemiology and Surveillance program.* Arthritis Rheumatol, 2014. **66**(2): p. 369-78.
87. Lim, S.S., et al., *The incidence and prevalence of systemic lupus erythematosus, 2002-2004: The Georgia Lupus Registry.* Arthritis Rheumatol, 2014. **66**(2): p. 357-68.
88. Catherino, W.H., H.M. Eltoukhi, and A. Al-Hendy, *Racial and ethnic differences in the pathogenesis and clinical manifestations of uterine leiomyoma.* Semin Reprod Med, 2013. **31**(5): p. 370-9.
89. Bhushan, A. and L.R. Covey, *CD40: CD40L interactions in X-linked and non-X-linked hyper-IgM syndromes.* Immunologic Research, 2001. **24**(3): p. 311-324.
90. Migeon, B.R., *X-linked diseases: susceptible females.* Genetics in Medicine, 2020. **22**(7): p. 1156-1174.
91. Sauteraud, R., *XCIR*. 2021, Bioconductor.
92. OMIM, *Online Mendelian Inheritance in Man, OMIM®*, in *McKusick-Nathans Institute of Genetic Medicine*. 2021.
93. Lappalainen, T., et al., *Transcriptome and genome sequencing uncovers functional variation in humans.* Nature, 2013. **501**(7468): p. 506-11.
94. The 1000 Genomes Project Consortium, *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
95. GTEx Consortium, et al., *Genetic effects on gene expression across human tissues.* Nature, 2017. **550**(7675): p. 204-213.
96. Buniello, A., et al., *NHGRI-EBI GWAS Catalog.* Nucleic Acids Research, 2019. **47**(D1005-D1012).
97. Sudlow, C., et al., *UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age.* PLoS Med, 2015. **12**(3): p. e1001779.
98. Zoë May Pendlington, P.R., Edward Mountjoy, Gautier Koscielny, Helen Parkinson, Simon Jupp, *Mapping UK Biobank to the Experimental Factor Ontology*. GitHub Repository, 2019.