The Pennsylvania State University
The Graduate School

STATISTICAL METHODS FOR THE FUNCTIONAL GENOMIC ANALYSIS OF THE X CHROMOSOME

A Dissertation in
Biostatistics
by
Renan Sauteraud

© 2021 Renan Sauteraud

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2021

The dissertation of Renan Sauteraud was reviewed and approved by the following:

Dajiang Liu
Dissertation Co-Adviser, Chair of Committee
Associate Professor, Department of Public Health Sciences

Laura Carrel
Dissertation Co-adviser
Associate Professor, Department of Biochemistry and Molecular Biology

Arthur Berg
Associate Professor, Division of Biostatistics and Bioinformatics
Department of Public Health Sciences
Head of Biostatistics Program

David Mauger
Professor, Division of Biostatistics and Bioinformatics
Department of Public Health Sciences

Nancy Olsen
Chief, Division of Rheumatology
Department of Medicine
Special Member

Ziaur Rahman
Associate Professor, Department of Microbiology and Immunology

Abstract

The X Chromosome plays an important role in human development and disease. However, functional genomic and disease association studies of X genes greatly lag behind autosomal gene studies.Several analytical challenges arise from the unique biology of X including chromosome copy number differences between males and females and X chromosome inactivation (XCI) in females with two copies of the X. Because of XCI, most genes are only expressed from one allele. Yet, 30% of X genes "escape" XCI and are transcribed from both alleles, many only in a proportion of the population. Such inter-individual differences are likely to be disease-relevant, particularly for sex-biased disorders.

In the first chapter, we introduce XCIR (X-Chromosome Inactivation for RNA-Seq), a novel statistical method to identify escape genes using bulk RNA-sequencing data. Our approach jointly models the probability of errors common to the study of XCI along with the sample mosaicism. In simulations, we show improvement in power to detect escape genes over existing methods. We further validate the data in controlled experiment and apply XCIR to publicly available data. Finally, we address limitations specific to expression based approaches and quantify their impact in the context of XCI and the analysis of X-linked genes.

In the second chapter, we apply our novel method to real data in order to understand the functional biology for X-linked genes. Using annotated XCI states, we examined the contribution of X-linked genes to the disease heritability in the UK Biobank dataset.We show that escape and variable escape genes explain the largest proportion of X heritability, which is in large part attributable to X genes with Y homology. Finally, we investigated the role of each XCI state in sex-biased diseases and found that while XY homologous gene pairs have a larger overall effect size, enrichment for variable escape genes is significantly increased in female-biased diseases. These results, for the first time, quantify the importance of variable escape genes for the etiology of sex-biased disease. Our method, available as an R package, is more powerful than alternative approaches and is computationally efficient to handle large population-scale datasets allowing the analysis of a broad range of phenotypes.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank the many people that made my graduation possible. First and foremost, my Academic advisers Dr. Dajiang Liu and Dr. Laura Carrel for their mentorship and continued support throughout my time at Penn State. I am extremely grateful to Dr. Liu for allowing me to join his research group early on and making use of my abilities while helping me transition into statistics. I am deeply thankful to Dr. Carrel for providing expert insight and an outside perspective on my work, ensuring its usefulness to a broader audience. I want to thank all committee members for their assistance and collaboration as well as my teachers in the Public Health Sciences department who made enrolling in the Biostatistics PhD program one of the best decision I ever made.

I extend my deepest gratitude to Dr. Dan McGuire, Dr. Vishal Midya and Dr. Lin Qiu for the the time we spent together and their patience as friends classmates, teammates, coworkers and teachers. Many thanks to the ones who made Hershey and Penn State the best place to be, in particular, Aditi Sharma and Xiangyu Cai along with many more.

Finally, I want to recognize the contributions of, my Mother, Father, Kevin, as well as la vraie famille and the homies.

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

Chapter 1
Statistical Models for X Chromosome Inactivation Inference

1.1 Introduction

The human X Chromosome harbors 1000 genes that perform diverse functions and play key roles in human development and disease etiology [1–4]. Yet, the study of X-linked genes in disease genetics and functional genomics greatly lags compared to those on autosomes [5]. Several analytical challenges arise from the unique biology of X, including chromosome copy number differences between males and females and X-Chromosome inactivation (XCI) in XX females. To balance dosage between sexes, XCI transcriptionally silences one X in female cells (Xi), and most genes only express the active X (Xa) allele. However, about 10% of X-linked genes consistently escape XCI and are expressed from both the Xa and Xi. Moreover, as many as 30% of human X genes show variable XCI escape and exhibit inter- and intra-individual differences; that is, they escape XCI in a subset of individuals or tissues within an individual but remain X inactivated in others [6–9]. As most genes that escape or variably escape XCI lack functionally equivalent Y homologs, Xi expression can result in dosage imbalance between the sexes. A role for XCI escape is emerging for sex-biased traits and disorders [10–19]. For example, XCI escape in immune cells, due to altered XCI maintenance that results in partial gene reactivation, may in part explain the severe sex bias in autoimmune disorders

such as Systemic Lupus Erythematosus (male:female ratio 1:9) [11,13,16,20]. Nonetheless, a more complete understanding of the role that XCI escape plays in disorders, particularly those that are sex-biased, has yet to be fully appreciated.

In order to better understand the biology of X-linked genes and unveil functional and clinical consequences of XCI escape, a critical first step is to identify genes that escape XCI in a particular tissue or disease state. The chromosome-wide analysis is complicated by the random nature of XCI in early development, as the determination of Xa and Xi chromosomes differs from cell to cell (cellular mosaicism). Single-cell RNA-seq circumvents complications of XCI mosaicism and allows monoallelic/biallelic XCI state assessments [8]. However, most single-cell RNA-seq datasets are limited to a very small number of individuals [9, 21], making them inadequate for dissecting subtle inter-individual differences.

1.1.1 Existing Methods

In humans, computational and experimental methods have been developed to identify escape genes from samples with mosaic XCI. Indirect assessment of XCI escape genes using DNA methylation can be applied to large datasets [22, 23], yet whether methylation accurately proxies Xi expression of variable escape genes has not been directly assessed at a whole X level. Male/female differential gene expression (DGE) analyses has also been employed to identify escape genes (e.g. [8]), as genes that escape XCI are often expressed higher in females than males. A limitation of DGE is that it cannot characterize whether a gene escapes in a given individual, which is key for understanding genotype-phenotype associations. Allele-specific expression (ASE) analyses have also been used to identify escape genes [24, 25]. ASE can be quantified using heterozygous transcribed SNPs to compare the expression ratio between the two alleles. The ASE levels from silenced genes mirror the level of XCI mosaicism in the sample whereas escape genes are expressed more evenly between the two alleles. Comparing the observed allelic ratio between the candidate gene and commonly silenced genes, statistical tests can be conducted to infer XCI escape genes.

Among approaches that identify escape genes using expression datasets, ASE-based methods are desirable as they allow the assessment of XCI states of genes in individual samples and can maximize the utility of existing bulk RNA-seq data. Among ASE-based methods, Cotton, et al. [24] estimate Xi expression level by comparing ASE ratios for a candidate gene against the ASE ratios for a set of commonly silenced genes. Genes with Xi expression level >10% of Xa expression levels are deemed to escape XCI. We refer to this method as Xi-threshold. Larson et al [25] use a Bayesian mixture model (BayesMix) to cluster the genic ASE and calculate a posterior probability of escaping XCI (PPE) for each gene. To determine if a gene escapes XCI, a threshold on the PPE (50%) is set. Genes with PPE greater than the threshold are deemed to escape. Therefore, both Xi-threshold and BayesMix require a pre-specified threshold (on Xi expression or the PPE) to call escape genes. In practice, such a threshold can be hard to determine and both methods can have unacceptably high false positive rates under the default threshold (i.e. they incorrectly call silenced genes as escape genes). Also, neither method properly considers whether a heterozygous SNP used in the ASE calculation may be due to a sequencing error or if the set of commonly silenced genes includes a gene that escapes XCI in a given individual. Failure to account for these artifacts may lead to inflated type 1 errors and reduced power for inferring XCI escape genes.

1.2 Estimating Individual Mosaicism From RNA Sequencing data

1.2.1 Data Structure

For a gene g in a female individual, let N be the total number of reads, N_1 and N_2 and be the number of reads mapped to each haplotype. When haplotype information is available, N_1 and N_2 represent the total read count from haplotypes 1 and 2 summed across heterozygous SNPs. If haplotype information is not available, N_1 and N_2 may represent the read counts from the most highly expressed SNP within the transcribed

region. Let N_a, N_i be the number of reads expressed from X_a and X_i respectively. Given that the bulk RNA-seq data consists of a mosaic of cells with different Xa/Xi assignment, we further denote the pairs N_{a1}, N_{a2} and N_{i1}, N_{i2} as the number of reads on the first and second active and inactive chromosomes respectively to get the following data structure.

Table 1.1. Data structure

	X_1	X_2	
X_a	N_{a1}	N_{a2}	N_a
X_i	N_{i1}	N_{i2}	N_i
	$\boldsymbol{N_1}$	$\boldsymbol{N_2}$	$\boldsymbol{N}$

Where only N_1, N_2 and N are observed. The read counts $N_{a1}, N_{a2}, N_{i1}, N_{i2}$ are not directly observed in bulk RNA-seq data and need to be statistically inferred. For genes that escape XCI, we expect $N_{i1} > 0, N_{i2} > 0$. The relationship between these read counts satisfies

$$N_j = N_{aj} + N_{ij}, j = 1, 2 \tag{1.1}$$

A key parameter of interest for XCI inference is XCI skewing (denoted as f), which can be represented by the fraction of cells where a given haplotype (e.g., the first haplotype) is actively expressed. The observed allelic expression and the number of reads from Xa and Xi satisfy

$$E(N_1) = E(N_a)f + E(N_i)(1 - f) \tag{1.2}$$

For genes that are silenced by XCI, $E(N_{i1}) = 0$, and the above equation reduces to

$$f = \frac{E(N_1)}{E(N_a)} \tag{1.3}$$

1.2.2 Joint modelling of Sample Skewing and Error Sources

In theory, the sample skewing f can be estimated using the ratio of N_1/N. Yet, it should be noted that the training set of commonly silenced genes may include genes which escape XCI in a particular sample. The contamination can be extensive as the original

training set was obtained using relatively small datasets [6, 22, 24]. Variable escape genes that escape in a small fraction of individuals may be incorrectly included as a commonly silenced gene. The observed read counts may also be sequence errors.

In order to account for these potential artifacts and infer XCI states rigorously, we adopted a likelihood-based approach. If the observed read counts are due to sequence errors [with probability p_{err}], we assume that the read count follows $Bin(N, p_{err})$. If the read counts come from an silenced gene [with probability $p_s(1-p_{err})$ and p_s is the fraction of silenced genes], we assume that they follow a beta-binomial (BB) distribution, which allows for over-dispersion [26], i.e., $BB(N, \alpha_s, \beta_s)$. Finally, if the reads come from an escape gene [with probability $(1-p_s)(1-p_{err})$], we assume that they follow $BB(N, \alpha_{esc}, \beta_{esc})$.

Together, the observed read counts are assumed to follow the full mixture model (M_{full}) below:

$$N_1 \sim \begin{cases} BB(N, \alpha_s, \beta_s) & \text{with probability } (1-p_{err}) \times p_s \\ BB(N, \alpha_{esc}, \beta_{esc}) & \text{with probability } (1-p_{err}) \times (1-p_s) \\ Bin(N, p_{err}) & \text{with probability } p_{err} \end{cases} \tag{1.4}$$

The model parameters for M_{full} are estimated for each sample separately.

1.2.3 Determination of the Mixture Components Using Akaike Information Criterion

While the full model M_{full} incorporates all possibilities, it is of practical interest to determine for each sample if any commonly silenced gene escapes in the sample and if the reads contain sequencing errors. In addition, the downstream XCI inference would also benefit from a more parsimonious model with fewer mixture components, as the skewing estimates can be more accurate.

We determine the optimal number of mixture components using Akaike Information Criterion (AIC) [27] based variable selection. We consider 4 possible models for the read count from one haplotype allele. Specifically, the 3 models that we consider along with

M_{full} are:

$$N_1 \sim \begin{cases} BB(N, \alpha_s, \beta_s) & \text{with probability } p_s \\ BB(N, \alpha_{esc}, \beta_{esc}) & \text{with probability } 1 - p_s \end{cases} \quad M_2$$

$$N_1 \sim \begin{cases} BB(N, \alpha_s, \beta_s) & \text{with probability } (1 - p_{err}) \\ Bin(N, p_{err}) & \text{with probability } p_{err} \end{cases} \quad M_1$$

$$N_1 \sim \{ BB(N, \alpha_s, \beta_s) \quad M_0$$

The model with the smallest AIC will be selected. Based upon the selected model, the parameters of interest can be estimated using a maximum likelihood approach. The sample skewing estimate are given by

$$\hat{f} = \frac{\hat{\alpha}_s}{\hat{\alpha}_s + \hat{\beta}_s} \tag{1.5}$$

$$Var(\hat{f}) = \frac{\hat{\alpha}_s \hat{\beta}_s (\hat{\alpha}_s + \hat{\beta}_s + N)}{(\hat{\alpha}_s + \hat{\beta}_s)^2 (\hat{\alpha}_s + \hat{\beta}_s + 1) N} \tag{1.6}$$

1.3 Inference of XCI Escape States

In order to perform hypothesis testing and infer the XCI states, we compare the observed ASE of each gene to the sample skewing.

For a given gene g, we test the hypothesis:

$$H_0 : f_g = \hat{f} vs f_g > \hat{f} \tag{1.7}$$

using the t-statistic

$$T = \frac{f_g - \hat{f}}{\sqrt{var(f_g)}} \tag{1.8}$$

Where $\hat{f}$ is the skewing estimated in the first step and $f_g = \frac{N_{g1}}{N_{g1}+N_{g2}}$ is the observed ASE

ratio for gene g.

Under H_0, the variance for f_g satisfies

$$Var(f_g) = \frac{N_g\hat{\alpha}_s\hat{\beta}_s(\hat{\alpha}_s + \hat{\beta}_s + N_g)}{(\hat{\alpha}_s + \hat{\beta}_s)^2(\hat{\alpha}_s + \hat{\beta}_s + 1)N_g^2} \tag{1.9}$$

The p-value can be approximated from normal distribution. We also calculate the exact p-value based upon the beta-binomial distribution as

$$p = Pr(f_g > \hat{f}|\hat{f}) = \frac{1}{2}\sum_{k>\hat{f}} d_{BB}(k;\hat{\alpha}_s,\hat{\beta}_s) + \frac{1}{2}\sum_{k\geq\hat{f}} d_{BB}(k;\hat{\alpha}_s,\hat{\beta}_s) \tag{1.10}$$

The mid-p procedure was used in the above formula to account for discreteness in the exact p-values.

1.4 Simulation Studies

Briefly, we first performed simulations to evaluate sample skewing estimates. We then compared estimates of the skewing from XCIR, BayesMix, and Xi-threshold. Finally, we obtain predicted XCI states for all genes in the test set and computed type 1 error and power across methods.

1.4.1 Simulation Scenario

For our simulation scenarios, we use distributions and parameters that closely match what we observe in real data such as the GEUVADIS dataset [28].

First, we simulated samples true skewing parameters using a range of skewing mean $\mu \in (0.15, 0.25, 0.30)$ and variance $\sigma^2 \in (4 \times 10^{-8}, 2 \times 10^{-4}, 1 \times 10^{-3})$ where the skewing mean represents the true level of mosaicism in an individual and the variance represents how much variability there is in the observed ASE around the skewing mean. Such that we get a representative range of individuals, from very skewed with low variance of ASE

($\mu = 0.15,\ \sigma^2 = 4 \times 10^{-8}$) that are typically easier to predict, to individuals that have a much more balanced skewing with large variation in their ASE ($\mu = 0.35,\ \sigma^2 = 1 \times 10^{-3}$) that are much harder to predict and better separate methods on their ability to study a larger part of the population. Although the skewing mean ranges from 0 to 0.5, at both extremes, the power to infer XCI-states is either too high or too low for all methods, and as a result, not useful to compare the accuracy of different approaches.

To evaluate the skewing estimation step of XCIR and BayesMix under different conditions, we simulated 60 samples for each of the 9 combinations above and split the samples evenly based on the quality of their set of 40 training genes: The first 20 samples have 40 properly silenced genes, the next 20 have 10% of their training genes that are sequencing errors where the SNP used is actually homozygous. The last 20 have 15% of the training genes that escape. To evaluate the type 1 error and power associated with each method, we also generated 100 silenced and 100 escape test genes, not included in the training set.

The simulation scenario does not assume the availability of phasing information. For every gene, ASE ratios are observed from a single, highly expressed SNP. We generate total read counts and ASE ratios as follow:

- The read depth N is simulated according to a negative-binomial distribution $N \sim NB(\mu = 113, \theta = 0.83)$
- For silenced genes in both training and testing, the allelic expression is simulated according to a beta-binomial (BB) distribution $N_1 \sim BB(N, \mu, \sigma^2)$ where μ and σ^2 are the sample specific true skewing mean and variance as described above.
- For escape genes in the training (errors) and test set, the allelic expression is simulated according to $N_1 \sim BB(N, \alpha_{esc}, \beta_{esc})$
- Training genes with sequencing errors allelic expression are sampled from a binomial distribution $N_1 \sim Bin(N, 0.01)$

While we chose a constant rate of sequencing errors to facilitate the interpretation of the results, additional simulations show that the results of XCIR remain valid and XCIR retains its advantages when assuming a variable sequencing error rate (**Figure A.3**).

1.4.2 Simulation Results

Simulation of skewing

Overall, the XCIR method gave much more accurate estimates of the sample skewing compared to the BayesMix model (**Figure 1.1a**). The mean squared error of Xi-threshold's and BayesMix's skewing estimates is on average 6.7 and 7.4 times that of XCIR, respectively. In bulk RNA-Seq, the presence of sequencing errors in some training genes may decrease average ASE and this can bias the estimates of the skewing towards 0 (**Figure A.1a**). On the other hand, if the training set contains genes that escape XCI in the sample, the skewing estimate can be biased toward 0.5 (**Figure A.2a**). In our simulations, we considered the possibility of both artifacts with one-third of the samples containing sequencing errors and one-third containing genes that escape XCI in the training set. Xi-threshold and BayesMix do not account for these sources of bias, leading to the decreased accuracy.

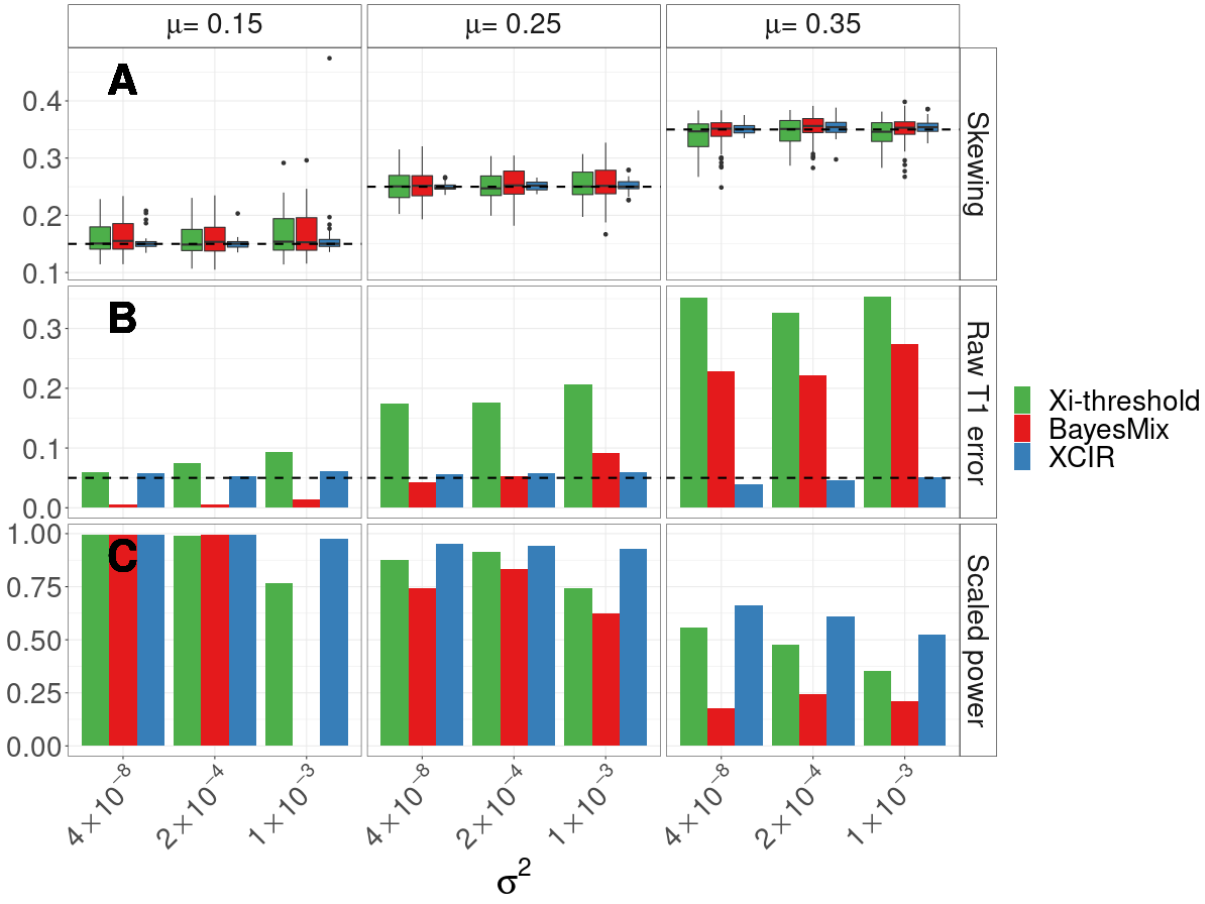

Figure 1.1. Comparison of skewing and XCI state estimates in XCIR, BayesMix, and Xi-threshold for different XCI skewing means (μ) and variances (σ^2) of the true skewing. **(A)** Skewing estimates. The dashed line indicates the true mean. **(B)** Type 1 error. The dashed line indicates the significance threshold 0.05. **(C)** Rescaled power. The Xi-threshold and BayesMix posterior probability of escape cutoffs are adjusted until a type 1 error of 5% is achieved. Power is then computed at the recalibrated thresholds for all three methods. Scaled power of 0 for BayesMix indicates that a type 1 error of 0.05 or less in the training set can only be achieved using a very high PPE threshold, thus classifying every gene as silenced.

Simulation of the Type 1 Error and Power

In addition to assessing XCI skewing, we also performed simulations to evaluate the ability of the three models to correctly identify XCI states in samples with varied XCI skewing. We simulated 100 silenced genes as well as 100 escape genes with different levels

of Xi expression in order to compare the type 1 error and power for detecting XCI escape genes.

XCIR has well-controlled type 1 error across all nine scenarios with different skewing and variance parameters (**Figure 1.1b**). In contrast, using default cutoffs for the Xi-threshold (the Xi expression >10% of the Xa expression) and BayesMix (PPE >50%) models, the type 1 error depends on the sample skewing and becomes unacceptably high for less skewed samples. For example, at a significance threshold of α = .05 when the sample skewing mean is .35, and its variance is 1×10^{-3}, the type 1 errors for Xi-threshold and BayesMix are 35% and 27% respectively (4.7% for XCIR).

In order to evaluate power with controlled false positive rates, we recalibrated the cutoffs used in the Xi-threshold and BayesMix models such that both methods have a controlled type 1 error on the training set with $T1 \leq \max(T1_{XCIR}, 0.05)$ (**Figure 1.1c**). Power is consistently higher in XCIR compared to the other approaches. All methods are adversely affected by decreased sample skewing and increased variance of ASEs across commonly silenced genes in the training set; however, XCIR retains higher power than the other approaches even for samples with less skewing or large variance. We further compared type I errors and power in samples with only sequencing or only training errors (**Figure A.1, A.2**), in order to separately assess their impact on the performance of different methods. Finally, we allowed for randomness in sequencing errors across genes (**Figure A.3**). XCIR outperforms other approaches under these scenarios as well.

1.5 Validation in Single Cell Clone Lines

To complement our simulations, we evaluated the methods using experimental data from mixes of single-cell derived clonal lines isolated from the lymphoblastoid cell line (LCL). Cell lines derived from a single-cell all have the same Xa/Xi assignment, that is, they are non-randomly X inactivated, and hence allow direct assessment of XCI status, as any gene that escapes XCI will show bi-allelic expression. To reflect different levels of XCI skewing, we experimentally generated mixes of two single-cell derived clonal lines

with different Xa/Xi assignments (**Figure 1.2**). For each mixed sample we performed RNA-seq and evaluated ASE . This approach allows us to estimate the type 1 error and power empirically, as escape genes are inferred in the mixes and the accuracy of results can be evaluated using the XCI states observed in the non-randomly inactivated single-cell derived lines. We applied XCIR, BayesMix, and Xi-threshold methods to the data. For BayesMix, we considered both flat and informative prior for single sample analysis.

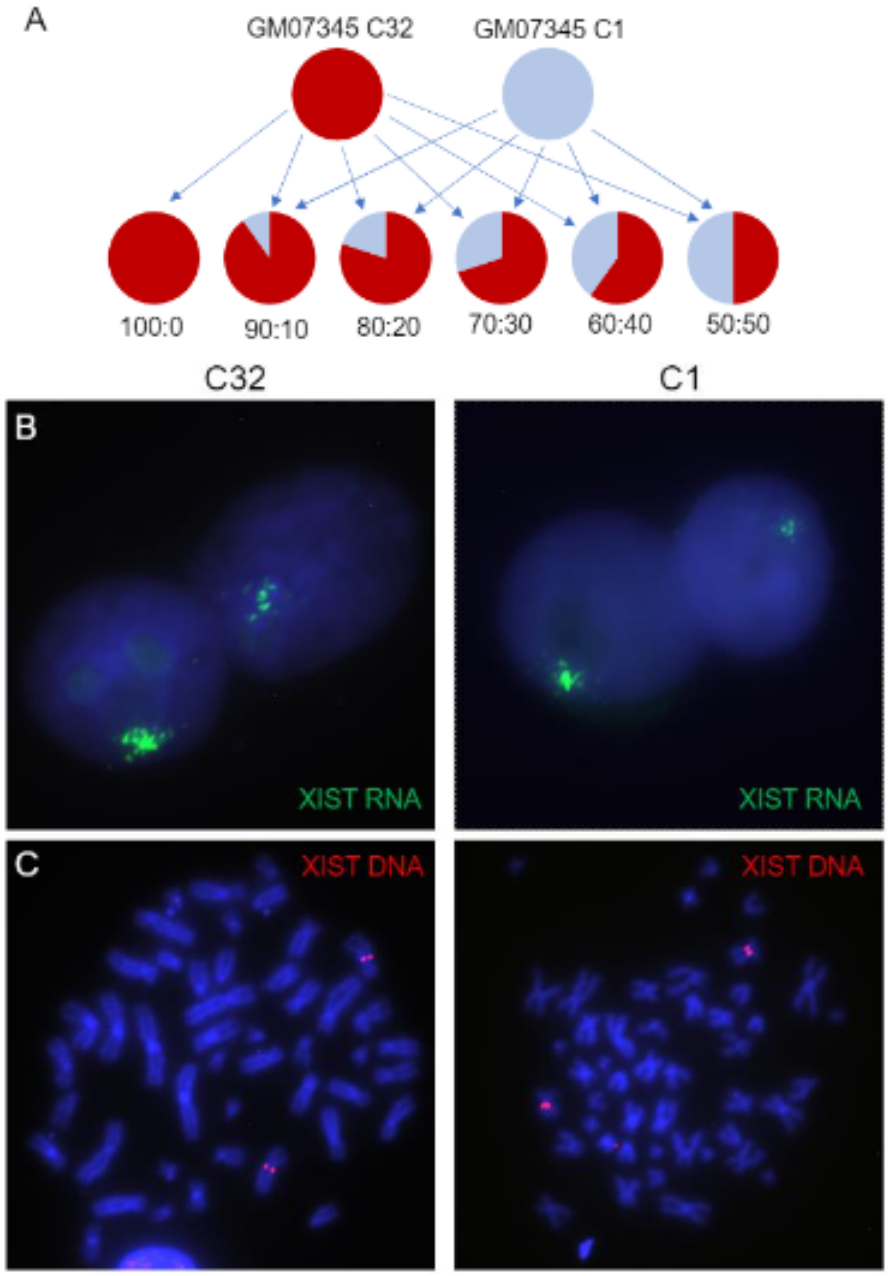

Figure 1.2. Single cell-derived lymphoblast cell line mixing experiments. (A) Overview of clonal cell line mixing strategy to generate "mosaic" samples. RNAseq was then performed and independently analyzed for each “mosaic” mixed sample. **(B)** XIST RNA FISH of clonal LCLs to verify X chromosome content. For clones c32 and c1, 89% and 88% of nuclei, respectively, were positive for an XIST signal (at least 100 cells scored). **(C)** FISH on metaphase chromsomes using an XIST cosmid probe to verify X chromosome content of clonal LCLs. For clones c32 andc1, 95% and 100%, respectively, contain two X chromosomes (>20 metaphases scored)

1.5.1 Cell Lines Preparation

Single-cell derived clonal lines were isolated from the mosaic LCL GM07345 by plating into 0.7% Methocel (Dow Chemical Company) essentially as described [29, 30]. About 400 cells per 35mm dish were diluted in methocel:RPMI media and plated over irradiated mouse embryonic fibroblasts. Individual colonies were picked after 4-6 weeks with sterile pipet tips under an inverted microscope, transferred to a 48-well plate and expanded. Clonality was established by methylation at the Androgen Receptor locus [31] and validated by RNA-seq, confirming that individual lines with maternal or paternal Xa were isolated.

To facilitate the evaluation of XCIR, BayesMix and Xi-threshold in mosaic cell lines, the single-cell clonal lines with differing Xa were mixed to generate five mosaic lines with an expected sample skewing of 50:50, 60:40, 70:30, 80:20 and 90:10. RNA-seq was performed on each of the single-cell clonal lines and on the five "mosaic" samples.

The reads were aligned using *HISAT2* [32] and adjusted for potential reference bias with *WASP* [33]. Allelic expression level at each heterozygous SNP position was quantified using *samtools mpileup* (v1.3.1). The reference-based haplotype phasing was performed with *SHAPEIT* (v2.r837), incorporating the 1000 Genome Project's phase 3 panel as the reference. ASE was computed on genes with at least 20 read counts. Based upon the quantified ASE, we applied XCIR, Xi-threshold, and BayesMix to infer the XCI states for each mosaic mixes. The true XCI states for these synthetic mosaic cell lines can be directly observed from the two single-cell clonal lines.

1.5.2 Performance in Single Cell Clone Lines

The XCI status for each informative, well-expressed gene was confirmed by the RNA-seq data from the non-randomly inactivated single-cell derived lines. 83 genes are expressed from only one X and were deemed X-inactivated, and 28 genes were expressed from both

Xs and therefore escape XCI (**TableB.1**). Of the silenced genes, 57 (68%) were part of the training set of commonly silenced genes [34](Balaton et al. 2015).

Evaluations based upon the clonal cell line mixes are concordant with the simulation experiment, with XCIR outperforming existing approaches. XCIR has well-controlled type 1 error in all four mixes including the less skewed 60:40 and 70:30 mixes (**Figure 1.3a**). Based upon the default cutoff values, both Xi-threshold and BayesMix methods have very high type 1 errors in all samples, particularly those that are less skewed. Using single sample analysis with the suggested flat prior, the observed type-I error rate for BayesMix lies between 22% and 40%. Similarly, the type 1 error rate for Xi-threshold lies between 7% and 37%.

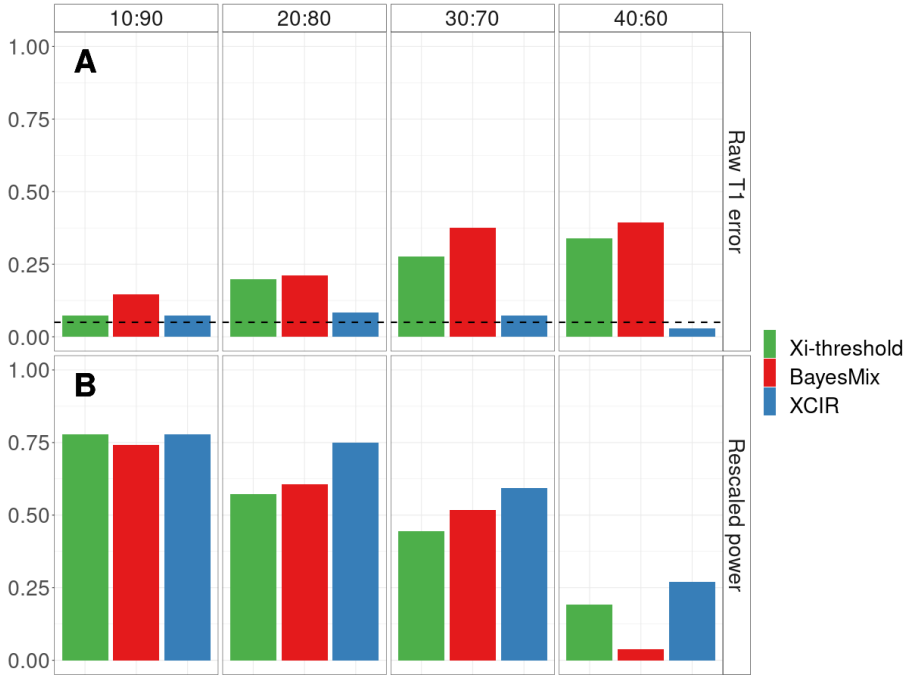

Figure 1.3. XCI inference in single–cell clone mixing experiment. The proportion of each single-cell clonal line in the mixed sample is indicated at the top of the panel and is equivalent to the true skewing of each sample. **(A)** Raw type 1 error. **(B)** Rescaled power at the empirical threshold for 5% type 1 error

Method	Raw Type 1 Error	Recalibrated Power	Recalibrated Threshold	Experimental clone mix
Xi-threshold	0.338	0.192	0.0368	40:60
	0.275	0.444	0.1216	30:70
	0.197	0.571	0.1451	20:80
	0.072	0.778	0.3297	10:90
BayesMix	0.394	0.038	0.9582	40:60
	0.377	0.519	0.7517	30:70
	0.211	0.607	0.8631	20:80
	0.145	0.741	0.7874	10:90
XCIR	0.028	0.269		40:60
	0.072	0.593		30:70
	0.085	0.750		20:80
	0.072	0.778		10:90

Table 1.2. Single cell-derived lymphoblast cell line mixing experiments. Type 1 error and power comparison across methods for the single cell-derived lymphoblast cell line mixing experiment. The Recalibrated thresholds show the values of Xi-threshold or BayesMix's posterior probability of escape that yield a type 1 error of less than 5% in the training set. Power is reported after recalibration.

1.6 Application to GEUVADIS

To further quantify XCI states in multiple samples using a population-scale dataset, we applied the pipeline to the GEUVADIS dataset, which contains RNA-seq data for 217 female LCLs. DNA genotypes for the same set of individuals were obtained from the 1000 Genomes Project phase 3 (The 1000 Genomes Project 2015).

1.6.1 Determining XCI States with XCIR

Processing of GEUVADIS Dataset

We utilized the same procedure to process the sequence data as in the analysis of the single-cell clones, including the alignment, the adjustment of reference bias, and the pileups. The phased haplotype information was extracted from the 1000 Genomes data. Reads that covered multiple heterozygous SNPs on each haplotype are aggregated for the inference of XCI states. On average, each gene is covered by 1.65 SNPs (1.66 when including only skewed samples). Sufficient coverage was available for an average 22.22

genes per sample out of the 177 available in the training set (22.07 per sample for skewed samples alone).

XCI Inference in the GEUVADIS dataset

Given that ASE-based methods such as XCIR have maximal power to detect escape genes in relatively skewed samples, we restricted our analysis to the 136 samples with skewing greater than 25:75. As a result, while the full dataset contains 351 genes with at least one heterozygous SNP covered at a sufficient depth, only the 215 genes that could be scored in at least 10 of the 136 skewed samples were used for the final classification. The GEUVADIS samples were part of the 1000 Genomes Project that included DNA-sequence genotypes and phased haplotypes. While the default input for our XCIR analysis takes the read depths of a single heterozygous SNP, aggregating reads from multiple heterozygous SNPs on phased haplotypes can potentially improve power.

We classified genes using previously established criteria [6]; escape genes were expressed from the inactive X in >75% of individuals, genes that escape in less than 25% of the individuals were deemed X-inactivated, and those that escape in 25% to 75% of the individuals were classified as variable escape genes. Following these criteria, 165 (76.7%) genes were found to be X-inactivated, 20 (9.3%) were predicted to escape XCI and 30 (14%) showed variable XCI escape in the dataset.

As described above, XCIR includes features to identify if the observed ASE ratio for a SNP is due to sequencing errors or the inclusion of an escape gene within the commonly silenced training set genes. Subsequently, the likelihood of each scenario can be quantified, and the model that best fits the data was used to infer XCI states. Applying the method, we noted that 28% of the samples were fitted with a two or three-component mixture model, which emphasizes the necessity of correcting for both confounding errors.

Using XCIR's ability to reliably identify inter-individual variability, we refined the previous classification of X-linked genes (Table S4, S5). Compared to the consensus [34], we identified five new genes variably escaping XCI that were previously classified as strictly X inactivated (CXorf40B, DMD, EMD, OTUD5, SASH3). Finally, our results

confirm known escapers as we did not reclassify any of the escape genes identified in all prior studies.

While the underlying XCI landscape for all GEUVADIS samples is not known, results were confirmed with multiple lines of evidence. Because of their presence on both the X and Y Chromosomes, pseudoautosomal (PAR) genes are expected to escape XCI. Indeed, the 10 PAR1 genes with sufficient read counts were predicted as escape XCI in over 80% of the subjects. Furthermore, of the 124 training set genes that could be scored in this dataset, only 3 escape in more than 25% of the subjects (SEPT6, MST4, and MAP7D2). It is notable, however, that other studies also classified these three genes as variably escaping XCI [24, 34], which confirms both our findings and the necessity to account for potential errors in the training list. While such findings could support refining the list of training set genes, they perhaps more importantly underscore that the tools integrated into XCIR have tremendous merit for defining the XCI landscape in larger datasets and cell types where XCI states are less well described.

1.6.2 Differential Gene Expression for X-linked Genes

We also examined expression between sexes using the 244 male lines included in the GEUVADIS project. The alignment-free method *kallisto* [35] was used to estimate transcript abundance, which is measured by transcript per million (TPM). The R package limma [36] (v3.30.13) was used to conduct differential expression analysis of males vs. females for the GEUVADIS dataset. Using a false discovery rate threshold of 0.05, we found 33 differentially expressed genes on the X Chromosome.

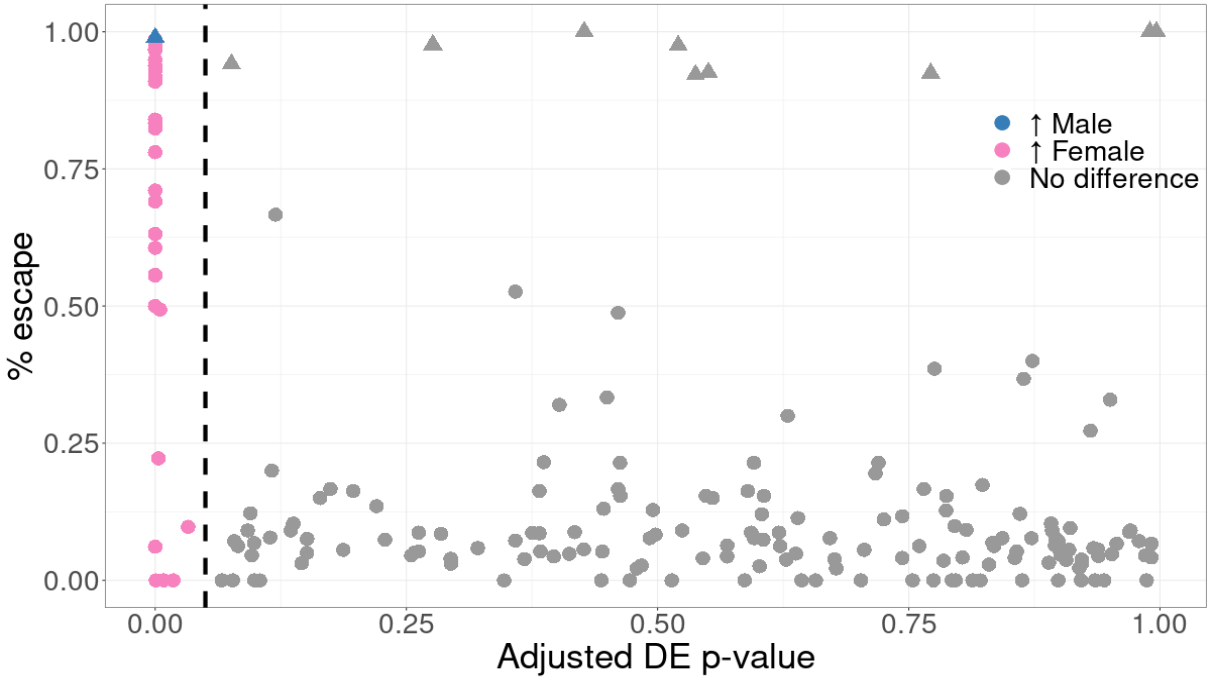

Figure 1.4. Genes that escape XCI are differentially expressed. Adjusted differential expression (DE) p-values as reported by *limma* for 215 genes against the frequency of samples that escape XCI as predicted by XCIR in GEUVADIS. The dashed line indicates the 5% significance cutoff and the significantly differentially expressed genes are colored based on the sex where increased expression is observed. Escape genes are mostly female-biased, reflecting expression from both X copies. PAR genes (triangles) are correctly identified as escaping XCI despite most of them not being significantly differentially expressed, as the expression on the Y is similar or higher to that on the Xi.

Because most escape genes lack functionally equivalent Y homologs, we sought to examine whether escape genes are more likely to be differentially expressed between sexes. We asked if genes with significant differential expression were those that escape in the most individuals. Our analyses were focused on the 215 X-linked genes for which XCI calls were available in at least 10 skewed samples (**Figure 1.4**). As expected, we found that genes predicted to escape XCI are up-regulated in females and show a significant difference in expression vs. males. When a Y homolog exists, such as for genes in the pseudoautosomal regions (PAR), there is expression from both sex chromosomes. Although most of these genes fail to show significantly increased expression in females,

XCIR consistently predicted a very high percentage of escape for the ten PAR1 genes available in the dataset (triangles in Figure 1.4). While differential gene expression analysis cannot reliably assess individual-level XCI states, XCIR can be successfully applied to infer XCI states of genes in each individual, provided the presence of an informative expressed SNP. Moreover, we found that five of these PAR1 genes show male-biased expression, including ZBED1 which was significantly differentially expressed. This observation is consistent with a previous report that escape from XCI in these genes is only partial and can lead to higher expression on the Xa and Y than on the Xi [8, 37].

1.7 Additional Methodology

1.7.1 Impact of phasing on prediction accuracy

When multiple heterozygous SNPs are available As described in table 1.1, when multiple intragenic heterozygous SNPs are available, we can sum the allelic read counts N_1, N_2 across SNPs to improve the precision of the ASE ratios f_g and eventually improve the accuracy of the XCI states prediction.

In order to quantify this improvement, we generate new simulations.

- Sample individual skewing and errors using the same setup as in section 1.4.
- For each gene g in both training and testing
    1. Sample an additional SNP
    2. If g contains a sequencing error, set one of the two SNP as homozygous
- Process the data with XCIR using either the read count of top SNP or summing allelic expression over SNPs.

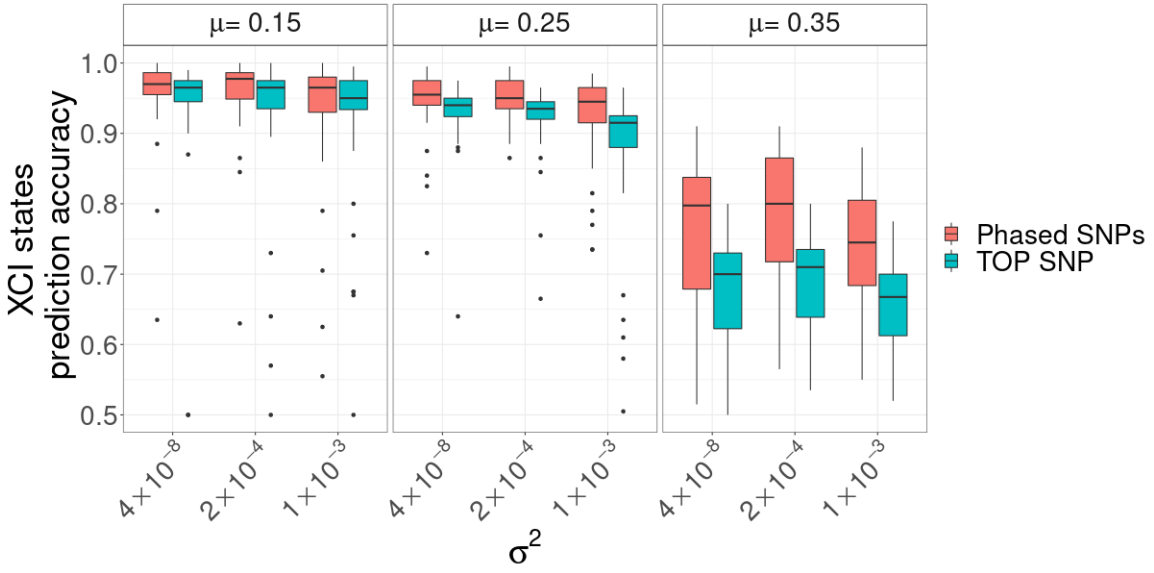

Figure 1.5. Effect of phasing on prediction accuracy. Boxplots show the distribution of the prediction accuracy across skewings (μ) and ASE variance (σ^2) using either the read count of the top SNP (Orange) or using aggregated SNPs on the haplotype (Blue)

We show that using phasing information can increase the accuracy of XCI state predictions, especially in less skewed samples with an average increase of 11% (**Figure 1.5**). Specifically, In GEUVADIS samples, using haplotype information increases the average per-gene read depth by 48% from 81 to 120.

We further note that because XCIR gracefully handles the introduction of sequencing errors in the training set, using haplotype level data does not significantly improve skewing estimates and the performance gains are due to the decreased bias ASE ratios.

1.7.2 Impact of eQTLs on XCI Inference

Similar to other expression-based approaches, XCIR may be influenced by the presence of eQTL regulatory variants, which may also contribute to allelic imbalance. In the estimation of skewing, multiple genes are used and hence the eQTL effects may be cancelled out across genes and have negligible effects on the estimates. However, at the individual XCI calls level, eQTLs may influence ASE.

In order to study the impact of regulatory variants on XCI inference, we show that

under realistic conditions, the effect of regulatory variants is limited and that by regressing the allelic expression over the eQTL genotypes and using the residuals as input for XCIR, we are able to recover most of the lost accuracy (**Figure 1.3**).

Simulation Scenario

The sampling of individual skewing follows the same setup as in section 1.4.

- Sample minor allele frequency (MAF) from the SNPs identified in whole blood in GTEx project.
- For each test gene we simulate an eQTL based on
- In sample that carry an effect allele, we assume it can increase or decrease the ASE ratio by modifying the true allelic expression (N_2) as follow

$$\begin{aligned} C &= N_1(\beta_{eQTL} - 1) \\ N_2^O &= N_2 \pm C \end{aligned} \tag{1.11}$$

  Where β_{eQTL} is the median allelic fold change (aFC) effect size of SNPs in GTEx. so that the eQTL effects reflect realistic fold changes. We obtain a new 'observed' allelic count N_2^O

- XCIR is then used to analyze the gene expression N_1,N_2'.
- When a list of known eQTL is available, we can adjust their impact by regressing the observed allelic expression on the number of alternate alleles (ALT).

$$\begin{aligned} N_2^O &\sim P(\mu) \\ \log(\mu) &= \beta_0 + \beta_{ALT} \cdot \text{ALT} \end{aligned} \tag{1.12}$$

- XCIR is then used to perform XCI inference with observed or adjusted expression.

eQTL simulation results

We find that under realistic conditions, the effect of regulatory variants is limited and that by regressing the allelic expression over the eQTL genotypes and using the residuals as input for XCIR, we are able to recover most of the lost accuracy (**Table 1.3**). Importantly, how and whether cis-regulatory variants influence XCI escape have never been addressed but may be revealed by such analyses.

eQTL effect	Type 1 Error		Power	
	No adjustment	Adjusted	No adjustment	Adjusted
Up	0.071	0.059	0.816	0.826
No eQTL	0.057	0.061	0.840	0.840
Down	0.048	0.056	0.800	0.835

Table 1.3. Effect of eQTL on XCI inference. Type 1 error and power comparison in the presence of cis-eQTL with and without adjusting for the genotypes effect.

In summary, we showed that while eQTLs may affect the estimation of skewing and XCI status inference, the overall impact of eQTLs on allelic imbalance can be mitigated for X-linked genes and the impact on type I errors and power was minimal. If we have a comprehensive catalog of eQTLs for X-linked genes, we can adjust for the effect using simple off-the-shelf methods such as Poisson regression, which can help further reduce the bias.

Chapter 2
Application of XCI Inference

2.1 Introduction

In Chapter 1 we saw that improvement in the modelling of sample skewing lead to significant improvements in the inference of XCI states. In Chapter 2, we attempt to derive biological insight from our new inference approach. With high confidence, sample specific, classification of X-linked genes we can compare XCI states across disease or treatment status.

2.2 Software Implementation

Our method is implemented in an R [38] package complete with examples and full documentation available

BioConductor : `https://www.bioconductor.org/packages/release/bioc/html/XCIR.html`

GitHub : `https://github.com/SRenan/XCIR`

While having a more complex statistical model than some existing approaches, XCIR remains computationally efficient. Analysis of the full GEUVADIS dataset is done under 1 minute using a standard computer server (with Intel(R) Xeon(R) CPU E5-2680 v2 and 128GB RAM), including reading the data, fitting all mixture models for the skewing estimates, and classification of X-linked genes. In comparison, using BayesMix, the

estimation of the sample skewing alone exceeded 2 hours and the PPE could not be computed on a single core. Likewise, the analysis of 720 simulated samples with 107 genes each takes under two minutes with XCIR. BayesMix required massive parallelization and the processing of a subset of 36 samples takes over 2 hours.

2.3 Heritability Analysis of XCI States in UK Biobank

We next sought to determine whether XCI state influences X-linked disease heritability. Disease heritability of X genes was estimated by extending LD score regression. We first annotated the entire X Chromosome by XCI states based upon our analysis of GEUVADIS data, and quantified the heritability for 319 self-reported phenotypes in pooled (while adjusting for sex as a covariate) and sex-specific GWAS of the UK Biobank [39].

To examine the contribution of inactive and escape genes, we partitioned heritability across XCI states. Enrichment is defined as the ratio of the proportion of heritability explained over the proportion of SNPs for a specific functional category. Overall, in the self-reported phenotypes, the analysis of traits in combined sexes indicates that the heritabilities from variable escape and escape genes are significantly enriched (mean enrichment of 5x and 3.8x) relative to silenced genes (1.3x) and intergenic SNPs (0.5x). These enrichment patterns are also observed when analyzing each sex separately (**Figure 2.1**).

2.3.1 Partitioning Heritability on the X Chromosome

LD score regression

The original LD score regression software did not support X Chromosome (as of version 1.0.0). We extend the method to analyze X-linked genes. We first estimate the LD score by pooling males and females, so that the sample size can be maximized. As a majority of genes on X is XCI-silenced, for each SNP, we encode male genotypes as 0 and 2 or 0,1 and 2 for the PAR regions, and female genotypes as 0, 1 and 2. We call this XCI coding.

The coding for the dosage of the expressed alternative alleles for escape gene is complicated. When the gene fully escape XCI and both the Xa and Xi have equal dosage as the male X, we may encode males as 0, 1 and females as 0,1,2. When the gene partially escape XCI, i.e., the Xi only expresses partially and the genes may only escape in a subset of the individuals, it is hard to encode the genotype to reflect the actual allelic dosage. However, as we will show below, the additive XCI coding will consistently underestimate the genetic effect per expressed allele for E/VE genes, which will in turn lead to underestimated heritability by E/VE genes. The enrichment of heritability for escape and variable escape genes would only be stronger if the correct coding were known.

Specifically, we assume that the sample skewing is f which is the fraction of cells where the reference allele is on the Xa. We denote the Xi expression level as E_{Xi}. The dosage of the expressed alternative allele equals to

$$E_{alt}\begin{cases} 0 & \text{if genotype is REF/REF} \\ fE_{Xi} + (1-f) & \text{if genotype is REF/ALT} \\ (1+E_{Xi}) & \text{if genotype is ALT/ALT} \end{cases}$$

We would like to estimate the genetic effect as the change of phenotype means per unit of change in the expressed alternative allele dosage. If we regress the phenotype over the XCI coding, the estimated genetic effect is biased downward. Specifically, assume the genotype frequencies for REF/REF, REF/ALT and ALT/ALT are a_{00}, a_{01} and a_{11}, and the underlying genetic model is

$$Y = \beta_{alt} E_{alt} + \epsilon$$

If we calculate the genetic effect based upon the XCI coding, G_{XCI} from the model

$Y = \beta_{G_{XCI}} + \epsilon$ using least square estimate:

$$\hat{\beta}_{XCI} = \frac{\sum_i G_{XCI,i} Y_i}{\sum_i G^2_{XCI,i}}$$

The estimate satisfies

$$E[\hat{\beta}_{XCI}] = \frac{a_{01} f(E_{Xi} + (1-f)) + a_{11}(1 + E_{Xi})}{a_{01} + 2a_{11}} \beta_{ALT}$$

To compute the LD score for the 3.3 million SNPs (with MAF>1%) on Chromosome X, we use 503 European samples from the 1000 genome project. For each SNP, we calculate its correlations with the SNPs within the 1 million basepair window and estimate the LD score as:

$$l_j = \sum_k r^2_{jk}$$

Assuming no confounding factors in the GWAS dataset (such as cryptic relatedness or population structure), the expected χ^2 statistic of variant j given its LD score is approximately

$$E[\chi^2 | l_j] = \frac{N h^2 l_j}{M} + 1$$

Where N is the sample size and M is the number of SNPs.

While some phenotypes have very imbalanced case:control ratios, our analysis is limited to common variants and as such does not result in inflated type 1 error [40].

Partitioning heritability

This LD score regression can be generalized to allow the estimation of heritability explained by SNPs in different functional annotation categories. We stratify the LD score regression by XCI states. Using the updated classification obtained from the GEUVADIS dataset, we map each SNP to the nearest gene within 200kb. Genes with at least 5 skewed samples are used to annotate SNPs as Escape, Silenced or Variable-Escape while SNPs mapped to un-annotated genes are annotated as NoCall. Finally, Intergenic SNPs

not within 200kb of the start or end of a gene are annotated as Intergenic. Together, these annotations cover the entire X Chromosome. With these annotations, we can model the mean value of the chi-square statistic as

$$E[\chi^2] = N \sum_C \tau_C l(j, C) + 1$$

Where τ_C is the total contribution to heritability of SNPs in category C and $l(j, C)$ is the LD score of SNP j with respect to neighboring SNPs in category C. Because error terms are correlated for SNPs in LD, standard error estimates were obtained via a block jack-knife over blocks of 2000 adjacent SNPs, providing robust estimates.

Enrichment of Heritability

In order to make the estimates comparable across each XCI state, we compute the heritability enrichment as the fraction of total heritability explained by category C over the proportion of SNPs mapped to C.

$$e_C \frac{h_C^2/h^2}{M_C/M}$$

GWAS summary results

We use GWAS results from the UK Biobank dataset as generated by Neale group. For each phenotype, GWAS summary statistics are available for the combined analysis and sex specific analysis. For all analyses, age and the first 20 principal components are included as covariates. The combined analyses of males and females also include sex and age × sex interaction as additional covariates.

2.3.2 Heritability Enrichment of XCI States

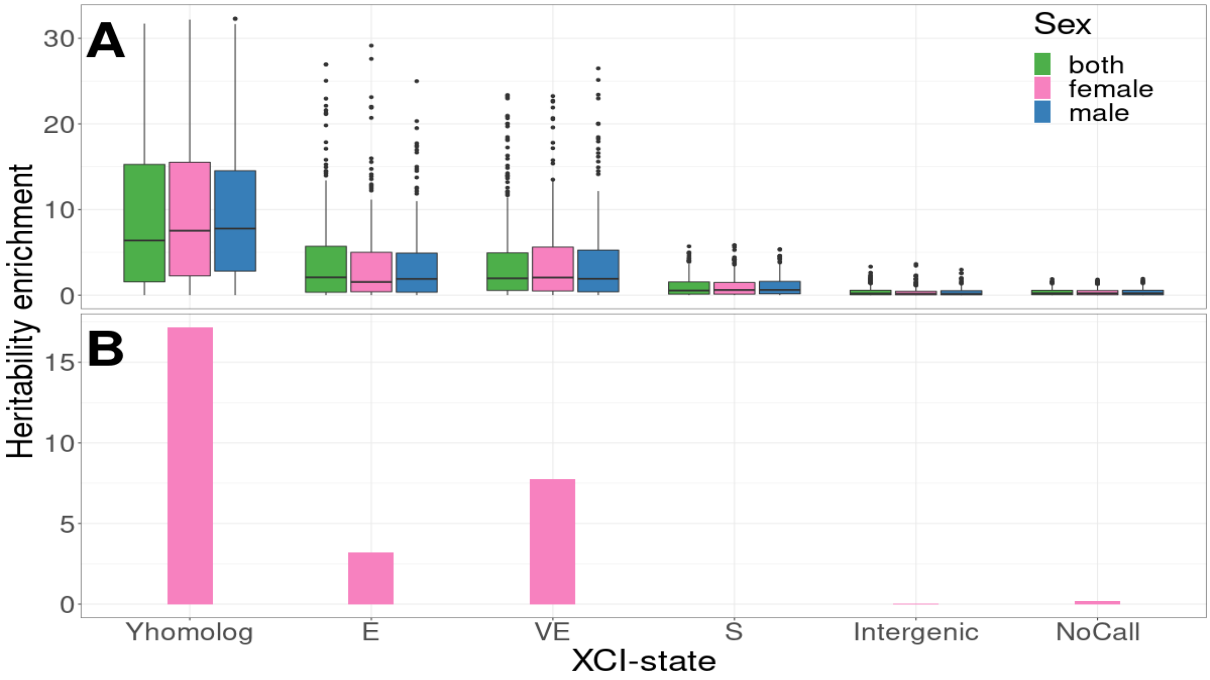

Figure 2.1. Genes that escape and variably escape XCI are enriched for heritability. (A) Distribution of heritability enrichment for all self-reported phenotypes available in females (240 phenotypes), males (218 phenotypes), or both (280 phenotypes). Heritability was independently assessed for each XCI state: escape (E), variably escape (VE), silenced (S). X genes with Y homologs that escape or variably escape are differentiated (Yhomolog). **(B)** Lupus heritability enrichment measured in females (392 cases, 182316 controls).

Notably, we observe that heritability is particularly enriched in escape or variable escape genes with functional Y homologs, including SNPs associated with genes in the PAR regions (9.5x). Such enrichment may reflect evolutionary constraints on the dosage of these genes [41]. This is supported by previous observations that these genes are widely expressed in multiple cell types and function broadly to regulate targets throughout the genome [42]. We noted slightly weaker enrichment for escape genes compared to variable escape genes. Constitutive escape genes are expected to have consistent bi-allelic expression across all females. Therefore, there is little inter-individual differences in

the escape status and the escape genes do not explain the inter-individual phenotypic differences. On the other hand, escape of a typically silenced gene (i.e., Variable Escape) will lead to more dramatic expression changes between individuals, which is likely to affect phenotypes.

While alterations in the XY-homologous genes may lead to broad effects that underlie their role in many disease phenotypes, the more modest overall enrichment observed in escape and variable escape genes that lack functional Y homologs could reflect significant enrichment in a subset of diseases. Moreover, as escape and variable escape genes without Y homologs can result in gene dosage imbalance between males and females, we hypothesize that heritability at these gene loci may be enriched in disorders with a female bias. Indeed, for Systemic Lupus Erythematosus, an autoimmune disease that predominantly affects females, we find that silenced genes explain virtually none of the heritability observed in females, while genes that escape or variably escape XCI lacking Y-homologs are significantly enriched for heritability (3.2x and 7.7x, respectively).

2.3.3 Permutation tests

To test for changes in enrichment between sex-biased and non-biased phenotypes, we use permutation tests. For each XCI state, we randomly shuffle the state of sex-biased and sex-nonbiased for all diseases and calculate the z-scores of the difference of mean enrichment values between the two groups. Repeating the shuffling 5000 times, we obtain an empirical null distribution for the Z-scores of enrichment difference. P-values are obtained by calculating the fraction of resampled Z-scores that are more extreme than the true observed value.

We identified sex-biased phenotypes from the UK Biobank with a female/male ratio greater than 2 among affected individuals. In this subset of 51 phenotypes, we observe a significantly higher enrichment of heritability for variable escape genes than for the 154 traits with a balanced sex ratio (with p-value = 0.007 based upon permutation testing). Importantly, for both female and male-biased diseases (Table 2.1), escape and variable escape genes are the only XCI states that exhibit significant enrichment in these

sex-biased diseases, strongly supporting a role for XCI escape in many female-biased diseases [3].

Analysis	XCI state	Enrichment Ratio: Male-biased / Non-biased	P-value
Differential enrichment in males: Male-biased vs. non-biased	Y homolog	1.08	0.306
	Escape	1.13	0.29
	Variable Escape	1.06	0.387
	Silenced	0.68	0.953
Analysis	XCI state	Enrichment Ratio: Female-biased / Non-biased	P-value
Differential enrichment in females: Female-biased vs. non-biased	Y homolog	0.91	0.761
	Escape	1.14	0.276
	Variable Escape	1.65	0.007
	Silenced	0.91	0.695

Table 2.1. Permutation tests to assess heritability enrichment differences between sex-biased diseases and non-biased phenotypes for each XCI state. Across all categories, only genes identified as variable escape by XCIR show significant enrichment in female-biased diseases.

These results may suggest that overexpression of some X-linked genes, due to XCI escape in a subset of females, plays a previously unappreciated role in diseases that merits further examination

2.4 X-QTL

As we saw in Chapter 1, expression of X-linked genes is a mixture of expression from the active X and inactive X. In section 1.7.2 we quantified the impact of eQTLs on XCI inference.

2.4.1 GWAS eQTL model

In genome wide association studies, we regress a phenotype y on the genotype G

$$y = G\beta + \epsilon \tag{2.1}$$

Using genes expression as a phenotype, we can measure the impact of eQTLs

$$\begin{bmatrix} X_1 \\ \vdots \\ X_N \end{bmatrix} = \begin{bmatrix} X_{11} + X_{12} \\ \vdots \\ X_{N1} + X_{N2} \end{bmatrix} = \beta G + \epsilon \tag{2.2}$$

Where $X_{.1}$ and $X_{.2}$ are expression of each haplotype.

While an alternate allele can modify the expression on either autosomal or sex chromosome, there are multiple ways in which the expression of X can be changed. A variant can alter the overall expression, but it could also affect the probability of a gene escaping XCI. Finally, the expression change may only be enabled in the context of either Xa or Xi.

Importantly, pooled analysis of males and female does not allow to distinguish between sources of variation and has reduced power to detect changes that may only occur in females.

2.4.2 X-QTL models

From bulk RNA-Seq data, XCIR estimates $\tau = \frac{Xi}{X}$, the contribution of Xi to the total gene expression. Using this estimate, we can differentiate Xa/Xi-expression and use the values in regression.

$$\begin{bmatrix} X_{11}+X_{12} \\ \vdots \\ X_{N1}+X_{N2} \end{bmatrix} = \beta G + \epsilon$$
$$\begin{bmatrix} X_1(1-\tau)+X_1\tau \\ \vdots \\ X_N(1-\tau)+X_N\tau \end{bmatrix} = \begin{bmatrix} X_{a1}+X_{i1} \\ \vdots \\ X_{aN}+X_{iN} \end{bmatrix} = \beta G + \epsilon \tag{2.3}$$

In its simplest form, we can run an eQTL analysis on the Xi-expression to compute an effect size specific to the Xi

$$\begin{bmatrix} X_{i1} \\ \vdots \\ X_{iN} \end{bmatrix} = \beta_{Xi} G + \epsilon \tag{M0}$$

In order to account for all possible ways an eQTL may change expression as detailed in section 2.4.1 above, we propose an approach that explicitly models each source of variantion (parameters of interest colored in the model specifications).

$$\underset{N\times 2}{\begin{bmatrix} Xa_1 & Xi_1 \\ \vdots & \vdots \\ Xa_N & Xi_N \end{bmatrix}} \sim \begin{cases} N\left(\begin{pmatrix} \mu_{Xa} \\ \mu_{Xi} \end{pmatrix}, \overset{\Omega_{eXa,eXi}}{\begin{bmatrix} \sigma^2_{Xa} & \sigma_{Xa.Xi} \\ \sigma_{Xa.Xi} & \sigma^2_{Xi} \end{bmatrix}}\right) \text{ with } P(I_{esc}=1) \\ N\left(\begin{pmatrix} \mu_{Xa} \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma^2_{Xa} & \sigma_{Xa.Xi} \\ \sigma_{Xa.Xi} & \sigma^2_{Xi} \end{bmatrix}\right) \text{ with } P(I_{esc}=0) \end{cases} \tag{M1}$$

With

$$\mu_{Xa} = \beta_{Xa,0} + \beta_{Xa} G$$
$$\mu_{Xi} = \beta_{Xi,0} + \beta_{Xi} G \tag{2.4}$$

And

$$I_{esc} \sim Bern(\gamma_{esc})$$
$$logit(\gamma_{esc}) = \beta_{esc,0} + \beta_{esc}G \tag{2.5}$$

Using M1 above, we can test for each parameters of interest and identify different mechanism of action

- Xi-QTL that only affect the expression on the Xi, with β_{Xi}
- Xa-QTL that only affect the expression on the Xa, with β_{Xa}
- esc-QTL that affect the likelihood of the gene escaping, with β_{esc}

M1 can only be used when XCIR calls are available to differentiate between Xa and Xi-expression. This is particularly problematic as most samples do not have informative SNPs for every gene. Furthermore, in any given study, the amount of samples that are skewed enough for XCI inference leads to discarding a significant portion of the data. While the discarded samples do not allow modelling of X-QTLs, they contain information relevant to the models above.

In an effort to increase power, we retain as much information as possible by jointly modelling the samples where Xa/Xi-expression is unknown:

$$\begin{bmatrix} X_1 \\ \vdots \\ X_N \end{bmatrix} \sim \begin{cases} N(\mu_{Xa} + \mu_{Xi}, \sum \Omega) \text{ with } P(I_{esc} = 1) \\ N(\mu_{Xa} + 0, \sum \Omega) \text{ with } P(I_{esc} = 0) \end{cases} \tag{M2}$$

In order to specifically look for esc-QTL, we fit model M3, defined as M2 with the additional assumption that $\beta_{esc} = 0$ such that

$$\mu_{Xi} = \beta_{Xi,0}$$

Finally, we test a model defined as M2 with two degrees of freedom where we jointly test $\beta_{xi} = 0$ &$\beta_{esc} = 0$.

Model	Assumptions	Test (H_0)	All samples
M0		$\beta_{xi} = 0$	×
M1		$\beta_{xi} = 0$	×
M2		$\beta_{xi} = 0$	✓
M3	$\beta_{xi} = 0$	$\beta_{esc} = 0$	✓
2df		$\beta_{xi} = 0$ &$\beta_{esc} = 0$	✓

Table 2.2. Summary of XQTL models assumptions

2.4.3 Simulations

- Simulate data according to M1
- 1000 Samples with Xi/Xa-expression (i.e: samples where XCI calls are available)
- 1000 Samples with overall expression only
- MAF = 0.25
- $\beta_{esc,0}$ = -1 s.t $P(I_{esc} = 1) \approx 25\%$ in the absence of genetic effect on escape (β_{esc}).
- effect sizes are fixed to $\beta \in (0, 0.25, 1)$ to investigate type 1 error and power under realistic or optimistic conditions.
- 1000 Simulations per setting

We then fit all models defined in section 2.4.2 and test the hypothesis summarized in Table 2.2.

Model	β_{xi}	β_{esc}	T1
M0	0	0	0.02
M1	0	0	0.15
M2	0	0	0.20
M3	0	0	0.03
2df	0	0	0.02

Table 2.3. Type 1 error in X-QTL analysis

We find that the more complex models lead to inflated type 1 error but fixing the escape parameters (M5,2df) helps bring it back to acceptable levels.

Model	β_{xi}	β_{esc}	Power
M0	0.25	0	0.18
M1	0.25	0	0.28
M2	0.25	0	0.31
M3	0.25	0	0.43
2df	0.25	0	0.51
M0	1.00	0	0.99
M1	1.00	0	1.00
M2	1.00	0	1.00
M3	1.00	0	1.00
2df	1.00	0	1.00

Table 2.4. Power to detext Xi-QTLs

As expected, under realistic conditions (where $\beta_{xi} = 0.25$ is the median eQTL effect size in whole blood reported in the GTEx project) the more complex model increases the

power to detect Xi-QTLs compared to the naive approach (M0). Including samples that do not have XCI calls in the inference (M2,M3,2df) leads to further increase in power as we make use of all information available in the dataset.

Model	β_{xi}	β_{esc}	Power
M0	0	0.25	0.15
M1	0	0.25	0.09
M2	0	0.25	0.04
M3	0	0.25	0.20
2df	0	0.25	0.19
M0	0	1.00	0.98
M1	0	1.00	0.03
M2	0	1.00	0.05
M3	0	1.00	1.00
2df	0	1.00	1.00

Table 2.5. Power to detect esc-QTLs

Because models M1 and M2 specifically test for Xi-QTLs, their power is negligible and should rather interpreted as type 1 error since β_{xi} is set to 0. On the other hand, we can see that model M3 and 2df which test for esc-QTL show a slight increase in power compared to M0.

Overall, the simulations show that while XQTL models provide clear improvements in the power to detect Xi and esc-QTLs over the naive approach, the results are sensitive to the assumptions as well as the effect sizes.

We note that while these results show some clear improvements over naive approaches, the increased power does not translate very well into real data analysis due to the limited sample sizes and additional sources of variation (noise in expression quantification,

genotyping errors, limited number of skewed samples, etc...). Reaching actionable levels of power will require using the full data, including samples where the Xa/Xi-expression breakdown is not available and possibly even adding males.

The XQTL models have been implemented in R [38] and are available on GitHub (`https://github.com/SRenan/XQTL`).

Appendix A
Additional Simulations

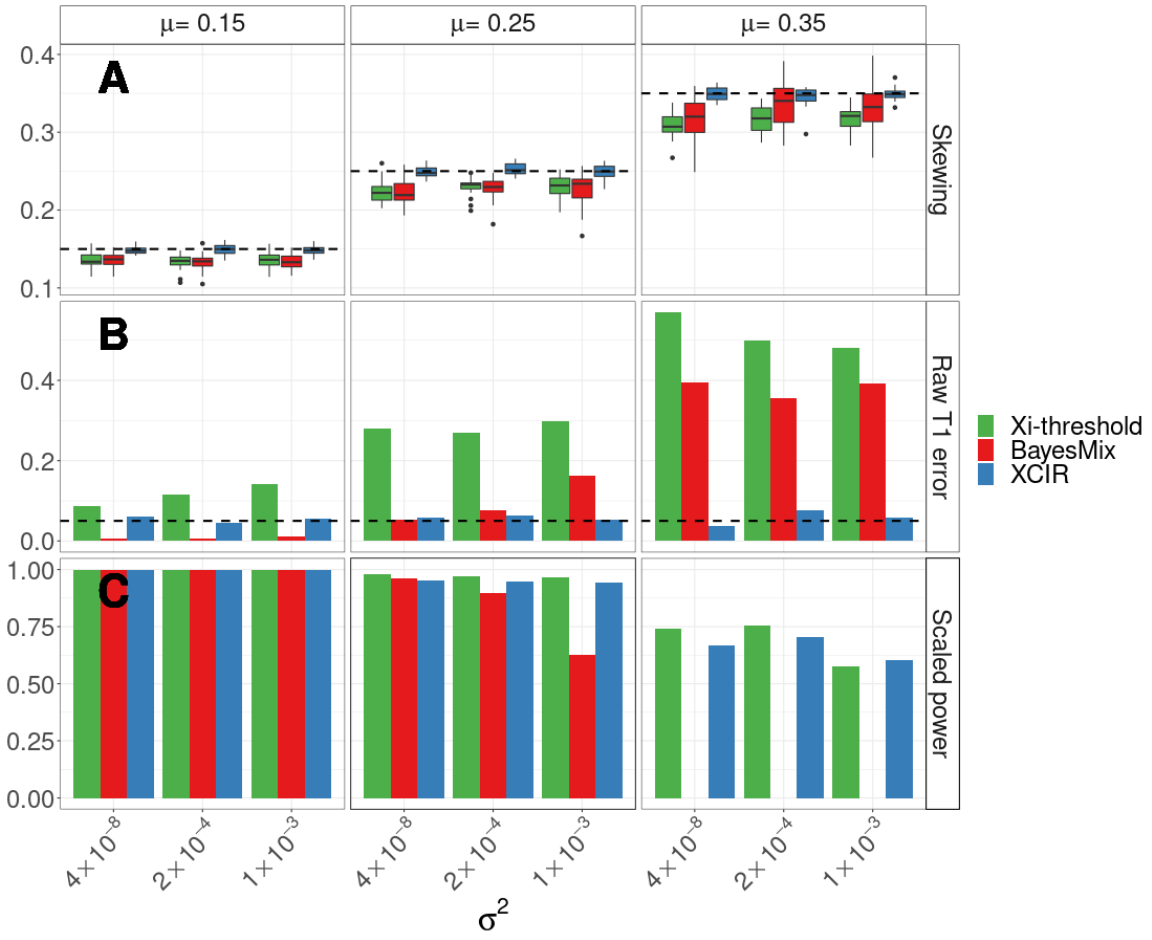

Figure A.1. Simulation results for samples with sequencing error models. The simulation scenario presented here is the same as 1.1, but the type I error and power are estimated using the 20 samples simulated with sequencing error, in order to separately evaluate the impact of sequencing errors on type I errors and power. **(A)** Distribution of the skewing estimates. **(B)** Raw type 1 error. **(C)** Power after rescaling thresholds until type 1 error of 0.05 is achieved in the training set. Scaled power of 0 for BayesMix indicates that a type 1 error of 0.05 or less in the training set can only be achieved using a very high PPE threshold, thus classifying every gene as silenced.

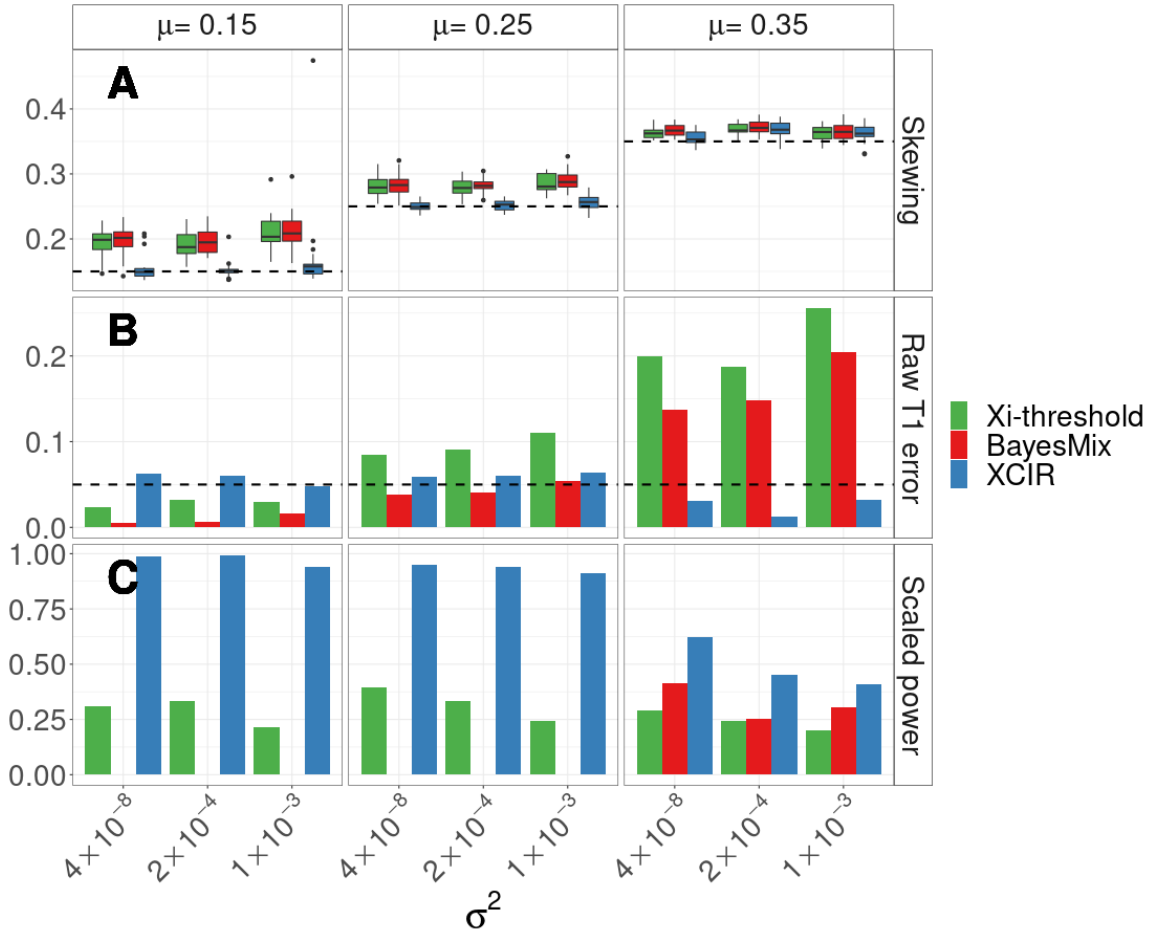

Figure A.2. Simulation results for samples with training error models. The simulation scenario presented here is the same as Fig 1, but the type I error and power were estimated using the 20 samples simulated with training error, where a portion of the genes in the training set escapes XCI. This figure separately evaluates the impact of training errors on type I error and power.This figure is identical to Fig 1. but limited to the 20 samples simulated with sequencing error. **(A)** Distribution of the skewing estimates. **(B)** Raw type 1 error. **(C)** Power after rescaling thresholds until type 1 error of 0.05 is achieved in the training set. Scaled power of 0 for BayesMix indicates that a type 1 error of 0.05 or less in the training set can only be achieved using a very high PPE, thus classifying every gene as silenced. When a portion of the genes in the training data escape inactivation, the XCIR estimates of the skewing are less biased than other approaches, leading to higher power in all scenarios.

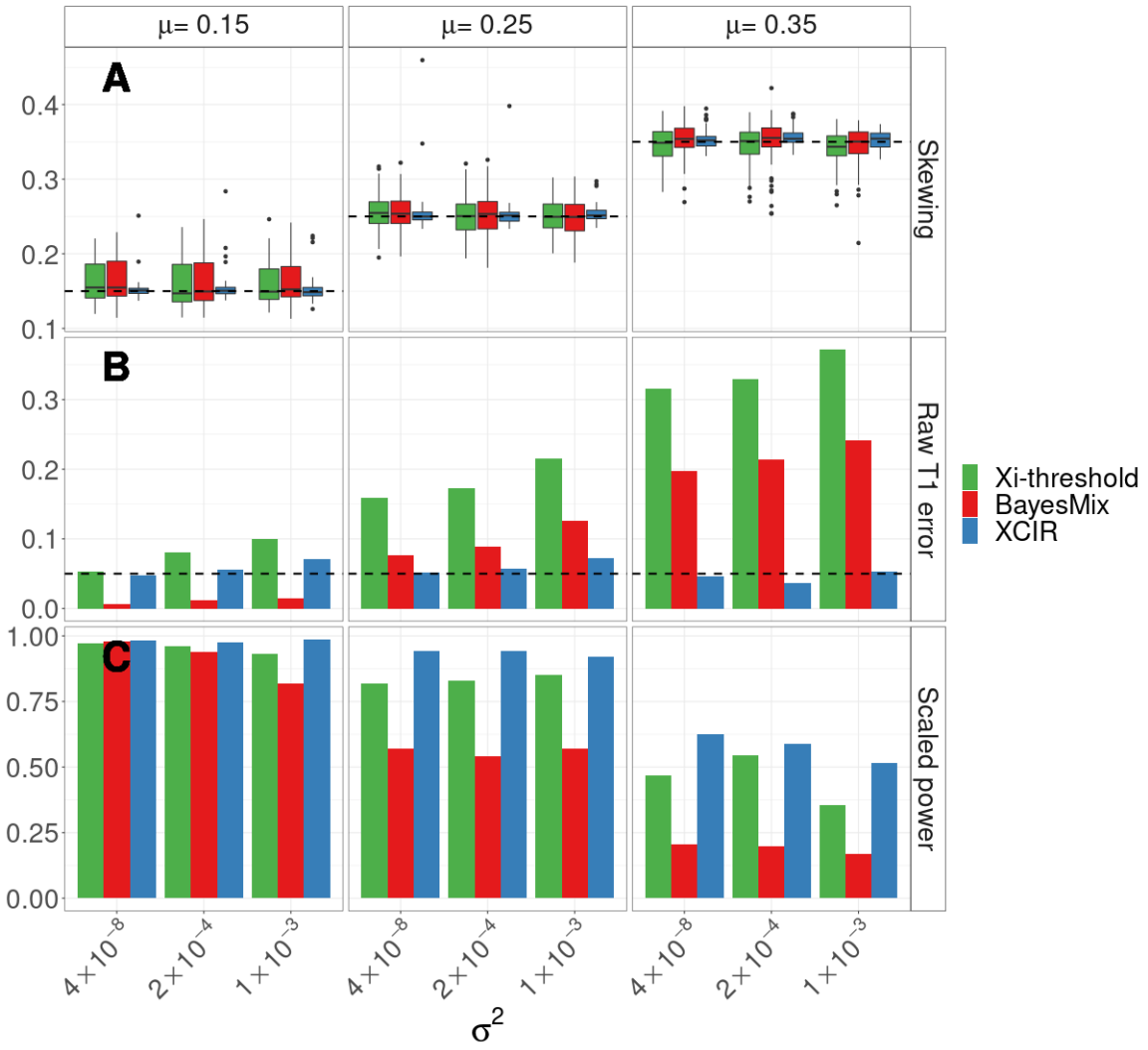

Figure A.3. Simulation results assuming a random sequencing error model. The simulation scenario is similar to Fig 1. but here we allow the sequencing error to vary between genes. For each gene, we simulate the sequencing error as:

$$\begin{aligned} \pi_{err} &\sim Unif(0, .1) \\ N_{err} &\sim Bin(N_g, \pi_{err}) \end{aligned} \tag{A.2}$$

Where N_{err} is the number of sequencing error generated for each gene and N_g is the total number of reads for the gene. **(A)** Distribution of the skewing estimates. **(B)** Raw type 1 error. **(C)** Power after rescaling thresholds until type 1 error of 0.05 is achieved in the training set.

Appendix B
XCI States Observed in Single-Cell Clone Lines

Table B.1: XCI States Observed in Single-Cell Clone Lines. The consensus is the gene's XCI state as determined in (Balaton 2015). Set indicates whether the gene was used in the set of known inactive genes to estimate skewing. Status in non-random samples is the true inactivation status observed in the two non-random cell lines (100:0 and 0:100). Genes where the observed true status were not consistent between the two cell lines were discarded.

gene	consensus	set	status	0:100	100:0	50:50	60:40	70:30	80:20	90:10
ABCD1	Mostly S	train	S	0:23	29:0	18:7	33:6	31:1	21:2	26:0
ACOT9	S	train	S	0:28	40:0	26:21	16:5	36:7	10:15	
AIFM1	S	train	S	272:0	0:393	120:220	83:251	33:295	44:305	17:355
AKAP17A	PAR	test	E	147:133	153:98	146:132	137:92	111:110	136:93	113:135
AMER1	S	test	unknown	22:0			12:15			
AMOT	S	train	unknown		53:0	42:2	36:3	39:0	46:8	51:0
APOO	S	train	S	63:0	0:61	24:59	24:41	8:68	19:55	4:74
ARHGAP4	Discordant	train	unknown	55:0		22:19	8:53	0:30	6:44	0:46
ARHGAP6	Mostly S	train	unknown	37:0		20:9	25:18	9:16	6:28	11:20
ARMCX3	S	train	S	0:52	88:1	60:32	51:11	43:7	83:5	66:0
ASMTL	PAR	test	E	119:110	119:155	171:202	152:173	185:113	145:103	150:147
ATP7A	Mostly S	test	S	0:85	61:0	33:53	45:27	48:12	57:20	55:7
ATRX	S	test	S	0:94	66:2	72:54	66:24	83:26	84:32	117:0
BCOR	Mostly S	train	unknown	3:53	0:36	0:58	0:70	0:52	0:51	0:36
BEX2	S	test	unknown			3:26	5:27	2:21	3:39	0:43
BEX4	S	test	S	98:0	0:100	55:78	37:74	34:61	8:96	22:77
BRCC3	Mostly S	train	S	0:38	50:0		24:16	48:18	45:1	36:1
BRWD3	S	train	S	0:54	0:51	0:75	0:63	0:59	3:62	0:64

BTK	S	train	S	0:152	510:0	314:61	367:32	474:40	512:29	498:18
C1GALT1C1	S	test	S	38:0	0:55	31:18	5:26	13:18	3:52	11:32
CA5BP1	E	test	E	50:16	11:31	9:35	16:46	9:25	14:31	3:37
CCDC22	S	train	S	85:0	0:80	35:40	25:55	14:23	9:69	1:62
CD99	PAR	test	E	313:304	314:150	353:260	354:201	455:222	418:254	351:246
CD99L2	S	train	S	0:105	79:0	71:51	59:37	75:15	79:6	112:7
CENPI	S	train	S	0:37	34:0	29:30	27:19	25:13	43:3	33:4
CHIC1	S	train	S	0:29	35:0	37:18	43:8	46:7	39:4	48:2
CHM	Discordant	train	S	0:37	20:1	22:5	27:12	28:13	73:0	22:0
CLIC2	Mostly S	train	unknown		0:37		5:23	0:24	0:30	0:43
CXorf21	VE	test	E	49:225	83:8	84:61	65:71	83:55	46:50	52:12
CXorf23	S	train	unknown				17:11			
CXorf40B	S	test	unknown			21:7	12:19	6:18	5:29	1:35
CXorf56	Mostly S	train	unknown		45:1	36:3	39:2	42:4	20:12	24:4
CYBB	S	train	S	154:0	0:287	194:26	51:211	37:191	20:200	5:283
DDX26B	S	train	S	42:0	0:43	23:22	22:48	17:36	8:47	9:28
DDX3X	E	test	E	98:48	16:126	76:85	51:92	41:86	29:110	30:99
DHRSX	PAR	test	E	31:70	36:62	49:47	59:74	36:60	22:72	23:89
DKC1	S	test	S	224:0	1:329	96:194	66:219	37:167	40:273	18:276
DNASE1L1	S	train	unknown				13:25			
DOCK11	Mostly S	train	S	69:0	69:1	71:38	22:63	23:79	55:10	7:123
DYNLT3	S	train	unknown	0:57	52:4	35:24	19:20	31:18	40:15	56:1
EBP	Mostly S	test	S	175:0	0:209	68:127	69:205	86:208	13:223	10:224
EIF2S3	E	test	E	794:196	190:796	403:594	365:611	364:624	293:705	249:739
ELF4	S	train	S	0:73	165:0	108:63	131:51	130:28	157:18	117:24
FAAH2	S	train	unknown					5:18	1:26	
FAM199X	S	test	unknown		0:22		7:20	2:29	0:46	2:34
FAM3A	S	test	S	90:0	0:84	37:70	26:103	9:72	15:71	2:87
FAM50A	S	train	unknown	0:150	143:9	93:35	127:17	66:14	130:16	144:5
FAM58A	S	train	S	0:72	73:0	77:47	46:28	83:12	71:12	92:11
FTSJ1	S	train	unknown				10:11			
FTX	S	test	unknown	3:23	42:0		15:6	14:18		25:0
G6PD	S	train	S	289:1	0:274	144:133	90:168	53:180	42:189	21:222
GPR174	VE	test	unknown	116:0		49:7	42:13	33:30	24:15	
GTPBP6	PAR	test	E	77:73	86:50	87:94	84:49	62:75	73:90	67:61
HCFC1	Mostly VE	test	unknown	420:26	0:329	147:285	142:353	87:324	77:328	15:285
HDHD1	E	test	S	0:182	0:130	0:186	0:166	3:203	1:182	1:170
HMGB3	S	test	S	113:0	0:91	72:108	44:56	34:60	12:81	9:102
IDS	S	train	S	355:0	428:0	294:118	395:90	338:74	504:49	513:19
IL1RAPL1	Discordant	test	unknown						19:3	
IL3RA	PAR	test	E	57:71	57:90	67:79	75:74	93:83	82:88	78:80
ITM2A	Discordant	train	S	164:0	87:0	79:87	66:61	46:81	24:117	103:13
JPX	E	test	E	37:13	65:8	67:26	49:22	37:30	35:24	71:11
KDM5C	Mostly E	test	E	202:241	279:144	267:220	288:244	256:193	371:183	324:147

KIAA2022	S	train	unknown		21:2		20:1			24:0
KLHL15	S	train	S	46:0	0:52	13:26	8:37	19:33	7:38	5:43
LAMP2	Mostly S	train	S	238:0	2:203	137:126	110:190	69:168	40:212	21:241
MAGEH1	S	test	S	82:0	0:77	33:33	16:50	16:64	16:64	3:93
MAP7D2	Discordant	test	unknown	12:18						
MBTPS2	S	train	S	32:0	0:22	24:27	18:32	7:17	6:26	4:42
MED12	S	train	S	0:121	82:0	61:42	92:37	86:7	89:14	97:15
MED14	VE	test	S	36:2	0:40	20:26	13:32	4:34	6:28	4:55
MORF4L2	Mostly S	train	S	209:0	0:256	106:169	82:196	47:181	45:201	25:248
MPP1	S	train	S	0:84	79:0	51:51	51:44	46:28	55:24	59:10
MSL3	Discordant	test	unknown		12:18					
MST4	Mostly S	train	unknown	155:0	13:159	52:100	64:147	29:145	31:179	23:127
MTCP1	S	test	unknown		59:0	23:9	25:4	20:2	22:2	36:13
MTMR1	S	train	S	1:106	118:1	99:51	102:47	94:15	115:19	112:12
NONO	Mostly S	test	S	0:95	54:0	48:34	71:27	88:20	37:13	73:11
NSDHL	S	train	S	0:91	0:126	4:115	0:100	0:84	0:123	0:140
NXT2	Mostly S	train	S	0:138	120:0	50:64	65:42	66:19	63:8	96:8
OGT	Mostly S	train	unknown						31:5	24:0
P2RY8	PAR	test	E	241:252	168:129	240:219	272:217	191:216	183:172	175:170
PDK3	S	train	S	0:64	64:0	47:37	49:33	51:13	51:17	83:1
PGK1	S	train	S	0:56	92:0	51:43	63:27	89:15	95:7	87:7
PIR	Discordant	train	S	56:0	0:61	21:44	25:58	10:44	4:51	5:47
PJA1	S	train	S	37:0	0:21	30:17	18:14	0:37	5:19	0:28
PLCXD1	PAR	test	E	111:89	52:67	107:96	77:109	69:73	69:92	54:82
PLXNA3	S	train	S	118:0	0:124	48:72	40:80	13:77	23:101	9:105
POF1B	VE	test	S	0:50	0:65	0:52	2:57	0:48	0:69	3:78
PPP2R3B	PAR	test	E	37:46	15:24	32:29	32:41	31:51	27:27	24:29
PRICKLE3	S	train	S	44:0	0:64	22:35	21:38	17:35	16:53	9:59
PRKX	E	test	E	122:123	72:60	138:115	106:130	93:157	79:136	86:75
PRPS1	S	test	S	185:0	0:181	69:90	49:77	44:111	26:148	16:131
PRPS2	S	train	S	0:183	191:0	119:75	137:80	170:39	172:20	169:9
RAB39B	S	train	S	25:0	0:24	13:20	6:16		3:42	0:34
RBM10	S	train	unknown				10:16			
RBM3	S	test	unknown	4:737	2:804	5:814	10:794	2:827	19:787	2:833
RBMX	S	test	unknown	569:46	1:494	316:279	195:371	193:355	104:421	46:464
RENBP	Discordant	test	unknown			19:6			15:6	22:0
RPS4X	Mostly E	test	E	641:354	423:572	483:510	458:538	473:526	427:571	456:536
RRAGB	Mostly S	train	S	64:0	0:56	31:33	8:53	24:45	20:45	8:50
SAT1	S	test	S	283:0	0:197	111:91	83:173	62:155	40:151	19:171
SEPT6	Mostly S	test	unknown	114:0	15:81	56:39	54:49	56:47	36:32	30:45
SLC25A43	Mostly S	train	S	0:58	58:1	50:20	43:24	72:2	52:3	69:7
SLC25A6	PAR	test	E	10:19	15:11	15:20		11:20	16:9	
SLC38A5	S	train	S	0:153	200:0	125:52	102:36	98:28	137:20	131:6
SLC9A7	S	train	S	88:0	0:138	58:72	31:110	11:70	18:120	7:93

SNX12	S	train	S	0:82	86:0	58:42	67:16	82:8	88:19	101:6
SPIN2B	Mostly S	test	unknown		0:22				5:18	
SPIN3	Mostly S	train	S	49:0	0:25	7:38	16:22	9:45	1:54	4:48
SUV39H1	S	test	S	75:0	0:55	21:34	23:37	20:40	14:29	2:62
TAB3	S	test	S	0:92	0:78	0:125	0:116	0:63	3:109	0:88
TAF1	S	train	S	0:94	59:0	55:23	53:21	68:16	66:11	58:6
TBC1D25	S	train	unknown	56:0	3:70	29:62	24:63	26:39	14:60	2:91
TCEAL4	S	train	S	0:259	317:0	205:104	237:136	265:68	289:44	290:23
TCEANC	Mostly E	test	unknown		7:16					
TMEM164	S	train	S	0:24	52:0	48:20	64:8	30:6	41:2	54:7
TMEM187	Mostly VE	train	S	1:46	40:0	38:22	24:9	45:9	22:9	22:0
TMSB15B	Mostly S	test	unknown	33:0		23:22	18:8		20:9	32:1
TMSB4X	No call	test	unknown	46:2	14:35	23:21	19:28	12:34	15:37	11:34
TRMT2B	S	train	S	27:0	0:28	17:15			4:24	
TRO	Mostly S	train	unknown	0:24				30:5	37:0	
TSPYL2	S	train	S	0:24	31:0	40:26	22:8	43:9	48:1	28:4
TXLNG	E	test	E	14:75	115:44	64:27	54:28	72:28	53:28	78:13
UBL4A	S	train	S	131:0	0:172	70:55	62:80	16:94	20:81	9:131
UBQLN2	S	test	unknown	196:0	4:121	80:120	35:101	45:78	25:127	6:122
UPF3B	S	train	S	78:0	0:64	16:37	20:51	12:39	25:67	8:77
USP9X	Mostly E	test	E	145:8	25:171	88:87	70:141	55:112	62:145	42:117
VAMP7	S	train	unknown	0:150	249:5	145:79	129:48	160:54	163:23	165:6
VBP1	S	train	S	199:0	1:212	81:111	55:131	41:109	36:172	9:177
VMA21	S	train	S	145:0	0:130	58:78	55:64	53:81	27:78	3:119
WDR13	S	train	unknown			12:14			12:17	
WDR44	S	train	S	68:0	0:65	9:41	13:57	6:36	7:53	4:56
WDR45	S	test	unknown			9:12			22:0	18:3
WWC3	Mostly S	train	unknown	7:140	48:0	51:59	44:42	54:39	41:49	56:17
XIAP	S	train	S	0:113	115:0	69:28	76:21	63:19	73:9	108:3
XIST	Mostly S	test	S	954:0	1:865	420:517	312:621	199:686	125:797	62:833
ZBED1	PAR	test	E	145:151	123:181	166:164	181:194	168:167	150:183	111:256
ZMAT1	Mostly S	test	S	0:21	33:0					
ZNF275	S	train	S	0:99	73:0	36:25	41:17	64:13	78:28	62:5
ZNF280C	Mostly S	test	unknown	39:0		11:14	15:13	1:26	6:24	
ZNF41	S	train	unknown	0:33			20:9			
ZNF75D	S	train	S	0:49	25:0	11:18	23:7	8:29	20:11	1:27
ZRSR2	E	test	E	7:22	17:22	11:11	26:11		6:26	5:23
ZXDB	S	test	S	24:0	0:22					0:28

Bibliography

[1] Ross, M. T., D. V. Grafham, A. J. Coffey, S. Scherer, K. McLay, D. Muzny, M. Platzer, G. R. Howell, C. Burrows, C. P. Bird, A. Frankish, F. L. Lovell, K. L. Howe, J. L. Ashurst, R. S. Fulton, R. Sudbrak, G. Wen, M. C. Jones, M. E. Hurles, T. D. Andrews, C. E. Scott, S. Searle, J. Ramser, A. Whittaker, R. Deadman, N. P. Carter, S. E. Hunt, R. Chen, A. Cree, P. Gunaratne, P. Havlak, A. Hodgson, M. L. Metzker, S. Richards, G. Scott, D. Steffen, E. Sodergren, D. A. Wheeler, K. C. Worley, R. Ainscough, K. D. Ambrose, M. A. Ansari-Lari, S. Aradhya, R. I. Ashwell, A. K. Babbage, C. L. Bagguley, A. Ballabio, R. Banerjee, G. E. Barker, K. F. Barlow, I. P. Barrett, K. N. Bates, D. M. Beare, H. Beasley, O. Beasley, A. Beck, G. Bethel, K. Blechschmidt, N. Brady, S. Bray-Allen, A. M. Bridgeman, A. J. Brown, M. J. Brown, D. Bonnin, E. A. Bruford, C. Buhay, P. Burch, D. Burford, J. Burgess, W. Burrill, J. Burton, J. M. Bye, C. Carder, L. Carrel, J. Chako, J. C. Chapman, D. Chavez, E. Chen, G. Chen, Y. Chen, Z. Chen, C. Chinault, A. Ciccodicola, S. Y. Clark, G. Clarke, C. M. Clee, S. Clegg, K. Clerc-Blankenburg, K. Clifford, V. Cobley, C. G. Cole, J. S. Conquer, N. Corby, R. E. Connor, R. David, J. Davies, C. Davis, J. Davis, O. Delgado, D. Deshazo, et al. (2005) "The DNA sequence of the human X chromosome," *Nature*, **434**(7031), pp. 325–37.

[2] Zhang, Y., K. Klein, A. Sugathan, N. Nassery, A. Dombkowski, U. M. Zanger, and D. J. Waxman (2011) "Transcriptional profiling of human liver identifies sex-biased genes associated with polygenic dyslipidemia and coronary artery disease," *PLoS One*, **6**(8), p. e23506.
URL `https://www.ncbi.nlm.nih.gov/pubmed/21858147`

[3] Khramtsova, E. A., L. K. Davis, and B. E. Stranger (2019) "The role of sex in the genomics of human complex traits," *Nat Rev Genet*, **20**(3), pp. 173–190.
URL `https://www.ncbi.nlm.nih.gov/pubmed/30581192`

[4] Sidorenko, J., I. Kassam, K. E. Kemper, J. Zeng, L. R. Lloyd-Jones, G. W. Montgomery, G. Gibson, A. Metspalu, T. Esko, J. Yang, A. F. McRae, and P. M. Visscher (2019) "The effect of X-linked dosage compensation on complex trait variation," *Nat Commun*, **10**(1), p. 3009.

[5] Medicine, N. (2017) "Accounting for sex in the genome," *Nat Med*, **23**(11), p. 1243.
URL https://www.ncbi.nlm.nih.gov/pubmed/29117171

[6] Carrel, L. and H. F. Willard (2005) "X-inactivation profile reveals extensive variability in X-linked gene expression in females," *Nature*, **434**(7031), pp. 400–4.
URL http://www.ncbi.nlm.nih.gov/pubmed/15772666

[7] Carrel, L. and C. J. Brown (2017) "When the Lyon(ized chromosome) roars: ongoing expression from an inactive X chromosome," *Philos Trans R Soc Lond B Biol Sci*, **372**(1733).
URL https://www.ncbi.nlm.nih.gov/pubmed/28947654

[8] Tukiainen, T., A. C. Villani, A. Yen, M. A. Rivas, J. L. Marshall, R. Satija, M. Aguirre, L. Gauthier, M. Fleharty, A. Kirby, B. B. Cummings, S. E. Castel, K. J. Karczewski, F. Aguet, A. Byrnes, G. T. Consortium, D. A. Laboratory, G. Coordinating Center Analysis Working, G. Statistical Methods groups Analysis Working, G. g. Enhancing, N. I. H. C. Fund, Nih/Nci, Nih/Nhgri, Nih/Nimh, Nih/Nida, N. Biospecimen Collection Source Site, R. Biospecimen Collection Source Site, V. Biospecimen Core Resource, B. Brain Bank Repository-University of Miami Brain Endowment, M. Leidos Biomedical-Project, E. Study, I. Genome Browser Data, E. B. I. Visualization, I. Genome Browser Data, U. o. C. S. C. Visualization-Ucsc Genomics Institute, T. Lappalainen, A. Regev, K. G. Ardlie, N. Hacohen, and D. G. MacArthur (2017) "Landscape of X chromosome inactivation across human tissues," *Nature*, **550**(7675), pp. 244–248.
URL https://www.ncbi.nlm.nih.gov/pubmed/29022598

[9] Garieri, M., G. Stamoulis, X. Blanc, E. Falconnet, P. Ribaux, C. Borel, F. Santoni, and S. E. Antonarakis (2018) "Extensive cellular heterogeneity of X inactivation revealed by single-cell allele-specific expression in human fibroblasts," *Proc Natl Acad Sci U S A*, **115**(51), pp. 13015–13020.
URL https://www.ncbi.nlm.nih.gov/pubmed/30510006

[10] Tukiainen, T., M. Pirinen, A. P. Sarin, C. Ladenvall, J. Kettunen, T. Lehtimaki, M. L. Lokki, M. Perola, J. Sinisalo, E. Vlachopoulou, J. G. Eriksson, L. Groop, A. Jula, M. R. Jarvelin, O. T. Raitakari, V. Salomaa, and S. Ripatti (2014) "Chromosome X-wide association study identifies Loci for fasting insulin and height and evidence for incomplete dosage compensation," *PLoS Genet*, **10**(2), p. e1004127.
URL https://www.ncbi.nlm.nih.gov/pubmed/24516404

[11] Wang, J., C. M. Syrett, M. C. Kramer, A. Basu, M. L. Atchison, and M. C. Anguera (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**(14), pp. E2029–38.
URL http://www.ncbi.nlm.nih.gov/pubmed/27001848

[12] Dunford, A., D. M. Weinstock, V. Savova, S. E. Schumacher, J. P. Cleary, A. Yoda, T. J. Sullivan, J. M. Hess, A. A. Gimelbrant, R. Beroukhim, M. S. Lawrence, G. Getz, and A. A. Lane (2017) "Tumor-suppressor genes that escape from X-inactivation contribute to cancer sex bias," *Nat Genet*, **49**(1), pp. 10–16.
URL https://www.ncbi.nlm.nih.gov/pubmed/27869828

[13] Souyris, M., C. Cenac, P. Azar, D. Daviaud, A. Canivet, S. Grunenwald, C. Pienkowski, J. Chaumeil, J. E. Mejia, and J. C. Guery (2018) "TLR7 escapes X chromosome inactivation in immune cells," *Sci Immunol*, **3**(19).
URL https://www.ncbi.nlm.nih.gov/pubmed/29374079

[14] Harris, V. M., I. T. W. Harley, B. T. Kurien, K. A. Koelsch, and R. H. Scofield (2019) "Lysosomal pH Is Regulated in a Sex Dependent Manner in Immune Cells Expressing CXorf21," *Front Immunol*, **10**, p. 578.
URL https://www.ncbi.nlm.nih.gov/pubmed/31001245

[15] Natri, H., A. R. Garcia, K. H. Buetow, B. C. Trumble, and M. A. Wilson (2019) "The Pregnancy Pickle: Evolved Immune Compensation Due to Pregnancy Underlies Sex Differences in Human Diseases," *Trends Genet*, **35**(7), pp. 478–488.
URL https://www.ncbi.nlm.nih.gov/pubmed/31200807

[16] Syrett, C. M., B. Paneru, D. Sandoval-Heglund, J. Wang, S. Banerjee, V. Sindhava, E. M. Behrens, M. Atchison, and M. C. Anguera (2019) "Altered X-chromosome inactivation in T cells may promote sex-biased autoimmune diseases," *JCI Insight*, **4**(7).
URL https://www.ncbi.nlm.nih.gov/pubmed/30944248

[17] Foresta, C., M. S. Rocca, and A. Di Nisio (2020) "Gender susceptibility to COVID-19: a review of the putative role of sex hormones and X chromosome," *J Endocrinol Invest*, pp. 1–6.

[18] Hagen, S. H., F. Henseling, J. Hennesen, H. Savel, S. Delahaye, L. Richert, S. M. Ziegler, and M. Altfeld (2020) "Heterogeneous Escape from X Chromosome Inactivation Results in Sex Differences in Type I IFN Responses at the Single Human pDC Level," *Cell Rep*, **33**(10), p. 108485.

[19] Mousavi, M. J., M. Mahmoudi, and S. Ghotloo (2020) "Escape from X chromosome inactivation and female bias of autoimmune diseases," *Mol Med*, **26**(1), p. 127.

[20] Yu, B., Y. Qi, R. Li, Q. Shi, A. T. Satpathy, and H. Y. Chang (2021) "B cell-specific XIST complex enforces X-inactivation and restrains atypical B cells," *Cell*, **184**(7), pp. 1790–1803 e17.
URL https://www.ncbi.nlm.nih.gov/pubmed/33735607

[21] Wainer Katsir, K. and M. Linial (2019) "Human genes escaping X-inactivation revealed by single cell expression data," *BMC Genomics*, **20**(1), p. 201.

[22] Cotton, A. M., E. M. Price, M. J. Jones, B. P. Balaton, M. S. Kobor, and C. J. Brown (2014) "Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation," *Hum Mol Genet*.
URL `http://www.ncbi.nlm.nih.gov/pubmed/25381334`

[23] Schultz, M. D., Y. He, J. W. Whitaker, M. Hariharan, E. A. Mukamel, D. Leung, N. Rajagopal, J. R. Nery, M. A. Urich, H. Chen, S. Lin, Y. Lin, I. Jung, A. D. Schmitt, S. Selvaraj, B. Ren, T. J. Sejnowski, W. Wang, and J. R. Ecker (2015) "Human body epigenome maps reveal noncanonical DNA methylation variation," *Nature*, **523**(7559), pp. 212–216.

[24] Cotton, A. M., B. Ge, N. Light, V. Adoue, T. Pastinen, and C. J. Brown (2013) "Analysis of expressed SNPs identifies variable extents of expression from the human inactive X chromosome," *Genome Biol*, **14**(11), p. R122.
URL `http://www.ncbi.nlm.nih.gov/pubmed/24176135`

[25] Larson, N. B., Z. C. Fogarty, M. C. Larson, K. R. Kalli, K. Lawrenson, S. Gayther, B. L. Fridley, E. L. Goode, and S. J. Winham (2017) "An integrative approach to assess X-chromosome inactivation using allele-specific expression with applications to epithelial ovarian cancer," *Genet Epidemiol*, **41**(8), pp. 898–914.
URL `https://www.ncbi.nlm.nih.gov/pubmed/29119601`

[26] Pickrell, J. K., J. C. Marioni, A. A. Pai, J. F. Degner, B. E. Engelhardt, E. Nkadori, J. B. Veyrieras, M. Stephens, Y. Gilad, and J. K. Pritchard (2010) "Understanding mechanisms underlying human gene expression variation with RNA sequencing," *Nature*, **464**(7289), pp. 768–72.
URL `http://www.ncbi.nlm.nih.gov/pubmed/20220758`

[27] Akaike, H., B. N. Petrov, and F. Csaki (1973), "Second international symposium on information theory," .

[28] Lappalainen, T., M. Sammeth, M. R. Friedlander, P. A. t Hoen, J. Monlong, M. A. Rivas, M. Gonzalez-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, M. Barann, T. Wieland, L. Greger, M. van Iterson, J. Almlof, P. Ribeca, I. Pulyakhina, D. Esser, T. Giger, A. Tikhonov, M. Sultan, G. Bertier, D. G. MacArthur, M. Lek, E. Lizano, H. P. Buermans, I. Padioleau, T. Schwarzmayr, O. Karlberg, H. Ongen, H. Kilpinen, S. Beltran, M. Gut, K. Kahlem, V. Amstislavskiy, O. Stegle, M. Pirinen, S. B. Montgomery, P. Donnelly, M. I. McCarthy, P. Flicek, T. M. Strom, C. Geuvadis, H. Lehrach, S. Schreiber, R. Sudbrak, A. Carracedo, S. E. Antonarakis, R. Hasler, A. C. Syvanen, G. J. van Ommen, A. Brazma, T. Meitinger, P. Rosenstiel, R. Guigo, I. G. Gut, X. Estivill, and E. T.

Dermitzakis (2013) "Transcriptome and genome sequencing uncovers functional variation in humans," *Nature*, **501**(7468), pp. 506–11.
URL http://www.ncbi.nlm.nih.gov/pubmed/24037378

[29] Kriegler, M. (1990) "Gene transfer and expression: A Laboratory Manual," .

[30] Freshney, R. I. (2000) *Culture of Animal Cells: A Manual of Basic Technique*, (4th edition).

[31] Allen, R. C., H. Y. Zoghbi, A. B. Moseley, H. M. Rosenblatt, and J. W. Belmont (1992) "Methylation of HpaII and HhaI sites near the polymorphic CAG repeat in the human androgen-receptor gene correlates with X chromosome inactivation," *Am J Hum Genet*, **51**(6), pp. 1229–39.

[32] Kim, D., B. Langmead, and S. L. Salzberg (2015) "HISAT: a fast spliced aligner with low memory requirements," *Nat Methods*, **12**(4), pp. 357–60.
URL https://www.ncbi.nlm.nih.gov/pubmed/25751142

[33] van de Geijn, B., G. McVicker, Y. Gilad, and J. K. Pritchard (2015) "WASP: allele-specific software for robust molecular quantitative trait locus discovery," *Nat Methods*, **12**(11), pp. 1061–3.
URL http://www.ncbi.nlm.nih.gov/pubmed/26366987

[34] Balaton, B. P., A. M. Cotton, and C. J. Brown (2015) "Derivation of consensus inactivation status for X-linked genes from genome-wide studies," *Biol Sex Differ*, **6**, p. 35.
URL http://www.ncbi.nlm.nih.gov/pubmed/26719789

[35] Bray, N. L., H. Pimentel, P. Melsted, and L. Pachter (2016) "Near-optimal probabilistic RNA-seq quantification," *Nat Biotechnol*, **34**(5), pp. 525–7.
URL https://www.ncbi.nlm.nih.gov/pubmed/27043002

[36] .

[37] Gershoni, M. and S. Pietrokovski (2017) "The landscape of sex-differential transcriptome and its consequent selection in human adults," *BMC Biol*, **15**(1), p. 7.

[38] R Core Team (2019) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL https://www.R-project.org/

[39] Bycroft, C., C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, A. Cortes, S. Welsh, A. Young, M. Effingham, G. McVean, S. Leslie, N. Allen, P. Donnelly, and J. Marchini (2018) "The UK Biobank resource with deep phenotyping and genomic data," *Nature*, **562**(7726), pp. 203–209.
URL https://www.ncbi.nlm.nih.gov/pubmed/30305743

[40] Zhou, W., J. B. Nielsen, L. G. Fritsche, R. Dey, M. E. Gabrielsen, B. N. Wolford, J. LeFaive, P. VandeHaar, S. A. Gagliano, A. Gifford, L. A. Bastarache, W. Q. Wei, J. C. Denny, M. Lin, K. Hveem, H. M. Kang, G. R. Abecasis, C. J. Willer, and S. Lee (2018) "Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies," *Nat Genet*, **50**(9), pp. 1335–1341.
URL https://www.ncbi.nlm.nih.gov/pubmed/30104761

[41] Slavney, A., L. Arbiza, A. G. Clark, and A. Keinan (2016) "Strong Constraint on Human Genes Escaping X-Inactivation Is Modulated by their Expression Level and Breadth in Both Sexes," *Mol Biol Evol*, **33**(2), pp. 384–93.
URL https://www.ncbi.nlm.nih.gov/pubmed/26494842

[42] Bellott, D. W., J. F. Hughes, H. Skaletsky, L. G. Brown, T. Pyntikova, T. J. Cho, N. Koutseva, S. Zaghlul, T. Graves, S. Rock, C. Kremitzki, R. S. Fulton, S. Dugan, Y. Ding, D. Morton, Z. Khan, L. Lewis, C. Buhay, Q. Wang, J. Watt, M. Holder, S. Lee, L. Nazareth, J. Alfoldi, S. Rozen, D. M. Muzny, W. C. Warren, R. A. Gibbs, R. K. Wilson, and D. C. Page (2014) "Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators," *Nature*, **508**(7497), pp. 494–9.
URL https://www.ncbi.nlm.nih.gov/pubmed/24759411

Vita

Renan Sauteraud

Education

August, 2021	Ph.D. in Biostatistics, **Penn State College of Medicine** **Pennsylvania State University**
August, 2011	Master of Bioinformatics and Biostatistics **Universite Paris-Sud XI**
August, 2009	Bachelor of Science (B.Sc.) in Genetics **Universite Rennes I, Rennes, France**
August, 2008	Technical diploma in Biology **IUT Saint-Brieuc (Universite Rennes I), Saint-Brieuc, France**

Publications

1. **R. Sauteraud**, Jill M. Stahl, Jesica James, Marisa Englebright, Fang Chen, Xiaowei Zhan, Laura Carrel and Dajiang J. Liu. Inferring Genes that Escape X chromosome Inactivation with XCIR Reveals Important Contribution of Variable Escape Genes to Sex-biased Diseases. *Genome Research*. Under review

2. D. McGuire, **R. Sauteraud** and V. Midya. Window-Based Feature Extraction Method Using XGBoost for Time Series Classification of Solar Flares. *2019 IEEE International Conference on Big Data*. 10.1109/BigData47090.2019.9006212. Dec 2019

3. Imholte, G., **Sauteraud, R.** and Gottardo, R. Analyzing Peptide Microarray Data with the R pepStat Package. *Methods Mol. Biol.* 2016

4. L. Vojtech, S. Woo, S. Hughes, C. Levy, L. Ballweber, **Sauteraud, R.**, J. Strobl, K. Wester-berg, R. Gottardo, M. Tewari, and F. Hladik. Exosomes in human semen carry a distinctive repertoire of small non-coding RNAs with potential regulatory functions. *Nucleic Acids Res.*, 42(11):7290–7304, Jun 2014

5. G. C. Imholte, **Sauteraud, R.**, B. Korber, R. T. Bailer, E. T. Turk, X. Shen, G. D.Tomaras, J. R. Mascola, R. A. Koup, D. C. Montefiori, and R. Gottardo. A computational framework for the analysis of peptide microarray antibody binding data with application to HIV vaccine profiling. *J. Immunol. Methods*, 395(1-2):1–13, Sep 2013.

6. S. Woo, X. Zhang, **Sauteraud, R.**, F. Robert, and R. Gottardo. PING 2.0: an R/Bioconductor package for nucleosome positioning using next-generation sequencing data. *Bioinformatics*, 29(16):2049–2050, Aug 2013.