Supplemental material

Renan Sauteraud, Jill M. Stahl, Jesica James, Marisa Englebright,
Fang Chen, Xiaowei Zhan, Laura Carrel, Dajiang J. Liu

April 2021

Contents

Figure S1 - Simulation results for sequencing error models alone

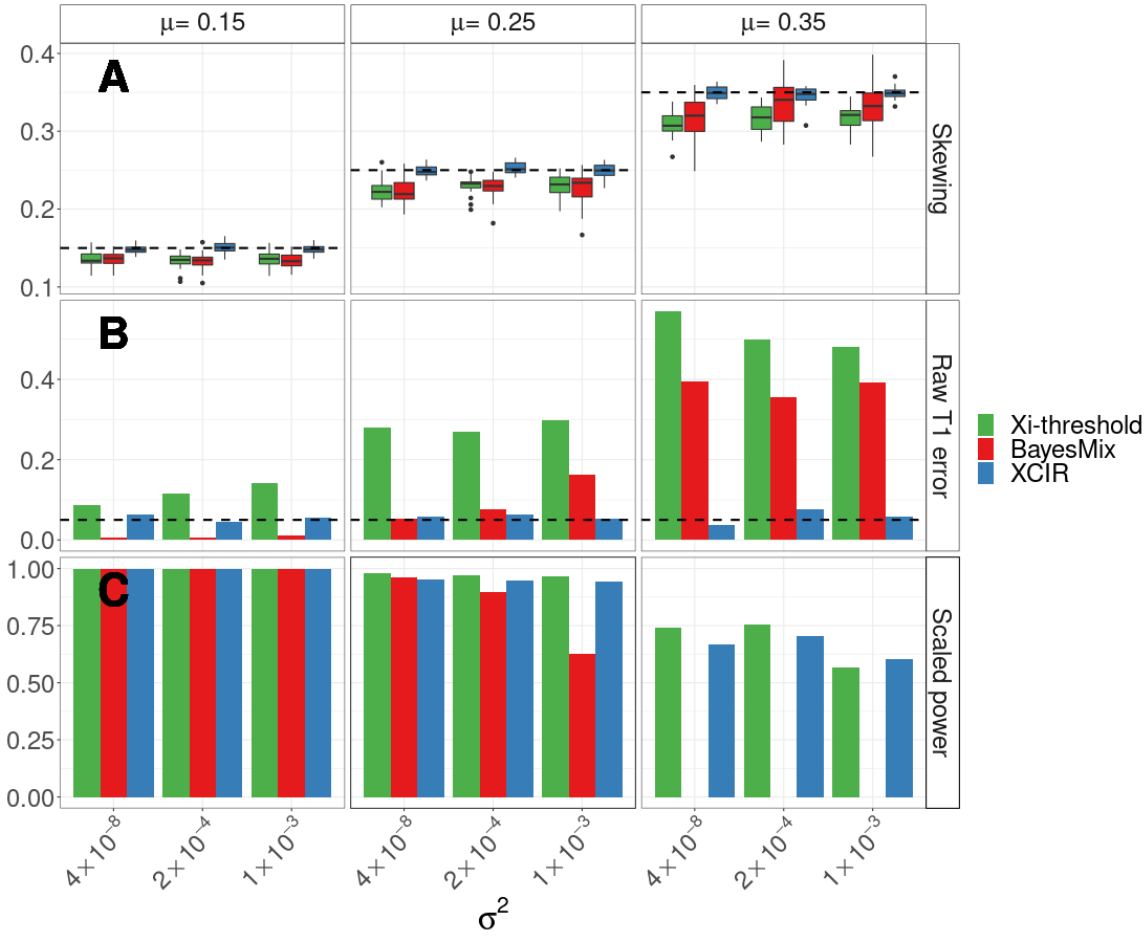

Fig S1. Simulation results for samples with sequencing error models. The simulation scenario presented here is the same as Fig 1, but the type I error and power are estimated using the 20 samples simulated with sequencing error, in order to separately evaluate the impact of sequencing errors on type 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 S2 - Simulation results for training error models alone

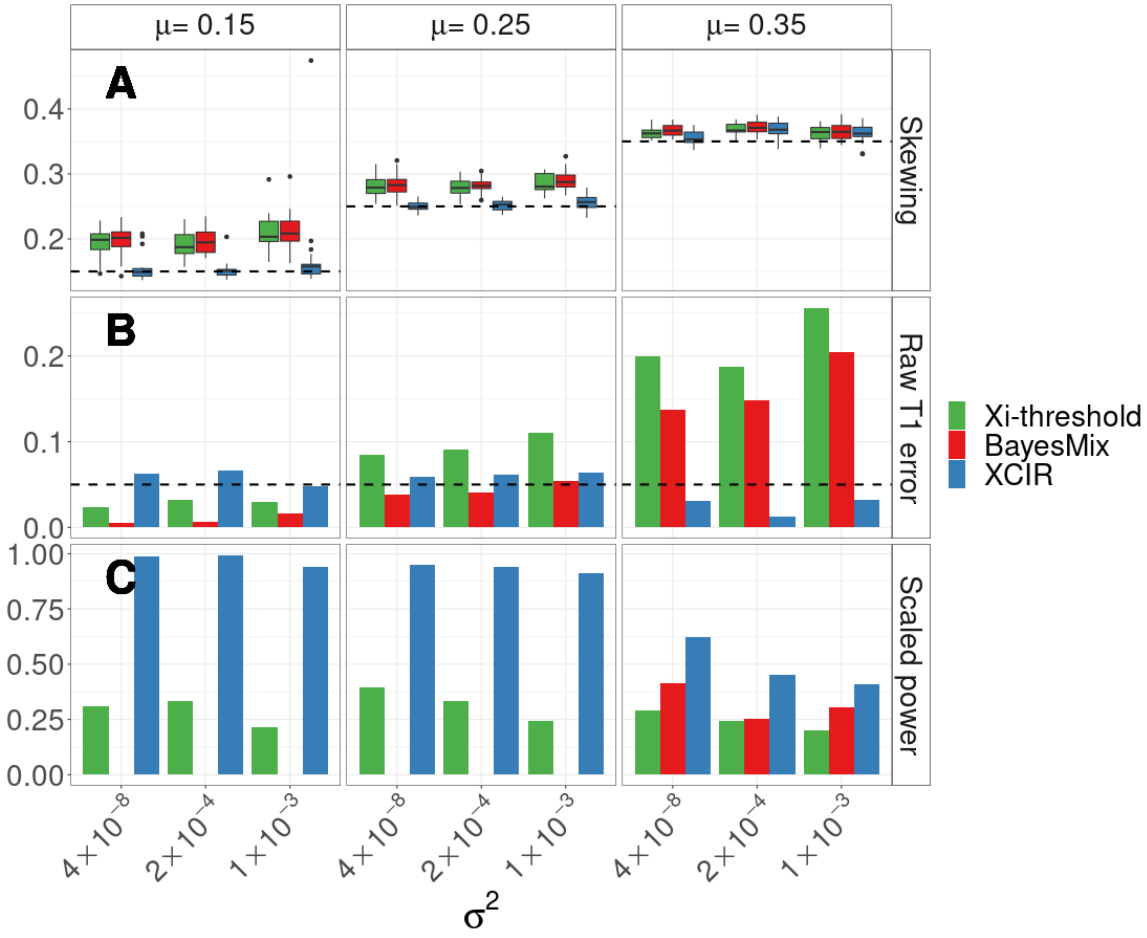

Fig S2. Simulation results for samples with training error models. The simulation scenario presented here is the same as Fig 1, but the type I error and power were estimated using the 20 samples simulated with training error, where a portion of the genes in the training set escapes XCI. This figure separately 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 S3 - Simulation results under a variable sequencing error model

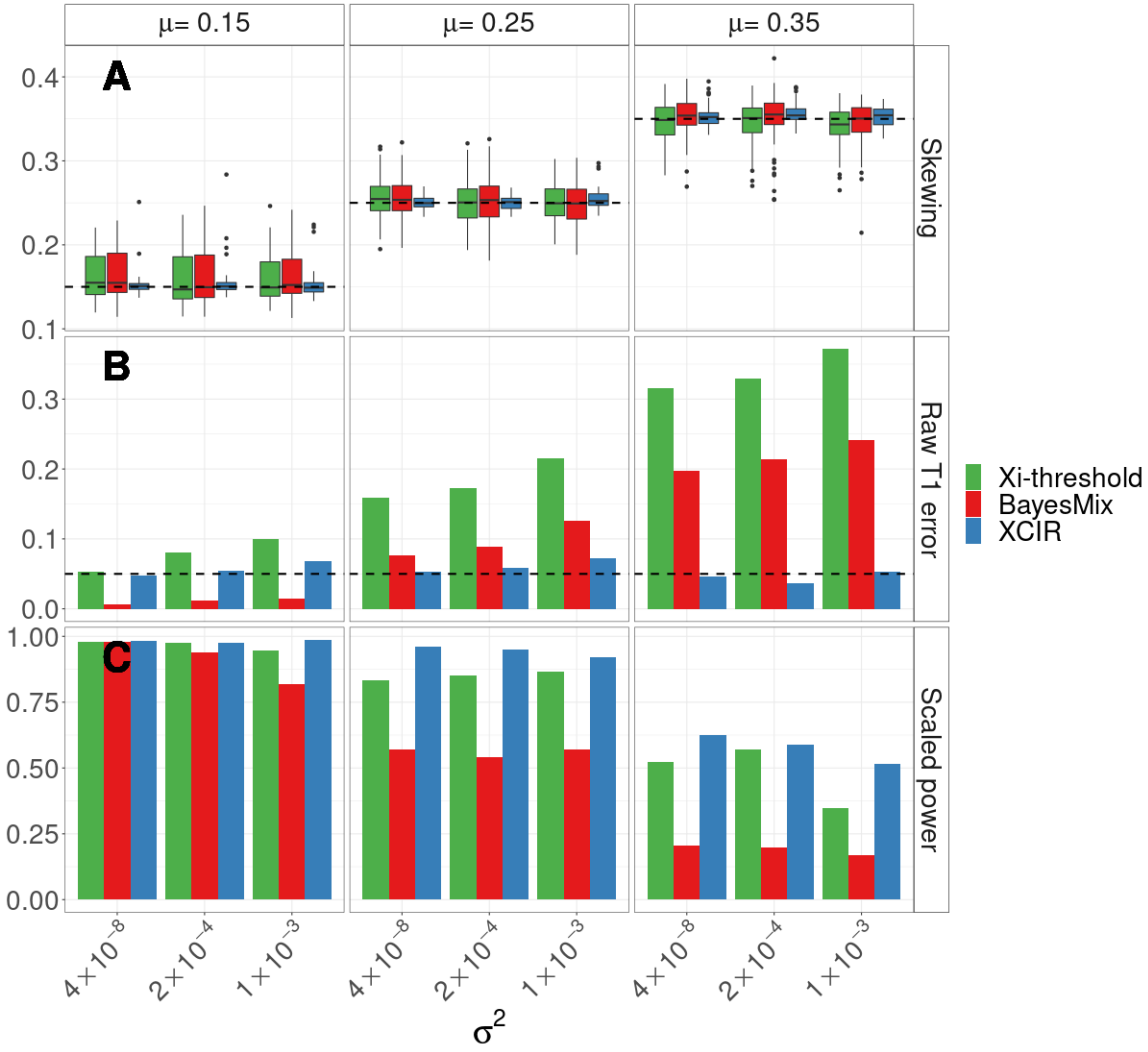

Fig S3. Simulation results assuming a random sequencing error model. The simulation scenario is similar to Fig 1. but here we allow the sequencing error to vary between genes. For each gene, we simulate the sequencing error as:

$$\pi_{err} \sim Unif(0, .1)$$

$$N_{err} \sim Bin(N_g, \pi_{err})$$

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.

Figure S4 - Single cell-derived lymphoblast cell line mixing experiments

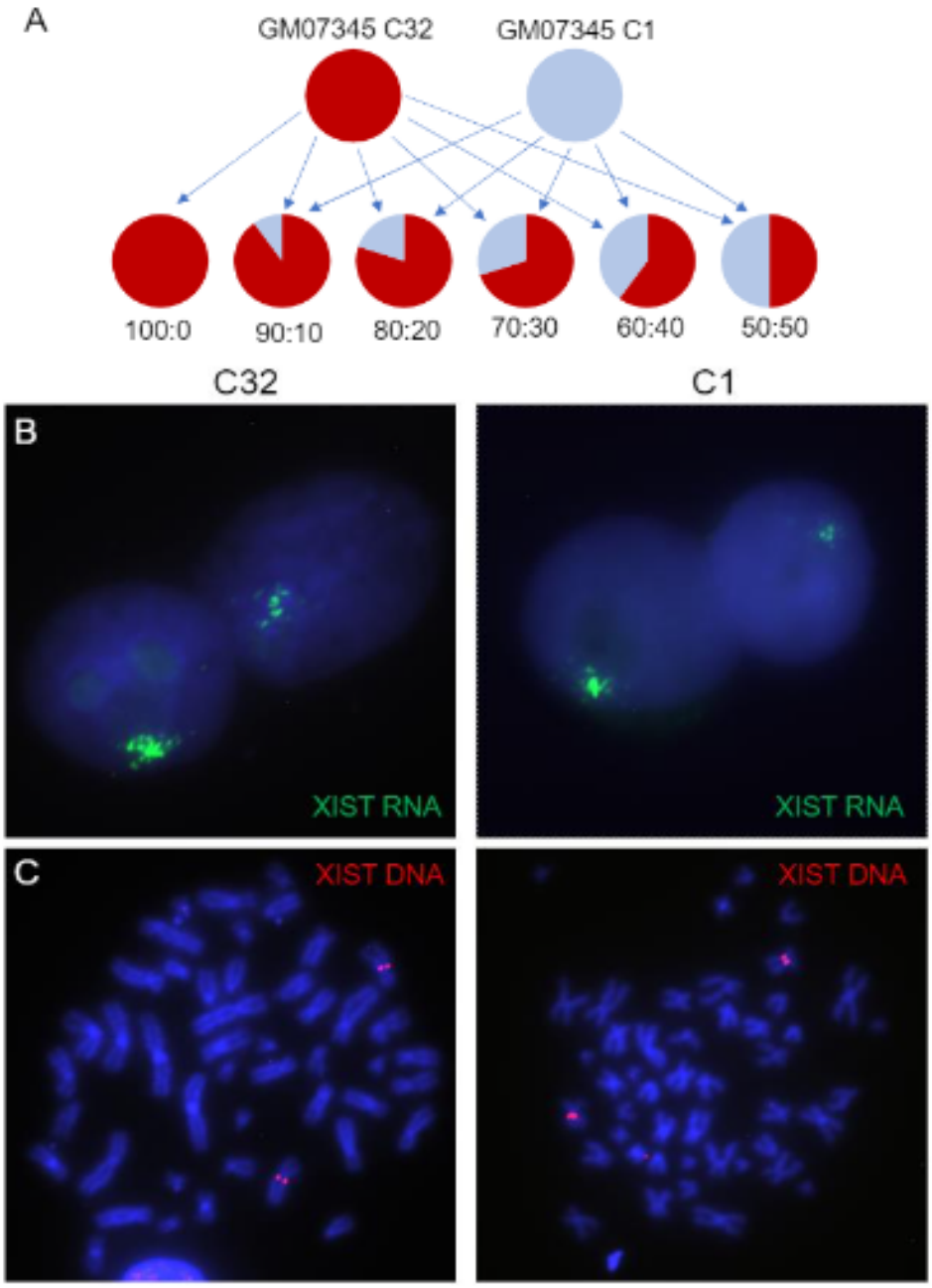

Fig S4. Single cell-derived lymphoblast cell line mixing experiments. (A) Overview of clonal cell line mixing strategy to generate "mosaic" samples. RNA-seq was then performed and independently analyzed for each "mosaic" mixed sample. **(B)** XIST RNA FISH of clonal LCLs to verify X chromosome content. 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)

Figure S5 - Accuracy change using phasing information

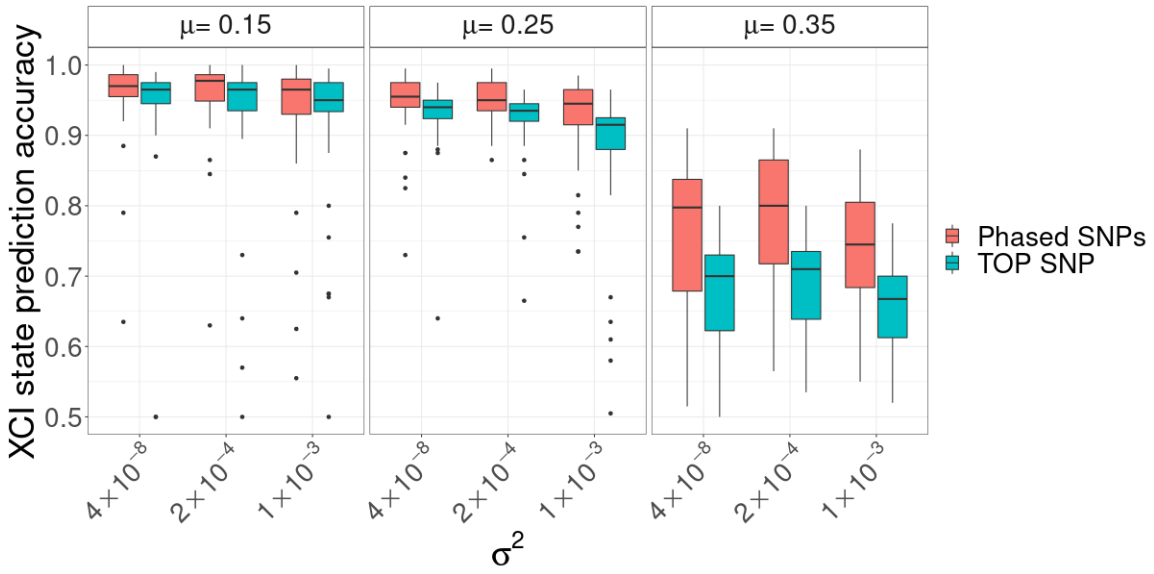

Fig S5. Effect of phasing on prediction accuracy. Using the same simulation setup as in Figure 1, we allow sampling of an additional SNP on the haplotype in both training and test set. XCIR is then run on the data using either the read count of the top SNP or using aggregated SNPs on the haplotype. We compute the accuracy of the predicted XCI states for each sample. Boxplots show the distribution of the prediction accuracy across skewings (μ) and ASE variance (σ^2).

Figure S6 - Quantile-Quantile plot of Lupus GWAS

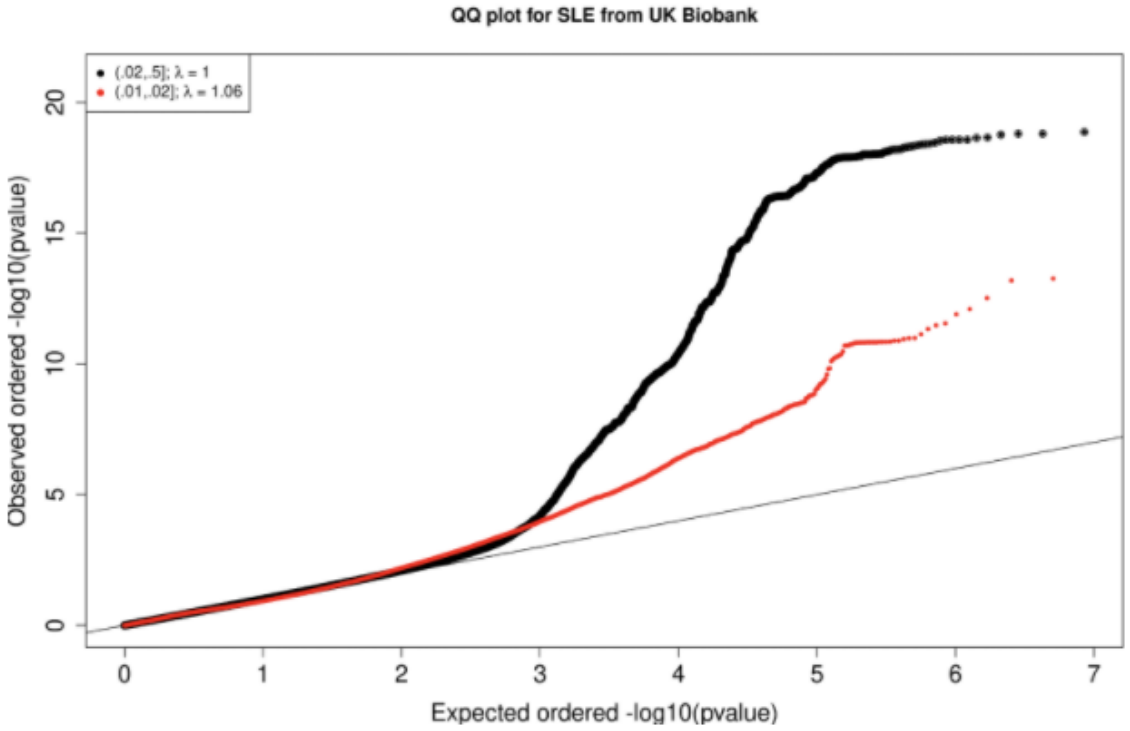

Fig S6. Quantile-Quantile plot of Lupus GWAS. GWAS significances extracted from UK Biobank summary statistics in self reported Lupus among female participants. Stratified for rare and common variants (MAF $< 2\%$ and MAF $\geq 2\%$, respectively).

Figure S7 - Manhattan plot of lupus GWAS

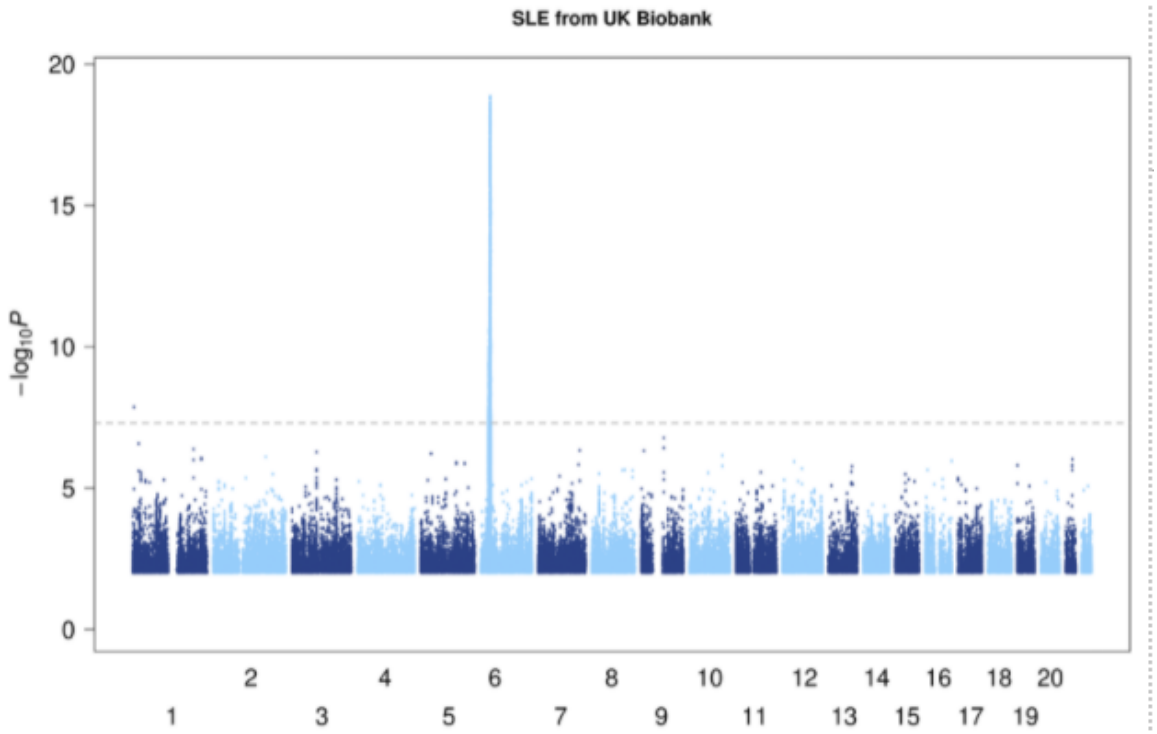

Fig S7. Manhattan plot of lupus GWAS. GWAS effect sizes and significances extracted from UK Biobank summary statistics in self reported Lupus among female participants.

Table S3 - Single cell-derived lymphoblast cell line mixing experiments results

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 S3. 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. Due to the relatively small number of genes that are consistently silenced, discreteness makes it impossible to calibrate type I errors to exactly 0.05. Therefore, we chose thresholds so that the type I error is closest to 0.05.

Table S7 - Effect of eQTL on XCI inference

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 S7. 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. The simulation scenario follows the settings used for Fig 1, including both sequencing and train errors. For each test gene, we simulate an eQTL based upon the MAF sampled from GTEx whole blood SNPs. Effect allele of the eQTL is placed on one haplotype. In samples that carry an effect allele, we assume that it can increase or decrease the ASE ratios by modifying the allelic expression (N_2) as follow

$$\begin{aligned} C &= N_1(\beta_{eQTL} - 1) \\ N_2' &= N_2 \pm C \end{aligned} \tag{2}$$

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. 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 genotype (G) of the eQTL and use the residuals with XCIR to perform XCI inference.

Supplementary Methods

Simulation of allele specific expression

For our simulation scenarios (**Figure 1, S1**), we use distributions and parameters that closely match what we observe in real data such as the GEUVADIS dataset.

First, we simulated samples true skewing parameters using a range of skewing mean $\mu \in (0.15, 0.25, 0.30)$ and variance $\sigma^2 \in (4 \times 10^{-8}, 2 \times 10^{-4}, 1 \times 10^{-3})$ 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 the following:

- 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 $N_1 \sim Bin(N, 0.01)$

All simulated parameters and values are reported in **Table S1**. 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 S3**).

Experimental Validation Using RNA-seq Data from Single-cell derived Clonal Cell Lines

In addition to simulation evaluation, we experimentally established the validity and power of proposed method using single-cell derived lines with identical Xa/Xi assignment. Single-cell derived clonal lines were isolated from the mosaic LCL GM07345 by plating into 0.7% Methocel (Dow Chemical Company) essentially as described [Kriegler, 1990, Freshney, 2000]. 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 [Allen et al., 1992] 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* [Kim et al., 2015] and adjusted for potential reference bias with *WASP* [van de Geijn et al., 2015]. 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.

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).

Male Female Differential Gene Expression Analysis

The alignment-free method *kallisto* was used to estimate transcript abundance, which is measured by transcript per million (TPM). The R package limma (v3.30.13) was used to conduct differential expression analysis of males vs. females for the GEUVADIS dataset.

Heritability estimates

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 escapes 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 and 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 downwardly biased. 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:

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

The estimate satisfies

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

As $0 \leq E_{Xi} \leq 1$ and $0 \leq f \leq 1$ it is easy to check that $E[\hat{\beta}_{XCI}] \leq_{ALT}$. So the genetic effect for each expressed alternative allele is always underestimated.

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:

$$l_j = \sum_k r_{jk}^2$$

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.

Some phenotypes, including Lupus, have very imbalanced case:control ratios, which could result in inflated type 1 error in low frequency variatns [Zhou et al., 2018]. For variants with MAF between 0.01 and 0.02, the QQ plot is well behaved and shows no inflation (**Figure S6**). The manhattan plot is also well-behaved, with clear peaks with a strand of signals. It shows no sign of standalone significant hits, which typically indicate spurious associations with rare variants (**Figure S7**).

We found no inflation for variants with MAF between 0.01 and 0.02 and no signs of spurious associations in rare variants. Therefore, we conclude that our results are reliable.

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 $\times$ sex interaction as additional covariates.

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.

References

[Allen et al., 1992] Allen, R. C., Zoghbi, H. Y., Moseley, A. B., Rosenblatt, H. M., and Belmont, J. W. (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):1229–39.

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

[Kim et al., 2015] Kim, D., Langmead, B., and Salzberg, S. L. (2015). Hisat: a fast spliced aligner with low memory requirements. *Nat Methods*, 12(4):357–60.

[Kriegler, 1990] Kriegler, M. (1990). Gene transfer and expression: A laboratory manual.

[van de Geijn et al., 2015] van de Geijn, B., McVicker, G., Gilad, Y., and Pritchard, J. K. (2015). Wasp: allele-specific software for robust molecular quantitative trait locus discovery. *Nat Methods*, 12(11):1061–3.

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