Predicting genetically regulated gene expression on the X chromosome

Xueyi Zhang, Piper Below, Adam Naj, Brian Kunkle, Eden Martin, William S Bush

Abstract

Despite the potential importance of genetic variation on the X chromosome, it is often omitted in disease association studies. The exclusion of the X chromosome has also propagated into the post-GWAS era, as transcriptome-wide association studies (TWAS) also ignore the X due to the lack of adequate models of X chromosome gene expression. In this work, we trained elastic net penalized models in the brain cortex and whole blood using whole genome sequencing (WGS) and RNA-seq data. To make generalizable recommendations, we evaluated multiple modeling strategies in a homogeneous study population of 175 whole blood samples for 600 genes, and 126 brain cortex samples for 766 genes. SNPs (MAF>0.05) within the gene's two megabase flanking window were used to train the tissue-specific model of each gene. We tuned the shrinkage parameter and evaluated the model performance with nested cross-validation. Across different mixing parameters, sample sex, and tissue types, we trained 511 significant gene models in total, predicting the expression of 229 genes (98 genes in whole blood and 144 genes in brain cortex). The average model coefficient of determination (R^2) was 0.11 (range from 0.03 to 0.34). We tested a range of mixing parameters (0.05, 0.25, 0.5, 0.75, 0.95) for the elastic net regularization, and compared the sex-stratified and sex-combined modeling on the X chromosome. We further investigated genes escaping X chromosome inactivation to determine if their genetic regulation patterns were distinct. Based on our findings, sex-stratified elastic net models with a balanced penalty (50% LASSO and 50% ridge) are the most optimal approach to predict the expression levels of X chromosome genes, regardless of X chromosome inactivation status. The predictive capacity of the optimal models in whole blood and brain cortex was confirmed through validation using DGN and MayoRNAseq temporal cortex cohort data. The R^2 the tissue-specific prediction models ranges from 9.94×10^{-5} to 0.091.

These models can be used in Transcriptome-wide Association Studies (TWAS) to identify putative causal X chromosome genes by integrating genotype, imputed gene expression, and phenotype information.

Introduction

The human X chromosome is composed of 155 million base pairs and contains nearly 1700 genes, and its copy number defines biological sex. While females have two copies of the X chromosome, gene promoters of one X chromosome are randomly silenced via DNA methylation in a process called X-chromosome inactivation (XCI) (Carrel and Willard 2005; Jégu, Aeby and Lee 2017). There is variation in this process as well, with a small proportion of genes that escape this silencing process (Wainer Katsir and Linial 2019; Navarro-Cobos, Balaton and Brown 2020), and potential skewing of X-inactivation based on the inheritance of the X chromosome (Plenge *et al.* 2002; Minks, Robinson and Brown 2008). Functionally, X-linked genes are responsible for a variety of processes, most notably human cognition and the development of multiple tissues, including neural and bone (Schwartz 2013). Numerous human phenotypes show sex disparities – for example, nearly two-thirds of individuals with Alzheimer's Disease are women (Eimer *et al.* 2018). While these differences may be due to systematic exposure to sex hormones or other endogenous factors, X chromosome genetic variation and its expression are also potential contributing factors.

Despite its potential importance, X chromosome variants are typically removed from genome-wide association studies (GWAS) during quality control, and they are often never subsequently analyzed due to the complexity of accounting for differences in chromosomal copy number and/or expression between males and females. In general, methods used for autosomal genotypes usually cannot be applied directly to X chromosome variants as they do not account for the genetic imbalance between sexes or expected X-inactivation. (Clayton 2009) (Zhu *et al.* 2015) (Wang, Talluri and Shete 2017) As a result, only one-third of genome-wide association studies include

evaluations of variants on the X (Wise, Gyi and Manolio 2013). The number of reported GWAS associations on the X chromosome is smaller than that of autosomes (Hay, Kumar and Ricaño-Ponce 2022), and Kukurba et al. discovered a relative depletion of cis-eQTLs on the X chromosome using a general eQTL-mapping pipeline on both autosomes and the X chromosome (Kukurba *et al.* 2016). These differences can either result from the inefficient power of the tests or the underlying different biological mechanisms and genetic architecture between autosomes and the X chromosome.

The omission of X chromosome variants also propagates to other GWAS-based study designs. Expression QTL and other studies of molecular phenotypes also often exclude the sex chromosomes. Popular extensions of GWAS, such as transcriptome-wide association studies (TWAS) (Li and Ritchie 2021), similarly omit the X. For example, PredictDB (predictdb.org) and the existing PrediXcan pipeline (Gamazon *et al.* 2015) does not fit genes on sex chromosomes. In this study, we evaluate multiple approaches to fitting expression prediction models for X chromosome genes, and produce a set of sex-stratified X-chromosome prediction panels. We then applied these panels to estimate the expression of X-chromosome genes for a large meta-analysis of Alzheimer's disease.

Methods

Data resources

We retrieved WGS and RNA-seq expression levels data (phs000424.v7.p2) from the brain cortex and whole blood sample from GTEx data (Lonsdale *et al.* 2013) to train the prediction models. For validation, we used whole-genome imputed genotype data and whole blood sample RNA-seq data from the DGN (Depression susceptibility Genes and Network) expression cohort (Battle *et al.* 2014) and WGS and temporal cortex (TCX). RNA-seq data from the MayoRNAseq study (Allen *et al.* 2016) has already been adjusted with AD diagnosis, sex, and age at death.

The GTEx donors were from post-mortem autopsy or organ and tissue transplantation settings. The DGN cohort reported no differences in eQTL detection between depression cases and controls. (Battle *et al.* 2014) MayoRNAseq study included patients with Alzheimer's disease (AD) and related dementias and health control on donors from the Mayo Clinic Brain Bank and Banner Health. (Allen *et al.* 2016) We believe these datasets can represent typical patterns of gene expression.

The key features of the three study populations are listed in Table 1. DGN and MayoRNAseq only recruited subjects of European descent from the United States. To evaluate different modeling strategies in a homogeneous study population and make generalizable recommendations, we restricted the samples to the GTEx participants who were self-identified as non-Hispanic white (NHW).

The age data in GTEx and MayoRNAseq is the age at death of the donors, while that for DGN is the age at sampling. MayoRNAseq subjects over 90 years old (recorded as 90+) were treated as age 91. It is noteworthy that the three studies exhibit different sex ratios, with males comprising two-thirds of the GTEx NHW cohort, females comprising two-thirds of the DGN cohort, and the MayoRNAseq TCX cohort displaying gender balance. Additionally, MayoRNAseq subjects are significantly older than GTEx and DGN.

Table 1. Overview of the study datasets.

	GTEx blood	GTEx cortex	DGN blood	MayoRNAseq cortex
Sample size	175	126	917	255
Female	58	46	648	130
Male	117	80	269	125
Age	56.1(10.5)	57.3(10.6)	44.7(10.9)	80.5(8.5)

Female	57.4(9.0)	57.9(11.4)	45.0(10.5)	81.5(8.5)
Male	55.5(11.2)	57.0(10.3)	44.1(11.7)	79.4(8.3)
Platform	WGS RNA-seq	WGS RNA-seq	Genotype RNA-seq	WGS RNA-seq
Genome Assembly	GRCh37	GRCh37	GRCh37	GRCh37
Number of expressed genes	600	766	486	587

Data processing

To keep the analysis concise, we dropped pseudoautosomal regions (PARs) (GRCh37) (X:60,001-2,699,520 and X:154,931,044-155,260,560) for this study. (https://www.ncbi.nlm.nih.gov/grc/human) Genetic variants with missing call rates exceeding 5% and samples with missing call rates exceeding 10% were removed. A minor allele frequency (MAF) threshold of >5% was applied to the overall dataset. In both sex-stratified and sex-combined modeling, the genotypes of female subjects were encoded as 0/1/2, which assumed entirely random X chromosome inactivation, and male genotypes were encoded as 0/2. (Keur *et al.* 2022) For sex-stratified analyses, the normalized expression was adjusted for covariates including top 3 PCs, PEER factors (Stegle *et al.* 2012), and sequencing platform. We also adjusted for sex in sex-combined analysis. We fit linear regression models of expression against the covariates and kept the residuals as the adjusted expression. (Supplementary Figure 1)

We also obtained a list of 59 genes (Supplementary Table 1) that escape the XCI (XCI escapees) from Navarro-Cobos et al. (Navarro-Cobos, Balaton and Brown 2020), and investigated whether a different modeling strategy would improve the predictions of these escaping genes.

Elastic net regression

We selected the elastic net regression to model gene expression levels on the X chromosome (Gamazon *et al.* 2015). While other approaches are available, elastic net regression allows for comparisons to both the PrediXcan autosomal models (Gamazon *et al.* 2015) and published sex-stratified models of autosomal genes (Mahoney *et al.* 2021). Elastic net regression is a regularized method that combines the LASSO (L_1) and ridge (L_2) penalties to select variables while fitting a linear regression model. Both penalty terms shrink regression coefficients by penalizing the regression model. The difference is that the LASSO penalty L_1 often reduces some regression coefficients to 0 when the ridge penalty L_2 does not.

The estimate from elastic net regression is $\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(\|y - G\beta\|^2 + \lambda[(1-\alpha)L_1 + \alpha L_2])$, where y is the vector of the measured expression of one gene in all the individuals, and G is the genotype matrix of all the individuals. The genotypes of female subjects were encoded as 0/1/2, and male genotypes were encoded as 0/2.

The shrinkage parameter λ controls the amount of penalization applied to the regression. Theoretically, an elastic net model favoring the LASSO penalty (mixing parameter $\alpha = 0$) will lead to a sparse model with a few SNPs with non-zero effects, while favoring the ridge penalty (mixing parameter $\alpha = 1$) will lead to a polygenic model. We implemented k-fold nested cross-validation to tune each gene's shrinkage parameter λ at any given mixing parameter α and evaluate the model performance for how the model predicts on the hold-out fold. The best shrinkage parameter λ was selected by the lowest mean squared error (MSE). And the model assessment metrics include the coefficient of determination R^2 and p-value, where the null hypothesis is that there is no correlation between predicted and measured expression. We only kept models with $R^2 > 0.01$ and p-value < 0.05, the significant models, for the downstream analyses.

PrediXcan implements the elastic net regression model with a mixing parameter $\alpha = 0.5$, where the penalty is defined as a balanced (50%-50%) combination of L_1 and L_2. This selection of $\alpha = 0.5$ has been shown to be optimal by previous studies of autosomes. (Gamazon *et al.* 2015; Wheeler *et al.* 2016) Since we are unclear about

the underlying regulatory function on the X chromosome, we fit elastic net regression models with different mixing parameters $\alpha = 0.05, 0.25, 0.5, 0.75$, and 0.95. The models were fit in both sex-stratified and sex-combined manners. (Figure 1).

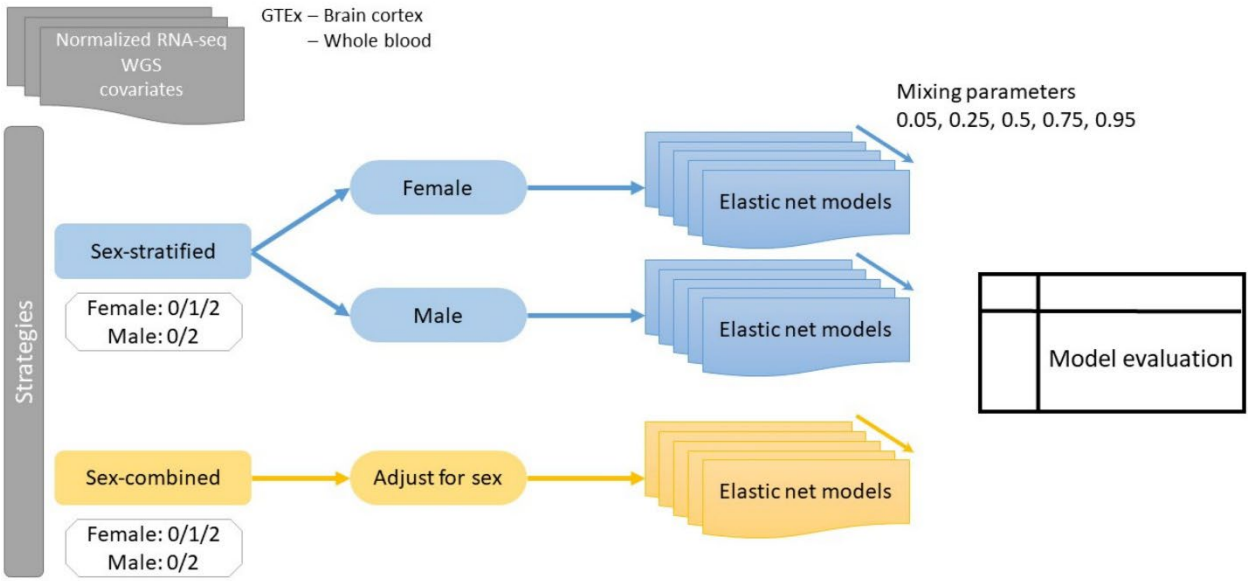

Figure 1. The working flow of model training using GTEx data. Elastic net regression models of the X chromosome gene expressions were trained in whole blood and brain cortex with different mixing parameters, in bother sex-stratified and sex-combined manner.

Model comparisons with two-way ANOVA

We compared the performance of the fitted prediction models to determine (1) whether the widely-adopted mixing parameter $\alpha = 0.5$ also works the best in X chromosome models; (2) whether sex-stratified models outperform the sex-combined models. Two-way ANOVA serves perfectly for our goal. Compared to one-way ANOVA, two-way ANOVA allows us to examine the combined effects of modeling strategies and biological factors.

We applied a data transformation $R^{2*} = \log R^2$ to the model's coefficient of determination R^2 to meet the normality assumption of two-way ANOVA since the distribution is right-skewed with a long tail. The density plots and QQ plots of the original R^2 and the transformed R^2 are in Supplementary Figure 2. Though a light tail still exists, the transformed prediction performance metrics exhibit a more normal distribution than the original R^2. The transformed R^2 shows homoscedasticity (homogeneity of variance) (see Supplementary Table 2). The design of GTEx and most other studies measuring human transcriptome provide independent observations of tissue-specific gene expression levels in each individual. Therefore, there is no violation of the assumptions of the two-way ANOVA test.

After transformation, we compare the model performance in different settings (mixing parameters, tissue types, female- vs. male-specific models, and sex-stratified vs. sex-combined analysis). We drew conclusions from the F statistics and p-values from the two-way ANOVA.

Model Validation

We predicted the gene expression levels in the whole blood and brain cortex by applying the sex-stratified GTEx gene prediction models to DGN and MayoRNAseq data, respectively (Gamazon *et al.* 2015). Genotypes of the DGN cohort were imputed to the HRC reference panel using the Michigan Imputation Server. We estimated the genetically regulated expression for each gene and tested its correlation with the measured expression with R^2 statistics.

Results

Given that there are numerous possible approaches to the analysis of X chromosome variants, we evaluated multiple options to determine the best modeling strategy for estimating the genetically regulated component of expression for genes on the X chromosome. We assessed whether the model parameters may differ by sex, and also attempted sex-combined and sex-stratified analyses. Given the unique biological features of the X chromosome, we also examined possible differences in the modeling of genes based on their X chromosome inactivation status.

Evaluation of mixing parameters within sex-stratified models

We trained sex-stratified models with GTEx data using five different mixing parameters: $\alpha = 0.05, 0.25, 0.5, 0.75$, and 0.95. We used a nested cross-validation approach (similar to PrediXcan: https://github.com/hakyimlab/PredictDB-Tutorial) that tunes the shrinkage parameter λ in an inner loop while model performance was evaluated in an outer loop. The models were fit in four tissue-sex sets: female-cortex, female-blood, male-cortex, and male-blood. We dropped intercept-only models since we are only interested in models that included genetic variants. As described in Methods, models with $R^2 < 0.01$ and $p > 0.05$ were removed for downstream analyses.

Across different α, sex, and tissue types, we trained 511 significant gene models for 98 genes in whole blood and 144 genes in brain cortex. The average number of SNPs in each model is 20.9, and the average $R^2 = 0.11$ (range from 0.03 to 0.34) (Table 2).

Table 2. Summary of sex-specified models in blood and cortex with different mixing parameters. Models with $R^2 < 0.01$ and $p > 0.05$ were removed.

Sex	α	Whole blood (600 genes)				Brain Cortex (766 genes)			
		Models number	Average SNPs	Average R^2	R^2 range	Models number	Average SNPs	Average R^2	R^2 range
Female	0.05	19	49.6	0.14	(0.064, 0.37)	27	64.7	0.18	(0.086, 0.35)
	0.25	21	29	0.14	(0.083, 0.39)	34	29.6	0.20	(0.074, 0.54)
	0.5	30	16.2	0.11	(0.058, 0.21)	25	17.2	0.18	(0.098, 0.34)
	0.75	22	23.1	0.16	(0.068, 0.49)	25	15.2	0.21	(0.11, 0.38)
	0.95	17	15.5	0.16	(0.062, 0.38)	28	10.3	0.19	(0.11, 0.37)
Male	0.05	24	36.1	0.057	(0.031, 0.16)	33	35.6	0.078	(0.052, 0.15)
	0.25	23	12.6	0.066	(0.035, 0.13)	34	13.9	0.076	(0.049,0.17)
	0.5	22	11.3	0.068	(0.034, 0.24)	28	8.25	0.089	(0.046, 0.17)
	0.75	20	10.2	0.073	(0.035, 0.24)	28	8.61	0.090	(0.051, 0.24)
	0.95	21	7.14	0.074	(0.038, 0.17)	30	5.73	0.083	(0.052, 0.15)

In both whole blood and brain cortex, the transformed R^2 of female-only models present a wider spread and higher central tendency than male-only models (Supplementary Figure 3). We checked the distributions of the

original R^2 in the sex-stratified models. For each sex-tissue set, we compared the original R^2 of the significant gene models with the R^2 of the same genes in the other sex-tissue set. (Figure 2 for the comparison of $\alpha = 0.5$ and Supplementary Figure 4-7 for the other mixing parameters)

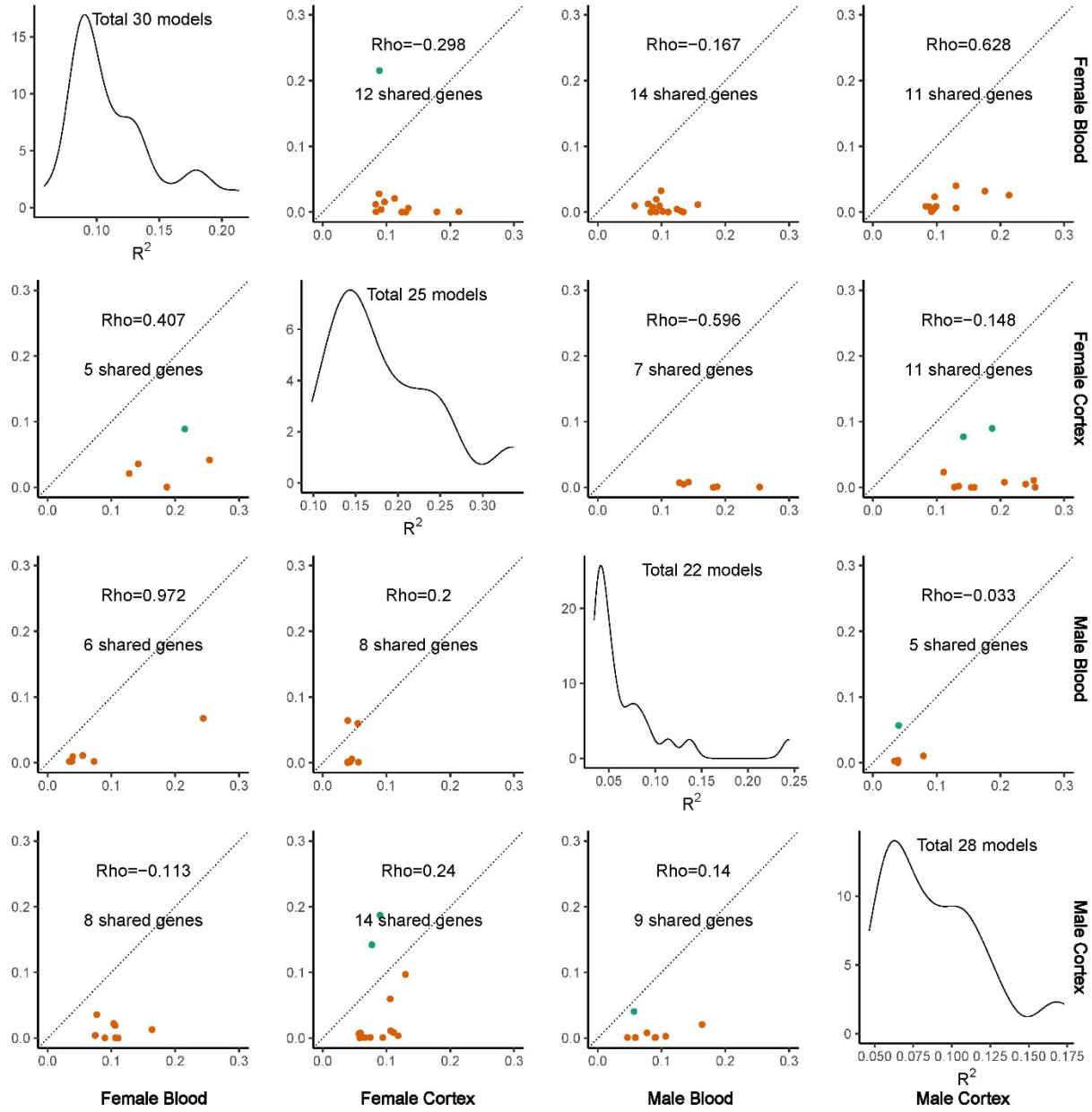

Figure 2. Comparison matrix of the balanced sex-stratified elastic net models trained in GTEx whole blood and brain cortex data. The four diagonal panels are the distributions of model R^2. The number of genes being modeled in each tissue-sex set is listed in the distribution plot. The off-diagonal panels compared the R^2 of the models in pairs of tissue-sex sets. The four rows are for the female-blood, female-cortex, male-blood, and male-cortex. For each row, the genes of the significant models in the tissue-sex set were matched in the other sets for the matching R^2. The green dots represent genes with a significant model ($R^2 < 0.01$ and $p > 0.05$) in another tissue-sex set, while the orange dots represent genes with a non-significant model in another tissue-sex set. The number of gene models and the Pearson correlation coefficients were listed in the comparison plots.

We performed a two-way ANOVA under the null hypothesis that there are no differences in the transformed R^2 among sex-stratified elastic net models with different mixing parameters $\alpha = 0.05, 0.25, 0.5, 0.75$, and 0.95. In both cortex and whole blood models, we found no significant differences in the average R^2 among different α's in

sex-stratified models ($p = 0.14$ for blood, $p = 0.46$ for cortex), but did observe significant differences in the average R^2 by the sample sex ($p = 6.10 \times 10^{-27}$ for blood, $p = 5.52 \times 10^{-56}$ for cortex). This result is driven by lower average R^2 in male-only models conditioned on tissue and α, as shown in Table 2. Sex by α interaction was also non-significant ($p = 0.68$ for blood, $p = 0.72$ for cortex). This finding indicates that different mixing parameters generally do not impact the performance of gene expression prediction models. In contrast, the models trained in female- and male-only samples have different coefficients of determination. While cortex models do capture more genes with $\alpha = 0.25$, to keep consistent with the literature and existing PredictDB autosomal models, we used a mixing parameter of $\alpha = 0.5$ (balanced penalty) for all subsequent analyses.

Differences in the tissue-specific sex-stratified models

For 105 models with a mixing parameter of $\alpha = 0.5$, the average number of SNPs in each model is 13.3, and the average $R^2 = 0.12$ (range from 0.03 to 0.54). There are 30, 22, 25, and 28 significant models for female-only blood, male-only blood, female-only cortex, and male-only cortex samples, respectively (see Figure 2). Among all the genes modeled, no blood-specific expression was modeled both in males and females, while two cortex-specific expressions (ENSG00000005893 *LAMP2*, ENSG00000101955 *SRPX*) were modeled in both sexes, and one gene was modeled in both tissues in male- or female-only data (ENSG00000197582 *GPX1P1*, ENSG00000013563 *DNASE1L1*).

We further compared the differences between models trained with data from different tissue types. We found a statistically significant difference in the model's average R^2 by sex ($p = 6.18 \times 10^{-14}$), and as expected, we observed a statistically significant difference between whole blood and brain cortex ($p = 1.53 \times 10^{-7}$). As shown in Table 2, the average R^2 of blood-specific models is lower than cortex-specific models, and the average R^2 of male-only models is lower than female-only models. Notably, there were no significant differences noted between sex-stratified models of autosomal genes from the same dataset (reported in Table 1 of (Mahoney *et al.* 2021)), regardless of sample size differences between males and females. The interaction between sex and tissue, however, was not significant ($p = 0.43$). We highlight that sex differences are just as crucial as tissue differences for X chromosome gene expression.

Sex-combined models with a balanced penalty

The tissue-specific sex-combined models were trained in blood and cortex, respectively, with a balanced mixing parameter $\alpha = 0.5$. After removing intercept-only models and models with $R^2 < 0.01$ and $p > 0.05$, we have 45 significant sex-combined models in whole blood and cortex. The average number of SNPs in the sex-combined models across different tissues is 10, while the average $R^2 = 0.052$ (range from 0.023 to 0.11). (Table 3) The average R^2 of the sex-combined models is smaller than that of sex-stratified models (single-sided t-test $p < 2.2 \times 10^{-16}$).

Table 3. Summary of sex-combine, balanced penalty models in blood and cortex. models with $R^2 < 0.01$ and $p > 0.05$ were removed.

Tissue	Escape XCI	Models number	Average SNPs	Average R^2	R^2 range
Blood[1]	×	23	8.96	0.045	(0.023, 0.087)
Cortex	×	20	11	0.059	(0.03, 0.11)
	√	2	11.5	0.051	(0.034, 0.067)

We compared different modeling strategies (stratified-female, stratified-male, and sex-combined models) to determine **whether the sex-combined analysis would improve the overall fit** for X chromosome genes by increasing the sample size. Considering the potential different underlying regulatory effects, we tested the genes escaping X chromosome inactivation (XCI escapees) and the other genes separately. Besides, there are some XCI

[1] None of the genes escaping XCI (escapees) were significant in blood.

escapees in the sex-stratified balanced penalty models (3 for female-blood, 2 for male-blood, and 1 for male-cortex).

For the models of XCI escapees (n=8), we found no significant differences in the average R^2 in the modeling strategies ($p = 0.15$) nor tissue types ($p = 0.55$). Contrarily, in both blood and cortex, genes not escaping from the XCI show significantly different R^2 between any pair of modeling strategies (one-way ANOVA $p = 3.4 \times 10^{-11}$ for blood and $p = 1.01 \times 10^{-15}$ for cortex).

Model Validation

The sex-stratified GTEx gene prediction models in blood and cortex were applied to DGN and MayoRNAseq validation datasets, respectively. 65 SNPs from GTEx whole boole models were not available in DGN whole-genome imputed genotype data, and 15 SNPs from GTEx brain cortex models were not available in MayoRNAseq WGS data (Figure 3 A and B). In DGN whole blood validation, 20 and 15 genes were predicted by stratified-female and stratified-male significant models, respectively. There is no gene predicted in both sexes. In MayoRNAseq TCX validation, we predicted 21 and 20 genes with stratified-female and stratified-male significant models, respectively, among which two genes were predicted in both females and males. (Figure 3 C and D). In addition, sex-combined significant models were applied to genes that escape from XCI in MayoRNAseq data ($n = 2$).

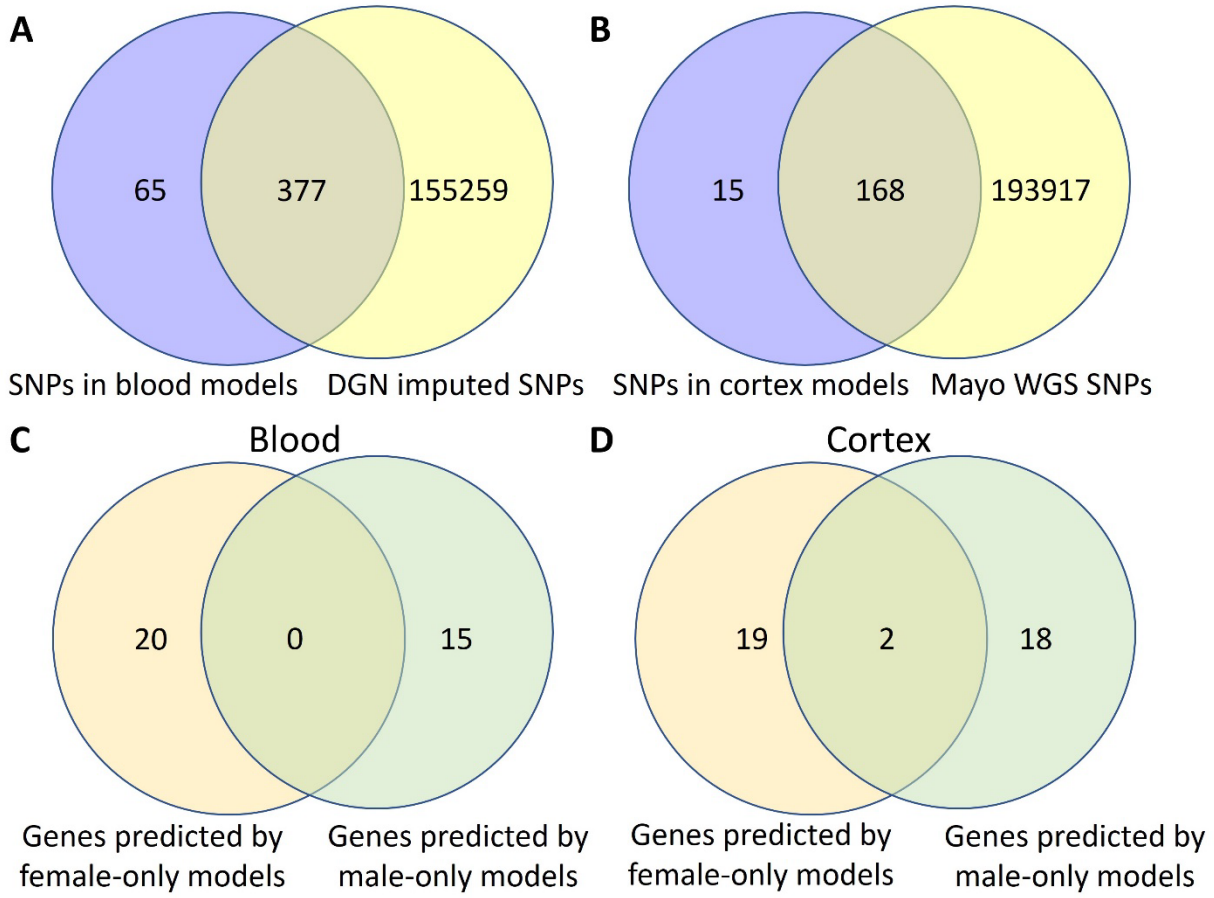

Figure 3. The number of SNPs in the validation datasets and the number of genes predicted.

We calculated prediction R^2 by regressing the estimated expressions against the measured expression via RNA-seq from the validation datasets (range from 9.94×10^{-5} to 0.091). Table 4 lists the prediction R^2 of the sex-stratified and sex-combined models in each tissue.

Table 4. Prediction R^2 in validation datasets. Note that there is no significant blood-specific sex-combined models for XCI escapees.

	Blood – DGN			Cortex – MayoRNAseq		
	Models Number	Average R^2	R^2 range	Models Number	Average R^2	R^2 range
Female-only model	20	0.014	$(9.99 \times 10^{-6}, 0.074)$	21	0.0097	$(9.94 \times 10^{-5}, 0.091)$
Male-only model	15	0.034	(0, 0.44)	20	0.010	$(1.05 \times 10^{-4}, 0.036)$
Sex-combined model (XCI escapees)	-	-	-	2	0.0010	$(1.28 \times 10^{-4}, 0.0020)$

A quantile-quantile plot showing the expected and observed distribution of prediction R^2 in blood samples (DGN) and cortex samples (MayoRNAseq) for all models (regardless of significance) is shown in Figure 4. These comparisons show a deviation from the expected null distribution, demonstrating that models are generally explaining variability in X chromosome gene expression.

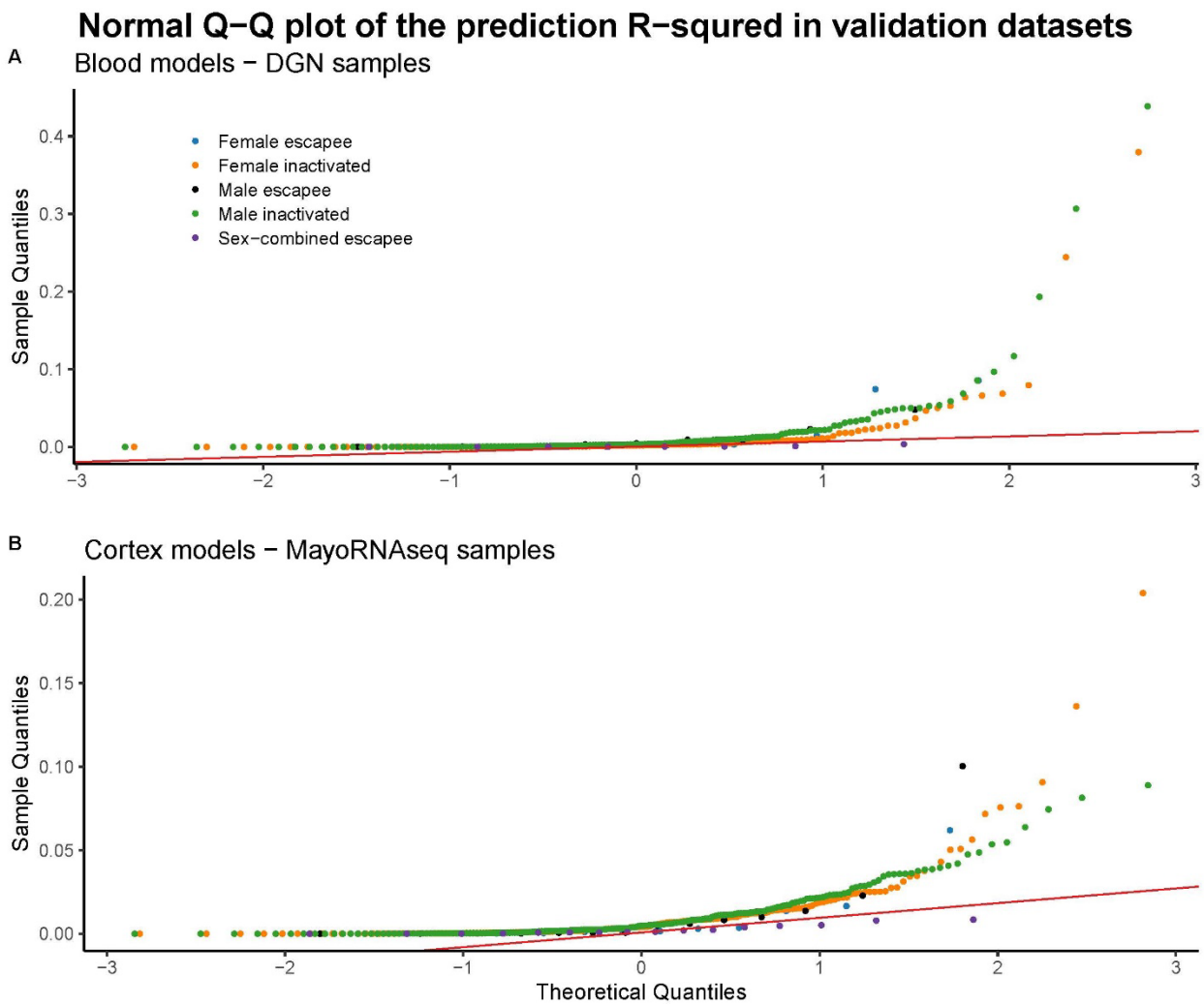

Figure 4. Q-Q plot of the prediction R^2 of blood (A) and cortex (B) models in validation datasets. All GTEx models were evaluated in these plots.

Discussion

In this work, we examined multiple modeling strategies for generating X chromosome gene expression prediction models. Based on our cross-validation results from analyses of the GTEx data for whole blood and brain cortex, we concluded that sex-stratified models (with $\alpha = 0.5$) provided the optimal R^2. We validated these models in two external tissue-matched datasets and found an average prediction performance of 0.012, which deviates from the expected distribution under a null hypothesis. Based on these models, we can estimate the genetically regulated expression of 20 genes for females and 15 genes for males on the X chromosome in whole blood, and 21 for females and 20 for males on the X chromosome in brain cortex.

DNA methylation silences one copy of X chromosome genes in females via X inactivation, and there are noted non-random patterns associated with this phenomenon, particularly skewed (Plenge *et al.* 2002; Minks, Robinson and Brown 2008) and escaped (Wainer Katsir and Linial 2019; Navarro-Cobos, Balaton and Brown 2020) X inactivation. We were surprised to observe that modeling genes escaping XCI with different strategies produced no substantial differences compared to non-escaped genes. In our sex-stratified models, we noted that our modeling approach assumes equal impact of each allele, which for the X chromosome would also assume totally random X chromosome inactivation. Accounting for skewed X-inactivation may be possible with availability of DNA methylation data or allele-specific expression analyses.

We note that the heterogeneity in gene expression prediction between males and females was substantial, and that these differences were not overcome by the power gain from increasing the sample size when combining males and females. Our X chromosome gene models fit within males and females are remarkably different from prior reports of similar sex-stratified analyses of the autosomes, also of the GTEx dataset (Mahoney *et al.* 2021). This finding provides support that differences between male and female models are not driven by sample size differences, and may also point to biological differences in the evolution of the X chromosome. For example, prior reports have shown depletion of regulatory variants on the X chromosome (Kukurba *et al.* 2016), etc.

Our strategy of selecting *statistically significant* models only is conservative, and the general distribution of prediction R^2 suggests that models not meeting our statistical significance criteria may still be useful in TWAS predictions (the full set of models is available in supplemental materials). We also chose an $\alpha = 0.5$ for all subsequent analyses as the R^2 distribution was not significantly different from our ANOVA analyses; however, we note that in cortex specifically, slightly more genes are modeled using an $\alpha = 0.25$. Therefore, for some TWAS applications, a lower alpha threshold may be preferred to capture additional genes.

There are several limitations to our study. First, we have a limited sample size; while eQTL sample sizes are generally substantially smaller than other types of genetic studies due to the expected larger effect sizes, our sex-stratified analyses are effectively halving the total sample. We also restricted analyses to whole blood and cortex tissues in this study. This was due to the availability of validation datasets, and our ultimate goal of applying these models toward TWAS of neurodegenerative and neuroinflammatory disorders. Another potential approach to allow for sex-specific effects while conserving statistical power is meta-analysis. Therefore, we explored the possibility of performing meta-analysis across sex-stratified models using the LASSOSum approach (Mak *et al.* 2017) (see supplemental methods); however, these models were difficult to fit due to model convergence issues and were not included in our evaluations.

Despite these limitations, this work provides a set of prediction models to enable X chromosome TWAS analyses of blood and cortex. Based on our results, we recommend using sex-stratified models for all X chromosome genes.

Acknowledgments

This work was supported by the National Institute on Aging [RF1AG061351 to P.B, A.N, W.B, R01AG059716 (Hohman), RF1AG070935 to (Griswold and) W.B].

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from dbGaP accession number phs000424.v7.p2 on 04/03/2018.

The results published here are in part based on data obtained from the AD Knowledge Portal (https://adknowledgeportal.org). The Mayo RNAseq study data was led by Dr. Nilüfer Ertekin-Taner, Mayo Clinic, Jacksonville, FL as part of the multi-PI U01 AG046139 (MPIs Golde, Ertekin-Taner, Younkin, Price). Samples were provided from the following sources: The Mayo Clinic Brain Bank. Data collection was supported through funding by NIA grants P50 AG016574, R01 AG032990, U01 AG046139, R01 AG018023, U01 AG006576, U01 AG006786, R01 AG025711, R01 AG017216, R01 AG003949, NINDS grant R01 NS080820, CurePSP Foundation, and support from Mayo Foundation. Study data includes samples collected through the Sun Health Research Institute Brain and Body Donation Program of Sun City, Arizona. The Brain and Body Donation Program is supported by the National Institute of Neurological Disorders and Stroke (U24 NS072026 National Brain and Tissue Resource for Parkinsons Disease and Related Disorders), the National Institute on Aging (P30 AG19610 Arizona Alzheimers Disease Core Center), the Arizona Department of Health Services (contract 211002, Arizona Alzheimers Research Center), the Arizona Biomedical Research Commission (contracts 4001, 0011, 05-901 and 1001 to the Arizona Parkinson's Disease Consortium) and the Michael J. Fox Foundation for Parkinsons Research.

Thanks Makaela Mews for the meta-analysis with LASSOSUM

Allen M, Carrasquillo MM, Funk C *et al.* Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. *Sci Data* 2016;**3**:160089.

Battle A, Mostafavi S, Zhu X *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Research* 2014;**24**:14–24.

Carrel L, Willard HF. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* 2005;**434**:400–4.

Clayton DG. Sex chromosomes and genetic association studies. *Genome Medicine* 2009;**1**:110.

Eimer WA, Vijaya Kumar DK, Navalpur Shanmugam NK *et al.* Alzheimer's Disease-Associated β-Amyloid Is Rapidly Seeded by Herpesviridae to Protect against Brain Infection. *Neuron* 2018;**99**:56-63.e3.

Gamazon ER, Wheeler HE, Shah KP *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics* 2015;**47**:1091–8.

Hay M, Kumar V, Ricaño-Ponce I. The role of the X chromosome in infectious diseases. *Briefings in Functional Genomics* 2022;**21**:143–58.

Jégu T, Aeby E, Lee JT. The X chromosome in space. *Nat Rev Genet* 2017;**18**:377–89.

Keur N, Ricaño-Ponce I, Kumar V *et al.* A systematic review of analytical methods used in genetic association analysis of the X-chromosome. *Briefings in Bioinformatics* 2022;**23**:bbac287.

Kukurba KR, Parsana P, Balliu B *et al.* Impact of the X chromosome and sex on regulatory variation. *Genome Research* 2016;**26**:768–77.

Li B, Ritchie MD. From GWAS to Gene: Transcriptome-Wide Association Studies and Other Methods to Functionally Understand GWAS Discoveries. *Front Genet* 2021;**12**:713230.

Lonsdale J, Thomas J, Salvatore M *et al.* The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* 2013;**45**:580.

Mahoney E, Janve V, Hohman TJ *et al.* Evaluation of Sex-Aware PrediXcan Models for Predicting Gene Expression. *Biocomputing 2022*. Kohala Coast, Hawaii, USA: WORLD SCIENTIFIC, 2021, 361–72.

Mak TSH, Porsch RM, Choi SW *et al.* Polygenic scores via penalized regression on summary statistics. *Genetic Epidemiology* 2017;**41**:469–80.

Minks J, Robinson WP, Brown CJ. A skewed view of X chromosome inactivation. *J Clin Invest* 2008;**118**:20–3.

Navarro-Cobos MJ, Balaton BP, Brown CJ. Genes that escape from X-chromosome inactivation: Potential contributors to Klinefelter syndrome. *Am J Med Genet* 2020;**184**:226–38.

Plenge RM, Stevenson RA, Lubs HA *et al.* Skewed X-Chromosome Inactivation Is a Common Feature of X-Linked Mental Retardation Disorders. *The American Journal of Human Genetics* 2002;**71**:168–73.

Schwartz CE. X Chromosome. *Brenner's Encyclopedia of Genetics*. Elsevier, 2013, 352–4.

Stegle O, Parts L, Piipari M *et al.* Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols* 2012;**7**:500–7.

Wainer Katsir K, Linial M. Human genes escaping X-inactivation revealed by single cell expression data. *BMC Genomics* 2019;**20**:201.

Wang J, Talluri R, Shete S. Selection of X-chromosome Inactivation Model. *Cancer Inform* 2017;**16**:1176935117747272.

Wheeler HE, Shah KP, Brenner J *et al.* Survey of the Heritability and Sparse Architecture of Gene Expression Traits across Human Tissues. Montgomery SB (ed.). *PLoS Genet* 2016;**12**:e1006423.

Wise AL, Gyi L, Manolio TA. eXclusion: Toward Integrating the X Chromosome in Genome-wide Association Analyses. *The American Journal of Human Genetics* 2013;**92**:643–7.

Zhu Z, Liang Z, Liany H *et al.* Discovery of a novel genetic susceptibility locus on X chromosome for systemic lupus erythematosus. *Arthritis Research & Therapy* 2015;**17**:349.