Title: Biological sex affects functional variation across the human genome

Authors: Angela G. Jones[1,2], Guinevere G. Connelly[1,2], Trisha Dalapati[1], Liuyang Wang[1], Benjamin H. Schott[1,2], Adrianna K. San Roman[1,2], and Dennis C. Ko[1,2,3]*

Affiliations:

[1]Department of Molecular Genetics and Microbiology, School of Medicine, Duke University; Durham, NC, USA.

[2]Duke University Program in Genetics and Genomics, Duke University; Durham, NC, USA.

[3]Division of Infectious Diseases, Department of Medicine, School of Medicine, Duke University; Durham, NC, USA.

*Corresponding author. Email: dennis.ko@duke.edu

Abstract: Humans display sexual dimorphism across many traits, but little is known about underlying genetic mechanisms and impacts on disease. We utilized single-cell RNA-seq of 480 lymphoblastoid cell lines to reveal that the vast majority (79%) of sex-biased genes are targets of transcription factors that display sex-biased expression. Further, we developed a two-step regression method that identified sex-biased expression quantitative trait loci (sb-eQTL) across the genome. In contrast to previous work, these sb-eQTL are abundant (n=10,754; FDR 5%) and reproducible (replication up to π_1=0.56). These sb-eQTL are enriched in over 600 GWAS phenotypes, including 120 sb-eQTL associated with the female-biased autoimmune disease multiple sclerosis. Our results demonstrate widespread genetic impacts on sexual dimorphism and identify possible mechanisms and clinical targets for sex differences in diverse diseases.

Humans display sexual dimorphism across a variety of traits. Anthropometric and physiological traits (e.g., body height, waist-to-hip ratio, etc.) are obvious examples of differences between males and females, but sex differences are also evident in the incidence, prevalence, severity, and treatment response in human disease. For example, females are more likely to develop autoimmune disease(*1, 2*) but show greater resistance to many infectious agents than males(*3*). Sex disparities are also apparent across cancers, with kidney(*4*), liver(*5*), skin(*6*), and laryngeal (*7*) cancers displaying a male bias, while breast and thyroid cancers (*8*) are more common in females. Phenotypic sex differences have been attributed to various factors including sex-specific hormone activity, sex chromosome complement, sex and gender influences on behavior, and complex differences in environment. However, despite these widespread differences, little is known about the mechanisms and biological functions that underlie sexual dimorphism in humans.

Previous studies have shown that gene expression differs by sex throughout the genome(*9, 10*), and several genome-wide association studies (GWAS) have identified genetic variants that show sex-specific effects in human disease(*11-14*). However, while previous studies have identified genes with sex-biased expression (sb-Genes)(*10, 15, 16*), how well sb-Genes replicate across diverse datasets and the underlying molecular mechanisms of why individual genes display sex-biased expression are not well characterized. Further, recent attempts to link gene expression differences to single genetic variants through sex-biased expression quantitative trait loci (sb-eQTL) discovery have reported few significant associations, with little replication across independent datasets(*10, 17-19*), leading to the dogma that sb-eQTL are very uncommon and tissue-specific(*10*). The lack of sb-eQTL previously discovered and replicated may be due to relatively small differences in the genetic control of gene expression by sex, low power in previous studies, or tissue-type specificity of genetic effects.

Here, we performed single-cell RNA-seq of 240 male and 240 female lymphoblastoid cell lines (LCLs) to identify transcriptomic sex differences and sb-eQTL across the genome. We identified 1,200 sb-Genes, replicated these in multiple datasets across tissues, and categorized their effects by underlying mechanism. Notably, we were able to assign ~97% of all sb-Genes to mechanisms involving sex chromosome copy number or transcription factor expression. We then developed a two-step method to identify 10,754 sb-eQTL, ~30 times more than previous reports(*10, 17-19*). Finally, we investigated the potential contribution of these sb-eQTL in human traits and disease by integrating sb-eQTL and GWAS data for a variety of sex-biased traits and found sb-eQTL are enriched in GWAS summary statistics for >600 phenotypes collected in the NHGRI-EBI GWAS catalog, catalogs of plasma and urine metabolites, and infectious disease phenotypes. Examining sb-eQTL in these GWAS datasets revealed many individual loci with concordant effects on sex-biased gene expression and effects on sex-biased disease risk, revealing a genetic basis for sexual dimorphism in human disease.

Sex-biased gene expression is reproducible and a subset is generalizable across tissues

Previous work in identifying sex-biased gene expression has identified thousands of genes displaying sex-biased expression (sb-Genes)(*10, 16*). The GTEx study examined conservation of sb-Genes across tissues and found largely tissue-specific effects, although a fraction of sb-Genes enriched for XCI-escapees were consistent across tissues. However, it is important to determine how well the sb-Genes identified in this ground-breaking work replicate across different individuals, particularly using a dataset with diverse genetic ancestry. Therefore, we sought to identify genes with sex-biased expression (sb-Genes) in a single cell type that would allow for testing replication in multiple datasets.

We previously developed a rapid and scalable pooled scRNA-seq method to uncover transcriptomic differences in a diverse cohort of 96 LCLs called scHi-HOST (single-cell highthroughput human in vitro susceptibility testing)(*20*). We expanded this method to 480 LCLs in twelve worldwide populations (**Fig. 1A**). Through integration of phased whole genome sequencing, we are able to deconvolute pooled transcriptomics to identify gene expression from each individual LCL, allowing for rapid screening of hundreds of cell lines. Importantly for this study, scHi-HOST was conducted with equal numbers of male and female LCLs, as defined by sex chromosome complement. Through this controlled method, we can isolate sex-specific transcriptomic signatures due to genetic differences without the impact of differential hormones and environmental conditions. In total, we sequenced 196,338 single cells that mapped to a single individual, with a mean of 409 cells per individual (**Table S1; Fig. S1**). As expected, we detected highly significant differences of X and Y chromosomal genes that matched the reported sex of these LCLs (**Fig. S1C**) but also observed significant changes in autosomal genes. In total, we discovered 1,200 genes with significant differential expression by sex (sb-Genes, FDR<0.05; **Fig. 1B**; **Table S2**). Also as expected, the effect size of sex-bias was largest for Y chromosomal genes (mean |log2FC|=6.59, 3.56-7.96), followed by X chromosomal genes (mean |log2FC|=0.34, 0.04-7.22), with autosomal genes showing the smallest effect (mean |log2FC|=0.02, 0.02-1.88; **Fig. 1C**). KEGG enrichment revealed sb-Gene functional enrichment in pathways involving immune responses and basic cellular functions (**Fig. S1**).

We assessed the level of replication of sb-Genes in an independent LCL dataset of comparable power (MAGE(*21*); n=635 LCLs) and across human tissues with the GTEx dataset (**Fig. 1D**). Autosomal and X chromosomal scHi-HOST sb-genes were tested for enrichment using a Fisher's Exact Test, restricting analysis only to genes expressed in both LCLs and the tissue being compared. The greatest fold-enrichment was observed with the MAGE and GTEx LCLs. Whole blood and spleen, tissues with high lymphocyte counts, were also highly enriched for scHi-HOST sb-Genes. Interestingly, we additionally detected enrichment across nearly all tested tissues. Thus, sex biased differences in gene expression are highly reproducible across three LCL datasets and a subset is generalizable across tissues. In total, 849 (71.5%) of tested scHi-HOST sb-Genes were reproduced in at least one other dataset, with a core subset of 21 primarily X chromosomal genes previously identified to escape XCI (*10*) showing sex-biased expression in all tested tissues. Interestingly, a single autosomal gene, *DDX43*, also showed conserved sex-biased expression in all tested tissues. This gene has previously been implicated in male fertility and spermatogenesis(*22*), and displays sex-biased methylation in newborn (*23*) and adult (*24*) individuals.

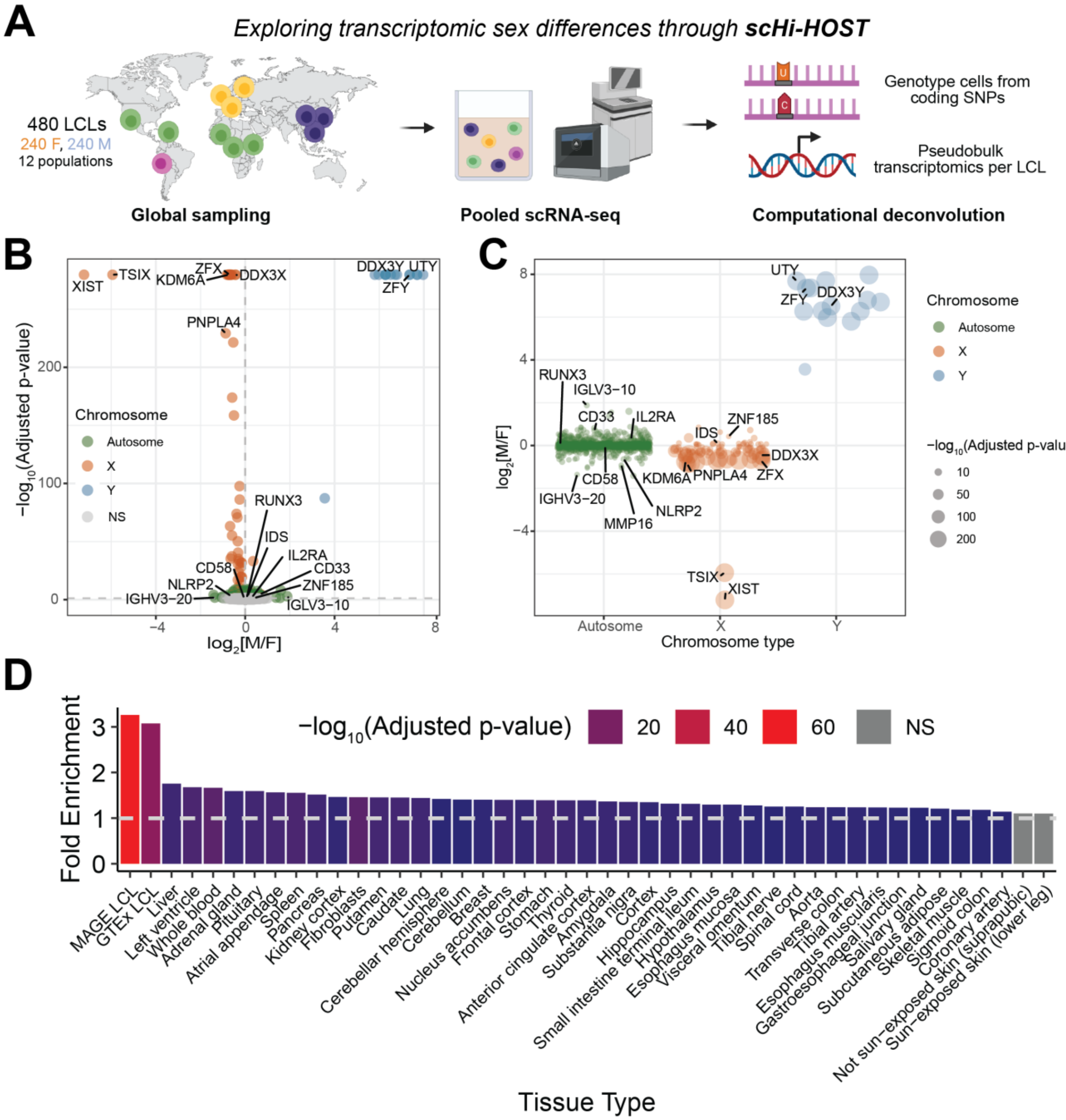

Figure 1. Sex-biased expression in LCLs is replicable and partly conserved across tissues. (**A**) Schematic of scHi-HOST pipeline. (**B**) Volcano plot of differential expression by sex colored by chromosomal location (FDR<0.05). (**C**) Sex-biased genes (FDR<0.05) displayed by chromosomal location. (**D**) Enrichment of sex-biased genes (FDR<0.05) across tissue types.

Partitioning of sb-Genes by mechanism

The variable effect size of sb-Genes points to different underlying mechanisms of sex-biased gene expression. The largest effects are due to the Y chromosomal genes that are entirely absent in females. We detected 14 Y chromosomal genes expressed in male LCLs. The next largest effect are genes that vary by copy number due to their presence on the X chromosome. In total, 92 X chromosomal genes display significant sex-biased expression, none of which lie in the pseudoautosomal regions shared by the X and Y chromosomes. Of these, 78 have previously been annotated as escaping X chromosome inactivation in at least one tissue (*10*) and display a larger sex difference in expression (mean |log2FC|=0.38) than both sex-biased autosomal genes (mean |log2FC|=0.02) and sex-biased X chromosomal genes not annotated as XCI-escapees (n=14, mean |log2FC|=0.15). In total, 7.7% of sb-Genes can be explained due to a direct difference in copy number.

However, the vast majority of sb-Genes occur on the autosomes where they exhibit more modest effect sizes. One step upstream of these differences in gene expression is activation or repression by transcription factors (TFs), which have been previously shown to be involved in sex differences across mammalian lineages(*25*). Thus, we hypothesized that some fraction of the remaining sb-Genes could be traced to variable expression of upstream TFs that lead to downstream targets exhibiting sex-biased expression. To determine if TF expression is predictive of sex-biased expression, we utilized a deep neural network (DNN) model previously trained and evaluated using >100,000 randomly drawn RNA-seq samples from the ARCHS4 resource(*26*). This DNN predictor uses the combinatorial effects of TFs to predict the expression of all genes. We assigned TF expression in females as the baseline condition before applying the TF fold-changes determined from the differential expression analyses described above. Through this approach, we discovered that DNN-predicted target expression (based on 1600 TFs) agreed with our empirical dataset (Spearman's ρ=0.25, p=1.72×10^{-15}). We further hypothesized that only significantly sex-biased TF expression would be sufficient to predict genome-wide sex-biased expression. We thus restricted the DNN prediction to only include effects of sex-biased TFs (n=127, FDR<0.05) and discovered that DNN-predicted expression still correlated with our empirical data (Spearman's ρ=0.20, p=4.53×10^{-11}).

We then sought to identify and model the effects of the TFs that could be driving sex-biased expression. TF enrichment analyses through ChEA3 (which ranks likely causal TFs based on RNA-seq co-expression and ChIP-Seq datasets(*27*)) revealed 608 TFs with significant (FDR<0.05) enrichment for sex-biased targets (**Table S3**) in at least one dataset. Out of these enriched TFs, 81 display sex-biased expression themselves (FDR<0.05). By investigating the targets of enriched TFs, we discovered that the majority (79%) of sb-Genes are targets of at least one enriched sex-biased TF. An additional 10% of sb-Genes are targets of non-sex-biased TFs that were also enriched for sb-Gene targets. Thus, ~89% of sb-Genes can be classified as possibly secondary to a difference in TF expression or activity (**Fig. 2A**).

To further identify which TFs primarily contribute to sex-biased expression, we integrated enriched TFs with the DNN described above. By removing the effect of a single enriched TF at a time from the model, we were able to determine that the sb-Gene *FOSL1* has the largest impact on sex-biased expression (**Fig. 2B**). *FOSL1* encodes the Fos-like 1 (FOSL1, also called FRA1) protein which can act as both a transcriptional activator and repressor(*28*). ScHi-HOST transcriptomics revealed that *FOSL1* displays significant male-biased expression (log2FC=0.37, p=7.39×10^{-6}, **Fig. 2C**) and primarily male-biased sb-Gene targets (**Fig. 2D**).

Three other sb-TFs (*ZNF730, ZFX, ZNF726*) notably affected DNN model fit. In each case, we broadly see concordant direction of effect in the sex-biased expression of the TF and the sb-Gene targets (**Fig. 2C, D**). While little is known about *ZNF730* and *ZNF726*, *ZFX* is a highly female-biased gene that escapes X chromosome inactivation and has previously been shown to drive X-responsive expression across the genome(*29, 30*). Thus, for *ZFX*, the mechanism for female-biased TF expression is copy number variation. Previous work has demonstrated that much of the effect of additional copies of the X chromosome are mediated by ZFX and that a homologous TF encoded on the Y, ZFY, activates a similar set of targets, with ZFX having a larger effect(*30*). In fact, *FOSL1*, *ZNF730*, and *ZNF726* all display at least nominal significance for correlation of expression with the number of X chromosomes in individuals with 1 to 4 X chromosomes (**Fig.**

2E; data from(*29, 30*)). Notably for all 81 enriched sb-TFs that we identified, 49 (60.5%) are also correlated with X chromosome copy number (p<0.05; **Table S4**). Additionally, 29 sb-TFs (35.8%) are significantly differentially expressed in ZFX CRISPRi knockdown fibroblasts (p<0.05; **Table S4**), which supports previous work highlighting ZFX as a key transcriptional regulator of genome-wide sex differences(*30*). However, the top autosomal sb-TFs previously discussed (*FOSL1, ZNF730, ZNF726*) are not significantly impacted during ZFX knockdown and display only limited correlation in expression (**Table S4**, **Fig. S2**). This suggests that the sex-biased expression of most TFs that appear to mediate the sex-biased expression of autosomal genes may arise secondary to the number of X chromosomes, although this effect may be ZFX-independent.

Thus, through integration of ChIP-Seq signatures, coexpression datasets, and DNN modelling, we identified sb-TF expression as a likely mechanism for sex-biased expression across the genome and highlight four primary TFs driving these sex differences. Of these four, *ZFX* has higher expression in females due to females having two copies, while the sex-biased expression of the other three TFs can be traced back to X-chromosome copy number effects independent of ZFX.

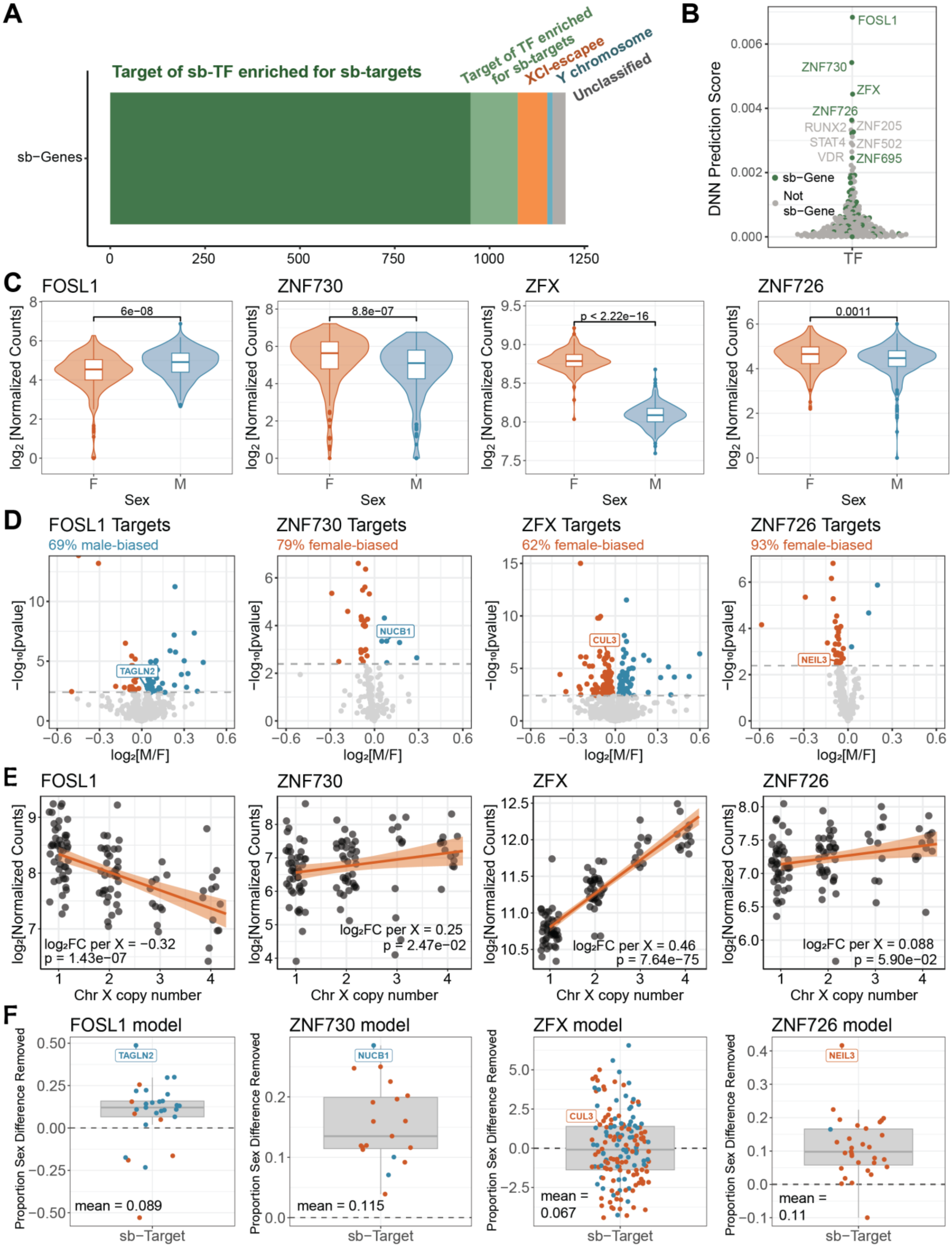

Figure 2. Nearly all sb-Genes can be explained by differences in sex chromosomes and their impact on sex-biased transcription factors. (**A**) Partitioning of sex-biased genes by location on the sex chromosomes (1.2% are on the Y; 6.5% are X-chromosome inactivation (XCI) escapees) and

as enriched targets of transcription factors. (**B**) Ranking of ChEA3-enriched transcription factors by DNN score where removal of a single transcription factor affects the correlation of the model. Transcription factors are colored by sex-bias. (**C**) Differential expression of the top four enriched transcription factors by sex. P-value from Wilcox rank-sum test. (**D**) Volcano plot of scHi-HOST differential expression by sex results of all ChEA3 targets of each transcription factor. Orange and blue reflect significantly female- and male-biased targets respectively (FDR<0.05). (**E**) Expression of each transcription factor from LCLs with varying X chromosome copy number (data from(*30*)). (**F**) Linear modelling results per transcription factor where each target is plotted as a proportion of total sex difference removed by the model. Orange and blue reflect significantly female- and male-biased targets respectively (FDR<0.05).

Modeling and testing the effect of a TF on sex-biased expression throughout the genome

We tested how well sb-Genes could be explained by the effects of sb-TFs through linear modeling and through examining experimental perturbation of sb-TF expression. First, the effects of each of the 81 enriched sb-TFs can be modeled based on two metrics from our dataset: 1) the relative expression in males vs. females for the TF (see Fig. 2C) and 2) the beta of the regression of TF on downstream target (see Fig. S2). Specifically, the expression level of the target gene can be modeled based on the relative increase in TF expression in one sex vs. the other ($TF_{sexeffect}$) and the linear regression coefficient of the effect of TF expression level on target expression level within one sex (β) while accommodating the regression intercept (α) and residual (ε): $Target = \alpha + \beta \times \left(TF_{measured} \times TF_{sex\_effect}\right) + \varepsilon$. The effect of sex-biased expression can then be simply modeled as a set of linear equations for all downstream targets of that TF.

We tested this model on the top four sb-TFs highlighted in the DNN models. We discovered that removing the sex difference in TF expression accounts for a mean of 11.8% of sex-biased expression in targets of FOSL1, ZNF730, and ZNF726 (**Fig. 2F**, **Fig. S2**). Modeling results for the sb-TF ZFX revealed that ZFX sex-biased expression still accounted for 6.7% of target sex differences, but the model tended to over- or under-correct expression (**Fig. 2F, Fig. S3**). This is likely due to the influence of ZFY, the Y chromosome homolog of ZFX.

We then tested the same models in the independent MAGE dataset. We again show that remediating the sex differential expression of *FOSL1* reduces the sex-bias in target expression (**Fig. S2**), with parallel results also found for *ZNF730* and *ZNF726*.

Finally, for two of these top sb-TFs, RNA-seq datasets have been published, defining how gene expression is affected by RNAi knockdown of *FOSL1* in Th17 cells (*31*) and CRISPRi knockdown of *ZFX* in fibroblasts(*30*). We determined that targets of FOSL1 that display sex-biased expression in scHi-HOST are enriched in differentially expressed genes (FDR<0.1) during *FOSL1* knockdown in Th17 cells (fold enrichment=1.87, Fisher's Exact Test p=0.032; **Table S5**). Similarly, we identified significant enrichment of sex-biased targets in differentially expressed genes (FDR<0.1) following *ZFX* knockdown (fold enrichment=1.25, Fisher's Exact Test p=0.001).

Thus, our categorization of sb-Genes appears accurate following computational and experimental testing. Modeling the effects of sb-TFs based on our scHi-HOST dataset accurately predicts the expression of downstream target genes in two independent datasets and experimental knockdown supports these findings. The largest sex differences in gene expression are secondary to copy number variation, while most sex differences are found in autosomal genes and can be traced back to sex-biased TFs, some of which are influenced by copy number of the sex chromosomes.

Identification of sb-eQTL with a two-step approach

While the large effects seen with sb-Genes on the X and Y are easily explained by their variation in copy number, and our evidence indicates that many of the smaller autosomal differences are secondary to these differences, some of these smaller differences might also be due to sb-eQTL. However, previous work has identified limited evidence of sb-eQTL through genome-wide interaction models(*17, 32*) or difference in slopes methods(*19*). These methods support the discovery of large-effect sb-eQTL but are hampered by low power due to high multiple-testing burden. In addition, previous studies that have identified sb-eQTL have sampled populations with primarily European ancestry, further limiting discoveries of functional genetic variation. Therefore, we developed a two-step pipeline in the diverse scHi-HOST cohort that reduces the multiple-test burden by first restricting analyses to eQTL that are significant in males or females separately before testing for sex interaction (**Fig. 3A**). We identified 364,480 eQTL in females and 375,445 eQTL in males in 6,721 and 6,923 eGenes (target genes of eQTL) respectively ($q<0.05$). Interestingly, we identified 7 eQTL on the Y chromosome for the eGene *TTTY14*, demonstrating the first example of genome-wide significant eQTL on the human Y chromosome (**Fig. S4**)(*33*). Combined, there were 502,384 unique eQTL for 9,042 eGenes. We then tested all unique eQTL for a SNP × Sex interaction and identified 10,754 sb-eQTL for 1,282 eGenes (FDR<0.05), including 35 X-chromosomal genes (**Table S6**). The majority (76.4%) of discovered sb-eQTL display a significant effect in only one sex, although differences in amplification (22.4%) as well as rare instances of reverse effect between sexes (1.2%) is also evident (**Fig. 3B**). The majority of sb-eQTL are present in intronic regions, although several are also found in 3' and 5' UTRs and other annotated regulatory regions of the genome (**Fig. S4**).

Crucially, we detect moderate replication of sb-eQTL in an independent LCL dataset (MAGE, $\pi_1=0.20$, **Fig. S4**). However, with a more stringent FDR threshold (FDR<0.01), this replication increased dramatically ($\pi_1=0.56$; **Fig. S4**), suggesting that while small effect sb-eQTL are difficult to detect in replication datasets, those with strong effects can be found across datasets. We then tested if sb-eQTL are associated with global changes in sex-biased expression. We identified a minor overall overlap of sb-eGenes and sb-Genes (7.4%), although this enrichment was not significant (Fisher's Exact Test p=1). Thus, in contrast to prior studies, we demonstrate that sb-eQTL are abundant throughout the genome and reproducible. However, they do not explain the majority of sex-biased gene expression in the human genome.

Examining the distribution of sb-eGenes across the genome, chromosomes 8 and 19 displayed a significantly increased proportion of sb-eGenes to non sb-eGenes (**Fig. 3C**). We asked whether differences in the abundance of specific transcription factor binding sites might contribute to the skewed genomic distribution. TF motif enrichment analyses revealed that sb-eQTL were significantly enriched in ZFX binding motifs ($\text{FDR}=2.49\times10^{-4}$; **Fig. 3D**, **Table S5**)(*34*), and in ZFX ChIP-Seq peaks in K-562 cells ($\text{FDR}=2.77\times10^{-9}$; **Table S5**)(*35*). Further, the density of ZFX ChIP-Seq peaks was correlated with the proportion of sb-eGenes on each chromosome (**Fig. 3E**). Therefore, differences in the abundance of ZFX binding sites may contribute to differences in the density of sb-eQTL among chromosomes. Further, the fact that ZFX expression is inherently sex-biased suggests that SNPs residing in ZFX binding sites can result in downstream sex differences across the genome.

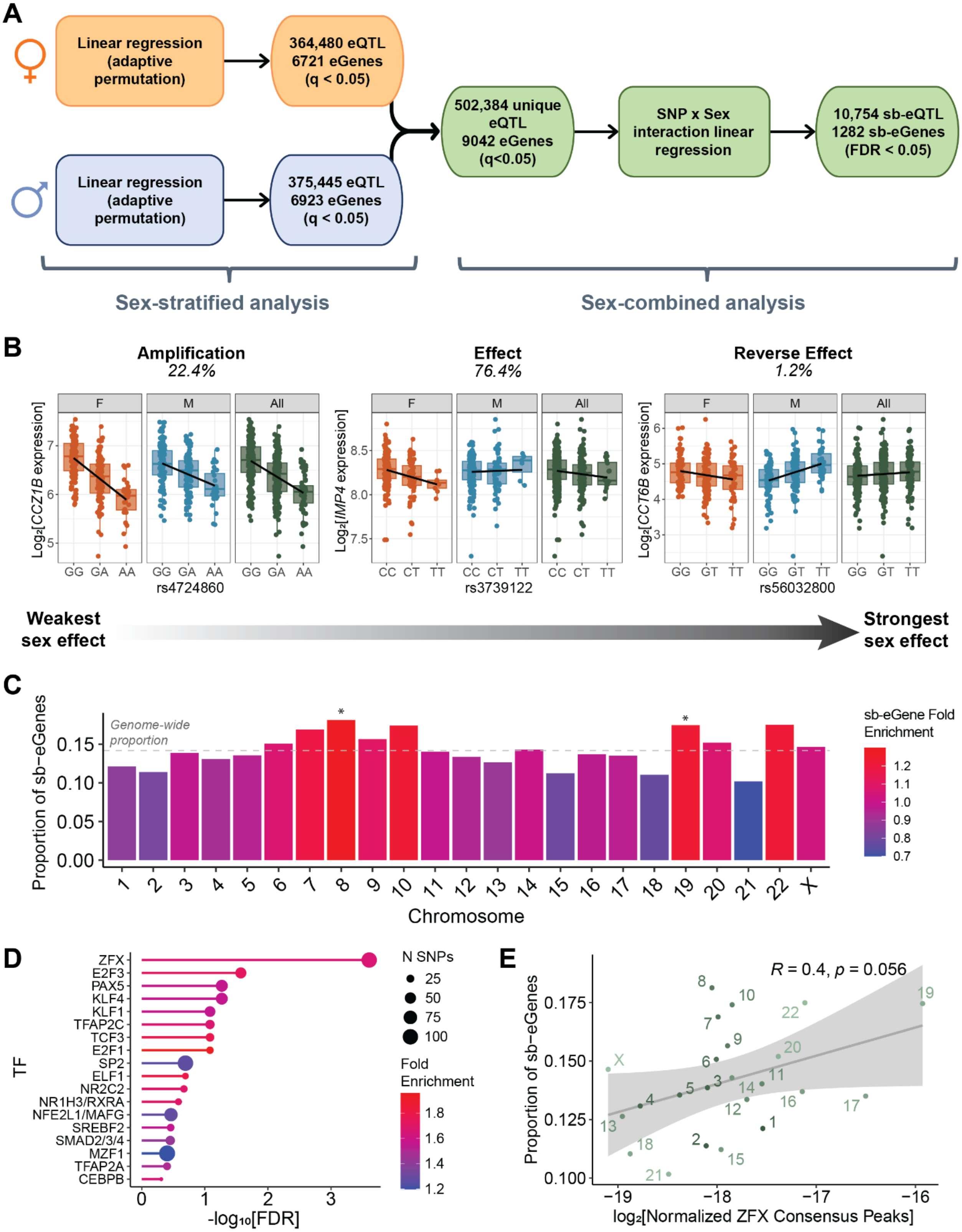

Figure 3. sb-eQTL are abundant throughout the genome but are of small effect size. (**A**) Two-step regression pipeline for the efficient discovery of sb-eQTL. (**B**) Significant sb-eQTL classified into three main modes of action. (**C**) Proportion of sb-eGenes to total eGenes by chromosome. *: p<0.05 two-proportion z-test. (**D**) Enrichment of sb-eQTL in transcription factor motif-altering

sites. (**E**) Correlation of *ZFX* ChIP-seq consensus peaks normalized by chromosome lengths to proportion of sb-eGenes to total eGenes with Pearson's correlation values.

sb-eQTL underlie some sex-bias in human traits

Although most sex-biased gene expression is not due to sb-eQTL, we hypothesized that the cases where it does occur might underlie some of the sex differences seen in human phenotypes and disease risk. This is most pronounced in autoimmunity where diseases such as systemic lupus erythematous and multiple sclerosis (MS) display severe sex-bias. To identify sb-eQTL associated with human disease risk, we utilized the genetic architecture tool iCPAGdb(*36*). iCPAGdb integrates the NHGRI-EBI GWAS catalog (*37*) with large datasets of plasma (*38*) and urine (*39*) metabolites and cellular host-pathogen phenotypes (*40*) to determine shared genetic factors. Briefly, iCPAGdb overlaps lead SNPs from the integrated GWAS data with SNPs from a user-submitted list, accounts for linkage between signals, and calculates enrichment. We discovered that 617 phenotypes significantly overlap with our sb-eQTL (FDR<0.05, **Table S7**). The most enriched phenotypes consist of highly sex-biased anthropometric traits such as height and body mass index (**Fig. 4A**). Specifically, for height, out of 2984 GWAS loci, 69 are or are in linkage disequilibrium ($r^2>0.4$) with leading sb-eQTL, a highly significant enrichment based on Fisher's exact test in iCPAGdb ($p=1.46\times10^{-69}$).

For each sb-eQTL associated with a human trait, we would predict that the association would be stronger in the sex with the larger eQTL effect. Therefore, we tested this prediction using UK Biobank sex-stratified GWAS data(*41*). We discovered 458 sb-eQTL (63 sb-eGenes) with genome-wide significant association ($p<5\times10^{-8}$) with height in either males or females and an additional 254 associated sb-eQTL (44 sb-eGenes) at the suggestive threshold of $p<1\times10^{-5}$. We can also detect improved colocalization of eQTL and GWAS loci by utilizing sex-stratified analyses. We identified a sb-eQTL (rs12948394) associated with expression of the gene *SOCS3* (**Fig. 4B**) and with body height (male $p=9.67\times10^{-8}$, female $p=7.68\times10^{-4}$). Upon sex-stratified colocalization analysis of scHi-HOST eQTL (**Fig. 4C**) and UK Biobank body height (**Fig. 4D**), we found strong colocalization in males (PP4=0.93) but not in females (PP4=0.05), suggesting that the effect of *SOCS3* expression on height is a male-specific effect.

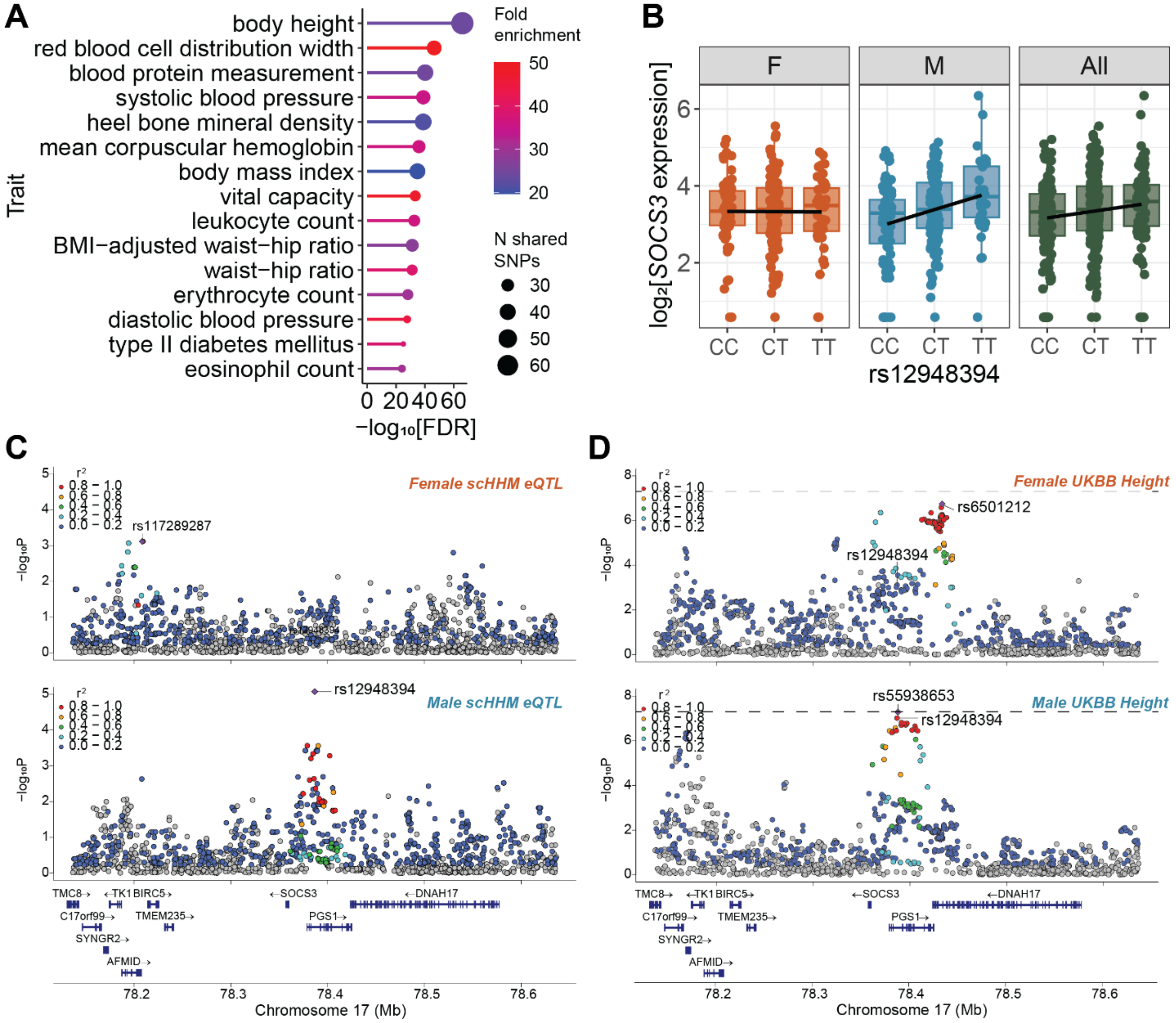

Figure 4. sb-eQTL are associated with human phenotypes in a sex-biased manner. (**A**) iCPAGdb enrichment of sb-eQTL in human phenotypes. (**B**) sb-eQTL rs12948394 shows association with *SOCS3* expression in males but not in females. (**C**) LocusZoom plots of female and male eQTL results demonstrate an association peak only in males. (**D**) LocusZoom plots of female and male UKBB body height GWAS results revealed a peak that colocalized with the eQTL signal only in males.

When results were restricted to disease susceptibility risk, we identified 132 significant enrichments with sex-biased autoimmune diseases predominantly enriched (FDR < 0.05, **Fig. 5A**). Due to the high enrichment of sb-eQTL in MS susceptibility loci (fold-enrichment=56.5, FDR= 4.54×10^{-19}) and the relevance of lymphoblastoid cells to this disease (as EBV infection of B cells has been demonstrated to play a causal role(*42, 43*)), we further investigated sb-eQTL effects in MS. Through integration of MS GWAS data from the International Multiple Sclerosis Consortium(*44*), we identified 80 sb-eQTL (16 sb-eGenes) with genome-wide significant association with MS, and an additional 40 (9 sb-eGenes) with suggestive association. We identified a variant on chromosome 8 (rs1466526) that is associated with sex-biased expression of the nearby gene *ZC2HC1A* (**Fig. 5B**), where the effect is larger in female individuals than male. This gene has previously been implicated with MS through integration of GWAS and eQTL data(*45*).

Because this genetic variant is an example of a sex difference in amplification, we hypothesized that we would see similar colocalization of GWAS signals with sex-stratified eQTL. We discovered that this GWAS variant does significantly colocalize with both female (PP4=0.93) and male (PP4=0.91) eQTL (**Fig. 5C**) as well as sex-combined eQTL (**Fig. S5**). In addition, we identified a variant on chromosome 10 (rs1790120) that shows association with expression of *DDX55* in only male individuals (**Fig. 5D**) and is associated with reduced risk of MS (**Fig. 5E**, $p=3.98\times10^{-8}$, OR=0.90). We then tested the sex-stratified eQTL and sex-combined MS GWAS for colocalization at this locus and found significant colocalization only in male individuals (PP4=0.90) and not in female individuals (PP4=0.06) suggesting a protective effect of *DDX55* expression in MS pathogenesis. With sex-combined eQTL, we do not see significant colocalization (PP4=0.68, **Fig. S5**), highlighting the importance of integrating sex-stratified analyses into genome-wide discoveries. Although DDX55 has not previously been implicated with MS pathogenesis, other DEAD-box helicases have been shown to have a protective effect(*46-48*).

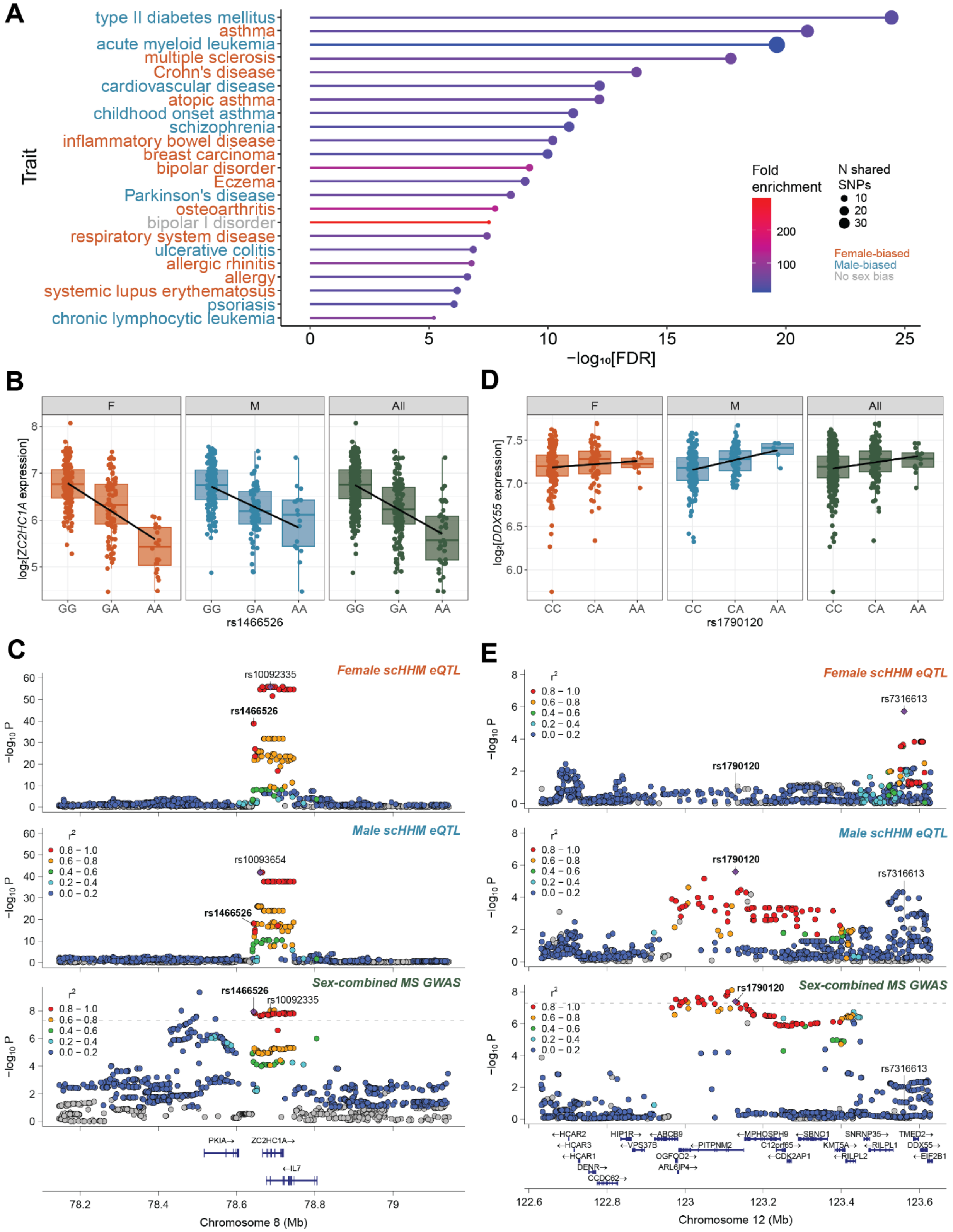

Figure 5. sb-eQTL are associated with disease-risk loci, particularly in autoimmunity. (**A**) iCPAGdb enrichment of sb-eQTL in human disease-risk loci. (**B**) sb-eQTL rs1466526 displays greater association with *ZC2HC1A* expression in females than in males. (**C**) LocusZoom plots of
5 female and male eQTL as well as sex-combined multiple sclerosis GWAS results for the region

surrounding *ZC2HC1A*. (**D**) sb-eQTL rs1730120 shows association with *DDX55* expression in males but not in females. (**E**) LocuzZoom plots of female and male eQTL as well as sex-combined multiple sclerosis GWAS results for the region surrounding *DDX55*.

Discussion

We identified sex-biased gene expression across the genome with 9% of genes displaying significant sex-bias. The largest sex differences in gene expression are the product of copy number variation due to the sex chromosomes, but these account for only 8% of sb-Genes in this study. However, the vast majority (79%) of sex-biased genes are targets of at least one sex-biased transcription factor, suggesting a possible mechanism of global sex differences in gene expression. One such enriched transcription factor*, ZFX*, is known to escape X chromosome inactivation and suggests that sex chromosome copy number is the underlying mechanism leading to sex-biased expression of *ZFX* targets. In fact, sex chromosome number may also underlie the sex-biased expression of autosomal TFs and drive sex-biased expression of targets. Thus, our data demonstrate that much of the observed sex-biased expression of human genes can be traced directly or indirectly to the number of X chromosomes, although our model cellular data represents cell-autonomous sex differences and does not account for the influence of sex hormones.

Beyond the effects of the X chromosome, the initial mechanism of sex-bias of autosomal genes may have additional contributing factors, such as differential chromatin accessibility between biological sexes(*32*). In addition, a small portion of sex-biased genes (10%) are not targets of sex-biased transcription factors. These examples of sex-biased expression may be due to additive effects of multiple transcription factors or a possible amplification of minor sex differences in transcription factor expression. Lastly, not included in this study was the consideration of transcription-associated RNA-binding proteins or large-scale 3D-genome structure which may explain some residual sex differences in gene expression. But beyond categorizing sex-biased transcriptional effects, we have modeled the effects of the most important sex-biased TFs on their target genes and experimentally examined their effects in loss-of-function transcriptomic datasets. In combining this categorization of sex-biased effects with modeling and empirical testing, we move to the level of molecular detail necessary to understand why individual genes have sex-biased expression and how this might be manipulated to impact human health.

While it is well-accepted that sb-Genes are abundant throughout the genome, with many conserved across tissues(*10*), the existence and relevance of sb-eQTL is much more contentious. In contrast to previous reports, we identified extensive evidence of sex differences in *cis*-acting genetic regulation (10,754 sb-eQTL) that replicated in an independent LCL dataset. Due to their subtle effects, sb-eQTL are difficult to identify in genome-wide analyses. Previous studies have mainly tested genome-wide for SNP × sex interaction(*17, 32*), considered sex differences in effect size(*19, 49*), or only considered variants with sex-combined association with gene expression(*10*). The prior approaches suffer from large multiple correction burden while the latter is biased towards eQTL with sex-differences in amplification. Our utilization of a two-stage pipeline reduces multiple testing burden while focusing analysis on expression-associated variants. In addition to increased power, the two-stage approach allows for well-powered detection of all three kinds of sb-eQTL, and in our analysis, sb-eQTL that display an effect only in one sex are the most common. Our finding that sb-eQTL identified in this way are abundant and replicable across datasets sets a precedent for using this method across other tissues and disease contexts.

Enrichment of sb-eQTL in *ZFX* binding sites again suggests sex chromosome complement number and transcription factor activity as an underlying mechanism of genome-wide sex differences. This is supported by *ZFX* binding sites predicting sex-biased eQTL distribution across the genome. However, there are additional factors that may impact sb-eQTL activity that were not investigated including differential chromatin accessibility and additional RNA-binding transcriptional regulators. Further investigation is required to identify tissue-specific and conserved regulation. In addition, this data represents only a single timepoint and does not consider how sex differences change during development or in response to stimuli to impact human biology.

While not accounting for large-scale differences in sex-biased gene expression, sb-eQTL provide additional context for complex traits and diseases and may even reveal mechanisms that explain sex differences in disease risk and severity. Indeed, we identified sb-eQTL that show stronger association with human traits and better colocalization preferentially in one sex. In MS, a highly female-biased autoimmune disease, we identified sex-bias in a known risk gene (*ZC2HC1A*) as well as discovered a novel gene connection (*DDX55*). Further stratification of genome-wide discoveries by sex will continue to reveal novel biological mechanisms and support the inclusion of sex as a biological variable. These results serve as a blueprint to discover mechanisms of sex differences across the genome and can even be applied more broadly to uncovering any gene × environment interaction that affects human health.

METHODS

Lymphoblastoid cell lines

240 male and 240 female 1000 Genomes LCLs from 12 worldwide populations (Table S1) were purchased from the Coriell Institute. LCLs were maintained at 37°C in a 5% CO2 atmosphere and grown in RPMI 1640 media (Invitrogen) supplemented with 10% fetal bovine serum (FBS), 2 mM glutamine, 100 U/mL penicillin-G, and 100 mg/mL streptomycin. Equal numbers of pooled LCLs are stored in liquid nitrogen vapor phase.

Single-cell transcriptomics

Equal numbers of LCLs were pooled and added to a 24-well non-tissue culture treated plate in PBS with 0.4% BSA, Mg^{2+}, Ca^{2+}, 100 U/mL penicillin-G, and 100 mg/mL streptomycin. Uninfected controls from the previously described infection (*20*) were spiked after 3 h with 500 μL of the LCL growth media described above. After 24 h, pooled cells were collected, spun down, and resuspended in PBS with 0.04% BSA for single-cell cDNA library preparation. This process was repeated three times with 48 LCLs used in both scHi-HOST Neo (previously scHH-EIK in(*20*)) and scHi-HOST Morpheus (previously scHH-LGC in(*20*)) and 384 LCLs used in scHi-HOST Trinity (not previously described; Table S1).

Cell counts and viability were collected on a Guava EasyCyte HT system by 7-AAD staining before dilution to a solution of 1 million cells/mL with intended capture of 10,000 cells/well for scHi-HOST Neo and Morpheus and to 1.25 million cells/mL with intended capture of 20,000 cells/well for scHi-HOST Trinity. The 10x Chromium Single Cell 3' platform version 3.1 (Pleasanton, CA) was used to generate individual barcoded cDNA libraries for each well following the manufacturer's protocol. Briefly, the 10x Chromium Controller separates individual cells into nanoliter-scale gel beads in emulsion (GEMS) where cell-specific barcoding and oligo-dT-primed reverse-transcription occurs. For scHi-HOST Neo, 13,675 uninfected droplets were captured in 1 Chromium well. For scHi-HOST Morpheus, 36,716 uninfected droplets were captured across 2 Chromium wells. For scHi-HOST Trinity, 85,558 uninfected droplets were captured across 4 Chromium wells.

cDNA samples from scHi-HOST Neo were single-indexed and sequenced on an Illumina HiSeq system with a target depth of 50,000 reads per barcoded droplets. Reads were sequenced with read 1 length of 150 base pairs (bp) and read 2 length of 150 bp. This resulted in a mean depth per cell of 37,815 reads. scHi-HOST Morpheus samples were dual-indexed and sequenced on one Illumina NovaSeq S4 flow cell with target depth 100,000 reads per barcoded droplet. Reads were sequenced with read 1 length of 28 bp and read 2 length of 150 bp resulting in a mean depth per cell of 67,415 reads. Lastly, cDNA samples from scHi-HOST Trinity were dual-indexed and sequenced on a NovaSeq S4 flow cell with target depth of 50,000 reads per droplet. Reads were sequenced with read 1 length of 28 bp and read 2 length of 150 bp which resulted in a mean depth per cell of 69,933 reads.

Single-cell RNA-seq alignment and LCL assignment

As previously described(*20*), we processed all raw sequencing results using the 10X Genomics CellRanger 7.0.1 with default parameters unless otherwise indicated. Reads from each sample were mapped to GRCh38. For the single-indexed scHi-HOST Neo, we used the 10X Genomics Index-hopping-filter to removed index-hopped reads (https://github.com/10XGenomics/index_hopping_filter) before mapping to the human genome. To remove possible ambient RNA contamination, we used CellBender (*50*) v.0.3.0 with –fpr 0.01 and –epohcs 150. CellBender uses a deep generative model to learn a background noise profile to delineate cell-containing and empty droplets and provide noise-free gene count quantification.

We then used Demuxlet (*51*) to assign each barcoded read from cell-containing droplets to an LCL. Demuxlet utilizes a CellRanger bam file with barcoded sample reads and a VCF containing all autosomal genotypes (obtained from the 1000 Genomes Project, NYGC 30x high-coverage release) to computational deconvolute each droplet into its corresponding LCL identity. After assigning each droplet to its corresponding LCL, CellBender gene counts were pseudobulked by sum to produce gene counts for each LCL. To reduce any batch effects in expression between the scHi-HOST Neo, Morpheus, and Trinity datasets, we conducted ComBat-seq(*52*) correction and merged all LCL gene counts into the combined dataset scHi-HOST Matrix which was used for all downstream analyses.

Sex-biased differential expression

The R package "DESeq2" (*53*)was then used to perform all differential expression analyses. Feature counts were normalized according to the DESeq2 default "Median of Ratios" method which accounts for differences in sequencing depth and RNA composition between samples. Genes with at least 5 normalized counts across 10% of individuals were retained for further analyses. Differentially expressed genes by sex were then discovered in DESeq2 with LCL population as an additional covariate to account for possible population effects. KEGG pathway enrichment was conducted using the R package "clusterProfiler"(*54*).

A parallel analysis was conducted to discover sb-Genes in the MAGE dataset(*21*). Raw counts from individuals not included in scHi-HOST Matrix were normalized as described above with LCL population as a covariate. Sb-Genes in GTEx were previously discovered using a linear model with covariates accounting for known sample and donor characteristics, as well as surrogate variables that capture hidden factors of expression variability(*10*).

Transcription factor enrichment and deep neural network modeling

Transcription factor enrichment of sb-DEGs was accomplished in ChEA3 (*27*) through the utilization of two co-expression datasets (ARCHS4 (*55*) and GTEx(*56*)) and two ChIP-Seq datasets (ENCODE (*57*) and Literature(*27*)). Briefly, the overlap of sb-DEGs and known TF targets from these distinct datasets was compared and enrichment calculated with a Fisher's Exact Test. Enriched TFs were then prioritized for additional analyses based on the Chea3 Mean Rank Score.

The deep neural network (DNN) model developed by Magnusson, et al. (*26*)was used to predict gene expression from TF expression. This neural network was trained on the ARCHS4 expression database (*55*) using two hidden layers with 250 hidden nodes each. Using this model, we predicted sex-biased expression in scHi-HOST Matrix from TF expression and female/male fold-change results from the above DESeq2 analyses. We then compared the predicted values of sb-Genes to observed quantifications using Spearman's correlation. To determine if sb-TFs are sufficient to predict sb-Gene expression, we retained only those TFs with significant sex-biased expression (n=127) and repeated the model as outlined above. Lastly, to prioritize TFs that highly impact sex-biased expression, we iteratively removed the effect of one TF at the time and evaluated the difference in model performance.

Transcription factor linear modeling

We utilized linear models to determine the effect of removing the observed sex-bias in TF expression on target genes. We tested all sb-Genes annotated as targets of a particular TF in Chea3 for co-expression by generating a linear model of target expression as a function of TF expression. Target and TF expression matrices were generated as outlined above in the DESeq2 differential expression analyses. We then modeled how the removal of sex differential TF expression would impact expression of those targets with significant ($p<0.05$) co-expression. We utilized a model based on the relative increase in TF expression in one sex vs. the other ($\mathrm{TF}_{\mathrm{sexeffect}}$) and the linear regression coefficient of the effect of TF expression level on target expression level within one sex (β) while accommodating the regression intercept (α) and residual (ε): $Target = \alpha + \beta \times \left(TF_{measured} \times TF_{sex\_effect}\right) + \varepsilon$. To determine which targets were most impacted by sb-TF expression, we compared the modulated log2(M/F) fold changes of each target to the baseline values.

Enrichment of sb-targets in transcription factor knockdown datasets

We utilized raw counts from *FOSL1* and scrambled control RNAi (72hr) in Th17 cells from GSE174809 (*31*) to identify enrichment of sb-targets. We then calculated differential expression in DESeq2 using default parameters in genes with at least 5 raw counts in at least 3 replicates. For *ZFX* CRISPRi knockdown, we utilized DESeq2 differential expression results previously published in(*30*).

Sex-biased eQTL discovery

Sex-biased eQTL were discovered through a two-step linear regression method executed in the enhanced version of FastQTL (*58*) utilized in the GTEx studies(*10, 59*). First, scHi-HOST Matrix counts in each sex separately and sex-combined were normalized through the median of ratios method in DESeq2. Genes with at least 6 unnormalized counts across 20% of individuals were retained for eQTL discovery. Gene counts were then further normalized using rank-based inverse normal transformation as previously described(*59*). To account for population structure, we utilized the top 5 genotypic principal components (PCs) from the autosomal genome as covariates. In addition, the top 15 probabilistic estimation of expression residuals (PEER) factors (*60*) were used to account for hidden confounding variables in the expression matrix. PEER factors were tested for collinearity with known covariates such as known genotypic PCs and sex (sex-combined analysis only). *Cis*-eQTL in each sex were discovered through the nominal mode of FastQTL using the covariates described above and a window of 1 Mb from the transcriptional start site (TSS) of each gene. The adaptive permutation mode (--permute 1000 10000) of FastQTL with the same covariates was used to account for multiple tests within each gene. Multiple test correction across the genome was then conducted using the Storey's Q (*61*) method on the top eQTL for each gene to determine a significance threshold ($q<0.05$) for each gene.

All unique significant *cis*-eQTL from each sex were then tested for a SNP × sex interaction through the interaction mode of FastQTL. Because the interaction mode does not support adaptive permutation, multiple test correction for tests within a gene was conducted using eigenMT(*62*), which estimates the number of unique tests per gene based on linkage blocks. Multiple test correction across all tested genes was then conducted using the Benjamini-Hochberg FDR method on the top sb-eQTL for each gene to determine a significance threshold (FDR<0.05) for each gene.

Enrichment of sb-eQTL in transcription factor motifs and binding sites

We utilized the SNP2TFBS database (*34*) to identify TFs impacted by sb-eQTL. The SNP2TFBS database estimates a SNP's effect on TF binding based on a position weight matrix (PWM) model from JASPAR database motifs(*63*). We identified enrichment of sb-eQTL in TF ChIP-Seq peaks through human blood cell ChIP-Seq datasets collected in ChIP-Atlas(*35*). We utilized 100 permutations to estimate background signatures and a peak threshold of $q<1\times10^{-5}$. Lastly, for targeted investigation of *ZFX* involvement with sb-eQTL, we downloaded ENCODE-processed ZFX ChIP-Seq peaks for K562 (ENCFF840NZE), HCT116 (ENCFF858YWR), C42B (ENCFF160AXQ), and MCF7 (ENCFF861DOL) cells(*64-66*). We then called consensus peaks as overlapping regions found in at least 2 cell types using the R package "DiffBind"(*67*).

Phenotype association of sex-biased eQTL

To investigate possible phenotypic associations of sb-eQTL, we utilized the command-line execution of the interactive Cross-Phenotype Analysis of GWAS database (iCPAGdb(*36*)). Briefly, iCPAGdb trims an input list of SNPs (here sb-eQTL) to leading variants based on linkage disequilibrium information from 1000 Genomes European populations(*68*). These variants were then queried against cataloged GWAS variants from the NHGRI-EBI GWAS catalog(*37*), urine (*39*) and plasma metabolites(*38*), and Hi-HOST infectious disease phenotypes(*40*). Both direct overlap and overlap with SNPs in linkage disequilibrium ($r^2>0.4$) with lead variants were considered. Enrichment of query SNPs within GWAS phenotypes was calculated based on observed vs. expected overlap and significance calculated with Fisher's Exact Test. Phenotypes with an enrichment FDR<0.05 were considered to share significant genetic architecture with sb-eQTL. Phenotypes were filtered to include only single effect phenotypes and behavioral GWAS were removed due to their likelihood to be biased by misreporting(*69*).

Colocalization of phenotype-associated sex-biased eQTL

We applied Giambartolomei et al.'s colocalization analysis (COLOC), using the R package "coloc" (*70*)to determine if GWAS and eQTL signals were due to the same causal SNP. COLOC uses a Bayesian framework to calculate the posterior probabilities that two traits are not associated in the locus of interest (PP0), only one trait is associated in the locus (PP1 and PP2), both traits are associated at the locus but with different, independent causal variants (PP3), or both traits are associated with a single causal variant in the locus (PP4). For GWAS summary statistics, we filtered SNPs within a 400 kilobase (kb) window from the SNP of interest. For eQTL datasets, we filtered all eQTL for a candidate eGene and then filtered SNPs within a 400 kb window from the SNP of interest. We ran the COLOC "coloc.abf" function using the default prior parameters, $p1=1\times10^{-4}$, $p2=1\times10^{-4}$, and $p12=1\times10^{-5}$ for all analyses. PP4 between 0.700 to 0.900 was interpreted as likely to share a single causal variant, while PP4>0.900 was interpreted as sharing a single causal variant. The PP4/PP3 measured the intensity of the colocalization signal with values >5.00 indicating further support for colocalization and >3.00 suggesting likely colocalization(*71, 72*). LocusZoom plots were created using "locuszoomr"(*73*).

References and Notes:

1. N. Lasrado *et al.*, Mechanisms of sex hormones in autoimmunity: focus on EAE. *Biology of Sex Differences* **11**, 50 (2020).
2. C. C. Whitacre, Sex differences in autoimmune disease. *Nature Immunology* **2**, 777-780 (2001).
3. L. Gay *et al.*, Sexual Dimorphism and Gender in Infectious Diseases. *Frontiers in Immunology* **12**, (2021).
4. A. J. Peired *et al.*, Sex and Gender Differences in Kidney Cancer: Clinical and Experimental Evidence. *Cancers* **13**, 4588 (2021).
5. E. M. Wu *et al.*, Gender differences in hepatocellular cancer: disparities in nonalcoholic fatty liver disease/steatohepatitis and liver transplantation. *Hepatoma Res* **4**, 66 (2018).
6. M. R. Schwartz, L. Luo, M. Berwick, Sex Differences in Melanoma. *Curr Epidemiol Rep* **6**, 112-118 (2019).
7. G. Edgren, L. Liang, H.-O. Adami, E. T. Chang, Enigmatic sex disparities in cancer incidence. *European Journal of Epidemiology* **27**, 187-196 (2012).
8. I. d. S. Silva, A. J. Swerdlow, Sex Differences in the Risks of Hormone-dependent Cancers. *American Journal of Epidemiology* **138**, 10-28 (1993).
9. J. J. Shen, Y. F. Wang, W. Yang, Sex-interacting mRNA-and miRNA-eQTLs and their implications in gene expression regulation and disease. *Frontiers in Genetics* **10**, 313 (2019).
10. M. Oliva *et al.*, The impact of sex on gene expression across human tissues. *Science* **369**, (2020).
11. L. Dumitrescu *et al.*, Sex differences in the genetic predictors of Alzheimer's pathology. *Brain: A Journal of Neurology* **142**, 2581-2589 (2019).
12. S. E. Graham *et al.*, Sex-specific and pleiotropic effects underlying kidney function identified from GWAS meta-analysis. *Nature Communications* **10**, 1847 (2019).
13. J. A. Hartiala *et al.*, Genome-wide association study and targeted metabolomics identifies sex-specific association of CPS1 with coronary artery disease. *Nature Communications* **7**, 10558 (2016).
14. J. Martin *et al.*, A Genetic Investigation of Sex Bias in the Prevalence of Attention-Deficit/Hyperactivity Disorder. *Biological Psychiatry* **83**, 1044-1053 (2018).
15. M. Gershoni, S. Pietrokovski, The landscape of sex-differential transcriptome and its consequent selection in human adults. *BMC Biology* **15**, 7 (2017).
16. B. T. Mayne *et al.*, Large Scale Gene Expression Meta-Analysis Reveals Tissue-Specific, Sex-Biased Gene Expression in Humans. *Frontiers in Genetics* **7**, (2016).
17. C. Yao *et al.*, Sex- and age-interacting eQTLs in human complex diseases. *Human molecular genetics* **23**, 1947-1956 (2014).
18. K. R. Kukurba *et al.*, Impact of the X Chromosome and sex on regulatory variation. *Genome Research*, (2016).
19. E. Porcu *et al.*, Limited evidence for blood eQTLs in human sexual dimorphism. *Genome Medicine* **14**, 1-13 (2022).
20. B. H. Schott *et al.*, Single-cell genome-wide association reveals that a nonsynonymous variant in ERAP1 confers increased susceptibility to influenza virus. *Cell Genomics* **2**, 100207 (2022).
21. D. J. Taylor *et al.*, Sources of gene expression variation in a globally diverse human cohort. *Nature* **632**, 122-130 (2024).

22. H. Tan *et al.*, Single-cell RNA-seq uncovers dynamic processes orchestrated by RNA-binding protein DDX43 in chromatin remodeling during spermiogenesis. *Nature Communications* **14**, 2499 (2023).
23. P. Yousefi *et al.*, Sex differences in DNA methylation assessed by 450 K BeadChip in newborns. *BMC Genomics* **16**, 911 (2015).
24. E. Gatev *et al.*, Autosomal sex-associated co-methylated regions predict biological sex from DNA methylation. *Nucleic Acids Research* **49**, 9097-9116 (2021).
25. S. Naqvi *et al.*, Conservation, acquisition, and functional impact of sex-biased gene expression in mammals. *Science* **365**, eaaw7317 (2019).
26. R. Magnusson, J. N. Tegner, M. Gustafsson, Deep neural network prediction of genome-wide transcriptome signatures - beyond the Black-box. *NPJ Syst Biol Appl* **8**, 9 (2022).
27. A. B. Keenan *et al.*, ChEA3: transcription factor enrichment analysis by orthogonal omics integration. *Nucleic Acids Res* **47**, W212-W224 (2019).
28. A. Shetty *et al.*, Interactome Networks of FOSL1 and FOSL2 in Human Th17 Cells. *ACS Omega* **6**, 24834-24847 (2021).
29. A. K. San Roman *et al.*, The human inactive X chromosome modulates expression of the active X chromosome. *Cell Genom* **3**, 100259 (2023).
30. A. K. San Roman *et al.*, The human Y and inactive X chromosomes similarly modulate autosomal gene expression. *Cell Genom* **4**, 100462 (2024).
31. A. Shetty *et al.*, A systematic comparison of FOSL1, FOSL2 and BATF-mediated transcriptional regulation during early human Th17 differentiation. *Nucleic Acids Res* **50**, 4938-4958 (2022).
32. K. R. Kukurba *et al.*, Impact of the X Chromosome and sex on regulatory variation. *Genome Res* **26**, 768-777 (2016).
33. K. Parker, A. M. Erzurumluoglu, S. Rodriguez, The Y Chromosome: A Complex Locus for Genetic Analyses of Complex Human Traits. *Genes (Basel)* **11**, (2020).
34. S. Kumar, G. Ambrosini, P. Bucher, SNP2TFBS - a database of regulatory SNPs affecting predicted transcription factor binding site affinity. *Nucleic Acids Res* **45**, D139-d144 (2017).
35. Z. Zou, T. Ohta, F. Miura, S. Oki, ChIP-Atlas 2021 update: a data-mining suite for exploring epigenomic landscapes by fully integrating ChIP-seq, ATAC-seq and Bisulfite-seq data. *Nucleic Acids Research* **50**, W175-W182 (2022).
36. L. Wang *et al.*, An atlas connecting shared genetic architecture of human diseases and molecular phenotypes provides insight into COVID-19 susceptibility. *Genome Med* **13**, 83 (2021).
37. A. Buniello *et al.*, The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005-D1012 (2019).
38. S. Y. Shin *et al.*, An atlas of genetic influences on human blood metabolites. *Nat Genet* **46**, 543-550 (2014).
39. J. Raffler *et al.*, Genome-Wide Association Study with Targeted and Non-targeted NMR Metabolomics Identifies 15 Novel Loci of Urinary Human Metabolic Individuality. *PLoS Genet* **11**, e1005487 (2015).
40. L. Wang *et al.*, An Atlas of Genetic Variation Linking Pathogen-Induced Cellular Traits to Human Disease. *Cell Host Microbe* **24**, 308-323 e306 (2018).
41. UK Biobank. *Neale lab.*
42. K. Bjornevik *et al.*, Longitudinal analysis reveals high prevalence of Epstein-Barr virus associated with multiple sclerosis. *Science* **375**, 296-301 (2022).

43. T. V. Lanz *et al.*, Clonally expanded B cells in multiple sclerosis bind EBV EBNA1 and GlialCAM. *Nature* **603**, 321-327 (2022).
44. C. International Multiple Sclerosis Genetics *et al.*, Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science* **365**, eaav7188 (2019).
45. X.-B. Mo *et al.*, Integrative analysis revealed potential causal genetic and epigenetic factors for multiple sclerosis. *Journal of neurology* **266**, 2699-2709 (2019).
46. G. Galarza-Munoz *et al.*, Human Epistatic Interaction Controls IL7R Splicing and Increases Multiple Sclerosis Risk. *Cell* **169**, 72-84 e13 (2017).
47. M. Hirano *et al.*, The RNA helicase DDX39B activates FOXP3 RNA splicing to control T regulatory cell fate. *Elife* **12**, (2023).
48. R. Zhan *et al.*, A DEAD-box RNA helicase Ddx54 protein in oligodendrocytes is indispensable for myelination in the central nervous system. *Journal of neuroscience research* **91**, 335-348 (2013).
49. Y. Huang *et al.*, Deciphering genetic causes for sex differences in human health through drug metabolism and transporter genes. *Nature Communications* **14**, 175 (2023).
50. S. J. Fleming *et al.*, Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. *Nature methods* **20**, 1323-1335 (2023).
51. H. M. Kang *et al.*, Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature biotechnology* **36**, 89-94 (2018).
52. Y. Zhang, G. Parmigiani, W. E. Johnson, ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genomics and Bioinformatics* **2**, lqaa078 (2020).
53. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Bio.* **15**, 550 (2014).
54. G. Yu, L. G. Wang, Y. Han, Q. Y. He, clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics* **16**, 284-287 (2012).
55. A. Lachmann *et al.*, Massive mining of publicly available RNA-seq data from human and mouse. *Nat Commun* **9**, 1366 (2018).
56. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-585 (2013).
57. C. A. Davis *et al.*, The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res* **46**, D794-d801 (2018).
58. H. Ongen, A. Buil, A. A. Brown, E. T. Dermitzakis, O. Delaneau, Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479-1485 (2016).
59. F. Aguet *et al.*, The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318-1330 (2020).
60. O. Stegle, L. Parts, M. Piipari, J. Winn, R. Durbin, Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature protocols* **7**, 500-507 (2012).
61. J. D. Storey, R. Tibshirani, Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100**, 9440-9445 (2003).
62. J. R. Davis *et al.*, An Efficient Multiple-Testing Adjustment for eQTL Studies that Accounts for Linkage Disequilibrium between Variants. *Am J Hum Genet* **98**, 216-224 (2016).
63. A. Mathelier *et al.*, JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **44**, D110-115 (2016).
64. B. C. Hitz *et al.*, The ENCODE Uniform Analysis Pipelines. *bioRxiv*, (2023).
65. Y. Luo *et al.*, New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res* **48**, D882-d889 (2020).

66. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
67. C. S. Ross-Innes *et al.*, Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, (2012).
68. A. Auton *et al.*, A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
69. A. Xue *et al.*, Genome-wide analyses of behavioural traits are subject to bias by misreports and longitudinal changes. *Nature Communications* **12**, 20211 (2021).
70. C. Giambartolomei *et al.*, Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* **10**, e1004383 (2014).
71. S. Pan *et al.*, COLOCdb: a comprehensive resource for multi-model colocalization of complex traits. *Nucleic Acids Res* **52**, D871-d881 (2024).
72. H. Guo *et al.*, Integration of disease association and eQTL data using a Bayesian colocalisation approach highlights six candidate causal genes in immune-mediated diseases. *Human Molecular Genetics* **24**, 3305-3313 (2015).
73. R. J. Pruim *et al.*, LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336-2337 (2010).

Acknowledgments: We thank the investigators and individuals from diverse populations genotyped as part of the 1000 Genomes Project who have made their LCLs available through the Coriell Institute. We thank the individuals who participated in the UK Biobank and International Multiple Sclerosis Genetics Consortium (IMSGC) studies as well as the investigators for helpful distribution of data. We thank members of the Ko lab for useful discussion. We acknowledge the assistance of the Duke Molecular Physiology Institute Molecular Genomics Core for the generation of data for the manuscript.

Funding:

National Institutes of Health grant R01AI170089 (AGJ, GC, LW, BS, DCK)

National Institutes of Health grant R01AI118903 (TD, LW, DCK)

National Institutes of Health grant R21AI153812 (AGJ, LW, BS, DCK)

Duke Precision Genomics Center student pilot grant (AGJ, BS, GC)

TriCEM Graduate Student Fellowship (AGJ, TD)

Whitehead Scholar Award (AKSR)

Author contributions:

Conceptualization: AGJ, DCK

Methodology: AGJ, GGC, TD, LW, BS, DCK

Investigation: AGJ, TD, DCK

Visualization: AGJ, AKSR, TD

Funding acquisition: AGJ, DCK

Supervision: DCK

Writing – original draft: AGJ, TD, DCK

Writing – review & editing: AGJ, AKSR, DCK

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: Data and scripts will be available at the Duke Research Repository (https://research.repository.duke.edu/) after publication. Individual LCL genotyping data is available through 1000 Genomes (FTP: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/). Single-cell RNA-seq data for scHi-HOST Neo and Morpheus are available from GEO (GSE205796). Single-cell RNA-seq data from scHi-HOST Trinity will be available from GEO as of the date of publication. All other data are available in the main text or the supplementary materials.

Supplementary Materials

Figs. S1 to S4

Tables S1 to S7