Title: Quantitative analysis of gene expression on the inactive human X chromosome

Authors: Adrianna K. San Roman[1], Alexander K. Godfrey[1,2], Helen Skaletsky[1,3], Daniel W. Bellott[1], Abigail F. Groff[1], Hannah L. Harris[1,2], Laura V. Blanton[1], Jennifer F. Hughes[1], Laura Brown[1,3], Sidaly Phou[1,3], Ashley Buscetta[4], Paul Kruszka[4]†, Nicole Banks[4,5], Amalia Dutra[6], Evgenia Pak[6], Patricia C. Lasutschinkow[7], Colleen Keen[7], Shanlee M. Davis[8], Nicole R. Tartaglia[8,9], Carole Samango-Sprouse[7,10], Maximilian Muenke[4]†, and David C. Page[1,2,3#]*

Affiliations:

[1]Whitehead Institute; Cambridge, MA 02142, USA.

[2]Department of Biology, Massachusetts Institute of Technology; Cambridge, MA 02139, USA.

[3]Howard Hughes Medical Institute, Whitehead Institute; Cambridge, MA 02142, USA.

[4]Medical Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda; MD 20892, USA.

[5]Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health; Bethesda, MD 20892 USA.

[6]Cytogenetics and Microscopy Core, National Human Genome Research Institute, National Institutes of Health; Bethesda, MD 20892 USA.

[7]Focus Foundation; Davidsonville, MD 21035, USA.

[8]Department of Pediatrics, University of Colorado School of Medicine; Aurora, CO 80045, USA.

[9]Developmental Pediatrics, eXtraOrdinarY Kids Program, Children's Hospital Colorado, Aurora, CO 80011, USA

[10]Department of Pediatrics, George Washington University; Washington, DC 20052, USA; Department of Human and Molecular Genetics, Florida International University, Miami, FL 33199, USA.

Author list footnotes:

[#]Lead author

†Present addresses: GeneDx, Gaithersburg, MD 20877, USA (P.K.); American College of Medical Genetics and Genomics, Bethesda, MD 20814, USA (M.M.).

Contact info:

*Correspondence dcpage@wi.mit.edu

Summary: [150 word limit, currently 150]

The human X chromosome exists in active (Xa) and inactive (Xi) states. To quantify Xi gene expression and explore its implications, we conducted RNA sequencing in two cell types from individuals with one Xa and zero to three Xi's. Linear models fit the data well, indicating equal contributions from each Xi and quantifying significant Xi-driven expression changes for 124 of 406 expressed X-chromosome genes. As revealed by allele-specific expression analyses, these Xi-driven changes reflect both i) Xi transcription and ii) Xi modulation – both positive and negative – of Xa transcript levels. Incorporating population and evolutionary measures of constraint on expression levels, we identified 13 X-chromosome genes most likely to drive the phenotypic consequences of natural variation in Xi copy number. These analytics and the underlying datasets offer a foundation for understanding Turner, Klinefelter, and other sex chromosome aneuploidy syndromes, and sex-chromosome effects on health and disease in euploid males and females.

Keywords: [up to 10]
Sex chromosomes, X chromosome inactivation, gene expression, sex differences, aneuploidy, Turner syndrome, Klinefelter syndrome, dosage sensitivity.

Introduction:

The X chromosome of eutherian mammals exists in two distinct epigenetic states that are referred to as "active" (Xa) and "inactive" (Xi) (Ohno and Hauschka, 1960, Lyon, 1961, Lyon, 1962). All human somatic cells possess one X chromosome in the active state (Xa), while all other copies of Chr X, if any, are transcriptionally repressed through a mechanism known as X chromosome inactivation (XCI). Despite the name, multiple lines of evidence indicate that Xi is functionally active, making critical contributions to human fitness and viability. For example, 99% of fetuses with only one X chromosome (45,X) abort spontaneously, suggesting that viability hinges on gene expression from a second sex chromosome – either Xi or Y (Hook and Warburton, 1983, Hook and Warburton, 2014). The rare survivors likely have a mixture of 45,X cells and cells with a second sex chromosome, and they display a constellation of anatomic features known as Turner syndrome (Turner, 1938, Ford et al., 1959).

Investigations undertaken across the past four decades have revealed that as many as a quarter of X-linked genes are expressed from Xi in humans; such genes are often said to "escape" X inactivation (Balaton et al., 2015). Early studies assayed expression from Xi directly, using human-rodent hybrid cell lines to genetically isolate Xi [for example, (Mohandas et al., 1980, Brown et al., 1997, Carrel et al., 1999)]. Subsequently, allele-specific methods in human cells were developed to distinguish transcripts deriving from Xa and Xi – either in cell lines with skewed XCI or in single cells (Carrel and Willard, 2005, Cotton et al., 2013, Tukiainen et al., 2017, Garieri et al., 2018, Wainer Katsir and Linial, 2019, Sauteraud et al., 2021). While conceptually superior to hybrid cell lines, allele-specific methods yield sparse data because they require the fortuitous presence of heterozygous SNPs to differentiate between alleles. Circumventing this issue, other studies approximated the contributions of Xi to X-linked gene

expression by comparing samples with varying Xi copy number, either between 46,XX and 46,XY or between sex chromosome aneuploid and euploid individuals (Sudbrak et al., 2001, Craig et al., 2004, Talebizadeh et al., 2006, Johnston et al., 2008, Trolle et al., 2016, Tukiainen et al., 2017, Raznahan et al., 2018, Zhang et al., 2020, Nielsen et al., 2020). These studies employed pairwise comparisons between individuals with different sex Chr constitutions, making it difficult to separate analytically the effect of Xi copy number from the possibly confounding effect of correlated factors like Chr Y copy number, hormonal differences, or tissue composition. Given these limitations, we believed that a deeper understanding of Xi gene expression was possible.

Here, we used a series of quantitative approaches to investigate gene expression on Xi and to assess any secondary effects on gene expression from Xa. Inspired by previous studies, we took advantage of the natural occurrence in human populations of diverse sex chromosome aneuploidies and performed RNA sequencing (RNA-seq) on two cell types (lymphoblastoid cell lines and primary skin fibroblasts) from 176 individuals spanning 11 different sex chromosome constitutions – from 45,X (Turner Syndrome) to 49,XXXXY. Unlike previous studies, we analyzed the resulting data from all sex Chr constitutions together – using linear regression models to identify significant changes in Chr X gene expression in identically cultured cells with zero, one, two, or three copies of Xi. We found that 30% of Chr X genes have expression changes *directly* driven by Xi; we quantified these changes using a metric called ΔE_X. By combining ΔE_X findings with an allele-specific analysis performed in the same cell lines, we discovered that Xi can positively or negatively modulate steady-state levels of transcripts of some genes on Xa. Finally, to understand the phenotypic impacts of natural variation in Xi copy

number, we intersected ΔE_X with external metrics of gene constraint and identified 13 genes most likely to be responsible for those phenotypic consequences.

Results:

A cell and gene-expression resource across sex chromosome constitutions

To conduct a robust, quantitative analysis of Xi's impacts on X-linked gene expression, we recruited individuals with a wide range of sex chromosome constitutions to provide blood samples and/or skin biopsies (**Fig. 1A**). We generated or received Epstein Barr Virus-transformed B cell lines (lymphoblastoid cell lines, LCLs) and/or primary dermal fibroblast cultures from 176 individuals with one to four X Chrs and zero to four Y Chrs. After culturing cells under identical conditions, we profiled gene expression by RNA sequencing (RNA-seq) in LCLs from 106 individuals and fibroblast cultures from 99 individuals (some individuals contributed both blood and skin samples; **Table 1 and Table S1**). To enable analysis at both the gene and transcript isoform levels, we generated 100-base-pair paired-end RNA-seq reads to a median depth of 74 million reads per sample. By our estimates, more individuals with sex chromosome aneuploidy would not substantially increase the number of differentially expressed genes detected in our analyses (**Fig. S1, Methods**).

Table 1. Samples included in sex chromosome aneuploidy analysis

Karyotype	# LCLs	# Fibroblast cultures
45,X	17	23
46,XX	22	20
46,XY	17	14
47,XXX	7	4

47,XXY	11	30
47,XYY	10	5
48,XXXX	1	0
48,XXXY	4	1
48,XXYY	3	0
49,XXXXY	12	1
49,XYYYY	2	1
Total:	*106*	*99*

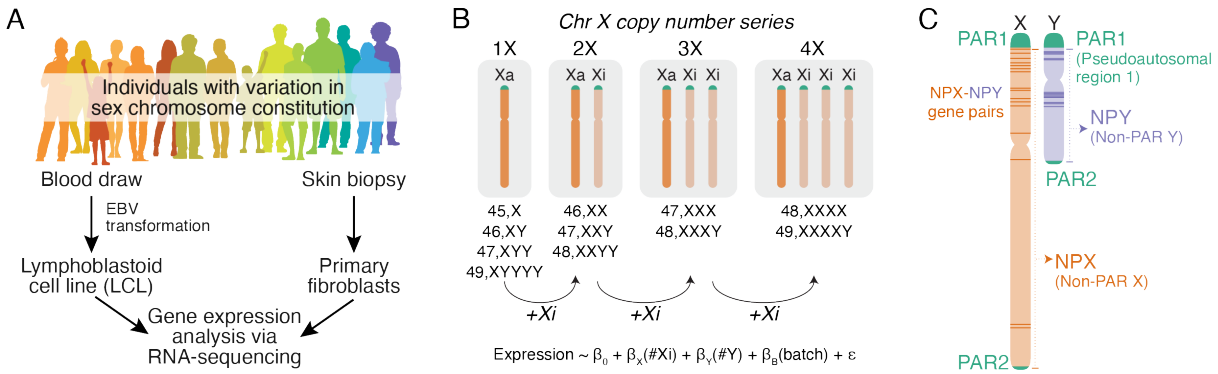

Figure 1. Gene expression analysis of cells from across the spectrum of sex chromosome constitution. (A) Collection and processing of samples from human subjects with variation in sex chromosome constitution. **(B)** Strategy for analyzing RNA-seq data from individuals with one to four copies of Chr X (zero to three copies of Xi). **(C)** Schematic of the sex chromosomes.

A metric for the impact of Xi on gene expression

To leverage the full power of our datasets, we assembled all of the RNA-seq data for each cell type into a single analysis. We included genes with a median expression in euploid samples of at least 1 transcript per million (TPM), resulting in 403 Chr X genes analyzed in LCLs and/or fibroblasts. These genes reside within structurally and evolutionarily distinct regions of Chr X

(**Fig. 1C**): two pseudoautosomal regions (PAR1 and PAR2), which are identical in sequence between Chr X and Y, and the non-pseudoautosomal region of the X (NPX), which has diverged in structure and gene content from the non-pseudoautosomal region of the Y [NPY, (Skaletsky et al., 2003, Ross et al., 2005)]. Despite this divergence, 17 homologous "NPX-NPY pair genes" with varying degrees of X-Y similarity in sequence and function remain (Skaletsky et al., 2003, Bellott et al., 2014).

We anticipated that each additional copy of Xi would incrementally increase expression of some or all genes on Chr X and therefore modeled gene expression as a linear function of Xi copy number, controlling for variation in Chr Y copy number and batch (**Fig. 1B**; **Methods**). To assess whether expression of Chr X genes changes linearly per Xi, we fit non-linear least square regression models to the expression data using power functions (**Methods**). We found that most NPX and PAR1 genes previously annotated as escaping XCI were best fit by linear models in which their expression increases by a fixed amount per Xi, while most genes previously annotated as subject to XCI were best fit by models that indicate no change in expression per Xi (**Fig. S2**).

Linear models allowed us to identify genes whose expression changed significantly with additional copies of Xi and to quantify the *absolute* changes in expression (i.e., the changes in read counts per additional Xi). To compare genes expressed at different levels, we also quantified the *relative* changes in expression per additional Xi. Specifically, we divided the change in expression per additional Xi (slope of regression, $ß_X$) by the expression from the single active copy of Chr X (average intercept across batches, $ß_0$) – a metric we refer to as ΔE_X (**Fig. 2A**). $\Delta E_X=0$ indicates that adding one or more copies of Xi does not affect the level of expression (e.g., *PRPS2*, **Fig. 2B**); $\Delta E_X>0$ indicates that expression increases under these

circumstances (e.g., *KDM5C*, **Fig. 2C**), with ΔE_X=1 suggesting that expression from each additional Xi is equivalent to expression from Xa; and ΔE_X<0 indicates that expression decreases (e.g., *F8*, **Fig. 2D**). *XIST,* the long non-coding RNA that acts in *cis* to transcriptionally repress X chromosomes from which it is expressed (Brown et al., 1991, Penny et al., 1996), was the only gene without detectable expression in cells with one copy of Chr X (Xa) that was expressed robustly in cells with a second, inactive copy (Xi). *XIST* expression increased linearly with each additional copy of Xi (**Fig. 2E**).

ΔE_X values vary widely among Chr X genes but not between cell types

Analyzing ΔE_X values across Chr X revealed that Xi contributions to expression vary widely among Chr X genes. Expression levels for most Chr X genes did not change significantly with the addition of one or more copies of Xi ($\Delta E_X \approx 0$) consistent with expression from a cell's first X chromosome (Xa) and silencing on all others (Xi). 67.2% of Chr X genes in LCLs and 76.9% in fibroblasts fit this pattern (**Fig. 2F-G**; full results in **Table S2**). The remaining 32.8% of Chr X genes in LCLs and 23.1% in fibroblasts had significantly negative or positive ΔE_X values, ranging from -0.39 to 1.48. We observed distinct expression signatures in different regions of Chr X: PAR1 genes had ΔE_X values near 1, PAR2 genes had ΔE_X values near zero, and NPX genes' ΔE_X values varied widely (**Table S2**). The difference between PAR1 and PAR2 genes likely reflects their distinct evolutionary origins: PAR1 was preserved through sex chromosome evolution and retains autosome-like features, while PAR2 was added later (Freije et al., 1992). Among NPX genes, those that retain an NPY homolog had higher median ΔE_X values than those with no NPY homolog (**Fig. 2F-G**).

We next asked whether these Chr X expression dynamics were consistent across the two cell types. For genes expressed in both LCLs and fibroblasts, the ΔE_X values were highly concordant (**Fig. 2H**). This is consistent with previous analyses of differential expression between 46,XX versus 46,XY tissues, which found correlated expression changes for Chr X genes over a wide array of tissues across the body (Tukiainen et al., 2017). Furthermore, our data suggests that this consistency derives from the cell-autonomous contributions of a copy of Xi, rather than systemic effects of hormones or other environmental factors.

We validated these results in two ways. To control for the effects of gonadal sex or Y chromosome copy number on our results, we re-analyzed the data from samples with zero (45,X; 46,XX; 47,XXX; 48,XXXX) or one copy of Chr Y (46,XY; 47,XXY; 48,XXXY; 49,XXXXY), modeling expression as a function of Xi copy number and batch. We found that ΔE_X values were unaffected by the presence or absence of a single Y chromosome (**Fig. S3**). To assess the reproducibility of our results, we turned to an independent dataset using microarrays to assay gene expression in LCLs spanning sex chromosome constitutions (Raznahan et al., 2018). Reanalyzing this dataset using linear models, we found that the resulting microarray ΔE_X values correlated well with the ΔE_X values calculated from our RNA-seq data (**Fig. S4; Methods**).

In all, these results provide quantitative insight into the transcriptional response to Xi copy number. For nearly all Chr X genes, expression contributed by each copy of Xi is less than that contributed by Xa ($\Delta E_X<1$). Only two NPX genes – *XIST* and *PUDP* – and a subset of PAR1 genes – *DHSRX, PLCXD1,* and *PPP2R3B* – showed near-equal or greater expression with each additional Xi compared to their expression from Xa ($\Delta E_X\approx1$) in both LCLs and fibroblasts. The results for *XIST* are unsurprising given its role in inactivating each copy of Xi.

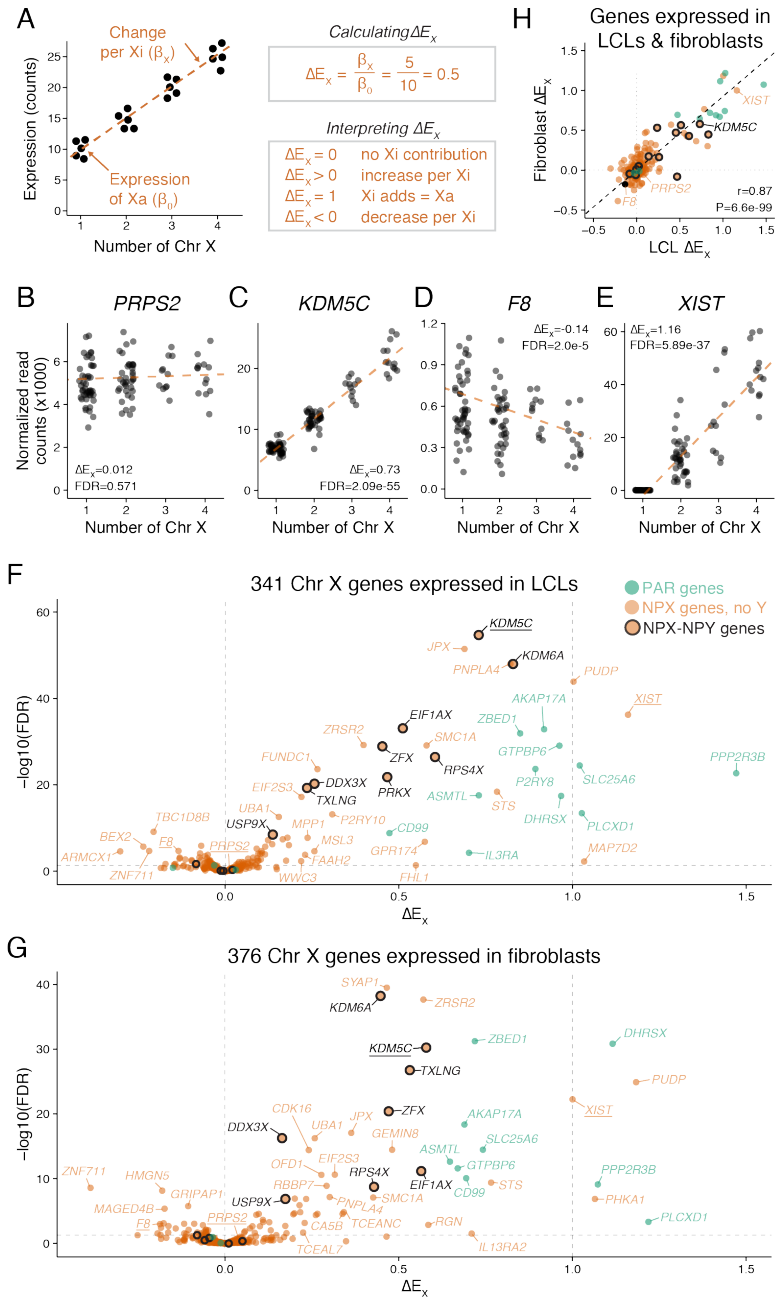

Figure 2. Quantitative assessment of Xi contributions to X chromosome gene expression. **(A)** Schematic scatterplot, linear regression line, and ΔE_X calculation for a hypothetical Chr X gene. Each point represents the expression level for an individual sample with the indicated number of copies of Chr X. **(B-E)** Actual scatterplots and regression lines for selected Chr X genes in LCLs, representing a range of ΔE_X values. Adjusted p-values (FDR) <0.05 indicate that

ΔE_X values are significantly different from 0. **(F-G)** Scatterplots of ΔE_X versus significance for all Chr X genes expressed in LCLs **(F)** and fibroblasts **(G)** illustrate variation in Xi contributions to Chr X gene expression; genes depicted in B-E are underlined. **(H)** Scatterplot comparing ΔE_X in LCLs and fibroblasts for Chr X genes expressed in both cell types. Colors as in F-G. Deming regression line and Pearson correlation are indicated.

Divergent ΔE_X values for transcript isoforms of an X-linked gene

Among genes expressed in LCLs and fibroblasts, about half had multiple expressed transcript isoforms (**Fig. S5A, Methods**), raising the question of whether different transcript isoforms for a given gene respond similarly to the addition of one or more copies of Xi. Using linear regression models, we calculated ΔE_X values for each transcript isoform. For most genes, ΔE_X values for different transcript isoforms were highly consistent (e.g., *CA5B*, **Fig. S5B**), indicating that the change in expression per Xi is similar for each gene's transcript isoforms (**Fig. S5C, Table S3**). However, we identified a set of genes with multiple transcript isoforms that had inconsistent ΔE_X values – that is, at least one transcript's ΔE_X differed significantly from zero (FDR<0.05) while another transcript's ΔE_X did not (19.5% and 13.1% of genes with multiple transcript isoforms in LCLs and fibroblasts, respectively).

UBA1 was a particularly interesting case because its two transcript isoforms, which differ in their ΔE_X values, also differ in their first exons and transcription start sites (TSS). The transcript from the 5' TSS had ΔE_X values near zero in LCLs and fibroblasts, while the transcript from the 3' TSS had significantly positive ΔE_X values (**Fig. S5D**). Three previous studies corroborate these results: 1) in human-mouse hybrid cell lines, the 5' *UBA1* TSS was only active in lines containing Xa, while the 3' TSS was active in lines containing either Xa or Xi (Goto and

Kimura, 2009); 2) a comparison of 46,XX and 46,XY expression found that transcripts from the 3' TSS were more abundant in 46,XX cells, while there was no difference for the 5' TSS (Chen et al., 2016); and 3) DNA methylation studies showed hypermethylation of the 5' but not the 3' TSS in XX female cells from human, chimp, and horse (Balaton et al., 2021). Moreover, we found that ΔE_X values correlated with those of the flanking genes: the 5' flanking gene, *RBM10*, had a ΔE_X near zero, and the 3' flanking gene, *CDK16*, had a positive ΔE_X. Analysis of ENCODE data reveals a strong CTCF binding site between the two *UBA1* TSSs in LCLs and fibroblasts (**Fig. S5E**), suggesting that the two TSSs reside in distinct topologically associated domains – one of which is more fully heterochromatinized than the other on Xi.

Supernumerary copies of chromosomes Y and 21 show little attenuation of gene expression

To determine whether the attenuated expression observed with extra copies of Chr X also occurs with additional copies of other chromosomes, we analyzed cells from individuals with additional copies of Chr Y and trisomy 21, the most common autosomal aneuploidy and the cause of Down syndrome (Megarbane et al., 2009). For Chr Y we used the same linear model as for Chr X: modeling expression as a function of Xi copy number, Chr Y copy number, and batch (**Fig. 3A, Methods**). We calculated ΔE_Y values separately for NPY genes and PAR genes, because NPY genes are not expressed in samples with zero Y Chrs, while PAR genes are expressed from Chr X. For NPY genes, we analyzed samples with one or more Y Chrs to observe expression differences, if any, between the first and additional Y Chrs. All NPY genes increased significantly, with ΔE_Y values close to 1, consistent with near-equal expression from each copy of Chr Y (e.g., *KDM5D*; **Fig. 3B-D,** full results in **Table S4**).

We next analyzed PAR genes utilizing samples with zero to four Y Chrs. As with Chr X, PAR1 gene expression increased with additional copies of Chr Y and had ΔE_Y values close to 1, whereas genes in PAR2 had ΔE_Y values near 0 (**Fig. 3C,D**). This implies, first, that PAR1 genes are expressed on each additional copy of Chr X or Y, while PAR2 genes are only expressed on the first copy of Chr X (Xa); and second, that the expression of PAR1 genes from each additional Chr X or Y is roughly equal to the expression from the first. We observed modestly greater increases in PAR1 gene expression with additional Y chromosomes than with additional X chromosomes, especially in LCLs (**Fig. 3H**). This Y-vs-X effect was most pronounced for *CD99*, located near the PAR1-NPX/Y boundary (**Tables S2 and S4**), consistent with suggestions that PAR1 gene expression is modestly attenuated by spreading of heterochromatinization on Xi (Tukiainen et al., 2017).

Finally, we examined Chr 21 gene expression as a function of Chr 21 copy number (**Table S5**). Three-quarters of expressed Chr 21 genes significantly (FDR<0.05) increased in expression with an additional copy of Chr 21 (e.g., *CCT8*; **Fig. 3F**), and no genes significantly decreased in expression (**Fig. 3G**). These results are consistent with independent studies of Chr 21 gene expression (Sullivan et al., 2016).

In sum, our analysis reveals that most genes on Chr Y and Chr 21 are expressed similarly on each copy of their respective chromosomes. The median ΔE values for Chr Y (including PAR1 ΔE_Y and NPY) and Chr 21 genes range from 0.76-1. In contrast, genes are not equally expressed from Xi and Xa. NPX gene expression is strongly reduced on Xi: NPX genes without NPY homologs had median $\Delta E_X \approx 0$, and NPX genes with NPY homologs had modestly higher median ΔE_X values (LCLs: 0.17, fibroblasts: 0.26; **Fig. 3H**). Even PAR1 genes which, as a group, had the highest median ΔE_X values, showed attenuation of expression on Xi compared to

Chr Y, especially in LCLs. These differences highlight the absence of a chromosome-wide mechanism attenuating (or otherwise altering) the expression of genes on supernumerary copies of Chr Y and Chr 21, in contrast to Chr X.

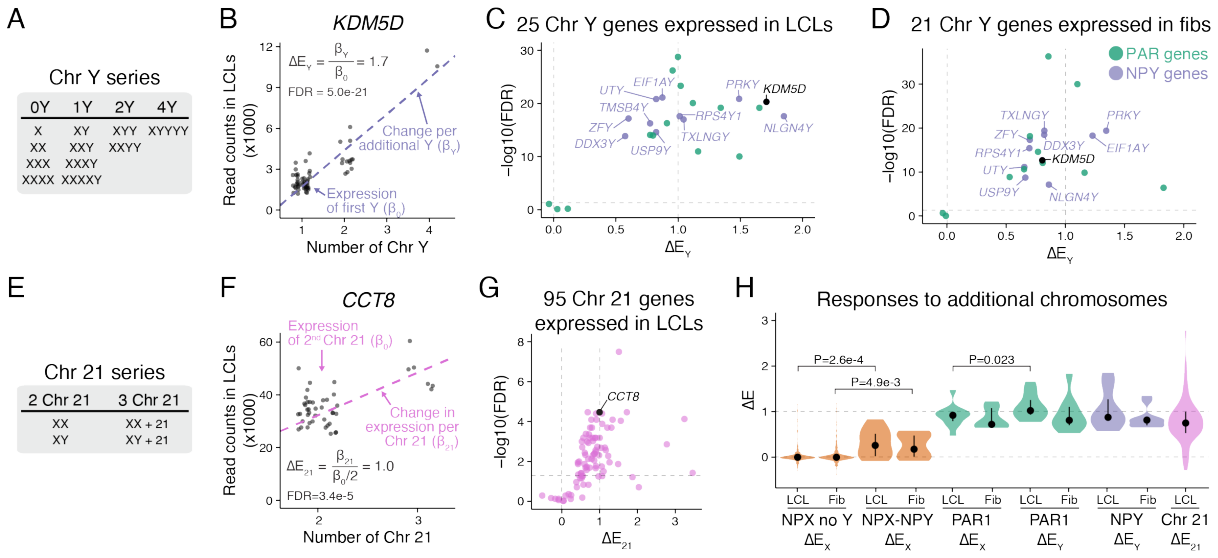

Figure 3. Contributions of Chr Y or 21 copy number to gene expression. (A) Chr Y copy number series with zero to four copies. **(B)** Scatterplot and regression line for NPY gene *KDM5D* in LCLs showing ΔE_Y calculation. **(C-D)** Scatterplot of ΔE_Y values versus significance for all Chr Y genes expressed in LCLs **(C)** or fibroblasts **(D)**; *KDM5D,* depicted in B, is shown in black. **(E)** Chr 21 copy number series with two to three copies. **(F)** Scatterplot, regression line, and ΔE_{21} calculation for Chr 21 gene *CCT8* in LCLs. **(G)** Scatterplot of ΔE_{21} versus significance for all Chr 21 genes expressed in LCLs. *CCT8*, depicted in F, is shown in black. **(H)** Violin plots with median and interquartile range for ΔE values of genes in NPX (without or with a NPY homolog), NPY, PAR, and Chr 21. P-values, Wilcoxon rank sum test.

Xi modulation of Xa transcript levels revealed by divergence of ΔE_X and allelic ratio

ΔE_X conveys the change in gene expression due to an additional Xi regardless of the mechanism(s) responsible for this change. We hypothesized that an X-linked gene's ΔE_X value could reflect the combined effects of two mechanisms: 1) transcription of Xi alleles and 2) modulation of steady-state transcript levels from Xa and/or Xi.

To dissociate these two mechanisms, we sought to directly measure expression from Xi and Xa through an allele-specific analysis of our RNA-seq data. If a gene's Xi and Xa expression are fully independent of each other, and therefore additive, we would expect the gene's ΔE_X to be equal to the expression ratio of Xi to Xa alleles (the allelic ratio, AR). Conversely, if Xi modulates steady-state transcript levels from Xa, and therefore is not strictly additive, we would expect ΔE_X and AR to differ.

To distinguish between Chr X alleles in samples with two X chromosomes, we first identified heterozygous single nucleotide polymorphisms (SNPs) in genes expressed in LCLs or fibroblasts (**Methods**). We then identified samples in our dataset with skewed XCI (22 LCL and 10 fibroblast samples; see **note S1**), and used these samples to compute the average AR for each gene. We required heterozygous SNPs (needed to calculate the AR) in at least three samples to ensure the results were reliable, resulting in AR values for 161 genes in LCLs and 119 in fibroblasts (**Table S6**). In LCLs and fibroblasts, respectively, 21.7% and 17.6% of these genes had AR values significantly greater than 0, indicating that they are expressed from Xi.

We next compared AR with ΔE_X values. Most X-linked genes, e.g., *EIF2S3*, had AR values that approximate their ΔE_X values, and AR and ΔE_X were highly correlated among informative genes in both LCLs and fibroblasts (**Fig. 4**). For these genes, the ΔE_X value directly reflects the level of transcription from Xi.

However, for 11% of informative genes in LCLs or fibroblasts, AR and ΔE_X were significantly different, indicating that Xi either negatively or positively modulated steady-state transcript levels (**Fig. 4**). Some genes, like *MPP1,* were not expressed from Xi (AR≈0), but nonetheless had ΔE_X values significantly different from zero – 0.24 in LCLs and 0.21 in fibroblasts, indicating that levels of Xa-derived transcripts are positively regulated by Xi. Other genes, like *DDX3X* and *PUDP*, had significant expression from Xi (AR>0, FDR<0.05) *and* evidence of modulation of steady-state expression levels. For example, *DDX3X* had an AR (LCLs: 0.55, fibroblasts: 0.48) that is significantly higher than its ΔE_X value (LCLs: 0.25, fibroblasts: 0.16) in both LCLs and fibroblasts, indicating both that *DDX3X* is expressed on Xi *and* that its steady-state transcript levels are negatively regulated by Xi. Conversely, *PUDP* had an AR (LCLs: 0.72, fibroblasts: 0.73) that is significantly lower than its ΔE_X value (LCLs: 1.0, fibroblasts: 1.2), indicating both that *PUDP* is expressed on Xi *and* that the gene's steady-state transcript levels are positively regulated by Xi.

These combined analyses reveal features of X-linked gene regulation that have not been demonstrated previously. We found that Xi alters the expression levels of X-linked genes through at least two mechanisms. Operating on a gene-by-gene basis, these mechanisms include transcription from a gene's allele on Xi and modulation of steady-state transcript levels. These mechanisms can operate independently of each other, or in conjunction. Finally, the observed modulation of steady-state transcript levels by Xi implies that Xi regulates the expression of genes on Xa in trans – by attenuating or elevating their expression through transcriptional or post-transcriptional mechanisms.

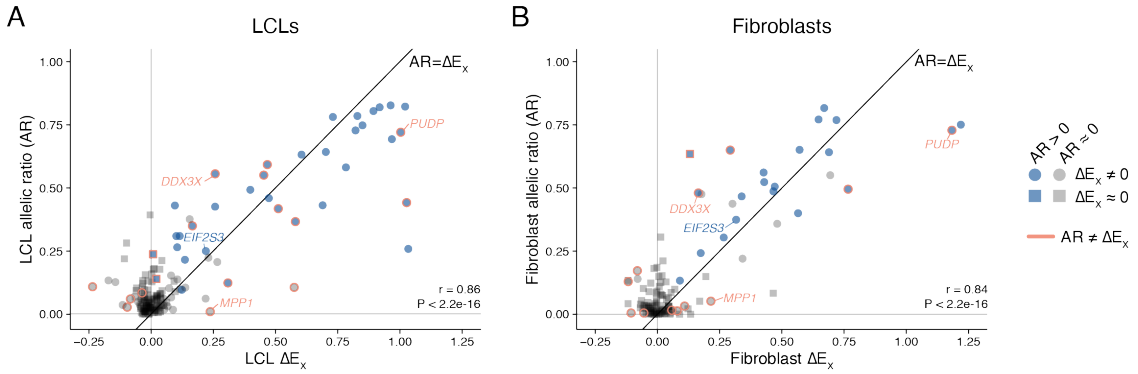

Figure 4. Bivariate analysis of allelic ratio and ΔE_X reveals that Xi modulates steady-state transcript levels. Scatterplots of allelic ratio (AR) and ΔE_X values for genes with heterozygous SNPs in at least three samples with skewed XCI in **(A)** LCLs – 161 genes – and **(B)** fibroblasts – 119 genes. Circles, genes with $\Delta E_X \neq 0$ (FDR<0.05); squares, genes with $\Delta E_X \approx 0$. Blue, genes with AR > 0 (FDR<0.05); grey, AR ≈ 0. Orange outline, genes with $\Delta E_X \neq$ AR (FDR<0.05). Black diagonal line, AR = ΔE_X. Pearson correlation coefficients (r) and P-values are indicated.

High tolerance of under- and over-expression for most PAR1 genes

We hypothesized that Turner, Klinefelter, and other phenotypes associated with Xi copy number are driven by a subset of the Chr X genes identified here as having significantly positive or negative ΔE_X values. We reasoned that these phenotypes would be driven by copy-number-sensitive genes on Chr X, and that these copy-number-sensitive genes should be more expression-constrained by natural selection than other genes on Chr X. But how can one pick out or enrich for this copy-number-sensitive subset from among the many Chr X genes with significant ΔE_X values? We reasoned that the expression levels of the copy-number-sensitive genes should be more constrained by natural selection than other genes on Chr X.

To quantitatively examine the constraints that natural selection imposes on each gene's expression level, we turned to metrics derived from population and evolutionary genetic studies.

We assessed gene-by-gene tolerance of under-expression using three independent metrics: 1) LOEUF, the ratio of observed to expected loss-of-function (LoF) variants in human populations (Karczewski et al., 2020)), 2) RVIS, the residual variation intolerance score (Petrovski et al., 2013), and 3) pHI, the probability of haploinsufficiency (Huang et al., 2010). Both LOEUF and RVIS use large-scale human genomic sequencing data to evaluate selection against accumulation of LoF variants, while pHI is calculated using a model that incorporates various evolutionary and functional metrics. LoF variants should be culled from the population in genes whose under-expression is deleterious (low LOEUF and RVIS, high pHI), while they would likely accumulate in genes whose under-expression has little or no effect on fitness (high LOEUF and RVIS, low pHI). To assess gene-by-gene tolerance of over-expression, we examined the conservation of targeting by microRNAs (miRNAs; P_{CT} score (Friedman et al., 2009)), which repress expression by binding to a gene's 3' untranslated region (Bartel, 2009). Genes whose over-expression is deleterious would be more likely to maintain their miRNA binding sites over evolutionary time via natural selection, while genes whose over-expression has little or no effect on fitness would show less conservation of these sites. To integrate these metrics, we calculated each gene's percentile rank for each metric, from most (high percentile) to least (low percentile) constrained. We calculated percentiles separately for autosomal (including PAR1) and NPX genes, and then, for each gene, averaged percentile rankings across the four metrics.

We began by investigating expression constraints for PAR1 genes, which have high ΔE_X values, and thus might be considered good candidates for contributing to phenotypes associated with Xi copy number. Compared to autosomal genes, PAR1 genes are less constrained on average, with most ranking in the least-constrained quartile (**Fig. 5A, Table S7**). This reveals that humans can tolerate both increases and decreases in the expression levels of PAR1 genes

without significant impacts on fitness or viability. Indeed, homozygous LoF mutations have been reported for 3 of 15 PAR1 genes, suggesting that they are dispensable (Karczewski et al., 2020). Only two PAR1 genes, *SHOX* and *SLC25A6,* rank in the more constrained half of the comparison group (**Table 2**). *SHOX* copy number has been linked to variation in height in individuals with sex chromosome anomalies (Ogata and Matsuo, 1993, Rao et al., 1997, Clement-Jones et al., 2000, Ottesen et al., 2010, Fukami et al., 2016). *SLC25A6* has not previously been linked with sex chromosome-related phenotypes but should be considered a candidate in future investigations. Apart from these two genes, the high tolerance of under- and over-expression for most PAR1 genes argues that they do not play prominent roles in phenotypes associated with Chr X (or X+Y) copy number.

Combining ΔE_X and over/under-expression metrics identifies NPX genes with NPY homologs as likely drivers of Xi-dependent phenotypes

We next examined expression constraints for NPX genes, comparing them with their ΔE_X values. NPX genes varied widely in expression constraint metrics, regardless of whether they had significantly non-zero ΔE_X values (**Fig. 5B-C, Table S8**). Thus, ΔE_X alone is a poor predictor of copy-number sensitivity among NPX genes.

To identify the NPX genes most likely to drive copy-number-dependent phenotypes, we considered those with both 1) expression constraints ranking in the top quartile and 2) $|\Delta E_X \geq 0.1|$ (FDR<0.05) in LCLs or fibroblasts. We found 13 genes that fit these criteria (**Table 2**). Five of the 13 NPX genes in this group had NPY homologs, a significant enrichment (LCLs: P=5.8e-6, fibroblasts: P=6.8e-6, hypergeometric test); all five genes had ΔE_X values > 0.1 in both cell types. Of the genes without NPY homologs, only two had ΔE_X values significantly greater than

zero in both cell types: *SMC1A* and *CDK16.* The remaining six genes were expressed, or had ΔE_X values significantly different from zero, in only one of the two cell types we analyzed, and they tended to have lower absolute ΔE_X values. All together, these 13 genes represent good candidates for driving Xi-dependent phenotypes in individuals with sex chromosome aneuploidies as well as in euploid individuals.

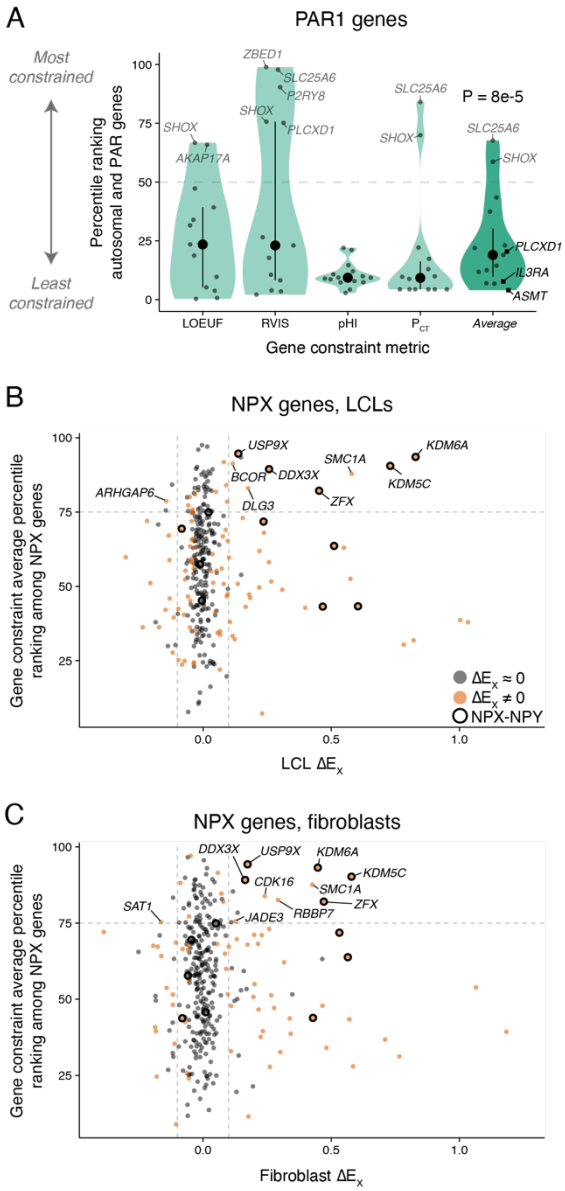

Figure 5. Evolutionary and population genetic measures of tolerance of under-expression or over-expression. (A) Violin plots, with median and interquartile ranges, of PAR1 gene

percentile rankings among all autosomal and PAR1 genes, where higher percentile indicates higher constraints on expression levels. P-value Wilcoxon rank sum test of PAR1 versus autosomal genes. The right-most violin is the average percentile ranking per gene across all constraint metrics. Genes above the median for all autosomal genes are labelled. Squares, genes with observed homozygous loss-of-function variants in gnomAD – these are inferred to be unconstrained. **(B-C)** Scatterplots of the expression constraint average percentile ranking among NPX genes versus ΔE_X values in LCLs (B) or fibroblasts (C). Genes annotated as "ampliconic" and genes with values for <2 of the gene constraint metrics were excluded. Orange points, NPX genes with ΔE_X significantly different from zero (FDR < 0.05); grey points, $\Delta E_X \approx 0$; black dot, NPX genes with NPY homologs. Dashed lines indicate absolute ΔE_X threshold of 0.1 and constraint percentile rank threshold of 75% for genes to be considered likely contributors to phenotypes driven by sex Chr constitution; these genes are labelled.

Table 2. Genes that may mediate the phenotypic impacts of Xi copy number

				LCL		Fibroblast		AR			
Region	**Gene symbol**	**Gene name**	**NPY gene symbol**	**ΔE_X**	**FDR**	**ΔE_X**	**FDR**	**LCL**	**Fibroblast**	**Published[a]**	**Gene constraint (average % ranking)[b]**
NPX	*KDM6A*	Lysine demethylase 6A	*UTY*	0.83	1.1e-48	0.45	6.0e-39	0.78	--	0.43	93.30
	KDM5C	Lysine demethylase 5C	*KDM5D*	0.73	2.1e-55	0.58	5.8e-31	--	--	0.54	90.32
	ZFX	Zinc finger protein X-linked	*ZFY*	0.45	1.2e-29	0.47	4.0e-21	0.55	0.51	0.64	82.07
	DDX3X	DEAD-box helicase 3 X-linked	*DDX3Y*	0.26	5.5e-21	0.16	5.6e-17	0.56	0.48	0.55	89.24
	USP9X	Ubiquitin specific protease 9 X-linked	*USP9Y*	0.14	3.4e-09	0.17	1.4e-07	0.10	0.24	0.20	94.39
	SMC1A	Structural maintenance of chromosomes 1A	--	0.58	7.8e-30	0.43	8.4e-08	0.37	0.56	0.30	87.68
	DLG3	Discs large MAGUK scaffold protein 3	--	0.18	3.6e-03	0.07	0.13	--	--	0.01	82.91
	BCOR	BCL6 corepressor	--	0.12	2.0e-06	0.01	0.83	0.10	--	0.01	91.12
	CDK16	Cyclin dependent kinase 16	--	0.09	5.9e-06	0.24	3.9e-15	--	--	0.22	83.91
	RBBP7	RB binding protein 7, chromatin remodeling factor	--	0.01	0.71	0.29	1.3e-09	0.24	0.65	0.09	82.65
	SAT1	Spermidine/spermine N1-acetyltransferase 1	--	-0.06	0.26	-0.16	0.022	0.10	--	0.03	75.33
	ARHGAP6	Rho GTPase activating protein 6	--	-0.14	0.020	--	--	0.13	--	0.01	78.62
	JADE3	Jade family PHD finger 3	--	-0.15	0.059	0.13	5.5e-04	--	--	0.14	75.54

PAR1	*SLC25A6*	Solute carrier family 25 member 6	n/a	1.0	3.3e-25	0.74	3.5e-15	0.82	--	0.88	67.69
	SHOX	Short stature homeobox	n/a	n/a[c]	n/a	n/a	n/a	n/a	n/a	n/a	58.59

[a] Average of published AR values from (Cotton et al., 2013, Tukiainen et al., 2017, Garieri et al., 2018, Sauteraud et al., 2021).
[b] Gene constraint percentile ranking is calculated for NPX genes relative to all annotated NPX genes, and for PAR1 genes relative to all PAR and autosomal genes.
[c] *SHOX* is not expressed in fibroblasts or LCLs but is included because its dosage has been conclusively linked to height in individuals with sex chromosome aneuploidy.

Discussion:

Assessing direct effects of Xi copy number on Chr X gene expression

The dataset presented here encompassed a broad spectrum of naturally occurring sex chromosome constitutions, allowing us to quantify the effects of adding one, two, or three copies of Xi on Chr X gene expression. Combining all samples into one analysis using linear models had several advantages over previous approaches employing pairwise comparisons between sex chromosome constitutions. First, this method allowed for greater statistical power to detect gene expression differences. Rather than limiting our analyses to pairwise comparisons – and thus to subsets of the data -- we leveraged the data from all samples to overcome the variability between human samples. This variability may arise at multiple biological and technical levels, including genetic variation apart from sex Chr copy number, and variation in procedures for deriving cell lines or cultures. We estimated that the size of our dataset allowed us to find all but the smallest differences in Chr X gene expression due to Xi copy number. Second, this approach *directly* attributed expression differences to the cell-autonomous effects of Xi copy number without being confounded by other correlated variables, such as sex hormones. Third, this approach revealed that the expression contributions made by each copy of Xi were surprisingly modular, indicating that each copy of Xi is more or less equivalent. Finally, by comparing samples that vary in Xi copy number with and without a Y chromosome, we found that expression from Xa is

quantitatively indistinguishable in phenotypic males and females. Thus, both Xa and Xi make strikingly modular contributions to X Chr gene expression.

Quantitative insights into gene expression on Xa and Xi

Our approach led to new quantitative insights regarding gene expression on Chr X. We found that few Chr X genes had $\Delta E_X = 1$, the value expected if expression from Xi equaled that from Xa. Instead, nearly all NPX genes – and even some PAR1 genes – had $\Delta E_X < 1$, indicating that expression from Xi is almost invariably attenuated relative to that from Xa. Similar analyses of samples with additional copies of Chr Y and 21 revealed ΔE values closer to one for Chr Y and 21 genes, confirming that a chromosome-wide mechanism of attenuation is specific to Xi and does not apply to supernumerary copies of other chromosomes.

Another advantage of our study was the utilization of two cell types, enabling us to understand to what extent the results were cell-type specific or potentially applicable across the body. For Chr X genes expressed in both cell types, we found that Xi had highly consistent effects (similar ΔE_X values), particularly for genes with evidence of robust expression directly from Xi ($\Delta E_X>0$ and AR>0). LCLs and skin fibroblasts both have mesodermal embryonic origins, indicating that – for widely expressed genes – the epigenetic state of Xi was established in or before their common mesodermal progenitor. A study that investigated sex-biased expression for Chr X genes across a wide range of tissues, including those derived from ectoderm and endoderm, showed consistent female-bias across tissues for many of the genes that we found to be robustly expressed from Xi (Tukiainen et al., 2017). This raises the possibility that the gene-by-gene attenuation on Xi that we observed in mesodermal derivatives may be established prior to gastrulation, and subsequently maintained across the germ layers.

By combining ΔE_X with allele-specific expression analyses, we found that multiple gene regulatory mechanisms underlie the spectrum of responses to an additional copy of Xi. For some genes, ΔE_X was explained entirely by expression from the Xi allele, while for others ΔE_X was explained entirely by the modulation – positive or negative – of steady-state RNA levels derived from the Xa allele. For a third set of genes, we found that ΔE_X was explained by the combined effects of expression from Xi *and* modulation of steady-state RNA levels. The mechanisms and biological consequences of this modulation of transcript levels with Xi copy number must now be explored.

In all, our results allow us to move beyond a qualitative classification of genes that either "escape" or are "subject" to XCI and towards a quantitative accounting of the effects of Xi copy number on gene expression. Applying a quantitative approach has already proved useful for obtaining new insights into Chr X gene regulatory mechanisms, and it provides a foundation for biochemical, cell biological, developmental, and evolutionary studies of Xi and Xa.

Implications of Xi expression on human health and disease

By combining ΔE_X with population and evolutionary measures of constraint on expression levels, we can inform our expectations for how Chr X genes contribute to i) phenotypes that differ between typical euploid males (46,XY) and females (46,XX) and ii) phenotypes associated with sex chromosome aneuploidies. We propose that phenotypes driven by Xi copy number are caused by genes whose expression levels i) change with copy number *and* ii) are critical for fitness or viability. Further, we propose that this holds in both euploid and aneuploid individuals. Here, we consider how genes in three classes – PAR1 genes, NPX-NPY pair genes, and NPX genes without Y pairs – impact phenotypes related to Xi copy number.

PAR1 genes, as a group, displayed the highest ΔE_X values, and have long been an appealing set of candidates to contribute to sex chromosome aneuploidy phenotypes. Some of the strongest evidence implicating PAR1 genes in sex chromosome aneuploidy comes from the study of Turner syndrome (TS), classically associated with a 45,X karyotype. Since 46,XY individuals do not display phenotypes associated with TS despite having a single copy of Chr X, the genes underlying TS somatic phenotypes must be shared between Chr X and Y, implicating PAR and/or NPX-NPY pair genes. However, we found that the expression levels of PAR1 genes were among the least-constrained genome-wide, with only *SLC25A6* and *SHOX* showing higher-than-average levels of selection against over- or under-expression. Overall, these results suggest that the expression levels of other PAR1 genes do not contribute to sex chromosome aneuploidy phenotypes.

NPX-NPY pair genes, like PAR1 genes, have high ΔE_X values. However, unlike PAR1 genes, they are highly expression-constrained, with five in the top quartile of NPX genes: *KDM6A*, *KDM5C, ZFX, DDX3X,* and *USP9X*. We propose that these NPX genes with NPY homologs represent promising candidates to contribute to phenotypes related to Xi copy number – including the 99% in utero lethality of monosomy X. They are expressed throughout the body and are involved in basic cellular functions including transcription, translation, and protein stability (Bellott et al., 2014).

An open question is the functional interchangeability of NPX-NPY pair genes, since they have been evolving independently on Chr X and Y for many millions of years. If the NPX and NPY homologs have maintained complete functional interchangeability, then their expression levels and patterns are the only factors that would affect phenotypes. In this case, we would anticipate larger effects in individuals with sex chromosome aneuploidies (with fewer or

additional copies of these genes) and more subtle effects between euploid females and males due to differences in expression from Xi or Y. Indeed, we observed that the NPX homologs' ΔE_X values were invariably lower than the NPY homologs' ΔE_Y values, and a recent analysis of the cumulative expression of the NPX and NPY pair genes revealed higher expression in 46,XY males (Godfrey et al., 2020). On the other hand, if NPX-NPY pair functions have diverged, we expect more significant effects due to the combinatorial impact of expression *and* functional differences. Focused comparative studies of the functions of NPX-NPY pair genes will be required to understand their roles in euploid sex differences and in sex chromosome aneuploidy phenotypes.

Finally, we considered the phenotypic contributions of NPX genes that have no NPY homologs. Expression levels of most of these genes are not affected by Xi copy number, but a subset had $\Delta E_X \neq 0$ and expression constraints ranking in the top quartile of NPX genes. In particular, expression of *SMC1A* and *CDK16* increased significantly with Xi copy number in both LCLs and fibroblasts. NPX genes without NPY homologs would be expected to contribute to phenotypes of sex chromosome aneuploidies where Xi copy number varies, and also between euploid males and females.

Comparisons of phenotypes among individuals with sex chromosome aneuploidies may be helpful in pinpointing the genes responsible. For example, 47,XXY and 47,XYY males display a higher prevalence of autism spectrum disorder (ASD) traits than 46,XY males, who in turn display a higher prevalence of ASD traits than 46,XX females (Ross et al., 2012, Tartaglia et al., 2017, van Rijn, 2019). Interestingly, among males, a second copy of Chr Y (47,XYY) leads to an even higher prevalence of ASD traits than does the addition of an Xi (47,XXY). This suggests that both Xi and Y influence certain phenotypes, but that the strengths of their effects

differ. We suspect that one or more NPX-NPY pairs play a role in ASD traits, with the higher prevalence in 47,XYY reflecting the higher ΔE_Y values for NPY genes as compared with the ΔE_X values of their NPX homologs, or functional differences between NPX-NPY pair genes, or both. Indeed, six NPX-NPY gene pairs have already been implicated in ASD traits [SFARI Gene database (Banerjee-Basu and Packer, 2010)]. In contrast to ASD traits, other neurodevelopmental phenotypes, such as verbal and language weaknesses, are more prevalent among individuals with extra copies of Xi, compared to those with additional copies of Chr Y (Green et al., 2019). For such phenotypes, we hypothesize important roles for NPX genes that are functionally differentiated from their NPY partners or have no such partners.

This work also allows us to better understand the features of Chr X, Y, and 21 that may account for the milder phenotypes associated with an extra X or Y compared to trisomy 21. As with the sex chromosomes, it is thought that dosage-sensitive genes on Chr 21 are responsible for the Down syndrome phenotype (Makino and McLysaght, 2010). Although we observed similar numbers of genes with $\Delta E>0$ on Chr X and Chr 21, repression of most genes on Xi resulted in lower median ΔE values compared to Chr 21. We conclude that dosage compensation on Chr X permits the viability of individuals with up to three additional chromosomes (e.g., 49,XXXXY). While median ΔE values were comparable for Chr Y and Chr 21, the smaller number of genes on Chr Y may account for the milder phenotypes associated with additional copies of Chr Y.

Altogether, the dataset and accompanying analyses presented here provide a foundation for studies of Turner, Klinefelter and other sex chromosome aneuploidy phenotypes, and for studies of sex-chromosome effects on health and disease in euploid (typical) males and females. In addition, this analysis provides opportunities to revisit many aspects of X chromosome biology, including its epigenetic regulation and evolution.

Acknowledgements:

We thank members of the Page laboratory, especially Lukas Chmatal, for helpful comments on the manuscript, Sahin Naqvi for advice on RNA-sequencing and analysis, Jorge Adarme and Susan Tocio for laboratory support, and the Whitehead Institute Genome Technology Core facility for library preparation and sequencing.

Funding:

National Institutes of Health grant F32HD091966 (AKSR)

National Institutes of Health grant U01HG0007587 (DCP and MM)

National Institutes of Health grant K23HD092588 (SMD)

Schmidt Science Fellows (AFG)

Lallage Feazel Wall Damon Runyon Cancer Research Foundation Fellowship (LVB)

Howard Hughes Medical Institute (DCP)

National Human Genome Research Institute Intramural Research Program (MM)

NIH/NCATS Colorado CTSA grant UL1 TR002535 (NRT)

Contents are the authors' sole responsibility and do not necessarily represent official NIH views.

Philanthropic gifts from:

Brit and Alexander d'Arbeloff

Arthur W. and Carol Tobin Brill

Matthew Brill

Charles Ellis

Author contributions:

Conceptualization: AKSR and DCP

Data curation: AKSR and HS

Formal analysis: AKSR, AKG, HS, DWB, AG, HLH, LVB, and JFH

Funding acquisition: AKSR, AFG, LVB, SMD, NRT, MM, and DCP

Investigation: AKSR, SP, LB, AD, and EP

Methodology: AKSR, AKG, HS, AFG, and HLH

Project administration: AKSR, JFH, LB, AB, PK, NB, PCL, CK, and SMD

Resources: AB, PK, NB, PCL, CK, SMD, NRT, CSS, and MM

Software: AKSR, AKG, HS, AFG, HLH, and LVB

Supervision: AKSR, JFH, NRT, CSS, MM, and DCP

Validation: AKG, HS, DWB, and LVB

Visualization: AKSR, AKG, and HLH

Writing - original draft preparation: AKSR and DCP

Writing - review and editing: AKSR, AKG, DWB, AFG, HLH, LVB, JFH, SD, and DCP

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

Methods

RESOURCE AVAILABILITY

Lead contact

Further information and request for resources and reagents should be directed to and will be fulfilled by lead contact, David C. Page (dcpage@wi.mit.edu).

Materials availability

Cell lines are available upon request to the lead contact.

Data and code availability

- Raw, de-identified RNA-sequencing data from human cell cultures has been deposited to dbGaP (accession # phs002481), and processed data has been deposited to github (https://github.com/aksanroman/Xi_expression.git).
- Original code has been deposited at github (https://github.com/aksanroman/Xi_expression.git).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human subjects

Adults (18+ years of age) with sex chromosome (Chr) aneuploidies or euploid controls were recruited through an IRB-approved study at the NIH Clinical Center (12-HG-0181) and Whitehead Institute/MIT (Protocol #1706013503). Informed consent was obtained from all study

participants. Individuals with a previous karyotype showing non-mosaic sex Chr aneuploidy were included in the study. From these individuals, blood samples and skin biopsies were collected at the NIH Clinical Center and shipped to the Page lab for derivation of cell lines. In addition, blood samples from individuals with sex Chr aneuploidies, and euploid family members, ranging in age from 4-44 years were contributed by the Focus Foundation. Additional LCLs and fibroblast cultures were obtained from the Colorado Children's Hospital Biobank and Coriell Research Institute, and cultured in the Page laboratory for at least two passages prior to collection for RNA-sequencing. Karyotyping of peripheral blood and fibroblast cell cultures was performed at the National Human Genome Research Institute Cytogenetics and Microscopy Core. To reduce the impact of sex Chr mosaicism on our sex Chr aneuploidy analysis, we excluded individuals with >15% mosaicism for other karyotypes. Metadata for cell lines represented in the RNA-sequencing dataset are provided in **Table S1**.

Method Details

Cell culture

Lymphoblastoid cell lines

Blood was collected in BD Vacutainer ACD tubes and shipped at room temperature to the Page Lab for processing 1-3 d after collection. The buffy coat was resolved by centrifuging blood at 3300 rpm for 10 min, transferred to a new tube with PBS, and subjected to density gradient centrifugation in 50% Percoll (Cytiva) at 3300 rpm for 10 min. Lymphocytes were transferred to a new tube and washed twice with PBS. Lymphocytes were resuspended in 3 mL complete RPMI medium (RPMI 1640 (Gibco), 25mM HEPES (SAFC), 15% FBS (Hyclone), Fungizone (Amphotericin B, Gibco), Gentamicin (Gibco), Penicillin/Streptomycin (Lonza), pH 7.2) per

tube of blood and transferred to a T25 flask, supplemented with 0.25mL EBV (produced by B95-8 marmoset lymphoblasts), and 0.2 mL of 1 mg/mL cyclosporine (LC Laboratories). They were incubated for one week at 37°C, fed 1-2 mL complete RPMI, and incubated for another week at 37°C. Once the media began to turn yellow (acidified), cultures were "half-fed" by removing half of the media and replacing it with double the volume. When cultures reached 15 mL, they were transferred to T75 flasks, and gradually expanded to 30 mL, while maintaining a concentration of <1 million cells/mL to ensure viability. Cells were viably frozen for future use by mixing with freezing media (LCL culture media + 5% DMSO), 1 million cells per vial. Cells were also preserved for RNA, DNA, and protein extraction (see below).

Primary fibroblast cultures

Our protocol for generating primary skin fibroblast cultures from a skin biopsy is based on (Vangipuram et al., 2013). From adults (18+ years of age) at the NIH Clinical Center we obtained two 4-mm skin punch biopsies from the upper arm, which were immediately placed into a 15 ml conical tube with 10 ml of media (DMEM/F12 (Gibco), 20% FBS, and 100 IU/ml Penicillin-Streptomycin. Tubes were shipped to the Page lab overnight on ice for processing. Each biopsy was used to generate a separate skin fibroblast culture. Biopsies were cut into 18 pieces of equal size and placed 3/well in gelatinized 6-well plates with 1 mL media (High Glucose DMEM (Gibco), 20% FBS, L-Glutamine (MP Biomedicals), MEM Non-Essential Amino Acids (Gibco), 100 IU/ml Penicillin/Streptomycin (Lonza)). Plates were gelatinized by incubating 1 mL sterile 0.1% gelatin (Sigma) solution per well for 30 minutes at room temperature.

Plates were incubated for 1 week at 37°C without disturbance to allow biopsies to attach to the plate and begin to grow out. During week 2, we added 200 µl of fresh media per well every 2-3 days, being careful not to disturb the biopsies. The following week (week 3), we aspirated the media and replaced with 1 mL fresh media per well every 2-3 days. During week 4, we aspirated the media and replaced with 2 mL fresh media per well every 2-3 days. At this point, the fibroblasts generally reached the edges of the wells and were expanded to two T75 gelatinized flasks per 6 well plate. After two days, we combined the cells from the two T75 flasks and split them to three T175 gelatinized flasks. After two days, cells were viably frozen with 1 million cells per vial in freezing media (fibroblast culture media + 5% DMSO). Cells were also preserved for RNA extraction (see below). During optimization of the protocol, cell culture purity was confirmed by immunofluorescence of SERPINH1, a fibroblast marker.

Cell collection for subsequent analysis

Cells were collected when LCL cultures reached 30mL, and fibroblasts were ~80% confluent in three T175 plates. All cell counting was performed using the Countess II cell counter (Life Technologies) and Trypan Blue exclusion. Cultures with >85% cell viability were used in subsequent experiments. To preserve cells for subsequent RNA extraction, 1 million cells were washed in PBS, pelleted, and resuspended in 500 µl TRIzol (Invitrogen) or 200 µl RNAprotect Cell Reagent (Qiagen). Cell suspensions were then frozen at -80°C. Cell cultures were maintained at low passage number; RNA-sequencing experiments were performed on samples at or below passage 4.

Periodically, and on each passage used for experiments, cell cultures were confirmed negative for mycoplasma contamination using either the MycoAlert Kit (Lonza) following the

manufacturer's instructions, or PCR using SapphireAmp Fast PCR Master Mix (Takara) and the following primers:

Myco2(cb): 5' ctt cwt cga ctt yca gac cca agg cat 3'
Myco11(cb): 5' aca cca tgg gag ytg gta at 3'

PCR for *GAPDH* was performed on the same sample, using the following primers:

hGAPDH-F: TGT CGC TGT TGA AGT CAG AGG AGA
hGAPDH-R: AGA ACA TCA TCC CTG CCT CTA CTG

Known mycoplasma positive and negative samples were used as a reference.

RNA extraction, library preparation, and sequencing

RNA was extracted from 1 million cells per experiment using the RNeasy Plus Mini Kit (Qiagen) following the manufacturer's instructions, with the following modifications: Cells in RNAprotect Cell Reagent were thawed on ice, pelleted, and lysed in buffer RLT supplemented with 10µl β-mercaptoethanol per mL. For most samples, ERCC control RNAs were added to the lysate based on the number of cells: 10µl of 1:100 dilution of ERCC control RNAs was added per 1 million cells. The lysate was then homogenized using QIAshredder columns (Qiagen), and transferred to a gDNA eliminator column. All subsequent optional steps in the protocol were performed, and RNA was eluted in 30 µl RNase-free water. RNA levels were measured using a Qubit fluorometer and the Qubit RNA HS Assay Kit (ThermoFisher). Before we switched to the per-cell spike-in protocol, we prepared 18 samples in which ERCC control RNAs were added based on amount of RNA after isolation: 2 µl of a 1:100 dilution of ERCC control RNAs was added per 1 µg of RNA. These samples are: #2237, 2245, 6312, 711, 4032, 706, 3429, 3430, 3442, 2690, 2703, 3107, 5297, 5566, 5755, 6029, 2547, and 525. RNA quality control was performed using the 5200 Fragment Analyzer System (Agilent); we consistently purified high-

quality RNA with RNA integrity numbers (RIN) near 10. We randomized the samples by karyotype into batches for RNA extraction, library preparation, and sequencing.

RNA sequencing libraries were prepared using the TruSeq RNA Library Preparation Kit v2 (Illumina) with modifications as detailed in (Naqvi et al., 2019), or using the KAPA mRNA HyperPrep Kit V2 (Roche). In both cases, libraries were size selected using the PippinHT system (Sage Science) and 2% agarose gels with a capture window of 300-600 bp. Paired-end 100x100 bp sequencing was performed on a HiSeq 2500 or NovaSeq 6000 (Illumina). **Table S1** lists the library preparation kit and sequencing platform for each sample.

RNA-seq data processing and analysis

All analyses were performed using human genome build hg38, and a custom version of the comprehensive GENCODE v24 transcriptome annotation (Godfrey et al., 2020). This annotation represents the union of the "GENCODE Basic" annotation *and* transcripts recognized by the Consensus Coding Sequence project (Pruitt et al., 2009). Importantly, the GENCODE annotation lists the PAR gene annotations twice – once on Chr X and once on Chr Y– which complicates analysis. We removed these annotations from Chr Y so the PAR genes are only listed once in our annotation, on Chr X. To analyze samples in which ERCC spike-ins were added, we merged our custom transcript annotation with the ERCC Control annotation.

Reads were pseudoaligned to the transcriptome annotation, and expression levels of each transcript were estimated using kallisto software v0.42.5 (Bray et al., 2016). We included the "--bias" flag to correct for sequence bias. The resulting count data (abundance.tsv file) were imported into R with the tximport package v1.14.0 (Soneson et al., 2015) for normalization using DESeq2 v1.26.0 (Love et al., 2014). For downstream analysis, we used only protein-coding

genes (as annotated in ensembl v104) with the following exceptions: we included genes annotated as pseudogenes on Chr Y that are members of X-Y pairs (*TXLNGY, PRKY*) and well-characterized long non-coding RNAs (lncRNAs) involved in X-inactivation or other processes (*XIST, JPX, FTX, XACT, FIRRE, TSIX).* We annotated genes distal to *XG*, which spans the pseudoautosomal boundary on Xp and is truncated on Chr Y, as part of PAR1 - 15 genes in total. PAR2 comprised the four most distal genes on Xq and Yq. Annotations of non-pseudoautosomal region of the X (NPX) genes with homologs on the non-pseudoautosomal region of the Y (NPY) were derived from (Bellott and Page, 2021). 224 protein-coding genes on Chr 21 (ensembl v104) were used as a starting point for our analyses. We excluded 21 annotated genes in several regions with high homology between the long and short arms of Chr 21 because the assembly was not fully validated in these regions (https://www.ncbi.nlm.nih.gov/grc/human/issues?filters=chr:21).

Identifying genes affected by changes in Chr X, Y, or 21 copy number

We first defined lists of expressed NPX, NPY, PAR, or Chr 21 genes as those with median TPM of at least 1 in 46,XX or 46,XY samples. To ensure that no genes with robust expression were excluded, we also analyzed LCL and fibroblast expression data from GTEx (Consortium et al., 2017), and included several genes that were just below our TPM cutoff but had median TPM of at least 1 in those datasets.

For each expressed NPX, NPY, or PAR gene we performed linear modeling using the lm() function in R. These calculations suppose that each additional Chr adds a consistent and equal increment to the total expression level of the gene in question.

For NPX and PAR genes we used the following equation:

$$E = \beta_0 + \beta_X(\#chrXi) + \beta_Y(\#chrY) + \beta_B(batch) + \in$$

E represents the expression (read counts) per gene, β_0 represents the intercept, β_X and β_Y are the coefficients of the effect of additional Chrs Xi or Y, respectively, and $\in$ is an error term. For this equation, the intercept represents the 45,X samples.

For NPY genes we employed the following equation, analyzing only those samples with one or more copies of Chr Y:

$$E = \beta_0 + \beta_X(\#chrXi) + \beta_Y(\#chrY - 1) + \beta_B(batch) + \in$$

For this equation, the intercept represents the 46,XY samples.

For Chr 21 genes we employed the following equation, analyzing only those samples with 46,XX; 46,XX; 47,XY,+21; or 47,XX,+21 karyotypes:

$$E = \beta_0 + \beta_{21}(\#chr21 - 2) + \beta_{Sex}(Sex) + \beta_B(batch) + \in$$

β_{21} and β_{Sex} are the coefficients of the effect of an additional copy of Chr 21 and sex (XY vs XX), respectively. For this equation, the intercept represents the 46,XX samples.

The resulting *P*-values were adjusted for multiple hypothesis testing using the p.adjust() function in R, specifying the Benjamini Hochberg method. Genes with a false discovery rate (FDR) < 0.05 were considered significant. To compute the normalized expression change per Chr Xi (ΔE_X) or Y (ΔE_Y), we divided the coefficient of interest (β_X or β_Y) by the average intercept across batches, which corresponds to the baseline expression of the gene in samples with only one X chromosome (for NPX and PAR genes) or one Y chromosome (in the case of NPY genes). For Chr 21, we computed ΔE_{21} by dividing the coefficient (β_{21}) by the average intercept across batches divided by two to obtain the average expression from one copy of Chr 21.

$$\Delta E_X = \frac{\beta_X}{\beta_0} \qquad \Delta E_Y = \frac{\beta_Y}{\beta_0} \qquad \Delta E_{21} = \frac{\beta_{21}}{\beta_0/2}$$

In the case of *XIST*, which is only expressed when two or more copies of Chr X are present, we used the following equations:

$$\Delta E_X = \frac{\beta_X}{\beta_0 + \beta_X} \qquad \Delta E_Y = \frac{\beta_Y}{\beta_0 + \beta_X}$$

We calculated the standard error of ΔE_X, ΔE_Y, and ΔE_{21} using the following equations:

$$S_{\Delta E_X} = \sqrt{\frac{{\beta_X}^2}{{\beta_0}^2}\left[\frac{{S_{\beta_X}}^2}{{\beta_X}^2} + \frac{{S_{\beta_0}}^2}{{\beta_0}^2}\right]} \qquad S_{\Delta E_Y} = \sqrt{\frac{{\beta_Y}^2}{{\beta_0}^2}\left[\frac{{S_{\beta_Y}}^2}{{\beta_Y}^2} + \frac{{S_{\beta_0}}^2}{{\beta_0}^2}\right]} \qquad S_{\Delta 21} = \sqrt{\frac{{\beta_{21}}^2}{(\beta_0/2)^2}\left[\frac{{S_{\beta_{21}}}^2}{{\beta_{21}}^2} + \frac{{S_{\beta_0}}^2}{(\beta_0/2)^2}\right]}$$

To confirm the validity of our approach, we used bootstrapping to sample our dataset with replacement 1000 times and obtained similar results. *BEX1* was removed from downstream analyses in fibroblasts because two samples (one 45,X and one 49,XXXXY) had high expression values for this gene resulting in > 25 times higher error values for ΔE_X and ΔE_Y compared to all other genes.

Saturation analysis for sex chromosome-encoded genes

For LCLs and fibroblasts, size-*n* subsets of available RNA-seq libraries were sampled randomly without replacement, 100 times for each sample size, *n*. After confirming that the model matrix would be full rank in each sampling (for example, that samples would not all be of the same karyotype or batch), we performed linear modeling on NPX, PAR, NPY, and Chr 21 genes as described above to identify genes whose expression changes significantly (FDR < 0.05) with copy number of Chr X, Y or 21.

Assessing linearity of sex-chromosome gene expression changes

To assess whether sex-chromosome gene expression changed linearly (i.e., by a fixed amount) with additional X or Y chromosomes, their expression levels across the LCL or fibroblast

samples were fit by non-linear least squares to the power curves shown below, using the "nlsLM" function from the R package "minpack.lm".

NPX genes:

$$y_j^* = 1 + b\left(xcount_j - 1\right)^a, \text{ where } y_j^* = \frac{y_j}{y_k}$$

PAR genes:

$$y_j^* = 1 + b\left(xcount_j + ycount_j - 1\right)^a, \text{ where } y_j^* = \frac{y_j}{y_k}$$

NPY genes:

$$y_j^* = 1 + b\left(ycount_j - 1\right)^a, \text{ where } y_j^* = \frac{y_j}{y_k}$$

In each of the equations above, y_j^* is the normalized RNA-seq read count for a given gene in sample j, given by the raw read count in sample j divided by the average read count in the set of samples $S_{\{\cdot\}}$ with only one chromosome of the relevant type: for NPX genes, 1 copy of Chr X (and any number of Y Chrs); for PAR genes, 45,X samples; for NPY genes, 1 copy of Chr Y (and any number of X Chrs). $b = 0.5$ and $a = 1$ were used as initial parameter values. Fitted values of $a \approx 1$ indicate a linear relationship between expression and sex-chromosome count. Fitted values of $a \approx 0$ or $b \approx 0$ indicate no change in expression with X or Y count.

ΔE_X calculations in samples with 0 Y chromosomes (females) and 1 Y chromosome (males)

We took subsets of the samples with either zero Y chromosomes (females) and with one Y chromosome (males) and performed the same linear modeling and ΔE_X calculations as above. We removed *MAP7D2* in female LCLs, *IL13RA2* in female fibroblasts, and *FHL1* in male fibroblasts because their error values (likely due to smaller sample size) were much higher than

those of other genes. To compare the linear modeling results, we performed Pearson correlations between the results using all samples, and those from male-only or female-only samples.

Reanalysis of array data and comparison to RNA-seq data

A previous study performed gene expression analysis, using Illumina oligonucleotide BeadArrays, of LCLs from 68 individuals of the following karyotypes: 45,X; 46,XX; 46,XY; 47,XXX; 47,XXY; 47,XYY; and 48,XXYY (Raznahan et al., 2018). Since this microarray dataset was generated from an independent set of samples, we sought to validate our results through a reanalysis of the data.

The raw data from the microarrays was not publicly available, but the authors provided us pre-processed data upon request, which we used to perform our analysis. To identify genes that cleared a minimum signal threshold to be considered expressed in the microarray data, we assessed the median signal in 46,XY samples for all Chr Y genes annotated on the microarray. We focused on Chr Y genes in this analysis because many are known to be expressed exclusively in testes, and therefore could provide us with an appropriate sense of the background signal expected for genes not expressed in LCLs. From this analysis, we concluded that a signal threshold of 111 would be appropriate for identifying expressed genes. We used this threshold to identify 278 expressed Chr X genes (including PAR and NPX genes) in the microarray dataset. This was fewer than the 341 expressed Chr X genes identified in our LCL RNA-seq data, but more than double the 121 expressed Chr X genes considered in Raznahan *et al.* (Supplemental Table 4 in (Raznahan et al., 2018)). This discrepancy could not be resolved by simply increasing the signal threshold in our analysis, as *TMSB4X,* one of the most highly expressed genes in LCLs, was excluded from Raznahan and colleagues' list of expressed genes.

Using our list of 278 expressed genes from the Raznahan *et al.* dataset, we analyzed the microarray signal values (in place of RNA-seq read counts) using linear models as a function of Xi copy number, controlling for Chr Y copy number. We calculated ΔE_X values from the microarray data and compared these to our RNA-seq dataset using a Pearson correlation, which revealed that the results were generally concordant. For genes that were lowly expressed, however, ΔE_X values tended to be much lower in the microarray dataset, consistent with the higher sensitivity of RNA-seq data.

Isoform-specific analysis of RNA-seq data

After estimating counts for each transcript using kallisto software (as described above, with 100 bootstraps), we used sleuth v0.30.0 to normalize those transcript counts (Pimentel et al., 2017). X chromosome transcripts were called as expressed if their corresponding gene was on the list of expressed genes (above) and median transcript counts were > 100. Linear regressions and ΔE_X calculations were performed as for genes (above) to identify transcripts whose abundance changes significantly with additional copies of Chr X.

The following ENCODE datasets were used for visualization in IGV software (Robinson et al., 2011) at the *UBA1* locus:

ChIP-seq	Cell type	Cell line	Karyotype	Accession
CTCF	LCL	GM12878	46,XX	ENCFF852CRG
		GM12864	46,XY	ENCFF593YIG
	Fibroblast	AG09309	46,XX	ENCFF640EZJ
		AG10803	46,XY	ENCFF694SQA

Allele-specific expression analysis

SNP calling

We called SNPs in each RNA-seq sample with two X chromosomes (46,XX, 46,XX,+21, 47,XXY, 48,XXYY) following the Broad Institute's "Best Practices" workflow for identifying short variants in RNA-seq data (https://gatk.broadinstitute.org/hc/en-us/articles/360035531192-RNAseq-short-variant-discovery-SNPs-Indels-). To perform our skewing analysis, we filtered for SNPs with the following properties: 1) annotated in the dbSNP database, 2) located within an exon of an expressed gene, 3) displaying a minimum coverage of 10 reads, and 4) heterozygous with at least 3 reads mapping to each of the reference and alternative alleles. We excluded SNPs where the presence of two alleles likely represented technical artifacts rather than biallelic expression, including in *WASH6P* (SNPs map to multiple near-identical autosomal paralogs), *ATRX* (SNP in a mutation-prone stretch of Ts), and *APOOL* (SNPs within an inverted repeat). For samples with a copy of Chr Y, we excluded SNPs mapping to PAR genes, to avoid measuring allelic contributions of Chr Y.

Identifying cell lines with skewed X chromosome inactivation

We classified genes as "Xa-only" (only expressed from Xa) if previously characterized as "silenced" (Balaton et al., 2015) *and* found here to have $\Delta E_X<0.05$ (FDR>0.5). We expect that in skewed cell lines, reads from Xa-only genes should be near or completely monoallelic. For each SNP in Xa-only genes, we calculated the "skewing coefficient" by dividing the number of reads from the dominant allele by the total number of reads covering the SNP. These coefficients range from 0.5 (equal expression of two alleles) to 1 (expression from a single allele). For each sample, we computed the median skewing coefficient across all SNPs in Xa-only genes, requiring a

threshold of 0.8 to classify as skewed. Using simulations, we find that this level of skewing is unlikely to occur by chance ($P<1x10^{-6}$), and we do not find evidence of such skewing for SNPs on Chr 8, an autosome with a similar number of expressed genes.

Several samples had few (≤ 5) informative SNPs in Xa-only genes, but many SNPs in XaXi genes. We interpret this to mean that these samples are highly skewed and that we do not observe enough RNA reads covering both alleles to count SNPs in Xa-only genes as informative. Between these highly-skewed samples and the samples with skewing coefficients of at least 0.8, we classified 22 LCLs and 10 fibroblast cultures with skewed XCI.

Determining allelic ratios for X chromosome genes

After identifying the skewed cell lines, we computed the allelic ratio (AR) at each informative SNP by dividing the number of reads from the more lowly expressed allele by the number of reads from the more highly expressed allele. In cell cultures that are partially skewed, genes will appear more biallelic than in completely skewed cell cultures since there are two populations of cells with different active X chromosomes present – the "major" and "minor" cell populations. Using our skewing estimates, we adjusted the AR on a per-sample basis using the following formula:

$$Adj_AR = \frac{AR - AR * t - t}{1 - t - AR * t}$$

Where t is the estimated percentage of cells in the "minor" population (i.e., with the other X chromosome active compared to the "major" cell population), calculated by: 1 – skewing coefficient. For highly-skewed samples, we were unable to calculate a stringent skewing coefficient due to too few SNPs, so we set skewing coefficient = 1. As a result, it is possible that allelic ratios in these samples may be slightly overestimated if the skewing coefficients are in

fact <1. Within each sample, we obtained the average AR for each gene by averaging across all SNPs in that gene's exons. We filtered for genes with AR values in at least three skewed samples of a given cell type and then calculated the mean AR across skewed samples to obtain a final per-gene AR estimate.

To assess whether AR values for each gene were significantly greater than zero, we performed paired, two-sample, one-sided t-tests using the t.test function in R, asking whether the raw AR values were greater than the baseline AR given the level of skewing in each sample ($baseline\ AR = \frac{1-skewing\ coefficient}{skewing\ coefficient}$); we corrected the resulting P-values for multiple comparisons with the p.adjust function in R using the Benjamini-Hochberg method. We also repeated this analysis excluding highly skewed samples, since the skewing coefficients cannot be stringently determined. This removed some informative genes but did not significantly affect the AR levels (**Table S6**).

To identify genes whose AR and ΔE_X values differ significantly, we performed one-sample, two-sided t-tests for the AR values across samples, setting mu= ΔE_X. We selected genes with Benjamini-Hochberg adjusted P-values <0.05 as having significantly different AR and ΔE_X values. From this list we excluded genes for which the 95% confidence interval of ΔE_X values (1.96*standard error) included the mean AR value, and those for which both ΔE_X and AR were not significantly different from zero (FDR≥ 0.05).

<u>Comparing allelic ratios with published datasets</u>

We compared five additional published datasets for AR with our AR values derived from LCLs or fibroblasts. The first dataset was from paired genomic and cDNA SNP-chips in skewed LCL

and fibroblast cell cultures (Additional file 7 in (Cotton et al., 2013)). We used the AR values provided for genes with at least five informative samples.

The second dataset was from bulk RNA-seq on LCLs from an individual in the GTEx dataset with 100% skewed XCI across the body (Table S5 in (Tukiainen et al., 2017)). We recalculated AR from this data using the reported read counts from the more lowly and highly expressed alleles.

The third dataset was from single-cell RNA-seq in LCLs from three individuals (Table S8 in Tukiainen *et al.* – we excluded data from one dendritic cell sample). We recalculated the median AR for each gene using the read counts from the lowly and highly expressed alleles across samples and required data from at least two individuals to be considered informative for a given gene.

The fourth dataset was derived from single-cell allelic expression in fibroblasts (Garieri et al., 2018). We converted their reported values (Xa reads/total reads) to AR values using this formula: $AR = \frac{1}{\frac{Xa\,reads}{total\,reads}} - 1$. The dataset includes five individuals (Dataset 3 in Garieri et al.), and we required data from at least 3 different samples to be considered informative for a given gene.

The fifth dataset was derived from allele-specific bulk RNA-seq performed on 136 samples from the set of GEUVADIS LCLs (Table S4 in (Sauteraud et al., 2021)). We used their data for "Mean.Xi.Expression" as the AR values.

Gene constraint analysis

To investigate sensitivity to a reduction in gene dosage, we used three metrics: LOEUF, RVIS, and pHI. We downloaded LOEUF (loss-of-function observed/expected upper fraction) scores

from gnomAD (v2.1.1.lof_metris.by_gene.txt; https://gnomad.broadinstitute.org/), only using scores with a minimum of 10 expected LoF variants. Updated RVIS (residual variation intolerance scores) from (Petrovski et al., 2013) including the ExAC dataset were downloaded from http://genic-intolerance.org/data/RVIS_Unpublished_ExAC_May2015.txt. Updated pHI (probability of haploinsufficiency) scores from (Huang et al., 2010) were downloaded from https://www.deciphergenomics.org/files/downloads/HI_Predictions_Version3.bed.gz. To complement these data, we obtained a list of genes with observed homozygous loss-of-function variants from (Karczewski et al., 2020). For sensitivity to an increase in gene dosage, we used the per-gene average probability of conserved miRNA targeting scores (P_{CT}) (Friedman et al., 2009).

For each metric, we computed a percentile rank score, ranking from most- to least-constrained. Because several of the metrics calculated scores separately for autosomal (including PAR genes) and NPX genes, we also ranked autosomal (and PAR) genes separately from NPX genes. All annotated genes, regardless of expression status in LCLs or fibroblasts, were included in the rankings, with the following exceptions: 1) NPX genes previously annotated as "ampliconic" in (Mueller et al., 2013, Jackson et al., 2021), since constraint metrics cannot be accurately applied to these highly similar genes. 2) Genes with <2 annotations across all metrics.

To obtain an aggregate sense of a gene's expression constraint across multiple metrics, we calculated the average. Among NPX genes, we considered those with an average gene constraint percentile ranking of at least 75% and with $|\Delta E_X \geq 0.1|$ (FDR<0.05) to be most likely to contribute to phenotypes mediated by Xi copy number. For PAR1 genes, we prioritized genes with an average gene constraint percentile ranking of at least 50%.

QUANTIFICATION AND STATISTICAL ANALYSES

Various statistical tests were used to calculate P-values as indicated in the methods section, figure legend, or text, where appropriate. Results were considered statistically significant when $P<0.05$ or $FDR<0.05$ when multiple hypothesis correction was applied. Unless otherwise stated, data are shown as median and interquartile range. All statistics were calculated using R software, version 3.6.3 (R Development Core Team, 2020).

References

Balaton, B. P., Cotton, A. M. & Brown, C. J. 2015. Derivation of consensus inactivation status for X-linked genes from genome-wide studies. *Biology of Sex Differences,* 6.

Balaton, B. P., Fornes, O., Wasserman, W. W. & Brown, C. J. 2021. Cross-species examination of X-chromosome inactivation highlights domains of escape from silencing. *Epigenetics Chromatin,* 14**,** 12.

Banerjee-Basu, S. & Packer, A. 2010. SFARI Gene: an evolving database for the autism research community. *Dis Model Mech,* 3**,** 133-5.

Bartel, D. P. 2009. MicroRNAs: target recognition and regulatory functions. *Cell,* 136**,** 215-33.

Bellott, D. W., Hughes, J. F., Skaletsky, H., Brown, L. G., Pyntikova, T., Cho, T.-J., Koutseva, N., Zaghlul, S., Graves, T., Rock, S., et al. 2014. Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature,* 508**,** 494-499.

Bellott, D. W. & Page, D. C. 2021. Dosage-sensitive functions in embryonic development drove the survival of genes on sex-specific chromosomes in snakes, birds, and mammals. *Genome Res*.

Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology,* 34**,** 525-527.

Brown, C. J., Ballabio, A., Rupert, J. L., Lafreniere, R. G., Grompe, M., Tonlorenzi, R. & Willard, H. F. 1991. A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature,* 349**,** 38-44.

Brown, C. J., Carrel, L. & Willard, H. F. 1997. Expression of genes from the human active and inactive X chromosomes. *Am J Hum Genet,* 60**,** 1333-43.

Carrel, L., Cottle, A. A., Goglin, K. C. & Willard, H. F. 1999. A first-generation X-inactivation profile of the human X chromosome. *Proc Natl Acad Sci U S A,* 96**,** 14440-4.

Carrel, L. & Willard, H. F. 2005. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature,* 434**,** 400-404.

Chen, C. Y., Shi, W., Balaton, B. P., Matthews, A. M., Li, Y., Arenillas, D. J., Mathelier, A., Itoh, M., Kawaji, H., Lassmann, T., et al. 2016. YY1 binding association with sex-biased transcription revealed through X-linked transcript levels and allelic binding analyses. *Sci Rep,* 6**,** 37324.

Clement-Jones, M., Schiller, S., Rao, E., Blaschke, R. J., Zuniga, A., Zeller, R., Robson, S. C., Binder, G., Glass, I., Strachan, T., et al. 2000. The short stature homeobox gene SHOX is involved in skeletal abnormalities in Turner syndrome. *Hum Mol Genet,* 9**,** 695-702.

Consortium, G., Laboratory, D. A., Coordinating Center -Analysis Working, G., Statistical Methods Groups-Analysis Working, G., Enhancing, G. G., Fund, N. I. H. C., Nih/Nci, Nih/Nhgri, Nih/Nimh, Nih/Nida, et al. 2017. Genetic effects on gene expression across human tissues. *Nature,* 550**,** 204-213.

Cotton, A. M., Ge, B., Light, N., Adoue, V., Pastinen, T. & Brown, C. J. 2013. Analysis of expressed SNPs identifies variable extents of expression from the human inactive X chromosome. *Genome Biology,* 14**,** R122.

Craig, I. W., Mill, J., Craig, G. M., Loat, C. & Schalkwyk, L. C. 2004. Application of microarrays to the analysis of the inactivation status of human X-linked genes expressed in lymphocytes. *Eur J Hum Genet,* 12**,** 639-46.

Ford, C. E., Jones, K. W., Polani, P. E., Dealmeida, J. C. & Briggs, J. H. 1959. A Sex-Chromosome Anomaly in a Case of Gonadal Dysgenesis (Turners Syndrome). *Lancet,* 1**,** 711-713.

Freije, D., Helms, C., Watson, M. S. & Donis-Keller, H. 1992. Identification of a second pseudoautosomal region near the Xq and Yq telomeres. *Science,* 258**,** 1784-7.

Friedman, R. C., Farh, K. K., Burge, C. B. & Bartel, D. P. 2009. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res,* 19**,** 92-105.

Fukami, M., Seki, A. & Ogata, T. 2016. SHOX Haploinsufficiency as a Cause of Syndromic and Nonsyndromic Short Stature. *Mol Syndromol,* 7**,** 3-11.

Garieri, M., Stamoulis, G., Blanc, X., Falconnet, E., Ribaux, P., Borel, C., Santoni, F. & Antonarakis, S. E. 2018. Extensive cellular heterogeneity of X inactivation revealed by single-cell allele-specific expression in human fibroblasts. *Proc Natl Acad Sci U S A,* 115**,** 13015-13020.

Godfrey, A. K., Naqvi, S., Chmatal, L., Chick, J. M., Mitchell, R. N., Gygi, S. P., Skaletsky, H. & Page, D. C. 2020. Quantitative analysis of Y-Chromosome gene expression across 36 human tissues. *Genome Res,* 30**,** 860-873.

Goto, Y. & Kimura, H. 2009. Inactive X chromosome-specific histone H3 modifications and CpG hypomethylation flank a chromatin boundary between an X-inactivated and an escape gene. *Nucleic Acids Res,* 37**,** 7416-28.

Green, T., Flash, S. & Reiss, A. L. 2019. Sex differences in psychiatric disorders: what we can learn from sex chromosome aneuploidies. *Neuropsychopharmacology,* 44**,** 9-21.

Hook, E. B. & Warburton, D. 1983. The distribution of chromosomal genotypes associated with Turner's syndrome: livebirth prevalence rates and evidence for diminished fetal

mortality and severity in genotypes associated with structural X abnormalities or mosaicism. *Human Genetics,* 64**,** 24-27.

Hook, E. B. & Warburton, D. 2014. Turner syndrome revisited: review of new data supports the hypothesis that all viable 45,X cases are cryptic mosaics with a rescue cell line, implying an origin by mitotic loss. *Human Genetics,* 133**,** 417-424.

Huang, N., Lee, I., Marcotte, E. M. & Hurles, M. E. 2010. Characterising and predicting haploinsufficiency in the human genome. *PLoS genetics,* 6**,** e1001154.

Jackson, E. K., Bellott, D. W., Cho, T. J., Skaletsky, H., Hughes, J. F., Pyntikova, T. & Page, D. C. 2021. Large palindromes on the primate X Chromosome are preserved by natural selection. *Genome Res,* 31**,** 1337-1352.

Johnston, C. M., Lovell, F. L., Leongamornlert, D. A., Stranger, B. E., Dermitzakis, E. T. & Ross, M. T. 2008. Large-scale population study of human cell lines indicates that dosage compensation is virtually complete. *PLoS genetics,* 4**,** e9.

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alfoldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., et al. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature,* 581**,** 434-443.

Love, M. I., Huber, W. & Anders, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology,* 15**,** 550.

Lyon, M. F. 1961. Gene action in the X-chromosome of the mouse (Mus musculus L.). *Nature,* 190**,** 372-3.

Lyon, M. F. 1962. Sex chromatin and gene action in the mammalian X-chromosome. *Am J Hum Genet,* 14**,** 135-48.

Makino, T. & Mclysaght, A. 2010. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc Natl Acad Sci U S A,* 107**,** 9270-4.

Megarbane, A., Ravel, A., Mircher, C., Sturtz, F., Grattau, Y., Rethore, M. O., Delabar, J. M. & Mobley, W. C. 2009. The 50th anniversary of the discovery of trisomy 21: the past, present, and future of research and treatment of Down syndrome. *Genet Med,* 11**,** 611-6.

Mohandas, T., Sparkes, R. S., Hellkuhl, B., Grzeschik, K. H. & Shapiro, L. J. 1980. Expression of an X-linked gene from an inactive human X chromosome in mouse-human hybrid cells: further evidence for the noninactivation of the steroid sulfatase locus in man. *Proc Natl Acad Sci U S A,* 77**,** 6759-63.

Mueller, J. L., Skaletsky, H., Brown, L. G., Zaghlul, S., Rock, S., Graves, T., Auger, K., Warren, W. C., Wilson, R. K. & Page, D. C. 2013. Independent specialization of the human and mouse X chromosomes for the male germ line. *Nat Genet,* 45**,** 1083-7.

Naqvi, S., Godfrey, A. K., Hughes, J. F., Goodheart, M. L., Mitchell, R. N. & Page, D. C. 2019. Conservation, acquisition, and functional impact of sex-biased gene expression in mammals. *Science,* 365.

Nielsen, M. M., Trolle, C., Vang, S., Hornshoj, H., Skakkebaek, A., Hedegaard, J., Nordentoft, I., Pedersen, J. S. & Gravholt, C. H. 2020. Epigenetic and transcriptomic consequences of excess X-chromosome material in 47,XXX syndrome-A comparison with Turner syndrome and 46,XX females. *Am J Med Genet C Semin Med Genet,* 184**,** 279-293.

Ogata, T. & Matsuo, N. 1993. Sex chromosome aberrations and stature: deduction of the principal factors involved in the determination of adult height. *Hum Genet,* 91**,** 551-62.

Ohno, S. & Hauschka, T. S. 1960. Allocycly of the X-chromosome in tumors and normal tissues. *Cancer Res,* 20**,** 541-5.

Ottesen, A. M., Aksglaede, L., Garn, I., Tartaglia, N., Tassone, F., Gravholt, C. H., Bojesen, A., Sorensen, K., Jorgensen, N., Rajpert-De Meyts, E., et al. 2010. Increased number of sex chromosomes affects height in a nonlinear fashion: a study of 305 patients with sex chromosome aneuploidy. *Am J Med Genet A,* 152A**,** 1206-12.

Penny, G. D., Kay, G. F., Sheardown, S. A., Rastan, S. & Brockdorff, N. 1996. Requirement for Xist in X chromosome inactivation. *Nature,* 379**,** 131-137.

Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. 2013. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet,* 9**,** e1003709.

Pimentel, H., Bray, N. L., Puente, S., Melsted, P. & Pachter, L. 2017. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat Methods,* 14**,** 687-690.

Pruitt, K. D., Harrow, J., Harte, R. A., Wallin, C., Diekhans, M., Maglott, D. R., Searle, S., Farrell, C. M., Loveland, J. E., Ruef, B. J., et al. 2009. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res,* 19**,** 1316-23.

R Development Core Team 2020. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rao, E., Weiss, B., Fukami, M., Rump, A., Niesler, B., Mertz, A., Muroya, K., Binder, G., Kirsch, S., Winkelmann, M., et al. 1997. Pseudoautosomal deletions encompassing a novel homeobox gene cause growth failure in idiopathic short stature and Turner syndrome. *Nature Genetics,* 16**,** 54-63.

Raznahan, A., Parikshak, N. N., Chandran, V., Blumenthal, J. D., Clasen, L. S., Alexander-Bloch, A. F., Zinn, A. R., Wangsa, D., Wise, J., Murphy, D. G. M., et al. 2018. Sex-

chromosome dosage effects on gene expression in humans. *Proc Natl Acad Sci U S A,* 115**,** 7398-7403.

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. & Mesirov, J. P. 2011. Integrative genomics viewer. *Nat Biotechnol,* 29**,** 24-6.

Ross, J. L., Roeltgen, D. P., Kushner, H., Zinn, A. R., Reiss, A., Bardsley, M. Z., Mccauley, E. & Tartaglia, N. 2012. Behavioral and Social Phenotypes in Boys With 47,XYY Syndrome or 47,XXY Klinefelter Syndrome. *PEDIATRICS,* 129**,** 769-778.

Ross, M. T., Grafham, D. V., Coffey, A. J., Scherer, S., Mclay, K., Muzny, D., Platzer, M., Howell, G. R., Burrows, C., Bird, C. P., et al. 2005. The DNA sequence of the human X chromosome. *Nature,* 434**,** 325-37.

Sauteraud, R., Stahl, J. M., James, J., Englebright, M., Chen, F., Zhan, X., Carrel, L. & Liu, D. J. 2021. Inferring genes that escape X-Chromosome inactivation reveals important contribution of variable escape genes to sex-biased diseases. *Genome Res,* 31**,** 1629-1637.

Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P. J., Cordum, H. S., Hillier, L., Brown, L. G., Repping, S., Pyntikova, T., Ali, J., Bieri, T., et al. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature,* 423**,** 825-837.

Soneson, C., Love, M. I. & Robinson, M. D. 2015. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res,* 4**,** 1521.

Sudbrak, R., Wieczorek, G., Nuber, U. A., Mann, W., Kirchner, R., Erdogan, F., Brown, C. J., Wöhrle, D., Sterk, P., Kalscheuer, V. M., et al. 2001. X chromosome-specific cDNA arrays: identification of genes that escape from X-inactivation and other applications. *Human molecular genetics,* 10**,** 77-83.

Sullivan, K. D., Lewis, H. C., Hill, A. A., Pandey, A., Jackson, L. P., Cabral, J. M., Smith, K. P., Liggett, L. A., Gomez, E. B., Galbraith, M. D., et al. 2016. Trisomy 21 consistently activates the interferon response. *Elife,* 5.

Talebizadeh, Z., Simon, S. D. & Butler, M. G. 2006. X chromosome gene expression in human tissues: male and female comparisons. *Genomics,* 88**,** 675-681.

Tartaglia, N. R., Wilson, R., Miller, J. S., Rafalko, J., Cordeiro, L., Davis, S., Hessl, D. & Ross, J. 2017. Autism Spectrum Disorder in Males with Sex Chromosome Aneuploidy: XXY/Klinefelter Syndrome, XYY, and XXYY. *J Dev Behav Pediatr,* 38**,** 197-207.

Trolle, C., Nielsen, M. M., Skakkebæk, A., Lamy, P., Vang, S., Hedegaard, J., Nordentoft, I., Ørntoft, T. F., Pedersen, J. S. & Gravholt, C. H. 2016. Widespread DNA hypomethylation and differential gene expression in Turner syndrome. *Scientific reports,* 6**,** 34220-14.

Tukiainen, T., Villani, A. C., Yen, A., Rivas, M. A., Marshall, J. L., Satija, R., Aguirre, M., Gauthier, L., Fleharty, M., Kirby, A., et al. 2017. Landscape of X chromosome inactivation across human tissues. *Nature,* 550**,** 244-248.

Turner, H. H. 1938. A syndrome of infantilism, congenital webbed neck, and cubitus valgus. *Endocrinology,* 23**,** 566-574.

Van Rijn, S. 2019. A review of neurocognitive functioning and risk for psychopathology in sex chromosome trisomy (47,XXY, 47,XXX, 47, XYY). *Curr Opin Psychiatry,* 32**,** 79-84.

Vangipuram, M., Ting, D., Kim, S., Diaz, R. & Schüle, B. 2013. Skin Punch Biopsy Explant Culture for Derivation of Primary Human Fibroblasts. *JoVE (Journal of Visualized Experiments)***,** e3779-e3779.

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

Zhang, X., Hong, D., Ma, S., Ward, T., Ho, M., Pattni, R., Duren, Z., Stankov, A., Bade Shrestha, S., Hallmayer, J., et al. 2020. Integrated functional genomic analyses of Klinefelter and Turner syndromes reveal global network effects of altered X chromosome dosage. *Proc Natl Acad Sci U S A,* 117**,** 4864-4873.