Supplementary Materials for

A gene-by-gene mosaic of dosage compensation strategies on the human X chromosome

Adrianna K. San Roman, Alexander K. Godfrey, Helen Skaletsky, Daniel W. Bellott, Abigail F. Groff, Laura V. Blanton, Jennifer F. Hughes, Laura Brown, Sidaly Phou, Ashley Buscetta, Paul Kruszka, Nicole Banks, Amalia Dutra, Evgenia Pak, Patricia C. Lasutschinkow, Colleen Keen, Shanlee M. Davis, Nicole R. Tartaglia, Carole Samango-Sprouse, Maximilian Muenke, and David C. Page*

*Correspondence to: dcpage@wi.mit.edu

This PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S12
Captions for Tables S1 to S18
Supplemental References

Other Supplementary Materials for this manuscript include the following:

Tables S1 to S18 (.xlsx)

Materials and Methods

Annotation of Chr X and Y genes

For the X and Y chromosomes, we used only protein-coding genes (as annotated in ensembl v104) with the following exceptions: we included genes annotated as pseudogenes on the Y chromosome that are part 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 listed genes in the pseudoautosomal regions (PARs), found on both Chr X and Y, only once – on the X chromosome. We annotated genes distal to *XG*, which is located on the PAR boundary and truncated on the Y chromosome, as part of PAR1 - 15 genes in total. PAR2 comprised the four most distal genes on the long arm of the X and Y chromosomes. Annotations of non-pseudoautosomal region of the X (NPX) genes with homologs on the non-pseudoautosomal region of the Y (NPY) across mammals were derived from (*40*). Copy number assignments were derived from (*49, 50*) for Chr X genes and (*26*) for Chr Y genes.

Annotation of Chr 21 genes

224 protein-coding genes on Chr 21 (ensembl v104) were used as a starting point for our downstream 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 properly validated in these regions (https://www.ncbi.nlm.nih.gov/grc/human/issues?filters=chr:21).

Dosage sensitivity analysis

To investigate sensitivity to a reduction in gene dosage, we downloaded the loss-of-function observed/expected upper fraction (LOEUF) scores from gnomAD

(v2.1.1.lof_metris.by_gene.txt; https://gnomad.broadinstitute.org/). Genes with homozygous loss-of-function variants were obtained from (*20*). For sensitivity to an increase in gene dosage, we used the per-gene average probability of conserved miRNA targeting scores (P_{CT}) (*21*). For genes with data for both scores, we divided genes into quadrants based on the 50th percentile score, with the highest dosage sensitivity in either direction for genes in quadrant one and the lowest for genes in quadrant four. Common essential genes were downloaded from the Broad Institute's Dependency Map (Achilles_common_essentials.csv, release 21Q1, depmap.org; (*32*)).

Expression breadth and testis specificity scores

To define expression breadth, we re-analyzed data from GTEx V7 (accessed from dbGaP accession phs000424.v7.p2; (*31*)), across tissues from male samples (so as to include germ cells, which are the majority of the cells in the testis, but a minority in the ovary, and so are not adequately captured in female samples). To obtain the expression breadth for each gene, we calculated the adjusted median TPM per tissue (see (*17*)), and used the following equation, where n is the number of GTEx tissues:

$$Expression\ breadth\ = \frac{\sum_1^n TPM}{max(TPM)\ \times\ n}$$

We used the following equation to calculate a testis-specificity score using GTEx tissues:

$$Testis\ specificity\ = \frac{TPM\ in\ testis}{\sum_1^n TPM}$$

Recruitment of study participants and karyotyping

Adults (18+ years of age) with sex chromosome 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 previous karyotype showing non-mosaic sex chromosome 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 chromosome aneuploidies, and euploid family members, ranging in age from 4-44 years were contributed by the Focus Foundation. Additional derived LCLs and fibroblasts 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 chromosome mosaicism on our sex chromosome aneuploidy analysis, we only included individuals with <15% mosaicism for other karyotypes.

Cell culture methods

Lymphoblastoid cell lines:

Blood was collected in BD Vacutainer ACT blood collection tubes, and shipped at room temperature to the Page Lab for processing 1-3 days after collection. Buffy coat was resolved by spinning blood for 10 minutes at 3300rpm, transferred to a new tube with PBS, and subjected to density gradient centrifugation with 50% Percoll for 10 minutes at 3300 rpm. Lymphocytes were transferred into a new tube and washed twice with PBS. Lymphocytes were resuspended in 3mL complete RPMI (RPMI 1640, 25mM HEPES, 15% FBS, Fungizone, Gentamycin, Pen/Strep, pH

7.2) per tube of blood and transferred to a T25 flask, supplemented with 0.25mL EBV, and 0.2mL 1mg/mL cyclosporine. They were incubated for one week at 37°C, fed 1-2mL complete RPMI, and incubated for another week at 37°C. Once the media begins to turn yellow (acidifies), cultures were "half-fed" by removing half of the media and replacing it with double the volume. When cultures reached 15mL, they were transferred to T75 flasks, and gradually expanded to 30mL, 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 (*51*). From adults at the NIH Clinical Center (18+ years of age) we obtained two 4mm skin punch biopsies from the upper arm, which were immediately placed into a 15ml conical tube with 10ml of media (DMEM/F12 (ThermoFisher), 20% FBS (ThermoFisher), and 100 IU/ml Penicillin-Streptomycin (ThermoFisher). 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 1mL media (High Glucose DMEM (Gibco), 20% FBS (Hyclone), L-Glutamine (MP Biomedicals), MEM Non-Essential Amino Acids (Gibco), Penicillin/Streptomycin (Lonza)). Gelatinized plates were prepared by incubating 1mL sterile 0.1% gelatin (Sigma) solution per well for 30 minutes at room temperature.

Plates were incubated for 1 week at 37°C without disturbing to allow biopsies to attach to plate and begin to grow out. During week 2, we added 200µl 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 1mL fresh media per well every 2-3 days. During week 4, we aspirated the media and replaced with 2mL 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, DNA, and protein 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 ul TRIzol (Invitrogen) or 200 ul RNAprotect Cell Reagent (Qiagen). Cell suspensions were then frozen at -80°C. To preserve cells for subsequent protein and/or DNA extraction, cells were counted and aliquoted, washed in PBS, pelleted, and snap-frozen in liquid nitrogen.

Cell culture quality control:

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.

Cell cultures were maintained at low passage number to minimize genetic drift; RNA-sequencing experiments were performed on samples at or below passage 4.

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 30µl 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: 2µl of a 1:100 dilution of ERCC control RNAs was added per 1µg of RNA. These samples are: #2237, 2245, 6312, 711, 4032, 706, 3429, 3430, 3442, 2690, 2703, 3107, 5297, 5566, 5755, 6029, 2547, and 525. We consistently purified high-quality RNA with RNA integrity numbers (RIN) near 10. To minimize batch effects, we randomized the order of RNA extraction, library preparation, and sequencing, while maximizing the number of samples we could sequence simultaneously (24-40 samples/flow cell on Illumina HiSeq 2500).

RNA sequencing libraries were prepared using the TruSeq RNA Library Preparation Kit v2 (Illumina) with modifications as detailed in (*52*). Briefly, libraries were size-selected using the PippinHT system (Sage Science) and 2% agarose gels with a capture window of 300-600bp. An additional round of PCR was performed using fresh reagents, and libraries were size selected again using the same settings to reduce the amount of small fragments in the final library. Paired-end 100x100 sequencing was performed on an Illumina HiSeq 2500 sequencer with 3-5 samples per lane to obtain ~50 million reads per library.

All RNA-seq data has been deposited to dbGaP, accession # phs002481.v1.p1.

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 as in (*17*). This annotation represents the union of the "GENCODE Basic" annotation and transcripts recognized by the Consensus Coding Sequence project (*53*). Importantly, the GENCODE annotation lists the PAR gene annotations twice – once on the X chromosome and once on the Y chromosome – which

complicates analysis. We removed these annotations from the Y chromosome so the PAR genes are only listed once in our annotation, on the X chromosome. 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) (*54*). 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) (*55*) for normalization using DESeq2 (1.26.0) (*56*).

To analyze reproducibility of our cell line derivation and sequencing workflow, we performed RNA-sequencing on samples that were obtained from the same individual, but underwent separate cell line derivation, RNA isolation, and sequencing processes. We used the normalized, variance stabilized transformed counts (from DESeq2), removed the effects of library preparation batch using the limma removeBatchEffect function (*57*), and then calculated the pairwise sample distances on the residuals (1-Spearman correlation).

Identifying genes on the sex chromosomes that are affected by sex chromosome dosage

We first defined a list 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 we did not exclude genes with robust expression, we compared with LCL and fibroblast expression data from GTEx, 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 assume that each additional chromosome provides an equal amount of additional expression to the total expression level of that gene.

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 Xi or Y chromosomes, respectively, and ∈ is an error term. For this equation, the intercept represents the 45,X samples.

For NPY genes we used the following equation, only using samples with at least one Y chromosome:

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

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

For Chr 21 genes we used the following equation, only using samples with 46,XX, 46,XX, 47,XY,+21 and 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 (for NPX and PAR) or one Y (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 in the case of two or more X chromosomes, 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 our downstream analyses in fibroblasts because two samples (one 45,X and one 49,XXXXY) had high expression values for this gene resulting in more than 25 times higher error values for ΔE_X or 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 randomly sampled without replacement, 100 times for each sample size, *n*. After checking that the model matrix would be full rank with each sampling (for example, all samples cannot be from the same karyotype or batch), we performed linear modeling on NPX, PAR, NPY, and Chr 21 genes as above to find the number of genes that significantly change in expression (FDR < 0.05) with dosage of Chr X, Y or 21.

X chromosome inactivation status annotations

We used annotations from (*11*), with several modifications. We did not include *XG* in our list of PAR genes, because it is located on the PAR boundary and truncated on the Y chromosome. We replaced its annotation with escape ("E") per previous evidence. We classified the status of *XIST* as escape ("E") as it is expressed from the heterochromatinized X chromosome. For genes not present in the Balaton study, we list "NA".

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 "informative" SNPs - that is, those in the dbSNP database, located within an exon, and heterozygous, with at least 3 reads covering the reference and alternative alleles to control for sequencing errors. We also excluded SNPs for which 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 in *APOOL* (SNPs are within an inverted repeat). For samples with a Y chromosome, we excluded SNPs mapping to PAR genes, to avoid measuring allelic contributions of the Y chromosome.

Identifying cell lines with skewed X chromosome inactivation

We classified a list of genes as likely only expressed from the active X chromosome ("Xa-only" genes) using the overlap of genes previously characterized as "silenced" (*11*) and those with $\Delta E_X<0.05$ and FDR>0.5. For comparison, we also considered genes as likely expressed from both the active and inactive chromosomes ("XaXi" genes) using the overlap of previously characterized as "escape" or "mostly escape" (*11*) and those with $\Delta E_X>0$ and FDR<0.05 (**Table S17**).

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 allele with the higher number of reads by the total number of reads covering the SNP. This value spans from 0.5 (equal expression of two alleles) to 1 (expression from a single allele). For each sample, we computed the skewing coefficient across all SNPs in Xa-only genes, requiring a threshold of 0.8 to classify as skewed (**Table S18**). 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 informative SNPs in Xa-only genes, but many SNPs in XaXi genes. We interpret this to mean that these samples are extremely skewed and that we do not observe enough RNA reads covering both alleles to count SNPs in Xa-only genes as informative even though at the DNA level, there likely are SNPs in these Xa-only genes.

Between these extremely skewed samples and the samples with skewing coefficients of at least 0.8 we classified 18 LCLs and 4 fibroblasts as having skewed XCI.

Determining allelic ratio for X chromosome genes

After identifying the skewed cell lines, we computed the allelic ratio (AR) of each informative SNP by dividing the number of reads from the minimum allele by the maximum 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 samples with extreme skewing, we set the skewing coefficient = 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 samples for LCLs and at least two samples for fibroblasts, and then calculated the median AR across samples to obtain a final per-gene AR estimate.

Analyzing allelic expression from published datasets

Additional published datasets for AR were used for comparison with our list of genes expressed in 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. (*38*)). We used the AR values provided for genes with at least five informative samples. There were 307 informative genes that overlapped with our expressed gene list and a gene was considered biallelic if AR≥0.1.

The second dataset was from bulk RNA-seq on LCLs from an individual in the GTEx dataset with 100% skewed XCI across the body (Supplementary Table 5 in Tukiainen et al. (*34*)). We recalculated AR from this data using the reported read counts from the minimum and maximum alleles. There were 116 informative genes that overlapped with our expressed gene set, and these were considered biallelic if their adjusted p-values reported in Tukiainen et al. were <0.05.

The third dataset was from single-cell RNA-seq in LCLs from three individuals (Supplementary Table 8 in Tukiainen et al. – we did not use the data from the one dendritic cell sample). We re-calculated the median AR for each gene using the read counts from the minimum and maximum alleles across samples and required data from at least two individuals to be considered informative, resulting in 40 genes that overlapped with our expressed gene set. In Tukiainen et al, biallelic expression was called on an individual sample basis. We compiled these p-values into "biallelic" or "monoallelic", according to the majority of calls, or "discordant" if there were two calls in opposite directions.

Finally, we used a fibroblast single-cell allelic expression dataset from Garieri et al (*39*). We converted their reported value (Xa reads/total reads) to AR 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 an informative gene, resulting in 125 informative genes. A gene was considered biallelic if it had a median AR≥0.052 (converted from Garieri et al cutoff value of 0.95 for Xa/total reads).

We classified a gene as monoallelic if the majority of observations across datasets was monoallelic. If there was an equal number of observations in each direction, or if the majority of observations biallelic, the gene was considered biallelic.

Adjacency analysis

We defined "adjacent" genes as the nearest expressed gene, either upstream or downstream, at any distance, of a gene of interest. To assess whether NPX genes without Y pairs that have $\Delta E_X > 0$ ("group 2" genes) are more likely to be adjacent to NPX-NPY pairs with $\Delta E_X > 0$ in eutherian mammals including humans ("group 1" genes), we performed a positional shuffling analysis. Using the list of expressed NPX genes in each cell type, we randomly shuffled gene positions and calculated the percentage of group 2 genes adjacent to group 1 genes, repeating this shuffling 1000 times. We then compared the shuffled distribution to the actual percentage of adjacent genes. We calculated an empiric P-value using the percentage of shuffled samples that meet or exceed the actual percentage of adjacent genes.

Code availability

All analyses were conducted in R (v3.6.3) unless otherwise indicated. Code and processed data for recreating these analyses are available at https://github.com/aksanroman/dosage_compensation_strategies.git.

Supplemental Text

In this manuscript, we separately considered genes in pseudoautosomal region 1, PAR1, on the short arms of X and Y (referred to simply as PAR in the main text), from those in a second PAR, PAR2. PAR1 genes are unique because they never lost the ability to exchange genetic information between X and Y during evolution of the sex chromosomes over the past 300 million years, thus retaining identical gene copies on each chromosome (*15*). The genes on PAR2, on the other hand, have not retained identical gene copies on X and Y throughout this evolutionary process. Of PAR2's four genes, two were present on the ancestral autosomes and lost their Y homologs, evolving down a similar path to other NPX genes, while two were added to Chr X in separate transposition events from autosomes (*58*). In humans, these four genes were transposed from the X to the Y, creating PAR2 (*59*). Another distinguishing feature between PAR1 and PAR2 is the frequency of crossing-over in male meiosis: crossing-over in the 2.6 Mb PAR1 occurs in every meiosis, having one of the highest recombination rates across the male genome (*60*), while crossing-over in the 320 kb PAR2 is infrequent (*25*).

In our analysis of samples with sex chromosome aneuploidy, we found that PAR2 genes have ΔE_X and $\Delta E_Y \approx 0$, consistent with them being exclusively expressed from the first X chromosome and silenced on all Y chromosomes (**fig. S12**). Thus, these genes operate more like NPX genes without Y homologs that have acquired complete silencing on the inactive X chromosome. We suspect that the transposition of PAR2 genes to the Y chromosome distal to a large heterochromatic region that separates PAR2 from all other Y chromosome genes resulted in a significant position effect that maintained their epigenetically repressed status.

Figure S1.

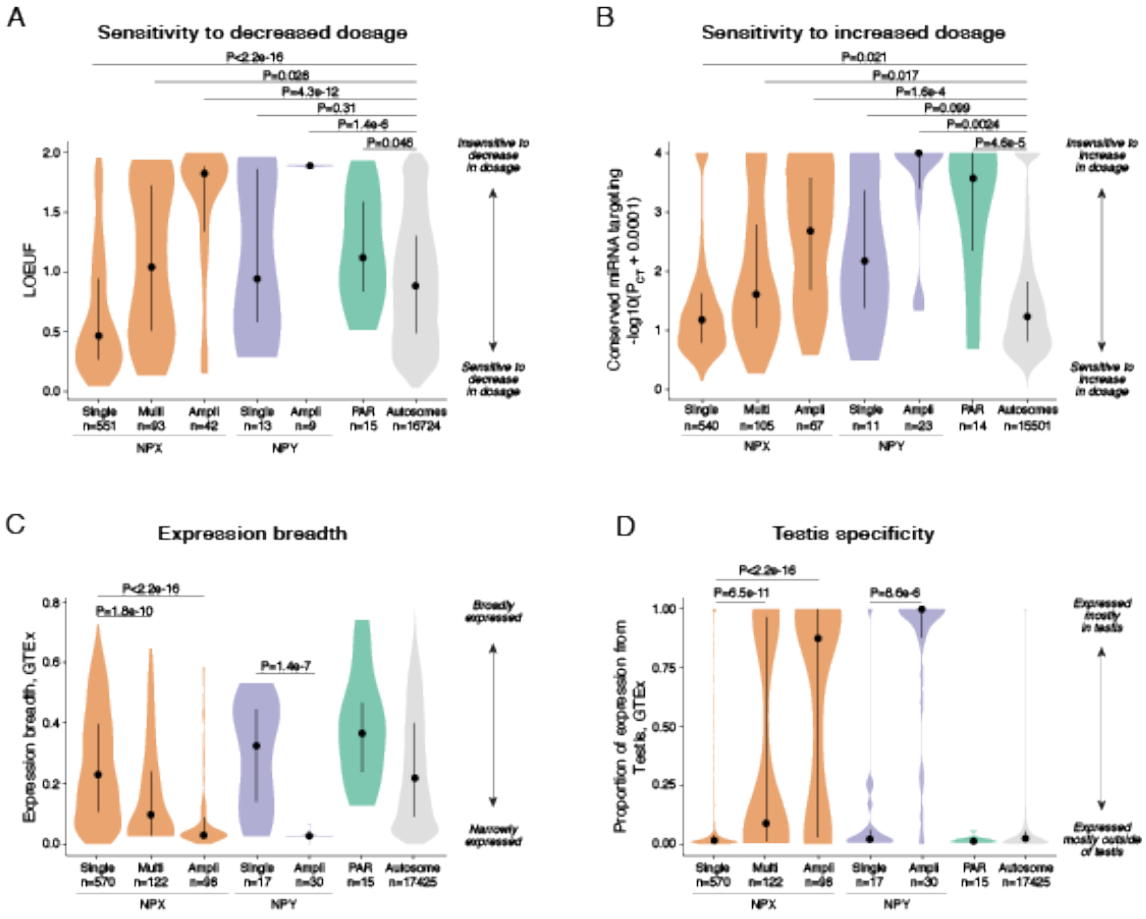

Figure S1. Single-copy genes accumulate fewer loss of function mutations, have more conserved miRNA targeting, and are more broadly-expressed than multi-copy or ampliconic genes. Violin plots with median (dots) and interquartile range (whiskers) of LOEUF scores **(A)**, transformed conserved miRNA targeting scores **(B)** expression breadth across GTEx samples **(C)**, and testis-specificity **(D)**, for single-copy, multicopy, and ampliconic genes on the X and Y chromosomes, PAR genes, and autosomal genes. P-values, Wilcoxon-rank sum test.

Figure S2.

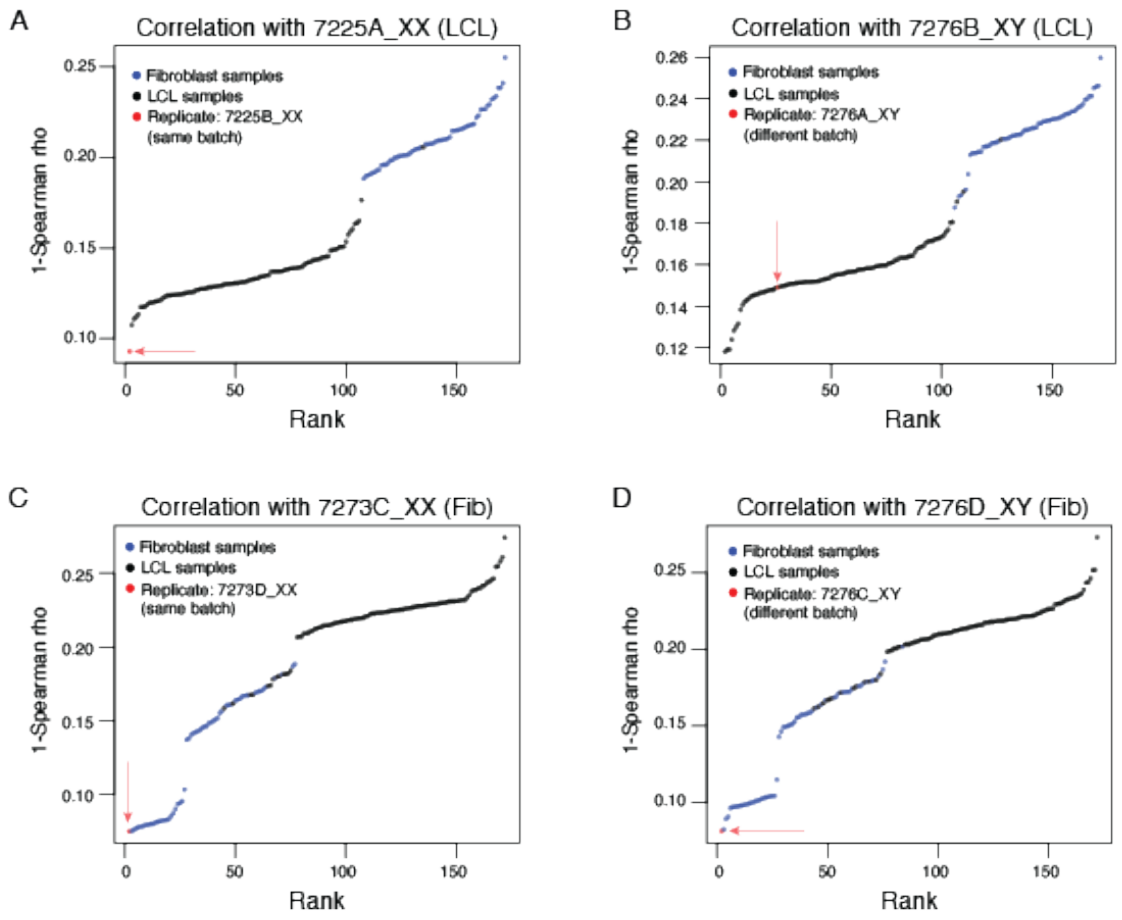

Figure S2. RNA-seq of independent samples from the same individuals are highly correlated. (A-D) The pairwise distances (1-Spearman correlation) between the indicated sample and all other samples in the sex chromosome aneuploidy dosage series were ranked from lowest to highest. Independent cultures of LCLs or fibroblasts from the same individual were sequenced and considered replicates (red dot and arrows). Replicates were highly correlated, and were more similar to samples of the same cell type. When replicates were within the same library preparation and sequencing batch **(A, C)** they tended to have the lowest pairwise sample-distances.

Figure S3.

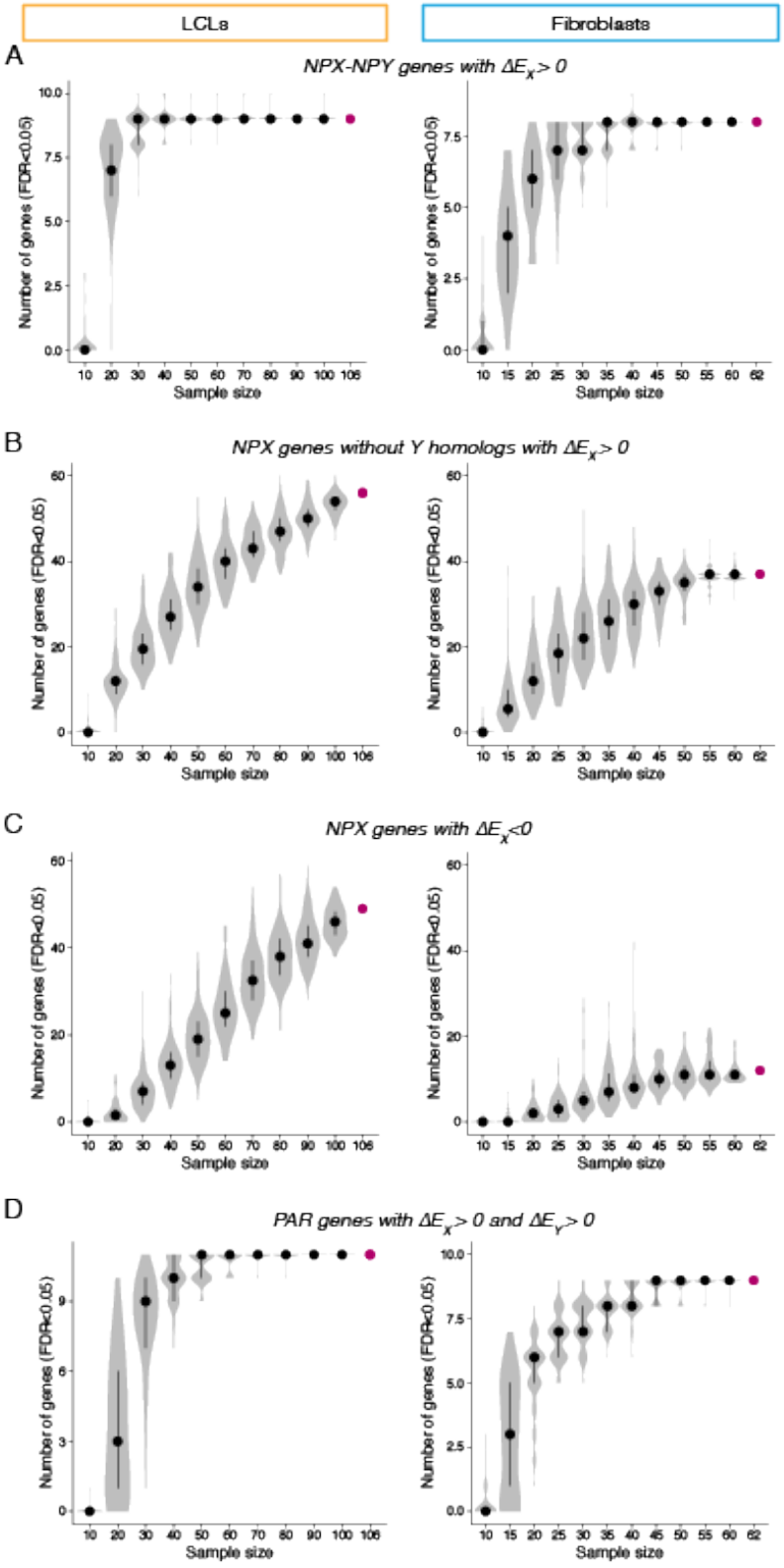
LCLs
Fibroblasts
A
NPX-NPY genes with $\Delta E_X > 0$
B
NPX genes without Y homologs with $\Delta E_X > 0$
C
NPX genes with $\Delta E_X < 0$
D
PAR genes with $\Delta E_X > 0$ and $\Delta E_Y > 0$
Number of genes (FDR<0.05)
Sample size

Figure S3, continued.

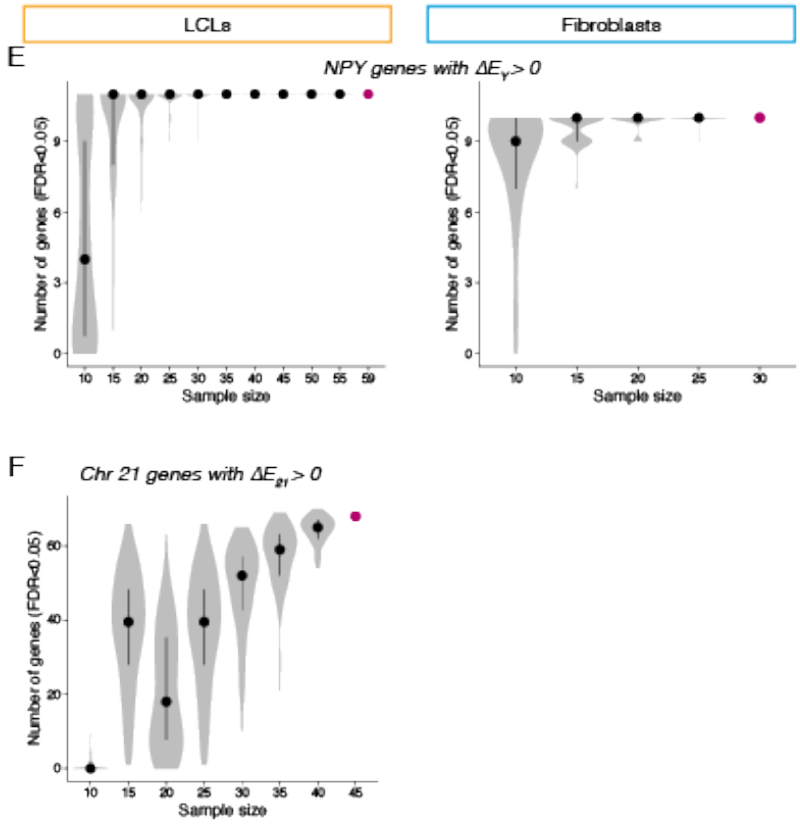

Figure S3. Bootstrapping analyses reveal that few additional significant genes would be observed with a larger sample size. Each analysis was performed by randomly choosing the indicated number of samples and performing the linear regression analysis, repeated 100 times for each sample size. The magenta point in each figure represents the number of significant genes in the final analysis using all of the samples. **(A)** NPX genes with Y homologs with $\Delta E_X>0$ reach saturation by a sample size of 40. (**B)** For NPX genes without Y homologs, the number of genes with $\Delta E_X>0$ increases rapidly at low sample sizes, and then levels off. **(C)** The trajectory of NPX genes with $\Delta E_X<0$ grows more slowly compared to genes with $\Delta E_X>0$, and appears to still be increasing in LCLs, while it has levelled off in fibroblasts. **(D)** PAR genes with both $\Delta E_X>0$ and $\Delta E_Y>0$ reach saturation by sample size of 50. **(E)** NPY genes with $\Delta E_Y>0$ reach saturation after 25 samples. **(F)** The trajectory for Chr 21 genes with $\Delta E_{21}>0$ increases rapidly between 10-30 samples, and then slows.

Fig. S4

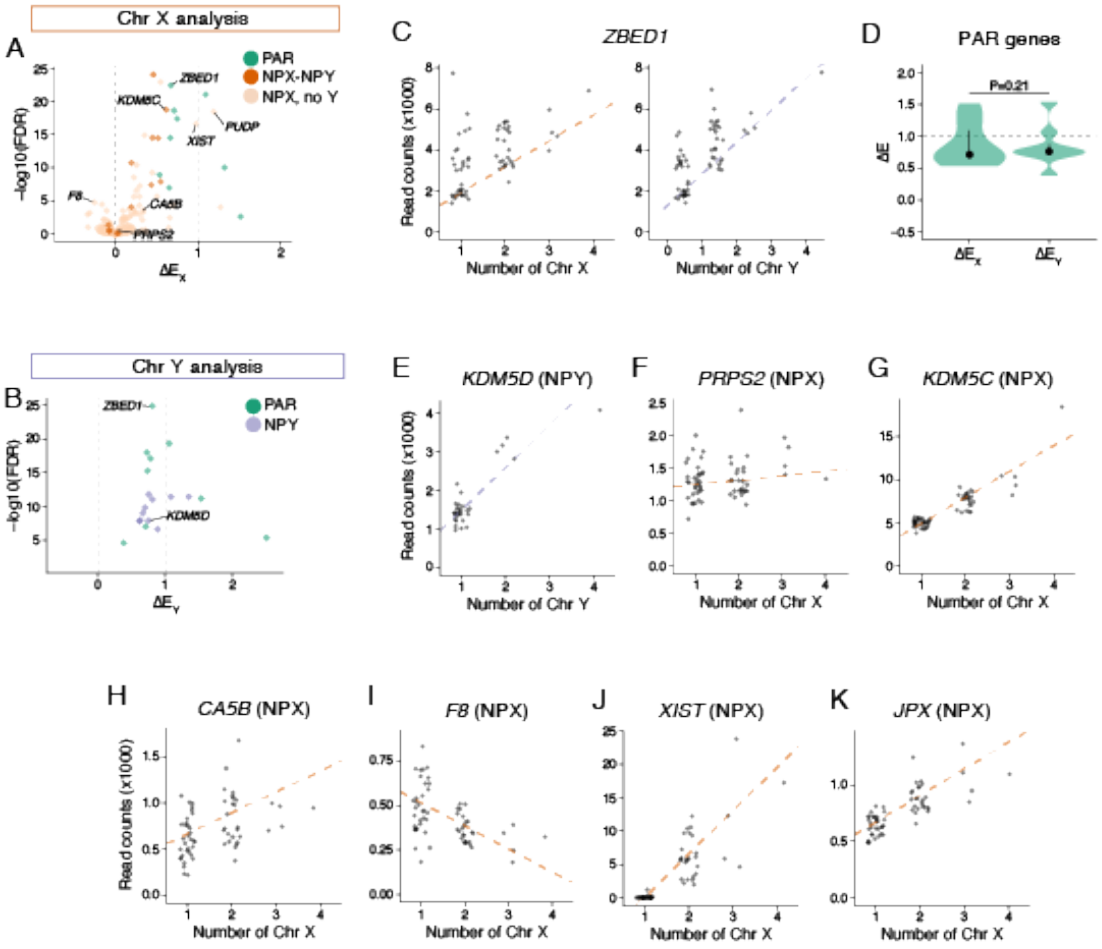

Figure S4. Gene expression analysis of sex chromosome aneuploid fibroblasts shows varying responses of PAR, NPX, and NPY genes to copy number changes. Volcano plots of ΔE_X **(A)** or ΔE_Y **(B)** versus significance show differences in the overall response of genes on Chr X or Y to increasing copy number. Scatterplots and regression lines for individual genes as a function of Chr X **(C, F-K)** or Y **(C, E)** copy number for all individuals in the sex chromosome dosage series. **(D)** Violin plots, with median (dots) and interquartile ranges (whiskers) indicated, comparing ΔE_X and ΔE_Y for PAR genes in fibroblasts. P-values, paired t-test. See **Fig. 3** for corresponding plots for LCLs.

Figure S5.

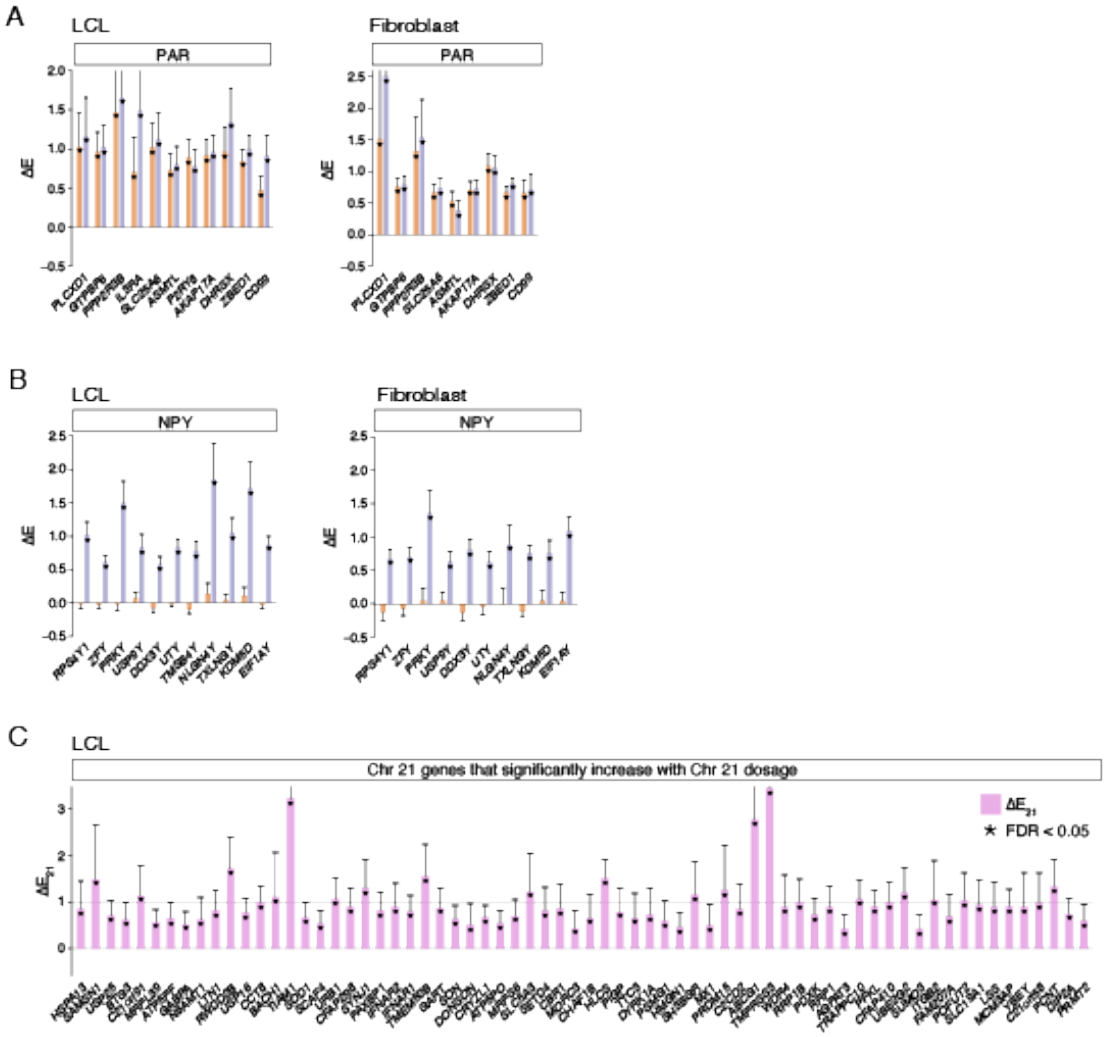
A
LCL
PAR
Fibroblast
PAR
ΔE
B
LCL
NPY
Fibroblast
NPY
ΔE
C
LCL
Chr 21 genes that significantly increase with Chr 21 dosage
ΔE21
ΔE21
★ FDR < 0.05

Figure S5, continued

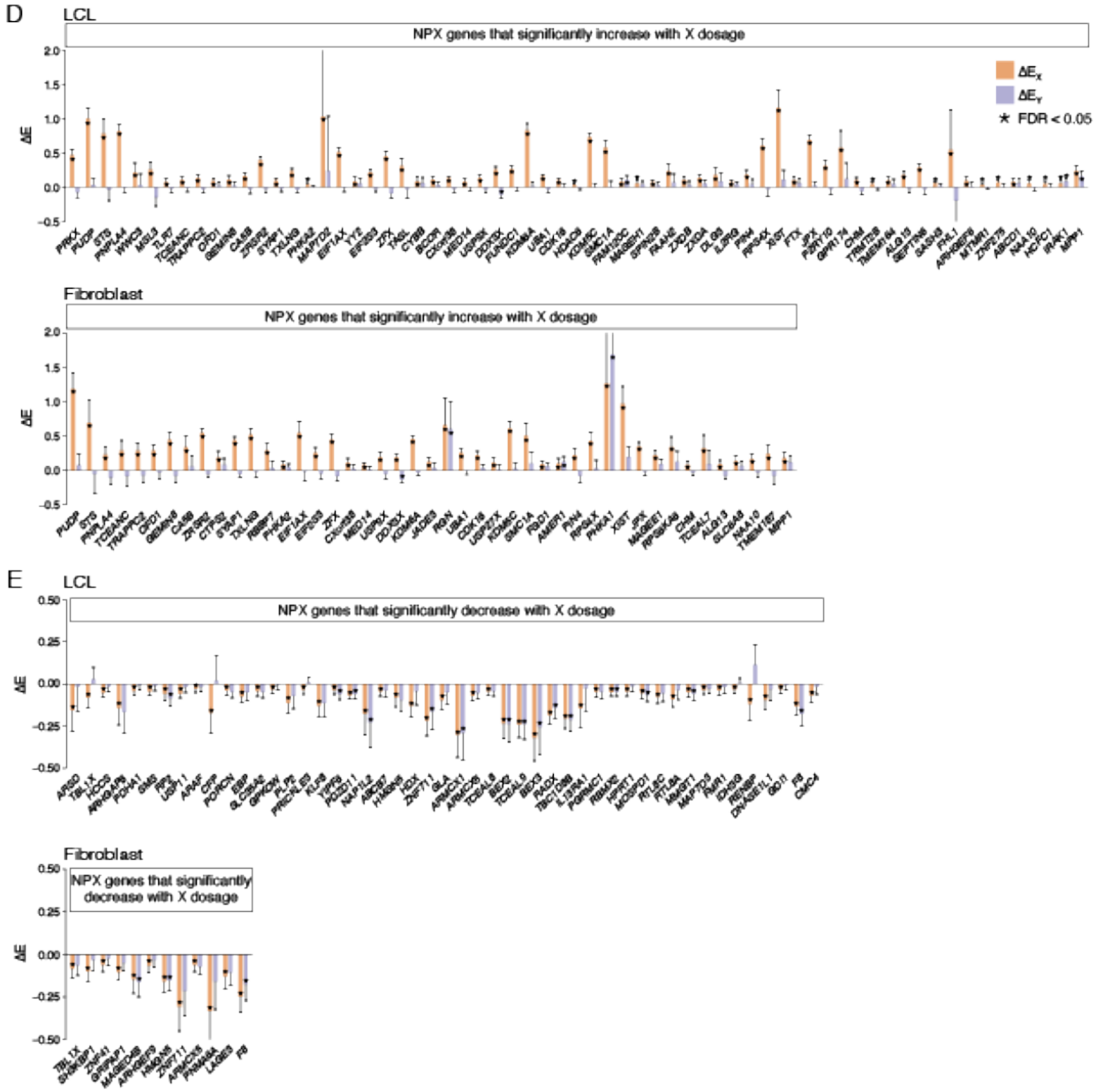

Figure S5. ΔE values for genes that are significantly up- or down-regulated with Chr X, Y, or 21 dosage. All plots show the normalized expression change per additional Chr X, Y, or 21 (ΔE_X, ΔE_Y, ΔE_{21}) for individual genes, plotted in genomic order from left to right. Asterisks indicate genes significantly differentially expressed (FDR<0.05). **(A)** All PAR genes significantly increase with both X and Y dosage. **(B)** All expressed NPY genes significantly increase with Y dosage, but not X dosage. **(C)** Chr21 genes that significantly increase with

Chr21 dosage in LCLs. **(D-E)** NPX genes that significantly increase **(D)** or decrease **(E)** with X chromosome dosage.

Figure S6

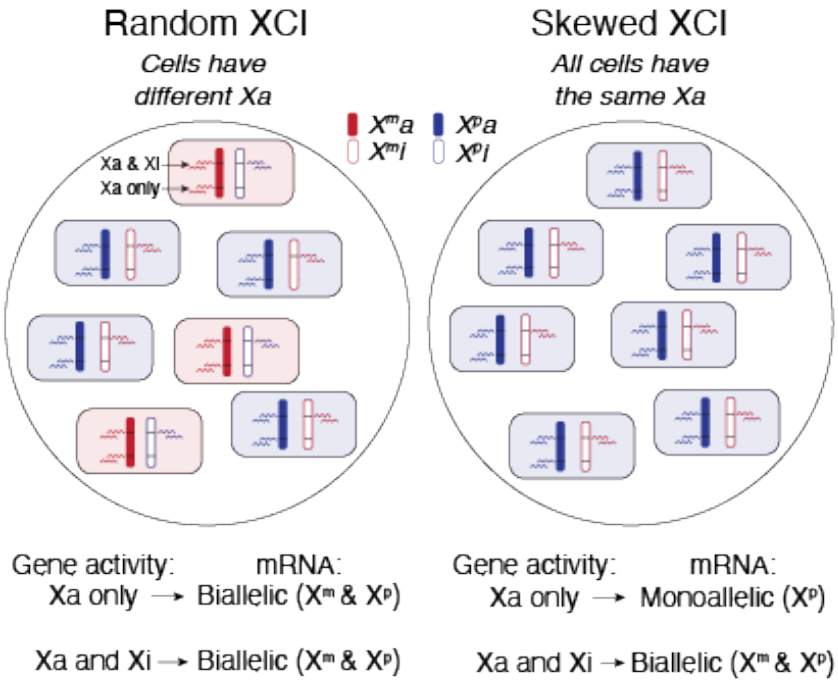

Figure S6. Cell lines with skewed XCI allow for identifying monoallelic versus biallelic expression. Schematic showing allele-specific expression results in 46,XX cell cultures with random XCI (left) or completely skewed XCI (right). In the random XCI culture, cells have either the maternal (X^m) or paternal (X^p) Chr X as their active X, meaning that even genes only expressed from Xa will appear biallelic. In the skewed XCI culture, all cells have X^p as their active X, so the resulting RNA will be monoallelic for genes expressed from Xa only, but biallelic for genes expressed from both Xa and Xi, allowing us to distinguish the two types of gene regulation.

Figure S7

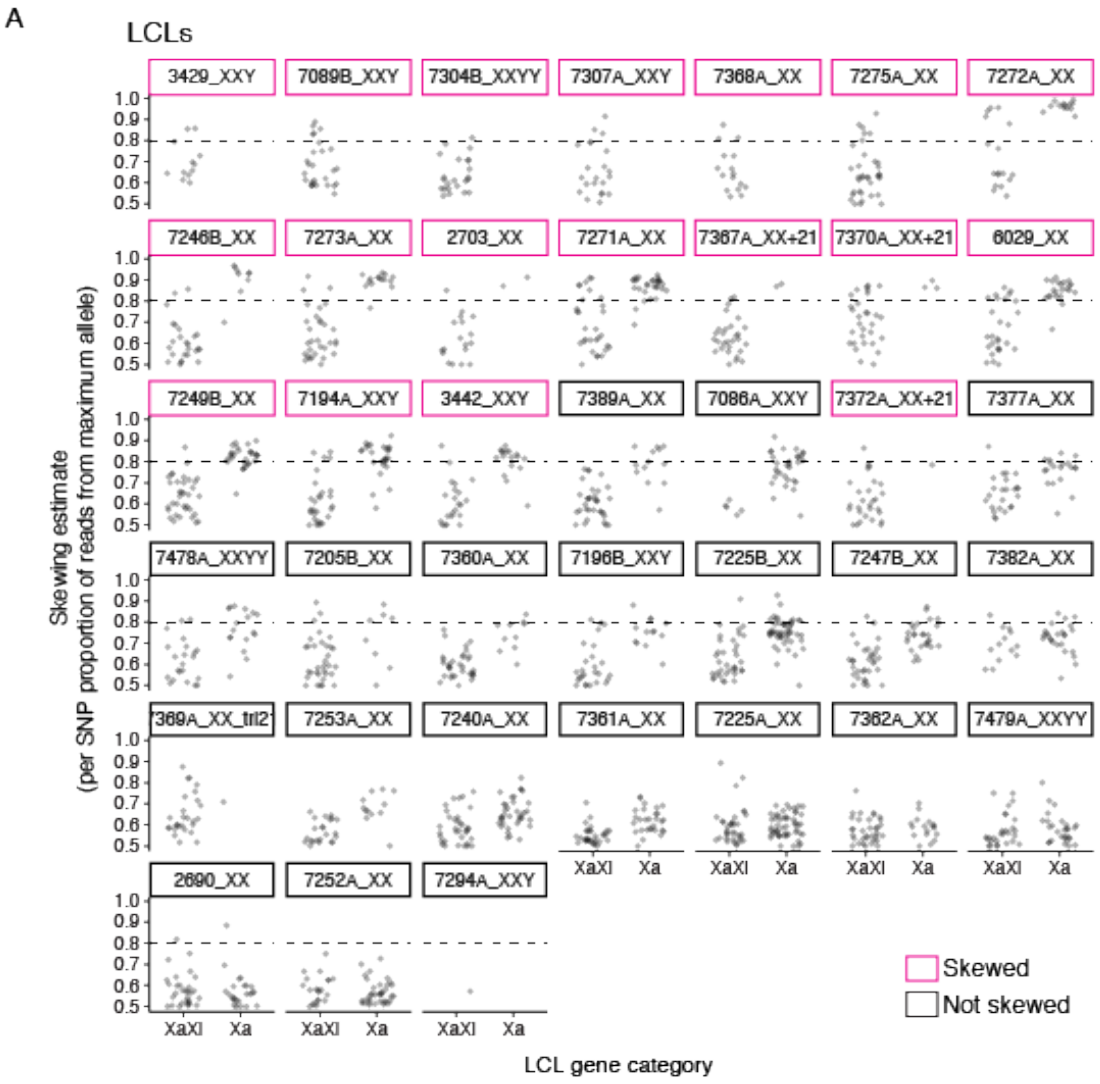
A
LCLs
3429_XXY
7089B_XXY
7304B_XXYY
7307A_XXY
7368A_XX
7275A_XX
7272A_XX
7246B_XX
7273A_XX
2703_XX
7271A_XX
7367A_XX+21
7370A_XX+21
6029_XX
7249B_XX
7194A_XXY
3442_XXY
7389A_XX
7086A_XXY
7372A_XX+21
7377A_XX
7478A_XXYY
7205B_XX
7360A_XX
7196B_XXY
7225B_XX
7247B_XX
7382A_XX
7369A_XX_tri2
7253A_XX
7240A_XX
7361A_XX
7225A_XX
7362A_XX
7479A_XXYY
2690_XX
7252A_XX
7294A_XXY
Skewing estimate
(per SNP proportion of reads from maximum allele)
1.0
0.9
0.8
0.7
0.6
0.5
XaXi
Xa
LCL gene category
Skewed
Not skewed

Figure S7, continued

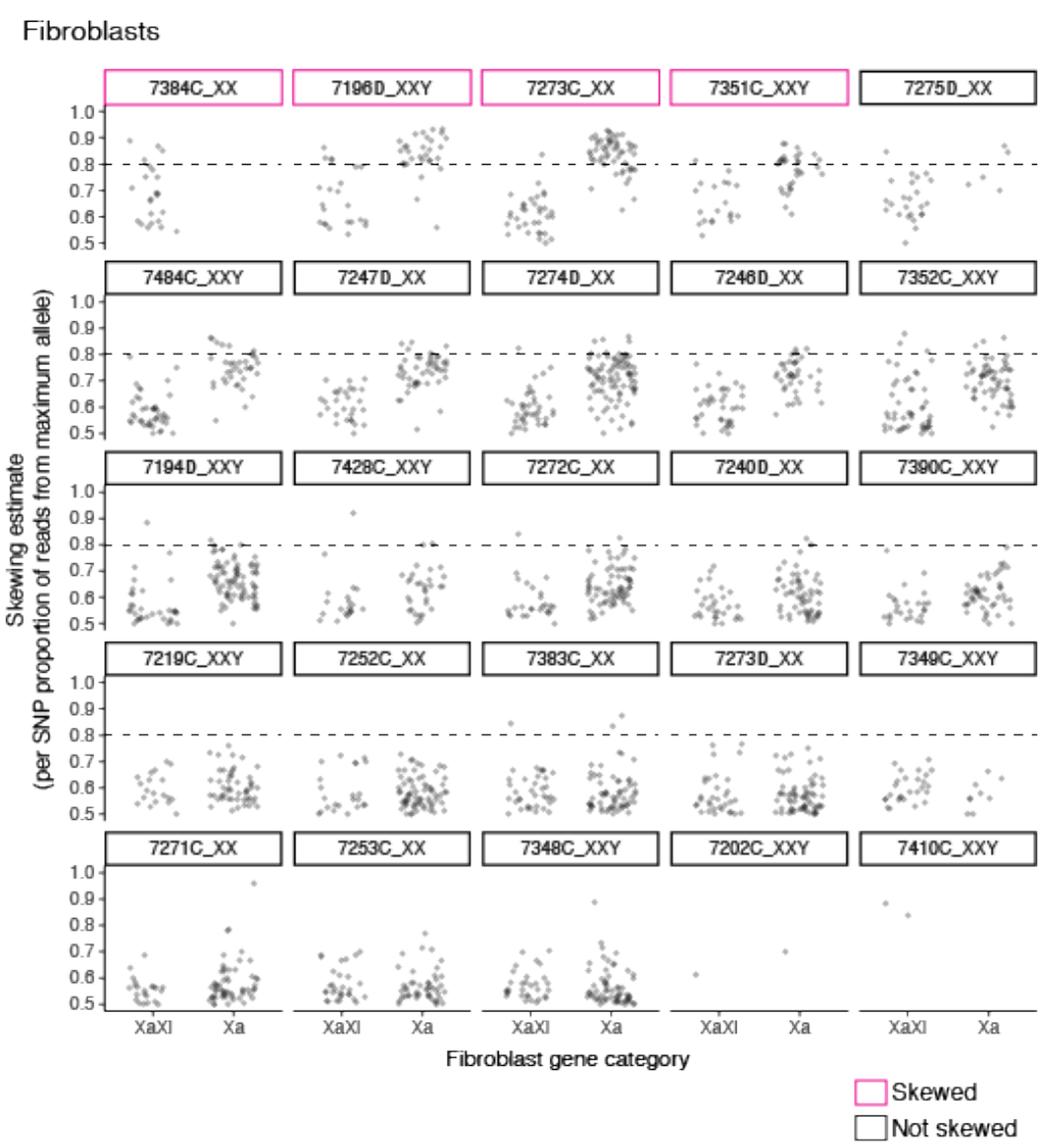

Figure S7. Identifying cell lines with skewed XCI using RNA-seq data. After identifying informative SNPs in RNA-seq data, we used SNPs within a stringent set of genes known to be expressed only from Xa to uncover the possibility of XCI skewing in our LCLs **(A)** or fibroblasts **(B)**. Samples were classified as skewed if they had few SNPs in Xa-only genes, or a median proportion of reads derived from the maximum allele among SNPs in Xa-only genes of 0.8. A set of genes that have been previously shown to be expressed from both Xa and Xi are included in these plots for comparison.

Figure S8

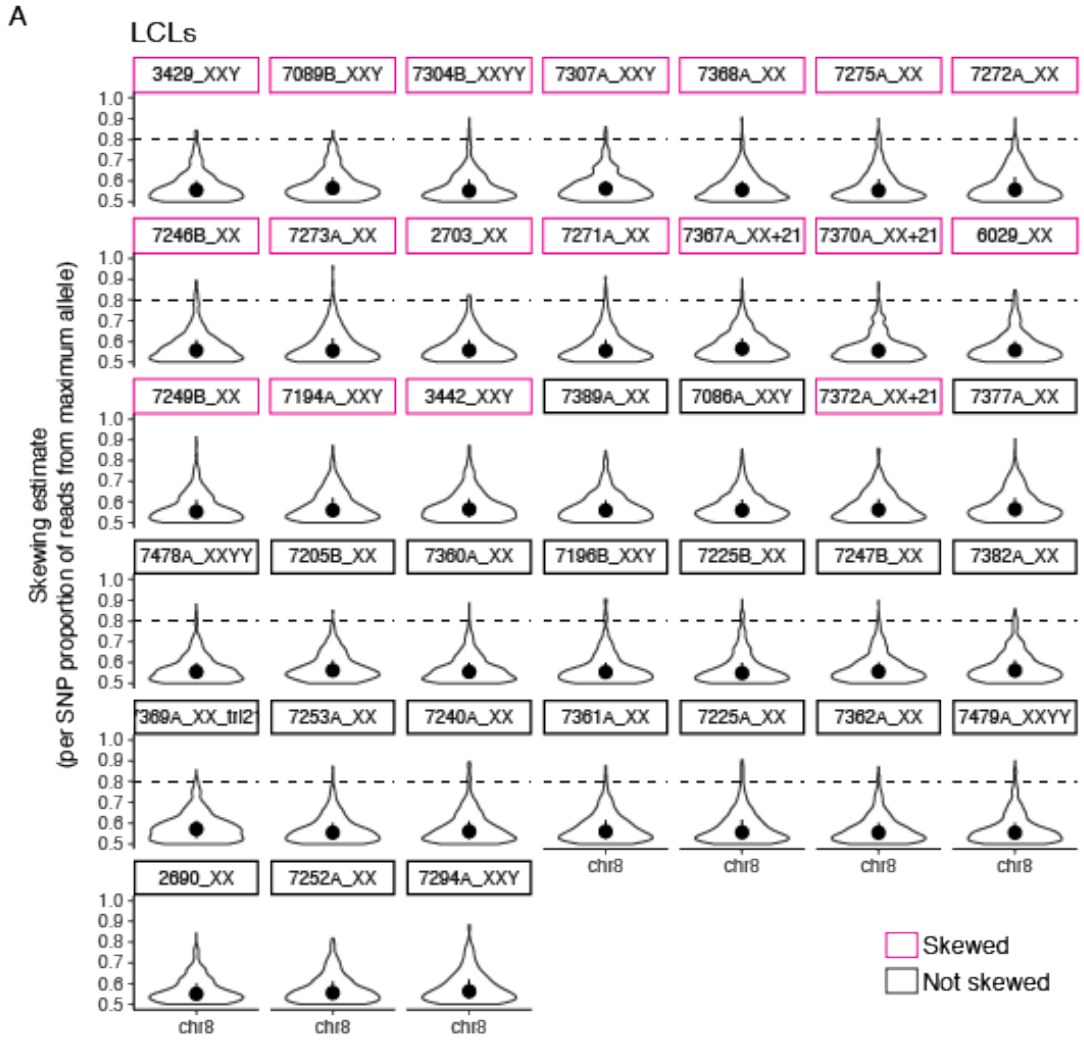
A
LCLs
3429_XXY
7089B_XXY
7304B_XXYY
7307A_XXY
7368A_XX
7275A_XX
7272A_XX
7246B_XX
7273A_XX
2703_XX
7271A_XX
7367A_XX+21
7370A_XX+21
6029_XX
7249B_XX
7194A_XXY
3442_XXY
7389A_XX
7086A_XXY
7372A_XX+21
7377A_XX
7478A_XXYY
7205B_XX
7360A_XX
7196B_XXY
7225B_XX
7247B_XX
7382A_XX
369A_XX_trl2
7253A_XX
7240A_XX
7361A_XX
7225A_XX
7362A_XX
7479A_XXYY
2690_XX
7252A_XX
7294A_XXY
Skewing estimate
(per SNP proportion of reads from maximum allele)
1.0
0.9
0.8
0.7
0.6
0.5
chr8
Skewed
Not skewed

Figure S8, continued

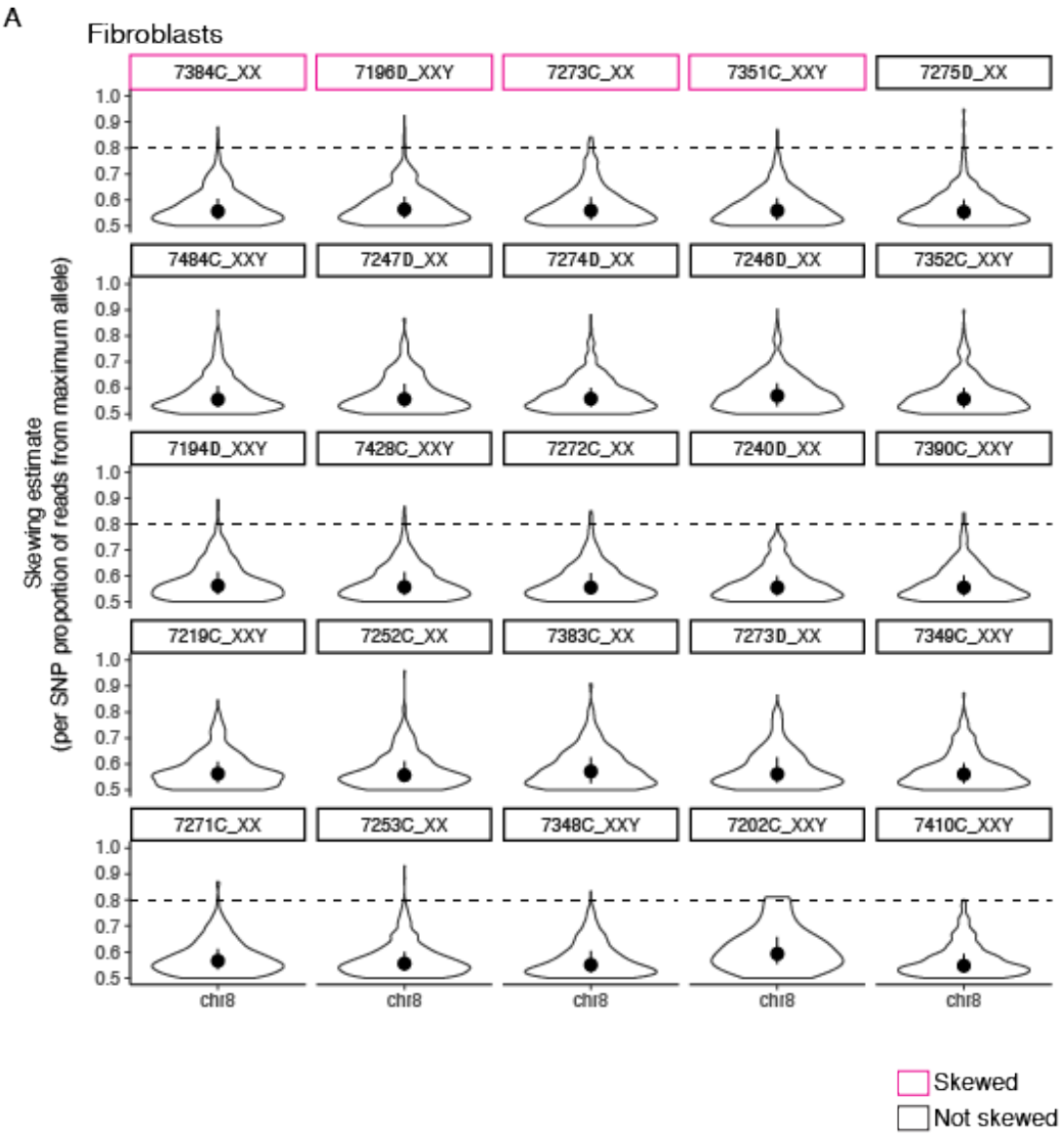

Figure S8. Autosomal SNPs do not show evidence of skewing. We performed a parallel skewing analysis using SNPs from genes on Chr 8, which has a similar number of expressed genes in these cell types as Chr X. We uniformly observed allele proportions near 0.5 across LCL and fibroblast samples.

Figure S9

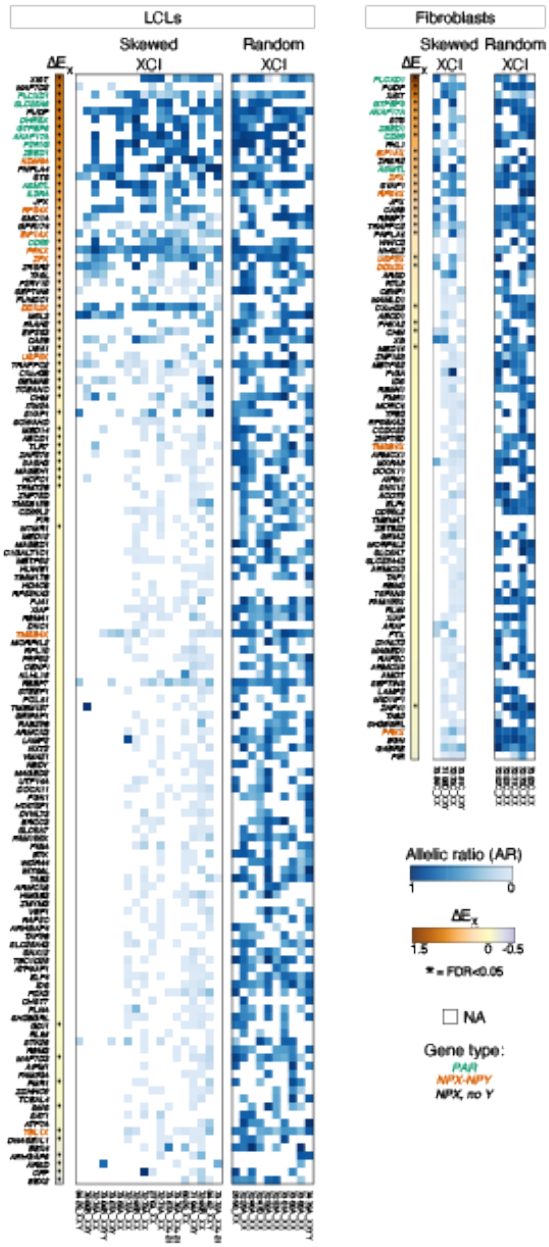

Figure S9. Allelic ratios across cell lines with skewed XCI for informative genes. Heatmaps of ΔE_X and the average allelic ratio (adjusted for the extent of skewing in each cell line) for genes with at least three observations in LCLs or two observations in fibroblasts across skewed samples. Samples with random XCI are provided for reference, and show biallelic expression (blue) across all genes. Heatmaps are ordered by ΔE_X values from highest to lowest.

Figure S10

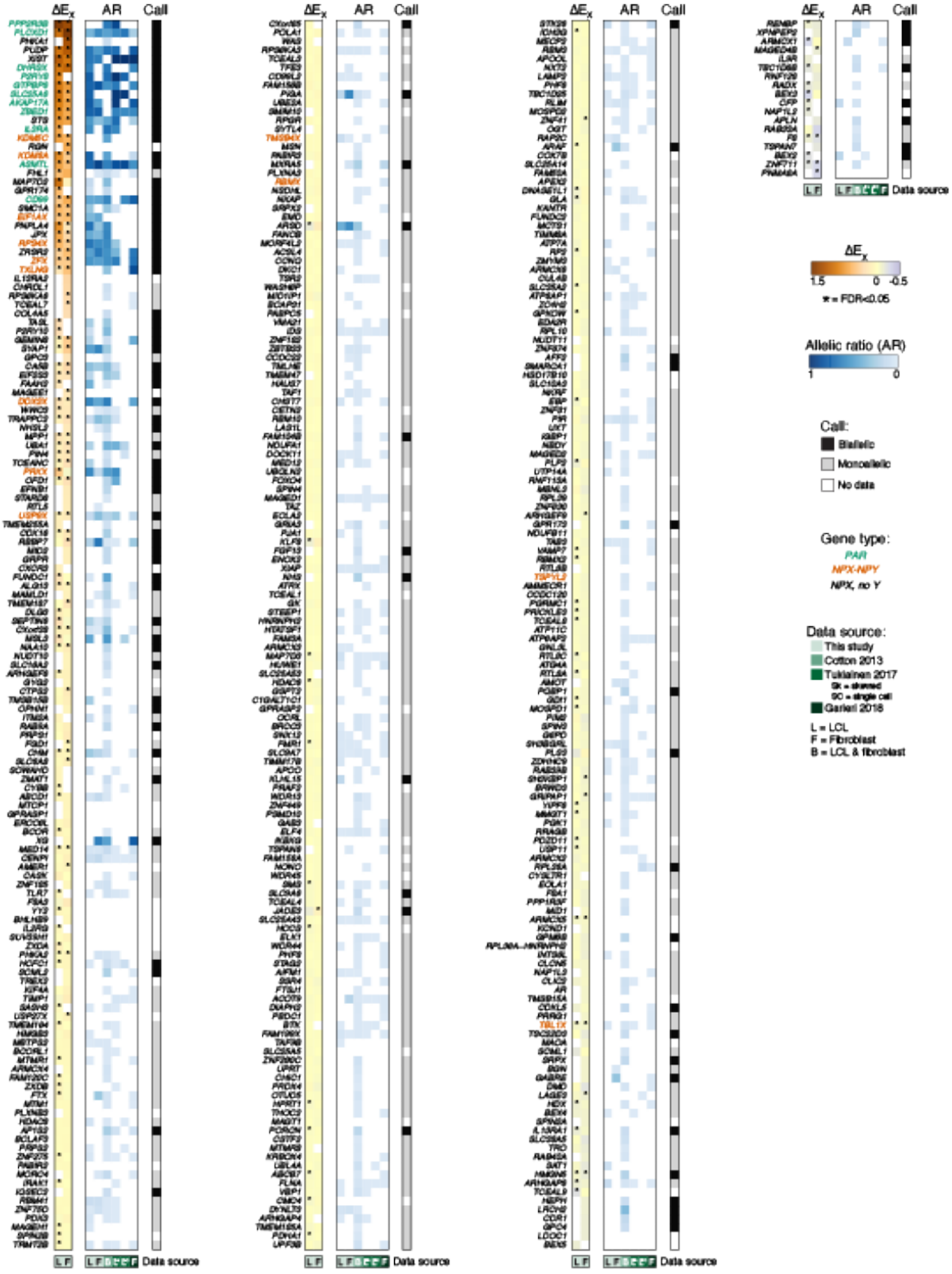

Figure S10. All AR data from this study and published datasets. Heatmaps for ΔE_X, individual AR values for each dataset, and our calls for monoallelic or biallelic expression for genes expressed in LCLs or fibroblasts. Rows are ordered by median AR across datasets.

Figure S11.

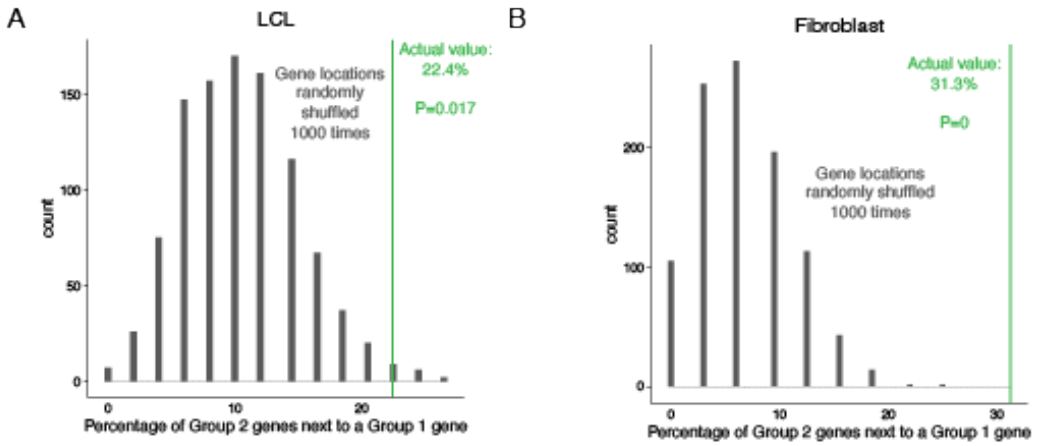

Figure S11. NPX genes without a Y homolog that have $\Delta E_X > 0$ are more likely to be adjacent to NPX-NPY genes with $\Delta E_X > 0$ than by chance. Histogram showing that the percentage of Group 2 genes (NPX genes without a Y homolog and $\Delta E_X > 0$) that are adjacent to Group 1 genes (NPX-NPY genes in humans or other eutherian mammals with $\Delta E_X > 0$) is larger for our actual data (green) compared to 1000 random shufflings of the gene positions.

Figure S12.

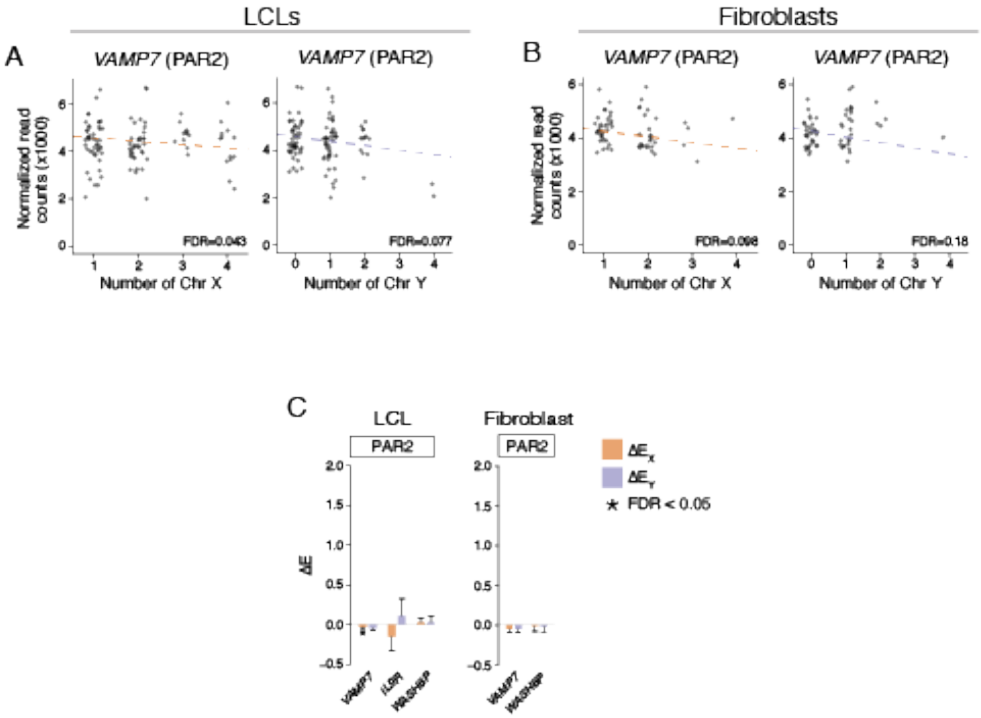

Figure S12. Expression of PAR2 genes does not change with additional copies of Chr X or Y. (A-B) Scatterplots showing normalized read counts for PAR2 gene *VAMP7* as a function of the number of copies of Chr X or Y for all LCL **(A)** or fibroblast **(B)** samples. **(C)** Boxplots of the relative change in expression with each X or Y chromosome indicates that PAR2 genes are exclusively expressed from the first X chromosome.

Captions for Supplemental Tables

All tables are contained as individual tabs in a single .xlsx file.

Table S1.

Features and metrics of NPX and PAR genes

Table S2.

Features and metrics of NPY genes

Table S3.

Metadata for euploid and aneuploid samples with RNA-sequencing

Table S4.

Linear regression results for expressed PAR and NPX genes in LCLs

Table S5.

Linear regression results for expressed PAR and NPX genes in fibroblasts

Table S6.

Linear regression results for expressed NPY genes in LCLs

Table S7.

Linear regression results for expressed NPY genes in fibroblasts

Table S8.

Linear regression results for expressed Chr 21 genes in LCLs

Table S9.

Chr X SNPs called in RNA-sequencing reads from samples with two X chromosomes

Table S10.

Chr 8 SNPs called in RNA-sequencing reads from samples with two X chromosomes

Table S11.

Allelic ratio per skewed LCL sample for informative Chr X genes

Table S12.

Allelic ratio per skewed fibroblast sample for informative Chr X genes

Table S13.

Meta-analysis of allelic ratios in informative Chr X genes from our study and published datasets

Table S14.

NPX-NPY pairs in mammals

Table S15.

Clusters of significantly up-regulated NPX genes in LCLs and fibroblasts

Table S16.

Genes adjacent to NPX-NPY homologs that have $\Delta E_X > 0$

Table S17.

Genes used in evaluation of XCI skewing in cell cultures.

Table S18.

Skewing coefficients and confidence intervals for skewed samples.

Supplemental References

49. J. L. Mueller *et al.*, Independent specialization of the human and mouse X chromosomes for the male germ line. *Nat Genet* **45**, 1083-1087 (2013).

50. E. K. Jackson *et al.*, Large palindromes on the primate X Chromosome are preserved by natural selection. *bioRxiv*, (2021).

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

52. S. Naqvi *et al.*, Conservation, acquisition, and functional impact of sex-biased gene expression in mammals. *Science* **365**, 249-+ (2019).

53. K. D. Pruitt *et al.*, The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res* **19**, 1316-1323 (2009).

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

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

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

57. M. E. Ritchie *et al.*, limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).

58. A. Ciccodicola *et al.*, Differentially regulated and evolved genes in the fully sequenced Xq/Yq pseudoautosomal region. *Hum Mol Genet* **9**, 395-401 (2000).

59. W. A. Bickmore, H. J. Cooke, Evolution of homologous sequences on the human X and Y chromosomes, outside of the meiotic pairing segment. *Nucleic Acids Res* **15**, 6261-6271 (1987).

60. F. Rouyer *et al.*, A gradient of sex linkage in the pseudoautosomal region of the human sex chromosomes. *Nature* **319**, 291-295 (1986).