SUPPLEMENTARY INFORMATION

Sex Chromosome Dosage Effects on Gene Expression in Humans

Armin Raznahan[1], Neelroop N. Parikshak[2], Vijay Chandran[3], Jonathan D. Blumenthal[1], Liv S. Clasen[1], Aaron F. Alexander Bloch[1], Andrew Zinn[4], Danny Wangsa[5], Jasen Wise[6], Declan G. M. Murphy[7], Patrick F. Bolton[7], Thomas Ried[5], Judith Ross[8], Jay N. Giedd[9], Daniel H. Geschwind[2]

Correspondence to: Armin Raznahan, Chief Developmental Neurogenomics Unit, Human Genetics Branch, National Institute of Mental Health, NIH, Bethesda, MD, USA, raznahana@mail.nih.gov

Contents:

SUPPLEMENTARY TEXT

Supplementary Text S1. Pre-processing of microarray data.

Gene expression was profiled using the Illumina HT-12 v4 Expression BeadChip Kit (Illumina Inc, San Diego, CA). Expression data were quantile normalized across arrays, and log2 transformed using the *limma* package in R (1). For each of 47323 probes, we estimated mean expression by karyotype group, and log2 expression fold change for each unique pairwise group contrast between karyotype groups (**Fig. 1a**). Estimation of expression fold-change differences between groups were conducted using standard robust linear regression methods as implemented by the lmFit function in *limma*. For each group contrast, differentially expressed probes survived correction for multiple comparison across all probes, with q (the expected proportion of falsely rejected nulls) set at 0.05, and showed a |log2| fold change greater than 0.26. This log 2 cut-off was selected empirically, by defining the log 2 fold increment associated with the greatest drop in DEG count for each SCA group, and averaging these thresholds all 5 groups (**SI Fig S3**). Probes were annotated using both the vendor manifest file and an independently published re-annotation that assigns a quality rating to each probe based on the specificity of its alignment to the purported transcript target (2). We filtered for all probes with “perfect” or “good” quality alignment to a known gene according to this reannotation, and then used the *collapseRows* function from the WGCNA (3) package in R (with default settings), to select one probe per gene. We also applied a further filter to remove any Y-linked probes that showed differential

expression between female karyotype groups. These steps resulted in high-quality measures of expression and estimates of differential expression for 19984 autosomal and 894 sex-chromosome genes in each of 68 independent samples from 7 different karyotype groups.

Supplementary Text S2. Supplementary methods for qRT-PCR Validation.

Selection of genes for qRT-PCR validation: We selected genes for qRT-PCR validation as follows. From the set of 10 sex-linked genes with patterns of "obligate" dosage sensitivity (**Fig 1b**), we selected XIST, the 2 most X-chromosome sensitive X-linked gametologs [EIF1AX, KDM6A(UTX)], and the 2 most Y-chromosome sensitive Y-linked gametologs (ZFY, DDX3Y). From the sets of dosage sensitive sex chromosome genes defined by k-means clustering (**Fig. 2a,b**), all genes selected for qPCR from showed (i) stable cluster membership in >95% of bootstrap draws (**SI Appendix Fig. S2b**), and (ii) consistent inclusion in the top 10 DEGs across multiple relevant group contrasts for that k-means cluster (Pink Y-linked cluster: XYY vs. XY and XXYY vs. XXY | Yellow XCI and Green XCIE clusters: XO vs. XXX, XO vs. XY, XXY vs. XX).

Fluidigm qRT-PCR protocol: Reverse Transcription reaction was performed using RT2 HT First Strand Kit (QIAGEN, 330411) with 1000 ng RNA input per sample. One-tenth of cDNA was preamplified using RT2 Microfluidics qPCR Reagent system (QIAGEN, 330431) in combination with custom RT2 PreAmp pathway primer mix Format containing 94 RT2 primer assays. Fourteen cycles of preamplification were performed using the manufacturer recommended

preamplification protocol. Amplified cDNA was diluted 5-fold using RNase-free water and assessed in real-time PCR using the RT2 Micrifluidics EvaGreen qPCR Master mix and a Custom RT2 profiler PCR array PCR Array containing 96 assays, including selected DEGs of interest, housekeeping genes, reverse-transcription controls, and positive PCR control. Real-time PCR was performed on a Fluidigm BioMark HD (Fluidigm, San Francisco, US) using the RT2 cycling program for the Fluidigm BioMark, which consists of an initial thermal mix stage (50°C for 2 minutes, 70°C for 30 minutes, and 25°C for 10 minutes) followed by a hot start at 95°C for 10 minutes and 40 cycles of 94°C for 15 seconds, and 60°C for 60 seconds. For data processing, an assay with CT > 23 was deemed to be not expressed.

Differential Expression Analysis of qRT-PCR data: The ΔΔCT method of relative quantification was used to analyze qRT-PCR data (4). To provide normalized estimates of expression for each gene we calculated ΔCT values, by subtracting the CT for each gene of interest from the mean CT of two housekeeping genes (GAPDH and RPLP0) which were not differentially expressed across groups in either microarray or qRT-PCR data. Thus, larger ΔCT values reflected greater normalized expression relative to mean expression of the reference housekeeping genes. These ΔCT values were used as input for calculation of all unique pairwise group differences in expression between karyotype groups represented in the independent qRT-PCR validation dataset. Group differences in expression were modeled using the limma R package with identical setting to those used in analysis of microarray data (see above). The resulting ΔΔCT represent fold-changes in

gene expression between groups, on a log scale with base determined by the effective qRT-PCR efficiency.

Validation Microarray Results Using qRT-PCR Results: All 21 unique pairwise SCD group contrasts in our microarray sample were also represented in the independent qRT-PCR dataset. We used the correlation across these 21 group contrasts for the qRT-PCR fold-change and microarray log2 fold change to quantify the degree of agreement between qRT-PCR and microarray findings (**SI Appendix Fig. S1a** and **SI Appendix Fig. S2e,f**).

Extension of Microarray Results Using qRT-PCR Results: The qRT-PCR dataset also included two SCD groups that were not represented in the microarray dataset – XXXY and XXXXY – allowing for a total of 15 novel pairwise SCD group contrasts ("XO-XXXY", "XO-XXXXY", "XXXY-XX", "XXXXY-XX", "XXX-XXXY", "XXX-XXXXY", "XXXY-XY", "XXXXY-XY", "XXY-XXXY", "XXY-XXXXY", "XYY-XXXY", "XYY-XXXXY", "XXYY-XXXY", "XXYY-XXXXY", "XXXY-XXXXY") sampling diverse disparities of X- and Y-chromosome dosage. These novel contrasts were used as a further test for the validity and reproducibility of our microarray findings. Each of the 15 novel pairwise SCD group contrasts was coded according to two effects of interest: difference in X-chromosome count and difference in Y-chromosome count. These coded SCD disparities were then correlated with observed fold-changes for unique pairwise group contrasts in the qRT-PCR dataset to test if patterns of fold-change observed in the microarray dataset could be extended into unseen karyotype groups (**SI Appendix Fig. S1**: X- and Y-linked genes with “obligatory” sex chromosome dosage sensitivity | **SI**

Appendix Fig. S2e-h: sex-linked genes from the Y-, XCIE- and XCI-enriched k means clusters that countered expectations of the classical Four Class Model).

Supplementary Text S3. Testing the Four Class Model

A priori classification of genes into the Four Class Model: PAR genes were defined as those lying distal to the PAR1 and PAR2 boundaries specified in hg18 build of the human genome. Y-linked and X-linked genes were defined as those lying proximal to these PAR boundaries on the Y- and X-chromosome respectively. X-linked genes were assigned to XCIE and XCI classes using consensus classifications from a systematic integration of XCI calls from 3 methodologically distinct large-scale surveys of the X-chromosome (5). According to the XCI categories of this consensus report we classified X-linked genes as X-inactivated ("Subject" or "Mostly Subject" categories), X-escape ("Escape" or "Mostly Escape" categories), or X-other (all other intermediate categories).

Unsupervised clustering of genes by dosage sensitivity: Mean expression values were calculated for all sex chromosome genes per karyotype group, and normalized to their mean expression across all SCD groups. The resulting 894 by 7 matrix was submitted to k-means clustering across a range of k-values using the *kmeans* function in R with nstart and iter.max set at 100. Scree plot inspection indicated an optimal 6-cluster solution (**SI Appendix Fig. S2a**), consisting of 5 clusters of expressed genes with SCD sensitivity, and a remainder cluster of genes with low or undetectable expression levels across all samples. Reproducibility of gene assignment to the 5 expressed clusters (**SI Appendix Fig. S2b**), and the

single large cluster of genes with low/undetectable expression by microarray (**SI Text S4**) was established by re-running k-means 1000 times with samples that were randomly drawn (with replacement) from each SCD group. Classifications of genes by the Four Cluster Model and k-means were compared using two-tailed Fishers tests for all pairwise cluster-grouping combinations (**Fig. 2b**).

Modelling X- and Y-chromosome dosage effects on k-means cluster expression. The mean expression of XCIE and XCI k-means clusters, and of individual genes within each cluster were related to X- and Y- chromosome dosage variation across samples using the following linear model: [Expression = β_0 (intercept) + β_1(X_count) + β_2(Y_count) +e]. Significant X- and Y-dosage effects were given by null hypothesis testing of β_1 and β_2 coefficients. We also estimated pairwise karyotype group differences in the mean expression of each k-means cluster (**Dataset S3**). Power calculations using the pwr package in R (6) indicated that these analyses were powered to detect a two-tailed group difference in expression of effect size > 1.32 at a significance level of 0.05 and a power of 0.8.

Comparing k-means clusters to genetic, evolutionary, and epigenetic X-chromosome annotations. The GenomicRanges package in R was used to align probes for genes in XCIE and XCI k-means clusters with independently published X-chromosome annotations for (i) “chromatin states” defined by computational analysis of coordinated changes in 10 distinct chromatin marks in LCLs (7), and (ii) “evolutionary strata” reflecting staged loss of recombination between the X- and Y-chromosome (8). We also aligned our clustering of X-lined genes with a previously published annotation of X-linked genes according to whether their

corresponding ancestral Y-linked homologue has been lost, converted to a pseudogene or maintained (5). Non-random associations between these three annotations and XCIE-/XCI-cluster membership were assessed using Chi-squared tests.

Supplementary Text S4. Reproducibility and convergent validity of sex chromosome gene grouping detected by k-means clustering analysis.

K-means partitioning of sex chromosome genes using mean expression in each of the 7 karyotype groups in our microarray dataset (normalized to mean expression over all samples) defined a large cluster of 773 sex chromosome genes, which were characterized by low/undetectable expression across samples (median detection rate of 4/68 samples by microarray). Mean expression of this cluster was not significantly associated with SCD ($p=0.7$ in F-test for omnibus effects of karyotype group on mean cluster expression, **SI Appendix Fig. S2c**). We took several steps to verify the reproducibility and convergent validity of this large cluster of genes with low/undetectable expression and little sensitivity to SCD.

First, we established that this large cluster was reproducible in our own data. We tested for reproducibility by repeating k-means clustering across 1000 sets of 68 samples generated by separately resampling individuals with replacement from each SCD group. Eighty-six percent of genes within this cluster (i.e. 665/773) maintained their cluster assignment in 90 percent or more of the 1000 bootstrap k-means analyses.

Second, we verified that genes assigned to this cluster show low/undetectable expression in LCLs using publicly-available RNA-seq measures of gene expression from 343 LCLs in the Genotype-Tissue Expression (GTEx) database. Seventy-five percent of the 773 genes within the low/undetectable expression cluster, also showed low expression – i.e. median Transcripts Per Million (TPM) < 10 - across the GTEX LCL set by RNA-seq. The 773 gene cluster detected by k-means analysis was also specifically and significantly enriched for genes with undetectable expression in GTEX data (i.e. TPM<0.1, enrichment odds ratio = 101, p<2.2*10-16).

Third, we also used GTEx data to verify that this cluster was significantly enriched for the set of 268 sex chromosome genes that show detectible levels of expression in LCLs (TMP>0.1 in GTEx), but lacked sensitivity to SCD (p>0.05 in F-test for omnibus effects of karyotype group on expression). Ninety-five percent of these 268 genes fell within the 773 cluster, representing a statistically-significant enrichment (Fisher test, odds ratio = 4.5, p=3 *10-8). Collectively, the above three sets of analyses verify that detection of this large gene cluster is not only reproducible within our own dataset, but converges with independent RNA-seq annotations of genes with little or no expression in LCLs.

We also tested and confirmed that repeating k-means analysis of microarray data - after first filtering out all genes that lacked detectible expression levels in LCLs by RNA-seq in the GTEx dataset - could recover the observed grouping of sex chromosome genes by our original k-means analysis. Specifically, repeating k-means on this GTEx-filtered subset partitioned genes into separate clusters which

each showed a unique, specific and statistically significant overlap with one of the gene clusters defined in our original unfiltered k-means analysis (all pairwise Fishers tests p< 4.5*10-8). Furthermore, the XCI status of X-linked genes in a published analysis of GTEx data was strongly associated with the grouping of genes by SCD-sensitivity profile in both our original k-means analysis (Chi-squared statistic=47, p=1.7*10-8) and the "GTEx-expressed k-means" analysis (Chi-squared statistic=44, p=9.2*10-8).

Finally, we confirmed that the observed clustering of sex-chromosome genes by k-means analysis was also reproducible if expression data for each SCA group was normalized to that in its respective gonadal control group (vs. normalization to expression across the full sample as used in our main analyses). The 6 clusters defined by k-means using this alternatively normalized data fully converged with our original k-means analysis by assigning XIST to its own "cluster" and defining 5 other gene clusters which each showed a unique, specific and statistically significant overlap with one of the gene clusters defined in our original k-means analysis (all pairwise Fishers tests $p< 8.9*10^{-15}$).

Supplementary Text S5. Assessing alternative explanations for the non-canonical patterns of dosage sensitivity displayed by Green (XCIE-enriched) and Yellow (XCI-enriched) clusters of dosage sensitive sex chromosome genes.

We observed that mean expression of XCIE-enriched cluster genes scaled sub-linearly with X-chromosome dosage, and sought to test if this observation could

potentially be explained by the recognized phenomenon of partial escape from XCI in genes known to be expressed from the inactivated X-chromosome. Seven of the 39 XCIE-cluster genes were included in a recently published survey of partial escape from X-inactivation based on allelic expression imbalance analysis in LCLs from a female with skewed X-chromosome inactivation (9). For each of these 7 genes, this survey provided details regarding the proportion of total expression in XX LCLs that derived from the inactive X-chromosome. We used these published proportions to compute gene-specific predictions for the magnitude of log 2 fold change in expression expected between XXY LCLs and XY LCLs – assuming that the ratio between expression of these genes from the inactivated second X-chromosome and active X-chromosome in XXY cells similar to the similar ratio between the inactivated and activate X-chromosome in XX cells. For 6 of the 7 genes examined, predicted log2 fold changes were in line with those observed in our microarray data (**SI Appendix Fig. S2c**) – suggesting that the partial XCIE may explain observed sublinear relationship between mean expression of XCIE-cluster genes as a group and X-chromosome count (**Fig. 2b**).

Supplementary Text S6. Testing if the observed magnitude of inverse X-chromosome dosage effects on expression of XCI-cluster genes is greater than chance, after controlling for sample size and gene expression variability within our microarray dataset.

The inverse effect of X-chromosome dosage on XCI-enriched cluster expression involved relatively modest log2 fold-change values (i.e. XO vs. XXX differences,

Fig 2d, Dataset S3). We therefore sought to test if the magnitude of these fold-changes was significantly greater than a distribution of "negative control" fold-change values generated from comparison of XX and XY controls within our own microarray dataset. To conduct this test we first identified a "control set" of 388 X-linked genes within our microarray dataset that had been independently annotated (9) as failing to show significant XX-XY expression differences based on RNA-seq analysis of gene expression in 343 LCL lines from the Genotype-Tissue Expression (GTEx) database. Next, we estimated XX-XY differences in the expression of these 388 genes across 10,000 XX-XY contrasts based on resampling XX and XY individuals from our microarray dataset with replacement. Then, for each of the 66 X-linked genes within the dosage sensitive XCI-enriched cluster, we compared the magnitude of expression fold-change observed in XO vs. XXX samples with the null distributions of 1000 fold-change values for all 388 X-linked genes within the "control set" from GTEx. For 59/66 (89%) X-linked genes in the XCI-enriched cluster, the average centile for observed fold-change in the XO vs. XXX contrast - relative to null-fold change distributions for the 388 "negative control set" genes - was >95 (i.e. mean empirical p value <0.05).

Supplementary Text S7. Weighted Gene Co-expression Network Analysis (WGCNA)

Defining Gene Co-expression Modules: Signed gene co-expression modules were generated using the R package Weighted Gene Co-expression Network Analysis (WGCNA). Briefly, this involved first calculating the Pearson correlation coefficient

between all 20978 genes across all 68 samples in our study. This correlation matrix was transformed using a signed power adjacency function with a threshold power of 12 (selected based on fit to scale-free topology), and then converted into Topological Overlap Matrix (TOM) by modifying the correlation between each pair of genes using a measure of the similarity in their respective correlations with all other genes (10). The resulting TOM was then converted to a distance matrix by subtraction from 1, and used to generate a dendrogram for clustering genes into modules. Gene modules were defined using the Dynamic Hybrid cutree function (11) [with the following parameter settings: deepSplit (control over sensitivity of module detection to module splitting) = 2, mergeCutHeight (distance below which modules are merged)= 0.25, minimum module size=30)]. Given the large number of genes included in our analyses, we implemented module detection using the "blockwise" WGCNA algorithm, which starts with a computationally inexpensive method (k-means) to assort genes into smaller co-expression blocks, and then completes the above steps within each block before merging module designations across blocks. This implementation of WGCNA defined 18 mutually exclusive co-expression gene modules within our data, which ranged from 45 to 1393 genes in size, and a left-over group of 14630 genes without module membership (**Dataset S5**). The expression of each module was summarized as a module eigengene value (ME: the right singular vector of standardized expression values for genes in that module) in every sample. These ME values were used to determine differential expression of modules across (omnibus F-tests) and between (T-tests) SCD

groups, as well as module co-expression across samples (Pearson correlation coefficient).

Further characterizing gene co-expression modules: We used module preservation analysis to establish that our defined co-expression modules were not dominated by (i) the large number of DEGs induced by X-monosomy (using expression data excluding XO samples), or (ii) other SCD group differences in mean expression levels (using expression data after residualization for the effects of SCD group and re-centering at a common mean). All modules showed high reproducibility based on a module-specific $Z_{summary}$ scores derived by comparing observed modular connectivity and density metrics with null values generated by 200 permutations of gene-level module membership(12). We focused further characterization of modules which passed two independent statistical criteria; (i) SCD sensitivity - quantified using F-tests for the omnibus effects of karyotype group on modular expression quantified as the ME, (ii) functional coherence as inferred by analysis of modular gene ontolology term enrichments using GO elite (13), and Gorilla (14).

Testing for enrichment of autoimmune disorder risk genes in WGCNA modules: A large-scale records-based study was used to define 10 Autoimmune Disorder (ADs) with clearly elevated prevalence rates in XXY vs. XY males (15), 9 of which were represented in the largest available catalog of Genome Wide Association Study (GWAS) findings (https://www.ebi.ac.uk/gwas/): Diabetes Mellitus type 1, Multiple Sclerosis, Autoimmune Hypothyroidism, Psoriasis, Rheumatoid Arthritis, Sjogren's Syndrome, Systemic Lupus Erythematosus, Ulcerative Colitis, and

Coeliac Disease. A total of 495 genes within our microarray sample were annotated for showing a significant association in GWAS with one or more of these 9 AD conditions. Overrepresentation of this AD gene set in the XXY upregulated Red gene co-expression module was tested for using both Fisher's exact test (p=0.01), and by comparing the observed representation of AD genes against a null distribution generated by 10,000 random gene samples of equal size to the red module.

Testing for patterned enrichment of dosage sensitive sex chromosome genes in WGCNA modules: We tested if any of the 8 SCD-sensitive and functionally enriched WGCNA modules showed enrichment for the previously derived k-means clusters of dosage sensitive sex chromosome genes (**Fig. 2**) by applying two-tailed Fishers tests to all pairwise module-cluster combinations (**Fig. 3e**). All observed associations survived Bonferroni correction for multiple comparisons.

Supplementary Text S8. Comparison of Autosomal Gene Fold-Change in SCA and Down Syndrome.

The transcriptomic effects on Trisomy 21 (T21) were characterized in LCLs by passing a publically available Illumina microarray gene expression dataset (GEO, Accession number GSE34458) through an identical analytic pipeline to that used in characterizing genome-wide fold changes in our SCA sample (see above). We first independently confirmed the previously reported finding that chromosome 21 was robustly enriched for genes showing differential expression in this T21 data set (Chi-squared=999, p<2*10-16 for enrichment of DEGs on chromosome 21) –

buttressing use of these data to assess transcriptomic effects of T21. We examined overlaps in genome-wide expression change between T21 and the three sex-chromosome trisomies in our samples (XXY, XYY and XXX) using 17671 genes with complete expression data in both microarray datasets (after exclusion of genes on chromosomes X, Y and 21). We tested for, and failed to find any evidence of significant overlap in DEGs using Chi-squared tests (**Dataset S6**). To test if T21 showed a similar shift in gene co-expression modules to sex chromosome trisomies, we used the designation of genes to modules in the SCA sample to recalculate module Eigengenes. We then calculated ME fold changes for T21 and three SCA trisomies (XXX, XXY and XYY). Trisomy of chromosome 21 was associated with a clearly distinct profile of ME expression change than all three of the SCA trisomies (**SI Appendix Fig. S4c**).

Supplementary Text S9. Transcription factor binding site (TFBS) enrichment analysis.

Transcription factor binding site analyses: Transcription factor binding site (TFBS) enrichment analysis was performed each of the 4 SCD-sensitive WGCNA modules - Blue, Green, Turquoise and Brown - that were enriched for inclusion of gene from one or more of the 5 SCD-sensitive clusters of sex chromosome genes. In each module, we scanned canonical promoter regions (1000bp upstream of the transcription start site) for the top 500 genes with strongest intramodular "connectivity" (based on kME - the magnitude of each gene's coexpression with its module's eigengene). Next we utilized TFBS position weight matrices (PWMs)

from JASPAR database (205 non-redundant and experimentally defined motifs) (16) to examine the enrichment for corresponding TFBS within each module. For TFBS enrichment all the modules were scanned with each PWMs using the Clover algorithm (17). To compute the enrichment analysis, we utilized three different background datasets (1000 bp sequences upstream of all human genes, human CpG islands and human chromosome 20 sequence). To increase confidence in the enrichment analyses, we considered TFBS to be over-represented based on the P-values (<0.05) obtained relative to all the three corresponding background datasets.

This analysis converged on a single TF - ZFX, encoded by the X-linked member of an X-Y gametolog pair – as the only SCD-sensitive TF showing significant TFBS enrichment in one or more modules of the 4 WGCNA modules examined. To provide an orthogonal experimental test for evidence of a regulatory role for ZFX within these 4 WGCNA modules, we used a list of human homologs for genes with significantly decreased expression due to ZFX knockout in murine lymphocytes (18). Two-tailed Fishers Tests were used to assess if these human genes were significantly enriched/impoverished in any of the 8 SCD sensitive WGCNA modules.

REFERENCES

1. Ritchie ME, et al. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43(7):e47.

2. Barbosa-Morais NL, et al. (2010) A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. *Nucleic Acids Res* 38(3):e17.

3. Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559.

4. Schmittgen TD, Livak KJ (2008) Analyzing real-time PCR data by the comparative C(T) method. *Nat Protoc* 3(6):1101–1108.

5. Balaton BP, Cotton AM, Brown CJ (2015) Derivation of consensus inactivation status for X-linked genes from genome-wide studies. *Biol Sex Differ* 6:35.

6. Stephane Champely (2018) *pwr: Basic Functions for Power Analysis*.

7. Ernst J, et al. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473(7345):43–49.

8. Pandey RS, Wilson Sayres MA, Azad RK (2013) Detecting evolutionary strata on the human x chromosome in the absence of gametologous y-linked sequences. *Genome Biol Evol* 5(10):1863–1871.

9. Tukiainen T, et al. (2017) Landscape of X chromosome inactivation across human tissues. *Nature* 550(7675):244–248.

10. Zhang B, Horvath S (2005) A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* 4:Article17.

11. Langfelder P, Zhang B, Horvath S (2008) Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinforma Oxf Engl* 24(5):719–720.

12. Langfelder P, Luo R, Oldham MC, Horvath S (2011) Is my network module preserved and reproducible? *PLoS Comput Biol* 7(1):e1001057.

13. Zambon AC, et al. (2012) GO-Elite: a flexible solution for pathway and ontology over-representation. *Bioinforma Oxf Engl* 28(16):2209–2210.

14. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 10:48.

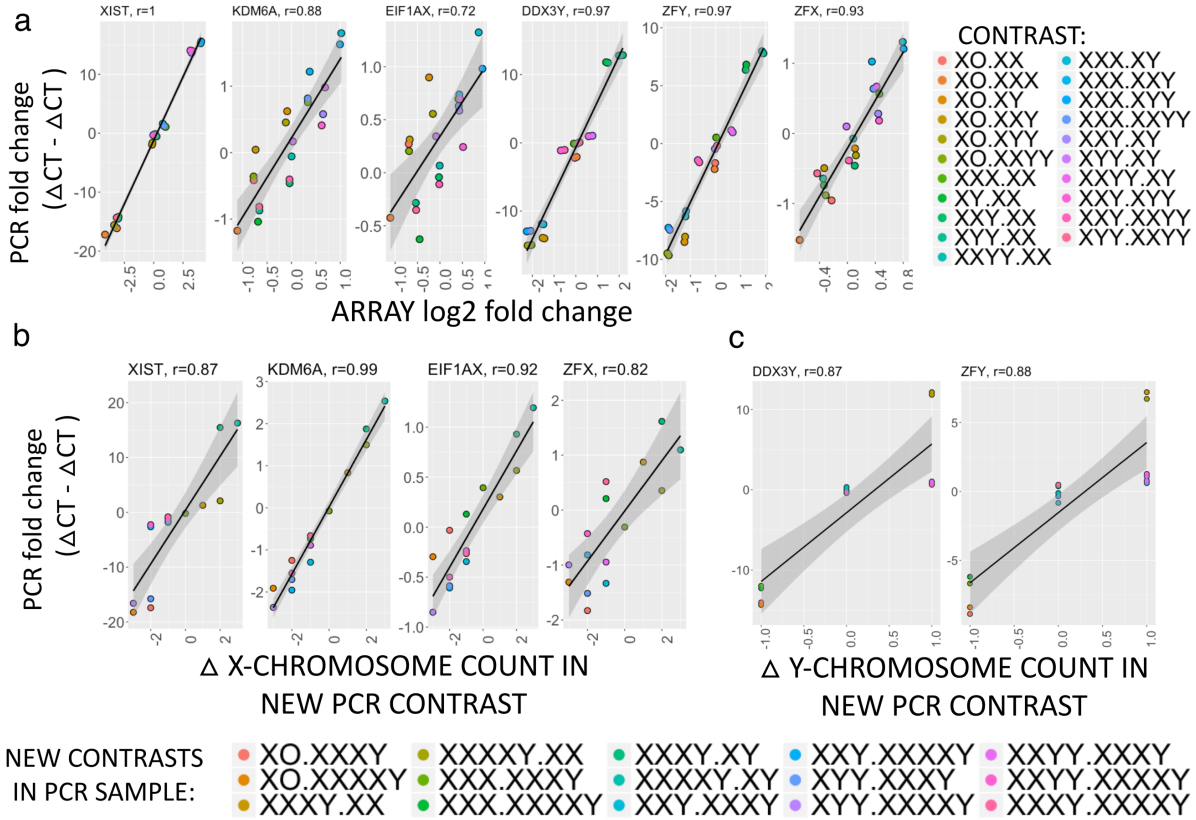
15. Seminog OO, Seminog AB, Yeates D, Goldacre MJ (2015) Associations between Klinefelter's syndrome and autoimmune diseases: English national record linkage studies. *Autoimmunity* 48(2):125–128.

16. Portales-Casamar E, et al. (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res* 38(Database issue):D105-110.

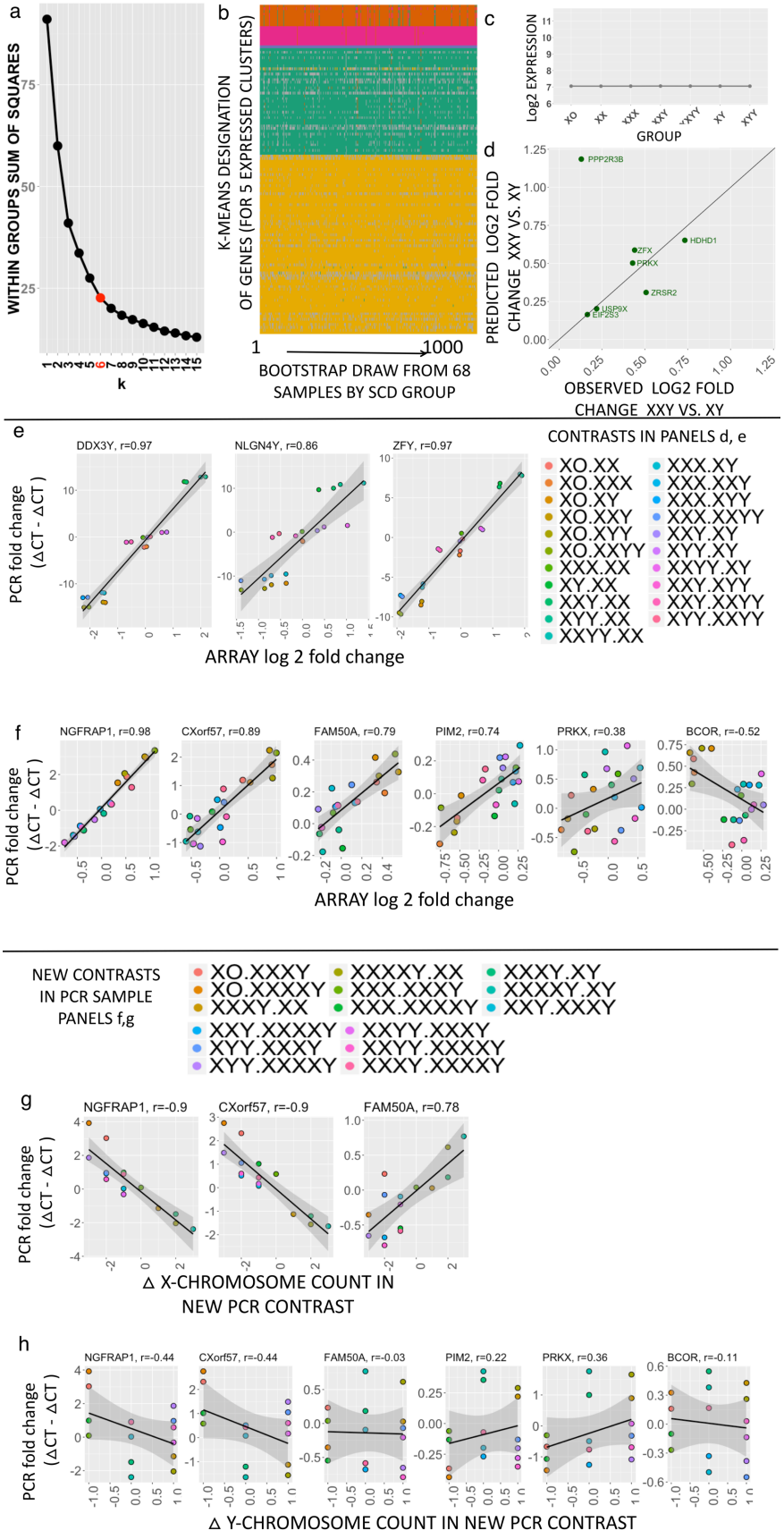
17. Frith MC, et al. (2004) Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res* 32(4):1372–1381.

18. Weisberg SP, et al. (2014) ZFX controls propagation and prevents differentiation of acute T-lymphoblastic and myeloid leukemia. *Cell Rep* 6(3):528–540.

SUPPLEMENTARY FIGURES LEGENDS



Supplementary Fig. S1. qRT-PCR Validation Results for Selected Genes Showing Consistent Microarray Expression Changes with Altered Sex Chromosome Dosage [XIST, KDM6A, EIF1AX, DDX3Y, ZFY] and ZFX. a) Scatterplots comparing observed log 2 fold-change in gene expression for each unique pairwise group contrast in microarray data (x-axis), vs. observed fold-change (ΔCT - ΔCT) for the identical group contrast in an independent validation sample by qRT-PCR (y-axis). Contrast names reflect the pair of SCD groups being compared (i.e. "XO.XX" means "XO vs. XX"). **b, c)** Scatterplots of observed fold-change (ΔCT - ΔCT) by qRT-PCR (y-axis) for new group contrasts provided by the additional karyotype groups that were uniquely available in the independent validation sample. The x-axes encode the difference in X-chromosome (**b**) and Y-chromosome (**c**) dose captured by each group contrast.

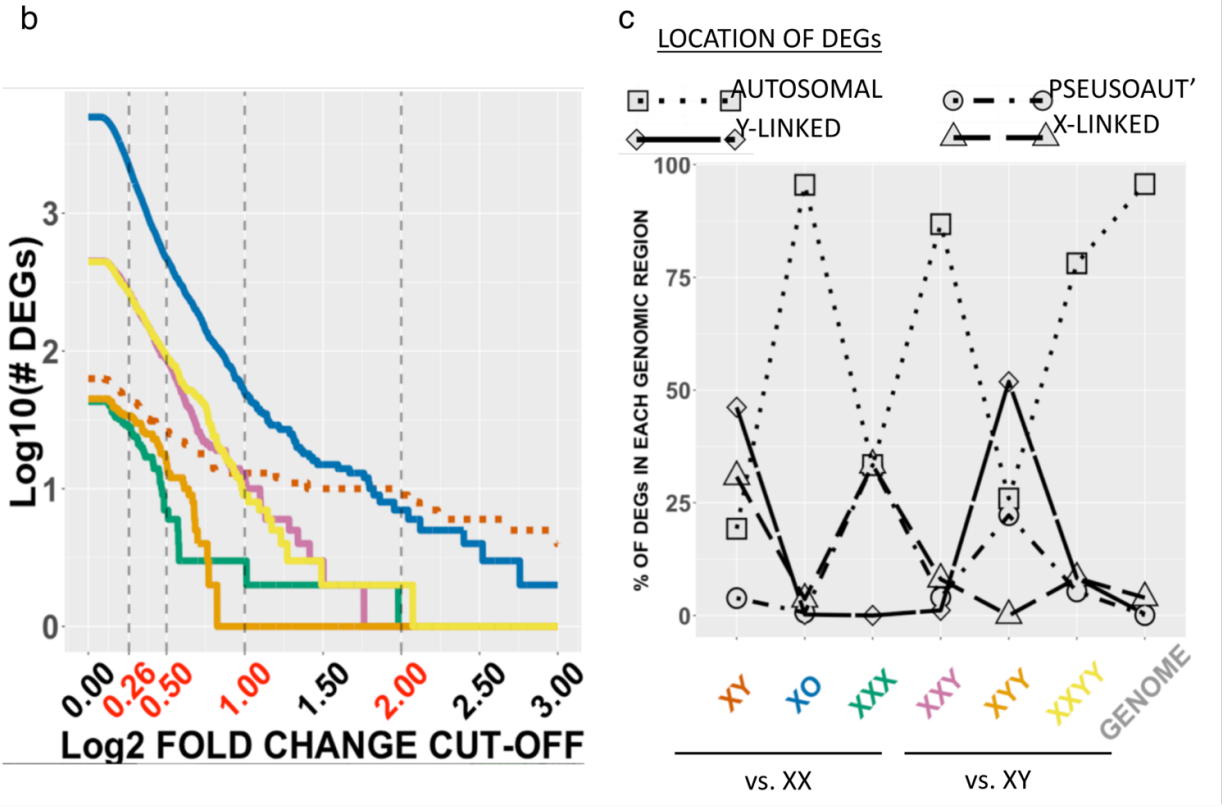

a
WITHIN GROUPS SUM OF SQUARES
k
b
K-MEANS DESIGNATION OF GENES (FOR 5 EXPRESSED CLUSTERS)
1
1000
BOOTSTRAP DRAW FROM 68 SAMPLES BY SCD GROUP
c
Log2 EXPRESSION
GROUP
XO
XX
XXX
XXY
XYY
XY
XXY
d
PREDICTED LOG2 FOLD CHANGE XXY VS. XY
OBSERVED LOG2 FOLD CHANGE XXY VS. XY
PPP2R3B
HDHD1
ZFX
PRKX
ZRSR2
USP9X
EIF2S3
e
DDX3Y, r=0.97
NLGN4Y, r=0.86
ZFY, r=0.97
PCR fold change (ΔCT - ΔCT)
ARRAY log 2 fold change
CONTRASTS IN PANELS d, e
XO.XX
XO.XXX
XO.XY
XO.XXY
XO.XYY
XO.XXYY
XXX.XX
XY.XX
XXY.XX
XYY.XX
XXYY.XX
XXX.XY
XXX.XXY
XXX.XYY
XXX.XXYY
XXY.XY
XYY.XY
XXYY.XY
XXY.XYY
XXY.XXYY
XYY.XXYY
f
NGFRAP1, r=0.98
CXorf57, r=0.89
FAM50A, r=0.79
PIM2, r=0.74
PRKX, r=0.38
BCOR, r=-0.52
PCR fold change (ΔCT - ΔCT)
ARRAY log 2 fold change
NEW CONTRASTS IN PCR SAMPLE PANELS f,g
XO.XXXY
XO.XXXXY
XXXY.XX
XXXXY.XX
XXX.XXXY
XXX.XXXXY
XXXY.XY
XXXXY.XY
XXY.XXXY
XXY.XXXXY
XYY.XXXY
XYY.XXXXY
XXYY.XXXY
XXXY.XXXXY
g
NGFRAP1, r=-0.9
CXorf57, r=-0.9
FAM50A, r=0.78
PCR fold change (ΔCT - ΔCT)
Δ X-CHROMOSOME COUNT IN NEW PCR CONTRAST
h
NGFRAP1, r=-0.44
CXorf57, r=-0.44
FAM50A, r=-0.03
PIM2, r=0.22
PRKX, r=0.36
BCOR, r=-0.11
PCR fold change (ΔCT - ΔCT)
Δ Y-CHROMOSOME COUNT IN NEW PCR CONTRAST

Supplementary Fig. S2 (above). Reproducible Grouping of Sex Chromosome Genes by K-means Clustering Analysis of Expression Profiles, and PCR Validation of Observed Expression Profiles for Selected Genes from Y-, XCIE- and XCI-enriched K-Means Clusters. a) Scree plot used for selection of 6 cluster solution in k-means analysis. One large cluster of genes with low/absent expression across all microarray samples (**SI Appendix Text S4**) was excluded from further analysis, and the 5 remaining SCD sensitive clusters were color-coded Orange (PAR-cluster), Pink (Y-cluster), Purple (XIST), Green (XCIE-cluster) and Yellow (XCI-cluster). **b)** Matrix of genes * bootstrap samples showing consistency of assignment of sex chromosome genes to these 5 clusters across 1000 sets of 68 samples generated by separately resampling individuals with replacement from each SCD group. **c)** Dot and line plot showing observed mean expression for the large cluster of 773 low-expressed genes across karyotype groups (see **SI Appendix Text S4**). **d)** Qualitative comparison between observed log 2 fold changes for XCIE genes in XXY vs. XY groups, and predicted fold-changes based on independently published Allelic Expression Imbalance study (9) of XX LCLs with skewed X-inactivation. General agreement for 6/7 genes with available data suggests the sub-linear scaling between expression of the XCIE-enriched Green cluster and X-chromosome dosage may be explained by the incomplete nature of escape from XCI. **e, f)** Scatterplots comparing observed log 2 fold-change in gene expression for each unique pairwise group contrasts in microarray data (x-axis), vs. observed fold-changes for the identical group contrasts by qRT-PCR (y-axis) in an independent validation sample. Contrast names reflect the pair of SCD groups being compared (i.e. “XO.XX” means “XO vs. XX”). Plots are shown for selected genes from the **e)** Y-cluster and **f)** X-linked , XCI-

and XCIE-clusters. **g, h)** Scatterplots for observed gene-expression fold-change by qRT-PCR (y-axis) in pairwise group contrasts involving XXXY and XXXXY karyotypes that were unique to the independent validation sample and not represented in the original microarray study. **g)** For 2 of 3 XCI-cluster genes (NGFRAP1 and CXorf57), qRT-PCR data extend the finding of greater gene expression with lower X-chromosome count. **h)** qRT-PCR data confirm microarray findings of Y-chromosome dosage sensitivity for 2 of 3 XCI-cluster genes (NGFRAP1 and CXorf57), and 2 of 3 XCIE-cluster (PIM2, PRKX) genes.

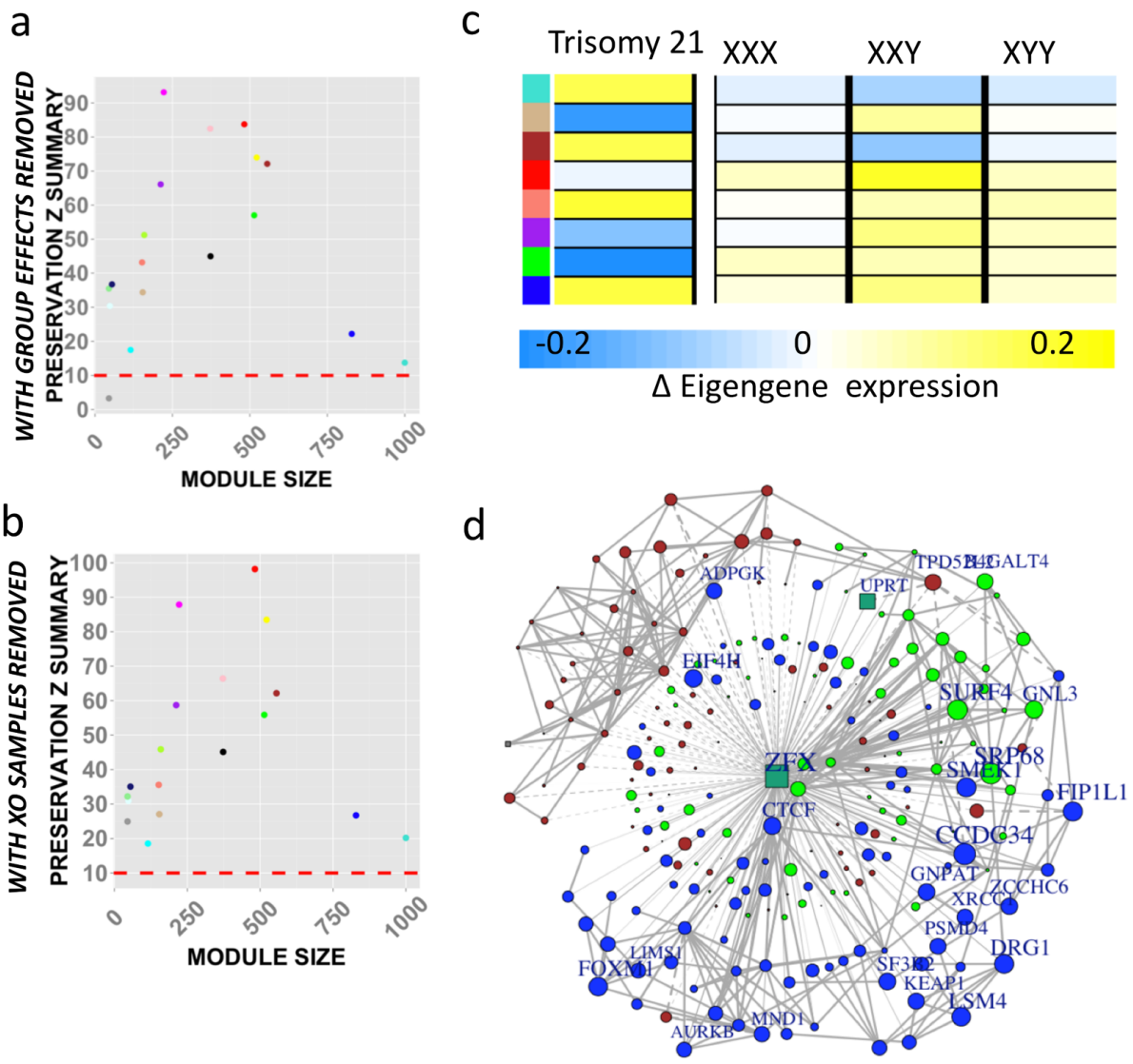
a

DOSAGE CHANGE	CONTRAST	LOG2 FOLD CHANGE CUT-OFF			
		0.26	0.5	1	2
Loss of X in female	XO vs. XX	2204	469	50	7
Addition of X in male	XXY vs. XY	268	85	13	1
	XXYY vs. XY	266	93	9	2
Addition of Y in male	XYY vs. XY	34	15	0	0
Addition of X in female	XXX vs. XX	28	7	3	1



Supplementary Figure S3. Genome-wide Effects of Sex Chromosome Dosage Variation. a-b) Table **a** and corresponding line-plot **b** showing number of genes with significant differential expression (after FDR correction with q<0.05) in different SCD contrasts at varying |log2 fold change| cut-offs. Note the order-of-magnitude differences between the number of Differentially Expressed Genes (DEGs) in XO ("removal of X from female") vs. XXY and XXYY ("addition of X to male") vs. XYY and XXX ("addition of Y and X to male and female, respectively"). Filters of q<0.05 and |log2 fold change| threshold

of 0.26 (~20% change in expression, and mean point of fastest DEG drop-off in Fig S3a) were applied to categorically define differential expression in other analyses. **c)** Dot-and-line plot showing the proportion of DEGs in each karyotype group that fell within different regions of the genome. The proportion of all genes in the genome within each genomic region is shown for comparison. All SCD groups showed non-random DEG distribution relative to the genome ($p<2*10^{-16}$), but DEG distributions differed significantly between SCD groups ($p<2*10^{-16}$). XO, XXX and XXYY are distinguished from all other SCDs examined by the large fraction of their overall DEG count that comes from autosomal genes.



Supplementary Figure S4. Supplementary Results from WGCNA Analyses. a) Scatterplot showing scaled module preservation score for each module when testing for the reproducibility of all 18 modules defined in our original dataset, after all expression data had been "de-meaned" by SCD group. Z-summary scores above 10 indicate highly significant module preservation – suggesting that the modules presented in our main paper are not artifacts of correlated inter-SCD group differences in gene expression between SCD groups, but capture meaningful co-expression patterns that can be

detected within each SCA group. **b)** Scatterplot showing scaled module preservation score for each module when testing for the reproducibility of all 18 modules defined in our original dataset, after exclusion of all XO samples. Z-summary scores above 10 indicate highly significant module preservation – suggesting that the composition of the co-expression modules presented in our main paper is not dominated by the extreme impact of XO on gene expression as compared to other SCD changes. **c)** Heatmap showing distinct profile of module DE with a supernumerary chromosome 21 vs. a supernumerary X-chromosome. **d)** ZFX and its target genes from Blue, Green and Brown modules with significant ZFX TFBS enrichment. Note that expression levels of ZFX (which increases in expression with mounting X-chromosome dosage) are positively correlated (solid edges) with SCD sensitive genes that are up-regulated by increasing X-chromosome dose (Blue and Green modules), but negatively correlated (dashed edges) with genes that are down-regulated by increasing X-chromosome dose (Brown module).

SUPPLEMENTARY DATASET LEGENDS
(Tables/Datasets provided as individual .xlsx data files)

Dataset S1. Sample Characteristics for Core Microarray Dataset (n=68) and PCR Validation Sample (n=401). Number of samples by karyotype group, with details of age and self-reported "race" distributions by group. *4 XO samples in PCR validation study are an independent sample of cells drawn from 4 of the 12 XO cell-lines used in the core microarray study (i.e. technical replicates). All other samples used in the PCR validation study were drawn from a separate set of participants to those included in the microarray study (i.e. biological replicates).

Dataset S2. Annotations for 10 Sex Chromosome Genes with "Obligate SCD Sensitivity".

Dataset S3. Data-Driven Grouping of Sex Chromosome Genes by Dosage Sensitivity. Each row relates to one of the 5 distinct gene groups detected by k-means clustering according to mean expression across karyotype groups. Details are provided group size, constituent genes, and observed profile of SCD Sensitivity. For clusters with >10 genes, we provide names for the 10 cluster genes with the smallest Euclidean distance from the cluster k-mean centroid. Heatmap colors in fold-change (FC) columns index the magnitude (color intensity) and direction of each k-mean cluster's change in mean expression (yellow - increased, blue- decreased) for selected pairwise SCD group contrasts. Note how (i) the Yellow cluster enriched for XCI genes is less expressed in groups with greater X-chromosome count, and (ii) Yellow and Green clusters enriched for

XCI and XCIE genes respectively are both differentially expressed between groups that differ in Y-chromosome status, but have identical X-chromosome count.

Dataset S4. Mean Expression by Karyotype Group for Genes in SCD-sensitive k-means Clusters.

Dataset S5. Characteristics of Gene-Coexpression Modules Generated by WGCNA. For all 19 modules defined by WGCNA, we provide information regarding: module size; "top" module genes [defined by strength of correlated expression with module eigengene (ME)]; proportion variance in module expression explained by ME; ME correlation with other ME of all other modules; F-test for effect of SCD on ME variance; results of t-tests for selected group differences in ME expression (bold cells survive Bonferroni correction for # modules); significant GO term enrichment by GOelite and GOrilla; binary statement regarding whether module shows both significant SCD sensitivity and significant GO terms enrichment

Dataset S6. Cross-tabulations and associated Chi-squared tests for overlap in differentially expressed genes between trisomy 21 and the 3 sex-chromosome trisomies in our sample. Note, these calculations were made after removal of genes located on sex chromosomes, or chromosome 21. For this comparison, differential expression was defined as a statistically significant fold change in expression of any magnitude that survived FDR correction for multiple comparisons at q=0.05 This liberal

fold-change cut-off was applied given the to accommodate the very small number of DEGs seen in XXX.