Sex Chromosome Dosage Effects on Gene Expression in Humans

Armin Raznahan*[1], Neelroop Parikshak[2], Vijayendran Chandran[2], Jonathan Blumenthal[1], Liv Clasen[1], Aaron Alexander-Bloch[1], Andrew Zinn[3], Danny Wangsa[4], Jasen Wise[5], Declan Murphy[6], Patrick Bolton[6], Thomas Ried[4], Judith Ross[7], Jay Giedd[8], Daniel Geschwind[2]

1. *Developmental Neurogenomics Unit, National Institute of Mental Health, NIH, Bethesda, MD, USA*
2. *Neurogenetics Program, Department of Neurology and Center for Autism Research and Treatment, Semel Institute, David Geffen School of Medicine, University of California Los SCVAngeles, Los Angeles, CA, USA*
3. *McDermott Center for Human Growth and Development and Department of Internal Medicine, University of Texas Southwestern Medical School, TX, USA*
4. *Genetics Branch, Center for Cancer Research, National Cancer Institute, NIH, Bethesda, MD, USA*
5. *Qiagen*
6. *Institute of Psychiatry, Psychology and Neuroscience, King's College London, University of London, UK*
7. *Department of Pediatrics, Thomas Jefferson University, Philadelphia, PA, USA*
8. *Department of Psychiatry, UC San Diego, La Jolla, CA, USA*

*Correspondence to: raznahana@mail.nih.gov

ABSTRACT

A fundamental question in the biology of sex-differences has eluded direct study in humans: how does sex chromosome dosage (SCD) shape genome function? To address this, we developed a systematic map of SCD effects on gene function by analyzing genome-wide expression data in humans with diverse sex chromosome aneuploidies (XO, XXX, XXY, XYY, XXYY). For sex chromosomes, we demonstrate a pattern of obligate dosage sensitivity amongst evolutionarily preserved X-Y homologs, and revise prevailing theoretical models for SCD compensation by detecting X-linked genes whose expression increases with decreasing X- and/or Y-chromosome dosage. We further show that SCD-sensitive sex chromosome genes regulate specific co-expression networks of SCD-sensitive autosomal genes with critical cellular functions and a demonstrable potential to mediate previously documented SCD effects on disease. Our findings detail wide-ranging effects of SCD on genome function with implications for human phenotypic variation.

INTRODUCTION

Disparity in SCD is fundamental to the biological definition of sex throughout much of the animal kingdom. In almost all eutherian mammals, females carry two X-chromosomes, while males carry an X- and a Y-chromosome: presence of the Y-linked SRY gene determines a testicular gonadal phenotype, while its absence allows development of ovaries[1]. Sexual differentiation of the gonads leads to hormonal sex-differences that have traditionally been considered the major proximal cause for extra-gonadal phenotypic sex-differences. However, diverse studies, including recent work in transgenic mice that uncouple Y-chromosome and gonadal status, have revealed direct SCD effects on several sex-biased metabolic, immune and neurological phenotypes[2].

These findings - together with reports of widespread transcriptomic differences between pre-implantation XY and XX embryos[3,4] - suggest that SCD has gene regulatory effects independently of gonadal status. However, genome-wide consequences of SCD remain poorly understood, especially in humans, where experimental dissociation of SCD and gonadal status is not possible. Understanding these regulatory effects is critical for clarifying the biological underpinnings of phenotypic sex-differences, and the clinical features of sex chromosome aneuploidy[5] [SCA, e.g. Turner (XO) and Klinefelter (XXY) syndrome], which can both manifest as altered risk for several common autoimmune and neurodevelopmental disorders (e.g. systemic lupus erythematosus and autism spectrum disorders)[6,7]. Here, we explore the genome

47 wide consequences of SCD through comparative transcriptomic analyses
48 amongst humans across a range of dosages including typical XX and XY
49 karyotypes, as well as several rare SCA syndromes associated with 1, 3, 4 or 5
50 copies of the sex chromosomes. We harness these diverse karyotypes to dissect
51 the architecture of dosage compensation amongst sex chromosome genes, and
52 to systematically map the regulatory effects of SCD on autosomal gene
53 expression in humans.

54 We examined gene expression profiles in a total of 469 lymphoblastoid
55 cell lines (LCLs), from (i) a core sample of 68 participants (12 XO, 10 XX, 9 XXX,
56 10 XY, 8 XXY, 10 XYY, 9 XXYY) yielding for each sample genome-wide
57 expression data for 19,984 autosomal and 894 sex-chromosome genes using the
58 Illumina oligonucleotide Beadarray platform (**Methods**), and (ii) an independent
59 set of validation/replication samples from 401 participants (3 XO, 145 XX, 22
60 XXX, 146 XY, 34 XXY, 16 XYY, 17 XXYY, 8 XXXY, 10 XXXXY) with quantitative
61 reverse transcription polymerase chain reaction (qPCR) measures of expression
62 for genes of interest identified in our core sample (**Table 1, Methods**).

63

64 **RESULTS**

65 **Evolutionarily Preserved X-Y Gametologs Show Extreme Dosage**
66 **Sensitivity**

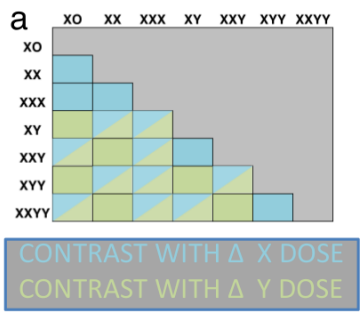
67 To first verify our study design as a tool for probing SCD effects on gene
68 expression, and to identify core SCD-sensitive genes, we screened all 20,878
69 genes in our microarray dataset to define which, if any, genes showed a

70 persistent pattern of significant differential expression (DE) across all unique
71 pairwise group contrasts involving a disparity in either X- or Y-chromosome
72 dosage (n=15 and n=16 contrasts respectively, **Fig. 1a**). Disparities in X-
73 chromosome dosage were always accompanied by statistically significant DE in
74 4 genes, which were all X-linked: *XIST* (the orchestrator of X-inactivation) and 3
75 other known genes known to escape X-chromosome inactivation (*PUDP*,
76 *KDM6A*, *EIF1AX*)[8]. Similarly, disparities in Y-chromosome dosage always led to
77 statistically-significant DE in 6 genes, which were all Y-linked: *CYorf15B*, *DDX3Y*,
78 *TMSB4Y*, *USP9Y*, *UTY,* and *ZFY*. Observed expression profiles for these 10
79 genes perfectly segregated all microarray samples by karyotype group (**Fig. 1b**),
80 and could be robustly replicated and extended in the independent sample of
81 LCLs from 401 participants with varying SCD (**Supplementary Fig. 1, Methods**).
82 Strikingly, 8 of the 10 genes showing obligatory SCD sensitivity (excepting
83 XIST and PUPD) are members of a class of 16 sex-linked genes with homologs
84 on both the X and Y chromosomes (i.e. 16 X-Y gene pairs, henceforth
85 gametologs)[9] that are distinguished from other sex-linked genes by (i) their
86 selective preservation in multiple species across ~300 million years of sex
87 chromosome evolution to prevent male-female dosage disparity, (ii) the breadth
88 of their tissue expression from both sex chromosomes; and (iii) their key
89 regulatory roles in transcription and translation[9,10] (**Fig. 1c**). Broadening our
90 analysis to all 14 X-Y gametolog pairs present in our microarray data found that
91 these genes as a group exhibit a heightened degree of SCD-sensitivity that
92 distinguishes them from other sex-linked genes (**Fig. 1d, Methods**). These

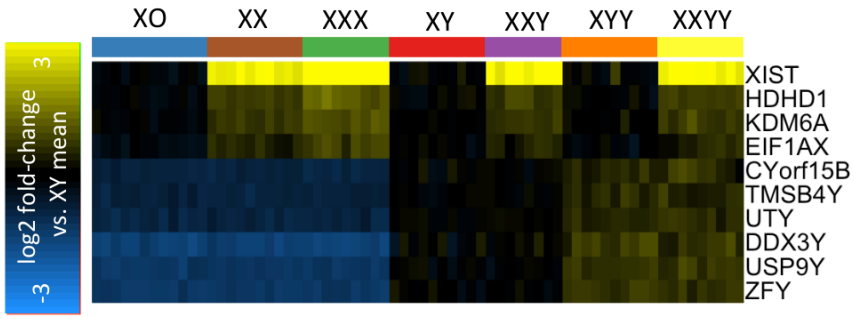
findings provide the strongest evidence to date that the evolutionary maintenance, broad tissue expressivity and enriched regulatory functions of X-Y gametologs[10] are indeed accompanied by a distinctive pattern of dosage sensitivity, which firmly establishes these genes as candidate regulators of SCD effects on wider genome function.

Figure 1 (next page). Consistent Gene Expression Changes with Altered Sex Chromosome Dosage. a) Cross table showing all unique pairwise SCD group contrasts within our microarray dataset, color coded for their involvement of changes in X- and Y-chromosome count. **b)** Two-dimensional expression heat-map for the 10 genes showing differential expression across all contrasts that involve disparity in X or Y-chromosome count. Column colors encode SCD group membership for each sample. Rows detail gene expression across all SCD samples as a log 2 fold-change relative to the mean expression in XY males. **c)** Table providing Gene ID, location, function and homolog annotations for the 10 genes that showed obligate SCD sensitivity. Eight genes in this set are members of X-Y gametolog gene pairs. **d)** Density plots showing observed mean SCD sensitivity of the 14 gametolog genes in our study (red line), vs, distribution (black line) of 10,000 randomly sampled sets of non-gametolog sex-linked genes of equal size. Results are provided separately for X- and Y-chromosomes. For both chromosomes, the mean SCD sensitivity of the gametolog gene set greater is than that of all 10k permuted gene sets.

a



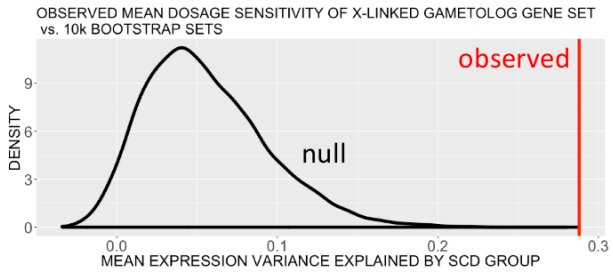
b

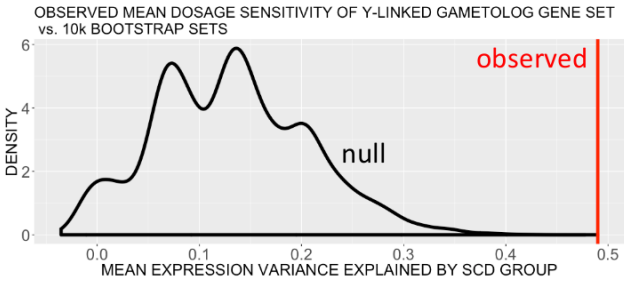


c

Gene_Name (Alias)	Entrez_ID	Chromosome	Probe_Start	Probe_End	Functions	Gametolog_Name (Alias)	Functions
PUDP (HDHD1)	8226	X	6967201	6967250	encodes protein of unknown biological function	-	-
KDM6A (UTX)	7403	X	44970808	44970857	catalyzes the demethylation of tri/dimethylated histone H3	UTY	protein-protein interactions
XIST	7503	X	73040921	73040970	essential for initiation and spread of X inactivation	-	-
EIF1AX	1964	X	20150300	20150326	essential translation initiation factor	EIF1AY	translation initiation
Gene_Name (Alias)	**Entrez_ID**	**Chromosome**	**Probe_Start**	**Probe_End**	**Functions**	**Gametolog_Name (Alias)**	**Functions**
ZFY	7544	Y	2849596	2849645	transcription factor	ZFX	transcription factor
USP9Y	8287	Y	14971985	14972034	ubiquination	USP9X	ubiquination
DDX3Y	8653	Y	15032184	15032233	RNA and ATP binding, hydrolysis	DDX3X	RNA helicase, transcriptional regulation and translation
UTY	7404	Y	15360382	15360431	protein-protein interactions	KDM6A (UTX)	catalyzes the demethylation of tri/dimethylated histone H3
TMSB4Y (TB4Y)	9087	Y	15817595	15817644	uncharacterized	TMSB4X (TB4X)	actin polymerization, cell proliferation/differentiation
TXLNGY (CYorf15B)	84663	Y	21765838	21765887	pseudogene	TXLNG	transcriptional regulation, intracellular trafficking

d





120 **Observed Sex Chromosome Dosage Effects on X- and Y-chromosome**
121 **Genes Revise Current Models of Dosage Compensation**

122 We next harnessed our study design to test the canonical model for SCD
123 compensation which defines four mutually-exclusive classes of sex chromosome
124 genes that are predicted to have differing responses to changing SCD[11]: (i)
125 pseudoautosomal region (PAR) genes, (i) Y-linked genes, (ii) X-linked genes that
126 undergo X-chromosome inactivation (XCI), and (iv) X-linked genes that "escape"
127 XCI (XCIE). To test this "Four Class Model" we considered all sex chromosome
128 genes and performed unsupervised *k*-means clustering of genes by their mean
129 expression in each of the 7 karyotype groups represented in our microarray
130 dataset, and compared this empirically-defined grouping with that the given *a*
131 *priori* by the Four Class Model (**Methods**). *k*-means clustering distinguished 5
132 reproducible (color-coded) clusters of SCD-sensitive sex-chromosome genes
133 that overlapped strongly with gene groups predicted by the Four Class Model,
134 from a large left-over cluster of 773 genes that were not consistently expressed
135 across samples (median detection rate of 4/68 samples) and did not vary in their
136 expression as a function of SCD (**Table 2, Fig. 2a,b , Supplementary Fig. 2a,b**).

137 The 5 SCD-sensitive groups of sex chromosome genes detected by k-
138 means were: an Orange cluster of PAR genes, a Pink cluster of Y-linked genes
139 (especially enriched for Y gametologs, odds ratio=5213, $p=1.3*10^{-15}$), a Green
140 cluster enriched for known XCIE genes (especially X gametologs, odds
141 ratio=335, $p=3.4*10^{-11}$), and a Yellow cluster enriched for known XCI genes. The
142 X-linked gene responsible for initiating X-inactivation - XIST - fell into its own

143 Purple “cluster” (**Fig. 2b**). For all but the Orange cluster of PAR genes, observed
144 patterns of gene-cluster dosage sensitivity across karyotype groups deviated
145 from those predicted by the Four Class Model in several substantial ways (**Fig.**
146 **2c**).

147 Mean Expression for the Pink cluster of Y-linked genes increased in a
148 stepwise fashion with Y-chromosome dosage, but countered the Four Class
149 Model prediction by showing a sub-linear relationship with Y-chromosome count
150 – indicting that these Y-linked genes may be subject to active dosage
151 compensation. Fold-changes observed by microarray for 5/5 of these Y-linked
152 genes were highly replicable by qPCR in an independent sample of 401
153 participants with varying SCD (**Methods, Supplementary Fig. 2c).**

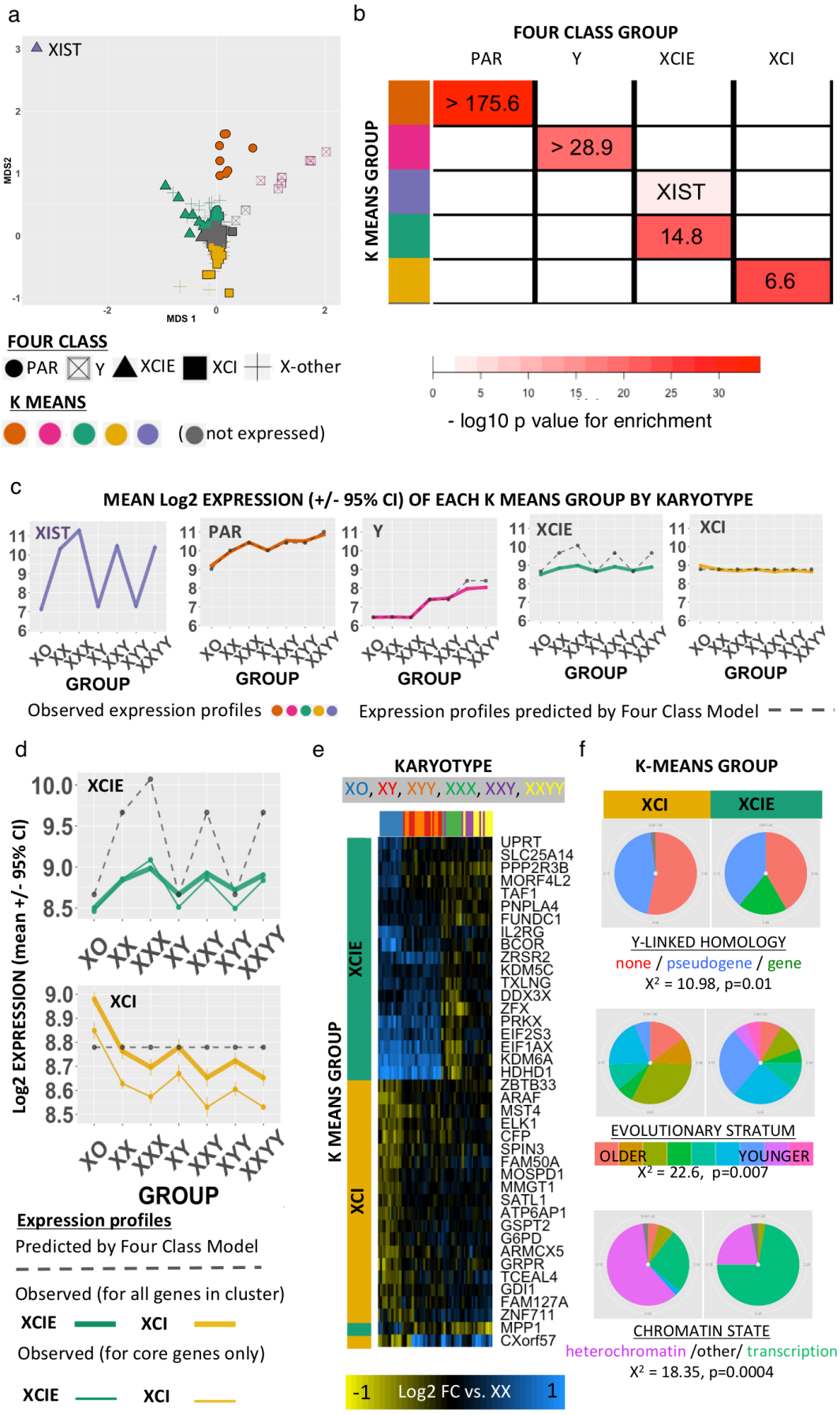
154 Four Class Model predictions were also challenged by observed
155 expression profiles for the Yellow and Green clusters of X-linked genes (**Table 2,**
156 **Fig. 2c,d**). Linear models for X and Y chromosome dosage effects on expression
157 (**Methods**) indicated that the XCI-enriched Yellow cluster was highly sensitive to
158 SCD ($F=47.7$, $p<2.2*10^{-16}$), and expression of genes in this cluster was
159 significantly *inversely* related to X-chromosome dosage at the level of both mean
160 cluster expression (coefficient for linear effect of X-chromosome count on
161 expression = -0.12, $p=3.8*10^{-15}$) and the individual expression profile of 60/66
162 genes within the cluster ($p<0.05$ for negative linear effect of X count on
163 expression). This observation indicates that increasing X copy number does not
164 solely involve silencing of these genes from the inactive X-chromosome, but a
165 further repression of their expression from the additional active X-chromosome.

166 Remarkably, mean expression of the XCI gene cluster was also
167 significantly decreased by presence of a Y-chromosome at the level of both
168 mean cluster expression (coefficient for linear effect of Y-chromosome count on
169 expression = - 0.09, $p= 1.4*10^{-14}$) and expression profiles of 48/66 individual
170 cluster genes ($p<0.05$ for negative linear effect of Y count on expression). The
171 Green XCIE cluster manifested an inverted version of this effect whereby
172 increases in Y chromosome dosage were associated with increased gene
173 expression ($p< 6.2*^{10-11}$ for mean cluster expression and $p<0.05$ for 23/39 cluster
174 genes) - providing the first evidence that Y-chromosome status can influence the
175 expression level of X-linked genes independently of circulating gonadal factors.
176 Finally, mean fold change for XCIE-cluster genes scaled sub-linearly with X-
177 chromosome dosage. Importantly, mean expression profiles for XCIE- and XCI-
178 enriched gene clusters still deviated from predictions of the Four Class Model
179 when analysis was restricted to genes with the highest-confidence (i.e.
180 independently replicated across 3 independent studies[8]) XCIE and XCI status
181 (respectively) in each cluster (**Fig. 2d)**.

182 To determine the reproducibility and validity of these unexpected modes of
183 dosage sensitivity in XCI and XCIE gene, we first confirmed that the distinct
184 expression profiles for these two clusters across karyotype groups were
185 reproducible at the level of individual genes and samples. Indeed, unsupervised
186 clustering of microarray samples based on expression of XCI and XCIE cluster
187 genes relative to XX controls distinguished three broad karyotype groups:
188 females with one X-chromosome (XO), males with one X-chromosome (XY,

189 XYY), and individuals with extra X-chromosome (XXX, XXY, XXYY) (**Fig. 2e**).
190 We were also able to validate our data-driven discovery of XCI and XCIE gene
191 clusters against independently generated X-chromosome annotations (**Fig. 2f**),
192 which detail 3 distinct genomic predictors of inactivation status for X-linked
193 genes. Specifically, XCI cluster genes were relatively enriched (and XCIE cluster
194 genes relatively impoverished) for (i) having lost a Y-chromosome homolog
195 during evolution[12] ($X^2 = 10.9$, p=0.01), (ii) being located in older evolutionary
196 strata of the X-chromosome[13] ($X^2 = 22.6$, p=0.007), and (iii) bearing
197 heterochromatic markers[14] ($X^2 = 18.35$, p=0.0004). Finally, qPCR assays in LCLs
198 from an independent sample of 401 participants with varying SCD (**Methods,**
199 **Supplementary Fig. 2c**) validated the fold changes observed in microarray data
200 for 6/7 of the most SCD-sensitive Yellow and Green cluster genes. To
201 independently validate these predictions, we assayed novel karyotype groups
202 (XXXY, XXXXY) not used to test the original model and were able to confirm
203 reduction in expression with greater X-chromosome dosage (*ARAF*, *NGFRAP1*,
204 *CXorf57*, *TIMP1*), and Y-chromosome dosage effects upon expression of X-
205 linked genes (*NGFRAP1*, *CXorf57*, *PIM2*, *IL2RG*, *MORF4L2*) (**Methods,**
206 **Supplementary Fig. 2d,e**). Our findings update the canonical Four Class Model
207 of SCD compensation for specific Y-linked and X-linked genes, and expand the
208 list of X-linked genes capable of mediating wider phenotypic consequences of
209 SCD variation.
210
211

212 **Figure 2 (next page). Data-Driven Partitioning of Sex Chromosome Genes by Dosage**
213 **Sensitivity. a)** 2D Multidimensional scaling (MDS) plot of sex chromosome genes by their mean
214 expression profiles across all 7 SCD groups. Genes are coded by both Four Class Model and k-
215 means cluster grouping. **b)** Cross table showing enrichment of k-means clusters (rows) for Four
216 Class Model gene groups. Lower-bounds for enrichment odds-ratios are given where mean
217 enrichment = ∞ . **c)** Dot and line plots showing observed and predicted mean expression values
218 for each k-means gene cluster across karyotype groups. **d)** Close-up of observed (solid, color-
219 coded) vs. predicted (dashed, gray) expression profiles of Green and Yellow gene clusters.
220 Observed mean expression profiles counter still counter predictions when analysis is restricted to
221 core genes in each cluster with XCIE/XCI status that has been confirmed across three
222 independent reports (Balaton et al). **e)** Heatmap showing normalized (vs. XX mean) expression of
223 dosage sensitive genes in the XCIE and XCI k-means groups (rows, color-coded green and
224 yellow respectively), for each sample (columns, color coded by SCD group). **f)** Pie-charts
225 showing how genes within XCIE and XCI-enriched k-means clusters (green and yellow columns
226 respectively), display mirrored over/under-representation for three genomic features of X-linked
227 gene that have been linked to XCEI in prior research (i) persistence of a surviving Y-linked
228 homolog, (ii) location of the gene within "younger" evolutionary strata of the X-chromosome, and
229 (iii) presence of euchromatic rather than heterochromatic epigenetic markers.
230
231


a
XIST
MDS2
MDS 1
FOUR CLASS
PAR
Y
XCIE
XCI
X-other
K MEANS
(not expressed)
b
FOUR CLASS GROUP
PAR
Y
XCIE
XCI
K MEANS GROUP
> 175.6
> 28.9
XIST
14.8
6.6
- log10 p value for enrichment
c
MEAN Log2 EXPRESSION (+/- 95% CI) OF EACH K MEANS GROUP BY KARYOTYPE
XIST
PAR
Y
XCIE
XCI
GROUP
Observed expression profiles
Expression profiles predicted by Four Class Model
d
XCIE
XCI
Log2 EXPRESSION (mean +/- 95% CI)
GROUP
Expression profiles
Predicted by Four Class Model
Observed (for all genes in cluster)
XCIE
XCI
Observed (for core genes only)
XCIE
XCI
e
KARYOTYPE
XO, XY, XYY, XXX, XXY, XXYY
K MEANS GROUP
XCIE
XCI
UPRT
SLC25A14
PPP2R3B
MORF4L2
TAF1
PNPLA4
FUNDC1
IL2RG
BCOR
ZRSR2
KDM5C
TXLNG
DDX3X
ZFX
PRKX
EIF2S3
EIF1AX
KDM6A
HDHD1
ZBTB33
ARAF
MST4
ELK1
CFP
SPIN3
FAM50A
MOSPD1
MMGT1
SATL1
ATP6AP1
GSPT2
G6PD
ARMCX5
GRPR
TCEAL4
GDI1
FAM127A
ZNF711
MPP1
CXorf57
-1
Log2 FC vs. XX
1
f
K-MEANS GROUP
XCI
XCIE
Y-LINKED HOMOLOGY
none / pseudogene / gene
X² = 10.98, p=0.01
EVOLUTIONARY STRATUM
OLDER
YOUNGER
X² = 22.6, p=0.007
CHROMATIN STATE
heterochromatin /other/ transcription
X² = 18.35, p=0.0004

233

234 **Context-Specific Disruption of Autosomal Expression by Sex Chromosome**
235 **Aneuploidy**

236 We next leveraged the diverse SCAs represented in our study to assess
237 how SCD variation shapes expression on a genome-wide scale. By counting the
238 total number of differentially expressed genes (DEGs, **Methods**) in each SCA
239 group relative to its respective euploidic control (i.e XO and XXX compared with
240 XX; XXY, XYY, XXYY compared with XY), we detected order of magnitude
241 differences in DEG count amongst SCAs across a range of log2 fold change
242 (log2FC) cut-offs (**Fig. 3a,b**). We observed an order of magnitude increase in
243 DEG count with X-chromosome supernumeracy in males vs. females, which
244 although previously un-described, is congruent with the more severe phenotypic
245 consequences of X-supernumeracy in males vs. females[15]. Overall, increasing
246 the dosage of the sex chromosome associated with the sex of an individual (i.e.
247 X in females and Y in males) had a far smaller effect than other types of SCD
248 changes. Moreover, the ~20 DEGs seen in XXX contrasted with >2000 DEGs in
249 XO – revealing a profoundly asymmetric impact of X-chromosome loss vs. gain
250 on the transcriptome of female LCLs, which echoes the asymmetric phenotypic
251 severity of X-chromosome loss (Turner) vs. gain (XXX) syndromes in females[6].

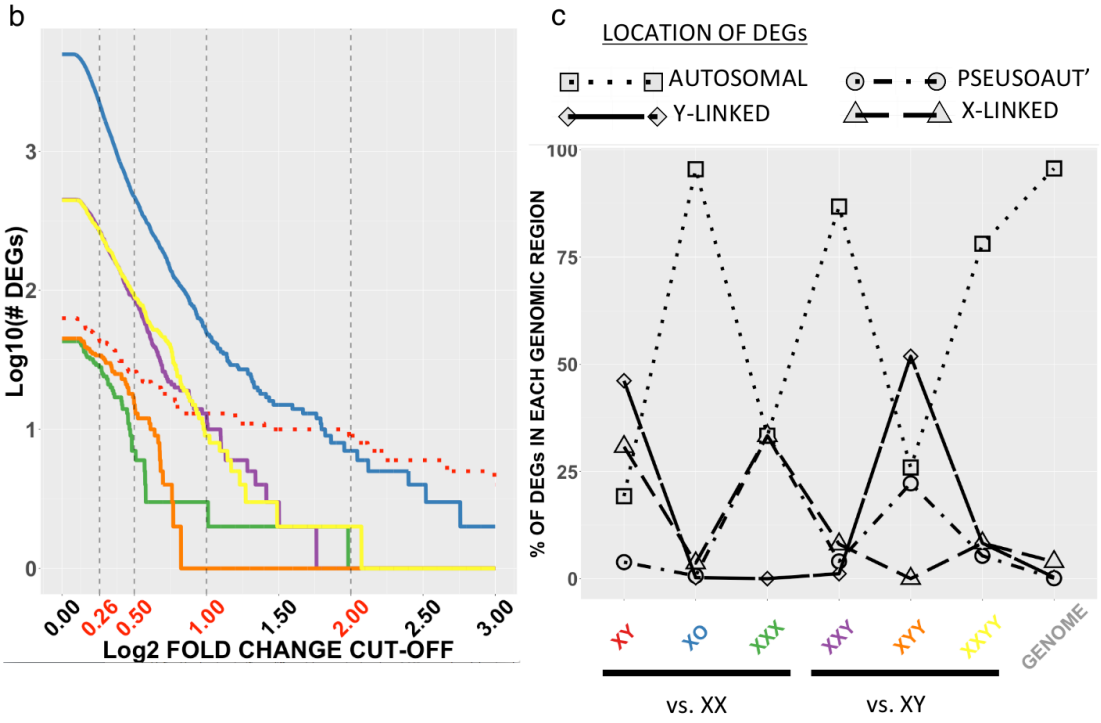
252 To clarify the relative contribution of sex chromosome vs. autosomal
253 genes to observed DEG counts with changes in SCD, we calculated the
254 proportion of DEGs in every SCD group (comparing SCAs to their "gonadal
255 controls", and XY males to XX females) that fell within each of four distinct

genomic regions: autosomal, PAR, Y-linked and X-linked (**Fig. 3c**). Autosomal genes accounted for >75% of all DEGs in females with X-monosomy (XO) and males with X-supernumeracy (XXY, XXYY), but <30% DEGs in all other SCD groups (**Methods**). These results reveal that SCD changes vary widely in their capacity to disrupt genome function, and demonstrate that differential involvement of autosomal genes is central to this variation. Moreover, associated SCA differences in overall DEG count broadly recapitulate SCA differences in phenotypic severity.

Figure 3 (next page). 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"). A |log2 fold change| threshold of 0.26 (~20% change in expression) was applied to categorically define differential expression in other analyses, by identifying the log2 fold change threshold increase causing the greatest drop in DEG count for each karyotype group, and then averaging this value across karyotype groups. **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.

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



283

284

285

286

287

288

289

290

291

292 **Sex Chromosome Dosage Regulates Large-Scale Gene Co-expression**
293 **Networks**

294 To provide a more comprehensive systems-level perspective on the
295 impact of SCD on genome-wide expression patterns, we leveraged Weighted
296 Gene Co-expression Network Analysis[16] (WGCNA, **Methods**). This analytic
297 approach uses the correlational architecture of gene expression across a set of
298 samples to detect sets (modules) of co-expressed genes. Using WGCNA, we
299 identified 18 independent gene co-expression modules in our dataset (**Table 3**).
300 We established that these modules were not artifacts of co-differential expression
301 of genes between groups by demonstrating their robustness to removal of all
302 group effects on gene expression by regression (**Supplementary Fig. 3a**), and
303 after specific exclusion of XO samples (**Supplementary Fig. 3b**) given the
304 extreme pattern of DE in this karyotype (**Fig. 3b**). We focused further analysis on
305 modules meeting 2 independent statistical criteria after correction for multiple
306 comparisons: (i) significant omnibus effect of SCD group on expression, (ii)
307 significant enrichment for one or more gene ontology (GO) process/function
308 terms (**Methods, Table 3, Fig. 4a-b**). These steps defined 8 functionally
309 coherent and SCD-sensitive modules (Blue, Brown, Green, Purple, Red, Salmon,
310 Tan and Turquoise). Notably, the SCD effects we observed on genome wide
311 expression patterns appeared to be specific to shifts in sex chromosome gene
312 dosage, as application of our analytic workflow to publically available genome-
313 wide Illumina beadarray expression data from LCLs in patients with trisomy 21
314 (Down syndrome) revealed a highly dissimilar profile of genome-wide expression

315 change to that observed in sex chromosome trisomies (**Methods, Fig. 3c, Table**
316 **4**).

317 To specify SCA effects on module expression, we compared all
318 aneuploidy groups to their respective "gonadal controls" (**Fig. 3d**). Statistically
319 significant differences in modular eigengene expression were seen in XO, XXY
320 and XXYY groups - consistent with these karyotypes causing larger total DE
321 gene counts than other SCD variations (**Fig. 3a**). The largest shifts in module
322 expression were seen in XO, and included robust up-regulation of protein
323 trafficking (Turquoise), metabolism of non-coding RNA and mitochondrial ATP
324 synthesis (Brown), and programmed cell death (Tan) modules, alongside down-
325 regulation of cell cycle progression, DNA replication/chromatin organization
326 (Blue, Salmon), glycolysis (Purple) and responses to endoplasmic reticular stress
327 (Green) modules. Module DE in those with supernumerary X chromosomes on
328 an XY background, XXY and XXYY, involved "mirroring" of some XO effects –
329 i.e. *down-regulation* of protein trafficking (Turquoise) and *up-regulation* of cell-
330 cycle progression (Blue) modules – plus a more karyotype-specific up-regulation
331 of immune response pathways (Red).

332 The distinctive up-regulation of immune-system genes in samples of
333 lymphoid tissue from males carrying a supernumerary X-chromosome carries
334 potential clinical relevance for one of the best-established clinical phenotypes in
335 XXY and XXYY syndromes: a strongly (up to 18-fold) elevated risk for
336 autoimmune disorders (ADs) such as Systemic Lupus Erythmatosus, Sjogren
337 Syndrome, and Diabetes Mellitus[7]. In further support of this interpretation, we

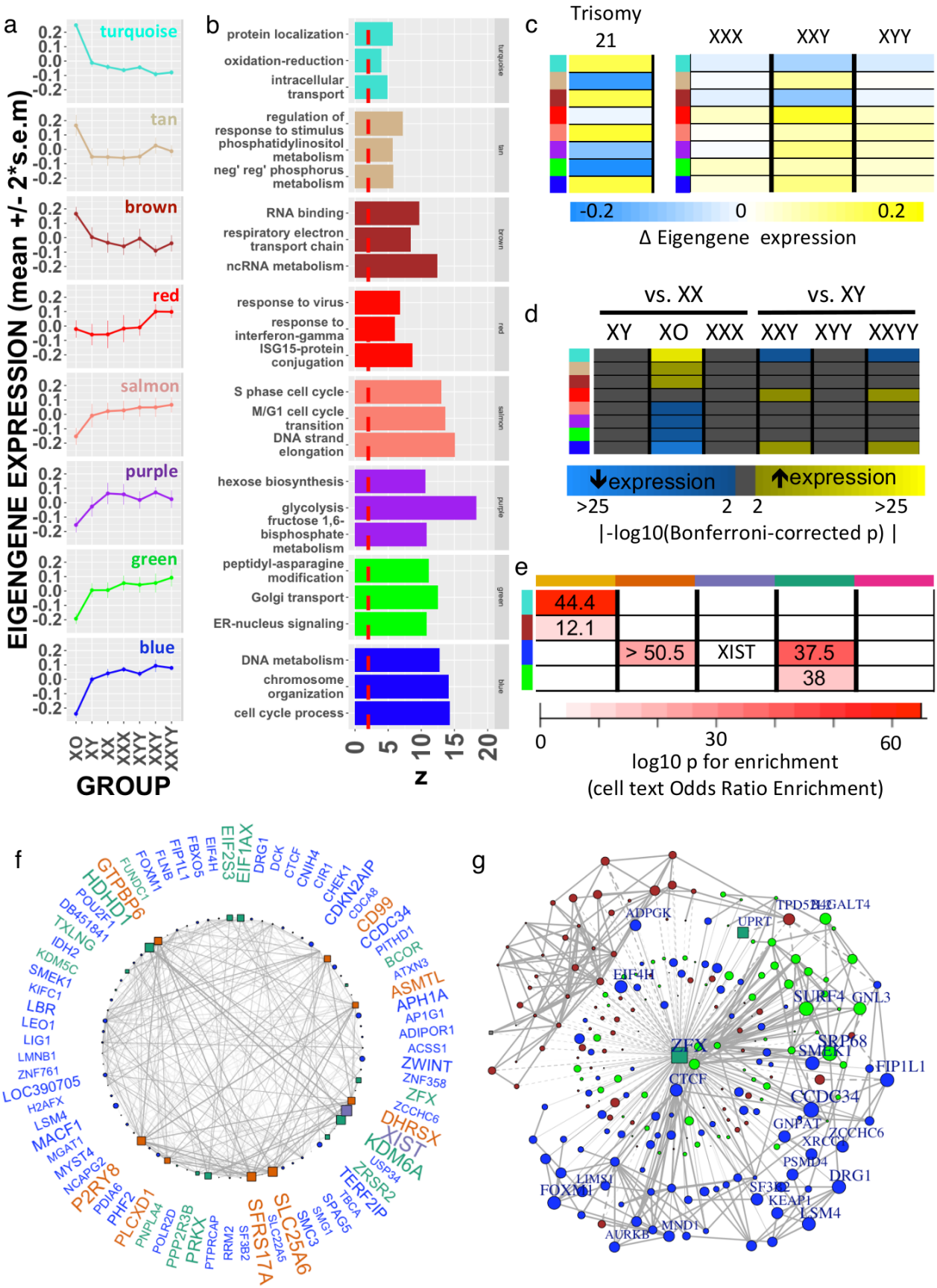
338 found the Red module to be significantly enriched (p=0.01 by Fisher's Test, and
339 p=0.01 by gene set permutation) for a set of known AD risks compiled from
340 multiple large-scale Genome Wide Association Studies (GWAS**, Methods**). The
341 two GWAS implicated AD risk genes showing strongest connectivity within the
342 Red module and up-regulation in males bearing an extra X-chromosome were
343 *CLECL1* and *ELF1* – indicating that these two genes should be prioritized for
344 further study in mechanisms of risk for heightened autoimmunity in XXY and
345 XXYY males. Collectively, these results represent the first systems-level
346 characterization of SCD effects on genome function, and provide convergent
347 evidence that increased risk for AD risk in XXY and XXYY syndrome may be
348 arise due to an up-regulation of immune pathways by supernumerary X-
349 chromosomes in male lymphoid cells.

350 To test for evidence of coordination between the changes in sex-
351 chromosome genes imparted by SCD (**Fig. 2**), and the genome-wide
352 transcriptomic variations detected through WGCNA (**Fig. 4a**), we asked if any
353 SCD-sensitive gene co-expression modules were enriched for one or more of the
354 5 SCD-sensitive clusters of sex chromosome genes (i.e. "PAR", "Y-linked",
355 "XCIE", "XCI" and the gene XIST). Four WGCNA modules - all composed of
356 >95% autosomal genes - showed such enrichment (**Fig. 4e**): The Turquoise and
357 Brown modules were enriched for XCI cluster genes, whereas the Green and
358 Blue modules were enriched for XCIE cluster genes. The Blue module was
359 unique for its additional enrichment in PAR genes, and its inclusion of XIST. We
360 generated network visualizations to more closely examine SCD-sensitive genes

361 and gene co-expression relationships within each of these four sex-chromosome
362 enriched WGCNA modules (**Fig 3f** Blue and **Supplementary Figure 3c-e** for
363 others**, Methods)**. The Blue module network highlights XIST, select PAR genes
364 (*SLC25A6*, *SFRS17A*) and multiple X-linked genes from X-Y gametolog pairs
365 (*EIF1AX*, *KDM6A* (*UTX*), *ZFX*, *PRKX*) for their high SCD-sensitivity, and shows
366 that these genes are closely co-expressed with multiple SCD-sensitive
367 autosomal genes including *ZWINT*, *TERF2IP* and *CDKN2AIP*.

368 Our detection of highly-organized co-expression relationships between
369 SCD sensitive sex-linked and autosomal genes hints at specific regulatory effects
370 of dosage sensitive sex chromosome genes in mediating the genome-wide
371 effects of SCD variation. To test this, and elucidate potential regulatory
372 mechanisms, we performed an unbiased transcription factor binding site (TFBS)
373 enrichment analysis of genes within Blue, Green, Turquoise and Brown WGCNA
374 modules (Methods). This analysis converged on a single TF - ZFX, encoded by
375 the X-linked member of an X-Y gametolog pair – as the only SCD sensitive TF
376 showing significant TFBS enrichment in one or more modules. Remarkably, the
377 gene *ZFX* was itself part of the Blue module, and ZFX binding sites were not only
378 enriched amongst Blue and Green module genes (increased in expression with
379 increasing X-chromosome dose), but also amongst Brown module genes that are
380 downregulated as X-chromosome dose increases (**Fig. 4g)**. To directly test if
381 changes in ZFX expression are sufficient to modify expression of Blue, Green or
382 Brown modules genes in immortalized lymphocytes, we harnessed existing
383 gene-expression data from murine T-lymphoblastic leukemia cells with and

384 without ZFX knockout[17] (GEO GSE43020). These data revealed that genes
385 downregulated by ZFX knockout in mice have human homologs that are
386 specifically and significantly over-represented in Blue (p=0.0005) and Green
387 (p=0.005) modules (p>0.1 for each of the other 6 WGCNA modules) – providing
388 experimental validation of our hypothesized regulatory role for ZFX.
389
390 **Figure 4 (next page). Weighted Gene Co-expression Network Analysis of Sex Chromosome Dosage**
391 **Effects. a)** Dot and line plots detailing mean expression (+/- 95% confidence intervals) by SCD group for 8
392 SCD-sensitive and functionally-coherent gene co-expression modules. **b)** Top 3 GO term enrichments for
393 each module. **c)** Heatmap showing distinct profile of module DE with a supernumerary chromosome 21 vs. a
394 supernumerary X-chromosome. **d)** Heatmap showing statistically significant differential expression of gene
395 co-expression modules between karyotype groups. **e)** Cross tabulation showing enrichment of each module
396 for the dosage-sensitive clusters of sex-chromosome genes detected by k-means. Lower-bounds for
397 enrichment odds-ratios are given where mean enrichment = ∞. **f)** Gene co-expression network for the Blue
398 module showing the top decile of co-expression relationships (edges) between the top decile of SCD-
399 sensitive genes (nodes). Nodes are positioned in a circle for ease of visualization. Node shape distinguishes
400 autosomal (circle) from sex chromosome (square) genes. Sex chromosome genes within the blue module
401 are color coded by their SCD-sensitivity grouping as per **Figure 2a** (PAR-Orange, XCIE – dark green, XIST
402 –purple). Larger node and gene name sizes reflect greater SCD sensitivity. Edge width indexes the strength
403 of co-expression between gene pairs. **g).** ZFX and its target genes from Blue, Green and Brown modules
404 with significant ZFX TFBS enrichment. Note that expression levels of ZFX (which increases in expression
405 with mounting X-chromosome dosage) are positively correlated (solid edges) with SCD sensitive genes that
406 are up-regulated by increasing X-chromosome dose (Blue and Green modules), but negatively correlated
407 (dashed edges) with genes that are down-regulated by increasing X-chromosome dose (Brown module).
408


a
EIGENGENE EXPRESSION (mean +/- 2*s.e.m)
turquoise
tan
brown
red
salmon
purple
green
blue
XO
XY
XX
XXX
XYY
XXY
XXYY
GROUP
b
protein localization
oxidation-reduction
intracellular transport
regulation of response to stimulus
phosphatidylinositol metabolism
neg' reg' phosphorus metabolism
RNA binding
respiratory electron transport chain
ncRNA metabolism
response to virus
response to interferon-gamma
ISG15-protein conjugation
S phase cell cycle
M/G1 cell cycle transition
DNA strand elongation
hexose biosynthesis
glycolysis
fructose 1,6-bisphosphate metabolism
peptidyl-asparagine modification
Golgi transport
ER-nucleus signaling
DNA metabolism
chromosome organization
cell cycle process
z
c
Trisomy 21
XXX
XXY
XYY
-0.2
0
0.2
Δ Eigengene expression
d
vs. XX
vs. XY
XY
XO
XXX
XXY
XYY
XXYY
expression
expression
>25
2
2
>25
|-log10(Bonferroni-corrected p) |
e
44.4
12.1
> 50.5
XIST
37.5
38
0
30
60
log10 p for enrichment
(cell text Odds Ratio Enrichment)
f
g
ADPGK
UPRT
EIF4H
SURF4
GNL3
ZFX
SRP68
SMEK1
FIP1L1
CTCF
CCDC34
GNPAT
ZCCHC6
XRCC1
PSMD4
DRG1
LIMS1
FOXM1
SF3B2
KEAP1
LSM4
AURKB
MND1

409

410

411 **DISCUSSION**

412 In conclusion, our study – which systematically examined gene expression
413 data from over 450 individuals representing a total of 9 different sex chromosome
414 karyotypes - yields several new insights into sex chromosome biology with
415 consequences for basic and clinical science. First, our discovery and validation of
416 X-linked genes that are upregulated by reducing X-chromosome count – so that
417 their expression is elevated in XO vs. XX for example - overturns classical
418 models of sex chromosome dosage compensation in mammals, and signals the
419 need to revise current theories regarding which subsets of sex chromosome
420 genes can contribute to sex and SCA-biased phenotypes[18]. Our findings also
421 challenge classical models of sex-chromosome biology by identifying X-linked
422 genes that vary in their expression as a function of Y–chromosome dosage –
423 indicating that the phenotypic effects of normative and aneuploidic variations in
424 Y-chromosome dose could theoretically be mediated by altered expression of X-
425 linked genes. Moreover, the discovery of Y chromosome dosage effects on X-
426 linked gene expression provides novel routes for competition between maternally
427 and paternally inherited genes beyond the previously described mechanisms of
428 parental imprinting and genomic conflict– with consequences for our mechanistic
429 understanding of sex-biased evolution and disease[19].

430 Beyond their theoretical implications, our data help to pinpoint specific
431 genes that are likely to play key roles in mediating sex chromosome dosage
432 effects on wider genome function. Specifically, we establish that a distinctive

433 group of sex-linked genes - notable for their evolutionary preservation as X-Y
434 gametolog pairs across multiple species, and the breadth of their tissue
435 expression in humans[10] – are further distinguished from other sex-linked genes
436 by their exquisite sensitivity to SCD, and exceptionally close co-expression with
437 SCD-sensitive autosomal genes. These results add critical evidence in support of
438 the idea that X-Y gametologs play a key role in mediating SCD effects on wider
439 genome function. In convergent support of this idea we show that (i) multiple
440 SCD-sensitive modules of co-expressed autosomal genes are enriched with
441 TFBS for an X-linked TF from the highly dosage sensitive *ZFX-ZFY* gemetolog
442 pair, and (ii) ZFX deletion causes targeted gene expression changes in such
443 modules.

444 Gene co-expression analysis also reveal the diverse domains of cellular
445 function that are sensitive to SCD – spanning cell cycle regulation, protein
446 trafficking and energy metabolism. These effects appear to be specific to shifts in
447 SCD as they are not induced by trisomy of chromosome 21. Furthermore, gene
448 co-expression analysis of SCD effects dissects out specific immune activation
449 pathways that are upregulated by supernumerary X-chromosomes in males, and
450 enriched for genes known to confer risk for autoimmune disorders that
451 overrepresented amongst males bearing an extra X-chromosome. Thus, we
452 report coordinated genomic response to SCD that could potentially explain
453 observed patterns of disease risk in SCA. Collectively, these novel insights serve
454 to refine current models of sex-chromosome biology, and advance our

455 understanding of genomic pathways through which sex chromosomes can shape
456 phenotypic variation in health, and sex chromosome aneuploidy.
457
458 **MATERIALS AND METHODS**
459
460 **Acquisition and preparation of biosamples**
461 RNA was extracted by standard methods (Qiagen, MD, USA) from
462 lymphoblastoid cell lines (LCLs) for 469 participants recruited through studies of
463 SCA at the National Institute of Health Intramural Research Program, and
464 Thomas Jefferson University [20]. All participants with X/Y-aneuploidy were non-
465 mosaic, and stability of karyotype across LCL derivation was confirmed by
466 fluorescent in situ hybridization in a randomly selected subset of samples used
467 for microarray analysis. Sixty-eight participants provided RNA samples for
468 microarray analyses (12 XO, 10 XX, 9 XXX, 10 XY, 8 XXY, 10 XYY, 9 XXYY),
469 and 401 participants provided RNA samples for a separate qPCR
470 validation/extension study (3 XO, 145 XX, 22 XXX, 146 XY, 34 XXY, 16 XYY, 17
471 XXYY, 8 XXXY, 10 XXXXY). The microarray and qPCR samples were fully
472 independent of each other (biological replicates), with the exception of 3 XO
473 participants in the microarray study, who each also provided a separate LCL
474 sample for the qPCR study (**Table 1**).
475
476

477 **Microarray data preparation, differential expression analysis, annotation**
478 **and probe selection**

479 Gene expression was profiled using the Illumina HT-12 v4 Expression
480 BeadChip Kit (Illumina Inc, San Diego, CA). Expression data were quantile
481 normalized across arrays and log2 transformed using the *limma* package in R[21].
482 For each of 47323 probes, we estimated mean expression by karyotype group,
483 and log2 fold change in gene expression for each unique pairwise group contrast
484 between karyotype groups (**Fig. 1a**), along with their associated false-discover-
485 rate (FDR) corrected p-values. For each pairwise karyotype group comparison,
486 we identified probes with significant log2 fold-changes that survived FDR
487 correction for multiple comparisons across all 47323 probes with q (the expected
488 proportion of falsely rejected nulls) set at 0.05. We also calculated a single
489 summary estimate of SCD effects for each probe, by calculating the proportion of
490 variance (r^2) in probe expression that was accounted for by the 7-level factor of
491 SCD group.

492 All 47323 microarray probes were annotated using both the vendor
493 manifest file and an independently published re-annotation that assigns a quality
494 rating to each probe based on the specificity of its alignment to the purported
495 transcript target [22]. We filtered for all probes with "perfect" or "good" quality
496 alignment to a known gene according to this reannotation, and then used the
497 *collapseRows* function from the WGCNA[23] package in R (with default settings), to
498 select one probe per gene. These steps resulted in high-quality measures of
499 expression and estimates of differential expression for 19984 autosomal and 894

500 sex-chromosome genes in each of 68 independent samples from 7 different
501 karyotype groups.

502 To select a log2 threshold for use in categorical definition of differentially
503 expressed genes (DEGs), we estimated DEG count across a range of absolute
504 log2 fold-change cut-offs for 6 contrast of primary interest: karyotypically normal
505 males vs. females (i.e. XY vs. XX), and each SCA group vs. its respective
506 "gonadal control" (i.e XX was the control for XO and XXX groups / XY was the
507 control for XXY, XYY and XXYY). An absolute log2 fold change of 0.26
508 (equivalent to a ~20% increase/decrease in expression) was empirically selected
509 as the cut-off to define differential expression by (i) separately identifying the
510 increase in log2FC threshold that cause the greatest drop in DEG count for each
511 SCA group, (ii) averaging these log2FC thresholds across all 5 SCA groups.

512

513 **Identifying Genes Showing Significant Differential Expression Across All**
514 **Sampled Changes in X- and Y-Chromosome Count**

515 The 21 unique pairwise karyotype group comparisons in our microarray
516 dataset included 15 contrasts involving a disparity in X-chromosome count, and
517 16 contrasts involving a disparity in Y-chromosome count (**Fig. 1a**). Using the
518 empirically-defined |log2 fold change| cut-off of 0.26 (see above), we screened all
519 20878 genes for evidence of significant differential expression across all
520 instances of X- or Y-chromosome disparity.

521

522 **Comparing SCD Sensitivity of Gametolog vs. non-Gametolog Sex-Linked**
523 **Genes**

524 Fourteen of 16 established[9] X-Y gametolog gene pairs were represented
525 in our microarray dataset. To compare the SCD sensitivity of this gene set to that
526 of non-gametolog sex-linked genes by we first quantified the mean effect of SCD
527 group on gene expression within the gametolog gene by averaging gene-wise r-
528 squared values for the effect of SCD group on expression. We then determined
529 the centile of this observed gene set mean r-squared against a distribution of r-
530 squared values for 10,000 similarly sized sets of randomly samples non-
531 gametolog sex-linked genes (**Fig. 1d**). This procedure was conducted separately
532 for X- and Y-chromosomes.

533

534 **A Priori Assignment of Sex Chromosome Genes to Four Class Model**
535 **Categories**

536 PAR genes were defined as those lying distal to the PAR1 and PAR2
537 boundaries specified in hg18 build of the human genome. Y-linked and X-linked
538 genes were defined as those lying proximal to these PAR boundaries on the Y-
539 and X-chromosome respectively. X-linked genes were assigned to XCIE and XCI
540 using consensus classifications from a systematic integration[8] of XCI calls from 3
541 large-scale assays: expression from the inactivated human X-chromosome in 9
542 hybrid human-mouse cell lines[24], allelic-imbalance analysis of expression data in
543 cell-lines from females with skewed X-inactivation[25], and X-chromosome
544 methylation data from microarray[26]. According to the XCI categories of this

545 consensus report we classified X-linked genes as X-inactivated ("Subject" or
546 "Mostly Subject" categories), X-escape ("Escape" or "Mostly Escape" categories),
547 or X-other (all other intermediate categories).
548
549 **Clustering of Sex Chromosome Genes by Dosage Sensitivity**
550 For all 894 sex chromosome genes within our dataset we calculated the
551 mean fold-change per SCD group relative to mean expression across all SCD
552 groups. The resulting 894 by 7 matrix was submitted to k-means clustering
553 across a range of k-values using the *kmeans* function in R with nstart and
554 iter.max set at 100. Visual inspection of a scree plot of mean within partition sum-
555 of-squared residuals against k indicated an optimal 6-cluster solution
556 (**Supplementary Fig. 2a**). The largest of these 6 clusters (Gray cluster) gathered
557 genes with low or undetectable expression levels across all samples, and was
558 excluded from further analysis.
559 Reproducibility of 6 cluster solution was established using a bootstrap
560 method whereby individuals were randomly drawn (with replacement) from each
561 SCD group within our microarray dataset to derive 1000 bootstrap sets of 68
562 samples. k-means clustering was repeated for each of these 1000 sets to define
563 a 6-cluster solution in each draw (**Supplementary Fig. 2b**). Consistency of
564 clustering was quantified for each gene as the proportion of bootstrap draws in
565 which it was assigned to the same cluster as it had been in the original sample.
566 The median consistency score for cluster designation was >93% for all 5 clusters
567 of SCD-sensitive sex chromosome genes.

568 The observed grouping of sex chromosome genes from k-mean
569 clustering, was compared with the predicted Four Class Model groupings using
570 two-tailed Fishers tests for all potential cluster-grouping combinations (**Fig. 2b**).
571

572 **Modelling X- and Y-chromosome Dosage Effects on Expression of Sex**
573 **Chromosome Gene Clusters**

574 We used the following linear models to estimate the combined influence of
575 X and Y chromosome dosage in cluster and gene-level expression of Yellow
576 (XCI enriched) and Green (XCIE enriched) gene clusters:
577 Expression ~ β_0 (intercept) + β_1(X_count) + β_2(Y_count) + error

578 We computed p-values for comparisons of both β_1 and β_2 coefficient
579 estimates against the null (0), and used these to test for significant directional
580 effects of sex chromosome dosage on the mean expression of each gene
581 cluster, as well the expression of individual genes within each cluster.
582

583 **Aligning Sex Chromosome Expression, Epigenetic and Evolutionary Data**

584 We validated our data-driven clustering of X-linked genes into Yellow (XCI
585 enriched) and Green (XCIE enriched) groups, by overlapping the genomic
586 coordinates of gene probes with segmentations of the X-chromosome according
587 to (i) "chromatin states" defined by computational analysis of coordinated
588 changes in 10 distinct chromatin marks in LCLs [14], (ii) "evolutionary strata"
589 reflecting staged loss of recombination between the X- and Y-chromosome [13].
590 Overlaps of our probe coordinate with these two annotations were defined using

591 the GenomicRanges package in R. As a third validation we also aligned our
592 clustering of X-lined genes with a previously published annotation of X-linked
593 genes according to whether their corresponding ancestral Y-linked homologue
594 has been lost, converted to a pseudogene or maintained [8]. Non-random
595 associations between these three annotations and Yellow vs. Green k-means
596 cluster membership were assessed using Chi-squared tests.
597

598 **Comparison of DEG Count and Genomic Distribution Across SCA Groups**

599 Total DEG counts were compared across SCD groups across a range of
600 log 2 fold change cut-offs as described above and reported in **Fig 3 a.b**. To test
601 for non-random distribution of DEGs across the genome in each SCD group **(Fig.**
602 **3c**), we compared observed DEG counts across 4 genomic regions - autosomal,
603 PAR, Y-linked and X-linked - to the background distribution of total gene counts
604 across these regions using the prop.test function in R. All SCD groups showed a
605 high non-random distribution of DEGs across the genome – reflecting preferential
606 involvement of sex chromosome genes ($p < 7.2 * 10^{-13}$).
607

608 **Quantitative rtPCR validation of Differentially-Expressed Genes in**
609 **Microarray**

610 *Selecting genes of interest:* For selected genes showing significant differential
611 expression between karyotype groups in our core sample, we used qPCR to
612 validate and extent observed fold-changes in an independent sample of 401
613 participants representing all the karyotypes in our core sample plus two

614 additional SCAs: XXXY, and XXXXY. For any given differential expression
615 analysis with microarray data, we selected genes for qPCR validation by
616 identifying which genes showed the three largest absolute fold-changes on
617 microarray in each relevant group contrast, and taking the union of these gene
618 lists.

619 *Fluidigm qPCR protocol:* Reverse Transcription reaction was performed using
620 RT2 HT First Strand Kit (QIAGEN, 330411) with 1000 ng RNA input per sample.
621 One-tenth of cDNA was preamplified using RT2 Microfluidics qPCR Reagent
622 system (QIAGEN, 330431) in combination with custom RT2 PreAmp pathway
623 primer mix Format containing 94 RT2 primer assays. Fourteen cycles of
624 preamplification were performed using the manufacturer recommended
625 preamplification protocol. Amplified cDNA was diluted 5-fold using RNase-free
626 water and assessed in real-time PCR using the RT2 Micrifluidics EvaGreen
627 qPCR Master mix and a Custom RT2 profiler PCR array PCR Array containing
628 96 assays, including selected DEGs of interest, housekeeping genes, reverse-
629 transcription controls, and positive PCR control. Real-time PCR was performed
630 on a Fluidigm BioMark HD (Fluidigm, San Francisco, US) using the RT2 cycling
631 program for the Fluidigm BioMark, which consists of an initial thermal mix stage
632 (50°C for 2 minutes, 70°C for 30 minutes, and 25°C for 10 minutes) followed by a
633 hot start at 95°C for 10 minutes and 40 cycles of 94°C for 15 seconds, and 60°C
634 for 60 seconds. For data processing, an assay with Ct > 23 was deemed to be
635 not expressed.

636 *Differential Expression Analysis of qPCR data*: For data analysis, the
637 ΔΔCtmethod of relative quantification was used[27]. The Ct values used for
638 calculating the fold changes were normalized using averaged expression of the
639 two housekeeping genes ACTB and B2M, which were not differentially
640 expressed across groups in either microarray or rtPCR data. All unique pairwise
641 group differences in expression were modeled using the limma R package with
642 identical setting to those used in analysis of microarray data (see above).
643 *Validation and Extension of Microarray Results Using qPCR Results:* All 21
644 unique pairwise SCD group contrasts in our microarray sample could be
645 reproduced in the independent qPCR dataset. We used the correlation in log 2
646 fold-change across these 21 contrasts to quantify the degree of agreement
647 between qPCR and microarray findings (**Supplementary Fig. 1a and 2d**). The
648 qPCR dataset also included two SCD groups that were not represented in the
649 microarray dataset – XXXY and XXXXY – allowing for a total of 15 novel pairwise
650 SCD group contrasts ("XO-XXXY", "XO-XXXXY", "XXXY-XX", "XXXXY-XX",
651 "XXX-XXXY", "XXX-XXXXY", "XXXY-XY", "XXXXY-XY", "XXY-XXXY",
652 "XXY-XXXXY", "XYY-XXXY", "XYY-XXXXY", "XXYY-XXXY", "XXYY-XXXXY",
653 "XXXY-XXXXY") sampling diverse disparities of X- and Y-chromosome dosage.
654 These novel contrasts were used as a further test for the validity and
655 reproducibility of our microarray findings. Each of the 15 novel pairwise SCD
656 group contrasts was coded according to three effects of interest: difference in X-
657 chromosome count, difference in Y-chromosome count, and difference in Y-
658 chromosome count given identical X-chromosome count. These coded SCD

659 disparities were then correlated with observed fold-changes in the qPCR dataset
660 for two gene sets of interest derived from microarray analyses: (i) the 10 sex-
661 linked genes with "obligatory" SCD-sensitivity (**Supplementary Fig. 1b**), (ii) X-
662 linked genes from the "Yellow" and "Green" k-means clusters which showed fold-
663 changes in microarray data that violated expectations of the classical Four Class
664 Model (**Supplementary Fig. 2e,f**).
665
666 **Weighted Gene Co-expression Network Analysis (WGCNA)**
667 *Defining Gene Co-expression Modules*: Gene co-expression modules were
668 generated using the R package Weighted Gene Co-expression Network Analysis
669 (WGCNA). Briefly, this involved first calculating the Pearson correlation
670 coefficient between all 20978 genes across all 68 samples in our study. This
671 correlation matrix was transformed using a signed power adjacency function with
672 a threshold power of 12 (selected based on fit to scale-free topology), and then
673 converted into Topological Overlap Matrix (TOM) by modifying the correlation
674 between each pair of genes using a measure of the similarity in their respective
675 correlations with all other genes[28]. The resulting TOM was then converted to a
676 distance matrix by subtraction from 1, and used to generate a dendrogram for
677 clustering genes into modules. Gene modules were defined using the Dynamic
678 Hybrid cutree function[29] [with the following parameter settings: deepSplit (control
679 over sensitivity of module detection to module splitting) = 2, mergeCutHeight
680 (distance below which modules are merged)= 0.25, minimum module size=30)].
681 Given the large number of genes included in our analyses, we implemented

682 module detection using the “blockwise” WGCNA algorithm, which starts with a
683 computationally inexpensive method to assort genes into smaller co-expression
684 blocks, and then completes the above steps within each block before merging
685 module designations across blocks. This implementation of WGCNA defined 18
686 mutually exclusive co-expression gene modules within our data, which ranged
687 from 45 to 1393 genes in size, and a left-over group of 14630 genes without
688 module membership (**Table 3**). The expression of each module was summarized
689 as a module eigengene value (ME: the right singular vector of standardized
690 expression values for genes in that module) in every sample. These ME values
691 were used to determine differential expression of modules across (omnibus F-
692 tests) and between (T-tests) SCD groups, as well as module co-expression
693 across samples (Pearson correlation coefficient).
694 *Further characterizing gene co-expression modules*: We used module
695 preservation analysis to establish that our defined co-expression modules were
696 not dominated by (i) the large number of DEGs induced by X-monosomy (using
697 expression data excluding XO samples), or (ii) other SCD group differences in
698 mean expression levels (using expression data after residualization for the
699 effects of SCD group and re-centering at a common mean). All modules showed
700 high reproducibility based on a module-specific $Z_{summary}$ scores derived by
701 comparing observed modular connectivity and density metrics with null values
702 generated by 200 permutations of gene-level module membership[30]. We focused
703 further characterization of modules which passed two independent statistical
704 criteria; (i) SCD sensitivity - quantified using F-tests for the omnibus effects of

705 karyotyope group on modular expression quantified as the ME, (ii) functional
706 coherence as inferred by analysis of modular gene ontolology term enrichments
707 using GO elite[31], and GOrilla[32].
708 *Testing for enrichment of autoimmune disorder risk genes in WGCNA modules:*
709 A large-scale records-based study was used to define 10 Autoimmune Disorder
710 (ADs) with clearly elevated prevalence rates in XXY vs. XY males[7], 9 of which
711 were represented in the largest available catalog of Genome Wide Association
712 Study (GWAS) findings (https://www.ebi.ac.uk/gwas/): Diabetes Mellitus type 1,
713 Multiple Sclerosis, Autoimmune Hypothyroidism, Psoriasis, Rheumatoid Arthritis,
714 Sjogren's Syndrome, Systemic Lupus Erythematosus, Ulcerative Colitis, and
715 Coeliac Disease. A total of 495 genes within our microarray sample were
716 annotated for showing a significant association in GWAS with one or more of
717 these 9 AD conditions. Overrepresentation of this AD gene set in the XXY
718 upregulated Red gene co-expression module was tested for using both Fisher's
719 exact test (p=0.01), and by comparing the observed representation of AD genes
720 against a null distribution generated by 10,000 random gene samples of equal
721 size to the red module.
722 *Testing for patterned enrichment of dosage sensitive sex chromosome genes in*
723 *WGCNA modules:* We tested if any of the 8 SCD-sensitive and functionally
724 enriched WGCNA modules showed enrichment for the previously derived k-
725 means clusters of dosage sensitive sex chromosome genes (**Fig. 2**) by applying
726 two-tailed Fishers tests to all pairwise module-cluster combinations (**Fig. 4 e**). All
727 observed associations survived Bonferroni correction for multiple comparisons.

728 *Module Visualization:* WGCNA co-expression modules were visualized by
729 selecting genes within the top decile of SCD-sensitivity (indexed using r-squared
730 for proportion of expression variance explained by group), and edges (co-
731 expression links between genes) in the top decile of edge strengths. All
732 visualizations were constructed using the igraph R package in R.
733 *Transcription factor binding site analyses:* Transcription factor binding site
734 (TFBS) enrichment analysis was performed each of the 4 SCD-sensitive
735 WGCNA modules - Blue, Green, Turquoise and Brown - that were enriched for
736 inclusion of gene from one or more of the 5 SCD-sensitive clusters of sex
737 chromosome genes. In each module, we scanned canonical promoter regions
738 (1000bp upstream of the transcription start site) for the top 500 genes with
739 strongest intramodular "connectivity" (based on kME - the magnitude of each
740 gene's coexpression with its module's eigengene). Next we utilized TFBS
741 position weight matrices (PWMs) from JASPAR database (205 non-redundant
742 and experimentally defined motifs)[33] to examine the enrichment for
743 corresponding TFBS within each module. For TFBS enrichment all the modules
744 were scanned with each PWMs using Clover algorithmhansen[34]. To compute the
745 enrichment analysis, we utilized three different background datasets (1000 bp
746 sequences upstream of all human genes, human CpG islands and human
747 chromosome 20 sequence). To increase confidence in the enrichment analyses,
748 we considered TFBS to be over-represented based on the P-values (<0.05)
749 obtained relative to all the three corresponding background datasets.

750 *Enrichment of SCD sensitive modules for genes with DE due to experimental*
751 *ZFX knockout.* To provide an orthogonal experimental test for evidence of a
752 regulatory role for ZFX within Blue, Green or Brown WGCNA modules, we used
753 a list of genes with significantly decreased expression due to ZFX knockout in
754 murine lymphocytes[17]. Human homologs were found for these genes
755 (http://www.informatics.jax.org/downloads/reports/index.html), and two-tailed
756 Fishers Tests were used to assess if these human genes were significantly
757 enriched/impoverished in any of the 8 SCD sensitive WGCNA modules.
758

759 **Comparison of Autosomal Gene Fold-Change in SCA and Down Syndrome**

760 The transcriptomic effects on Trisomy 21 (T21) were characterized in
761 LCLs by passing a publically available Illumina microarray gene expression
762 dataset (GEO, Accession number GSE34458) through an identical analytic
763 pipeline to that used in characterizing genome-wide fold changes in our SCA
764 sample (see above). We first independently confirmed the previously reported
765 finding that chromosome 21 was robustly enriched for genes showing differential
766 expression in this T21 data set (Chi-squared=999, p<2*10-16 for enrichment of
767 DEGs on chromosome 21) – buttressing use of these data to assess
768 transcriptomic effects of T21. We examined overlaps in genome-wide expression
769 change between T21 and the three sex chromosome trisomes in our samples
770 (XXY, XYY and XXX) using 17671 genes with complete expression data in both
771 microarray datasets (after exclusion of genes on chromosomes X, Y and 21).
772 We tested for, and failed to find any evidence of significant overlap in DEGs

773 using Chi-squared tests (**Table 4**). To test if T21 showed a similar shift in gene
774 co-expression modules to sex chromosome trisomies, we used the designation
775 of genes to modules in the SCA sample to recalculate module Eigengenes. We
776 then calculated MR fold changes for T21 and three SCA trisomies (XXX, XXY
777 and XYY). Trisomy of chromosome 21 was associated with a clearly distinct
778 profile of ME expression chance than all three of the SCA trisomies (**Fig. 4c**).

REFERENCES

1. Hughes, J. F. & Rozen, S. Genomics and genetics of human and primate y chromosomes. *Annu Rev Genomics Hum Genet* **13,** 83–108 (2012).
2. Arnold, A. P. The end of gonad-centric sex determination in mammals. *Trends Genet* **28,** 55–61 (2012).
3. Bermejo-Alvarez, P., Rizos, D., Rath, D., Lonergan, P. & Gutierrez-Adan, A. Sex determines the expression level of one third of the actively expressed genes in bovine blastocysts. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 3394–3399 (2010).
4. Petropoulos, S. *et al.* Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. *Cell* **165,** 1012–1026 (2016).
5. Belling, K. *et al.* Klinefelter syndrome comorbidities linked to increased X chromosome gene dosage and altered protein interactome activity. *Hum. Mol. Genet.* **26,** 1219–1229 (2017).
6. Hong, D. S. & Reiss, A. L. Cognitive and neurological aspects of sex chromosome aneuploidies. *Lancet Neurol.* **13,** 306–318 (2014).
7. Seminog, O. O., Seminog, A. B., Yeates, D. & Goldacre, M. J. Associations between Klinefelter's syndrome and autoimmune diseases: English national record linkage studies. *Autoimmunity* **48,** 125–128 (2015).

8. Balaton, B. P., Cotton, A. M. & Brown, C. J. Derivation of consensus inactivation status for X-linked genes from genome-wide studies. *Biol. Sex Differ.* **6,** 35 (2015).
9. Skaletsky, H. *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423,** 825–37 (2003).
10. Bellott, D. W. *et al.* Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* **508,** 494–499 (2014).
11. Deng, X., Berletch, J. B., Nguyen, D. K. & Disteche, C. M. X chromosome regulation: diverse patterns in development, tissues and disease. *Nat. Rev. Genet.* **15,** 367–378 (2014).
12. Wilson Sayres, M. A. & Makova, K. D. Gene survival and death on the human Y chromosome. *Mol. Biol. Evol.* **30,** 781–787 (2013).
13. Pandey, R. S., Wilson Sayres, M. A. & Azad, R. K. Detecting evolutionary strata on the human x chromosome in the absence of gametologous y-linked sequences. *Genome Biol. Evol.* **5,** 1863–1871 (2013).
14. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473,** 43–49 (2011).
15. Aksglaede, L. & Juul, A. Testicular function and fertility in men with Klinefelter syndrome: a review. *Eur. J. Endocrinol. Eur. Fed. Endocr. Soc.* **168,** R67-76 (2013).
16. Parikshak, N. N., Gandal, M. J. & Geschwind, D. H. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nat. Rev. Genet.* **16,** 441–458 (2015).

17. Weisberg, S. P. *et al.* ZFX controls propagation and prevents differentiation of acute T-lymphoblastic and myeloid leukemia. *Cell Rep.* **6,** 528–540 (2014).
18. Disteche, C. M. Dosage compensation of the sex chromosomes and autosomes. *Semin. Cell Dev. Biol.* **56,** 9–18 (2016).
19. Cocquet, J. *et al.* A genetic basis for a postmeiotic x versus y chromosome intragenomic conflict in the mouse. *PLoS Genet* **8,** e1002900 (2012).
20. Zinn, A. R. *et al.* A Turner syndrome neurocognitive phenotype maps to Xp22.3. *Behav. Brain Funct. BBF* **3,** 24 (2007).
21. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43,** e47 (2015).
22. Barbosa-Morais, N. L. *et al.* A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. *Nucleic Acids Res.* **38,** e17 (2010).
23. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9,** 559 (2008).
24. Carrel, L. & Willard, H. F. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434,** 400–4 (2005).
25. Cotton, A. M. *et al.* Analysis of expressed SNPs identifies variable extents of expression from the human inactive X chromosome. *Genome Biol.* **14,** R122 (2013).
26. Cotton, A. M. *et al.* Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation. *Hum. Mol. Genet.* **24,** 1528–1539 (2015).

27. Schmittgen, T. D. & Livak, K. J. Analyzing real-time PCR data by the comparative C(T) method. *Nat. Protoc.* **3,** 1101–1108 (2008).
28. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4,** Article17 (2005).
29. Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinforma. Oxf. Engl.* **24,** 719–720 (2008).
30. Langfelder, P., Luo, R., Oldham, M. C. & Horvath, S. Is my network module preserved and reproducible? *PLoS Comput. Biol.* **7,** e1001057 (2011).
31. Zambon, A. C. *et al.* GO-Elite: a flexible solution for pathway and ontology over-representation. *Bioinforma. Oxf. Engl.* **28,** 2209–2210 (2012).
32. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10,** 48 (2009).
33. Portales-Casamar, E. *et al.* JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **38,** D105-110 (2010).
34. Frith, M. C. *et al.* Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res.* **32,** 1372–1381 (2004).

TABLE LEGENDS (Tables provided as individual .xlsx files due to size)

Table 1. 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. *XO samples in PCR validation study are an independent sample of cells drawn from 3 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).

Table 2. 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. 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.

Table 3. 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

Table 4. 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