Strong Constraint on Human Genes Escaping X-Inactivation Is Modulated by their Expression Level and Breadth in Both Sexes

Andrea Slavney,[1,2] Leonardo Arbiza,[1] Andrew G. Clark,*[1,2] and Alon Keinan*[1]

[1]Department of Biological Statistics and Computational Biology, Cornell University

[2]Department of Molecular Biology and Genetics, Cornell University

***Corresponding author:** E-mail: ak735@cornell.edu; ac347@cornell.edu.

Associate editor: Joshua Akey

Abstract

In eutherian mammals, X-linked gene expression is normalized between XX females and XY males through the process of X chromosome inactivation (XCI). XCI results in silencing of transcription from one ChrX homolog per female cell. However, approximately 25% of human ChrX genes escape XCI to some extent and exhibit biallelic expression in females. The evolutionary basis of this phenomenon is not entirely clear, but high sequence conservation of XCI escapers suggests that purifying selection may directly or indirectly drive XCI escape at these loci. One hypothesis is that this signal results from contributions to developmental and physiological sex differences, but presently there is limited evidence supporting this model in humans. Another potential driver of this signal is selection for high and/or broad gene expression in both sexes, which are strong predictors of reduced nucleotide substitution rates in mammalian genes. Here, we compared purifying selection and gene expression patterns of human XCI escapers with those of X-inactivated genes in both sexes. When we accounted for the functional status of each ChrX gene's Y-linked homolog (or "gametolog"), we observed that XCI escapers exhibit greater degrees of purifying selection in the human lineage than X-inactivated genes, as well as higher and broader gene expression than X-inactivated genes across tissues in both sexes. These results highlight a significant role for gene expression in both sexes in driving purifying selection on XCI escapers, and emphasize these genes' potential importance in human disease.

Introduction

In eutherian mammals, X chromosome inactivation (XCI) (Lyon 1962) silences transcription from one ChrX homolog in each female somatic cell. This phenomenon is the result of the need to achieve dosage compensation between XX females and XY males due to extensive divergence and degeneration of ChrY over the course of mammalian evolution (Ohno 1967; Charlesworth 1991; Skaletsky et al. 2003). At a glance, the relationship between Y gametolog functionality and X gametolog inactivation seems straightforward: XCI rarely occurs in the pseudoautosomal regions (PARs) (Berletch et al. 2011), which contain most of the remaining functional ChrY genes, and recent lineage-specific Y gametolog loss within the eutherian clade correlates with XCI of the corresponding X gametolog (Jegalian and Page 1998). However, this pattern is disrupted by a set of genes outside of the PARs that show low but significant expression from the inactive ChrX in female cells, and are therefore said to "escape" XCI. The exact loci and extent of XCI escape varies between and within species (Yang et al. 2010; Wang et al. 2012, 2014), with at least 100 human ChrX genes having been observed to escape XCI to some extent in at least one tissue (Carrel and Willard 2005; Cotton et al. 2015; Schultz et al. 2015). Although XCI escape is expected for genes that possess a functional Y gametolog, only 12 human non-PAR XCI escapers have a functional Y gametolog (Lahn and Page 1999; Wilson-Sayres and Makova 2013). As such, co-occurrence with a functional Y gametolog provides at best a partial characterization of the evolutionary basis of XCI escape.

High sequence conservation across primates in genes that escape XCI (Park et al. 2010) strongly suggests that XCI escape is directly or indirectly a consequence of natural selection rather than neutral mechanisms, such as inefficient establishment or maintenance of X-inactivation (Brown and Greally 2003; Midgeon 2014). In addition to their observation of stronger sequence conservation in XCI escapers, Park et al. (2010) showed that human XCI escaper gene expression levels were more conserved across primates than those of X-inactivated genes. Expression level (Drummond et al. 2005; Wall et al. 2005) and, to a greater extent, expression breadth (Duret and Mouchiroud 2000; Park and Choi 2010) are known to correlate strongly with nucleotide substitution rates in mammals: Highly, broadly expressed genes evolve more slowly than weakly expressed or tissue-specific genes. Therefore, it is likely that gene expression level and/or breadth contribute to the high sequence conservation observed among XCI escapers. Importantly, this is a distinct hypothesis from that of selection on female-biased expression of XCI escapers resulting from XCI escape, which may contribute to sex-specific phenotypes (Trabzuni et al. 2013; Deng et al. 2014).

Although previous studies have provided essential insights into the evolution of XCI, none have specifically focused on the relationship between purifying selection and gene expression across all XCI escapers compared with other X-linked genes. In investigating this relationship, it is important to consider that the functional status of a ChrX gene's Y gametolog is a strong predictor of its evolutionary rate (Park et al. 2010). In particular, XCI escapers with functional Y gametologs are clear outliers on ChrX in terms of their high sequence conservation across mammals (Bellott et al. 2014). Therefore, to capture the effect of XCI status on patterns of selection independent of Y gametolog status, genes with functional Y gametologs should be treated as distinct subgroups within each XCI category.

In order to investigate the possibility that gene expression patterns in both sexes contribute to the differential signals of purifying selection between XCI escapers and X-inactivated genes, we used divergence data inferred from publicly available primate genomes, human polymorphism data, and human RNA-seq data to examine the relationship between purifying selection and gene expression within XCI categories (XCI escapers vs. X-inactivated genes), controlling for Y gametolog functionality. We first separated genes with functional Y gametologs from the rest so that we could determine the extent to which the stronger signal of purifying selection in XCI escapers is dependent on the inclusion of these highly conserved genes. Then, we compared gene expression level and breadth across XCI and Y gametolog status combinations in females and males in multiple primary tissues.

Results

Determining XCI Status

We assigned XCI statuses to unique protein-coding ChrX genes outside of the PARs that met several quality control criteria (see Methods and supplementary figure S1, Supplementary Material online) compiled from three studies that used different methods to determine XCI status: Carrel and Willard (2005), Cotton et al. (2015), and Schultz et al. (2015). A summary of the assays and XCI classifications used in each study is shown in supplementary figure S2, Supplementary Material online, and XCI status calls in for all genes across all three studies are available in supplementary table S1, Supplementary Material online. As each study used different criteria to define XCI escape, the exact genes assayed and their XCI status assignments differed across these data sets, with many genes assayed in only one of the three. As such, we elected to combine the three data sets in order to maximize the number of genes in our analysis. Aiming for robust XCI classifications, we drew on two important observations from Cotton et al. (2015), which reported XCI statuses across many individuals in 27 tissues: 1) For most genes, XCI status is highly consistent across tissues—only three of approximately 500 transcripts escaped XCI in one tissue but were X-inactivated in all others and 2) variable escape is tissue specific, as most genes that escape in only some individuals (called "heterogeneous escapers" in Carrel and Willard 2005 and "variable" escapers in Cotton et al. 2015) either escape XCI or are inactivated in most other tissues.

Based on these observations, we limited the number of XCI status categories within each data set to two—XCI escapers and X-inactivated genes—either by excluding variable (for Carrel and Willard 2005 and Cotton et al. 2015) and tissue-specific (for Schultz et al. 2015) escapers, or by splitting these genes between the XCI escaper and X-inactivated categories based on their behavior in the majority of individuals and/or tissues (supplementary table S2, Supplementary Material online). We considered each data set individually, but also determined a consensus XCI status for each gene across the three studies using a more inclusive definition of XCI escape. In this "combined" data set, we classified genes as XCI escapers if they showed significant evidence of escape in any number of tissues or individuals in any study. We classified all other genes as X-inactivated. This inclusive definition left us with a list of 248 ChrX genes that have been shown to escape XCI in at least one individual in one or more tissues, and 238 that showed no signs of XCI escape in any data set. We report results for the combined data set throughout the main text.

XCI Escapers with Functional Y Gametologs Do Not Drive Greater Signals of Purifying Selection in XCI Escapers compared with X-Inactivated Genes

For each XCI/Y status combination, we calculated two statistics derived from a McDonald and Kreitman (1991) framework: *N*, the fraction of sites evolving neutrally; and *S*, the fraction of sites evolving under strong purifying selection (Mackay et al. 2012). Weak purifying selection and population expansion can both produce excesses of rare polymorphisms (Keinan and Clark 2012; Gao and Keinan 2014), which bias the McDonald–Kreitman (MK) test. However, these statistics provide some differentiation between weak purifying selection and neutral polymorphism inflation, and all gene categories in this analysis are expected to share the same degree of neutral polymorphism inflation due to their shared history on ChrX. We calculated the XCI escaper-to-X-inactivated ratios of *S* and *N* within each Y gametolog category by pooling polymorphism and divergence data across genes within each category, and resampling gene category membership with replacement 1,000 times to account for variability in the extent of selection across genes. In cases where XCI escapers are undergoing more purifying selection than X-inactivated genes, these ratios are expected to be >1 for *S* and <1 for *N*. The three Y gametolog categories we considered were "functional Y" (the ChrX gene has retained a functional, but not identical Y gametolog), "pseudogenized Y" (the ChrX gene has a Y gametolog that is nonfunctional but detectable by sequence homology), and "lost Y" (no Y gametolog can be detected by sequence homology). There are a total of 15 non-PAR genes in the functional Y gametolog category, 11 of which escape XCI and only 4 of which undergo X-inactivation, which severely limits the power of any statistical tests used to detect differences between them. However, the direction of the relationship between XCI escapers and X-inactivated genes in this group is still somewhat informative, and in light of their high biological significance (Bellott et al. 2014) we report results for this class throughout the main text.

Using XCI status assignments from the combined data set (described above), the *S* ratio was significantly greater than 1 and the *N* ratio was significantly lesser than 1 in all but the

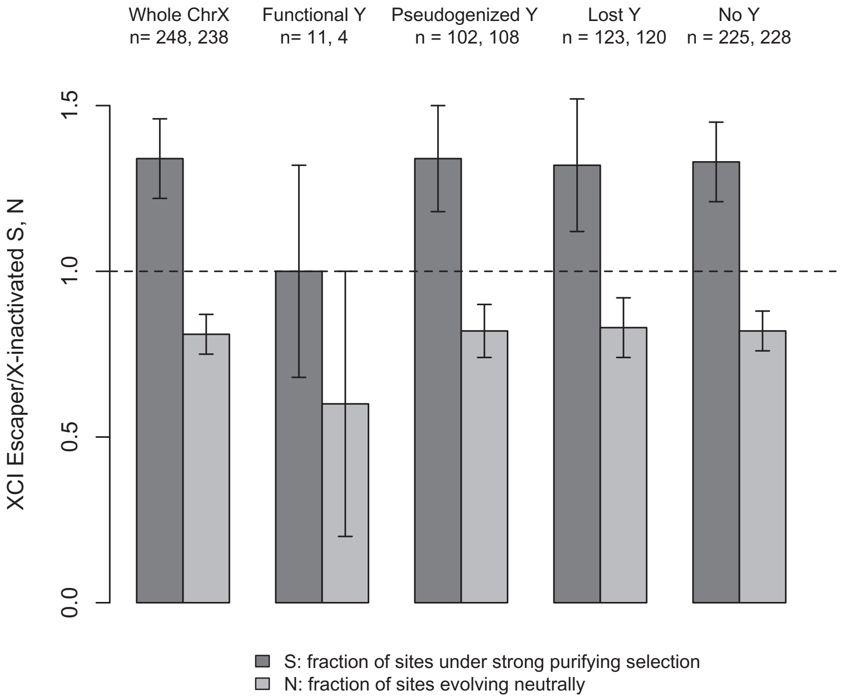

Fig. 1. XCI escapers show larger fractions of sites under strong purifying selection (*S*) and smaller fractions of neutral sites (*N*) than X-inactivated genes with the same Y gametolog status. XCI escaper-to-X-inactivated ratios of the values of *N* and *S* are presented, with standard deviations based on 1,000 bootstrap samples of genes in each XCI/Y category (error bars denote ±1 standard deviation). The numbers (*n*) of XCI escaper and X-inactivated genes in each Y gametolog category are shown below the category names at the top of the plot. The whole ChrX category includes genes from all other categories, as well as 18 genes that did not have a Y gametolog status assignment (12 XCI escapers and 6 X-inactivated genes). The no Y category is the union of the pseudogenized and lost Y categories.

Table 1. Resampling Mean and Standard Deviation of *S* and *N* for XCI/Y Categories

Y category	XCI category	*n* genes	*S*	*N*
All	XCI escapers	248	0.52 ± 0.02	0.46 ± 0.02
	X-inactivated	238	0.39 ± 0.03	0.57 ± 0.03
Functional Y	XCI escapers	11	0.60 ± 0.06	0.37 ± 0.05
	X-inactivated	4	0.63 ± 0.13	0.32 ± 0.14
Pseudogenized Y	XCI escapers	102	0.52 ± 0.03	0.47 ± 0.03
	X-inactivated	108	0.39 ± 0.04	0.58 ± 0.04
Lost Y	XCI escapers	123	0.50 ± 0.04	0.47 ± 0.04
	X-inactivated	120	0.38 ± 0.04	0.57 ± 0.04
No Y (Pseudo. Y + Lost Y)	XCI escapers	225	0.52 ± 0.02	0.47 ± 0.02
	X-inactivated	228	0.39 ± 0.03	0.57 ± 0.03

functional Y gametolog category (fig. 1 and table 1). To ensure that these results were not dependent on the inclusion of variable and/or tissue-specific XCI escapers, we calculated the same ratios by either excluding or including variable and tissue-specific XCI escapers in each individual XCI status data set and in the combined data set (supplementary table S2, Supplementary Material online). We observed that the *S* and *N* ratios were significantly greater and less than 1 (respectively) in at least one of the pseudogenized or lost Y gametolog categories in each XCI status data set. Additionally, the *S* and *N* ratios for each XCI/Y category did not change significantly within each data set when variable and tissue-specific escapers were excluded. We also observed that *S* and *N* ratios were rarely significantly different between the categories of genes with pseudogenized and lost Y gametologs (table 1 and supplementary table S2, Supplementary Material online). Therefore, for the remainder of our analyses, we combined the pseudogenized and lost Y gametolog categories within each XCI category into a single "no Y" gametolog category. Combined, the results presented in this section show that the signal of stronger purifying selection in XCI escapers compared with X-inactivated genes 1) is evident across the different methods and definitions used to define XCI escape in each of the previous studies, 2) is not solely driven by XCI escapers with functional Y gametologs, and 3) is generally robust to the exclusion of variable or tissue-specific XCI escapers.

XCI Escapers Show Higher Gene Expression in Both Sexes than X-Inactivated Genes

We compared the expression of XCI escapers and X-inactivated genes across 23 broad tissue types in females and males separately using RNA-seq data from the GTEx Consortium (2015). We first obtained a single reads per kilobase per million mapped reads (RPKM) value for each gene in each tissue by averaging significant expression values across individuals (see Methods), and observed that XCI escapers were generally more highly expressed than X-inactivated genes in individual tissues in both sexes (fig. 2*a* and supplementary table S3, Supplementary Material online). We then obtained a global expression value for each gene by taking the average of its expression values across tissues (fig. 2*b* and supplementary table S4, Supplementary Material online). To assess the significance of the differences between XCI escapers and X-inactivated genes across all tissues, we calculated the average global expression value across genes in each XCI/Y category, and calculated the XCI escaper-to-X-inactivated ratio of these values in each Y gametolog category. We permuted gene membership between the XCI groups in each Y gametolog category to obtain a *P* value for this ratio. Within the no Y gametolog category, the XCI escaper global expression mean was significantly higher than the X-inactivated global expression mean in both sexes. The XCI escaper-to-X-inactivated global expression mean ratio was 1.97 ($P << 10^{-3}$) in females and 2.06 in males ($P << 10^{-3}$). In the functional Y gametolog category, there was no significant difference between the global expression mean of significantly expressed X-inactivated genes and XCI escapers: The XCI escaper/X-inactivated global expression mean ratio was 0.867 ($P = 0.529$) in females and 1.204 in males ($P = 0.434$).

Comparing the global expression ratios of XCI/Y categories across various minimum expression level cutoffs (supplementary fig. S4, Supplementary Material online) showed that the XCI escaper-to-X-inactivated global expression ratio in the functional Y class was highly sensitive to the minimum expression cutoff. The purpose of using these cutoffs was to avoid using RPKM values that are indistinguishable from background signal, but this results in a substantial loss of information in some XCI/Y categories (this is readily apparent in supplementary fig. S5, Supplementary Material online, which is described in the expression breadth section of the Results). Therefore, to examine the impact of these weak expression values on global expression trends without making assumptions about what constitutes significant expression, we performed a nonparametric rank-based comparison of XCI/Y category expression.

For each XCI/Y category, we obtained a single expression distribution by averaging unfiltered RPKM values across individuals for each gene in each tissue, resulting in a distribution of length $n \times 23$ (where n is the number of genes in the XCI/Y category and 23 is the number of tissues). We then performed the Mann–Whitney *U* test to contrast the XCI escaper and X-inactivated RPKM distributions in each Y gametolog category. In females, the global Mann–Whitney *U* test *P* values were 0.010 for XCI escapers > X-inactivated in the small functional Y gametolog category, and 2.2×10^{-16} for XCI escapers > X-inactivated in the much larger no Y gametolog category. In males, the XCI escapers > X-inactivated *P* values were 0.047 and 2.2×10^{-16} for the functional and no Y gametolog categories. This test highlights the fact that X-inactivated genes with functional Y gametologs are a unique group showing extremely weak expression in many tissues but unusually high expression in others, whereas the XCI escapers with functional Y gametologs are more uniformly highly expressed. These results further support our overall conclusion that XCI escapers tend to be more highly expressed than X-inactivated genes.

XCI Escapers Show Broader Gene Expression in Both Sexes than X-inactivated Genes

XCI escapers showed significantly lower values of the tissue-specificity index τ (Yanai et al. 2005; Methods) than X-inactivated genes in both sexes and in both Y gametolog categories (fig. 3 and table 2), indicating that they are more broadly expressed than X-inactivated genes. As an additional expression breadth measurement, we calculated the proportion of tissues for which each gene showed significant expression (Methods) and observed that XCI escapers were significantly expressed in a higher proportion of tissues than X-inactivated genes with the same Y gametolog status in both sexes (supplementary fig. S5, Supplementary Material online). Consistent with the previous work demonstrating that XCI escapers with functional Y gametologs are enriched for highly conserved functions (Bellott et al. 2014), these genes showed the lowest τ values in both sexes across all XCI/Y categories.

Discussion

In summary, we report two findings: First, XCI escapers exhibit significantly greater degrees of purifying selection than X-inactivated genes, even when the highly conserved XCI escapers with functional Y gametologs are excluded. Second, XCI escapers show higher and broader average gene expression than X-inactivated genes across tissues in both sexes. This is, to our knowledge, the first study to examine the possibility of a connection between purifying selection on XCI escapers and gene expression across all XCI escapers, particularly those lacking a functional Y gametolog. Our results suggest that purifying selection on human XCI escapers is partially driven by high and broad expression of these genes (relative to other genes on ChrX). The sensitivity of XCI escapers to reductions in gene expression level can be directly tested with experiments in model organisms or human cell lines.

Although our analyses indicate that gene expression and purifying selection on XCI escapers are likely connected, the nature of the relationship between XCI status and expression is not completely clear. Another facet of this relationship which we have not investigated in this study is sex-biased expression. Specifically, female-specific dosage sensitivity for highly constrained XCI escapers, such as those involved in fertility and cognition, may partially explain the correlation between expression level and XCI escape. In humans, reduced expression of several XCI escapers has been implicated in phenotypes associated with Turner syndrome (45, X), including ovarian failure and various neurological abnormalities (Ellison et al. 1997; Bione et al. 1998; Zinn and Ross 1998). Another study identified several XCI escapers that were inferred to be highly dosage sensitive to be promising candidates for various X aneuploidy syndromes (Pessia et al. 2012). However, to date, few human XCI escapers have been shown to actually exhibit large female expression biases. For instance, in one study, only 5% of 299 ChrX genes surveyed showed significant female expression biases in at least one of 11 human tissues, and only 6 of these escape XCI (Talebizadeh et al. 2006). However, further investigation using improved technologies and larger numbers of individuals may reveal

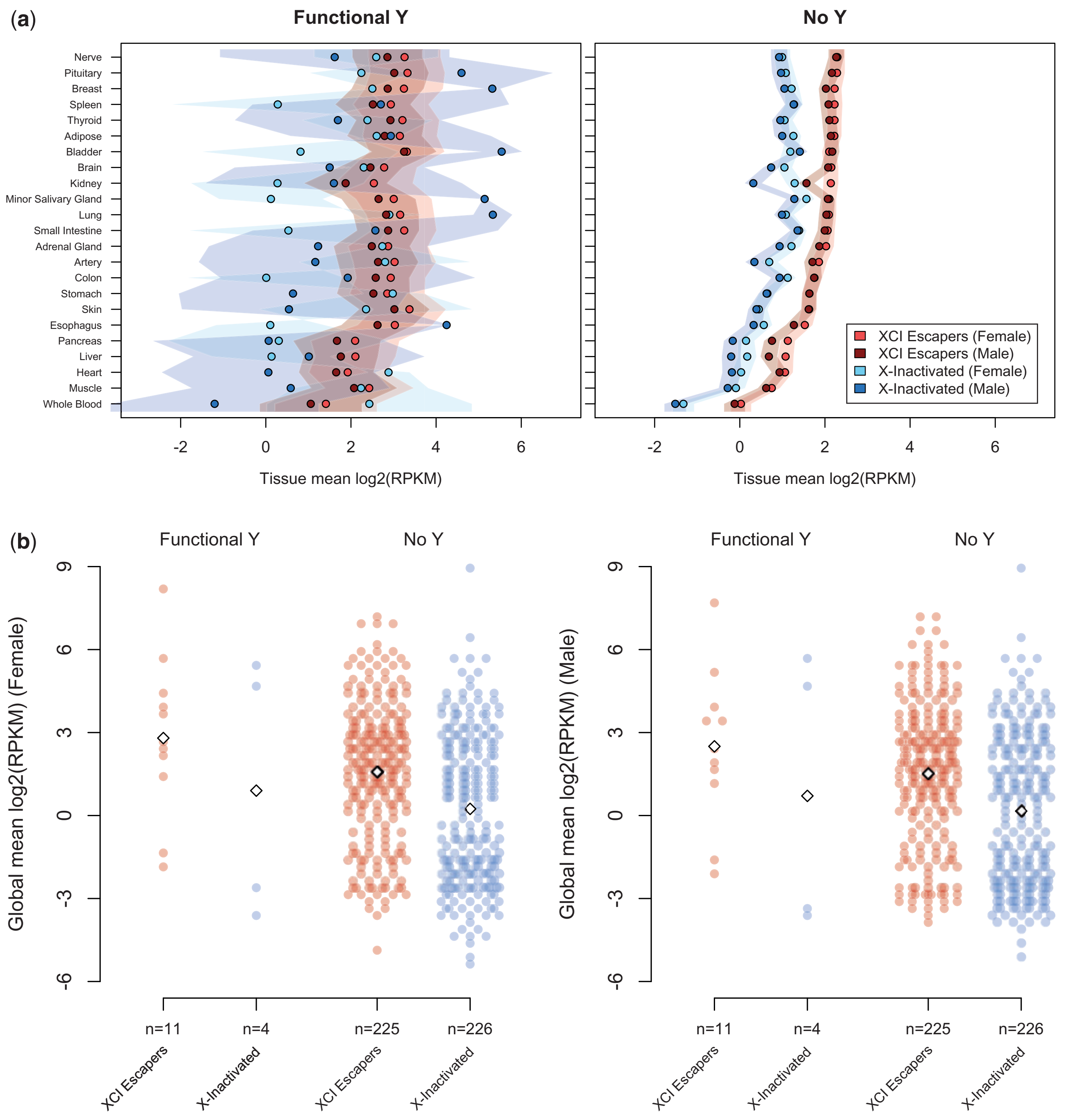

Fig. 2. (*a*) XCI category average gene expression by Y gametolog status, tissue, and sex. Scatterplots show the XCI/Y category averages of gene expression values for each sex in each GTEx tissue. Shaded areas show the standard deviation of the average across genes. In both plots, individual expression values were filtered by tissue-specific cutoffs (see Methods) before being incorporated into gene averages. (*b*) Global mean expression for each gene by XCI status, Y gametolog status, and sex. Each point in these one-dimensional scatterplots shows the global mean expression value of a single gene in female (left) and male (right) samples. The white diamonds indicate the mean expression value of the genes in that XCI/Y category. Areas of darker color indicate overlap of two or more points. In both plots, individual expression values were filtered by tissue-specific cutoffs (see Methods) before being incorporated into gene averages.

female expression biases in more XCI escapers. Alternatively, XCI escape may be an indirect consequence, rather than a driver, of high expression via any of a number of molecular mechanisms, such as local perturbations in chromatin state.

In this work, we have collected and analyzed data from three of the most recent and comprehensive studies with information on human XCI escape. Our results highlight the variability and uncertainty in the definition of variable XCI escapers. We observed that many genes annotated as variable escapers in one data set were annotated as consistently escaping or X-inactivated in one of the others (supplementary table S2, Supplementary Material online). This may

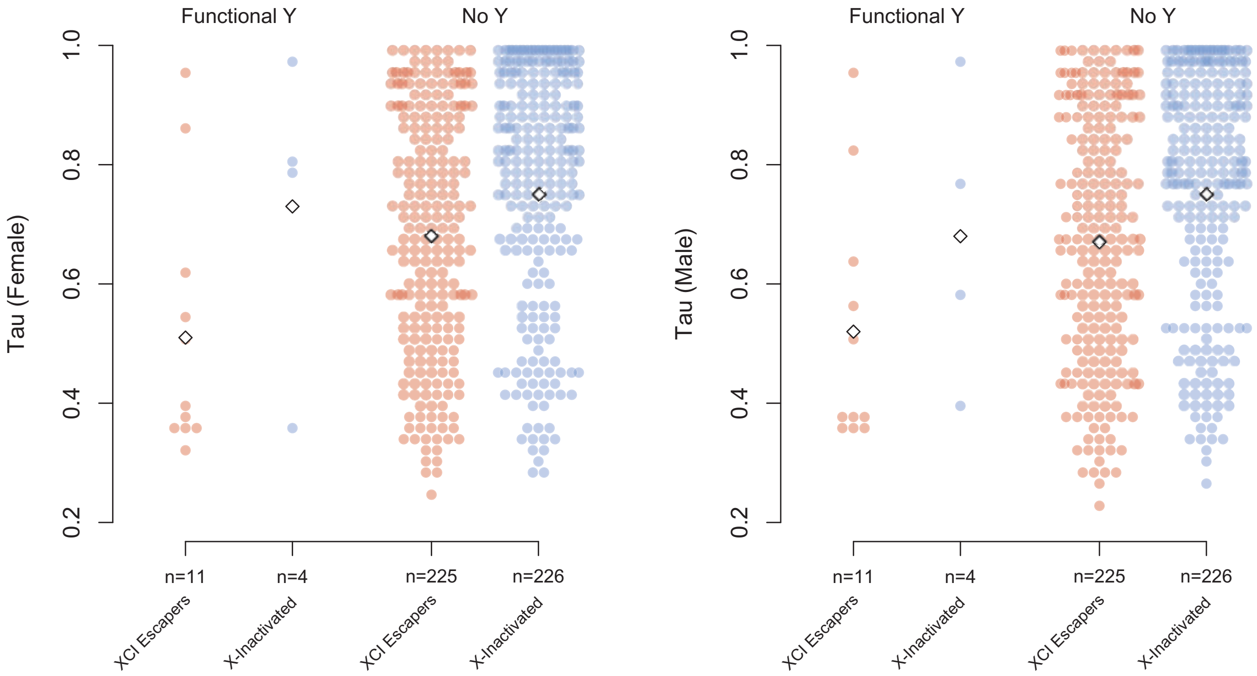

Fig. 3. XCI escapers are more broadly expressed than X-inactivated genes with the same Y gametolog status in both sexes: Points show τ statistic values for all genes in each XCI/Y category for female (left) and male (right) samples. White diamonds show the mean of τ across all genes in that XCI/Y category. Areas of darker color indicate overlap of two or more points. Values of τ range from 0 to 1, with higher values corresponding to greater tissue specificity.

Table 2. Resampling Mean and Standard Deviation of the Tissue Specificity Index (τ) for XCI/Y Status Categories in Females and Males

Y category	XCI category	*n* genes	Female τ	Male τ
Functional Y	XCI escapers	11	0.51 ± 0.06	0.52 ± 0.06
	X-inactivated	4	0.73 ± 0.11	0.60 ± 0.11
No Y (Pseudo. + Lost Y)	XCI escapers	225	0.68 ± 0.01	0.67 ± 0.01
	X-inactivated	226	0.75 ± 0.01	0.75 ± 0.01

be a consequence of the high prevalence of tissue-specific variable escape observed by Cotton et al. (2015), in which case it is likely that expanding the number of tissues and individuals in studies of XCI status will reveal more variable XCI escapers. The evolutionary basis for variable escape is unclear, but it should be a priority to determine whether variable escape reflects true polymorphism for underlying mechanistic differences among individuals in escape status, or whether it is technical noise that arises from catching the tail of a distribution of allele-specific expression or methylation after studying sufficiently many samples across sufficiently many tissues.

Finally, in light of their marked selection and expression patterns as indicative of their biological importance, we concur with many previous studies that human XCI escapers are promising candidates for investigation as contributing to human genetic disorders that exhibit significant sex differences in incidence, severity, or symptoms, as well as those that show high phenotypic variability among females. In particular, large-scale disease studies are likely to benefit from incorporating information about a gene's XCI status in association analyses, as in Tukiainen et al. (2014), Chang et al. (2014), and Ma et al. (2015).

Methods

XCI Status Data

We obtained XCI statuses for 758 unique protein-coding human ChrX loci from 3 data sets: Carrel and Willard (2005), Cotton et al. (2015), and Schultz et al. (2015) (supplementary table S2, Supplementary Material online). These XCI statuses were determined based on allele-specific expression from the inactivated X (Carrel and Willard 2005), female versus male transcription start site CG dinucleotide (mCG) level (Cotton et al. 2015), and female versus male gene body non-CG methylation (mCH) level (Schultz et al. 2015). Brief descriptions of the assays and categorizations used in each study are shown in supplementary figure S2, Supplementary Material online.

To ascertain the effects of the different methods of determining XCI status and the impact of variable and tissue-specific XCI escapers on the overall patterns of purifying selection across XCI categories, we repeated the analysis of strong purifying selection (described below) in each individual data set. For the two data sets with information about heterogeneous/variable escape, we performed these calculations by either excluding variable escapers or splitting them into the consistent escape and consistently X-inactivated categories. For the Schultz et al. data set, we calculated these statistics using the set of genes that escaped in any proportion of 11 tissues as XCI escapers, and using only genes that escaped in 6 or more tissues as escapers (treating those that escaped in 5 or fewer tissues as X-inactivated). We observed that excluding variable or tissue-specific escapers did not result in a significant change in the results within the individual data sets, nor did it have a significant effect on the "combined" data set defined by aggregating XCI statuses across all three studies using an inclusive definition of XCI escape: All genes showing escape in any number of tissues and/or individuals in any study were considered XCI escapers, and all others were considered X-inactivated (supplementary table S2, Supplementary Material online). This definition of XCI is expressly somewhat more inclusive, and meant as an appropriate contrast, being less vulnerable to false negatives in the identification of XCI escape associated to the nuances of the particular methods used in each of the individual studies.

Polymorphism Data

For our polymorphism analyses, we took advantage of a large, high-depth whole-exome sequence data set from the NHLBI GO Exome Sequencing Project (ESP), which includes a large amount of rare (minor allele frequency <0.5%) single nucleotide variants (SNVs) (Fu et al. 2013). We included only biallelic variants in the European American subsample with an average sequencing depth >20X, and that passed the original (Tennessen et al. 2012) quality control filters. The number of copies *n* available in the ESP data set varies across SNVs, with a range of 2,379–6,728, a mean of 6,558, and a standard error of the mean of 2.51 across non-PAR ChrX sites. To make all sites comparable, we downsampled derived allele counts at each site to a total of 6,056 (the 90th percentile of n), according to the method described in Marth et al. (2004), and excluded variants for which data were available for fewer than 6,056 copies.

Divergence Data

To facilitate expression analyses (described below), we considered only the transcript with the greatest total exonic length in the RefSeq database for each unique protein-coding gene, yielding 748 nonoverlapping ChrX transcripts. Of the 1,048,661 unique ESP SNVs across the entire human exome, 24,795 were associated with these 748 transcripts, with the majority falling in nonexonic regions. Polymorphic, divergent, and conserved monomorphic sites were identified as either synonymous (Syn) or nonsynonymous (NonSyn) based on the genetic code. Counts of divergent and polymorphic sites were calculated for each transcript. Divergent sites were those for which the human reference (hg19) allele differed from the ancestral allele, inferred by the reference chimpanzee (panTro4) allele and confirmed by one or both of the reference orangutan (ponAbe2) and macaque (rheMac3) alleles. Sites for which the ancestral state could not be confirmed in this manner were excluded.

McDonald–Kreitman Test Data Filters

Transcripts were filtered to remove overlap with the UC Santa Cruz Genome Browser (Kent et al. 2002) segmental duplications and simple sequence repeat tracks. Additionally, we used the primate syntenic net track to exclude sites with uncertain ancestral states due to poor synteny. Finally, we excluded potential CpG dinucleotides in either the ancestral or derived state. After filtering, we were left with 525 ChrX transcripts that retained at least one exon with usable data, existed in a single copy, and were inferred to be present on the ancestral mammalian ChrX. These 525 loci were the final set of genes we considered in our analysis. Across these genes, there were 3,581 NonSyn polymorphic sites, 2,279 Syn polymorphic sites, 392 NonSyn divergent sites, 558 Syn divergent sites, 1,431,973 conserved NonSyn sites, and 507,404 conserved Syn sites. The numbers of genes, following filtering, in each XCI and Y category, and XCI/Y combinations, are shown in supplementary figure S1, Supplementary Material online. All site counts for each genes are available in supplementary table S5, Supplementary Material online.

Quantifying Ancient/Strong Purifying Selection via Polymorphism and Divergence

The MK test (McDonald and Kreitman 1991) uses the ratios of selected and neutral sites among intraspecies polymorphisms compared with interspecies divergences to detect departures from neutral evolution. We used NonSyn and Syn sites of protein-coding genes as our selected and neutral site classes, respectively, and calculated three statistics based on the MK test across XCI/Y categories. In these statistics, m_S and m_N are the Syn and NonSyn total site counts, respectively. $P_{N\,neut}$ and $P_{N\,weak}$ are the neutral and weakly deleterious fraction of polymorphic sites out of P_N NonSyn polymorphisms overall. These are estimated by partitioning the NonSyn and Syn polymorphic site counts P_N and P_S into arbitrarily defined bins of high and low derived allele frequency, and using the neutral low:high count ratio to calculate the expected fractions of the rarest selected polymorphisms that are weakly deleterious and neutral. Throughout this study, we report values calculated using a derived allele frequency cutoff of $> 1\%$ to classify SNVs as high frequency, which corresponded to a downsampled derived allele count of 61. For a more detailed description of the calculation of these statistics, refer to the supplementary material of Mackay et al. (2012). The statistics we calculated for each XCI/Y group were as follows: $N = (m_S P_{N\,neut})/(m_N P_S)$, the fraction of sites that are neutral (called f in Mackay et al. 2012); $W = (m_S P_{N\,weak})/(m_N P_S)$, the fraction of sites that are evolving under weak purifying selection (called b in Mackay et al. 2012); and $S = 1 - (N + W)$, the fraction of sites evolving under strong purifying selection (called d in Mackay et al. 2012). The fraction of sites that are weakly deleterious, W, was the only statistic for which the relationships between XCI/Y categories changed across cutoff DAFs (supplementary fig. S3, Supplementary Material online). However, because W was an order of magnitude smaller than either N or S in all categories, we concluded that these fluctuations are most likely due to chance, and we did not further analyze this statistic.

Y Gametolog Status Data

Y gametolog statuses for 723 unique protein-coding ChrX genes were obtained from Wilson-Sayres and Makova (2013), who identified Y gametologs based on interspecies comparisons and human X/Y homology. The Y gametolog categories described in this study were defined as follows: 594 genes that were inferred to be ancestral in the eutherian lineage (conserved across 4 species among the set including mouse, rat, rabbit, cow, horse, dog, opossum, and chicken) were classified as functional Y ($n = 19$), pseudogenized Y ($n = 265$), and lost Y ($n = 312$). Finally, we removed three genes located in the human-specific X-transposed region from consideration, of which two had functional Y gametologs and one had a pseudogenized Y gametolog. For some analyses, based on our initial results, we combined the two categories of pseudogenized Y and lost Y into a single category ("no Y") corresponding to no functional Y gametolog (see Results).

ChrX Gene Copy Number Data

Human ChrX gene copy classifications were obtained from Mueller et al. (2013). In this data set, genes were annotated as single copy, X multicopy, or ampliconic, where X multicopy genes had at least one *cis* paralog but were not in ampliconic regions, and ampliconic genes were those in or near ampliconic regions. Because ampliconic and multicopy genes are poorly conserved across the great apes, we only considered genes with a single copy on the human ChrX.

Gene Expression Level

To assess differences in human gene expression across XCI/Y categories, we used the publicly available NIH GTEx RNA-seq data set (GTEx Consortium 2015), which includes expression data in 51 primary tissue subtypes across 30 broad tissue types from American women (78) and men (138).

Many genes in certain XCI/Y categories are very weakly expressed, and retain no usable data if we use typical arbitrary minimum expression values (e.g. 1 RPKM) to remove potentially spurious measurements from the RNA-seq data. However, these weakly expressed genes are not typically statistical outliers in the expression distributions of their gene categories. We therefore elected to use an adaptive method for filtering background expression values in an unbiased manner while including low expression values that were not outliers. For each broad tissue type, we fit a half Gaussian distribution to the higher mode of the distribution of nonzero $\log_2$(RPKM) values across all individuals for all ChrX genes, and mirrored it to capture the expression distribution of actively transcribed genes (Hart et al. 2013). We then used three standard deviations below the means of these distributions as a minimum expression value to consider a gene significantly expressed in each tissue, while excluding the other genes from analysis. The sex- and tissue-specific expression cutoffs defined in this way are available in supplementary table S6, Supplementary Material online. After individual expression values were filtered according to these minima, the remaining values were averaged across 1,000 bootstrap samples of individuals for each gene in each tissue. For the sake of clarity, in this section we refer to each gene's average expression across individuals in each tissue as its expression value for that tissue.

For GTEx tissues with multiple subtypes, we considered only the subtype with the largest sample size, yielding one expression value for each broad tissue type for each gene. To ensure that all genes had data for equal numbers of tissues in females and males, we excluded sex-specific tissues (cervix, fallopian tube, ovary, uterus, vagina, prostate, and testis) from our expression analyses, leaving 23 broad tissue types in both sexes. For each tissue, we calculated the mean and standard deviation of expression values across genes in each XCI/Y category. These values are displayed in figure 2*a*. The mean global expression value of each gene in each XCI/Y category is displayed in figure 2*b*. All data displayed in figure 2 are available in supplementary table S3 and S4, Supplementary Material online. Figure 2*b* and 3 were generated using the beeswarm R package, which is available at https://cran.r-project.org/web/packages/beeswarm/index.html (last accessed November 2, 2015).

To assess the significance of global expression ratios between XCI escapers and X-inactivated genes in each Y gametolog category, we performed 1,000 permutations of gene membership in pairs of XCI/Y groups and calculated the difference in the mean expression of these genes across all tissues. *P* values reflect the proportion of permutations for which the ratio of average expression values between categories, averaged across all tissues, was greater than or equal to the observed. Repeating this analysis using various more and less lenient fixed and adaptive expression minima showed that the adaptive three standard deviation cutoff produced results that were closest to those generated using a fixed cutoff of 0.1 RPKM (supplementary fig. S4, Supplementary Material online), which is the same cutoff used by the GTEx Consortium in their own expression analyses (GTEx Consortium 2015) and similar to the cutoff value calculated by the analysis of Hart et al. (2013).

To calculate two-sided Mann–Whitney *U* test statistics and *P* values for the unfiltered expression distributions in each Y gametolog category in each sex, we first calculated the average RPKM values for each gene in each tissue across individuals. These values were pooled across genes and tissues in each XCI/Y category, and the test then applied for each pair of these pooled data sets using the wilcox.test function in R (for unpaired data).

Gene Expression Breadth

We calculated the tissue-specificity index τ (Yanai et al. 2005) for each gene to quantify the breadth of expression across each category. This statistic is defined for each gene as follows:

$$\tau = \frac{\sum_{i=1}^{N}(1 - x_i)}{N - 1}$$

where N is the number of tissues and x_i is the expression value of this gene in tissue i, normalized by its maximum expression value across all N tissues. High values of this statistic (e.g., >0.6) indicate that the gene's expression is skewed toward a small number of tissues, and low values indicate that it is expressed at similar levels across many tissues. Using expression estimates obtained as described in the previous section, the genes in each XCI/Y category were resampled with replacement 1,000 times, and the mean and standard deviation of the τ values across these resamples were obtained (table 2).

Finally, for each gene, we calculated the proportion of 23 broad tissue types for which it showed significant expression (greater than the minimum expression value for each tissue, defined as in the previous section) as an additional metric of expression breadth (supplementary fig. S5, Supplementary Material online). The gene membership of each XCI/Y category was resampled with replacement 1,000 times and used to compute the mean and standard deviation.

Supplementary Material

Supplementary figures S1–S11 and tables S1–S5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Xu Wang, Feng Gao, Aviv Madar, Aaron J. Sams, Philipp Messer, and Timothy Connallon for valuable discussions. This project was supported in part by R01 HG006849 to A.K., R01 HG003229 to A.G.C., an award from The Ellison Medical Foundation to A.K., and an award from The Edward Mallinckrodt, Jr. Foundation to A.K.

References

Bellott DW, Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Cho TJ, Koutseva N, Zaghlul S, Graves T, Rock S, et al. 2014. Mammalian Y chromosomes retain widely expressed dosage sensitive regulators. *Nature* 508:494–499.

Berletch JB, Yang F, Xu J, Carrel L, Disteche CM. 2011. Genes that escape from X inactivation. *Hum Genet.* 130:237–245.

Bione S, Sala C, Manzini C, Arrigo G, Zuffardi O, Banfi S, Borsani G, Jonveaux P, Philippe C, Zuccotti M, et al. 1998. A human homologue of the *Drosophila melanogaster diaphanous* gene is disrupted in a patient with premature ovarian failure: evidence for conserved function in oogenesis and implications for human sterility. *Am J Hum Genet.* 62:533–541.

Brown CJ, Greally JM. 2003. A stain upon the silence: genes escaping X inactivation. *Trends Genet.* 19:432–438.

Carrel L, Willard HF. 2005. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* 434:400–404.

Chang D, Gao F, Slavney A, Ma L, Waldman YY, Sams AJ, Billing-Ross P, Madar A, Spritz R, Keinan A. 2014. Accounting for eXentricities: analysis of the X chromosome in GWAS reveals X-linked genes implicated in autoimmune diseases. *PLoS One* 9(12):e113684.

Charlesworth B. 1991. The evolution of sex chromosomes. *Science* 251:1030–1033.

Cotton AM, Price EM, Jones MJ, Balaton BP, Kobor MS, Brown CJ. 2015. Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X chromosome inactivation. *Hum Mol Genet.* 24(6):1528–1539.

Deng X, Berletch JB, Nguyen DK, Disteche CM. 2014. X chromosome regulation: diverse patterns in development, tissues and disease. *Nat Rev Genet.* 15:367–378.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A.* 102:14338–14343.

Duret L, Mouchiroud D. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol Biol Evol.* 17:68–74.

Ellison JW, Wardak Z, Young MF, Gehron Robey P, Laig-Webster M, Chiong W. 1997. PHOG, a candidate gene for involvement in the short stature of Turner syndrome. *Hum Mol Genet.* 6:1341–1347.

Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Lael SM, Gabriel S, Rieder MJ, Altshuler D, Shendure J, et al. 2013. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* 493:216–220.

Gao F, Keinan A. 2014. High burden of private mutations due to explosive human population growth and purifying selection. *BMC Genomics* 15(Supp. 4):S3.

GTEx Consortium. 2015. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348:648–660.

Hart T, Komori HK, LaMere S, Podshivalova K, Salomon DR. 2013. Finding the active genes in deep RNA-seq gene expression studies. *BMC Genomics* 14:778–785.

Jegalian K, Page DC. 1998. A proposed path by which genes common to mammalian X and Y chromosomes evolve to become X inactivated. *Nature* 394:776–780.

Keinan A, Clark AG. 2012. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 336:740–743.

Kent WJ, et al. 2002. The human genome browser at UCSC. *Genome Res.* 12:996–1006.

Lahn B, Page D. 1999. Four evolutionary strata on the human X chromosome. *Science* 248:964–967.

Lyon M. 1962. Sex chromatin and gene action in the mammalian X-chromosome. *Am J Hum Genet.* 14:135–148.

Ma L, Hoffman G, Keinan A. 2015. X-inactivation informs variance-based testing for X-linked association of a quantitative trait. *BMC Genomics* 16:241.

Mackay T, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, et al. 2012. The *Drosophila melanogaster* genetic reference panel. *Nature* 482:173–178.

Marth GT, Czabarka E, Murvai J, Sherry ST. 2004. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* 166:351–372.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351(6328):652–654.

Midgeon BR. 2014. Females are mosaics: X inactivation and sex differences in disease. 2nd ed. Oxford: Oxford University Press.

Mueller JL, Skaletsky H, Brown LG, Zaghlul S, Rock S, Graves T, Auger K, Warren WC, Wilson RK, Page DC. 2013. Independent specialization of the human and mouse X chromosomes for the male germ line. *Nat Genet.* 45:1083–1087.

Ohno S. 1967. Sex chromosomes and sex-linked genes. Berlin: Springer.

Park C, Carrel L, Makova KD. 2010. Strong purifying selection at genes escaping X chromosome inactivation. *Mol Biol Evol.* 27(11):2446–2450.

Park SG, Choi SS. 2010. Expression breadth and expression abundance behave differently in correlations with evolutionary rates. *BMC Evol Biol.* 10: 241.

Pessia E, Makino T, Bailly-Bechet M, McLysaght A, Marais GAB. 2012. Mammalian X chromosome inactivation evolved as a dosage-compensation mechanism for dosage-sensitive genes on the X chromosome. *Proc Natl Acad Sci U S A.* 109:5346–5351.

Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, Rajagopal N, Nery JR, Urich MA, Chen H, et al. 2015. Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* 523:212–216.

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

Talebizadeh Z, Simon SD, Butler MG. 2006. X chromosome expression in human tissues: male and female comparisons. *Genomics* 88:675–681.

Tennessen J, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, et al. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 237:64–69.

Trabzuni D, Ramasamy A, Imran S, Walker R, Smith C, Weale ME, Hardy J, Ryten M, North American Brain Expression Consortium. 2013. Widespread sex differences in gene expression and splicing in the adult human brain. *Nat Commun.* 4:2771.

Tukiainen T, Pirinen M, Sarin AP, Ladenvall C, Kettunen J, Lehtimäki T, Lokki ML, Perola M, Sinisalo J, Vlachopoulou E, et al. 2014. Chromosome X-wide association study identifies loci for fasting insulin and height and evidence for incomplete dosage compensation. *PLoS Genet.* 10(2):e1004127.

Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, Eisen MB, Feldman MW. 2005. Functional genomic analysis of the rates of protein evolution. *Proc Natl Acad Sci U S A.* 102(15):5483–5488.

Wang X, Douglas KC, Vanderberg JL, Clark AG, Samollow PB. 2014. Chromosome-wide profiling of X-chromosome inactivation and epigenetic states in fetal brain and placenta of the opossum, *Monodelphis domestica. Genome Res.* 24(1):70–83.

Wang X, Miller DC, Clark AJ, Antczak DF. 2012. Random X inactivation in the mule and horse placenta. *Genome Res.* 22(10):1855–1863.

Wilson-Sayres MA, Makova KD. 2013. Gene survival and death on the human Y chromosome. *Mol Biol Evol.* 30(4):781–787.

Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E, et al. 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* 21(5):650–659.

Yang F, Babak T, Shendure J, Disteche CM. 2010. Global survey of escape from X inactivation by RNA-sequencing in mouse. *Genome Res.* 20:614–622.

Zinn AR, Ross JL. 1998. Turner syndrome and haploinsufficiency. *Curr Opin Genet Dev.* 8:322–327..