Supplementary Materials

Supplementary Notes:

Supplementary Tables:

Supplementary Figure:

Supplementary Note 1. Methods

Identification of ancestral X-linked genes

To find the set of X-linked genes that likely existed on the ancestral human X chromosome (the X-added region, XAR, and X-conserved region, XCR [1]), we analyzed all human X-linked Refseq CCDS genes, extracted from the UCSC Genome Browser [2]. For comparative analysis, we analyzed all non-primate species out to chicken with assembled genomes in the 46-way alignment (mouse, rat, rabbit, cow, horse, dog, opossum, and chicken). As a quality filter, we removed any segments that mapped non-uniquely in each species individually. After removing pseudoautosomal region genes, we conducted a comparative analysis. If an ortholog was identified on the X, or homologous X-linked region in at least four out of eight non-primate assembled genomes (on chrX of mouse, rat, rabbit, cow, horse, or dog, on chrX, chr4 or chr7 of opossum, or on chr4 or chr1 of chicken), then it was identified as likely existing on the ancestral human X chromosome (Figure S1). We choose to analyze non-primate genomes to avoid including recent duplications and retrotranspositions.

Y gametolog identification

To find the set of X-linked genes with nonfunctional Y-linked counterparts (i.e., pseudogenes), we used all human X-linked Refseq CCDS genes, extracted from the UCSC Genome Browser [2], as queries in a lastZ search [3] of the human Y chromosome, keeping those with hsp score greater than 1500. In the case of genes with more than two copies on the X, we only conclude evidence for a Y-linked pseudogene if each X-linked gene has a unique best hit on the Y chromosome. We cannot exclude the possibility that some Y pseudogenes are the result of segmental duplications or retrotranspositions from the X chromosome. We argue, however, that by requiring presence in the homolgous X-linked region in at least four out of eight non-primate species with assembled genomes, that these events occurred anciently (at least before the common ancestor of eutherians, and likely even earlier), and so, like the entire X-added region, which was added after marsupials diverged from eutherian mammals, we should consider these events in our analysis of the evolution of the X and Y chromosomes.

Alignment and divergence

For X-linked genes with gametologous Y-linked sequence (functional or pseudogenized) we aligned X- and Y-linked sequences using lastZ [4], curating them to delete frameshift mutations from the gene-pseudogene alignments and mask internal stop codons. We computed the number of synonymous substitutions per synonymous site (d_S) and the number of nonsynonymous substitutions per nonsynonymous site (d_N) in paml [5], and calculated the d_N/d_S ratio, which is used as a proxy for the strength of selective constraint for pairwise X-Y comparisons and for X-Y-outgroup comparisons.

Functional assay

The status of each gene's link to human disease was defined in the OMIM disease database and extracted using DAVID [6].

RNAseq expression analysis

We analyzed reported normalized RNAseq expression across both male and female human samples across six tissues (brain, cerebellum, heart, kidney, liver, testes), with expression intensity is measured in reads per kilobase of exon model per million mapped reads, RPKM, [7]. The median RPKM expression was computed for each class of X-linked genes (with functional, pseudogenized or lost Y homologs) for each tissue, for each individual. We chose not to pool samples so that we could observe the variations across individuals.

Microarray expression analysis

We considered only those genes that had one-probe-to-one-gene mapping [8] in a dataset that assayed expression in 79 human tissues [9]. Expression intensity is measured as the average difference in hybridization intensity, AD, with expression breadth defined as the sum of all tissues where expression was measured at greater than 200 AD. For both breadth and intensity we studied four partitions of tissues - 1) healthy, non-cancerous tissues; 2) two female-specific tissues (ovary and uterus); and, 3) five male-specific tissues (whole testis cells, and testis leydig, germ, interstitial and seminiferous tubule cells).

X-inactivation

Available data [10] present the XCI status (escape or inactivated) for human X-linked genes in nine individuals. For each gene we computed the proportion of individuals for which the gene escaped inactivation (e.g., if the gene escaped inactivation in nine out of nine analyzed individuals it was given a value of 9/9 = 1).

References

[1] Ross, M. T. *et al.* The DNA sequence of the human X chromosome. *Nature* **434**, 325-337 (2005).

[2] Fujita, P. A. *et al.* The UCSC Genome Browser database: update 2011. *Nucleic Acids Res* **39**, D876-882, doi:10.1093/nar/gkq963 (2011).

[3] Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. **13**, 103-107, doi:10.1101/gr.809403 (2003).

[4] Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res* **13**, 103-107 (2003).

[5] Yang, Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-1591 (2007).

[6] Dennis, G., Jr. *et al.* DAVID: Database for Annotation, Visualization, and Integrated Discovery. **4**, P3 (2003).

[7] Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343-348, doi: 10.1038/nature10532 (2011).

[8] Park, C. & Makova, K. D. Coding region structural heterogeneity and turnover of transcription start sites contribute to divergence in expression between duplicate genes. **10**, R10, doi:gb-2009-10-1-r10 (2009).

[9] Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101**, 6062-6067 (2004).

[10] Carrel, L. & Willard, H. F. X-inactivation profile reveals extensive variability in X-linked gene expression in females. **434**, 400-404, doi:10.1038/nature03479 (2005).

Supplementary Table 2. Expected and observed gene loss on human Y. We classified ancestral X-genes into those with functional (gene) and nonfunctional (pseudogene and lost) Y-linked gametologs. Our observed loss of functional Y-linked genes is consistent with estimates from a model of exponential decay of the human Y chromosome (Hughes et al. 2012).

Strata	Human Y gametologs*			Observed functional Y gene loss	Expected Y gene loss [11]
	gene	pseudogene	lost		
1	5	158	206	364	414
2	2	36	46	82	88
3	8	57	62	119	143
4	2	7	1	8	12
5	0	7	0	7	7
Total	*17*	*265*	*315*	*580*	*664*

* Sequences in the X-transposed region are not included in the total counts here.

Hughes, J. F. *et al.* Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature* **483**, 82-86, doi:nature10843 (2012).

Supplementary Table 3. Linear regression results. For each partition of the human X chromosome (whole X, X-added region, XAR, or X-conserved region, XCR), linear regressions are computed between the X-Y pairwise synonymous substitution rate (dS) versus position on the human X for X-Y pairs with functional Y homologs (Gene) and for X-Y pairs with pseudogenized Y homologs (Pseudo). Given the relatively small number of X-linked genes with functional Y homologs, nearly all power is lost when this class (Gene) is partitioned into XAR and XCR.

X partition	Y Class	Intercept	Slope	R^2	*P*
wholeX	Gene	-0.074414	0.016813	0.6724	***0.0001***
wholeX	Pseudo	0.573076	0.002904	0.02762	***0.0121***
XAR	Gene	0.138592	0.006384	0.3679	*0.0630*
XAR	Pseudo	0.387887	0.012518	0.06316	***0.0469***
XCR	Gene	-0.04648	0.01693	0.4009	*0.1770*
XCR	Pseudo	0.560098	0.002952	0.01487	*0.1198*

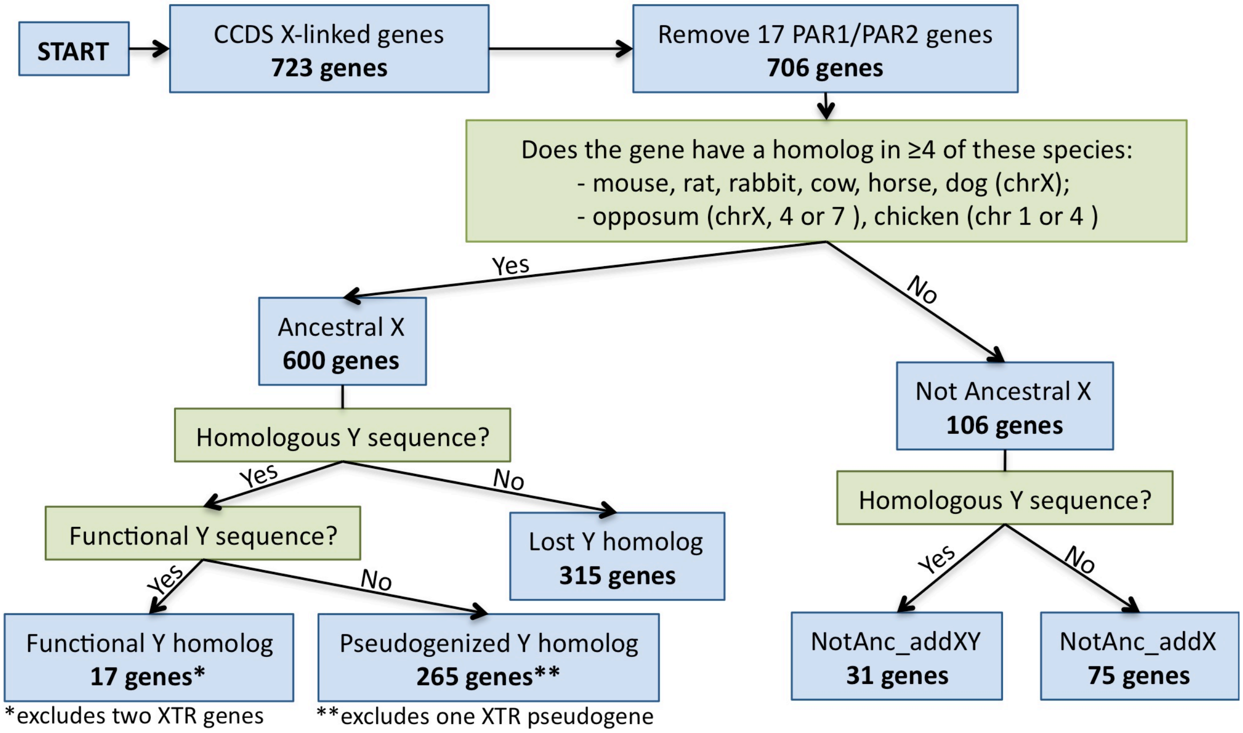
Supplementary Table 4. Human branch from three-way Human-Opossum-Platypus (H-O-P), Human-Dog-Opossum, (H-D-O) or Human-Chimp-Dog (H-C-D) comparisons. This set of analyses does not utilize Y-linked sequence, and so is able to analyze X-linked genes in all three classes; 1) X-linked genes with functional Y-linked genes (Gene); 2) X-linked genes with pseudogenized Y-linked genes (Pseudo); and, 3) X-linked genes whose Y homologs have been lost from the Y chromosome (Lost). *P* values from permutation tests with 10,000 replicates are shown for the tests for a significant difference between median dN, dS, and dN/dS values between each pairwise comparison of the classes described above. Partitions of the X (wholeX, X-added region, XAR and X-conserved region, XCR) are grouped together. *P* values that a re significant before Bonferroni correction are highlighted in bold.

X region	Y-linked	H-O-P		H-D-O		H-C-D	
	Class	d_S	d_Nd_S	d_S	d_Nd_S	d_S	d_Nd_S
wholeX	Gene	0.056	0.050	0.013	0.013	$5.14x10^{-4}$	$2.00x10^{-6}$
wholeX	Pseudo	0.477	0.487	0.139	0.129	$3.60x10^{-3}$	$2.93x10^{-3}$
wholeX	Lost	0.099	0.099	0.105	0.099	0.148	$1.95x10^{-4}$
wholeX	$P_{gene-pseudo}$	***0.0339***	*0.3405*	*0.1286*	*0.8496*	*0.4198*	***0.0077***
wholeX	$P_{gene-lost}$	*0.5544*	*0.8404*	*0.6262*	*0.2004*	*0.4390*	*0.4614*
wholeX	$P_{pseudo-lost}$	*0.1588*	*0.9660*	*0.3515*	*0.6029*	*0.9942*	***0.0036***
XAR	Gene	0.059	0.042	0.019	0.013	$8.99x10^{-4}$	$2.00x10^{-6}$
XAR	Pseudo	0.480	0.468	0.157	0.141	$4.37x10^{-3}$	$4.48x10^{-3}$
XAR	Lost	0.109	0.105	0.124	0.077	0.144	0.027
XAR	$P_{gene-pseudo}$	*0.1471*	*0.1693*	*0.3748*	*0.0769*	*0.7111*	***0.0335***
XAR	$P_{gene-lost}$	*0.8253*	*0.8138*	*0.8649*	*0.4290*	*0.4115*	*0.9346*
XAR	$P_{pseudo-lost}$	*0.0394*	*0.8186*	*0.8644*	*0.0384*	*0.8263*	*0.1440*
XCR	Gene	0.051	0.051	0.011	0.014	$4.57x10^{-4}$	$2.00x10^{-6}$
XCR	Pseudo	0.477	0.498	0.131	0.128	$3.18x10^{-3}$	$2.65x10^{-3}$
XCR	Lost	0.095	0.093	0.089	0.105	0.171	$1.77x10^{-4}$
XCR	$P_{gene-pseudo}$	*0.7214*	*0.9900*	*0.2819*	*0.3034*	*1*	***0.0263***
XCR	$P_{gene-lost}$	*0.5418*	*0.7438*	*0.5690*	*0.5465*	*0.3848*	*0.5532*
XCR	$P_{pseudo-lost}$	*0.6713*	*0.9255*	*0.3381*	*0.3766*	*0.9935*	***0.0109***

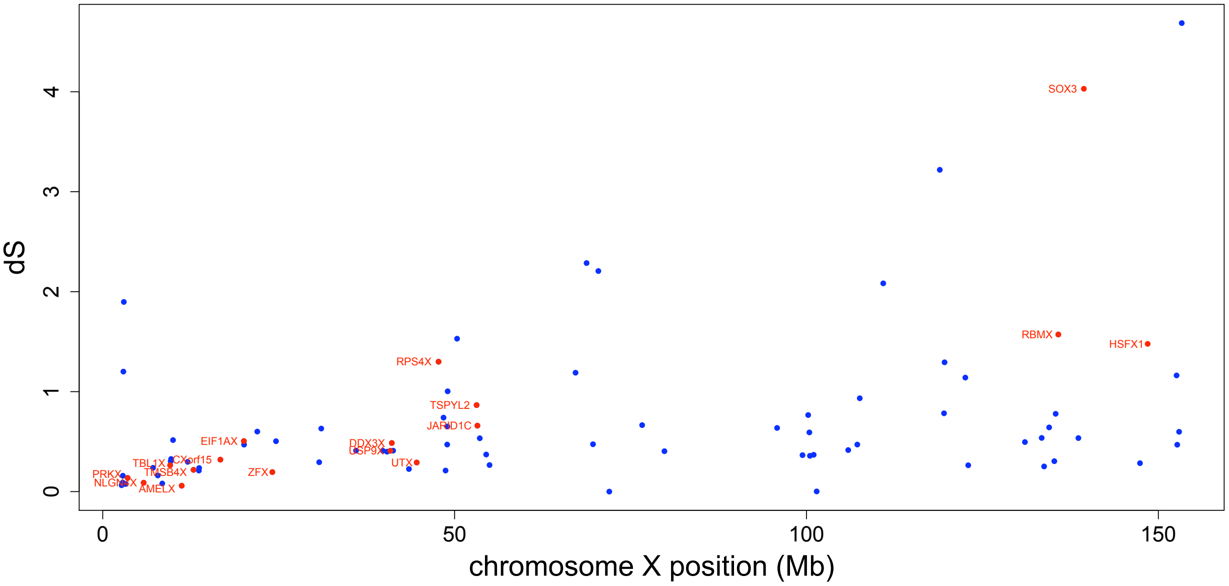
Supplementary Table 5. **X-inactivation comparisons.** Mean and median values of the proportion of individuals (out of nine assayed in (Carrel and Willard 2005)) are reported for which an X-linked gene escapes inactivation, observed for the sets X-linked genes with functional (gene) Y-linked gametologs, pseudogenized (pseudogene) Y-linked gametologs, or absent (lost) Y-linked gametologs, as well as for X-linked genes that appear to have been added to the X chromosome only (and so are "notAncestral" on the ancestral sex chromosomes), or are in the pseudoautosomal region (PAR). *P* values from permutation tests with 10,000 replicates, testing for significant differences between each pairwise class of X-linked genes (gene, pseudogene and lost), are labeled at the bottom of the table for the entire X (whole X), the X-added region (XAR) and the X-conserved region (XCR).

Observed	Mean			Median		
	Whole X	XAR	XCR	Whole X	XAR	XCR
gene	0.80	0.86	0.67	1.00	1.00	1.00
pseudogene	0.24	0.47	0.14	0.11	0.33	0.00
lost	0.16	0.24	0.14	0.00	0.00	0.00
PAR	0.90	-	-	1.00	-	-
notAncestral	0.22	-	-	0.11	-	-
$P_{Gene\text{-}Pseudo}$	***0***	***0.0276***	***0.0003***	***0***	*0.1392*	***0.0006***
$P_{Gene\text{-}Lost}$	***0***	***0.0001***	***0.0003***	***0***	***0.0042***	***0.0006***
$P_{Pseudo\text{-}Lost}$	*0.0621*	***0.0358***	*0.9001*	*0.3213*	*0.1560*	*1.0000*

Supplementary Figure 1. Gene classification workflow. Using comparative genomics and X-Y sequence comparisons, we assessed the status of the 723 consensus CDS (CCDS) genes listed for the human X chromosome, a set of consistently annotated and high quality genes (Figure S1). We first excluded the 17 pseudoautosomal (PAR) region genes, which still undergo X-Y recombination. Of the 706 non-PAR genes, 600 genes were classified as "ancestral X" due to the existence of sequence in homologous XAR or XCR regions in at least four out of eight assembled non-primate genomes (mouse, rat, rabbit, cow, horse, or dog, opossum or chicken; Table S1, Figure S1), and 106 are classified as "notAncestral". Of 600 ancestral X-linked genes, 19 have functional Y homologs (two of these are recently X-transposed, XTR, and so excluded from further analysis, so we are left with 17), 266 have evidence of a pseudogenized Y homolog (one is in the XTR, and so is excluded, so we are left with 265), and 315 have no evidence of a functional or pseudogenized Y homolog, so are classified as "lost" on the Y chromosome (Table S1, Figure S1). Of the 106 genes classified as "notAncestral", many are members of multi-gene families, or genes with a single exon, suggesting that they have been independently added (duplicated or retrotransposed) onto the X, or onto the X and Y chromosomes; 31 have evidence of homologous Y sequence, and 75 have no evidence of a Y homolog (Table S1). Only one degraded Y exon was recovered for 29 of the 31 notAncestral X-linked genes with evidence of homologous Y sequence, and only two exons were recovered for the remaining two genes. Because these 31 genes do not appear to be conserved from the ancestral X chromosome, are nearly all part of multi-gene families, and only one exon was recovered on the Y, the similarity search likely found degraded duplicated or retrotransposed copies added to the Y that should not be included in further analyses. Further, the 75 genes that did not pass our comparative genomics analysis, nearly all are part of multi-gene (often tandem) families, or single exon genes suggesting they were duplicated or retrotransposed after the eutherian common ancestor. We conservatively exclude all "notAncestral" genes from the downstream analysis.



Supplementary Figure 2. Pairwise X-Y synonymous substitution rate versus chromosome X position. To assess whether retrotransposition from the X onto the Y might be a large problem, we replotted pairwise dS versus position on the X chromosome excluding all X-linked genes with a single exon (which might be retrotransposed onto the X and to the Y), as well as excluding all Y-pseudogenes with evidence from only a single exon (multiple Y exons are not contiguous). It is possible that a retrotransposed pseudogene was broken apart by inversions on the Y, making it appear to be a pseudogenized gene, but we do not have power to detect such events. However, even without all of the potential retrotransposition events we do not observe high pairwise dS values for X-gene-Ypseudogene pairs. This suggests that the trend we observe, including all identified Y pseudogenes is likely not biased by a large fraction of retrotranspositions. We argue, then, that by requiring homology with dog, and also for most genes, opossum, that these events occurred anciently (at least before the common ancestor of human and dog), and so, like the entire X-added region, which was added after marsupials diverged from eutherian mammals, we should consider these events in our analysis of the evolution of the X and Y chromosomes.


dS
SOX3
RBMX
HSFX1
RPS4X
TSPYL2
JARID1C
EIF1AX
DDX3X
USP9X
UTX
ZFX
TBL1X
CXorf15
TMSB4X
PRKX
NLGN4X
AMELX
chromosome X position (Mb)