HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2014 March 01.

Published in final edited form as:

Nat Genet. 2013 September ; 45(9): 1083–1087. doi:10.1038/ng.2705.

Independent specialization of the human and mouse X chromosomes for the male germline

Jacob L. Mueller[1], **Helen Skaletsky**[1,2], **Laura G. Brown**[1,2], **Sara Zaghlul**[1], **Susan Rock**[4], **Tina Graves**[4], **Katherine Auger**[5], **Wesley C. Warren**[4], **Richard K. Wilson**[4], and **David C. Page**[1,2,3]

[1]Whitehead Institute, Cambridge, Massachusetts, USA

[2]Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

[3]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

[4]The Genome Institute, Washington University School of Medicine, St. Louis, Missouri, USA

[5]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom

Abstract

We compared the human and mouse X chromosomes to systematically test Ohno's law, which states that the gene content of X chromosomes is conserved across placental mammals[1]. First, we improved the accuracy of the human X-chromosome reference sequence through single-haplotype sequencing of ampliconic regions. This closed gaps in the reference sequence, corrected previously misassembled regions, and identified new palindromic amplicons. Our subsequent analysis led us to conclude that the evolution of human and mouse X chromosomes was bimodal. In accord with Ohno's law, 94–95% of X-linked single-copy genes are shared between human and mouse; most are expressed in both sexes. Strikingly, most X-ampliconic genes are exceptions to Ohno's law: only 31% of human and 22% of mouse X-ampliconic genes share orthologs. X-ampliconic genes are expressed predominantly in testicular germ cells, and many were independently acquired since the common ancestor of humans and mice, specializing portions of their X chromosomes for sperm production.

AUTHOR CONTRIBUTIONS

J.L.M., H.S., W.C.W., R.K.W. and D.C.P planned the project. J.L.M. and L.G.B. performed BAC mapping. J.L.M. performed RNA deep sequencing. T.G., S.R., K.A., and S.Z. were responsible for finished BAC sequencing. J.L.M. and H.S performed sequence analyses. J.L.M. and D.C.P. wrote the paper.

DATA ACCESS

Accession codes

Kit^{W}/*Kit*^{Wv} and *Kit*$^{+}$/*Kit*^{Wv} testis mRNA-seq reads have been deposited in GenBank under accession number SRA060831. SHIMS assemblies have been deposited in GenBank under accession numbers JH720451, JH720452, JH806589, KB021648, JH806590, JH806587, JH806591, JH806592, JH720453, JH720454, JH806593, JH806594, JH806595, JH806588, JH806601, JH806602, JH806603, JH806596, JH806597, JH806598, JH806599, JH806600, and JH159150. Specific information about BACs and fosmids used to generate the SHIMS assemblies is provided in Supplementary Table 1.

In 1967, Susumu Ohno predicted that catalogs of X-linked genes would differ little, if at all, among placental mammals[1]. Over the past 15 years, numerous comparative mapping studies across highly diverged mammals have supported what has become known as Ohno's law[2–11], although some individual gene exceptions have been noted[12,13]. We decided to perform a systematic and rigorous test of Ohno's law by comparing the human and mouse X chromosomes, including their gene contents. We chose these two X chromosomes because their reference sequences were determined via a high-quality, clone-based approach[14], were verified with high-resolution genetic maps[15,16], harbor significantly fewer gaps than all other sequenced X chromosomes (Table 1), and have been well-annotated[5,17].

A major difference between these two assemblies is that the mouse X-chromosome reference assembly is derived from a single haplotype[17], while the human X-chromosome reference represents a mosaic of X-chromosome sequences from at least 16 different individuals[5]. This mosaicism can lead to misassemblies in the human X-chromosome reference sequence, which, if left uncorrected, would confound our thorough testing of Ohno's law. It might explain why the human X-chromosome reference sequence does not contain seven large ampliconic regions (segmental duplications >10 kilobases in length that share >99% nucleotide identity) found in the mouse X-chromosome reference sequence (Fig. 1). Ampliconic regions are particularly prone to sequence misassembly[14], because the nucleotide identity between two amplicons (99.02% – 99.98%) is comparable to, if not greater than, the nucleotide identity between alleles (which can be as low as 99.40%[18]). Ampliconic regions assembled from multiple haplotypes may display expansions, contractions, or inversions that do not accurately reflect the structure of any extant haplotype. To thoroughly test Ohno's law, we constructed a more accurate assembly of the human X chromosome's ampliconic regions in order to compare the gene contents of the human and mouse X chromosomes.

We first identified all ampliconic regions of the human X chromosome, including those absent from the current reference sequence. We found 24 ampliconic regions present in the reference sequence by searching for duplicated segments >10 kb in length and exhibiting >99% nucleotide identity. To identify amplicons absent from the current reference sequence, we targeted regions surrounding gaps, which are generally enriched for amplicons[19], and regions where the reference sequence is discordant with a set of independent physical maps[20]. Together, these approaches yielded a total of 33 regions that merited scrutiny (Supplementary Table 1). Only four of the 33 regions were spanned by single-haplotype sequence, highlighting the mosaic nature of the human X-chromosome assembly. We chose to resequence the other 29 regions using an approach previously developed by our laboratories to sequence Y-chromosomal amplicons: single-haplotype iterative mapping and sequencing (SHIMS)[21–24]. This clone-based sequencing strategy utilizes single nucleotide differences between overlapping clones, all derived from a single haplotype, to accurately order and orient each clone across ampliconic sequences.

Using SHIMS, we generated 11.5 megabases (Mb) of non-overlapping sequence from 110 newly sequenced bacterial artificial chromosomes (BACs), 28 reassembled BACs, and 13 fosmids that collectively span all 29 regions (Supplementary Table 1). Of the 11.5 Mb of

sequence generated, 3.15 Mb was comprised of X-amplicons. We estimate the total size of the human X chromosome to be about 155.3 Mb, of which ~2% is ampliconic (Table 1).

Our SHIMS assembly substantially improved upon the current reference sequence (Supplementary Table 1). It closed four amplicon-associated gaps, corrected misassemblies of three large ampliconic regions (Fig. 2 and Supplementary Figs 1 and 2) and identified two previously unrecognized palindromic amplicons (Supplementary Fig. 3). As an example of the improved accuracy of this approach (Supplementary Note), our SHIMS assembly of one ampliconic region closed a gap, reduced the X-chromosome reference sequence by 236 kb, and turned an apparently complex collection of amplicons into a solitary palindrome (Fig. 2). This SHIMS assembly of X-amplicons will be incorporated into the reference sequence of the human X chromosome.

With our more accurate assembly and corresponding recalibration of the human X chromosome's gene content, we tested Ohno's law by systematically comparing the gene contents of the human and mouse X chromosomes. Contrary to Ohno's Law, 18% (144/800) of human and 23% (197/853) of mouse X-linked protein-coding genes are not shared between these two species (Fig. 3a and Supplementary Tables 2–4). In sum, this two-species comparison identified 341 genes that violate Ohno's law.

An exception to Ohno's law could arise through either gene loss or duplication of an ancestral X-linked gene, or independent acquisition of a novel gene. To identify cases of gene loss, we searched the following three outgroup species for orthologs of human and mouse X-linked genes that violate Ohno's law: dog (X chromosome)[25], horse (X chromosome)[26], and chicken (where autosomes 1 and 4 are homologous to the mammalian X chromosome)[27]. We concluded that a minority of the genes (55/144 in human, 34/197 in mouse) that violate Ohno's law are the result of lineage-specific gene loss (Fig. 3a and Supplementary Tables 3 and 4). To identify cases of duplication of an ancestral X-linked gene, we also used orthologous genes in dog, horse, and chicken for comparison. Only a small fraction of the genes (13/144 in human, 29/197 in mouse) that violate Ohno's law are due to duplication of an ancestral X-linked gene (Fig. 3a and Supplementary Tables 3 and 4). These findings indicate that in both lineages, the majority of genes (76/144 in human, 134/197 in mouse) that violate Ohno's law were independently acquired -- via transposition or retroposition from autosomes, or having arisen *de novo* on the X chromosome. Thus, surprisingly large fractions of X-linked genes (10% in human, 16% in mouse) have been acquired independently since the two lineages began to diverge from a common ancestor 80 million years ago.

We then counted the numbers of independently acquired and shared genes that are ampliconic (embedded in duplicated segments >10 kb in length and exhibiting >99% nucleotide identity), multicopy (only the gene structure is duplicated), or single-copy. Among independently acquired X-linked genes, roughly two-thirds are ampliconic (48/76 in humans, 102/134 in mice), while the remaining third are multicopy or single-copy (Fig. 3b and Supplementary Tables 3 and 4). Indeed, only 31% of human X-ampliconic genes (33/107) and 22% of mouse X-ampliconic genes (33/149) share orthologs (Supplementary Table 5). By contrast, 82% of shared X-linked genes are single-copy (548/656; Fig. 3b), and

an impressive 95% of human (548/575) and 94% of mouse (548/585) single-copy X-linked genes are shared (Supplementary Table 5). We conclude that, when comparing the X-linked genes of the human with those of the mouse, most exceptions to Ohno's law are ampliconic genes that were independently acquired in either the human or mouse lineage subsequent to their divergence from a common ancestor 80 million years ago (Fig. 3b). These exceptions provide a striking contrast to the shared, single-copy genes that follow Ohno's law.

We then compared the expression patterns of independently acquired and shared X-linked genes in eight human tissues and three mouse tissues, utilizing published[28–30] and newly generated RNA deep-sequencing (mRNA-seq) data. As a control, we analyzed all autosomal genes. We observed that most independently acquired human and mouse X-linked genes exhibit high expression in the testis and little or no expression in all other tissues examined (Fig. 3c, Supplementary Fig. 4 and Supplementary Tables 6–8). Since many of the independently acquired genes are members of multicopy or ampliconic gene families whose gene expression levels were averaged, it was important to rule out the possibility that only one family member is actively transcribed in the testis – which we did by scrutinizing the testis mRNA-seq data for sequence variants that differentiated between members of a gene family (Supplementary Table 9). The testis-predominant expression pattern of independently acquired genes is significantly different (Chi-square, $P < 0.0001$) from that of the shared, single-copy, X-linked genes (Fig. 3c, Supplementary Table 6). Notably, the proportion of shared, single-copy, X-linked genes that are expressed predominantly in the testis is much lower, and approximately the same as autosomal genes (Fig. 3c and Supplementary Tables 6, 10 and 11). In summary, we find that a common and distinguishing characteristic of most independently acquired X-linked genes is testis-predominant expression.

We next sought to determine whether independently acquired X-linked genes in mouse are expressed in germ cells or somatic cells of the testis. To do this we performed mRNA-seq analysis on adult testes of wild-type and Kit^{W}/Kit^{Wv} mice, the latter of which lack germ cells[31]. We found that most independently acquired genes are expressed specifically in testicular germ cells, regardless of whether they are single-copy, multicopy or ampliconic (Fig. 3c and Supplementary Tables 6–8). The proportion of independently acquired genes with high expression in wild-type testis and little or no expression in Kit^{W}/Kit^{Wv} testis is significantly higher (Chi-square, $P < 0.0001$) than that of either shared single-copy X-linked genes or autosomal genes (Fig. 3c and Supplementary Tables 6, 8 and 11). Additionally, in accordance with our previous studies[32], we find that most ampliconic genes, both shared and independently acquired, are also predominantly expressed in testicular germ cells (Fig. 3c and Supplementary Tables 6 and 8). Our findings underscore the importance of the male germline, relative to the soma, in promoting gene acquisition on a chromosome whose gene content is otherwise highly conserved.

Based on our present findings in human and mouse, we wonder whether the X chromosomes of other placental mammals (Supplementary Fig. 5) also harbor independently acquired ampliconic genes that are expressed predominantly in testicular germ cells. To answer this question in other species will require using a SHIMS approach to assemble their X-amplicons, and thus their reference sequences, completely and accurately (Table 1). If independently acquired, testis-expressed genes prove to be a general feature of mammalian

X chromosomes, then the acquisition of these genes may have contributed greatly to mammalian diversification and radiation, which began in the Paleocene epoch. This speculation is supported by a wealth of evidence that rapid evolution of hybrid male-sterility factors on animal X chromosomes has been an important driver of speciation[33]. In mouse, three strictly X-linked hybrid male-sterility loci[34–36] have been identified; all three map at or near X-ampliconic regions (Supplementary Fig. 6) harboring independently acquired genes expressed predominantly in spermatogenic cells.

The medical and biological significance of the independently acquired genes on the human X chromosome is essentially unexplored. To date, not a single X-linked phenotype (as catalogued in Online Mendelian Inheritance in Man) has been attributed, at the molecular level, to an independently acquired gene on the human X chromosome (Supplementary Table 12). By contrast, 238 X-linked traits have been traced, at the molecular level, to genes shared between human and mouse. Given that the independently acquired genes are expressed predominantly in spermatogenic cells, one might anticipate that loss-of-function mutations affecting these genes or gene families would perturb male gametogenesis – a possibility that can now be explored using the SHIMS reference sequence of the human X-ampliconic regions.

Our findings also provide a plausible explanation for how so many X-linked genes are able to defy Ohno's law. Ohno's law assumed that any given X-linked gene would be expressed in both sexes, and equally so. Consistent with this, we found that, in both humans and mice, >96% of genes that follow Ohno's law are expressed in both sexes (Supplementary Tables 13 and 14). However, not all genes function equivalently in males and females, and indeed some genes are expressed in one sex but not the other. As we have shown, the genes that violate Ohno's law are expressed in males but not females. The fact that many genes are expressed sex-specifically would not have been appreciated at the time of Ohno's writing, in the 1960's.

In summary, our study places Ohno's law within a larger context. Based on construction and analysis of a more complete and accurate human X reference sequence, our comparison between human and mouse X chromosomes enables us to characterize important exceptions to the law: in both species, large numbers of genes that are expressed in spermatogenic cells, and most of which are ampliconic or multicopy. We conclude that the gene repertoires of the human and mouse X chromosomes are products of two complementary, evolutionary processes: conservation of single-copy genes that serve functions shared by the sexes, and ongoing gene acquisition, usually involving formation of amplicons, which serves to differentiate and specialize X chromosomes toward functions in male gametogenesis.

URLs

LASTZ, http://www.bx.psu.edu/miller_lab/dist/README.lastz-1.02.00/README.lastz-1.02.00a.html; R-software for dot-plots, http://www.R-project.org; Custom perl script for triangular dot plots, http://pagelab.wi.mit.edu/material-request.html.

METHODS

Nucleotide sequence comparisons

Entire X-chromosome files for human (hg18), chimpanzee (panTro4), rhesus (rheMac3), mouse (mm9), dog (canFam3), horse (equCab2), and cow (bosTau7) were downloaded from the UCSC genome browser database[37]. Alignments of repeat-masked X-chromosome sequences were generated with BLASTZ[38], using non-gapped alignment settings and a step length of 20 nucleotides. Coordinates were obtained for stretches of alignable sequence that scored >3000 (the equivalent of a 30-bp perfect match), using default gap and mismatch penalties and rewards parameters. Square dot plots were generated with R software dot-plot package. Triangular dot plots were performed using a custom Perl script[23].

Selection of human X-chromosome regions for single-haplotype assembly

To identify human X-ampliconic regions, we considered regions of the human X-chromosome reference sequence falling into one of three categories:

1. Regions containing amplicons in the current reference sequence. A collection of previously detected segmental duplications[39] was filtered for duplications meeting the following criteria: repeat unit >10 kb with >99.0% identity between copies and <500 kb separation between copies.
2. Regions containing gaps. Ampliconic sequences are known to be associated with gaps in genome assemblies[14]. In the human reference sequence, gaps are marked by long stretches of N's, denoting missing sequence of unknown length. We scanned the non-pseudoautosomal reference sequence (hg18) for such large stretches of N's and identified the sequence coordinates of nine gaps.
3. Regions containing misoriented, physically mapped clones. We used position and orientation information for fosmid paired-end sequences (from eight different libraries[20]) previously aligned to the human X-chromosome reference sequence. X-chromosome regions where fosmid paired-end sequences, from at least three libraries, did not map in similar orientation to the reference sequence were considered putative ampliconic regions.

A total of 33 regions (Supplementary Table 1) were identified using these three approaches. For each of the 33 regions, we identified the library of origin for each reference-sequence clone in order to determine which portions of the reference assembly, if any, were composed of single-haplotype sequence.

Clone selection and sequencing

For the 29 regions not comprised of single haplotype sequence, we employed the single haplotype iterative mapping and sequencing (SHIMS) approach, as previously performed for Y and Z chromosome assemblies[21–24]. We used publicly available BAC fingerprint maps, fosmid- and BAC-end sequences and current X-chromosome reference sequence[5] as sources for generating markers.

We selected and sequenced BAC and fosmid clones that collectively spanned each region. For each region that included a gap in the current reference sequence, we selected a tiling path of clones stretching 500 kb to either side of the gap. Analysis of this ~1 MB of sequence allowed us to determine if sequences flanking the gap were ampliconic. For each region of the X chromosome that appeared to be ampliconic in the current reference sequence, or that contained misoriented fosmid ends, we selected a tiling path of clones stretching 100 kb to either side of the amplicon or misoriented fosmid-end sequence.

We primarily selected human X-chromosome BAC clones from the RP-11 male library[40]. In those instances where RP-11 BACs did not provide sufficient coverage of a region, we selected clones from the haploid CH-17 library. In some instances, amplicon repeat units were too short to be assembled accurately within a single BAC (average BAC insert size: 160 kb). For such cases, we selected clones from the ABC8 male fosmid library[20]; these clones have smaller inserts (~40 kb), which enabled us to order and orient amplicons with shorter repeat units. We used only one library (either RP-11 or CH-17 or ABC-8) to span each of the 29 ampliconic regions sequenced. In a few cases, we used alternative ABC fosmids (from libraries ABC-7, ABC-9, ABC-12, ABC-13, and ABC-14)[20] to extend into gaps that are not ampliconic.

BACs and fosmid sequences will be incorporated into the next update of the reference assembly (GRCh38). Supplementary Table 1 provides Genbank accession numbers for all BACs and fosmids as well as for SHIMS assemblies of the 29 regions sequenced.

Comparisons of human and mouse gene orthologs

Reference sequences for the human (hg18) and mouse (mm9) protein-coding gene sets were downloaded from the UCSC genome browser[37]. We selected the isoform yielding the longest peptide sequence for each gene, resulting in 821 and 865 genes for the human and mouse X chromosomes, respectively. These lists of genes were curated to provide an unbiased and comprehensive comparison of human and mouse X-linked gene content, as follows:

1. All pseudoautosomal genes were removed, because our analysis was limited to strictly X-linked genes. The case of the steroid sulfatase (*STS*) gene merits special mention. The human *STS* gene is X-linked. In mice, *Sts* is absent from the reference genome assembly, but multiple EST sequences have been reported. Previous studies[41] and our unpublished data (Supplementary Table 2) are consistent with the mouse *Sts* gene mapping to the X chromosome, within or near the pseudoautosomal region. We included *Sts* in the mouse gene set.
2. For 11 genes (Supplementary Table 2), we determined that the gene was multicopy in human but ampliconic in mouse, or vice versa. We excluded these genes from all tallies and analyses because we could not infer whether the genes were multicopy or ampliconic in the common ancestor of humans and mice.
3. We updated and corrected the human gene set to reflect our SHIMS sequence assembly across ampliconic regions. We searched novel genomic sequence generated in this study for genes using Genomescan[42] and BLAST[43] analyses of

human EST databases. In the case of ampliconic regions that were either expanded or contracted in our revised assembly, we recounted the numbers of genes for each gene family within the regions.

We arrived at 800 human and 853 mouse X-linked genes. These revised gene sets served as the basis for all subsequent comparative and expression analyses.

All X-linked genes determined to be shared between human and mouse were identified by having either a best reciprocal BLAST[43] alignment between the two species or a TBLASTN alignment to a syntenic, unannotated region of the compared X chromosome (Supplementary Table 2). Such regions were classified as unannotated genes when the predicted protein-coding gene sequence was free of nonsense mutations and there was evidence of transcription from either ESTs or mRNA-seq data[28,29].

Genes present on either the human X chromosome or the mouse X chromosome (but not both) could either have been lost in one lineage, duplicated in one lineage, or independently acquired in one lineage. To distinguish among these three possibilities, we determined, via TBLASTN, if X-linked genes present in either humans or mice (but not both) had orthologs on the dog X chromosome (canFam3), the horse X chromosome (equCab2), or syntenic regions of chicken chromosomes 1 and 4 (galGal4). Comparisons with these three outgroups helped us to infer whether a given gene was present on the X chromosome in the common ancestor of humans and mice. Each gene was classified as follows:

A. Lineage-specific gene loss: A gene with an ortholog in a syntenic chromosomal region in one or more of the three outgroups, or with a pseudogene ortholog in the syntenic region of the human or mouse X chromosome, was judged to have been lost.

B. Lineage-specific gene duplication: A gene duplicate (paralog) of a pre-existing X-linked gene that does not have an orthologous duplicate gene in the other species (human or mouse), or in a syntenic chromosomal region of in one or more of the three outgroups was judged to be a lineage-specific duplicate of a pre-existing X-linked gene.

C. Independently acquired: A gene not falling into either of these two categories was judged to have been independently acquired.

Our inferences regarding human and mouse X-linked gene losses and gains are based on comparisons with the current dog, horse, and chicken genome assemblies. As the assemblies of the dog, horse, and chicken genomes are not as complete as those of the human and mouse X chromosomes, our inferences should be reexamined in the future when more complete and accurate assemblies of the dog X chromosome, horse X chromosome, and chicken chromosomes 1 and 4 are available. In the Supplementary Note, we elaborate on these limitations and associated uncertainties.

Shared and species-specific genes were grouped into single-copy, multicopy or ampliconic. We defined multicopy genes as members of gene pairs/families exhibiting >50% amino acid identity across 80% of the protein and an e-value $<1x10^{-20}$ when protein sequences are aligned[44]. We defined ampliconic genes as genes located within a stretch of ampliconic

sequence (repeat unit >10 kb in length with >99% nucleotide identity and 500-kb separation).

mRNA-seq of testis cDNA

We crossed C57BL/6J-*Kit*^{Wv} (The Jackson Laboratory) males to WB/ReJ-*Kit*^{W} (The Jackson Laboratory) females to generate *Kit*^{W}/*Kit*^{Wv} compound-heterozygous males, which are germ-cell deficient, and control *Kit*$^{+}$/*Kit*^{Wv} males. Two biological replicate testes from *Kit*^{W}/*Kit*^{Wv} and *Kit*$^{+}$/*Kit*^{Wv} males were collected at ~3 months of age. Total RNA (1–2 ug) was extracted using Trizol (Invitrogen) according to manufacturer's instructions. Hemoglobin transcripts were selectively removed from total RNA by following GLOBINclear (Ambion) protocol recommendations. As per the Illumina kit protocol, poly-A-selected mRNA was used to generate mRNA-seq cDNA libraries using random-hexamer primers. cDNA fragments of ~200 nucleotides were isolated and modified for sequencing following the mRNA-seq protocol (Illumina). The Illumina Genome Analyzer II platform was used to sequence 36-mers from the mRNA-seq libraries by following the manufacturer's recommendations. *Kit*^{W}/*Kit*^{Wv} and *Kit*$^{+}$/*Kit*^{Wv} testis mRNA-seq reads have been deposited in GenBank under accession number SRA060831. The Massachusetts Institute of Technology's committee on animal care has approved all experiments involving mice.

RNA-seq analyses

Previously published mRNA-seq datasets from human[28,30] (adipose, colon, heart, liver, lymph node, skeletal muscle, ovary and testis) and mouse[29] (liver and skeletal muscle) were combined with our newly generated *Kit*$^{+}$/*Kit*^{Wv} testis and *Kit*^{W}/*Kit*^{Wv} testis datasets to determine the tissue expression pattern for each X-linked gene. For each tissue, mRNA-seq reads were aligned to the reference genome assembly using Tophat[45] with default settings. FPKM (fragments per kb of exon model per million mapped fragments) values were estimated using Cufflinks[46] with the reference sequence gene set used as an annotation file. Unannotated genes with orthologs in the reciprocal species (29 such cases) were excluded due to concerns regarding accurate estimates of FPKMs.

Cufflinks has difficulty accurately calculating FPKMs for multicopy and ampliconic genes, so we estimated FPKMs for these two gene classes using a customized method. FPKMs for multicopy and ampliconic genes were determined by aligning all reads to a representative gene family member. This total read count, per gene family, was then divided by the length of the gene, number of gene copies, and the number of reads mapped to the genome, resulting in an FPKM value for each ampliconic or multicopy gene family. To determine if multiple members of a multicopy or ampliconic gene family are expressed, we identified nucleotide variants that uniquely identified individual copies. We then counted the number of mRNA-seq reads, in human[30] and mouse testis samples[47], that aligned to each variant.

Genes with >1 FPKM value in testis and <1 FPKM in ovaries and all somatic tissues examined were considered to be expressed predominantly in testis. Similarly, genes with >1 FPKM value in *Kit*$^{+}$/*Kit*^{Wv} testes and <1 FPKM in *Kit*^{W}/*Kit*^{Wv} testis and in all other somatic tissues examined were considered to be expressed predominantly in testicular germ cells.

(Previous studies have used >1 FPKM as a cutoff for considering a gene to be expressed in a tissue[48].)

To determine if X-linked genes that follow Ohno's law are expressed in both sexes, we analyzed previously published mRNA-seq datasets from male and female human and mouse tissues[47]. We performed alignments to calculate FPKM values as described above. We considered a gene to be expressed in one sex but not the other if it met both of the following criteria:

1. FPKM >1 in one sex and FPKM <1 in the other sex.
2. At least three-fold higher expression in one sex as compared with the other sex.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank D. Albracht, J. Collins, M. Gill, N. Koutseva, C. Kremitzki, A. van der Veen, and J. Wood for technical assistance, and D. Bellott, R. Desgraz, G. Dokshin, T. Endo, A. Godfrey, Y. Hu, J. Hughes, M. Kojima, B. Lesch, L. Okumura, K. Romer, and Y. Soh for comments on the manuscript. Supported by the National Institutes of Health and the Howard Hughes Medical Institute.

References

1. Ohno, S. Sex Chromosomes and Sex-linked Genes. Springer; Berlin: 1967.
2. Kuroiwa A, et al. Conservation of the rat X chromosome gene order in rodent species. Chromosome Res. 2001; 9:61–7. [PubMed: 11272793]
3. Delgado CL, Waters PD, Gilbert C, Robinson TJ, Graves JA. Physical mapping of the elephant X chromosome: conservation of gene order over 105 million years. Chromosome Res. 2009; 17:917–26. [PubMed: 19789986]
4. Prakash B, Kuosku V, Olsaker I, Gustavsson I, Chowdhary BP. Comparative FISH mapping of bovine cosmids to reindeer chromosomes demonstrates conservation of the X-chromosome. Chromosome Res. 1996; 4:214–7. [PubMed: 8793206]
5. Ross MT, et al. The DNA sequence of the human X chromosome. Nature. 2005; 434:325–37. [PubMed: 15772651]
6. Veyrunes F, et al. Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. Genome Res. 2008; 18:965–73. [PubMed: 18463302]
7. Watanabe TK, et al. A radiation hybrid map of the rat genome containing 5,255 markers. Nat Genet. 1999; 22:27–36. [PubMed: 10319858]
8. Raudsepp T, et al. Exceptional conservation of horse-human gene order on X chromosome revealed by high-resolution radiation hybrid mapping. Proc Natl Acad Sci U S A. 2004; 101:2386–91. [PubMed: 14983019]
9. Band MR, et al. An ordered comparative map of the cattle and human genomes. Genome Res. 2000; 10:1359–68. [PubMed: 10984454]
10. Murphy WJ, Sun S, Chen ZQ, Pecon-Slattery J, O'Brien SJ. Extensive conservation of sex chromosome organization between cat and human revealed by parallel radiation hybrid mapping. Genome Res. 1999; 9:1223–30. [PubMed: 10613845]
11. Spriggs HF, et al. Construction and integration of radiation-hybrid and cytogenetic maps of dog Chromosome X. Mamm Genome. 2003; 14:214–21. [PubMed: 12647244]
12. Palmer S, Perry J, Ashworth A. A contravention of Ohno's law in mice. Nat Genet. 1995; 10:472–6. [PubMed: 7670497]

13. Rugarli EI, et al. Different chromosomal localization of the Clcn4 gene in Mus spretus and C57BL/6J mice. Nat Genet. 1995; 10:466–71. [PubMed: 7670496]
14. She X, et al. Shotgun sequence assembly and recent segmental duplications within the human genome. Nature. 2004; 431:927–30. [PubMed: 15496912]
15. Olivier M, et al. A high-resolution radiation hybrid map of the human genome draft sequence. Science. 2001; 291:1298–302. [PubMed: 11181994]
16. Dietrich WF, et al. A comprehensive genetic map of the mouse genome. Nature. 1996; 380:149–52. [PubMed: 8600386]
17. Church DM, et al. Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol. 2009; 7:e1000112. [PubMed: 19468303]
18. Tishkoff SA, Kidd KK. Implications of biogeography of human populations for 'race' and medicine. Nat Genet. 2004; 36:S21–7. [PubMed: 15507999]
19. Bovee D, et al. Closing gaps in the human genome with fosmid resources generated from multiple individuals. Nat Genet. 2008; 40:96–101. [PubMed: 18157130]
20. Kidd JM, et al. Mapping and sequencing of structural variation from eight human genomes. Nature. 2008; 453:56–64. [PubMed: 18451855]
21. Skaletsky H, et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature. 2003; 423:825–37. [PubMed: 12815422]
22. Hughes JF, et al. Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. Nature. 2010; 463:536–9. [PubMed: 20072128]
23. Kuroda-Kawaguchi T, et al. The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. Nat Genet. 2001; 29:279–86. [PubMed: 11687796]
24. Bellott DW, et al. Convergent evolution of chicken Z and human X chromosomes by expansion and gene acquisition. Nature. 2010; 466:612–6. [PubMed: 20622855]
25. Lindblad-Toh K, et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature. 2005; 438:803–19. [PubMed: 16341006]
26. Wade CM, et al. Genome sequence, comparative analysis, and population genetics of the domestic horse. Science. 2009; 326:865–7. [PubMed: 19892987]
27. Consortium ICGS. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 2004; 432:695–716. [PubMed: 15592404]
28. Wang ET, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008; 456:470–6. [PubMed: 18978772]
29. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008; 5:621–8. [PubMed: 18516045]
30. Bradley RK, Merkin J, Lambert NJ, Burge CB. Alternative splicing of RNA triplets is often regulated and accelerates proteome evolution. PLoS Biol. 2012; 10:e1001229. [PubMed: 22235189]
31. Handel MA, Eppig JJ. Sertoli cell differentiation in the testes of mice genetically deficient in germ cells. Biol Reprod. 1979; 20:1031–8. [PubMed: 476239]
32. Mueller JL, et al. The mouse X chromosome is enriched for multicopy testis genes showing postmeiotic expression. Nat Genet. 2008; 40:794–9. [PubMed: 18454149]
33. Coyne, JA.; Orr, HA. Speciation. Sinauer Associates; Sunderland, Mass: 2004.
34. Elliott RW, et al. Genetic analysis of testis weight and fertility in an interspecies hybrid congenic strain for Chromosome X. Mamm Genome. 2001; 12:45–51. [PubMed: 11178743]
35. Elliott RW, Poslinski D, Tabaczynski D, Hohman C, Pazik J. Loci affecting male fertility in hybrids between Mus macedonicus and C57BL/6. Mamm Genome. 2004; 15:704–10. [PubMed: 15389318]
36. Storchova R, et al. Genetic analysis of X-linked hybrid sterility in the house mouse. Mamm Genome. 2004; 15:515–24. [PubMed: 15366371]
37. Fujita PA, et al. The UCSC Genome Browser database: update 2011. Nucleic Acids Res. 2011; 39:D876–82. [PubMed: 20959295]

38. Schwartz S, et al. Human-mouse alignments with BLASTZ. Genome Res. 2003; 13:103–7. [PubMed: 12529312]
39. Bailey JA, et al. Recent segmental duplications in the human genome. Science. 2002; 297:1003–7. [PubMed: 12169732]
40. Osoegawa K, et al. A bacterial artificial chromosome library for sequencing the complete human genome. Genome Res. 2001; 11:483–96. [PubMed: 11230172]
41. Salido EC, et al. Cloning and expression of the mouse pseudoautosomal steroid sulphatase gene (Sts). Nat Genet. 1996; 13:83–6. [PubMed: 8673109]
42. Yeh RF, Lim LP, Burge CB. Computational inference of homologous gene structures in the human genome. Genome Res. 2001; 11:803–16. [PubMed: 11337476]
43. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215:403–10. [PubMed: 2231712]
44. Thornton K, Long M. Rapid divergence of gene duplicates on the Drosophila melanogaster X chromosome. Mol Biol Evol. 2002; 19:918–25. [PubMed: 12032248]
45. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25:1105–11. [PubMed: 19289445]
46. Trapnell C, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010; 28:511–5. [PubMed: 20436464]
47. Brawand D, et al. The evolution of gene expression levels in mammalian organs. Nature. 2011; 478:343–8. [PubMed: 22012392]
48. Deng X, et al. Evidence for compensatory upregulation of expressed X-linked genes in mammals, Caenorhabditis elegans and Drosophila melanogaster. Nat Genet. 2011; 43:1179–85. [PubMed: 22019781]

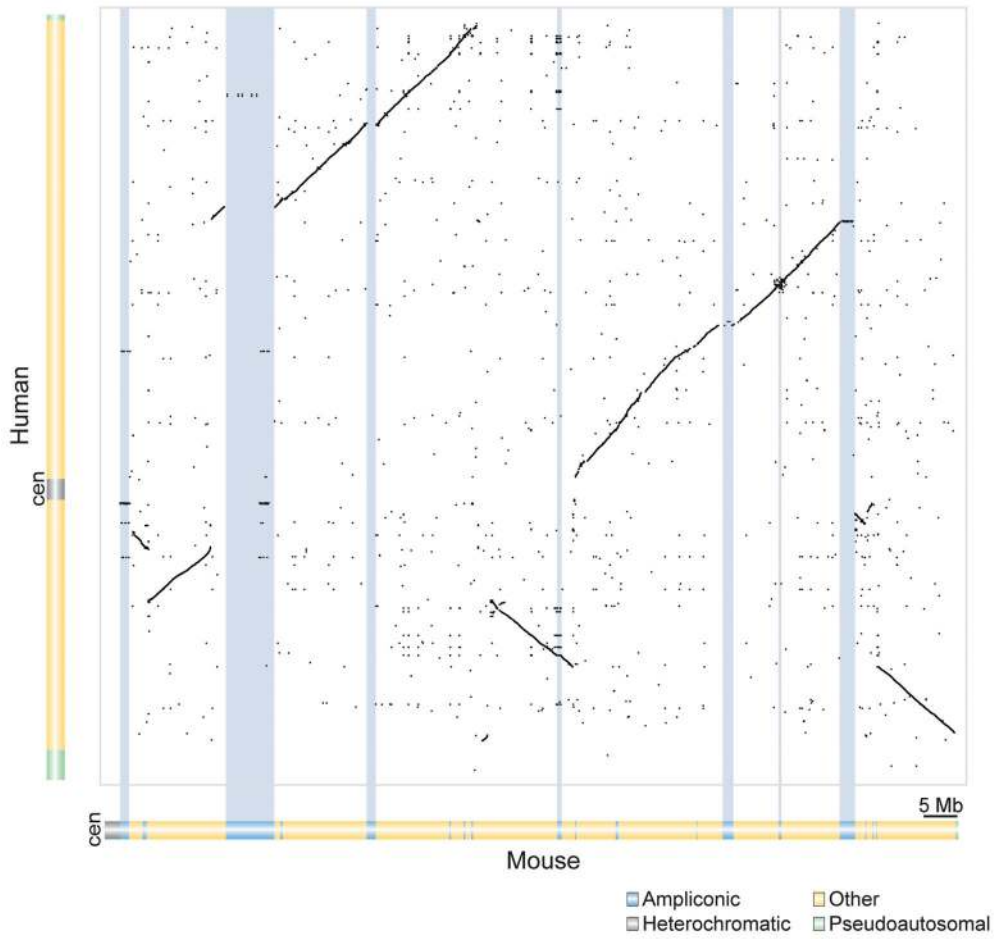

Figure 1. Dot-plot comparison of nucleotide sequences of human and mouse X chromosomes reveals large, divergent ampliconic regions on mouse X chromosome

Dot plot generated from BLASTZ nucleotide alignments of the human X chromosome assembly, prior to our SHIMS refinement (vertical axis), and the single-haplotype mouse X chromosome assembly (horizontal axis); each dot represents >70% nucleotide identity, within a 10-kb window, at that position. Within the plot, diagonal lines indicate syntenic blocks between the two chromosomes; regions lacking such diagonals are comprised of species-specific sequences. Blue shading highlights divergent ampliconic regions, each >500 kb in length, on the mouse X chromosome.

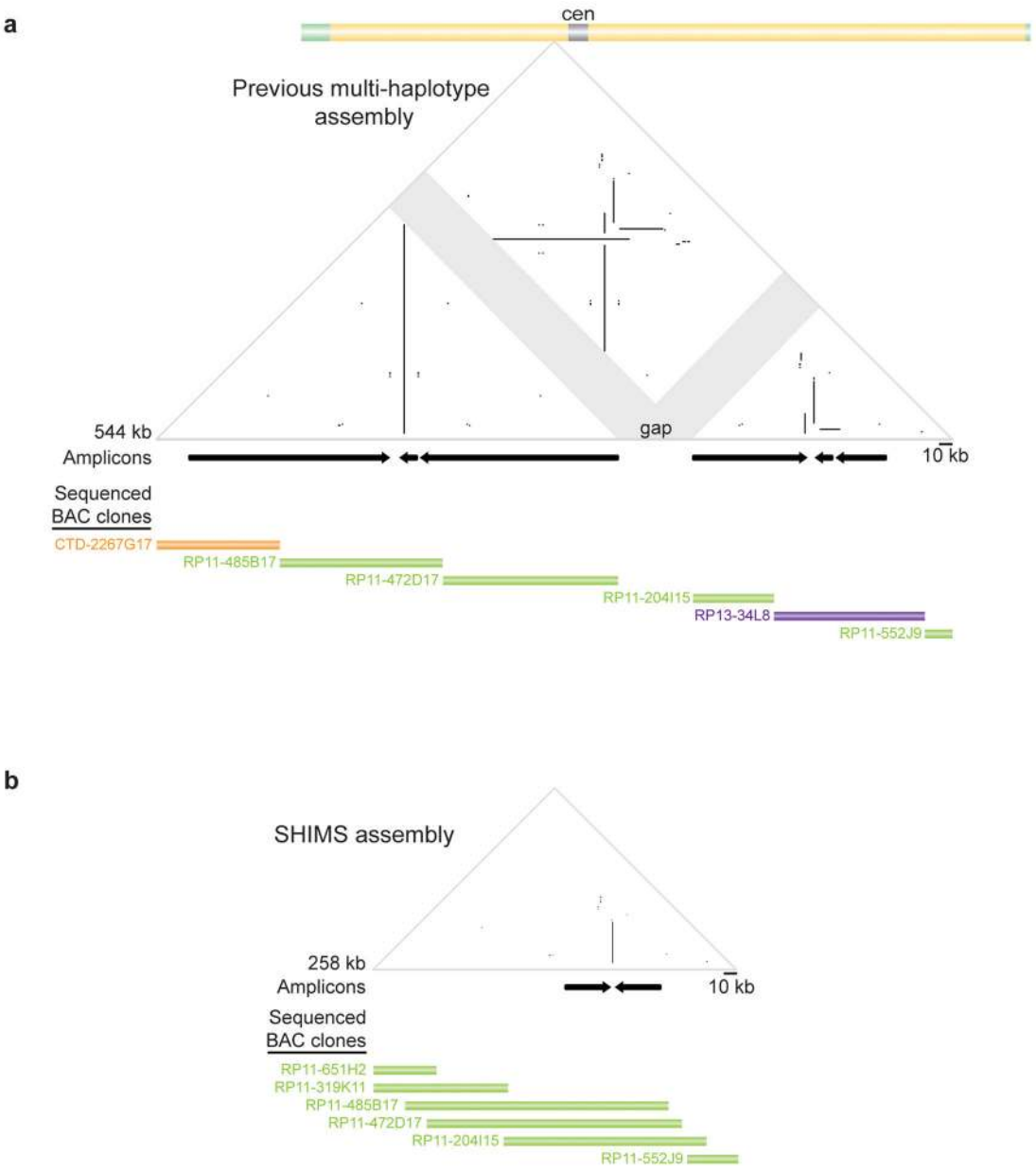

Figure 2. Comparison of mosaic and SHIMS sequence assemblies across one region of human X chromosome

a, Triangular dot-plot highlights sequence similarities within mosaic (multi-haplotype) assembly. Each dot represents 100% identity within a window of 100 nucleotides; direct repeats appear as horizontal lines, inverted repeats as vertical lines, and palindromes as vertical lines that nearly intersect the baseline; gaps are indicated by gray shading. Black arrows immediately below plots denote positions and orientations of amplicons. Further below, sequenced BACs from CTD, RP-11, and RP-13 libraries (each from a different individual) contributing to the assembly are depicted as orange, green, and purple bars, respectively; each bar reflects extent and position within assembly of finished sequence for that BAC. (As per the human genome assembly standard, finished-sequence overlaps

between adjoining BACs are limited to 2 kb.) GenBank accession numbers in Supplementary Table 1. **b,** SHIMS assembly of same region. All BACs derive from RP-11 library (one male) and are fully sequenced; each BAC's finished sequence extensively overlaps those of adjoining BACs.

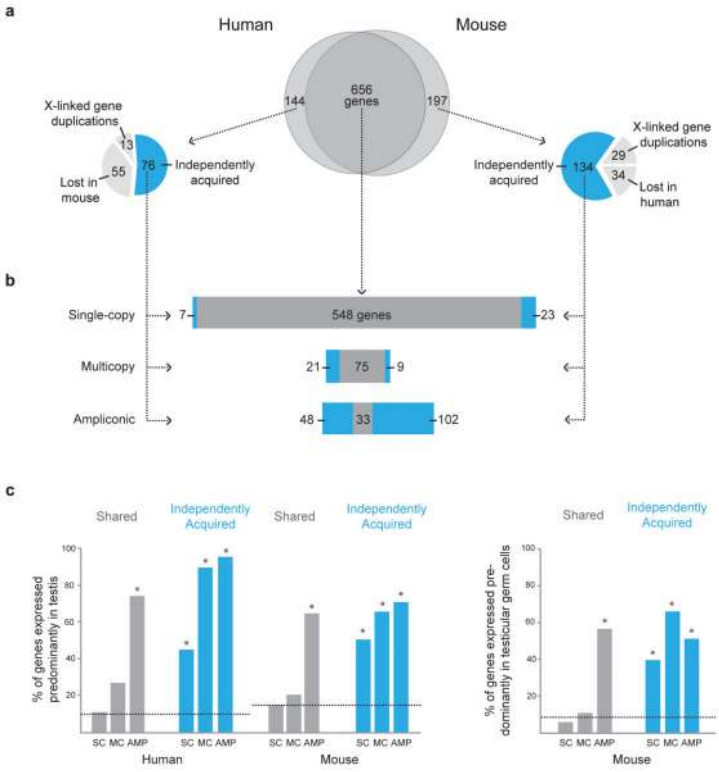

Figure 3. Comparison of X-linked gene classes between human and mouse

a, At center, Venn diagram depicts all human and mouse X-linked genes that are shared or not shared. To left and right, pie charts depict species-specific genes independently acquired in that lineage (blue), duplicated from an ancestral X-linked gene in that lineage (light gray), or lost in the opposite lineage (light gray). Venn diagram and pie charts are drawn to scale (by gene number). **b**, Horizontal bar stacks of single-copy, multicopy and ampliconic genes shared (dark gray) and independently acquired (blue) on human and mouse X chromosomes. Bar stacks are to scale (by gene number). **c**, Percentages of genes expressed predominantly in testis and in testicular germ cells. Horizontal dotted lines represent percentages of autosomal genes exhibiting testis- or testicular-germ-cell-predominant expression. SC, single-copy; MC, multicopy; AMP, ampliconic. Each asterisk indicates Chi-square test with Yates' correction $p < 0.0001$ (degrees of freedom = 1) when compared to either autosomal genes or X-linked single-copy genes.

Table 1

X-chromosome sequence assemblies in placental mammals

Organism	Sequencing strategy	# of gaps in X assembly	X-ampliconic sequence (Mb)	% of X chromosome composed of amplicons
Human	clone-based	5	3.15	2.0
Mouse	clone-based	25	19.42	11.6
Chimpanzee	whole-genome shotgun	10286	0.00*	0.00*
Rhesus	whole-genome shotgun	1996	0.16*	0.05*
Dog	whole-genome shotgun	215	0.05*	0.04*
Horse	whole-genome shotgun	2240	0.00*	0.00*
Cow	whole-genome shotgun	442	1.40*	0.93*

Human data reflects our revised, SHIMS-based assembly. Asterisks denote numbers based on whole-genome shotgun assemblies, which likely underestimate X-ampliconic content.