Materials and Methods

Retrogene Screen

To identify putatively functional retrogenes, we used FASTA to perform homology searches with each single exon peptide in the Ensembl (*S1*) gene sets for humans and mouse (homo_sapiens_10_30 and mus_musculus_10_3) against all genes in the same species set. We then kept only alignments that had at least 50 amino acids, aligned over at least 75% of both genes, and had at least 50% amino acid identity. For each single exon gene, we chose as the best hit the gene with the most identical amino acids for all alignments, reducing our data set to the one best hit per single exon gene. We then focused only on gene pairs where the single exon gene's best hit was a multiexon gene. We disregarded single exon genes that hit other single exon genes, as such pairs are not clearly retroposition events. To focus on genes likely to be functional, we kept pairs where the alignment was constrained at the amino acid level. We define constrained following a published procedure (*S2*), where $K_A/K_S < 0.5$ ($P < 0.05$) in the comparison between the parental and retrogene based on a likelihood ratio test (*S3, S4*). We also kept retrogenes for which unambiguous EST evidence was available (see Unigene Analysis). Of these genes, we only examined those alignments that resulted in chromosomal movement. Finally, we manually inspected each of the remaining alignments to discard pairs where the intron(s) of the putative parental genes did not align to an ungapped coding region in the putative retrogene. All intermediate steps in the screen were carried out with PERL scripts using BioEnsembl and BioPerl. Intermediate results were stored in a MySQL database. Final alignments were made using ClustalW using default parameters.

Affymetrix Expression Analysis

We downloaded expression data from http://expression.gnf.org/ for both human and mouse (*S5*). We 'clipped' the average difference values for each tissue such that the lowest value for any probe set was 20, representing a negligible level of expression. We then compared the expression of two replicate extractions of whole blood to test for data

quality. Each replicate was derived from an averaging of 3 replicate hybridizations for two male patients. In total, 99% of the genes in this comparison showed a lower than 2-fold difference in expression level, indicating that most genes outside of this threshold represent true differences in expression. We then compared the gonad data in the same way to identify genes that were expressed in ovaries and testis. Higher thresholds (4-fold and 10-fold) showed similar results.

Identification of Retropseudogenes

We retrieved 23,033 peptide sequences encoded by 18,459 genes from the Ensembl (*S1*) database (release 8.30a1) that contain at least one intron within the boundaries of their coding region. To screen for retroposed copies, these peptide sequences were used as queries in translated similarity searches against the complete human genome sequence using tBLASTn (*S6*). Adjacent homology matches were merged in a series of parsing steps using Perl scripts, only combining nearby matches (distance < 40 bp) that were likely not separated by introns. Furthermore, we required that query and merged target sequences had significant similarity on the amino acid level (amino acid identity > 50%) and aligned to one another over > 80% of the length of their sequence. Next, we performed similarity searches of the putative retroposed copies against themselves as well as all Ensembl genes using FASTA (*S7*). We kept only the copies where the closest hit was an Ensembl peptide with multiple coding exons. Based on the FASTA alignments, we identified 1,859 retroposed copies with frameshifts and/or stop codons and whose parental gene is on a different chromosome. We also confirmed the absence of introns in these retropseudogenes by mapping parental intron locations onto the alignments.

Unigene Analysis

To analyze expression patterns of human retrogenes, we performed similarity searches of retrogene and parental DNA sequences against all human ESTs (3,739,155) from the Unigene database (ftp://ftp.ncbi.nih.gov/repository/UniGene/, release 150) using Blast (*S6*). A Blast hit showing > 98% identity over 100 nucleotides or more between a retrogene and EST was counted as a sequence match, if the EST matched the retrogene

better than the parental gene according to the Blast score. In addition, we confirmed that the EST belongs to a Unigene cluster that maps to the same chromosome as the retrogene. In order to assess the overall spatial expression patterns of genes among chromosomes, we selected 15,666 human Unigene clusters, which were all supported by at least one mRNA and for which tissue and gene location information was available. Differences in expression between different sets of data were assessed using 2 x 2 contingency tables and c^2 and Fisher's Exact tests.

Dating of Retroposition Events

We estimated the age of retroposition events involving the X-chromosome both by comparing syntenic regions between the human and mouse genomes as well as by calculating sequence divergences (K_S). To analyze whether orthologs of retrogenes in one genome are present/absent in syntenic regions of the other, we compared chromosomal locations of retrogenes as provided by Ensembl and the coordinates of syntenic regions between human and mouse genomes (UCSC database, ftp://genome.cse.ucsc.edu/goldenPath/) using the same genome assembly (golden Path 28jun2002, NCBI release 30). Furthermore, we calculated K_S between parental and retrogene (K_{Spr}) as well as K_S between the parental gene and its ortholog in human or mouse (K_{Spo}) using PAML (*S3*). Presence of a retrogene homolog in a syntenic region and/or $K_{Spr}/K_{Spo} > 1$ was taken as evidence that the retrogene originated before the human-mouse split. In contrast, absence of a homolog in a syntenic region and $K_{Spr}/K_{Spo} < 1$ indicated that retroposition took place after the human-mouse split.

Statistical analyses

To test the patterns of retroposition, we compared the observations and the expectation of the random model of the movements that involve both autosomes and the X-chromosome. Expectations are given in Table S3 and S4. Because expected cell counts for several chromosomes are less than 5, we conducted Monte Carlo resampling analyses to compute the probabilities of the observed retroposed pattern under null hypothesis that they are all samples drawn from random retroposed instertions. The functions from the GNU Scientific Library was used to generate 10^6 multinomial deviates of the entire genome based on the expected

probabilities. The simulated statistic X^2 ($X^2=\Sigma[D_i-E_i]^2/E_i$ where i is the chromosome and D_i is the simulated multinomial deviate for chromosome i) was calculated for each multinomial simulation, and this distribution was compared against the observed X^2 value. P is the proportion of simulated X^2 statistics that exceed the observed X^2 statistic. For example, only 123 simulations involving expectations for functional export in humans exceeded the observed value of the statistic, leading to a P=123/1,000,000 = 0.000123. For comparison, we also conducted the analyses excluding the X as a parent and a target respectivelyin calculation of both expected and observated numbers. The *P* values were calculated for these autosomal groups by analyzing only the movements into or out of autosomes. We found that the autosomes had both generated and recruited retrogenes according to the random expectation (retrogenes leaving an autosome for another autosome: $P = 0.1243$ for humans and $P = 0.7931$ for mouse; retrogene entering an autosome from another autosome: P=0.2956 for humans and P = 0.1069 for mouse). Thus, any significant deviation detected in pooled datasets including the X-chromosome and autosomes should be attributed to the X-chromosome. Furthermore, we carried out the outlier tests in which outliers were determined by the Grubbs and Dixon methods (*S8*) on the ratio of observed/expected genes from the regression for all 6 data sets. Shapiro-Wilks normality tests (*S9*) showed that such ratios follow the requirement of both tests.

Supplementary References

S1. T. Hubbard *et al.*, *Nucleic. Acids. Res.* **30**, 38 (2002).

S2. E. Betrán, K. Thornton, M. Long, *Genome Res.* **12**, 1854 (2002).

S3. Z. Yang, *Comput. Appl. Biosci.* **13**, 555 (1997).

S4. Z. Yang, *Mol. Biol. Evol.* **15**, 568 (1998).

S5. R. B. Altman, S. Raychaudhuri, *Curr. Opin. Struct. Biol.* **11**, 340 (2001).

S6. S. F. Altschul *et al.*, *Nucleic. Acids. Res.* **25**, 3389 (1997).

S7. W. R. Pearson, *Methods Mol. Biol.* **132**, 185 (2000).

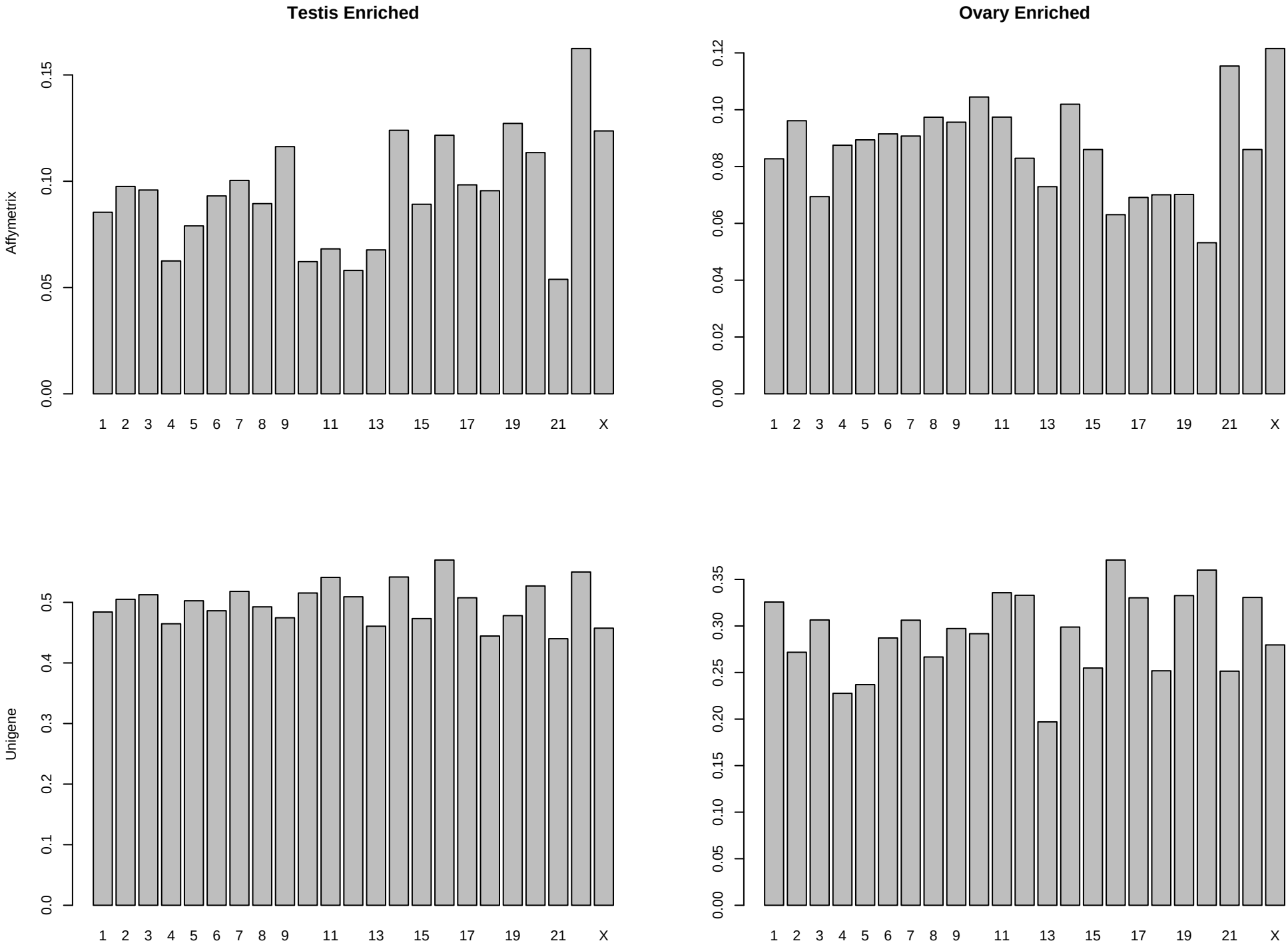
S8. R. R. Sokal, F. J. Rohlf. Biometry (third ed.) (Freeman, New York, 2000).

S9. P. Royston. *Applied Statistics* **44**, 547 (1995).

S10. E. M. Eddy, D. A. O'Brien, *Current topics in Developmental Biology* **37**, 141 (1998).

S11. A. Calenda, B. Allenet, D. Escalier, J. Bach, H. Garchon, *EMBO J.* **52**, 103 (1994).

Fig. S1. Proportions of sex-specific expression for each chromosome for humans as measured by Affymetrix oligonucleotide arrays and Unigene clusters.


Testis Enriched
Ovary Enriched
Affymetrix
Unigene
0.00
0.05
0.10
0.15
0.02
0.04
0.06
0.08
0.12
0.1
0.2
0.3
0.4
0.5
0.25
0.20
0.35
1 2 3 4 5 6 7 8 9 11 13 15 17 19 21 X