LETTERS

Conservation of Y-linked genes during human evolution revealed by comparative sequencing in chimpanzee

Jennifer F. Hughes[1], Helen Skaletsky[1], Tatyana Pyntikova[1], Patrick J. Minx[2], Tina Graves[2], Steve Rozen[1], Richard K. Wilson[2] & David C. Page[1]

The human Y chromosome, transmitted clonally through males, contains far fewer genes than the sexually recombining autosome from which it evolved. The enormity of this evolutionary decline has led to predictions that the Y chromosome will be completely bereft of functional genes within ten million years[1,2]. Although recent evidence of gene conversion within massive Y-linked palindromes runs counter to this hypothesis, most unique Y-linked genes are not situated in palindromes and have no gene conversion partners[3,4]. The 'impending demise' hypothesis thus rests on understanding the degree of conservation of these genes. Here we find, by systematically comparing the DNA sequences of unique, Y-linked genes in chimpanzee and human, which diverged about six million years ago, evidence that in the human lineage, all such genes were conserved through purifying selection. In the chimpanzee lineage, by contrast, several genes have sustained inactivating mutations. Gene decay in the chimpanzee lineage might be a consequence of positive selection focused elsewhere on the Y chromosome and driven by sperm competition.

The human X and Y chromosomes co-evolved from an ordinary pair of autosomes that existed in the mammalian ancestor roughly 300 million years ago[5]. Because most of the Y chromosome does not participate in sexual recombination, it has degenerated substantially, both in size and gene content, in comparison with the X chromosome[6]. Recent studies of the ampliconic region of the Y chromosome, which comprises almost half of the chromosome's euchromatin, revealed large palindromes where abundant gene conversion may forestall gene decay[3,4]. However, nearly all of the remainder of the Y chromosome's genes are found in the X-degenerate regions, which were once identical in sequence to the X chromosome but have since diverged substantially[3]. Unlike the ampliconic sequence, the X-degenerate sequence does not routinely undergo recombination of any sort, so rapid, ongoing gene loss might be expected there.

To understand better the recent evolution of the human X-degenerate sequence and the fate of its remaining genes, we determined the nucleotide sequence of the X-degenerate portion of the chimpanzee Y chromosome. The resulting sequence spans 9.5 megabases (Mb), is complete apart from two small gaps, and is accurate to about one nucleotide per 200,000.

Before using these sequence data to test the impending demise hypothesis, we compared basic characteristics of the X-degenerate sequences in chimpanzee and human, discovering that both the gross structures and nucleotide sequences of hominoid Y chromosomes have evolved rapidly. Counterpoint to the Y chromosome's rapid evolution is provided by human chromosome 21 and its orthologue, chimpanzee chromosome 22, the only autosomes fully sequenced in

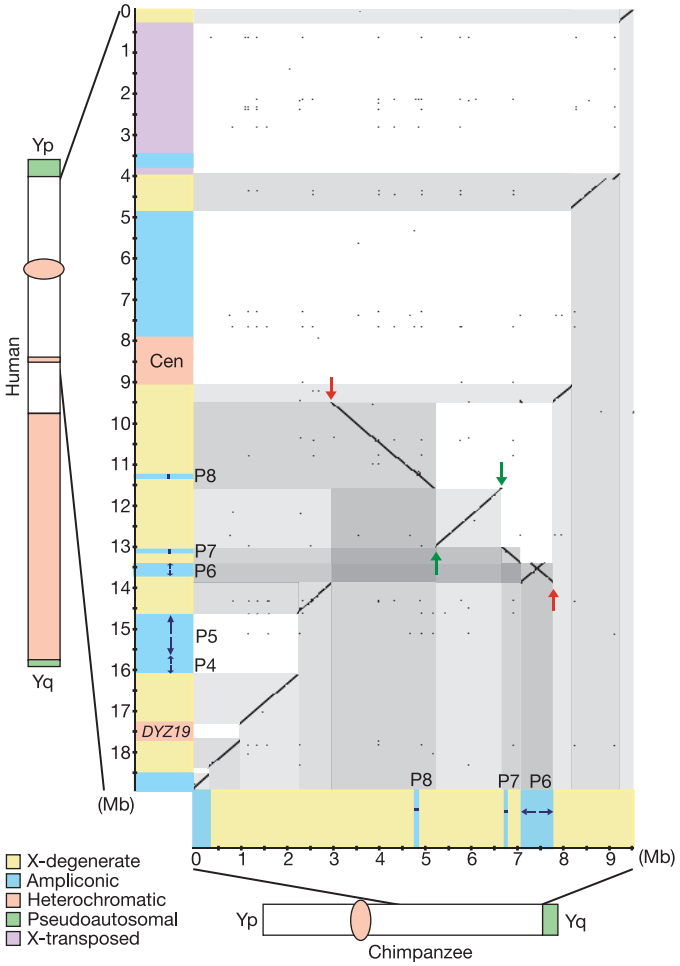

Figure 1 | Dot-plot comparison of the X-degenerate region of the chimpanzee Y chromosome (below) with the euchromatic region of the human Y chromosome (left). These chromosomal regions are shown schematically on the axes, where major features, including palindromes P4–P8, are indicated. Each dot within the plot represents 100% identity within a 200-bp window. Within the plot, grey shading indicates blocks of uninterrupted sequence alignment. Break points of two inversions (1.5 and 5.0 Mb) are indicated by arrows (green and red, respectively). Supplementary Fig. 1 provides a more highly annotated version of this plot, including gene and pseudogene positions. Cen, centromere.

[1]Howard Hughes Medical Institute, Whitehead Institute, and Department of Biology, Massachusetts Institute of Technology, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA. [2]Genome Sequencing Center, Washington University School of Medicine, 4444 Forest Park Boulevard, St Louis, Missouri 63108, USA.

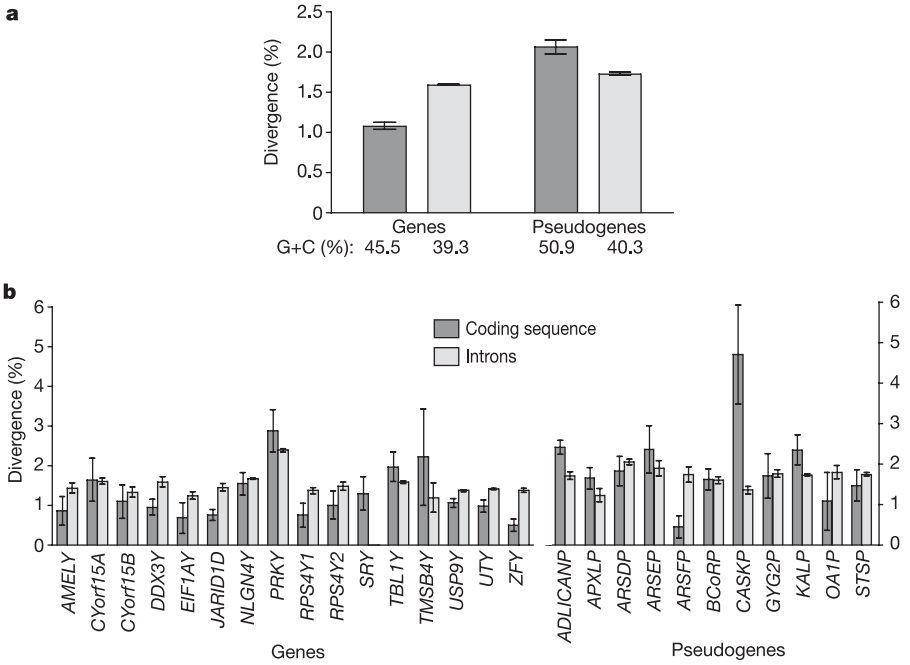

Figure 2 | Human–chimpanzee divergence in coding sequence and introns of X-degenerate genes and pseudogenes. a, Aggregate per cent divergences for coding and intron sequences of all genes and all pseudogenes. Listed below are G + C contents for each sequence class. See Supplementary Table 3 for numeric data and measures of statistical significance. **b**, Left: graph of coding and intron divergences for each gene, analysed separately and listed alphabetically. Right: an analogous graph of divergences for each pseudogene. Error bars depict standard errors for uncorrected per cent divergence calculated from 500 bootstrap replicates using MEGA2 software[24].

both species. Whereas the nucleotide sequences of human chromosome 21 and chimpanzee chromosome 22 are grossly co-linear[7], implying that little structural change has occurred, the Y chromosome has undergone significant restructuring since the chimpanzee and human lineages diverged (Fig. 1; see also Supplementary Fig. 1). In humans, the X-degenerate sequences are distributed along both arms of the Y chromosome, and they are interrupted at several points by large blocks of ampliconic, heterochromatic, or other sequences. In the chimpanzee, by contrast, the X-degenerate sequences are found in a single, nearly contiguous block on the long arm of the Y chromosome. In addition, the chimpanzee and human X-degenerate sequences differ by two large inversions (Supplementary Figs 2–4). The larger inversion, spanning nearly 5 Mb, occurred in the chimpanzee lineage. The smaller, 1.5-Mb inversion occurred in the human lineage. Such inversions may be of relatively little consequence in the male-specific region of the Y chromosome, which does not engage in crossing-over with a chromosomal homologue. (In heterozygotes for autosomal or X-linked inversions, crossing-over within the inverted segment results in genetically unbalanced offspring.) The nucleotide sequences of human chromosome 21 and chimpanzee chromosome 22 have been reported to display 1.44% divergence[7]. We observed significantly greater nucleotide divergence, 1.72%, between the X-degenerate sequences of the chimpanzee and human Y chromosomes. This difference is unsurprising given previous evidence of accelerated DNA sequence evolution on the Y chromosome, the result of its being transmitted exclusively through the substitution-prone male germ line[8].

The density of interspersed repetitive elements is nearly identical in the X-degenerate regions of the chimpanzee and human Y chromosomes—at least when one combines all such repeat classes (Supplementary Table 1). However, we found evidence of marked differences between the chimpanzee and human lineages in the levels of transposition activity of the three main classes of retroelements (Supplementary Table 2). We found that Alu elements were more active in the human lineage, whereas long interspersed nucleotide (L1) elements and especially endogenous retroviruses were more active in the chimpanzee lineage. Most notably, the chimpanzee sequence contains 21 copies of two novel endogenous retroviruses, CERV1 and CERV2, which are completely absent from the human genome.

Having completed these basic comparisons of the chimpanzee and human X-degenerate sequences, we then addressed a prediction of the impending demise model. The model's central premise is that, in recent times, the human Y chromosome has been losing genes once shared with the X chromosome at a pace approximating 5 genes per million years[1,2]. Assuming that Y-linked gene decay and loss occurred randomly, and that the chimpanzee and human lineages have been separate for about six million years, the chimpanzee Y chromosome should carry many genes that have no functional orthologue on the human Y chromosome.

To test this prediction, we characterized the gene content of the chimpanzee X-degenerate sequence by several means. First, we electronically searched the sequence for orthologues of all known human X-degenerate genes and pseudogenes. (The human pseudogenes, which do not seem to be transcribed, bear inactivating mutations that disrupt or delete splice sites and exons, or that interrupt or shift open reading frames[3].) We identified chimpanzee orthologues of all 16 such genes and all 11 such pseudogenes. Notably, the chimpanzee counterparts of the 11 human pseudogenes are also pseudogenes, with the great majority of inactivating mutations being shared between the two species, indicating that all 11 pseudogenes were inactivated before divergence of the chimpanzee and human lineages. This suggests that none of the 11 human X-degenerate pseudogenes has lost its functionality during the last six million years of human evolution. In addition, we conducted GenScan[9] and BLAST[10,11] searches of the chimpanzee X-degenerate sequence for transcription units that have no human Y counterpart. We found no such chimpanzee-specific transcription units. Thus, comparative cataloguing of X-degenerate genes and pseudogenes in the chimpanzee and human suggests that little or no X-degenerate gene loss or decay has occurred during the last six million years of human evolution. These findings contradict the model of the human Y chromosome's impending demise, and instead provide empirical support for mathematical models of sex chromosome evolution that predict a slowing of the rate of gene decay as Y chromosomes evolve[6].

These findings also suggest that purifying selection on the Y chromosome has been more effective during recent human evolution than previously supposed. To examine this hypothesis, we compared the degree of human–chimpanzee divergence in the X-degenerate genes' coding regions with the degree of divergence in their introns, which served as controls. As additional controls, we examined interspecies divergence in the X-degenerate pseudogenes. Genes for which the protein products are subject to purifying selection should exhibit less interspecies divergence in coding sequences than in introns. We found this to be the case for the X-degenerate genes as

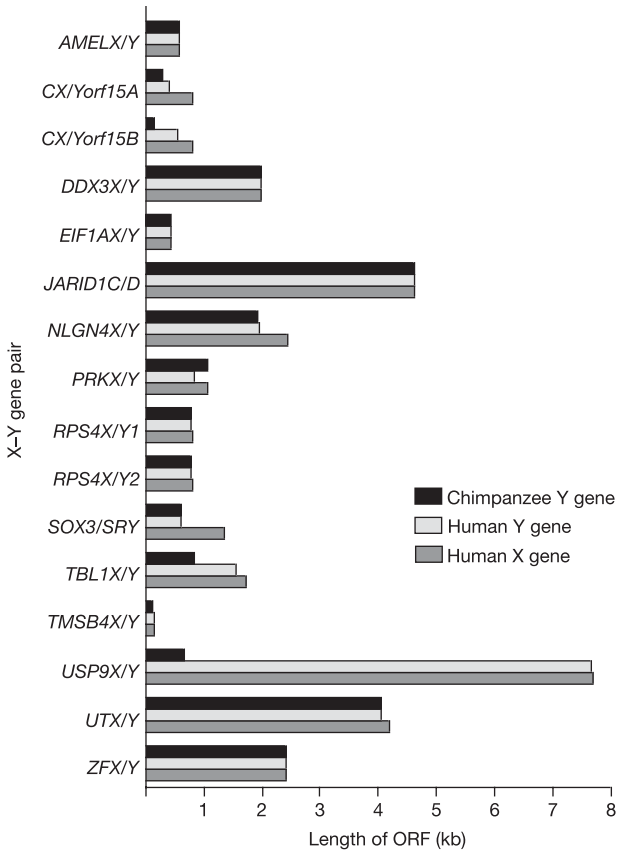

Figure 3 | Lengths of coding sequences of X-degenerate genes on chimpanzee and human Y chromosomes, and their human X-linked homologues. For each X–Y gene pair, three horizontal bars are shown, representing the predicted lengths (in kb) of the coding regions for the chimpanzee Y gene (black), its human Y-linked orthologue (light grey) and its human X-linked homologue (dark grey). Genes are listed in alphabetical order.

a group ($P < 0.00001$; Fig. 2a; see also Supplementary Table 3). As expected, this was not true for the X-degenerate pseudogenes, where human–chimpanzee divergence proved to be similar or even greater in (former) coding sequences than in (former) introns. (The elevated G + C content of these former coding sequences probably accounts for the elevated rate of sequence evolution[12].) When every gene was analysed independently, most exhibited the same trend towards greater conservation of coding sequence relative to introns (Fig. 2b). We conclude that purifying selection has been a potent force in maintaining X-degenerate gene function during recent human evolution.

Whereas the repertoire of X-degenerate genes present in the human/chimpanzee ancestor evidently remained functionally intact during subsequent human evolution, we discovered evidence of significant gene decay during chimpanzee evolution. We used three types of analysis: coding sequence divergence (Fig. 2), open reading frame (ORF) integrity (Fig. 3), and transcriptional activity (as assayed by polymerase chain reaction with reverse transcription (RT–PCR); Supplementary Fig. 6). Interspecies divergence of coding sequence differed substantially among the X-degenerate genes (Fig. 2), prompting us to rank all of the genes by this measure (Table 1). We reasoned that genes displaying relatively high interspecies divergence might have been subject to relaxed selective constraints, and possibly functional decay, in the chimpanzee or human lineage, or both. Of the eight X-degenerate genes displaying >1.0% divergence, five had ORFs that were substantially truncated in the chimpanzee Y chromosome as compared with the human Y chromosome (Fig. 3). These ORF truncations were due to point mutations that disrupted splice sites or introduced stop codons in the chimpanzee genes, but did not grossly alter their genomic size or structure (Table 1).

Intriguingly, the truncated ORFs in chimpanzee include orthologues of both the largest and smallest proteins predicted to be encoded by the human Y chromosome[3]. In human, *USP9Y* has a critical role in spermatogenesis[13] and is predicted to encode a 2,555-amino-acid ubiquitin protease that is 91% identical across its length to a 2,563-amino-acid protein encoded by the X chromosome. In chimpanzee, by contrast, the longest ORF in *USP9Y* would encode a protein of only 675 amino acids, with no intact catalytic domain (Supplementary Fig. 7). In human, *TMSB4Y* is expressed throughout the body and is predicted to encode a 44-amino-acid peptide that differs by three amino substitutions from an otherwise identical peptide encoded by its mammalian X-linked homologue. In chimpanzee, we detected no evidence of *TMSB4Y* transcription in any tested tissue or cell line (Supplementary Fig. 6), and the single splice donor site within the coding region has been lost through mutation (Supplementary Fig. 8). These examples illustrate the diversity of Y-linked, X-degenerate gene functions that have decayed in chimpanzee but not human.

Why have X-degenerate genes decayed in the chimpanzee lineage but not in the human lineage? We speculate that X-degenerate gene decay in the chimpanzee lineage may be a by-product of strong positive selection focused elsewhere on the Y chromosome, through a process known as genetic hitchhiking. Because the Y chromosome does not participate in sexual recombination with a chromosome

Table 1 | Chimpanzee gene characterization

Gene	Nucleotide divergence from human (%)	ORF length (% of human)	ORF-disrupting mutations in chimpanzee*
PRKY	2.89	128†	–
TMSB4Y	2.22	42	Splice donor (GT → GC)
TBL1Y	1.98	52	Splice acceptor (AG → GG); splice donor (GT → AT)
CYorf15A	1.65	72	−1 frame shift
NLGN4Y	1.55	99	–
SRY	1.31‡	100	–
CYorf15B	1.10	26	Nonsense (AAA → TAA)
USP9Y	1.07	9	Three splice donors (4-bp deletion; GT → AT; GT → AT)
RSP4Y2	1.01	100	–
UTY	0.99	100	–
DDX3Y	0.96	100	–
AMELY	0.86	100	–
JARID1D	0.76	100	–
RPS4Y1	0.76	100	–
EIF1AY	0.69	100	–
ZFY	0.50	100	–

Genes are ordered according to per cent coding sequence divergence from their human orthologues, from highest to lowest.
*We inferred the lineage in which each mutation occurred by comparison to the human X homologue.
†The ORF of *PRKY* is longer in the chimpanzee because of a 55-bp genomic deletion, near the 3′ end of the gene, that occurred in the human lineage.
‡The sex-determining gene *SRY* is subject to positive selection[23], which probably accounts for its relatively high interspecies divergence.

homologue, natural selection acts on the chromosome as a unit. Deleterious mutations in some Y-linked genes can be carried along, even to the point of fixation in a population, by physical linkage to strongly beneficial mutations in other Y-linked genes[6,14]. In addition to their X-degenerate genes, primate Y chromosomes contain many families of ampliconic genes, which have testes-restricted expression patterns and critical functions in sperm production[3,15]. Because of this central role in spermatogenesis, the Y chromosome's ampliconic genes may be subject to powerful selective pressures[16–18], especially in species such as chimpanzees where females usually mate with multiple males, the sperm of which then compete for a limited number of oocytes[19]. During chimpanzee evolution some X-degenerate genes may have been casualties of selective forces directed at the Y chromosome's ampliconic genes—forces that were not as intense during the evolution of our less promiscuous species. In the future, comparisons of sex chromosome variability[20] in chimpanzees and humans may provide a test of this speculative hypothesis.

METHODS

Mapping and sequencing. We mapped and sequenced a tiling path of 73 bacterial artificial chromosome (BAC) and 7 fosmid clones. Clones for sequencing were selected from two BAC libraries (CHORI-251, RPCI-43) and one fosmid library (CHORI-1251) (http://bacpac.chori.org). The CHORI-251 BAC and CHORI-1251 fosmid libraries originate from the same male chimpanzee. Only eight of the BAC clones in the tiling path are from the RPCI-43 library, which originates from a second male chimpanzee. We screened the BAC libraries with 11 pools of hybridization probes derived from 209 STS markers located within the X-degenerate region of the human Y chromosome[3,21]. Additional probes and markers were obtained from chimpanzee BAC end sequences. Fosmid end sequences were used to identify appropriate clones for filling gaps in the BAC contigs. See Supplementary Fig. 5 for the complete clone contig map. The accuracy of the sequence was estimated using all available CHORI-251 BAC overlaps in the assembled tiling path. There were a total of 26 errors in over 5.31 Mb of aligned sequence, which correlates to 1 error per 204 kilobases (kb). Two gaps remain in the sequence and their sizes were estimated based on human sequence to be roughly 14 kb and 69 kb. These sizes were confirmed by fibre fluorescence *in situ* hybridization (FISH) analysis using a cell line derived from the same chimpanzee used in constructing the CHORI-251 and CHORI-1251 libraries (Supplementary Fig. 9).

Assessing the completeness of chimpanzee X-degenerate sequence coverage. We used the data set based on the 13 November 2003 assembly from the Chimpanzee Sequencing Consortium (http://www.ensembl.org/Pan_troglodytes) to search for the existence of chimpanzee X-degenerate sequence that is absent from the human sequence using two strategies. First, we used all chimpanzee unassigned contigs that were not identified as interspersed repeats (10,750 contigs) in a BLAST search of the human genome, presuming that those contigs that were a closest match to human X chromosome sequence but displayed less that 97% identity were candidate X-degenerate contigs. However, none was found, and instead all chimpanzee contigs that matched the human X sequence were 99% identical and are presumably from the X chromosome. Second, we used all known and predicted human X chromosome genes in a BLAST search of the chimpanzee unassigned contigs in an attempt to identify chimpanzee X-degenerate genes or pseudogenes that are not on the human Y chromosome. No significant matches were found.

Sequence alignment and dot-plot analysis. Chimpanzee and human sequences were aligned using Clustal W with default parameters[22]. Dot-plot analysis was performed using custom Perl code, which is available upon request.

Assigning insertions and deletions to the chimpanzee or human lineage. Insertions and deletions (indels) were identified by aligning the chimpanzee and human sequences. For indels that were at least 100 base pairs (bp) in length, we determined the nature of the mutation (insertion or deletion) and the lineage in which it occurred as follows. If the indel sequence consisted entirely of a known, full-length interspersed repeat, such as an Alu or L1, it was inferred to be the result of an insertion event. Because integrations of partial L1 elements are known to occur, if these were identified the corresponding sequence in the other species was examined to look for the presence of the same L1 element. This was done to avoid misclassifying partial L1 deletions as integrations of non-full-length elements. Indel sequences not classified as repetitive element insertions or tandem duplications were presumed to be the result of deletion events. Putative deleted sequences that were not composed entirely of interspersed repeats were used in a BLAST search of the non-redundant GenBank database to ensure that they were not transposed sequences, which would be evidenced by close matches to an autosome. For the majority of these sequences (47 of 61), the second best match (after Y chromosome sequence) was the human X chromosome. This is expected because the X chromosome represents the ancestral state of the X-degenerate sequences. The remaining indels matched only chimpanzee or human Y chromosome sequence. These results were interpreted as corroborating a deletion event.

RT–PCR analysis. The RNeasy kit (Qiagen) was used to isolate total RNAs from chimpanzee male tissues (testis, liver, lung and spleen) and a chimpanzee male lymphoblastoid cell line; all tissues were obtained from Yerkes National Primate Research Center. RT–PCR primer sequences and product sizes are listed in Supplementary Table 4.

Received 15 April; accepted 3 August 2005.

1. Aitken, R. J. & Marshall Graves, J. A. The future of sex. *Nature* **415,** 963 (2002).
2. Graves, J. A. The degenerate Y chromosome—can conversion save it? *Reprod. Fertil. Dev.* **16,** 527–534 (2004).
3. Skaletsky, H. *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423,** 825–837 (2003).
4. Rozen, S. *et al.* Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* **423,** 873–876 (2003).
5. Lahn, B. T. & Page, D. C. Four evolutionary strata on the human X chromosome. *Science* **286,** 964–967 (1999).
6. Charlesworth, B. & Charlesworth, D. The degeneration of Y chromosomes. *Phil. Trans. R. Soc. Lond. B* **355,** 1563–1572 (2000).
7. Watanabe, H. *et al.* DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature* **429,** 382–388 (2004).
8. Miyata, T., Hayashida, H., Kuma, K., Mitsuyasu, K. & Yasunaga, T. Male-driven molecular evolution: a model and nucleotide sequence analysis. *Cold Spring Harb. Symp. Quant. Biol.* **52,** 863–867 (1987).
9. Yeh, R. F., Lim, L. P. & Burge, C. B. Computational inference of homologous gene structures in the human genome. *Genome Res.* **11,** 803–816 (2001).
10. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
11. Hellmann, I. *et al.* Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Res.* **13,** 831–837 (2003).
12. Casane, D., Boissinot, S., Chang, B. H., Shimmin, L. C. & Li, W. Mutation pattern variation among regions of the primate genome. *J. Mol. Evol.* **45,** 216–226 (1997).
13. Sun, C. *et al.* An azoospermic man with a *de novo* point mutation in the Y-chromosomal gene *USP9Y*. *Nature Genet.* **23,** 429–432 (1999).
14. Rice, W. R. Genetic hitchhiking and the evolution of reduced genetic activity of the Y sex chromosome. *Genetics* **116,** 161–167 (1987).
15. Lahn, B. T. & Page, D. C. Functional coherence of the human Y chromosome. *Science* **278,** 675–680 (1997).
16. Parker, G. A. Sperm competition and its evolutionary consequences in the insects. *Biol. Rev.* **45,** 525–567 (1970).
17. Dorus, S., Evans, P. D., Wyckoff, G. J., Choi, S. S. & Lahn, B. T. Rate of molecular evolution of the seminal protein gene *SEMG2* correlates with levels of female promiscuity. *Nature Genet.* **36,** 1326–1329 (2004).
18. Wyckoff, G. J., Wang, W. & Wu, C. I. Rapid evolution of male reproductive genes in the descent of man. *Nature* **403,** 304–309 (2000).
19. Dixson, A. F. *Primate Sexuality: Comparative Studies of the Prosimians, Monkeys, Apes and Human Beings* (Univ. Chicago Press, Chicago, 1998).
20. Filatov, D. A., Moneger, F., Negrutiu, I. & Charlesworth, D. Low variability in a Y-linked plant gene and its implications for Y-chromosome evolution. *Nature* **404,** 388–390 (2000).
21. Tilford, C. A. *et al.* A physical map of the human Y chromosome. *Nature* **409,** 943–945 (2001).
22. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680 (1994).
23. Whitfield, L. S., Lovell-Badge, R. & Goodfellow, P. N. Rapid sequence evolution of the mammalian sex-determining gene *SRY*. *Nature* **364,** 713–715 (1993).
24. Kumar, S., Tamura, K., Jakobsen, I. B. & Nei, M. MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* **17,** 1244–1245 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by the National Institutes of Health and the Howard Hughes Medical Institute.

Author Information GenBank accession numbers for CERV1 and CERV2 are AY692036 and AY692037, respectively. GenBank accession numbers for all complementary DNA sequences are listed in Supplementary Table 5; accession numbers for all BAC and fosmid clones are listed in Supplementary Table 6. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.C.P. (page_admin@wi.mit.edu).