

The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes

Helen Skaletsky*, Tomoko Kuroda-Kawaguchi*, Patrick J. Minx†, Holland S. Cordum†, LaDeana Hillier†, Laura G. Brown*, Sjoerd Repping‡, Tatyana Pyntikova*, Johar Ali†, Tamberlyn Bieri†, Asif Chinwalla†, Andrew Delehaunty†, Kim Delehaunty†, Hui Du†, Ginger Fewell†, Lucinda Fulton†, Robert Fulton†, Tina Graves†, Shun-Fang Hou†, Philip Latrielle†, Shawn Leonard†, Elaine Mardis†, Rachel Maupin†, John McPherson†, Tracie Miner†, William Nash†, Christine Nguyen†, Philip Ozersky†, Kymberlie Pepin†, Susan Rock†, Tracy Rohlfing†, Kelsi Scott†, Brian Schultz†, Cindy Strong†, Aye Tin-Wollam†, Shiaw-Pyng Yang†, Robert H. Waterston†, Richard K. Wilson†, Steve Rozen* & David C. Page*

<http://jura.wi.mit.edu/rozen>

* Howard Hughes Medical Institute, Whitehead Institute, and Department of Biology, Massachusetts Institute of Technology, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA

† Genome Sequencing Center, Washington University School of Medicine, 4444 Forest Park Boulevard, St Louis, Missouri 63108, USA

‡ Center for Reproductive Medicine, Department of Gynaecology and Obstetrics, Academic Medical Centre, Amsterdam 1105 AZ, the Netherlands

The male-specific region of the Y chromosome, the MSY, differentiates the sexes and comprises 95% of the chromosome's length. Here, we report that the MSY is a mosaic of heterochromatic sequences and three classes of euchromatic sequences: X-transposed, X-degenerate and ampliconic. These classes contain all 156 known transcription units, which include 78 protein-coding genes that collectively encode 27 distinct proteins. The X-transposed sequences exhibit 99% identity to the X chromosome. The X-degenerate sequences are remnants of ancient autosomes from which the modern X and Y chromosomes evolved. The ampliconic class includes large regions (about 30% of the MSY euchromatin) where sequence pairs show greater than 99.9% identity, which is maintained by frequent gene conversion (non-reciprocal transfer). The most prominent features here are eight massive palindromes, at least six of which contain testis genes.

The history of human Y chromosome research can be divided into three eras. The first era focused on mendelian examination of human family trees. In the opening decades of the twentieth century, proponents of Mendel's concept of the gene observed three modes of inheritance in our species: autosomal recessive, autosomal dominant and X-linked recessive. Contemporaneously, other scholars sought to identify traits that exhibited Y-linked (father to son) transmission. These scholars erroneously claimed success, presenting family trees purported to demonstrate that Y-chromosomal genes were responsible for hairy ears, scaly skin and other traits. Meanwhile, light microscopic studies of human cells provided strong physical evidence of the existence of a male-specific chromosome¹. By 1950, studies of human pedigrees reported at least 17 Y-linked traits².

The second era was dominated by the view that the Y chromosome was a genetic wasteland, based on the debunking of earlier studies and a dearth of new evidence for genes. In the 1950s, Stern systematically exposed critical flaws in each of the preceding pedigree studies and dismissed them². In 1959, Jacobs' study of Klinefelter (XXY) males³ and Ford's research on Turner (XO) females⁴ demonstrated that the Y chromosome carries a pivotal sex-determining gene, but this gene was considered to be an exception on a generally desolate chromosome. In the 1960s, Ohno proposed that the mammalian X and Y chromosomes had evolved from an ordinary pair of autosomes⁵. Ohno speculated that the X chromosome had retained the ancestral autosome's gene content whereas the Y chromosome had lost all but perhaps one gene involved in sex determination. Thus emerged the understanding of the human Y chromosome as a profoundly degenerate X chromosome.

The hallmark of the third and present era has been the application of recombinant DNA and genomic technologies to the Y chromo-

some, culminating in molecularly based conclusions about its genes. In recent decades, an understanding of the Y chromosome's biological functions has begun to emerge from DNA studies of individuals with partial Y chromosomes, coupled with molecular characterization of Y-linked genes implicated in gonadal sex reversal, Turner syndrome, graft rejection and spermatogenic failure⁶. Genomic studies revealed that the Y chromosome contains a region, comprising 95% of its length, where there is no X–Y crossing over. This region came to be known as the non-recombining region, or NRY, although our discovery of abundant recombination, as reported here and in the accompanying manuscript, compels us to rename it the male-specific region, or MSY⁷. The MSY is flanked on both sides by pseudoautosomal regions, where X–Y crossing over is a normal and frequent event in male meiosis (see Supplementary Note 1).

Previous efforts to construct accurate, high-resolution physical maps of the MSY had been stymied by an abundance of lengthy, intrachromosomal repetitive sequences, or amplicons⁸. To overcome this difficulty, we identified minute variations between amplicon copies, and then highlighted these minute variants (sequence family variants⁹) as markers to be ordered with respect to one another, yielding a map amenable to iterative refinement. However, the minute variants could only be found by fully and accurately sequencing and comparing near-identical amplicon copies. Thus, in our effort to determine the nucleotide sequence of the MSY, mapping and sequencing activities were fused into a single, iterative analytic process. We have previously reported the physical map that emerged from these efforts¹⁰. Here we report the sequencing of the MSY.

We mapped and sequenced a tiling path of 220 bacterial artificial chromosome (BAC) clones, each containing a portion of the MSY from the same individual. We used only one man's Y chromosome

to prevent any allelic variation, or polymorphism, from confounding our search for minute sequence variation between amplicon copies. (MSY amplicon copies can differ as little in sequence as two Y chromosomes chosen at random from the population¹¹.) We chose to sequence highly redundant BACs, especially in amplicon-rich regions: about 12.7 million (roughly 60%) of the euchromatic nucleotides were sequenced in at least two independent BAC clones. This redundancy allowed us to refine and validate the MSY sequence by exhaustively investigating, and in most cases resolving, sequence discrepancies between overlapping BACs.

Sequencing of euchromatic and heterochromatic regions

We begin with a statistical synopsis of the MSY sequence, considering the euchromatic and heterochromatic portions separately. (In this analysis, we have equated satellite sequences with heterochromatin, and all other sequences with euchromatin.) The product of our present research is a ‘reference’ sequence from one man’s Y chromosome. A full description of the nature and extent of

Y chromosome variation in human populations must await future studies. We and our colleagues have previously reported the nucleotide sequence of two portions of the MSY (the *AZFa* and *AZFc* regions^{12,13}). We have incorporated this previously reported sequence data in our present analysis of the entire MSY.

The MSY’s euchromatic DNA sequences total roughly 23 megabases (Mb), including 8 Mb on the short arm (Yp) and 14.5 Mb on the long arm (Yq) (Fig. 1). We obtained finished sequence, with an estimated error rate of about 1 per 10⁵ nucleotides, for all MSY euchromatin, with two known exceptions. First, there remain two gaps, each of which is roughly 50 kilobases (kb) long as judged by chromosomal fluorescence *in situ* hybridization (FISH) (Supplementary Fig. 1). Second, we obtained representative but incomplete sequence for a tandem array that spans roughly 0.7 Mb on Yp. We estimate that we obtained finished nucleotide sequence for roughly 97% of the MSY euchromatin, and that we captured 99% of the sequence complexity of MSY’s euchromatin.

So far, efforts to gain sequence-based understanding of human chromosomes have largely by-passed heterochromatic regions (refs 14, 15; see also Supplementary Note 2), including a large block of heterochromatic sequences found in the centromeric region of every nuclear chromosome¹⁶. In addition to its centromeric heterochromatin (approximately 1 Mb, ref. 17), the Y chromosome was previously shown to contain a second, much longer heterochromatic block (roughly 40 Mb) that comprises the bulk of the distal long arm (Fig. 1; see also Supplementary Note 3). In the course of the present sequencing project, we discovered and characterized a third heterochromatic block—a sharply demarcated island that spans approximately 400 kb, comprises >3,000 tandem repeats of 125 base pairs (bp), and interrupts the euchromatic sequences of proximal Yq (Figs 1 and 2). The other two heterochromatic blocks also consist of massively amplified tandem repeats of low sequence complexity. We attempted to sequence BACs spanning the boundaries and representing the body of each of the three heterochromatic blocks. We succeeded, with the exception that the distal boundary of the major heterochromatic region, on distal Yq, was not identified with certainty (Supplementary Fig. 1). In total, we found that the heterochromatin of MSY encompasses at least six distinct sequence species (Table 1), each of which form long, homogeneous tandem arrays. Our findings are detailed in Supplementary Note 4 and Supplementary Fig. 3.

A catalogue of genes and transcription units

With a comprehensive reference sequence of the MSY in hand, we set out to catalogue systematically the genes of the MSY. We electronically identified and manually examined all matches to previously reported MSY genes. Furthermore, we used polymerase chain reaction with reverse transcription (RT–PCR) and/or sequencing of complementary DNA clones to evaluate electronic matches to publicly available expressed sequence tags (ESTs), as well as potential genes that were predicted using GenomeScan software¹⁸. For all experimentally verified genes whose expression patterns had not been reported previously, we tested for expression in diverse human tissues by RT–PCR and subsequent sequencing of RT–PCR products.

We found that the MSY includes at least 156 transcription units, half of which probably encode proteins (Table 2 and Figs 2, 3; see also Supplementary Tables 1 and 2). All 156 transcription units identified are located in euchromatic sequences. We have no evidence of transcription of the MSY heterochromatin. Of the approximately 78 protein-coding units, about 60 are members of nine different MSY-specific gene families, each characterized by >98% nucleotide identity among family members, in both exons and introns. The remaining 18 protein-coding genes are present in one copy each in the MSY. (These include two genes, *RPS4Y1* and *RPS4Y2*, that exhibit 93.6% nucleotide identity in coding exons but are much more diverged in introns.) Thus, the MSY seems to

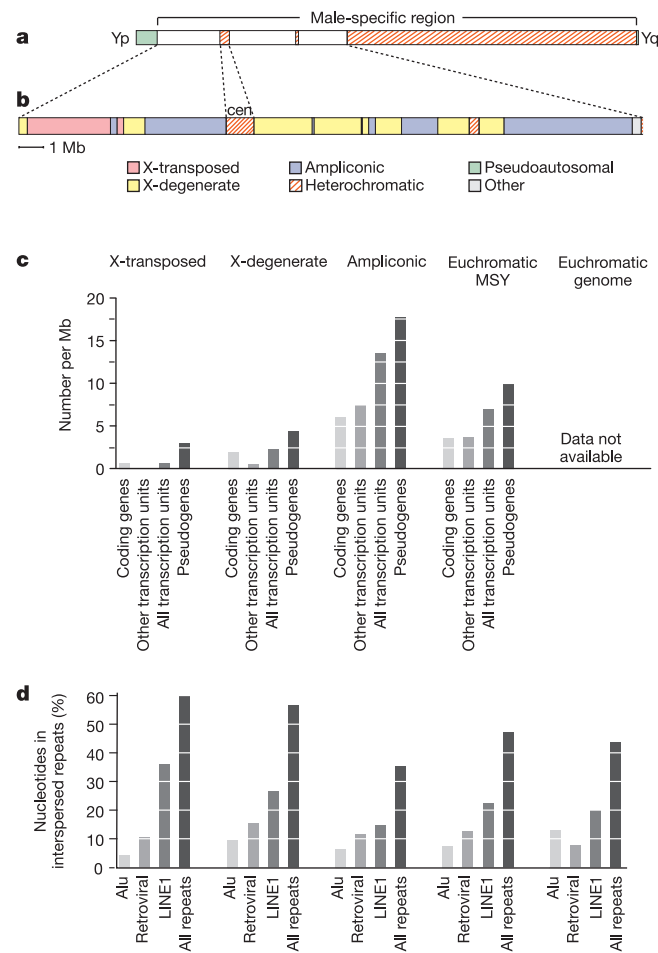


Figure 1 The male-specific region of the Y chromosome. **a**, Schematic representation of the whole chromosome, including the pseudoautosomal and heterochromatic regions. **b**, Enlarged view of a 24-Mb portion of the MSY, extending from the proximal boundary of the Yp pseudoautosomal region to the proximal boundary of the large heterochromatic region of Yq. Shown are three classes of euchromatic sequences, as well as heterochromatic sequences. A 1-Mb bar indicates the scale of the diagram. **c**, **d**, Gene, pseudogene and interspersed repeat content of three euchromatic sequence classes. **c**, Densities (numbers per Mb) of coding genes, non-coding transcription units, total transcription units and pseudogenes. **d**, Percentages of nucleotides contained in Alu, retroviral, LINE1 and total interspersed repeats. The data shown in **c** and **d** are available in numerical form in Supplementary Tables 6 and 7. Supplementary Table 6 also provides information about the size and (G + C) content of each sequence class.

encode at least 27 distinct proteins or protein families.

Furthermore, the MSY includes at least 78 transcription units for which strong evidence of protein coding is lacking; many of these transcription units are probably non-coding. Of these 78 transcription units, 13 occur in single copy in the MSY and the remaining 65 are members of 15 MSY-specific families. Considering together both coding and non-coding transcription units, the MSY appears to contain 24 MSY-specific families, which collectively account for 125 of the 156 MSY transcription units identified so far.

On the basis of earlier experiments, most of the genes of the MSY were thought to fall into two functional classes, with genes in the first group expressed throughout the body, in many organs, and genes in the second group expressed predominantly or exclusively in testes¹⁹. Our present catalogue of MSY genes and their patterns of tissue expression (Table 2) corroborate this model. Of the MSY's 27 distinct protein-coding genes or gene families identified so far, 12 are expressed ubiquitously and 11 are expressed exclusively or predominantly in testes.

Three classes of sequences in the MSY euchromatin

We find that nearly all of the euchromatic sequences fall into three classes, which we have named X-transposed, X-degenerate and ampliconic. As shown in Figs 1 and 2, the MSY euchromatin is a patchwork of these three sequence classes. The characteristics of the classes are summarized in Fig. 4.

The X-transposed sequences are 99% identical to DNA sequences in Xq21, a band in the midst of the long arm of the human X chromosome. The X-transposed sequences are so named because their presence in the human MSY is the result of a massive X-to-Y transposition that occurred about 3–4 million years ago, after the divergence of the human and chimpanzee lineages^{20–22}. Subsequently, an inversion within the MSY short arm cleaved the X-transposed block into two non-contiguous segments, as observed in the modern MSY (Figs 1 and 2)^{21,22}. The X-transposed sequences do not participate in X–Y crossing over during male meiosis, distinguishing them from the pseudoautosomal sequences found in the telomeric regions of the human X and Y chromosomes.

Within the X-transposed segments, which have a combined length of 3.4 Mb, we identified only two genes, both of which have homologues in Xq21 (Table 2). Thus the X-transposed sequences exhibit the lowest density of genes among the three sequence classes in the MSY euchromatin (Figs 1 and 3), as well as the highest density of interspersed repeat elements (Fig. 1). In particular, long interspersed nuclear element 1 (LINE1) elements account for 36% of all X-transposed sequence, or nearly twice the genome average of 20%^{14,15}. As expected, low gene density and high repeat density also characterize the homologous sequence block in Xq21.

In contrast to the X-transposed sequence blocks, the X-degenerate segments of the MSY are dotted with single-copy gene or pseudogene homologues of 27 different X-linked genes. These single-copy MSY genes and pseudogenes display between 60%

and 96% nucleotide sequence identity to their X-linked homologues, and they seem to be surviving relics of ancient autosomes from which the X and Y chromosomes co-evolved, as explained below. In 13 cases, the MSY homologue is a pseudogene with sequence similarity to exons and introns of the functional X homologue (Supplementary Table 3). In the remaining 14 cases, the MSY homologue seems to be a transcribed, functional gene, and the X- and Y-linked genes encode very similar but non-identical protein isoforms (Table 2 and Figs 2, 3). These include two cases in which a functional X-linked gene has two expressed homologues in the MSY. The Y-linked genes *RPS4Y1* and *RPS4Y2* are full-length homologues of the X-linked gene *RPS4X*, and they apparently encode two different, full-length isoforms of ribosomal protein S4. In contrast, the Y-linked genes *CYorf15A* and *CYorf15B* are homologous to, respectively, 5' and 3' portions of the X-linked gene *CXorf15*, and they apparently encode proteins homologous to, respectively, amino- and carboxy-terminal portions of the predicted CXORF15 protein (Supplementary Fig. 4). Together, the X-degenerate sequences encode 16 of the MSY's 27 distinct proteins or protein families.

Notably, all 12 ubiquitously expressed MSY genes reside in the X-degenerate regions; no such genes have been identified elsewhere in the MSY. Conversely, among the 11 MSY genes found to be expressed predominantly in testes, only one gene, the sex-determining *SRY*, is X-degenerate.

The third class of euchromatic sequences, the ampliconic segments, are composed largely of sequences that exhibit marked similarity—as much as 99.9% identity over tens or hundreds of kilobases—to other sequences in the MSY. We refer to these long, MSY-specific repeat units, of which there are many families, as amplicons. The amplicons are located in seven segments that are scattered across the euchromatic long arm and proximal short arm (Figs 1 and 2), and whose combined length is 10.2 Mb.

We identified these ampliconic regions through a comprehensive analysis of similarities within the sequenced portions of the MSY. We calculated the percentage nucleotide identity between all pairs of known MSY sequences and then plotted the data in two ways. First we determined, at each point along the length of the sequenced MSY, the highest intrachromosomal similarity. The resulting graph (Fig. 5c) identifies the ampliconic regions as those where intrachromosomal identity, over stretches of 50 kb or more, generally exceeds 50%. Notably, 60% (6.1 Mb) of the ampliconic sequences exhibit intrachromosomal identities of 99.9% or greater.

A more spatially detailed representation of intrachromosomal similarities is shown in Fig. 5a, which records the locations of all MSY sequence pairs characterized by at least 65% identity within a sliding window of 2,000 nucleotides. After heterochromatic and LINE1 repeats have been accounted for, the MSY is seen to contain many long stretches of sequence that are similar to those elsewhere in the MSY. As shown in the inset to Fig. 5a, the triangular plot can be broken down into two smaller triangles—one representing sequence comparisons within Yp, the other depicting comparisons

Table 1 Six sequence species in three MSY heterochromatic regions

Region	Locus symbol	Unit repeat	
		Primary	Secondary
Centromeric region	<i>DYZ3</i> <i>DYZ17*</i>	171 bp alphoid GGAAT	5,941 bp None observed
Proximal Yq (Yq11.22)	<i>DYZ19*</i>	125 bp	None observed
Distal Yq (Yq12)	<i>DYZ18*</i> <i>DYZ1</i> <i>DYZ2</i>	GGAAT GGAAT Insufficient sequence available; (A + T) rich	2,864 bp 3,584 bp ~2,470 bp†

See Supplementary Table 10 for a more detailed version of this Table, incorporating references and the sequences of repeat units.

*First reported in this study.

†On the basis of restriction endonuclease digestion, not sequence analysis.

within Yq—and a rectangle depicting comparisons between Yp and Yq. Scrutiny of these Yp, Yq and Yp–Yq components of the plot reveals a wealth of sequence similarities within and between ampliconic segments on both arms of the chromosome.

The ampliconic sequences exhibit by far the highest density of genes, both coding and non-coding, among the three sequence classes in the MSY euchromatin (Figs 1 and 3). We identified nine distinct MSY-specific protein-coding gene families, with copy numbers ranging from two (*VCY*, *XKRY*, *HSFY*, *PRY*) to three (*BPY2*) to four (*CDY*, *DAZ*) to six (*RBMY*) to approximately 35 (*TSPY*) (Table 2 and Figs 2, 3). (These copy numbers pertain to the particular Y chromosome that we sequenced; they may vary in human populations.) In aggregate, these nine coding families encompass roughly 60 transcription units. Furthermore, the ampliconic sequences include at least 75 other transcription units for which strong evidence of protein coding is lacking (Figs 2 and 3; see also Supplementary Table 2). Of these 75 putative non-coding transcription units, 65 are members of 15 MSY-specific families, and the remaining 10 occur in single copy. Considering together both coding and non-coding elements, the ampliconic sequences contain 135 of the 156 MSY transcription units identified so far.

In contrast to the ubiquitous expression of most X-degenerate genes, the ampliconic genes and transcription units show highly restricted expression (Table 2). All nine protein-coding families in the ampliconic regions are expressed predominantly or exclusively in testes, as are most of the regions' non-coding transcription units.

Among the three euchromatic sequence classes, the ampliconic sequences exhibit by far the lowest densities of LINE1 and total interspersed repeat elements (Fig. 1). Indeed, the interspersed repeat content of the MSY's ampliconic sequences (35%) is far below the mean for the human genome (44%; *z*-test yields $P \ll 0.000001$).

Eight palindromes comprising 25% of MSY euchromatin

The most pronounced structural features of the ampliconic regions of Yq are eight massive palindromes (Table 3). In the dot plot of Fig. 5a, the longer palindromes are visible as vertical blue lines that approach the baseline. An MSY map highlighting all eight palindromes is shown in Fig. 3a. In all eight palindromes, the arms are highly symmetrical, with arm-to-arm nucleotide identities of 99.94–99.997%. (By convention, these percentage identities refer only to nucleotide substitutions and do not take account of insertions and deletions by which palindrome arms differ.) The palindromes are long, their arms ranging from 9 kb to 1.45 Mb in length. They are imperfect in that each contains a unique, non-duplicated spacer, 2–170 kb in length, at its centre. Palindrome P1 is particularly spectacular, having a span of 2.9 Mb, an arm-to-arm identity of 99.97%, and bearing two secondary palindromes (P1.1 and P1.2, each with a span of 24 kb) within its arms¹³. The eight palindromes collectively comprise 5.7 Mb, or one-quarter of the MSY euchromatin.

Six of the eight palindromes carry recognized protein-coding genes, all of which seem to be expressed specifically in testes (Fig. 3b). In all known cases of genes on MSY palindromes, identical or nearly identical gene copies exist on opposite arms of the palindrome. Of the nine multi-copy, protein-coding gene families identified so far in the MSY, eight have members on palindromes. Indeed, six families are located exclusively in palindromes. These include the *DAZ* genes, which exist in four copies—two in palindrome P1 and two in P2—and the *CDY* genes, which also occur in four copies—two in P1 and two in P5 (Fig. 3b). In addition, the palindromes contain at least seven families of apparently non-coding transcription units, all expressed exclusively or predominantly in testes (Fig. 3e).

In addition to the eight palindromes, the ampliconic regions of Yq and Yp contain five sets of more widely spaced inverted repeats with repeat lengths of 62–298 kb (Fig. 2; see also Supplementary Table 4). Three of these inverted repeat pairs (IR1, IR2 and IR3)

exhibit nucleotide identities of 99.66–99.95%. Inversion of the IR3 repeats, both located on Yp, was probably a direct consequence of the molecular evolutionary event that cleaved the X-transposed sequences into two non-contiguous segments (Supplementary Fig. 5). Subsequent homologous recombination between inverted IR3 repeats was responsible, we suspect, for a 3.6-Mb inversion polymorphism observed on the short arm of the modern Y chromosome (Supplementary Figs 5 and 6)¹⁰.

Transcriptionally active tandem arrays

In addition to palindromes and inverted repeats, the ampliconic regions of Yq and Yp contain a variety of long tandem arrays. Prominent among these are the newly identified NORF (no long open reading frame) clusters, which in aggregate account for about 622 kb on Yp and Yq, and the previously reported TSPY clusters, which comprise about 700 kb of Yp (Fig. 2). Triangular dot plots that highlight the regularities and relatively crisp borders of the NORF and TSPY arrays are shown in Supplementary Fig. 7 (see also Supplementary Note 5).

The NORF arrays are based on a repeat unit of 2.48 kb. A consensus sequence for the repeat is readily identifiable (Supplementary File 2), but the sequence of individual repeat elements typically diverges from that consensus by 14–20%. The NORF arrays are so named because they harbour a great diversity of spliced but apparently non-coding transcription units, including the *TTTTY1*, *TTTTY2*, *TTTTY6*, *TTTTY7*, *TTTTY8*, *TTTTY18*, *TTTTY19*, *TTTTY21* and *TTTTY22* families. Both strands of the NORF arrays are transcribed; 3' portions of the *TTTTY1* and *TTTTY2* transcripts are complementary (Supplementary Fig. 8).

The TSPY arrays are based on a 20.4-kb repeat unit²³ that encodes, on one strand, a previously identified protein, TSPY. A newly identified transcription unit, *CYorf16*, is found on the opposite

Figure 2 Sequence-based map of the MSY; a detailed view of the 24-Mb region shown in Fig. 1b. Background colours indicate the three classes of MSY euchromatic sequences: X-transposed (pink), X-degenerate (yellow) and ampliconic (blue), as well as heterochromatic (red stripes) and pseudoautosomal (green) sequences and NORF arrays (grey stripes). Two gaps in the sequence are indicated at the top edge of the diagram. **a**, Eight primary palindromes (P1–P8) and two secondary palindromes (P1.1 and P1.2). Diverging black arrows mark the left and right arms of each palindrome. Gaps between diverging arrows represent non-palindromic spacers at the centres of these structures. **b**, Near-perfect inverted repeats (non-palindromic), three in all (IR1 to IR3; Supplementary Table 4). In each case, the left and right arms exhibit >99.5% nucleotide identity. **c**, Other inverted repeats (non-palindromic). Grey arrows (IR4) denote two regions of >93% identity, one on Yp and one on Yq. Yellow arrows (IR5) denote four regions of >92% identity, all on Yq. **d**, Deletions of any of the four indicated regions—*AZFa*, P5/proximal P1 (*AZFb*), *AZFc*, or P5/distal P1—cause spermatogenic failure^{13,43–46}. **e**, Previously reported genes and new, experimentally verified transcription units for which cDNA sequencing suggests protein-coding potential (Table 2). Plus (+) and minus (–) strand are indicated by the top or bottom row, respectively. **f**, See Fig. 5c. **g**, Scale (Mb). **h**, Sequences whose transcription has been verified (in this or previous studies) but for which there is little or no evidence of protein-coding potential (Supplementary Table 2). **i**, Previously reported pseudogenes and new, apparently non-transcribed homologues of known coding genes (Supplementary Table 3). **j**, (G + C) content (%) calculated in a 100-kb sliding window with 1-kb steps. **k**, Alu, LINE1 and human endogenous retroviral (HERV) repeat content, expressed as percentage of nucleotides, calculated in a 200-kb sliding window with 1-kb steps. **l**, 220 BAC clones completely or partially sequenced. Each bar represents size and position of one BAC clone, identified by the numeric portion of its GenBank accession number (in each case beginning with the prefix AC). Black bars represent finished sequences deposited in GenBank, where finished sequences are trimmed to retain only 200 bp of overlap with adjoining BACs. Grey bars represent the 'trimmings' of those BACs, not deposited in GenBank. Striped bars represent BACs whose sequence has not been finished but has been deposited in GenBank. See Supplementary Fig. 2 for a more detailed version of this figure. The composite sequence of the 24-Mb region studied is available as Supplementary File 1.

strand; its protein coding potential remains to be tested. Approximately 35 copies of this repeat unit—and hence 35 *TSPY* genes and 35 *CYorf16* transcription units—are found in a single, highly regular tandem array in proximal Yp (Fig. 2 and Supplementary Fig. 7d, e); here the sequences of individual repeat units rarely differ from the consensus by more than 1%. Furthermore, a single, isolated *TSPY* repeat unit, whose sequence diverges 3% from the consensus, is located more distally in Yp, embedded in the distal IR3 inverted repeat (Fig. 2). The 35-unit *TSPY* cluster is the largest and most homogeneous protein-coding tandem array identified so far in the human genome.

The evolution of the MSY

On the basis of our present findings and previous studies, we propose a model of MSY evolution that addresses all three euchromatic sequence classes (Figs 6 and 7). In developing the model, we will offer an evolutionary map of the MSY (Fig. 8). We will then consider the two largest and most gene-rich sequence classes—X-degenerate and ampliconic—arguing that two opposed evolutionary dynamics have been at work: gene decay versus gene acquisition and conservation. Throughout, we will propose decisive roles for

modulation of DNA recombination, both crossing over and gene conversion, in the evolution and on-going maintenance of the MSY (Fig. 9).

The human X and Y chromosomes are thought to have evolved from an ordinary pair of autosomes^{5,24}. Support for this hypothesis, and a proposed 300-million-year timeline for human sex chromosome evolution, have emerged from studies of modern X–Y gene pairs. In this context, investigators have interpreted the X–Y gene pairs as surviving ‘fossils’ where extensive sequence identity between ancestral X and Y chromosomes once existed^{25,26}. Our present sequencing of the MSY euchromatin expands the catalogue of known X–Y gene pair fossils, providing opportunity to re-examine models developed in earlier studies.

Evolutionary stratification of X–Y genes

Lahn and Page previously studied the evolutionary ages of X–Y gene pairs, as measured by synonymous X–Y nucleotide divergence, or K_s (ref. 26). They reasoned that X–Y differentiation would have begun only after X–Y crossing over ceased. They observed a strong correlation between the age (K_s) of individual X–Y gene pairs and the locations of their X members on the human X chromosome.

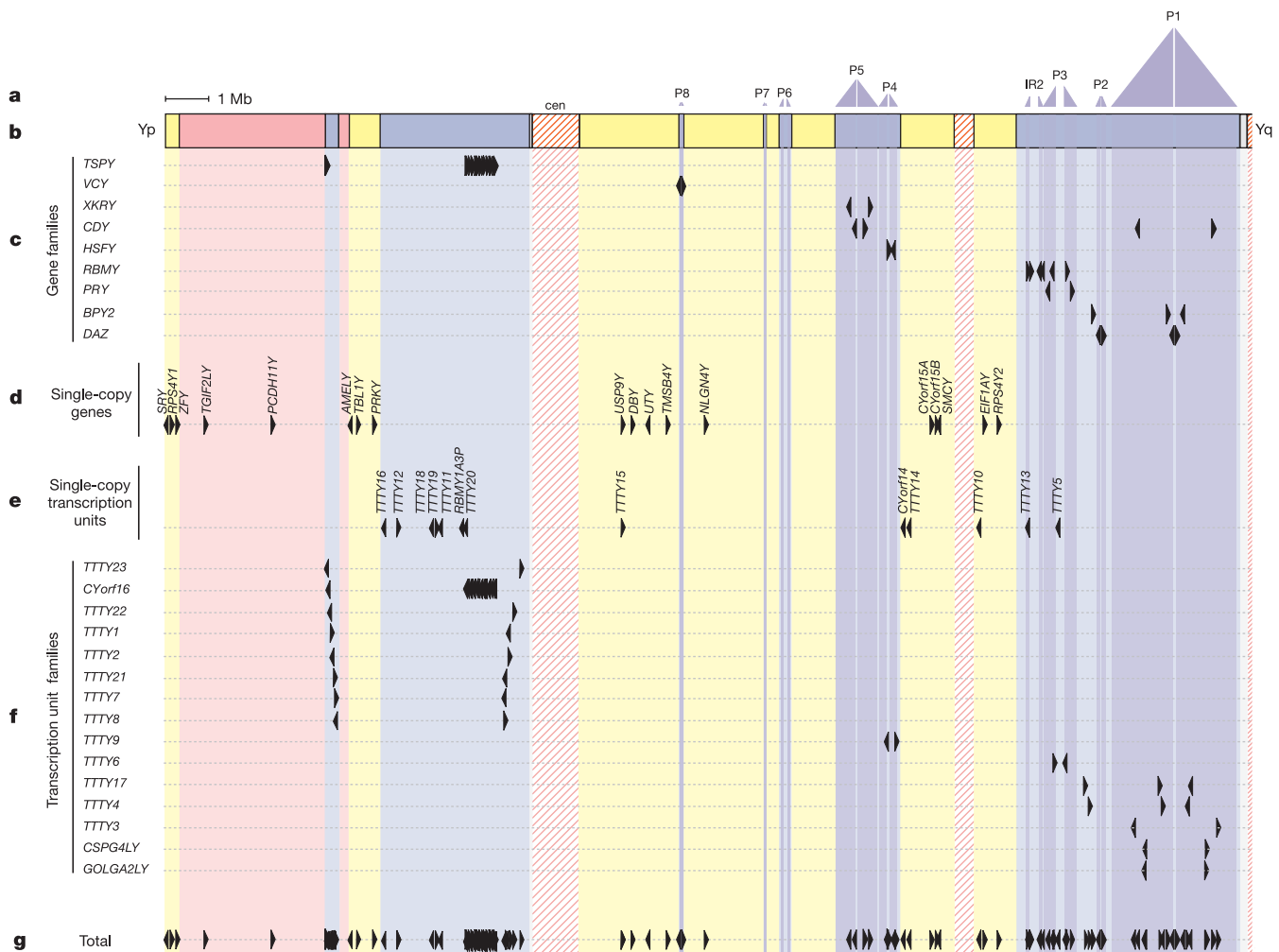


Figure 3 MSY genes, transcription units and palindromes. **a**, Triangles denote sizes and locations of arms of eight palindromes (P1–P8) and of IR2 inverted repeats (whose arms exhibit 99.95% identity). Gaps between opposed triangles represent the non-duplicated ‘spacers’ between palindrome arms. **b**, MSY schematic, as in Fig. 1b. **c**, Nine families of protein-coding genes. Solid triangles denote apparently intact genes (5’ to 3’ polarity

indicated); pseudogenes are not shown. **d**, Single-copy protein-coding genes. **e**, Single-copy transcription units. These give rise to spliced but apparently non-coding transcripts. **f**, Fifteen families of transcription units. **g**, Merged map of all genes and transcription units.

Among the 19 X–Y gene pairs studied, age increased in a stepwise fashion along the length of the X chromosome, in four ‘evolutionary strata’. This suggested that at least four events had punctuated human sex chromosome evolution, with each event suppressing X–Y crossing over in one stratum without grossly disturbing gene order in the X chromosome.

We re-analysed this published information and combined the results with K_s and map location data for 12 additional X–Y gene pairs, thus compiling data on 31 X–Y pairs in all (Supplementary Table 5). In each of 27 pairs, the Y member is an X-degenerate gene

or pseudogene. The other four pairs include two in which the Y member is an X-transposed gene and two in which the Y members are ampliconic gene families.

Among all X-degenerate pairs, and the two ampliconic pairs, the previously reported correlation between age (K_s) and X map position is readily apparent, with age increasing from the distal short arm to the long arm of the X chromosome (Fig. 7). Furthermore, as observed in the earlier study, the order of the homologous genes in the MSY appears to be scrambled with respect to K_s (Supplementary Fig. 9). These observations, together with the

Table 2 MSY genes and gene families demonstrated or hypothesized to encode proteins

MSY sequence class	Gene symbol	Gene name	Number of copies†	Tissue expression	X-linked homologue	Autosomal homologue	
X-transposed	<i>TGIF2LY*</i>	TGF (beta)-induced transcription factor 2-like Y	1	Testis	<i>TGIF2LX</i>	–	
	<i>PCDH11Y</i>	Protocadherin 11 Y	1	Fetal brain, brain	<i>PCDH11X</i>	–	
Total			2				
X-degenerate	<i>SRY</i>	Sex determining region Y	1	Predominantly testis	<i>SOX3</i>	–	
	<i>RPS4Y1</i>	Ribosomal protein S4 Y isoform 1	1	Ubiquitous	<i>RPS4X</i>	–	
	<i>ZFY</i>	Zinc finger Y	1	Ubiquitous	<i>ZFX</i>	–	
	<i>AMELY</i>	Amelogenin Y	1	Teeth	<i>AMELX</i>	–	
	<i>TBL1Y*</i>	Transducin (beta)-like 1 protein Y	1	Fetal brain, prostate	<i>TBL1X</i>	–	
	<i>PRKY</i>	Protein kinase Y	1	Ubiquitous	<i>PRKX</i>	–	
	<i>USP9Y</i>	Ubiquitin-specific protease 9 Y	1	Ubiquitous	<i>USP9X</i>	–	
	<i>DBY</i>	Dead box Y	1	Ubiquitous	<i>DBX</i>	–	
	<i>UTY</i>	Ubiquitous TPR motif Y	1	Ubiquitous	<i>UTX</i>	–	
	<i>TMSB4Y</i>	Thymosin (beta)-4 Y	1	Ubiquitous	<i>TMSB4X</i>	–	
	<i>NLGN4Y</i>	Neurologin 4 isoform Y	1	Fetal brain, brain, prostate, testis	<i>NLGN4X</i>	–	
	<i>CYorf15A*</i>	Chromosome Y open reading frame 15A	1	Ubiquitous	<i>CXorf15</i>	–	
	<i>CYorf15B*</i>	Chromosome Y open reading frame 15B	1	Ubiquitous	<i>CXorf15</i>	–	
	<i>SMCY</i>	SMC (mouse) homologue, Y	1	Ubiquitous	<i>SMCX</i>	–	
	<i>EIF1AY</i>	Translation initiation factor 1A Y	1	Ubiquitous	<i>EIF1AX</i>	–	
	<i>RPS4Y2*</i>	Ribosomal protein S4 Y isoform 2	1	Ubiquitous	<i>RPS4X</i>	–	
	Total			16			
	Ampliconic	<i>TSPY</i>	Testis-specific protein Y	~35	Testis	–	–
		<i>VCY</i>	Variable charge Y	2	Testis	<i>VCX</i>	–
<i>XKRY</i>		XK related Y	2	Testis	–	–	
<i>CDY</i>		Chromodomain Y	4	Testis	–	<i>CDYL</i>	
<i>HSFY*</i>		Heat shock transcription factor Y	2	Testis	–	–	
<i>RBMX</i>		RNA-binding motif Y	6	Testis	<i>RBMX</i>	–	
<i>PRY</i>		PTP-BL related Y	2	Testis	–	–	
<i>BPY2</i>		Basic protein Y 2	3	Testis	–	–	
<i>DAZ</i>		Deleted in azoospermia	4	Testis	–	<i>DAZL</i>	
Total				~60			
Grand total			~78				

See Supplementary Table 1 for a more detailed version of this Table, incorporating references and GenBank accession numbers.

*Genes first reported in this study.

†Excluding pseudogenes.

Sequence class	Defining characteristics	Evolutionary origins	Distribution	Aggregate length (Mb)	No. of coding genes	No. of non-coding transcription units	No. of transcription units per Mb	Nucleotides in interspersed repeats (%)
X-transposed	99% identity to X	Single transposition from X	2 blocks on Yp	3.4	2	0	0.6	60
X-degenerate	Single-copy gene or pseudogene homologues of X-linked genes	Relics of ancient autosomes from which X and Y evolved	8 blocks on Yp and Yq	8.6	16, most expressed widely	4	2.2	57
Ampliconic	Lengthy similarity to other MSY sequences	Acquired from diverse sources, then amplified	7 blocks on Yp and Yq	10.2	60 (in 9 families), expressed mainly or only in testes	74 (9 single-copy; 65 in 15 families), expressed mainly or only in testes	13.3	36

Figure 4 Three sequence classes in the MSY euchromatin. Colour scheme as in Fig. 2.

earlier arguments of Lahn and Page, suggest three conclusions. First, all MSY genes and pseudogenes identified here as X-degenerate seem to be products of a single molecular evolutionary process: the region-by-region suppression of crossing over in ancestral autosomes, with subsequent differentiation of the Y from the X chromosome (Fig. 6). Second, at least two of the MSY's ampliconic gene families, *VCY* and *RBMY*, also originated in this manner, but subsequently acquired the characteristics of ampliconic sequences (Fig. 6; for independent evidence concerning *RBMY* see refs 27 and 28). Third, as previously hypothesized, inversions in the Y chromosome may have suppressed crossing over with the X chromosome.

X-transposed genes as exceptions

A very different evolutionary model accounts for the X-transposed genes, as confirmed by our K_s analysis. If, as hypothesized, these MSY genes are the result of a single, recent transposition from the X chromosome (Fig. 6), then the K_s values of the two X-transposed X–Y gene pairs should be similar to each other but much lower than the K_s values of the nearby (X-degenerate) pairs in the X-chromosome long arm. This prediction is met (Fig. 7). The two X-transposed X–Y gene pairs seem to be orders of magnitude younger than the ancient pairs (group 1 in Fig. 7) among which they are physically situated in the X chromosome.

Table 3 MSY palindromes

Palindrome	Arm length (kb)	Arm-to-arm identity (%)	Spacer length (kb)	Palindrome span (kb)
P1	1,450	99.97	2.1	2,902
P1.1*	9.9	99.95	3.9	24
P1.2*	9.9	99.95	3.9	24
P2	122	99.97	2.1	246
P3	283	99.94	170	736
P4	190	99.98	40	419
P5	496	99.98	3.5	996
P6	110	99.97	46	266
P7	8.7	99.97	12.6	30
P8	36	99.997	3.4	75
Total span (kb)				5,670

*Palindromes P1.1 and P1.2 are located within, respectively, the distal and the proximal arms of palindrome P1 (Fig. 2).

Blurred boundaries

Our observations differ from those of Lahn and Page in that the boundaries between X–Y gene groups 2 and 3, and between groups 3 and 4, now seem less distinct (Fig. 7; compare with Fig. 2 in ref. 26). Whereas our present observations could be interpreted as evidence that suppression of X–Y crossing over evolved in more than four steps, such a conclusion would be premature. The apparent overlaps

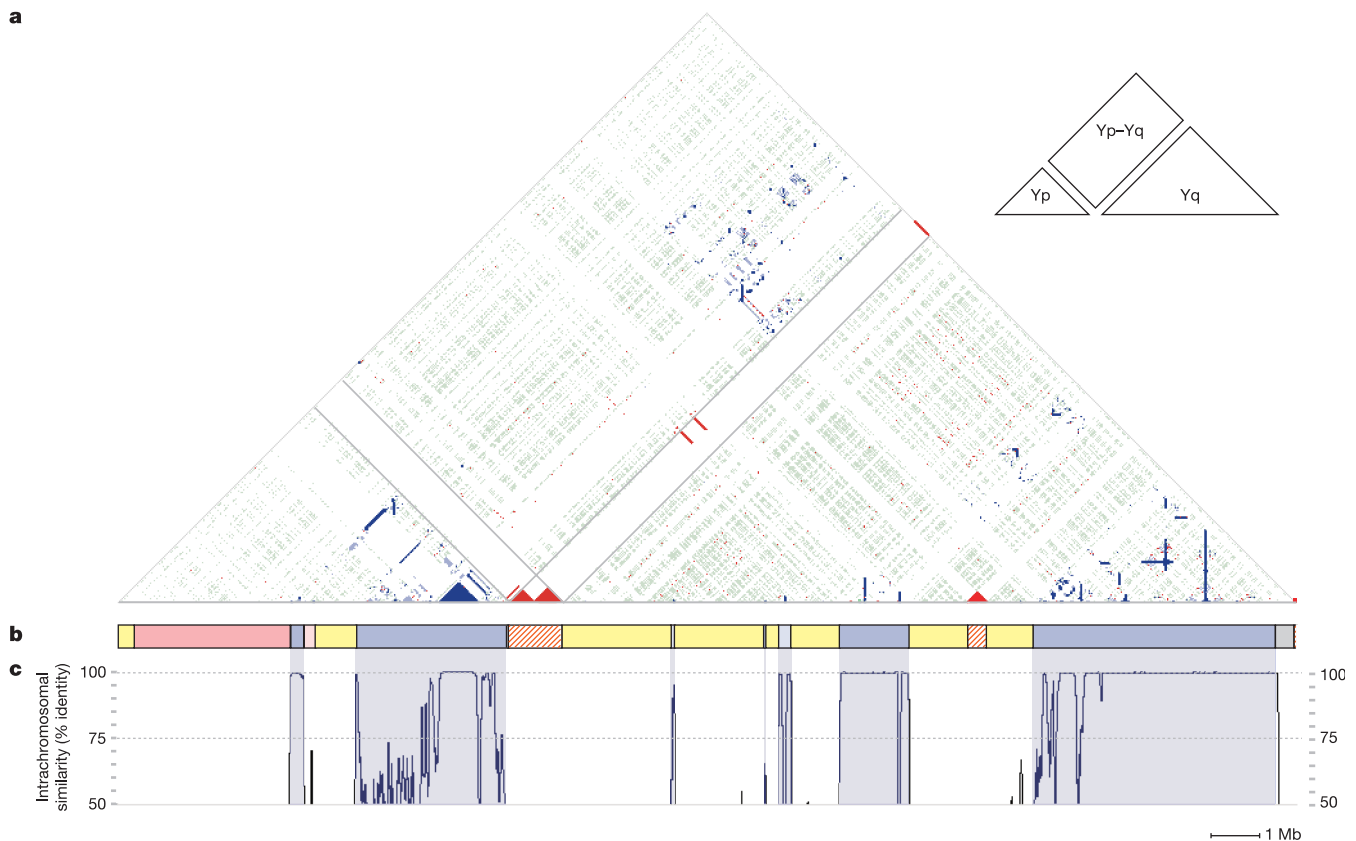


Figure 5 Sequence similarities within the MSY. **a**, Triangular dot plot in which the MSY's sequence is compared to itself. Within the plot, each dot represents a match of >65% within a window of 2,000 nucleotides. Green dots represent matches of this quality between LINE1 elements; red dots represent matches between heterochromatic sequences; blue dots represent matches between all other sequences. Direct repeats appear as horizontal lines, inverted repeats as vertical lines, and palindromes as vertical lines that nearly intersect the baseline. Long arrays of tandem repeats appear as pyramids. The inset indicates that the large triangular plot contains two smaller triangles (one revealing sequence similarities within Yp, and one revealing similarities within Yq)

and a rectangle (revealing similarities between Yp and Yq). **b**, MSY schematic, as in Fig. 1b. **c**, Plot of intrachromosomal sequence similarity, which serves to identify ampliconic sequences (blue). Using a 50-kb sliding window and 1-kb steps, each MSY euchromatic sequence was compared to all other available MSY euchromatic sequences. (Long interspersed repeats were excluded before analysis.) At each point along the length of the MSY, the highest sequence similarity (expressed as per cent nucleotide identity) was identified. All such values >50% are shown. An expanded version of this plot is shown in Fig. 2f.

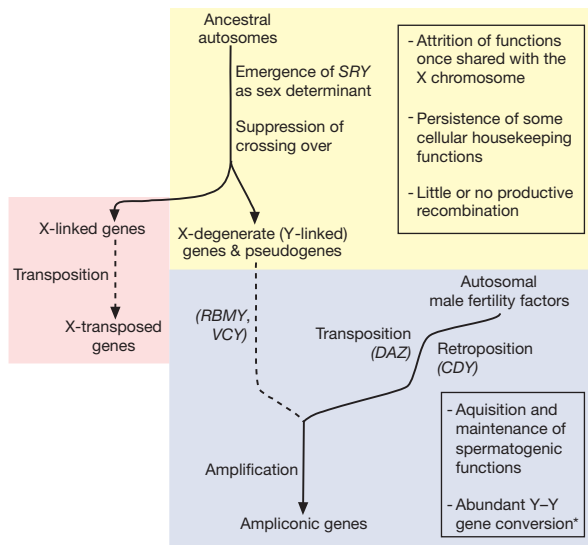


Figure 6 Molecular evolutionary pathways and processes that gave rise to genes in three MSY euchromatic sequence classes. X-degenerate genes and pseudogenes (yellow background) derived from an autosomal pair that was ancestral to both the X and Y chromosomes (and that was enlarged by subsequent fusion with other autosomes or autosomal segments⁵⁰). X-transposed genes (pink background) derived from X-linked genes, which in turn derived from the ancestral autosomal pair. Ampliconic genes (blue background) were derived through three converging processes: amplification of X-degenerate genes (for example, *RBMY*, *VCY*); transposition and amplification of autosomal genes (*DAZ*); and retroposition and amplification of autosomal genes (*CDY*). Boxes enumerate dominant themes in X-degenerate (yellow) and ampliconic (blue) gene evolution. The asterisk indicates that Y–Y gene conversion is apparently common in the 61% of ampliconic sequences that exhibit intrachromosomal identities of $\geq 99.9\%$.

between groups could be artefacts of local errors in ordering X-linked genes, these regions not yet having been fully sequenced, or simply of large standard errors for some K_s estimates (Fig. 7). Some changes in local gene order in the X chromosome may also have occurred during its evolution. Another potentially confounding factor is X–Y gene conversion, which would depress K_s values and estimated ages for gene-converted X–Y pairs. Gene conversion depends on high sequence similarity, and thus one might expect any such effect to be greater among the younger X–Y pairs, in groups 3 and 4. Indeed, comparisons of X and Y genomic sequences suggest that the *VCX/Y* pair and 3' portions of the *KALI/P* pair (both pairs in group 4) have engaged in extensive gene conversion (Supplementary Fig. 10), depressing their K_s values below those of the 5' portion of the *KALI/P* pair and of other group 4 pairs (Fig. 7).

A map of male-specific ages

Having examined the evolutionary ages of all 31 X–Y gene pairs, we used them to anchor an evolutionary map of the modern human MSY. The map displays the male-specific ages of many sequence segments (Fig. 8). Here, male-specific age is the estimated number of years that have passed since sequences ancestral to that segment were incorporated into the MSY (having previously been autosomal, pseudoautosomal, or X-linked). We estimated the age of each gene or segment using Lahn and Page's methods that combined K_s analysis (Supplementary Table 5) with comparative gene mapping data from other mammals. The resulting estimated ages are graphed on a logarithmic scale to accommodate a range that extends from approximately 4 million years (the X-transposed sequences; the youngest known sequences in the MSY) to approximately 300 million years (*SRY*, the sex determinant and arguably the oldest gene in the MSY).

As can be seen in Fig. 8, the MSY euchromatin is an elaborate patchwork of sequences of diverse male-specific ages. The result of a single, recent transposition from the X-chromosome, the MSY's X-transposed sequences are homogeneously youthful. The

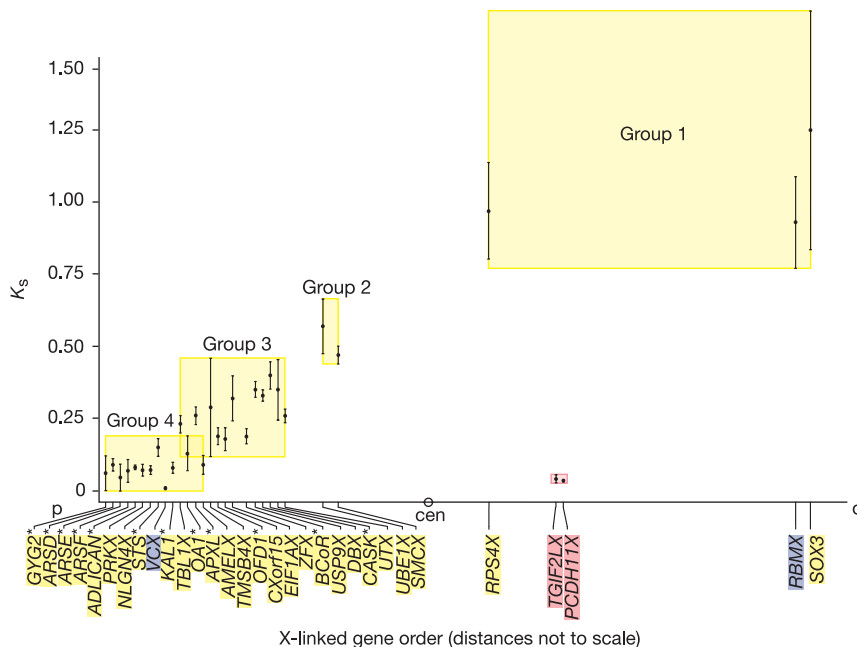


Figure 7 Plot of K_s (Supplementary Table 5) versus X-linked gene order for 31 X–Y gene (or gene/pseudogene) pairs. Colour highlighting of X-linked gene names indicates whether Y homologues are X-degenerate (yellow), ampliconic (blue) or X-transposed (pink). Within the plot, four yellow rectangles denote four previously defined 'evolutionary

strata', or groups of genes²⁶; a small pink rectangle highlights two X-transposed genes. Genes in the X chromosome are ordered according to the NCBI sequence assembly of November 2002; distances between genes are not drawn to scale. Standard errors for K_s values are shown.

sequences of both the X-degenerate and ampliconic classes are much older, and they display a wide range of male-specific ages (Fig. 8). As we will argue, it is in comparing and contrasting these two chronologically diverse classes that the central themes of MSY evolution and function are revealed most clearly.

Evolutionary dynamics of X-degenerate and ampliconic sequences

To appreciate the evolutionary dynamics of these two sequence classes, we need to consider both their similarities and differences. In many senses, the X-degenerate and ampliconic sequences together dominate the euchromatic MSY. The X-degenerate and ampliconic classes are physically intermingled in the MSY, and they are comparably large, constituting, respectively, 38% and 45% of the MSY's euchromatic sequences (Fig. 1 and Supplementary Table 6). Together, these two sequence classes carry all but two of the MSY's 78 known protein-coding transcription units (Table 2). The X-degenerate and ampliconic classes display comparable diversities of male-specific ages, from tens to hundreds of millions of years (Fig. 8). This implies that X-degenerate and ampliconic sequences evolved in parallel, as parts of a single DNA molecule, for as much as 300 million years. Moreover, we infer that the X-degenerate and ampliconic sequences evolved under similar, unusual circumstances: both were transmitted exclusively through the male germ line, and neither participated in meiotic crossing over with a homologous counterpart. However, a number of marked structural and functional differences between these two sequence classes

suggest that they followed different evolutionary trajectories. Palindromes are prevalent in ampliconic sequences. The density of transcription units is much higher and the density of interspersed repeats is much lower in ampliconic than in X-degenerate sequences (Fig. 1). The two sequence classes also diverge starkly with respect to gene-expression patterns. Most X-degenerate genes are expressed widely throughout the body, and many are probably involved in cellular housekeeping activities that are critical in both males and females. In contrast, most ampliconic genes are expressed predominantly or exclusively in testes, where they probably function in spermatogenesis.

Decay in the absence of sexual recombination

The X-degenerate sequences are adequately explained by the prevailing theory of sex chromosome evolution, which states that as the X and Y chromosomes evolved from an autosomal pair, the X chromosome maintained most of its ancestor's genes whereas the Y chromosome lost them^{5,24-26}. Our findings support the two major premises of this theory: the evolutionary genetic benefits of sexual recombination through meiotic crossing over, and the deleterious consequences of its absence. According to this theory, most ancestral genes remained functionally intact in the X chromosome, where the benefits of crossing over (in females) continued. In the Y chromosome, in contrast, the shutting down of X-Y crossing over during evolution triggered a monotonic decline in gene function. This model is corroborated by the presence, in the MSY's X-degenerate sequences, of decayed, intron-bearing pseudogenes of

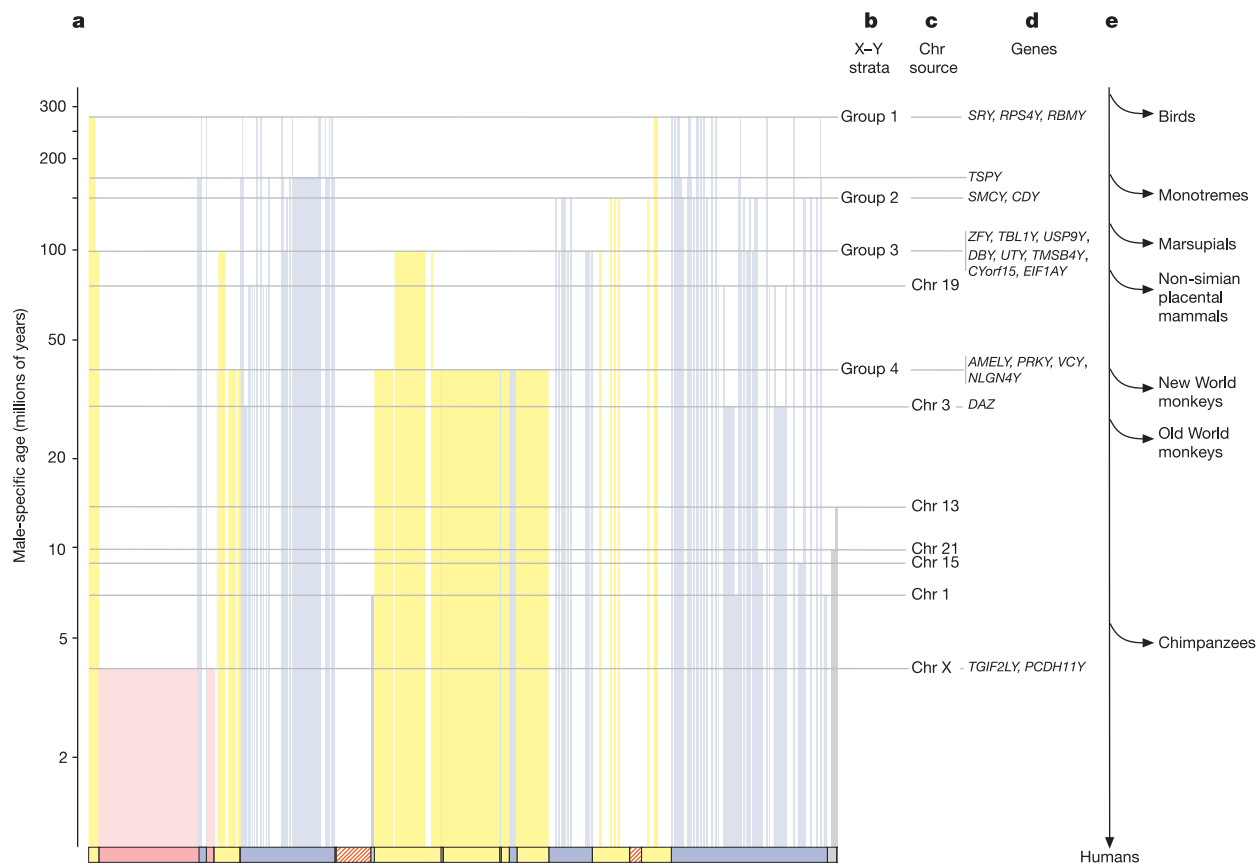


Figure 8 Evolutionary map of the MSY. At the bottom is an MSY schematic, as in Fig. 1b. Coloured rectangles extending above this schematic depict the estimated male-specific ages of the corresponding segments of the modern MSY. These ages are plotted on a logarithmic scale (a). b, X-Y strata 1, 2, 3 and 4 (ref. 26 and Fig. 7). c, The chromosomes (more properly, the modern human orthologues of the chromosomes) from which the

indicated X-transposed or ampliconic sequences apparently arose through transposition, during evolution. d, MSY genes that apparently arose at the indicated times. e, Approximate times of divergence between the human and certain other vertebrate lineages. The methods used to estimate the male-specific ages of each of the sequences and genes shown are listed in Supplementary Table 9.

13 different X-linked genes (Supplementary Table 3). Presumably, many hundreds of other X-homologous genes were deleted outright from the evolving MSY, leaving no trace in the DNA sequence of the modern human MSY. Seen in this light, the 16 protein-coding genes in the modern MSY's X-degenerate sequences (Table 2 and Fig. 3) appear as rare examples of persistence in the absence of sexual recombination.

Acquisition and conservation of spermatogenic functions

This evolutionary model of the Y chromosome as a decaying X chromosome, however, provides no explanation for central characteristics of the MSY's ampliconic sequences, including testis-specific gene expression, near-perfect palindromes, and an abundance of autosomal (as well as X-chromosomal) sequence similarities. To account for these characteristics, we propose that the MSY acquired, and evolved a means of conserving, genes that specifically enhanced male fertility.

Unlike the X-degenerate sequences, all of which trace to the MSY's shared ancestry with the X chromosome, the ampliconic sequences evolved from a great variety of genomic sources, and by a diversity of molecular mechanisms (Fig. 6). As mentioned previously, the ampliconic genes *VCY* and *RBMY* were, similar to the X-degenerate genes, derived from common ancestors of the X and Y chromosomes^{27,28}. In contrast, the *DAZ* genes arose, during primate evolution, by transposition and subsequent amplification of an autosomal transcription unit, *DAZL*, which still exists on human chromosome 3 (ref. 29). Indeed, systematic analysis of MSY/autosomal similarities suggests that a series of autosomal transpositions contributed to the MSY's ampliconic sequences during primate evolution (Fig. 8; see also ref. 13). Yet another molecular mechanism accounts for the *CDY* genes, which arose by retroposition (and subsequent amplification) of a processed messenger RNA derived from an autosomal gene³⁰. This retroposition event was previously thought to have occurred during primate evolution, but our present *K_s* analysis indicates a much older date, probably before the lineages of marsupials and placental mammals diverged (Fig. 8; see also Supplementary Table 5).

Despite the wide variety of genomic sources and molecular evolutionary mechanisms that gave rise to the ampliconic genes, they all came to exist in the MSY in multiple, nearly identical copies, and they evolved remarkably uniform patterns of tissue expression. Indeed, detailed studies of several ampliconic gene families have revealed that they are expressed predominantly or exclusively in one cell lineage: the spermatogenic cells of the testis. What accounts for this convergence of evolutionary outcomes? The genesis of XY sex chromosomes during mammalian evolution, and specifically the emergence of a male-specific domain, created a genomic niche where selection could operate to enhance male germ-cell development. Amplification of the testis genes might have enhanced sperm production through high levels of expression. However, in a region

devoid of crossing over, amplification might also have allowed another type of homologous recombination, gene conversion, to emerge as a means of conserving gene function.

Abundant Y–Y gene conversion in ampliconic regions

Gene conversion is the non-reciprocal transfer of sequence information from one DNA duplex to another³¹. This type of genetic recombination has been studied most extensively in fungi, where it was originally demonstrated to occur between chromosome homologues, or at lower frequency between sister chromatids, in meiosis. It was later shown that gene conversion could also occur between duplicated sequences on a single chromosome, and in mitosis³². Here we will argue that gene conversion (non-reciprocal recombination) is as frequent in the MSY as crossing over (reciprocal recombination) is in ordinary chromosomes.

Specifically, two major findings provide evidence that gene conversion occurs routinely in 30% of the MSY euchromatin, including nearly all of the MSY's testis-specific gene families. The accompanying study⁷ reports the identification and sequencing of chimpanzee Y-linked orthologues of human MSY palindromes and establishes that gene conversion between palindrome arms has occurred in both the human and chimpanzee lineages, and has continued to occur in human populations. Here we report that these palindromes are representative of a large, discrete fraction of MSY sequences, all of which bear at least 99.9% identity to other MSY sequences. These findings suggest that the entire fraction is subject to frequent gene conversion.

Above we described calculations of percentage nucleotide identity between all pairs of known MSY sequences. We defined and mapped the ampliconic regions by reporting, at each point along the length of the MSY euchromatin, the highest percentage identity to other MSY sequences (intrachromosomal similarity; Fig. 5). To view this data from another perspective, we electronically fractionated all MSY sequences according to intrachromosomal similarity. As seen in Fig. 9a, 30% of MSY euchromatic sequences display intrachromosomal identities of 99.9–100%. As intrachromosomal identity declines below 99.9%, the fractional representation of MSY sequences drops abruptly. Thus, the sequences displaying intrachromosomal identities of $\geq 99.9\%$ represent a large and distinct subset of the MSY euchromatin.

This $\geq 99.9\%$ subset comprises the eight palindromes as well as large portions of the IR2 and IR3 inverted repeats described above (Figs 2 and 3). Indeed, nearly all of the $\geq 99.9\%$ sequences exist as pairs in inverted orientation. Thus, the MSY palindromes in which gene conversion has been demonstrated⁷ are typical and representative of the $\geq 99.9\%$ fraction. We extrapolate that nearly all of the $\geq 99.9\%$ fraction is engaged in gene conversion on a routine basis, resulting in a degree of identity among MSY's inverted sequence pairs that rivals that of two autosomal homologues, or alleles, chosen at random from the human population^{15,33}.

Table 4 MSY testis gene family members in regions exhibiting $\geq 99.9\%$ or $< 99.9\%$ intrachromosomal identity

Gene family	Number of genes		Number of pseudogenes	
	Regions $\geq 99.9\%$ intrachromosomal identity	Regions $< 99.9\%$ intrachromosomal identity	Regions $\geq 99.9\%$ intrachromosomal identity	Regions $< 99.9\%$ intrachromosomal identity
<i>VCY</i>	2	0	0	0
<i>XKRY</i>	2	0	6	0
<i>CDY</i>	4	0	17	4
<i>HSFY</i>	2	0	0	0
<i>RBMY</i>	6	0	6	17
<i>PRY</i>	2	0	4	0
<i>BPY2</i>	3	0	Many	Many
<i>DAZ</i>	4	0	0	0
Total	25	0	>33	>21

The *TSPY* genes, the only MSY genes found in long tandem arrays, are excluded from this analysis.

Two modes of productive recombination in the human Y chromosome

Combined with previous discoveries in the pseudoautosomal regions, the present findings imply that two modes of homologous recombination occur regularly in the human Y chromosome. First, there is crossing over with the X chromosome in the pseudoautosomal regions (aggregate length 3.0 Mb) (Supplementary Note 6). Second, there is Y–Y gene conversion in the $\geq 99.9\%$ regions (aggregate length 6.1 Mb) dispersed throughout the MSY (Fig. 9b)⁷. We refer to both routine modes of Y chromosome recombination as ‘productive’ to distinguish them from the relatively rare, aberrant recombination events (typically Y–Y or X–Y) that perturb sex differentiation or fertility and thereby diminish the reproductive fitness of affected individuals.

Genetic mapping studies have shown that, typically, one X–Y crossover occurs per generation in the pseudoautosomal regions (Supplementary Note 7). As described in the accompanying report⁷, steady-state calculations suggest that, on average, multiple Y–Y gene conversion events take place per generation in the MSY. Thus, most homologous recombination events in the Y chromosome probably occur in the MSY.

In recent years, we and other investigators have referred to the MSY as the NRY, or ‘non-recombining region of the Y chromosome’. This usage reflected both awareness that productive X–Y crossing over did not occur in the MSY, and ignorance of the Y–Y gene conversion that is apparently commonplace there. We now refer to the NRY as the MSY, or ‘male-specific region of the Y chromosome’, because it is recombinogenic and unique to males.

Gene conversion and the MSY’s testis gene families

Examination of the MSY’s testis gene families provides additional insight into the potential biological significance of the $\geq 99.9\%$ fraction and the gene conversion associated with it. Eight of the

MSY’s nine identified testis gene families have members in the palindromes or inverted repeats that comprise the $\geq 99.9\%$ fraction just described. (The exceptional family is *TSPY*, most of whose members are found in a long tandem array.) Many of these family members are intact gene copies, but others are apparent pseudogenes with disrupted splice sites or reading frames. For each of the eight testis gene families, we counted the numbers of intact and pseudogene copies, both within and without the $\geq 99.9\%$ fraction (Table 4). Whereas large numbers of pseudogenes are present both inside and outside the $\geq 99.9\%$ fraction, the intact gene copies, 25 in all, are located exclusively in the $\geq 99.9\%$ fraction.

Thus, there is an evident association of intact testis genes with near-identical inverted sequence pairs that undergo gene conversion. What is the biological significance of this association? We envision two possibilities, which are not mutually exclusive. First, we note that in all cases examined so far, expression of these testis-specific gene families has been found to be limited to or most pronounced in cells of the spermatogenic lineage—in germ cells. Perhaps these near-identical sequence pairs are transcriptionally active in germ cells because there they generate cruciforms or other unusual chromatin configurations. Second, the occurrence of MSY gene pairs that are subject to frequent gene conversion might provide a mechanism for conserving gene functions across evolutionary time in the absence of crossing over.

Implications for future studies

We anticipate that the nucleotide sequence reported here, and the methods with which it was obtained, will find many applications in human biology and beyond.

Comparisons with other human Y chromosomes

The sequence of one man’s MSY, as reported here, provides a point

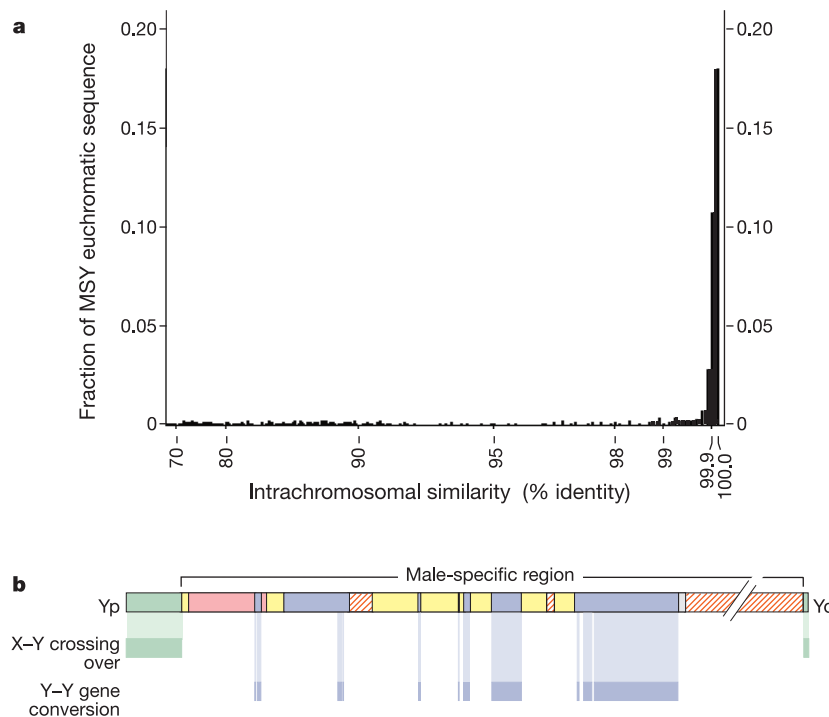


Figure 9 MSY sequences exhibiting $\geq 99.9\%$ intrachromosomal identity probably undergo Y–Y gene conversion. **a**, Electronic fractionation of MSY euchromatic sequences according to intrachromosomal similarity (per cent identity to other MSY sequences), plotted on a logarithmic scale. Values $< 70\%$ are not shown. **b**, Sites of productive recombination in the Y chromosome. Shown at the top is a schematic representation of

the entire Y chromosome, including the pseudoautosomal regions (green). The pseudoautosomal regions are sites of frequent X–Y crossing over. Within the MSY’s ampliconic sequences are many sites of apparently frequent Y–Y gene conversion; all of these sites display intrachromosomal identities of $\geq 99.9\%$.

of departure for systematic, comprehensive characterization of MSY sequence variation in human populations. The MSY's unique characteristics—male specificity, no crossing over and abundant gene conversion—suggest that its sequence variation might differ markedly from that of ordinary human chromosomes. Already the availability of MSY sequence information in public databases has accelerated the emergence of MSY sequence variation as a powerful tool in reconstructing the patrilineal origins of modern human populations^{11,34}.

Comparisons (or lack of) with other species

Little is known about the DNA sequences of Y chromosomes in other animals or plants, and thus it is not possible at present to compare systematically the human MSY with that of any other species. Both the *Drosophila* and mouse Y chromosomes contain genes required for spermatogenesis, but meagre Y chromosome sequence data is available in either species. In *Drosophila*, the sequences of autosomes and the X chromosome were assembled from whole-genome shotgun data. Unfortunately, this shotgun analysis was insufficient to assemble much Y chromosome sequence^{35,36}, confirming prior suspicions that, in *Drosophila* as in humans, the Y chromosome poses special challenges. In the mouse, a draft sequence of the female genome is available³⁷, but systematic efforts to sequence the male-specific region of the Y chromosome have yet to be initiated. If undertaken, Y chromosome sequencing projects in *Drosophila*, mouse and other species are likely to encounter special technical hurdles, but they are also likely to yield entirely unforeseen biological insights, as was the case here for the human MSY. The availability of human MSY sequence has already enabled new tests and rekindled debate of Haldane's hypothesis that mutations in the male germ line greatly outnumber those in the female germ line (Supplementary Note 8). This debate will surely be fuelled by sequencing of other primate and mammalian Y chromosomes.

Methods for sequencing difficult genomic regions

Our strategy of iterative mapping and sequencing was laborious but essential. Two faster, less costly strategies have been used recently in sequencing large genomes: whole-genome shotgun analysis^{15,35} and sequencing a tiling path of mapped clones (ref. 14 and Supplementary Note 9). Neither of these sequencing strategies would have yielded a coherent picture of the MSY. This is especially true of the MSY's ampliconic regions, and most particularly the 30% of the MSY euchromatin (including the eight palindromes) exhibiting intrachromosomal similarities of $\geq 99.9\%$. Large amplicons like those described here are not unique to the MSY, but as in the MSY, they have proven to be formidable obstacles to whole-genome methods^{38,39}. The iterative mapping and sequencing strategy used here should be considered by genome scientists wishing to determine the structure and sequence of amplicon-rich regions of human autosomes, the X chromosome and other genomes.

The medical relevance of the MSY

Propelled by advances in MSY genomics, the biomedical significance of the MSY has begun to surface in recent years, with evidence of roles in such diverse processes as gonadal sex determination, skeletal growth, germ-cell tumorigenesis and graft rejection⁶. Two research areas that should benefit from the present MSY sequence and gene catalogue are of particular note. First, one of the most common chromosomal disorders of girls and women is Turner syndrome, classically associated with a 45,X (X0) karyotype. Haploinsufficiency of particular genes common to the X and Y chromosomes may be responsible for somatic features of the syndrome^{40–42}. In most cases, the molecular identity of these Turner genes remains to be determined. One or more Turner genes are likely to be found within the catalogue of X-degenerate genes (and their X-linked homologues; see Table 2).

A highly active area of MSY research explores spermatogenesis and the genetic basis of male infertility. MSY deletions have emerged as the most common of the known genetic causes of spermatogenic failure in human populations^{13,43–46}. The availability of MSY sequence has already begun to transform our understanding, enabling investigators to precisely define four distinct classes of recurrent MSY deletions causing spermatogenic failure, identify the MSY genes absent as a result of these deletions (typically members of testis-specific families), and demonstrate that most such deletions are the result of homologous recombination between near-identical amplicons^{13,43–46}. Thus, the ampliconic structures that may help preserve testis gene function across evolutionary time (through gene conversion) also put individuals at risk of spermatogenic failure (again, through homologous recombination).

Genetic and biological differences between males and females

It is commonly stated that the genomes of two randomly selected members of our species exhibit 99.9% nucleotide identity. In reality, this statement holds only if one is comparing two males, or two females. If one compares a female with a male, the second X chromosome (160 Mb, or roughly 3% of the diploid DNA content) is replaced by the largely dissimilar Y chromosome (60 Mb, or 1% of the diploid DNA content). This common substitution of the Y chromosome for the second X chromosome dwarfs all other DNA polymorphism in the human genome. In decades past, and with the important exception of X-linked recessive diseases, biologists often judged this genomic dimorphism to be of limited functional consequence, especially because of inactivation of the second X chromosome in females and the presumed paucity of genes in the Y chromosome. Now we must begin to reconsider this position, given the unanticipated number and variety of MSY genes, many of which are expressed throughout the body, and the fact that many X-linked genes are expressed from both X chromosomes in female cells⁴⁷. The present sequence of the MSY, and the emerging sequence of the X chromosome, offer the near prospect of a comprehensive catalogue of genetic and sequence differences between human males and females. Translating this knowledge into an understanding of the myriad differences between the sexes in anatomy, physiology, cognition, behaviour and disease susceptibility presents a monumental challenge, but surely one of broad significance and interest. □

Methods

Iterative mapping and sequencing

The method of iterative mapping and sequencing used here has been described^{10,13}. All MSY BACs selected for sequencing were isolated from the RPCI-11 library⁴⁸, with the exception of 11 clones (nine spanning the *AZFa* region¹², and two used to narrow gaps¹⁰) from the CITB and CITC libraries. We made frequent use of publicly available BAC-end sequences as a source of markers during the final stages of map construction⁴⁹. Two gaps were closed by long-range PCR; see Supplementary Fig. 11.

Unfortunately, no cell line is available from the donor of the RPCI-11 BAC library. Thus, to confirm the large-scale organization of MSY sequences reported here, we PCR-amplified the inner and outer boundaries of all palindromes in ten men with genetically diverse Y chromosomes (PCR primers in Supplementary Table 8). We sequenced all resulting products. These experiments confirmed that each palindrome boundary is present in the great majority of human Y chromosomes.

Intrachromosomal sequence similarity

Analyses of intrachromosomal similarity were performed using custom Perl code. This code used BLAST (<http://blast.wustl.edu>) to compare all 5-kb sequence segments, in 2-kb steps, to the entire remainder of the MSY sequence.

Interspersed repeats

We electronically identified interspersed repeats with RepeatMasker (<http://repeatmasker.genome.washington.edu>).

Homology to other chromosomes

To identify sequence similarities to other human chromosomes, we conducted BLAST searches against GenBank databases with the sequence of each MSY clone. Interspersed repeats and low-complexity regions were masked using RepeatMasker. To experimentally verify the chromosomal origins of sequences similar to the MSY, we designed STSs from

those sequences and assayed them against the NIGMS human/rodent somatic cell hybrid mapping panels 1 and 2 (NIGMS Human Genetic Cell Repository, <http://locus.umd.edu/nigms/maps/mapping.html>).

Identification of new genes and transcription units

We identified potential transcripts from three sources: (1) BLAST matches to cDNA sequences (EST or full length). We pursued matches where the cDNA sequence showed evidence of polyadenylation or splicing, or where there were multiple matching cDNA sequences. (2) BLAST matches to fragments of putative MSY transcripts that had been cloned by cDNA selection of testis cDNA against a flow-sorted, genomic Y-chromosome library¹⁹. (3) GenomeScan¹⁸ predictions in the NCBI annotation of Y-chromosome contigs. We then tested for transcription by RT-PCR as previously described¹³.

Chromosomal FISH

One- or two-colour FISH to human chromosomes was performed as previously described⁶.

Calculation of K_s and K_a

We calculated the numbers of synonymous substitutions per synonymous site (K_s) and of non-synonymous substitutions per non-synonymous site (K_a) as follows. We used FASTA (<ftp://ftp.virginia.edu/pub/fasta>) to align the pairs of coding sequences in Supplementary Table 5. For non-transcribed MSY pseudogenes, we used FASTA to align the genomic sequence of pseudogene exons to the corresponding transcribed coding sequence (Supplementary Table 5 and File 3). Then, as is standard practice, insertions/deletions were manually removed from the alignments. We calculated K_s and K_a for these alignments using the *diverge* function in the Wisconsin Package (Version 10.2, Genetics Computer Group).

Received 7 March; accepted 8 April 2003; doi:10.1038/nature01722.

- Painter, T. S. The Y-chromosome in mammals. *Science* **53**, 503–504 (1921).
- Stern, C. The problem of complete Y-linkage in men. *Am. J. Hum. Genet.* **9**, 147–166 (1957).
- Jacobs, P. A. & Strong, J. A. A case of human intersexuality having a possible XXY sex determining mechanism. *Nature* **183**, 302–303 (1959).
- Ford, C. E., Miller, O. J., Polani, P. E., de Almeida, J. C. & Briggs, J. H. A sex-chromosome anomaly in a case of gonadal dysgenesis (Turner's syndrome). *Lancet* **1**, 711–713 (1959).
- Ohno, S. *Sex Chromosomes and Sex-linked Genes* (Springer, Berlin, 1967).
- Vogt, P. H. *et al.* Report of the third international workshop on Y chromosome mapping 1997. *Cytogenet. Cell Genet.* **79**, 1–20 (1997).
- Rozen, S. *et al.* Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* **423**, 873–876 (2003).
- Foote, S., Vollrath, D., Hilton, A. & Page, D. C. The human Y chromosome: Overlapping DNA clones spanning the euchromatic region. *Science* **258**, 60–66 (1992).
- Saxena, R. *et al.* Four DAZ genes in two clusters found in AZFc region of human Y chromosome. *Genomics* **67**, 256–267 (2000).
- Tilford, C. *et al.* A physical map of the human Y chromosome. *Nature* **409**, 943–945 (2001).
- Shen, P. *et al.* Population genetic implications from sequence variation in four Y chromosome genes. *Proc. Natl Acad. Sci. USA* **97**, 7354–7359 (2000).
- 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).
- Kuroda-Kawaguchi, T. *et al.* The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. *Nature Genet.* **29**, 279–286 (2001).
- Human Genome Sequencing Consortium Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
- Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
- Schueler, M. G., Higgins, A. W., Rudd, M. K., Gustashaw, K. & Willard, H. F. Genomic and genetic definition of a functional human centromere. *Science* **294**, 109–115 (2001).
- Tyler-Smith, C. *et al.* Localization of DNA sequences required for human centromere function through an analysis of rearranged Y chromosomes. *Nature Genet.* **5**, 368–375 (1993).
- 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).
- Lahn, B. T. & Page, D. C. Functional coherence of the human Y chromosome. *Science* **278**, 675–680 (1997).
- Page, D. C., Harper, M. E., Love, J. & Botstein, D. Occurrence of a transposition from the X-chromosome long arm to the Y-chromosome short arm during human evolution. *Nature* **311**, 119–123 (1984).
- Mumm, S., Molini, B., Terrell, J., Srivastava, A. & Schlessinger, D. Evolutionary features of the 4-Mb Xq21.3 XY homology region revealed by a map at 60-kb resolution. *Genome Res.* **7**, 307–314 (1997).
- Schwartz, A. *et al.* Reconstructing hominid Y evolution: X-homologous block, created by X-Y transposition, was disrupted by Yp inversion through LINE-LINE recombination. *Hum. Mol. Genet.* **7**, 1–11 (1998).
- Tyler-Smith, C., Taylor, L. & Muller, U. Structure of a hypervariably tandemly repeated DNA sequence on the short arm of the human Y chromosome. *J. Mol. Biol.* **203**, 837–848 (1988).
- Graves, J. A. & Schmidt, M. M. Mammalian sex chromosomes: Design or accident? *Curr. Opin. Genet. Dev.* **2**, 890–901 (1992).
- Jegalian, K. & Page, D. C. A proposed path by which genes common to mammalian X and Y chromosomes evolve to become X inactivated. *Nature* **394**, 776–780 (1998).
- Lahn, B. T. & Page, D. C. Four evolutionary strata on the human X chromosome. *Science* **286**, 964–967 (1999).
- Delbridge, M. L., Lingenfelter, P. A., Disteche, C. M. & Graves, J. A. M. The candidate spermatogenesis gene *RBMY* has a homologue on the human X chromosome. *Nature Genet.* **22**, 223–224 (1999).
- Mazeyrat, S., Saut, N., Mattei, M. G. & Mitchell, M. J. *RBMY* evolved on the Y chromosome from a ubiquitously transcribed X-Y identical gene. *Nature Genet.* **22**, 224–226 (1999).
- Saxena, R. *et al.* The DAZ gene cluster on the human Y chromosome arose from an autosomal gene that was transposed, repeatedly amplified and pruned. *Nature Genet.* **14**, 292–299 (1996).
- Lahn, B. T. & Page, D. C. Retroposition of autosomal mRNA yielded testis-specific gene family on human Y chromosome. *Nature Genet.* **21**, 429–433 (1999).
- Szostak, J. W., Orr-Weaver, T. L., Rothstein, R. J. & Stahl, F. W. The double-strand-break repair model for recombination. *Cell* **33**, 25–35 (1983).
- Jackson, J. A. & Fink, G. R. Gene conversion between duplicated genetic elements in yeast. *Nature* **292**, 306–311 (1981).
- The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
- Underhill, P. A. *et al.* Y chromosome sequence variation and the history of human populations. *Nature Genet.* **26**, 358–361 (2000).
- Adams, M. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
- Carvalho, A. B., Dobo, B. A., Vrbancovski, M. D. & Clark, A. G. Identification of five new genes on the Y chromosome of *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **98**, 13225–13230 (2001).
- Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
- Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
- Stankiewicz, P. & Lupski, J. R. Genome architecture, rearrangements and genomic disorders. *Trends Genet.* **18**, 74–82 (2002).
- Ferguson-Smith, M. A. Karyotype-phenotype correlations in gonadal dysgenesis and their bearing on the pathogenesis of malformations. *J. Med. Genet.* **2**, 142–155 (1965).
- Zinn, A. R., Page, D. C. & Fisher, E. M. C. Turner syndrome: The case of the missing sex chromosome. *Trends Genet.* **9**, 90–93 (1993).
- Rao, E. *et al.* Pseudoautosomal deletions encompassing a novel homeobox gene cause growth failure in idiopathic short stature and Turner syndrome. *Nature Genet.* **16**, 54–63 (1997).
- Sun, C. *et al.* Deletion of azoospermia factor A (*AZFa*) region of human Y chromosome caused by recombination between HERV15 proviruses. *Hum. Mol. Genet.* **9**, 2291–2296 (2000).
- Blanco, P. *et al.* Divergent outcomes of intrachromosomal recombination on the human Y chromosome: male infertility and recurrent polymorphism. *J. Med. Genet.* **37**, 752–758 (2000).
- Kamp, C., Hirschmann, P., Voss, H., Huellen, K. & Vogt, P. H. Two long homologous retroviral sequence blocks in proximal Yq11 cause *AZFa* microdeletions as a result of intrachromosomal recombination events. *Hum. Mol. Genet.* **9**, 2563–2572 (2000).
- Repping, S. *et al.* Recombination between palindromes P5 and P1 on the human Y chromosome causes massive deletions and spermatogenic failure. *Am. J. Hum. Genet.* **71**, 906–922 (2002).
- Carrel, L., Cottle, A. A., Goglin, K. C. & Willard, H. F. A first-generation X-inactivation profile of the human X chromosome. *Proc. Natl Acad. Sci. USA* **96**, 14440–14444 (1999).
- Osoegawa, K. *et al.* A bacterial artificial chromosome library for sequencing the complete human genome. *Genome Res.* **11**, 483–496 (2001).
- Zhao, S. *et al.* Human BAC ends quality assessment and sequence analyses. *Genomics* **63**, 321–332 (2000).
- Watson, J. M., Spencer, J. A., Riggs, A. D. & Graves, J. A. Sex chromosome evolution: Platypus gene mapping suggests that part of the human X chromosome was originally autosomal. *Proc. Natl Acad. Sci. USA* **88**, 11256–11260 (1991).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank R. Giardine, R. Oates and S. Silber for patient samples; and J. Alfoldi, C. Disteche, J. Koubova and J. Lange for comments on the manuscript. This work was supported by the National Institutes of Health and the Howard Hughes Medical Institute.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.C.P. (page_admin@wi.mit.edu). GenBank accession numbers are listed in Fig. 2 and the Supplementary Information.