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


Abstract

The human MSY, the male-specific region of the Y chromosome, differentiates the sexes and comprises 95 percent of the chromosome’s length. 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 percent 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 percent of the MSY) where sequence pairs show greater than 99.9 percent 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 20th 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 chromosome. By 1950, studies of human pedigrees reported at least 17 Y-linked traits.

The second era was dominated by the view that the Y chromosome was a genetic wasteland, based on 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 them. In 1959, Jacobs' study of Klinefelter (XXY) males and Ford's research on Turner (XO) females 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 autosomes. He speculated that the X chromosome had retained the ancestral autosome's gene content while 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 chromosome, 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 failure. Genomic studies revealed that the Y chromosome contains a region, comprising 95 percent 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 MSY. The MSY is flanked on both sides by pseudoautosomal regions, where X-Y crossing-over is a normal and frequent event in male meiosis (Supplementary Note 1).

Previous efforts to construct accurate, high-resolution physical maps of the MSY had been stymied by an abundance of lengthy, intra-chromosomal repetitive sequences,
or amplicons. To overcome this difficulty, we identified minute variations between amplicon copies, and then highlighted these minute variants (sequence family variants) 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 efforts. Here we report the sequencing of the MSY.

We mapped and sequenced a tiling path of 220 BAC (bacterial artificial chromosome) 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 population) 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). 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 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 $10^5$ nucleotides, for all MSY euchromatin, with two known exceptions. First, there remain two gaps, each of which is roughly 50 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. In sum, we estimate that we
obtained finished nucleotide sequence for roughly 97% of the MSY euchromatin, and that we captured 99% of the MSY euchromatin's sequence complexity.

Figure 1 A,B, 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 (short arm) pseudoautosomal region to the proximal boundary of the large heterochromatic region of Yq (long arm). 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, LINE-1, and total interspersed repeats. Note: The data graphed 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.
To date, efforts to gain sequence-based understanding of human chromosomes have largely bypassed heterochromatic regions (refs. 14, 15 and Supplementary Note 2), including a large block of heterochromatic sequences found in the centromeric region of every nuclear chromosome\textsuperscript{16}. 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) (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, is comprised of >3,000 tandem repeats of 125 bp, and interrupts the euchromatic sequences of proximal Yq (Figs. 1, 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 Figure 1). In total, we found that the MSY’s heterochromatin encompasses at least six distinct sequence species (Table 3), each of which form long, homogeneous tandem arrays. Our findings are detailed in Supplementary Note 4 and Supplementary Figure 3.

**Homogeneous families predominate in the gene and transcription unit catalogues**

With a comprehensive reference sequence of the MSY in hand, we set out to catalogue systematically the MSY’s genes. We electronically identified and manually examined all matches to previously reported MSY genes. In addition, we used RT-PCR and/or sequencing of cDNA clones to experimentally evaluate electronic matches to publicly available ESTs, as well as potential genes that were predicted using GenomeScan software\textsuperscript{18}. 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 appear very likely to encode proteins (Table 1; Figs. 2, 3; Supplementary Tables 1, 2). All 156 transcription units identified are located in euchromatic sequences. We have no evidence of transcription of the MSY heterochromatin. Of the ~78 protein-coding units, ~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, \textit{RPS4Y1} and \textit{RPS4Y2}, that exhibit 93.6% nucleotide identity in coding exons but are much more diverged in introns.) Thus, the MSY appears to encode at least 27 distinct proteins or protein families.
Table 1. MSY genes and gene families demonstrated or hypothesized to encode proteins.

<table>
<thead>
<tr>
<th>MSY sequence class</th>
<th>Gene symbol</th>
<th>Gene name</th>
<th>Number of copies</th>
<th>Tissue expression</th>
<th>X-linked homologue</th>
<th>Autosomal homologue</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-transposed</td>
<td>TGF2LY †</td>
<td>TGF (beta)-induced transcription factor 2-like Y</td>
<td>1</td>
<td>Testis</td>
<td>TGIF2LY</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PCDH11Y</td>
<td>Protocadherin 11 Y</td>
<td>1</td>
<td>Fetal brain, brain</td>
<td>PCDH11X</td>
<td></td>
</tr>
<tr>
<td>Total:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>X-degenerate</td>
<td>SRY</td>
<td>Sex determining region Y</td>
<td>1</td>
<td>Predominantly testis</td>
<td>SOX3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RPS4Y1</td>
<td>Ribosomal protein S4 Y isoform 1</td>
<td>1</td>
<td>Ubiquitous</td>
<td>RPS4X</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ZFY</td>
<td>Zinc finger Y</td>
<td>1</td>
<td>Ubiquitous</td>
<td>ZFX</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AMELY</td>
<td>Amelogenin Y</td>
<td>1</td>
<td>Teeth</td>
<td>AMELX</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TBL1Y †</td>
<td>Transducin (beta)like 1 protein Y</td>
<td>1</td>
<td>Fetal brain, prostate</td>
<td>TBL1X</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PRKY</td>
<td>Protein kinase Y</td>
<td>1</td>
<td>Ubiquitous</td>
<td>PRKX</td>
<td></td>
</tr>
<tr>
<td></td>
<td>USP9Y</td>
<td>Ubiquitin specific protease 9 Y</td>
<td>1</td>
<td>Ubiquitous</td>
<td>USP9X</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DBY</td>
<td>Dead box Y</td>
<td>1</td>
<td>Ubiquitous</td>
<td>DBX</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UTY</td>
<td>Ubiquitous TPR motif Y</td>
<td>1</td>
<td>Ubiquitous</td>
<td>UTX</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TMSB4Y</td>
<td>Thymosin (beta)4 Y</td>
<td>1</td>
<td>Ubiquitous</td>
<td>TMSB4X</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NLGN4Y</td>
<td>Neurilgin 4 isoform Y</td>
<td>1</td>
<td>Ubiquitous</td>
<td>NLGN4X</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cyorf15A †</td>
<td>Chromosome Y open reading frame 15A</td>
<td>1</td>
<td>Ubiquitous</td>
<td>CXorf15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cyorf15B †</td>
<td>Chromosome Y open reading frame 15B</td>
<td>1</td>
<td>Ubiquitous</td>
<td>CXorf15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SMCY</td>
<td>SMC (mouse) homologue, Y</td>
<td>1</td>
<td>Ubiquitous</td>
<td>SMCX</td>
<td></td>
</tr>
<tr>
<td></td>
<td>EIF1AY</td>
<td>Translation initiation factor 1A Y</td>
<td>1</td>
<td>Ubiquitous</td>
<td>EIF1AX</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RPS4Y2 †</td>
<td>Ribosomal protein S4 Y isoform 2</td>
<td>1</td>
<td>Ubiquitous</td>
<td>RPS4X</td>
<td></td>
</tr>
<tr>
<td>Total:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>16</td>
</tr>
<tr>
<td>Ampliconic</td>
<td>TSPY</td>
<td>Testis specific protein Y</td>
<td>35</td>
<td>Testis</td>
<td>TSPY</td>
<td></td>
</tr>
<tr>
<td></td>
<td>VCY</td>
<td>Variable charge Y</td>
<td>2</td>
<td>Testis</td>
<td>VCY</td>
<td></td>
</tr>
<tr>
<td></td>
<td>XKRY</td>
<td>XK related Y</td>
<td>2</td>
<td>Testis</td>
<td>XKRY</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CDY</td>
<td>Chromodomain Y</td>
<td>4</td>
<td>Testis</td>
<td>CDY</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HSFY †</td>
<td>Heat shock transcription factor Y</td>
<td>2</td>
<td>Testis</td>
<td>HSFY</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RBMY</td>
<td>RNA binding motif Y</td>
<td>6</td>
<td>Testis</td>
<td>RBMY</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PRY</td>
<td>PTP-BL related Y</td>
<td>2</td>
<td>Testis</td>
<td>PRY</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BPY2</td>
<td>Basic protein Y 2</td>
<td>3</td>
<td>Testis</td>
<td>BPY2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DAZ</td>
<td>Deleted in azoospermia</td>
<td>4</td>
<td>Testis</td>
<td>DAZ</td>
<td></td>
</tr>
<tr>
<td>Total:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>~60</td>
</tr>
<tr>
<td>Grand total:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>~78</td>
</tr>
</tbody>
</table>

† Excluding pseudogenes.
* Genes first reported in this study.

A more detailed version of this table, incorporating references and GenBank accession numbers, is available online as Supplementary Table 1.
Appendix

In addition, the MSY includes at least 78 transcription units for which strong evidence of protein coding is lacking; many of these transcription units are likely to be 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 to date.

Based on earlier experiments, the great majority of the MSY's genes 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 testes. Our present catalogue of MSY genes and their patterns of tissue expression (Table 1) corroborate this model. Of the MSY's 27 distinct protein-coding genes or gene families identified to date, 12 are expressed ubiquitously and 11 are expressed exclusively or predominantly in testes.

**X-transposed, X-degenerate and ampliconic 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 three to four million years ago, after the divergence of the human and chimpanzee lineages. 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, 2). 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 1). Thus the X-transposed sequences exhibit the lowest density of genes among the three sequence classes in the MSY euchromatin (Figs. 1, 3), as well as the highest density of interspersed repeat elements (Fig. 1). In particular, LINE-1 elements account for 36% of all X-transposed sequence, or nearly twice the genome average of 20%. As expected, low gene density and high repeat density also characterize the homologous sequence block in Xq21.
Figure 2. Sequence-based map of the MSY; a detailed view of the 24-Mb region shown in Fig. 1b. Indicated via background colouring are 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 through 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 centres of these structures. B, Near-perfect inverted repeats (non-palindromic), three in all (IR1 through 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 novel, experimentally verified transcription units for which cDNA sequencing suggests protein-coding potential (Table 1). Plus (+) strand above, minus (-) strand below. 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 novel, apparently non-transcribed homologues of known coding genes (Supplementary Table 3). J, G+C content (%) calculated in 100-kb sliding window with 1-kb steps. K, Alu, LINE-1, and human endogenous retroviral repeat content, expressed as percentage of nucleotides, calculated in 200-kb sliding window with 1-kb steps. L, 220 BAC clones that were completely or partially sequenced. Each bar represents size and position of one BAC clone, identified by the numeric portion of its GenBank accession number (which in each case begins 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. Note: A more detailed version of this figure, incorporating the locations of STSs and CpG islands, is available online as Supplementary Fig. 2. The composite sequence of the 24-Mb region studied is available as Supplementary File 1.
Figure 3. MSY genes, transcription units, and palindromes. A, Triangles denote sizes and locations of arms of eight palindromes (P1 through 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’ 3’ polarity indicated); pseudogenes are not shown. D, Single-copy, protein-coding genes. E, Single-copy transcription units. F, 15 families of transcription units. These give rise to spliced but apparently non-coding transcripts. G, Merged map of all genes and transcription units.

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 appear 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 appears to be a transcribed, functional gene, and the X- and Y-linked genes encode very similar but non-identical protein isoforms (Table 1; 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, N- and C-terminal portions of the predicted CXORF15 protein (Supplementary Fig. 4). All together, the X-degenerate sequences appear to encode 16 of the MSY's 27 distinct proteins or protein families.

<table>
<thead>
<tr>
<th>Sequence class</th>
<th>Defining characteristics</th>
<th>Evolutionary origins</th>
<th>Aggregate length (Mb)</th>
<th># of coding genes</th>
<th># of non-coding transcription units</th>
<th>% of nucleotides in interspersed repeats</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-transposed</td>
<td>99% identity to X</td>
<td>Single transposition from X</td>
<td>3.4</td>
<td>2</td>
<td>0</td>
<td>0.6</td>
</tr>
<tr>
<td>X-degenerate</td>
<td>Relics of ancient autosome from X and Y evolved</td>
<td>8 blocks on Yp and Yq</td>
<td>8.6</td>
<td>16, most expressed widely</td>
<td>4</td>
<td>2.2</td>
</tr>
<tr>
<td>Ampliconic</td>
<td>Acquired from diverse sources, then amplified</td>
<td>7 blocks on Yp and Yq</td>
<td>10.2</td>
<td>60 (in 9 families), expressed mainly or only in testes</td>
<td>74 (9 single-copy, 65 in 15 families), expressed mainly or only in testes</td>
<td>13.3</td>
</tr>
</tbody>
</table>

**Figure 4.** Three sequence classes in the MSY euchromatin.

Strikingly, 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 kb - 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, 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%. Remarkably, 60% (6.1 Mb) of the ampliconic sequences exhibit intrachromosomal identities of 99.9% or greater.
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 LINE-1 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. Inset in upper right hand corner indicates that the large triangular plot contains two smaller triangles (one revealing sequence similarities within the short arm, Yp, and one revealing similarities within the long arm, 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). Employing 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 prior to analysis.) At each point along the length of the MSY, the highest sequence similarity (expressed as % nucleotide identity) was identified. All such values >50% are shown. An expanded version of this plot is shown in Fig. 2f.

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 LINE-1 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 diagrammed 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 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, 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 1; 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. In addition, the ampliconic sequences include at least 75 other transcription units for which strong evidence of protein coding is lacking (Figs. 2, 3; 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 to date.

In contrast to the ubiquitous expression of most X-degenerate genes, the ampliconic genes and transcription units show highly restricted expression (Table 1). All nine protein-coding families in the ampliconic regions are expressed predominantly or exclusively in testes, as are the great majority of the regions’ non-coding transcription units.

Among the three euchromatic sequence classes, the ampliconic sequences exhibit by far the lowest densities of LINE-1 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 << 0.000001).

**Eight massive palindromes comprising 25% of MSY euchromatin**

The most pronounced structural features of Yq’s ampliconic regions are eight massive palindromes (Table 2). 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% to 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 to 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 arms. 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 appear 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 to date 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 to 298 kb (Fig. 2; Supplementary Table 4). Three of these inverted repeat pairs (IR1, IR2, and IR3) exhibit nucleotide identities of 99.66% to 99.95%. Inversion of the IR3 repeats, both located on Yp, was likely 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, 6)

Transcriptionally active tandem arrays, both non-coding and coding

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 TTTY1, TTTY2, TTTY6, TTTY7, TTTY8, TTTY18, TTTY19, TTTY21, and TTTY22 families. Both strands of the NORF arrays are transcribed; 3' portions of the TTTY1 and TTTY2 transcripts are complementary (Supplementary Fig. 8).
Table 2. MSY palindromes.

<table>
<thead>
<tr>
<th>Palindrome</th>
<th>Arm length (kb)</th>
<th>Arm-to-arm identity (%)</th>
<th>Spacer length (kb)</th>
<th>Palindrome span (kb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>1450</td>
<td>99.97</td>
<td>2.1</td>
<td>2902</td>
</tr>
<tr>
<td>P1.1’</td>
<td>9.9</td>
<td>99.95</td>
<td>3.9</td>
<td>24</td>
</tr>
<tr>
<td>P1.2’</td>
<td>9.9</td>
<td>99.95</td>
<td>3.9</td>
<td>24</td>
</tr>
<tr>
<td>P2</td>
<td>122</td>
<td>99.97</td>
<td>2.1</td>
<td>246</td>
</tr>
<tr>
<td>P3</td>
<td>283</td>
<td>99.94</td>
<td>170</td>
<td>736</td>
</tr>
<tr>
<td>P4</td>
<td>190</td>
<td>99.98</td>
<td>40</td>
<td>419</td>
</tr>
<tr>
<td>P5</td>
<td>496</td>
<td>99.98</td>
<td>3.5</td>
<td>996</td>
</tr>
<tr>
<td>P6</td>
<td>110</td>
<td>99.97</td>
<td>46</td>
<td>266</td>
</tr>
<tr>
<td>P7</td>
<td>8.7</td>
<td>99.97</td>
<td>12.6</td>
<td>30</td>
</tr>
<tr>
<td>P8</td>
<td>36</td>
<td>99.997</td>
<td>3.4</td>
<td>75</td>
</tr>
</tbody>
</table>

Total span: 5670

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

The TSPY arrays are based on a 20.4-kb repeat unit that encodes, on one strand, a previously identified protein, TSPY. A newly identified transcription unit, CYorf16, is found on the opposite 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%. In addition, 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 to date in the human genome.

The evolution of the MSY

Based on our present findings and previous studies, we propose a model of MSY evolution that addresses all three euchromatic sequence classes (Figs. 6, 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 hypothesize 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. 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. Our present sequencing of the MSY euchromatin expands the catalogue of known X-Y gene pair fossils, providing opportunity to reexamine 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 Ks. They reasoned that X-Y differentiation would have begun only after X-Y crossing-over ceased. They observed a strong correlation between the age (Ks) of individual X-Y gene pairs and the locations of their X members on the human X chromosome. Among the 19 X-Y gene pairs studied, age increased in 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.
Table 3. Six sequence species in three MSY heterochromatic regions.

<table>
<thead>
<tr>
<th>Locus symbol</th>
<th>Primary</th>
<th>Unit repeat</th>
<th>Secondary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Centromeric region</td>
<td>DYZ3</td>
<td>171 bp aliphoid</td>
<td>5941 bp</td>
</tr>
<tr>
<td></td>
<td>DYZ17*</td>
<td>GGAAT</td>
<td>none observed</td>
</tr>
<tr>
<td>Proximal Yq (Yq11.22)</td>
<td>DYZ19*</td>
<td>125 bp</td>
<td>none observed</td>
</tr>
<tr>
<td>Distal Yq (Yq12)</td>
<td>DYZ18*</td>
<td>GGAAT</td>
<td>2864 bp</td>
</tr>
<tr>
<td></td>
<td>DYZ1</td>
<td>GGAAT</td>
<td>3584 bp</td>
</tr>
<tr>
<td></td>
<td>DYZ2</td>
<td>insufficient sequence available; AT rich</td>
<td>~2470 bp *</td>
</tr>
</tbody>
</table>

* First reported in this study.
' Based on restriction endonuclease digestion, not sequence analysis.
A more detailed version of this table, incorporating references and the sequences of repeat units, is available online as Supplementary Table 10.

We re-analyzed this published information and combined the results with Ks 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 (Ks) and X map position is readily apparent, with age increasing from the distal short arm to the long arm of the X (Fig. 7). In addition, as observed in the earlier study, the order of the homologous genes in the MSY appears to be scrambled with respect to Ks (Supplementary Fig. 9). These observations, together with the earlier arguments of Lahn and Page, suggest three conclusions. First, all MSY genes and pseudogenes identified here as “X-degenerate” appear 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, 28). Third, as previously hypothesized, inversions in the Y chromosome may have suppressed crossing-over with the X chromosome.
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 genes\(^a\); 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.

**X-transposed genes as exceptions.** A very different evolutionary scenario 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 appear to be orders of magnitude younger than the ancient pairs (Group 1 in Fig. 7) among which they are physically situated in the X.
**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 appear less distinct (Fig. 7; compare with Fig. 2 in ref. 26). While 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 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 upon 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 $KAL1/P$ 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 $KAL1/P$ 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 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 the 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 1). 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 striking 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 likely 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 likely 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 while the Y chromosome lost them\(^5\). Our findings support the two major premises of this theory: the evolutionary genetic benefits of sexual recombination via 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 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 (Fig. 3, Table 1) 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, like the X-degenerate genes, derived from common ancestors of the X and Y chromosomes. 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/autosome similarities suggests that a series of autosomal transpositions contributed to the MSY’s ampliconic sequences during primate evolution (Fig. 8; ref. 13). Yet another molecular mechanism accounts for the CDY genes, which arose by retroposition (and subsequent amplification) of a processed mRNA derived from an autosomal gene. This retroposition event was previously thought to have occurred during primate evolution, but our present $K_s$ analysis indicates a much older date, likely before the lineages of marsupials and placental mammals diverged (Supplementary Table 5 and Fig. 8).

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.

**Hypothesis: abundant Y-Y gene conversion in ampliconic regions**

Gene conversion is the non-reciprocal transfer of sequence information from one DNA duplex to another. 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 mitosis. 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 percent of the MSY euchromatin, including nearly all of the MSY's testis-specific gene families. The accompanying study reports the identification and sequencing of chimpanzee Y-linked orthologs 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.

Earlier in this article, 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 - 100.0%. As intrachromosomal identity declines below 99.9%, the fractional representation of MSY sequences drops abruptly. Thus, the sequences displaying intrachromosomal identities of ≥99.9% represent a large and distinct subset of the MSY euchromatin.

This “≥99.9%” subset comprises the eight palindromes as well as large portions of the IR2 and IR3 inverted repeats described earlier in this article (Figs. 2, 3). Indeed, nearly all of the “≥99.9%” sequences exist as pairs in inverted orientation. Thus, the MSY palindromes in which gene conversion has been demonstrated are typical and representative of the “≥99.9%” fraction. We extrapolate that nearly all of the “≥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.
Appendix

### Table 4. Numbers of MSY testis gene family members (intact genes and pseudogenes) in regions exhibiting ≥99.9% or <99.9% intrachromosomal similarity*.

<table>
<thead>
<tr>
<th>Genes</th>
<th>Pseudogenes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>in regions of ≥99.9% in regions of &lt;99.9% intrachromosomal similarity</td>
</tr>
<tr>
<td>VCY</td>
<td>2</td>
</tr>
<tr>
<td>XKRY</td>
<td>2</td>
</tr>
<tr>
<td>CDY</td>
<td>4</td>
</tr>
<tr>
<td>HSFY</td>
<td>2</td>
</tr>
<tr>
<td>RBMY</td>
<td>6</td>
</tr>
<tr>
<td>PRY</td>
<td>2</td>
</tr>
<tr>
<td>BPY2</td>
<td>3</td>
</tr>
<tr>
<td>DAZ</td>
<td>4</td>
</tr>
<tr>
<td>Totals</td>
<td>25</td>
</tr>
</tbody>
</table>

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

---

**Figure 9.** MSY sequences exhibiting ≥99.9% intrachromosomal identity likely undergo Y-Y gene conversion. **A,** Electronic fractionation of MSY euchromatic sequences according to intrachromosomal similarity (% 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 top is 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 ≥99.9%.
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 “≥99.9%” regions (aggregate length 6.1 Mb) dispersed throughout the MSY (Fig. 9b). 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 report,7 steady-state calculations suggest that, on average, multiple Y-Y gene conversion events take place per generation in the MSY. Thus, the majority of homologous recombination events in the Y chromosome probably occur in the MSY.

Ironically, 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 “≥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 “≥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 “≥99.9%” fraction (Table 4). While large numbers of pseudogenes are present both inside and outside the ≥99.9% fraction, the intact gene copies, 25 in all, are located exclusively in the ≥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 to date, 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 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 dramatically 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 populations11,34.

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 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 sequence35,36, 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 novel 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 employed recently in sequencing large genomes: whole-genome shotgun analysis\textsuperscript{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 ≥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\textsuperscript{38,39}. The iterative mapping-and-sequencing strategy employed 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 rejection\textsuperscript{6}. 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 ("XO") karyotype. Haploinsufficiency of particular genes common to the X and Y chromosomes may be responsible for somatic features of the syndrome\textsuperscript{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; Table 1).

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\textsuperscript{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\textsuperscript{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).
Appendix

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 cells. 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 employed here has been described\textsuperscript{10,13}. All MSY BACs selected for sequencing were isolated from the RPCI-11 library\textsuperscript{48}, with the exception of 11 clones (nine spanning the AZFa region\textsuperscript{12}, and two used to narrow gaps\textsuperscript{10}) 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 construction\textsuperscript{49}. 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 10 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.umdnj.edu/nigms/maps/mapping.html).
Identification of novel genes and transcription units

We identified potential transcripts from three sources:

- 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.
- 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 library.
- GenomeScan predictions in the NCBI annotation of Y-chromosome contigs.

We then tested for transcription by RT-PCR as previously described.

Chromosomal fluorescence in situ hybridization (FISH)

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

Calculation of Ks and Ka

We calculated the numbers of synonymous substitutions per synonymous site (Ks) and of nonsynonymous substitutions per nonsynonymous site (Ka) 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 Supplementary File 3). Then, as is standard practice, insertions/deletions were manually removed from the alignments. We calculated Ks and Ka for these alignments using the Diverge function in the Wisconsin Package (Version 10.2, Genetics Computer Group, Madison, Wisconsin).

Supplementary Information

can be found on Nature's website (http://www.nature.com).

Acknowledgements

We thank R. Giardine, R. Oates, and S. Silber for patient samples; and J. Alfoldi, C. Distetche, J. Koubova, and J. Lange for comments on the manuscript. Supported by the National Institutes of Health and the Howard Hughes Medical Institute. Correspondence and requests for materials should be addressed to D.C.P. (e-mail: page_admin@wi.mit.edu).
References


