The *AZFc* region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men

Tomoko Kuroda-Kawaguchi[1], Helen Skaletsky[1], Laura G. Brown[1], Patrick J. Minx[2], Holland S. Cordum[2], Robert H. Waterston[2], Richard K. Wilson[2], Sherman Silber[3], Robert Oates[4], Steve Rozen[1] & David C. Page[1]

Published online: 22 October 2001, DOI: 10.1038/ng757

Deletions of the *AZFc* (azoospermia factor c) region of the Y chromosome are the most common known cause of spermatogenic failure. We determined the complete nucleotide sequence of *AZFc* by identifying and distinguishing between near-identical amplicons (massive repeat units) using an iterative mapping–sequencing process. A complex of three palindromes, the largest spanning 3 Mb with 99.97% identity between its arms, encompasses the *AZFc* region. The palindromes are constructed from six distinct families of amplicons, with unit lengths of 115–678 kb, and may have resulted from tandem duplication and inversion during primate evolution. The palindromic complex contains 11 families of transcription units, all expressed in testis. Deletions of *AZFc* that cause infertility are remarkably uniform, spanning a 3.5-Mb segment and bounded by 229-kb direct repeats that probably served as substrates for homologous recombination.

Introduction

At least one in every ten couples of reproductive age is unable to bear children despite an extended period of unprotected sexual intercourse[1]. In recent years, there has been an intensive search for genetic causes of infertility in both men and women. Spermatogenic failure is the most common form of male infertility. Here the most striking findings come from studies of the long arm of the Y chromosome (Yq). It is now widely accepted that deletion of any one of three Yq regions (*AZFa*, *AZFb* or *AZFc*) severely diminishes or extinguishes sperm production[2–6].

Among the three regions, deletions occur most frequently in *AZFc*, with *de novo* deletions arising in roughly 1 in 4,000 males[3–8] (L.G.B. *et al.*, unpublished data). *AZFc* is the region most commonly deleted in men with azoospermia, accounting for about 12% of cases of nonobstructive azoospermia (no sperm detected in semen) and about 6% of cases of severe oligozoospermia (sperm count less than 5 million/ml).

Despite the biological and medical importance of *AZFc*, efforts to develop physical maps are hindered by the region's unusually repetitive sequence composition. This difficulty was evident in the first STS-based physical map of the Y chromosome, where STSs could not be accurately ordered and YACs could not be assembled into robust contigs in *AZFc* and surrounding regions (deletion intervals 6D–6F)[9]. More recently, a painstaking effort to construct a long-range restriction map of *AZFc* left central questions unanswered, including the overall size of the region and the organization of gene-bearing repeats[10]. Similarly, studies of the *DAZ* genes in this region led to widely varying conclusions about their number and arrangement of these genes, which show 99.9% sequence identity to one another[4,10–13]. In the case of the *RBMY* gene family of the Y chromosome, there have been conflicting reports as to whether *AZFc* contains all, some or none of the functional copies. It has been unclear whether *RBMY* contributes to the spermatogenic failure caused by *AZFc* deletions[3,4,10,14–16]. Thus, the legacy of past studies suggested that identifying single-copy DNA markers, localizing deletion breakpoints, accurately sizing *AZFc* deletions (estimates ranged from 0.5–2 Mb[4,10]) and precisely cataloging the genes encompassed by *AZFc* deletions would be difficult or impossible because of the region's repetitive nature.

Here we examine key issues posed by this technically complex region of the Y chromosome: (i) what is the region's structure (ii) how are its repetitive sequences organized and (iii) how does this organization account for the region's role in fertility and its propensity to deletion? To address these questions, we employed an iterative mapping–sequencing strategy and examined the resulting structure in light of data from infertile patients.

Results

Analysis of sequence-family variants on a single Y chromosome

We and our colleagues previously sequenced portions of the *DAZ* genes in *AZFc* and showed that *DAZ* gene copies on the same Y chromosome are 99.9% identical in introns, exons and flanking sequences[11]. This indicated that *AZFc* contains repeats of considerable unit length (amplicons), and that individual

[1]Howard Hughes Medical Institute, Whitehead Institute, and Department of Biology, Massachusetts Institute of Technology, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA. [2]Genome Sequencing Center, Washington University School of Medicine, St. Louis, Missouri 63108, USA. [3]Infertility Center of St. Louis, St. Luke's Hospital, St. Louis, Missouri 63107, USA. [4]Department of Urology, Boston University School of Medicine, Boston, Massachusetts, 02228 USA. Correspondence should be addressed to D.P. (e-mail: page_admin@wi.mit.edu).

amplicon copies might be indistinguishable from one another by conventional mapping methods (such as analysis of sequence-tagged site (STS) content and fingerprint analysis of BACs) because they yield incomplete comparisons of DNA sequences. Indeed, *AZFc*'s amplicon copies might be as similar in sequence as two Y chromosomes chosen at random from the population (~99.97% identity)[17].

These lessons from the *DAZ* amplicons led to a strategy for sequencing the entire *AZFc* region. First, we would use BAC clones from one man's Y chromosome (the RPCI-11 library, BACPAC Resources)[18] to obtain all *AZFc* sequencing templates. We would thus avoid polymorphism, which might hinder our assembly of maps and sequences[19]. Second, we would map *AZFc* and sequence it as coupled and iterative processes, to distinguish and accurately place individual members of amplicon families. Sequencing of BACs selected from crude contigs assembled by conventional mapping methods would lead to map refinement, which in turn would enable selection of additional BACs for sequencing. Third, we would base the finished sequence of *AZFc* on a set of BACs that displayed substantial overlaps, which would allow us to verify that we had distinguished and accurately placed nearly identical amplicon copies.

Through the analysis of STS content, we initially organized 220 *AZFc* region BACs[19] into bins and hypothetical contigs and selected a hypothetical tiling path of BACs for sequencing. At many subsequent points, we revised the BAC map and selected additional BACs for sequencing based on our analysis of sequence-family variants (SFVs). SFVs are defined as subtle differences between closely related but nonallelic sequences[13,19], such as single-nucleotide substitutions or dinucleotide repeat–length variants. Because the *AZFc*-region BACs were derived from one man's Y chromosome, sequence differences

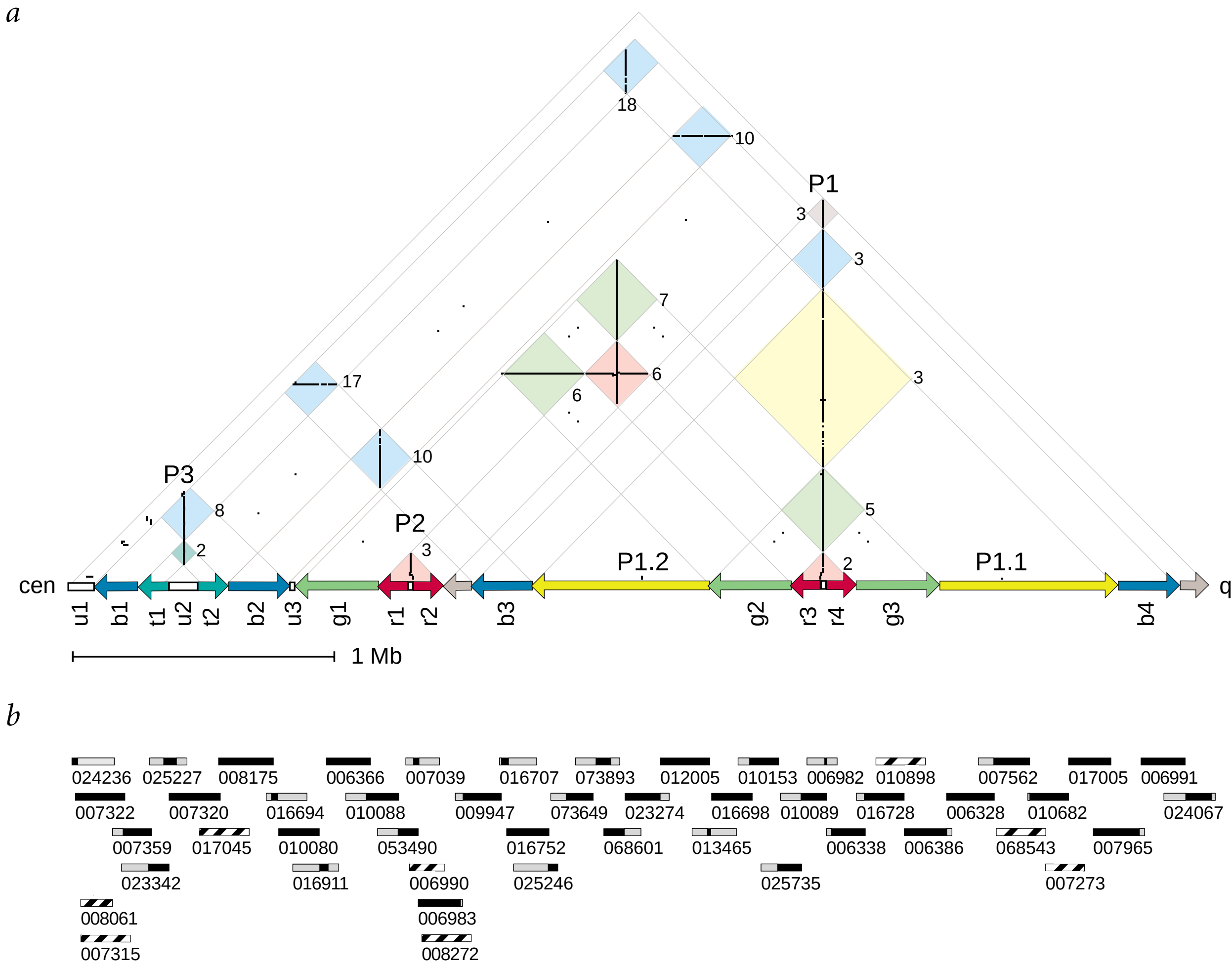

Fig. 1 Amplicons and palindromes in 4.5-Mb portion of the human Y chromosome that includes *AZFc*. ***a***, Dot-plot in which region's sequence is compared with itself. This compressed, triangular plot avoids redundancies and false symmetries that would appear in a conventional square plot. Amplicons of various families (blue, turquoise, green, red, gray and yellow arrows) are shown at the base of the plot. Each dot within the plot represents a perfect match of 500 bp. Direct repeats are shown as horizontal lines, inverted repeats as vertical lines and palindromes as vertical lines that nearly intersect the baseline. Palindromes P1, P2 and P3 are indicated, as are two smaller palindromes, P1.1 and P1.2, that lie within P1. Tinted squares reflect pairs of amplicons with nearly identical sequences; the adjacent numbers denote substitutions per 10 kb. Gray diagonal lines mark amplicon boundaries. Selected amplicons (b1, t1, t2, b2, g1, r1, r2, b3, g2, r3, r4, g3 and b4) are identified, as are sequences that occur once within the 4.5-Mb region (u1, u2 and u3). Orientation with respect to the centromere (cen) and long-arm telomere (qter) is shown. The full annotated sequence of the 4.5-Mb region is available in Web Tables A, B and C and Web Note A. ***b***, BAC clones (from RPCI-11 library[18]) that have been completely or partially sequenced. Each bar represents the size and position of one BAC clone, identified by the numeric portion of its GenBank accession number (the alphabetic portion is omitted in each case; all accession numbers begin with the prefix 'AC'). Black bars represent finished sequence, deposited in GenBank, where finished sequences are trimmed to retain only 200 bp of overlap with adjoining BACs. Gray bars represent the 'trimmings' of those BACs not deposited in GenBank. Complete BAC sequences, with no trimming of overlaps, were essential in establishing and validating the BAC order shown. Untrimmed sequences for all BACs are available at http://staffa.wi.mit.edu/page/Y/azfc.

between BACs could not represent allelic variants. We identified SFVs by comparing the sequences of BACs that initially appeared to overlap (as judged by STS content) but actually were derived from different amplicon copies. Using this approach, we disentangled lengthy amplicon copies that differed only slightly in DNA sequence. We ultimately assembled a 4.5-Mb sequence contig spanning the entirety of *AZFc*, based on 48 sequenced BACs (Fig. 1).

By design, the set of 48 sequenced BACs shows substantial redundancy, with overlaps between sequenced BACs totaling approximately 3 Mb (Fig. 1*b*). Nearly two-thirds of the nucleotides in the 4.5-Mb region were sequenced in two independent BAC clones. This allowed us to validate the 4.5-Mb sequence contig by exhaustively investigating all sequence discrepancies between overlapping BACs. We observed eight such discrepancies. In six of eight cases, re-sequencing of the discrepant site in new cultures of the same BACs resolved the discrepancies, suggesting that they were the result of mutations arising within the initial BAC cultures. In the remaining two cases, new BAC cultures failed to resolve the discrepancy. In these two cases, we examined all other RPCI-11 BACs predicted to contain the discrepant nucleotide, and in both cases the variant was restricted to a single BAC. We concluded that these two discrepancies were due either to mutations in the BACs or to somatic mutations in the human donor. These two possibilities cannot be experimentally distinguished because no cell line or genomic DNA is available from the RPCI-11 donor. This analysis of overlaps suggests that mutations in the BACs (and, possibly, somatic mutations in the donor) are the major source of error in the 4.5-Mb sequence and that undetected errors of this kind are on the order of 1–3 per megabase.

Amplicons and symmetries revealed by genomic sequence analysis

Examination of the *AZFc* sequence shows symmetries of unprecedented scale and precision. Fig. 1*a* is a compressed representation of a dot-plot analysis in which only perfect matches of at least 500 bp are scored. Inspection of the plot shows the following:

(i) There are six distinct families of amplicons (massive repeat units) in the region sequenced. The amplicon units range in length from 115 kb (gray) to 678 kb (yellow). The turquoise, gray and yellow amplicons each occur twice in the region sequenced, whereas the green amplicon occurs three times and the blue and red amplicons each occur four times. Together, the six amplicon families account for 93% (all but 313 kb) of the 4.5-Mb sequence shown in Fig. 1.

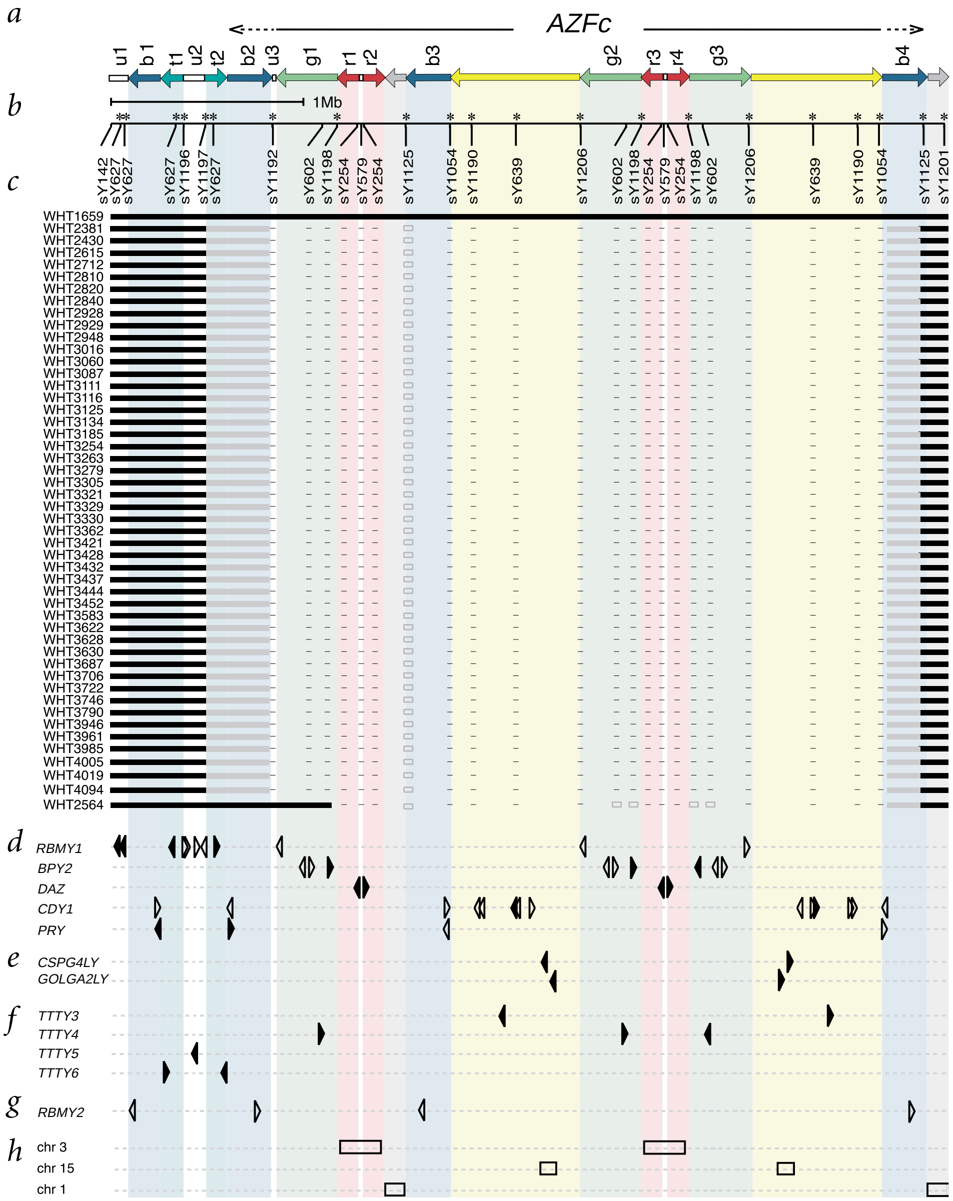

Fig. 2 Sequence-based map of STSs, deletions, transcription units and autosomal homologies. ***a***, Amplicon map (from Fig. 1*a*). Region of recurrent *AZFc* deletions is shown; precise locations of deletion endpoints within amplicons b2 and b4 are not known, as indicated by dotted lines. ***b***, STSs employed in characterizing deletions. Asterisks denote new STSs. ***c***, Results of testing genomic DNAs from a normal, fertile man (WHT1659) and 48 infertile men with *AZFc* deletions (WHT2381–WHT2564) for the presence or absence of STSs. Solid black bars encompass STSs that were present; minus signs represent absent STSs. Solid gray bars indicate that the presence or absence of sY627 in amplicon t2 cannot be determined with confidence because of cross-amplifying sY627 loci elsewhere in the palindromic complex. Open boxes represent positive PCR results that we dismiss because of (i) cross-amplifying loci elsewhere in the region and (ii) negative results with flanking STSs. ***d–g***, Transcription units (solid triangles) and selected pseudogenes (open triangles). Direction of triangles indicates 5′→3′ polarity. ***d***, Five gene families previously reported to have protein-coding potential. ***e***, Two families of transcription units with significant predicted open reading frames. ***f***, Four families of spliced but apparently noncoding transcripts. ***g***, *RBMY2* sequences. ***h***, Regions of homology to human chromosomes 3, 15 and 1. Homology to chromosome 1 extends distally beyond the region shown here.

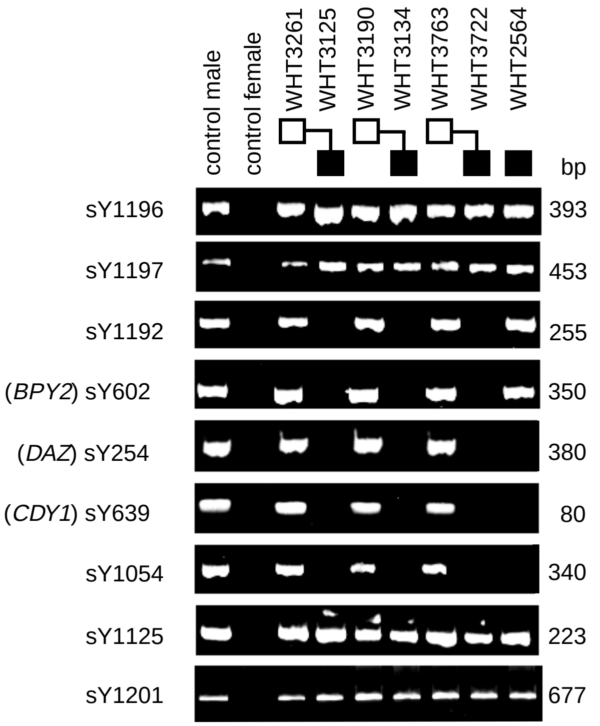

Fig. 3 PCR assays of selected STSs in the DNA of four men with *AZFc* deletions (and, in three cases, their fathers). Three STSs correspond to *AZFc* genes, as listed. PCR product sizes are indicated at right.

(ii) The members of each amplicon family are nearly identical, with sequence divergences ranging from 2 or 3 substitutions per 10 kb (99.98% identity in certain turquoise, red, yellow, blue and gray amplicons) to 17 or 18 substitutions per 10 kb (99.82% identity in some blue amplicons). By convention, these percentage identities refer only to nucleotide substitutions and do not take account of insertions and deletions by which amplicon family members differ.

(iii) The amplicons are arrayed to form six large inverted repeats (represented by vertical lines and with a total length of 2.8 Mb) and three large direct repeats (represented by horizontal lines and with a total length of 1.0 Mb).

(iv) Three of the six inverted repeats are palindromes, or near-palindromes, with large, inverted duplications bracketing much shorter intervening sequences. Notably, palindrome P1 has an arm length of 1.5 Mb, a span of 3 Mb and arm-to-arm identity of 99.97%. Within the arms of P1 lie two smaller palindromes, P1.1 and P1.2, each spanning 24 kb. Together, palindromes P1, P2 and P3 encompass 4.0 of the 4.5 Mb of sequence.

(v) There are no single-copy sequences in the 4.5-Mb region. The uncolored segments u1, u2 and u3 occur once each in this region. However, u1 shows 70–85% identity to a locus on Yp, u2 shows 70–90% identity to an interspersed Y-specific repeated locus and u3 falls within a 65-kb block of 99.7% identity to Yp. The uncolored 2-kb segments at the centers of the P1 and P2 palindromes are identical to each other.

Uniform recurrent deletions causing spermatogenic failure

Using the genomic sequence of *AZFc* and surrounding areas as a guide, we measured the length and localized the boundaries of *AZFc* deletions found in men with spermatogenic failure. We first developed several new PCR-based assays for the presence or absence of important sequence landmarks (STSs; Figs. 2*b* and 3). Using these new assays, we studied 48 infertile men with interstitial Yq deletions limited to *AZFc*. These 48 individuals had been identified during the course of screening azoospermic or severely oligospermic men for Y-chromosome deletions[4,20–22]. Specifically, these 48 individuals lack sY254 (a *DAZ* STS[4]) but possess both sY142 (proximal to *AZFc*) and sY160 (distal to *AZFc*).

Testing for the new STSs indicated that all but one of the 48 *AZFc* deletions share similar if not identical breakpoints, both distal and proximal. In 47 of the 48 men, the proximal boundary of the deletion falls within a 349-kb region bounded by STSs sY1192 and sY1197 (Figs. 2 and 3). In all 48 men, the distal boundary falls within a 229-kb region bounded by sY1054 and sY1125 (Figs. 2 and 3).

This clustering of breakpoints suggests a mechanism of *AZFc* deletion. On the Y chromosome and throughout the human genome, large deletions often seem to result from homologous recombination between direct repeat sequences[23–26]. This is probably true for *AZFc* deletions, because the region in which the proximal breakpoints cluster is strikingly similar in sequence to the region in which the distal breakpoints cluster. Indeed, the proximal and distal breakpoint regions largely correspond to two members (b2 and b4, respectively) of the blue amplicon family in Fig. 2. Amplicons b2 and b4 are separated by more than 3 Mb on the intact Y chromosome, where they are arrayed as direct, 229-kb repeats flanking *AZFc*. Amplicons b2 and b4 show 99.9% nucleotide identity. The high similarity among the blue amplicons b1, b2 and b4, and between the turquoise amplicons t1 and t2, precludes our placing more precisely the proximal and distal breakpoints in the 48 men with *AZFc* deletions. In light of previous reports[3–8,23–26], these findings suggest that homologous recombination between amplicons b2 and b4 is a frequent cause of spermatogenic failure in human populations. The sequence-based map and PCR assays described here should facilitate testing of this hypothesis by other investigators. Another mechanism may be responsible for the deletion in one of the 48 males (WHT2564), whose proximal breakpoint does not lie within the b2 amplicon (Fig. 2).

Assuming that homologous recombination occurs between amplicons b2 and b4, we estimate the size of the *AZFc* deletions (in 47 of the 48 men studied) to be 3.5 Mb. Without the assumption of homologous recombination, we estimate the size of those 47 *AZFc* deletions to be 3.3–3.8 Mb.

Intermingled families of testis-specific transcription units

Using the complete, 4.5-Mb DNA sequence of the palindromic complex including *AZFc*, we set out to catalog systematically the complex's genes. We electronically identified and scrutinized all matches to previously reported Y-chromosomal genes. In addition, we used RT–PCR analysis and/or sequencing of cDNA clones to experimentally investigate both (i) electronic matches to publicly deposited ESTs and (ii) potential genes predicted using GenScan software[27].

We found that the palindromic complex contains 11 families of transcription units, most of which seem to be expressed exclusively or most abundantly in testis (Fig. 2*d*–*f* and Web Table D).The complex contains 15 apparently intact genes that represent five families previously reported to have protein-coding potential: *RBMY1*, *PRY*, *BPY2*, *DAZ* and *CDY1* (refs. 4,14,28). The region also contains a total of 21 *RBMY*, *PRY*, *BPY2* or *CDY* pseudogenes. In addition, we identified six new families of transcription units, including two families with significant open reading frames (*CSPG4LY* and *GOLGA2LY*) and four families of spliced but apparently noncoding transcripts (*TTTY3*, *TTTY4*, *TTTY5* and *TTTY6*). The palindromic complex also carries four genes or pseudogenes of the previously

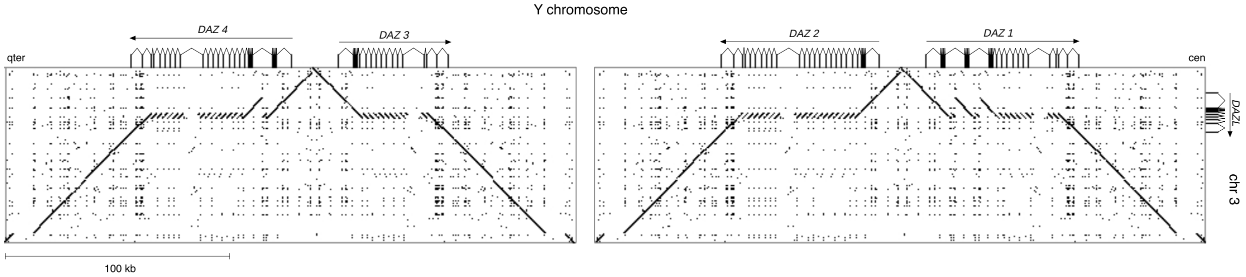

Fig. 4 Comparison of 78-kb segment of human chromosome 3 with two segments of Y chromosome totaling 540 kb. In this conventional (rectangular) dot-plot, each dot represents a perfect match of 15 bp. Structures of *DAZL* (on chromosome 3) and of the four *DAZ* genes are indicated. Orientation of Y-chromosomal sequences with respect to centromere (cen) and long-arm telomere (qter) is shown; orientation of chromosome 3 sequence is not known.

reported *RBMY2* family[14,29]; we have not detected transcription of these *RBMY2* sequences.

Of the 11 families of experimentally verified transcription units in the 4.5-Mb palindromic complex, 7 are located exclusively within the 3.5-Mb *AZFc* region proper (defined by *AZFc* deletions; Fig 2*c*): *BPY2*, *DAZ*, *CDY1*, *CSPG4LY*, *GOLGA2LY*, *TTTY3* and *TTTY4*. Spermatogenic failure in men with *AZFc* deletions is probably caused by the absence of one or more of these seven families, all of which are transcribed in testis. The palindromic complex's four other gene families, *RBMY1*, *PRY*, *TTTY5* and *TTTY6*, may be involved in spermatogenic failure caused by deletions of the *AZFb* region, which has been reported to lie just proximal to *AZFc*[5].

Discussion

Structure of the palindromic complex. We know of no other genomic region in any organism in which intermingled, hierarchically organized amplicons and palindromes have such scale, copy-to-copy uniformity and elaborate mosaicism (Fig. 1). Six distinct families of amplicons, with an aggregate nucleotide complexity of 1.6 Mb, comprise nearly the whole region. They are the building blocks of three palindromes, or near-palindromes, with an aggregate span of 4.0 Mb.

The region's most notable landmark, the P1 palindrome, shows 99.97% arm-to-arm identity across a span of 3.0 Mb, dwarfing all previously described inverted repeats. The next largest known inverted repeats are about one-third this size, and they are found on human chromosome 5, in the spinal muscular atrophy (*SMA*) region[30]. These repeats complicated positional cloning of the SMA gene in the mid-1990s[30,31] and continue to hamper sequencing of the *SMA* region by conventional methods[32]. Other, smaller inverted repeats have also been described in the human genome, but none rivals the size and symmetry of the P1 palindrome.

It is apparent why all previous efforts, relying on techniques that included STS content analysis, restriction mapping, Southern blot analysis and fluorescence *in situ* hybridization, failed to decipher the organization of *AZFc*. The techniques used in previous studies could not distinguish amplicons that have the size and sequence similarity found in *AZFc* (Fig 1*a*). The extraordinary structure of the *AZFc* complex can be perceived and understood only when accurate genomic sequencing of clones drawn from a single Y chromosome and analysis of SFVs identified through such sequencing are fully integrated with conventional mapping methods.

Uniform recurrent deletions and homologous recombination. The highly recurrent nature of *AZFc* deletions is apparently a consequence of *AZFc*'s amplicon structure. The arrangement, similarity and size of the b2 and b4 amplicons offer ample opportunity for deletion via homologous recombination (Fig. 2). Although both *AZFc* and *AZFa* deletions appear to result from homologous recombination between direct repeats on the Y chromosome, *AZFc* deletions seem to arise far more frequently in human populations around the world[3–8]. For example, in a study of 823 men with azoospermia or severe oligozoospermia (all ascertained through infertility clinics), we identified 54 *AZFc* deletions (48 of these deletions are shown in Fig. 2) but just one *AZFa* deletion[23,33]. These findings probably reflect the relative rates of *de novo* deletion of *AZFc* and *AZFa*. As both *AZFc* and *AZFa* deletions appear to affect spermatogenesis exclusively and dramatically, these deletions should be ascertained with similar efficiencies through infertility clinics. Nearly all *AZFc* and *AZFa* deletions examined have proven to be *de novo*[4,5,8,20,33]. The relative frequency of the deletions reflects the relative size of the corresponding targets for homologous recombination. The b2 and b4 amplicons that flank the *AZFc* region are 229 kb long (Figs. 1 and 2), whereas the direct repeats flanking the *AZFa* region are just 10 kb long[23–25].

The *AZFc* complex contains many amplicon pairs in addition to b2/b4 (Fig. 2). Some of these may prove to be targets for homologous recombination, resulting in deletions (or inversions or duplications) distinct from the canonical *AZFc* deletion described here. In addition, interstitial Yq deletions (associated with spermatogenic failure) that include but extend beyond *AZFc* have been reported[8,22,34]. Whether these '*AZFc*-plus' deletions are the result of homologous recombination between direct repeats remains to be investigated.

Functional specialization. The palindromic complex that encompasses *AZFc* is extraordinary not only in its structure but also in its functional specialization. Whereas roughly half of the genes on the Y chromosome are ubiquitously expressed[28], ubiquitously expressed genes seem to be excluded from the 4.5-Mb palindromic complex that includes *AZFc*. The complex includes approximately 27 transcription units belonging to 11 spatially intermingled families (Fig. 2). All of the identified transcription units seem to be expressed predominantly or exclusively in testis[4,14,28]. In males, nullisomy for the 3.5-Mb *AZFc* region proper (including 19 transcription units comprising seven distinct families) seems to affect only one cell lineage (spermatogenic) and only one aspect of male differentiation (spermatogenesis). Apart from spermatogenic failure, boys and men with *AZFc* deletions appear to be healthy[4,35]. Many contiguous gene deletions have been described for the human X chromosome and autosomes[26], but few if any nullisomic deletions of comparable size and genetic complexity display the phenotypic singularity of *AZFc*. In mammalian genomes, the combined transcriptional complexity and functional specialization of the *AZFc* complex is matched only in the major histocompatibility complex, a similarly sized (3.6-Mb) region with a large, diverse array of genes that function in immune recognition[36].

Reconstructing the evolution of the palindromic complex. The human X and Y chromosomes evolved from an ordinary pair of autosomes[37–39]. Until recently, it was thought that the Y chromosome evolved solely through monotonic decline in the function and density of the ancestral autosome's genes. In essence, the Y chromosome was understood to be a decaying autosome whose genes were preserved intact on the X chromosome[37,38]. Although this classical model has considerable explanatory power in those regions of the Y chromosome with extensive sequence similarity to the X chromosome, it does little to explain the palindromic complex described here. We propose, based on past and present findings, that the palindromic complex was molded during evolution through seven molecular processes, altogether reincarnating this region of the Y chromosome.

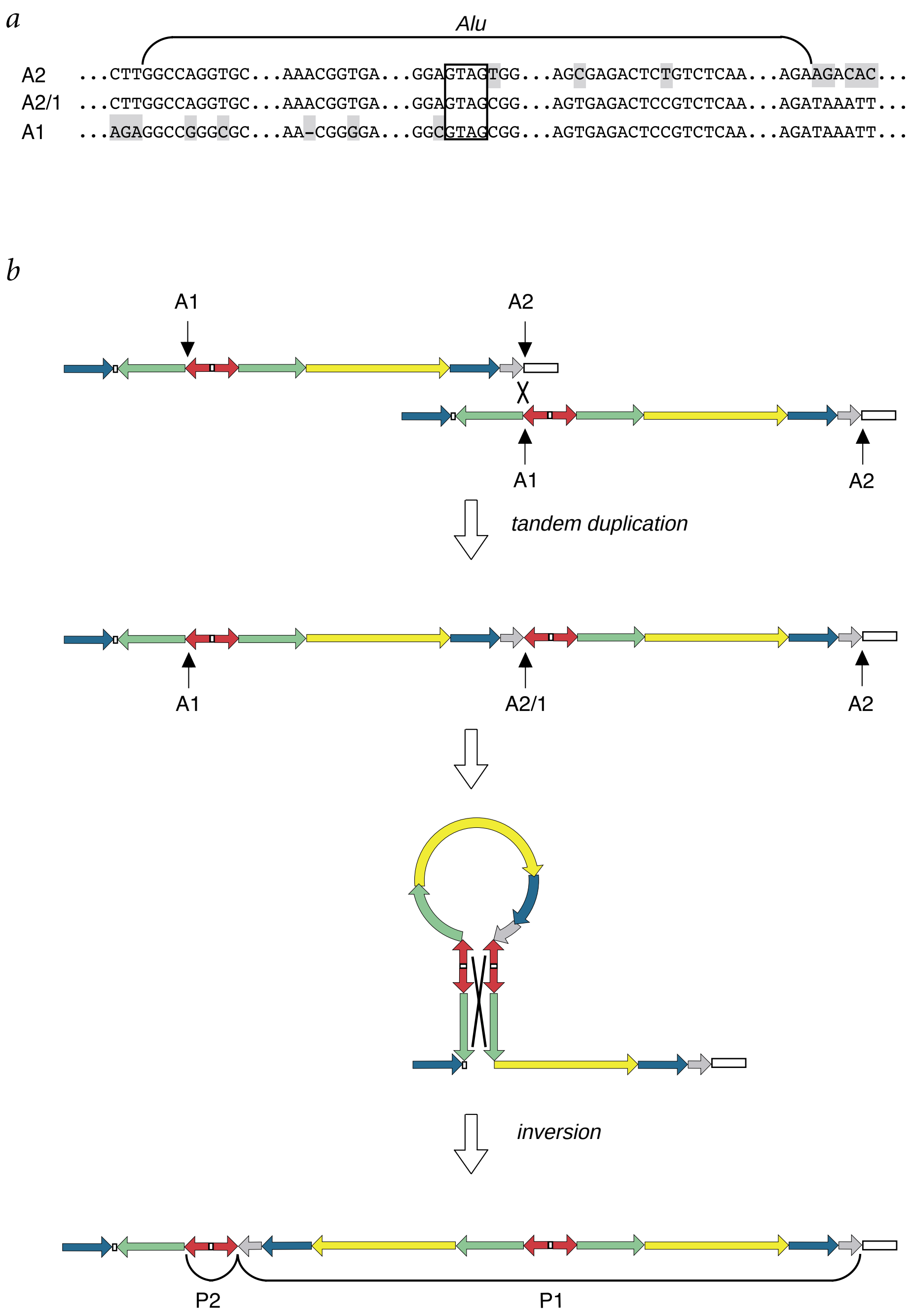
a
Alu
A2 ...CTTGGCCAGGTGC...AAACGGTGA...GGAGTAGTGG...AGCGAGACTCTGTCTCAA...AGAAGACAC...
A2/1 ...CTTGGCCAGGTGC...AAACGGTGA...GGAGTAGCGG...AGTGAGACTCCGTCTCAA...AGATAAATT...
A1 ...AGAGGCCGGGCGC...AA-CGGGGA...GGCGTAGCGG...AGTGAGACTCCGTCTCAA...AGATAAATT...
b
A1
A2
A1
A2
tandem duplication
A1
A2/1
A2
inversion
P2
P1

Fig. 5 Proposed model of origins of the P1 palindrome through *Alu*-mediated tandem duplication and subsequent inversion. ***a***, Alignment of nucleotide sequences of three *Alu* elements whose locations within palindromic complex are indicated in *b*. Eight nucleotides at which A2 or A1 differs from A2/1 are shaded. Box denotes 4-bp region to which recombination between A2 and A1 is localized. ***b***, Proposed tandem duplication (resulting from recombination between *Alu* repeats A1 and A2) followed by inversion (resulting from recombination between amplicons).

Three previously identified molecular evolutionary processes contribute to the palindromic complex's rich array of testis-specific genes: transposition of autosomal transcription units (such as *DAZ*), retroposition of autosomally encoded mRNAs (*CDY*) and persistence of genes previously shared with the X chromosome (*RBMY*)[11,40–42]. For example, the four *DAZ* genes arose during primate evolution through the transposition and subsequent amplification of a single-copy autosomal gene, *DAZL*, that is still extant on human chromosome 3 (refs. 11,13). By comparing the sequence of the palindromic complex with that of chromosome 3, we determined that the autosome-to-Y transposition unit was at least 78 kb long, including 12 kb 5′ of and 48 kb 3′ of the 18-kb *DAZL* transcription unit (Fig. 4). Amplification of portions of the 78-kb transposed segment ultimately accounted for 540 kb of the palindromic complex, including four *DAZ* transcription units with a combined length of 266 kb (Fig. 4).

Transposition and amplification of autosomal transcription units may also have given rise to the *CSPG4LY* and *GOLGA2LY* families. These transcription units are located within each arm of the P1 palindrome, in an 80-kb segment that shows 95% nucleotide identity to human chromosome 15 (Fig 2*h*). Transcribed homologs of *CSPG4LY* and *GOLGA2LY* are found in the corresponding region of human chromosome 15. A third autosomal transposition may have given rise, on each arm of the P1 palindrome, to 115-kb segments that have 97% nucleotide identity to human chromosome 1 (Fig. 2*h*).

In addition to transposition, retroposition and persistence, the palindrome complex evolved through intragenic tandem duplication and exon pruning. As previously reported, these two processes transformed an ancestral Y-borne *DAZ* gene with 11 exons (as in *DAZL*) into the modern-day *DAZ* genes, each containing as many as 28 exons and 19 pseudoexons[11,13]. Similarly, intragenic tandem duplication and exon pruning modified *RBMY1*. We determined this by comparing the *RBMY1* genes sequenced here with the recently sequenced *RBMX* locus[43]. We infer that an ancestral Y-borne gene with 9 exons (as in *RBMX*) was transformed into *RBMY1* with 12 exons and 1 pseudoexon (data not shown).

These molecular evolutionary processes by which genes were derived and modified, however, do not account for the region's most remarkable structures, the palindromes. Based on our examination of the region's sequence, we hypothesize that the P1 palindrome was generated during primate evolution through supragenic tandem duplication and subsequent inversion (Fig. 5). The region's other palindromes may have arisen through analogous two-step processes. In the case of P1, DNA sequence analysis suggests that a 1.6-Mb segment was tandemly duplicated as a result of homologous recombination between *Alu* repeats 'A1' and 'A2' on misaligned Y chromatids (Fig. 5). The A1 and A2 elements, separated by 3 Mb on the modern Y chromosome, differ at eight nucleotide positions. Both elements are diverged about 2% from the *Alu*Ya5 consensus (RepeatMasker database; A.F. Smit, personal communication). Homologous recombination between A1 and A2 readily and precisely accounts for the sequence of a third *Alu* element, A2/1. We have localized the probable point of recombination (and duplication) to a region of four base pairs (Fig. 5*a*). We hypothesize that this tandem duplication was followed by inversion of a large (1.3–2.4-Mb) segment. The inversion was apparently mediated by homologous recombination between 0.6-Mb inverted repeats. This economical, two-step hypothesis accounts not only for the existence of the 3-Mb P1 palindrome but also for the present boundaries of the smaller, immediately adjacent P2 palindrome (Fig. 5).

This elaborate and unusual evolutionary path, together with the vital spermatogenic functions encoded within the palindromic complex, raises the question of the adaptive pressures that favored this unusual structure. The evolutionary formation of the palindromic complex contrasts with the view of the Y chromosome as a decaying X chromosome. Analysis of the sequence of the entire Y chromosome may show these opposing evolutionary dynamics in a larger context.

Methods

Mapping, sequencing and analysis of SFVs. We studied *AZFc* region BACs that we had isolated previously[19] from the RPCI-11 library (BACPAC Resources)[18]. We sequenced individual BACs using previously described methods[43]. BACs that contained *DAZ* genes were difficult to assemble because of intragenic tandem repeats[11,13]; here we employed supplementary methods, including the sequencing of M13 short-insert libraries[44] derived from selected 7–10-kb plasmid subclones, and 'transposon bombing' of 5–7-kb plasmid subclones[45]. We obtained high-accuracy sequence for the 48 BACs shown in Fig. 1. The average overlap between consecutive sequenced BACs (Fig. 1) was 70 kb, with only two overlaps shorter than 20 kb (10 kb and 5 kb, both due to the absence within the RPCI-11 library of other BACs with larger overlaps). With amplicon copies differing at a minimum of 2 nt per 10,000 (Fig. 1*a*), we expected an incorrect overlap of 70 kb to be revealed by at least 14 single-nucleotide differences. We typed BACs for SFVs by PCR amplification of SFV sites and sequencing of the resulting PCR products. Web Table E provides detailed descriptions of 35 SFVs studied in this manner.

In addition, we identified and sequenced BACs containing segments of human chromosomes 1, 3 and 15 that showed sequence similarity to *AZFc*.

Dot plot and sequence analysis. We carried out dot-plot analyses using custom Perl code (http://staffa.wi.mit.edu/page/Y/azfc). When counting single-nucleotide differences per 10 kb, we used FASTA3 (ref. 46) to align sequences, excluding insertions, deletions and variable-copy-number tandem repeats, including microsatellites and polynucleotide tracts.

Characterization of naturally occurring deletions. By screening 823 men with idiopathic, nonobstructive azoospermia or severe oligozoospermia (less than 5 million sperm/ml) for Y-chromosome deletions, we identified 54 men with interstitial deletions limited to *AZFc*. These men lacked sY254 (a *DAZ* STS) but possessed both sY142 (proximal to *AZFc*) and sY160 (distal to *AZFc*)[4]. An additional 19 men (of the 823) lacked sY254 but did not possess sY142 and/or sY160. These men were not included in the present study. For 6 of the 54 men with deletions limited to *AZFc*, DNA stocks were depleted. We studied the remaining 48 men in detail (Fig. 2*c*). We have deposited at GenBank PCR conditions and primer sequences for all STSs employed.

Electronic prediction and laboratory validation of new transcription units. We used BLAST[47] to find matches between the 4.5-Mb sequenced region and all publicly available human ESTs. We experimentally investigated all high-similarity matches where the EST sequence showed evidence of polyadenylation or splicing, and identified the *TTTY4*, *TTTY5* and *TTTY6* transcription units in this way.

In addition, GenScan[27] predicted 19 new genes, all of which we tested by RT–PCR. For each predicted gene, we selected four consecutive predicted exons. For several predicted genes, we were able to select and study a second set of four consecutive predicted exons. For each exon set, we chose six primers: forward primers from exons 1, 2 and 3 and reverse primers from exons 2, 3 and 4 (ref. 48). We combined the primers in six pairings: the exon 1 forward primer with the exon 2, 3 and 4 reverse primers; the exon 2 forward primer with the exon 3 and 4 reverse primers; and the exon 3 forward primer with the exon 4 reverse primer. For RT–PCR, we prepared two cDNA pools (the first using oligo-dT priming and the second using random hexamer priming) from each of the following mRNA samples (Clontech), all from adults unless otherwise noted: (i) testis (ii) pooled salivary gland, stomach, small intestine, pancreas, fetal liver and spleen; (iii) pooled brain, fetal brain, placenta, prostate and thymus and (iv) pooled adrenal gland, heart, lung, skeletal muscle and trachea. We then tested the resulting cDNA pools using the six primer pairs. When a positive RT–PCR result indicated that a predicted gene might be transcribed, we sequenced the PCR product. If the sequence confirmed that the predicted gene was transcribed, we scrutinized the entire predicted gene through additional RT–PCR reactions and sequencing of PCR products, thus defining the transcript more fully. We identified the *CSPG4LY*, *GOLGA2LY* and *TTTY3* transcription units in this way.

Chromosomal hybrid panels. We used NIGMS human-rodent panels 1 and 2 (ref. 49) to confirm large-scale homology to chromosomes 1, 3 and 15 (Fig. 2*h*).

GenBank accession numbers. Sequenced Y-chromosomal BAC clones (all from RPCI-11 library; Fig. 1*b*); sequenced autosomal BAC clones (all from RPCI-11 library). Chromosome 1: 366C6, AC015973. Chromosome 3: 194G10, AC010139; 224E16, AC010727. Chromosome 15: 152F13, AC010724; 156N7, AC010725; 96J23, AC011295. Sequences of boundaries of amplicons as shown in Figs. 1 and 2: AF334526 through AF334546. Primer sequences and PCR conditions for previously reported STSs: sY142, G38345; sY160, G38343; sY254, G38349; sY579, G63909; sY602, G34986. Primer sequences and PCR conditions for new STSs: sY627, G67175; sY639, G67162; sY1054, G67163; sY1125, G67164; sY1190, G67165; sY1192, G67166; sY1196, G67167; sY1197, G67168; sY1198, G67169; sY1201, G67170; sY1206, G67171. New transcripts: *CSPG4LY* (chondroitin sulfate proteoglycan 4–like Y), AF332228; *GOLGA2LY* (golgi antigen 2–like Y), AF332229; *TTTY3* (testis transcript Y 3), AF332230; *TTTY4* (testis transcript Y 4), AF332231; *TTTY5* (testis transcript Y 5), AF332236; *TTTY6* (testis transcript Y 6), AF332237.

Note: Supplementary information is available on the Nature Genetics web site (http://genetics.nature.com/supplementary_info/).

Acknowledgments
We thank C. Tilford and R. Saxena for scientific contributions, L. Hillier for project management, M. Goodheart, L. Pooler, T.V. Pyntikova and M. Velez-Stringer for technical contributions, C. Bruning, V. Cardone, A. de la Chapelle,

B. Gilbert, W.A. Hogge, R. Newton, P. Patrizio, M. Summers and D. Van Dyke for patient samples and D. Berry, A. Bortvin, S. Repping and M. Royce-Tolland for comments on the manuscript. This work was supported by National Institutes of Health and Howard Hughes Medical Institute.

Received 22 May; accepted 10 September 2001.

1. Hull, M.G. *et al.* Population study of causes, treatment, and outcome of infertility. *Br. Med. J.* **291**, 1693–1697 (1985).
2. Ma, K. *et al.* Towards the molecular localisation of the *AZF* locus: mapping of microdeletions in azoospermic men within 14 subintervals of interval 6 of the human Y chromosome. *Hum. Mol. Genet.* **1**, 29–33 (1992).
3. Kobayashi, K. *et al.* PCR analysis of the Y chromosome long arm in azoospermic patients: evidence for a second locus required for spermatogenesis. *Hum. Mol. Genet.* **3**, 1965–1967 (1994).
4. Reijo, R. *et al.* Diverse spermatogenic defects in humans caused by Y chromosome deletions encompassing a novel RNA-binding protein gene. *Nature Genet.* **10**, 383–393 (1995).
5. Vogt, P.H. *et al.* Human Y chromosome azoospermia factors (*AZF*) mapped to different subregions in Yq11. *Hum. Mol. Genet.* **5**, 933–943 (1996).
6. Simoni, M. *et al.* Screening for deletions of the Y chromosome involving the *DAZ* (*Deleted in AZoospermia*) gene in azoospermia and severe oligozoospermia. *Fertil. Steril.* **67**, 542–547 (1997).
7. Nakahori, Y. *et al.* The Y chromosome region essential for spermatogenesis. *Horm. Res.* **46 (Suppl. 1)**, 20–23 (1996).
8. Girardi, S.K., Mielnik, A. & Schlegel, P.N. Submicroscopic deletions in the Y chromosome of infertile men. *Hum. Reprod.* **12**, 1635–1641 (1997).
9. 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).
10. Yen, P.H. A long-range restriction map of deletion interval 6 of the human Y chromosome: a region frequently deleted in azoospermic males. *Genomics* **54**, 5–12 (1998).
11. 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).
12. Glaser, B., Yen, P.H. & Schempp, W. Fibre-fluorescence *in situ* hybridization unravels apparently seven DAZ genes or pseudogenes clustered within a Y-chromosome region frequently deleted in azoospermic males. *Chromosome Res.* **6**, 481–486 (1998).
13. Saxena, R. *et al.* Four *DAZ* genes in two clusters found in *AZFc* region of human Y chromosome. *Genomics* **67**, 256–267 (2000).
14. Ma, K. *et al.* A Y chromosome gene family with RNA-binding protein homology: candidates for the azoospermia factor *AZF* controlling human spermatogenesis. *Cell* **75**, 1287–1295 (1993).
15. Elliott, D.J. *et al.* Expression of *RBM* in the nuclei of human germ cells is dependent on a critical region of the Y chromosome long arm. *Proc. Natl Acad. Sci. USA* **94**, 3848–3853 (1997).
16. Vogel, T., Speed, R.M., Teague, P. & Cooke, H.J. Mice with Y chromosome deletion and reduced *Rbm* genes on a heterozygous *Dazl1* null background mimic a human azoospermic factor phenotype. *Hum. Reprod.* **14**, 3023–3029 (1999).
17. Underhill, P.A. *et al.* Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Res.* **7**, 996–1005 (1997).
18. Osoegawa, K. *et al.* An improved approach for construction of bacterial artificial chromosome libraries. *Genomics* **52**, 1–8 (1998).
19. Tilford, C. *et al.* A physical map of the human Y chromosome. *Nature* **409**, 943–945 (2001).
20. Reijo, R., Alagappan, R.K., Patrizio, P. & Page, D.C. Severe oligozoospermia resulting from deletions of azoospermia factor gene on Y chromosome. *Lancet* **347**, 1290–1293 (1996).
21. Mulhall, J.P. *et al.* Azoospermic men with deletion of the *DAZ* gene cluster are capable of completing spermatogenesis: fertilization, normal embryonic development and pregnancy occur when retrieved testicular spermatozoa are used for intracytoplasmic sperm injection. *Hum. Reprod.* **12**, 503–508 (1997).
22. Silber, S.J., Alagappan, R., Brown, L.G. & Page, D.C. Y chromosome deletions in azoospermic and severely oligozoospermic men undergoing intracytoplasmic sperm injection after testicular sperm extraction. *Hum. Reprod.* **13**, 3332–3337 (1998).
23. 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).
24. Kamp, C., Hirschmann, P., Voss, H., Huellen, K. & Vogt, P. 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).
25. 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).
26. Lupski, J. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.* **14**, 417–422 (1998).
27. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
28. Lahn, B.T. & Page, D.C. Functional coherence of the human Y chromosome. *Science* **278**, 675–680 (1997).
29. Chai, N.N. *et al.* Structure and organization of the *RBMY* genes on the human Y chromosome: transposition and amplification of an ancestral autosomal *hnRNPG* gene. *Genomics* **49**, 283–289 (1998).
30. Lefebvre, S. *et al.* Identification and characterization of a spinal muscular atrophy–determining gene. *Cell* **80**, 155–165 (1995).
31. Lewin, B. Genes for SMA: *multum in parvo*. *Cell* **80**, 1–5 (1995).
32. Monani, U.R. *et al.* A single nucleotide difference that alters splicing patterns distinguishes the *SMA* gene *SMN1* from the copy gene *SMN2*. *Hum. Mol. Genet.* **8**, 1177–1183 (1999).
33. 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).
34. Kleiman, S.E. *et al.* Genetic evaluation of infertile men. *Hum. Reprod.* **14**, 33–38 (1999).
35. Page, D.C., Silber, S. & Brown, L.G. Men with infertility caused by *AZFc* deletion can produce sons by intracytoplasmic sperm injection, but are likely to transmit the deletion and infertility. *Hum. Reprod.* **14**, 1722–1726 (1999).
36. The MHC Sequencing Consortium. Complete sequence and gene map of a human major histocompatibility complex. *Nature* **401**, 921–923 (1999).
37. Ohno, S. *Sex Chromosomes and Sex-linked Genes* (Springer-Verlag, Berlin, 1967).
38. Graves, J.A.M. The origin and function of the mammalian Y chromosome and Y-borne genes—an evolving understanding. *BioEssays* **17**, 311–321 (1995).
39. Lahn, B.T. & Page, D.C. Four evolutionary strata on the human X chromosome. *Science* **286**, 964–967 (1999).
40. 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).
41. 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).
42. Sargent, C.A. *et al.* The critical region of overlap defining the *AZFa* male infertility interval of proximal Yq contains three transcribed sequences. *J. Med. Genet.* **36**, 670–677 (1999).
43. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
44. McMurray, A.A., Sulston, J.E. & Quail, M.A. Short-insert libraries as a method of problem solving in genome sequencing. *Genome Res.* **8**, 562–566 (1998).
45. Liu, L., Whalen, W., Das, A. & Berg, C.M. Rapid sequencing of cloned DNA using a transposon for bidirectional priming: sequence of the *Escherichia coli* K-12 avtA gene. *Nucleic Acids Res.* **15**, 9461–9469 (1987).
46. Pearson, W.R. Flexible sequence similarity searching with the FASTA3 program package. in *Bioinformatics Methods and Protocols* (eds. Krawetz, S. & Misener, S.) 185–219 (Humana Press, Totowa, New Jersey, 2000).
47. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
48. Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. in *Bioinformatics Methods and Protocols* (eds. Krawetz, S. & Misener, S.) 365–386 (Humana Press, Totowa, NJ, 2000).
49. Drwinga, H.L., Toji, L.H., Kim, C.H., Greene, A.E. & Mulivor, R.A. NIGMS human/rodent somatic cell hybrid mapping panels 1 and 2. *Genomics* **16**, 311–314 (1993).