The *SPANX* gene family of cancer/testis-specific antigens: Rapid evolution and amplification in African great apes and hominids

Natalay Kouprina[†], Michael Mullokandov[†], Igor B. Rogozin[‡], N. Keith Collins[†], Greg Solomon[§], John Otstot[§], John I. Risinger[†], Eugene V. Koonin[‡], J. Carl Barrett[†], and Vladimir Larionov[†¶]

[†]Laboratory of Biosystems and Cancer, National Cancer Institute, Bethesda, MD 20892; [‡]National Center for Biotechnology Information, Bethesda, MD 20892; and [§]Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709

Communicated by Ira Pastan, National Institutes of Health, Bethesda, MD, December 20, 2003 (received for review November 11, 2003)

Human sperm protein associated with the nucleus on the X chromosome (*SPANX*) genes comprise a gene family with five known members (*SPANX-A1*, *-A2*, *-B*, *-C*, and *-D*), encoding cancer/testis-specific antigens that are potential targets for cancer immunotherapy. These highly similar paralogous genes cluster on the X chromosome at Xq27. We isolated and sequenced primate genomic clones homologous to human *SPANX*. Analysis of these clones and search of the human genome sequence revealed an uncharacterized group of genes, *SPANX-N*, which are present in all primates as well as in mouse and rat. In humans, four *SPANX-N* genes comprise a series of tandem duplicates at Xq27; a fifth member of this subfamily is located at Xp11. Similarly to *SPANX-A/D*, human *SPANX-N* genes are expressed in normal testis and some melanoma cell lines; testis-specific expression of *SPANX* is also conserved in mouse. Analysis of the taxonomic distribution of the long and short forms of the intron indicates that *SPANX-N* is the ancestral form, from which the *SPANX-A/D* subfamily evolved in the common ancestor of the hominoid lineage. Strikingly, the coding sequences of the *SPANX* genes evolved much faster than the intron and the 5′ untranslated region. There is a strong correlation between the rates of evolution of synonymous and nonsynonymous codon positions, both of which are accelerated 2-fold or more compared to the noncoding sequences. Thus, evolution of the *SPANX* family appears to have involved positive selection that affected not only the protein sequence but also the synonymous sites in the coding sequence.

The sperm protein associated with the nucleus on the X chromosome (*SPANX*) multigene family encodes proteins whose expression is restricted to the normal testis and certain tumors (1, 2). These postmeiotically transcribed genes comprise one of the few examples of haploid expression from X-linked genes (3). Antibodies against SPANX recognized spermatozoa craters and cytoplasmic droplets in ejaculated spermatozoa. Spermatozoa craters correspond to indentations on the nuclear surface and to vacuoles within the condensed chromatin in spermatozoa nuclei. Nuclear vacuoles are believed to be derived from the nucleolus of spermatocytes and spermatids (ref. 4 and references therein). The presence of these craters usually is linked to reduced fertility in mammals (5). However, the correlation between fertility and large nuclear craters in human spermatozoa remains controversial (6, 7).

SPANX genes encode small proteins migrating as a broad band of 15–20 kDa under reducing electrophoresis conditions. In spermatozoa, SPANX proteins are found in the form of dimers or complexes with other proteins (1, 3). The *SPANX* cluster on chromosome X consists of five genes. These genes reside in the Xq26.3-Xq27.3 region, within ≈20 kb, highly similar tandem duplications. All *SPANX* genes consist of two exons separated by an ≈650-bp intron containing a solo retroviral LTR sequence (8). *SPANX* genes are divided into two groups, the *SPANX-A*- and *-B*-like subfamilies (8). Classification of *SPANX* genes is based on the presence of diagnostic amino acid substitutions. The *SPANX-A*-like subfamily consists of four members, *SPANX-A1, -A2, -C*, and *-D*. All these genes encode proteins consisting of 97-aa residues; the 23 N-terminal residues are encoded in exon 1, and the rest of the protein sequence is encoded in exon 2. The *SPANX-B*-like subfamily is represented by a single gene of the same name. This gene encodes a protein with 103-aa residues. The larger size of this protein compared to the *SPANX-A*-like subfamily proteins is due to the presence of an 18-bp insertion in exon 1. The sequences of the *SPANX-A* subfamily proteins share 90–98% identity to each other, whereas the *SPANX-B* sequence share 75–80% identity to the *SPANX-A* sequences.

The interest in the *SPANX* genes is mostly because they are specifically expressed in a variety of tumors. Expression profile analysis showed that at least four of the family members (*SPANX-A1*, *-A2*, *-B*, and *-D*) are expressed in cancer cells, including highly metastatic cell lines from melanomas, bladder carcinomas, and myelomas (1, 9–11). In transformed mammalian cells, SPANX proteins are associated with the nuclear envelope, a location similar to that in human spermatids and spermatozoa (1, 9, 10). Therefore SPANX represent a specific subgroup of cancer/testis-associated antigens (8), which are considered to be prime candidates for tumor vaccines.

So far, sequences of non-human *SPANX* homologs have not been reported. Recently, Zendman *et al*. (8) showed, by using blot hybridization, that *SPANX* homologs are present in the chimpanzee and gorilla genomes. In the same experiment, no signal was detected with orangutan and squirrel monkey DNAs. This result could be due to the recent emergence of *SPANX* genes in evolution, their rapid divergence in primates, or both.

In the present study, in an attempt to reconstruct the evolutionary history of the *SPANX* family, we isolated and sequenced *SPANX* genes from several primate species and additionally identified genes homologous to *SPANX* in the mouse and rat genomes. We describe a previously undetected subfamily of *SPANX* genes and show that *SPANX* genes evolve extremely rapidly and, apparently, under positive selection that acts on both synonymous and nonsynonymous sites of the coding sequences.

Materials and Methods

Amplification of *SPANX* Genes from Primates. Genomic DNAs from chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*), orangutan (*Pongo pygmaeus*), rhesus macaque (*Macaca mulatta*), and tamarin (*Saguinus labiatus*) (Coriell Institute for Medical Research, Camden, NJ) were used for amplification with the specific primers (Table 2, which is published as supporting information on the PNAS web site). PCR was performed by using 1 μl of genomic DNA (100 ng) in a 50-μl reaction volume under conditions: 94°C, 2 min (94°C, 30 s; 60°C, 10 s; 68°C, 9 min × 30 cycles); 72°C, 7 min; 4°C, hold.

Abbreviation: TAR, transformation-associated recombination.

¶To whom correspondence should be addressed. E-mail: larionov@mail.nih.gov.

RT-PCR. Total RNAs from mouse (brain, testis, liver, and heart) and human (brain, testis, liver, and skeletal muscle) tissues (Ambion, Austin, TX) were used for screening *SPANX* expression with the primers described in Table 2. cDNA was made from 1 μg of total RNA using the Superscript first-strand system kit (Invitrogen) and priming with oligo(dT) per their standard protocol. Human β-actin primers (BD Biosciences Clontech) were used as positive controls for both human and mouse RT-PCR. NCI-60 cancer cell lines were from the National Cancer Institute. RT-PCR was performed by using 1 μl of cDNA or 1 μl of genomic DNA in a 50-μl reaction volume. Standard reaction conditions were: 94°C, 5 min (94°C, 1 min; 55°C, 1 min; 72°C, 1 min $\times$ 35 cycles); 72°C, 7 min; and 4°C, hold.

Construction of Transformation-Associated Recombination (TAR) Vector and Cloning by *in Vivo* Recombination in Yeast. TAR cloning experiments were carried out as described (12). The TAR vector was constructed by using pVC604. The vector contains 5′ 164-bp and 3′ 187-bp targeting sequences, specific to the unique sequences flanking *SPANX-C*. They were amplified from human genomic DNA with specific primers (Table 2). The 5′ and 3′ targeting sequences correspond to positions 39,708–39,872 and 122,818–123,004 in the bacterial artificial chromosome (BAC) (AL109799). Before use in TAR cloning experiments, the vector was linearized with *Sph*I. Genomic DNAs were prepared from primate tissue culture lines (Coriell Institute for Medical Research). To identify clones positive for *LDOC1*, yeast transformants were examined by PCR by using a pair of diagnostic primers (Table 2). The yield of *LDOC1*-positive clones from African apes genomic DNAs (chimpanzee, gorilla, and bonobo) was the same as with human DNA (1%). The size, *Alu* profiles, and retrofitting of yeast artificial chromosomes (YACs) into BACs were determined as described (12). *Alu* profiles of three independent TAR isolates for each species were indistinguishable. These results strongly suggest that the isolated YACs contain nonrearranged genomic segments.

Sequencing. Bonobo and gorilla TAR clones were directly sequenced from bacterial artificial chromosome DNAs by Fidelity Systems (Gaithersburg, MD). Sequences of primate parologs were generated by TA subcloning of 81 PCR products, 1.2 or 1.4 kb, amplified from genomic DNAs (20 for chimpanzee, 20 for gorilla, 20 for orangutan, and 20 for tamarin) and one for rhesus macaque (9.0 kb). Sequence forward and reverse reactions were run on a 3100 Automated Capillary DNA Sequencer (PE Applied Biosystems). DNA sequences were compared by using the GCG DNA ANALYSIS Wisconsin Package (www.accelrys.com/support/bio/faqs_wis_pkg.html) and National Center for Biotechnology Information BLAST. Non-human sequences were deemed paralogous if more than two sequence differences were observed. All clones were named and numbered according to the clone/accession identifier (Table 3, which is published as supporting information on the PNAS web site).

Sequence Analysis. Database searches were performed by using the versions of the BLAST program appropriate for different types of sequence comparisons: BLASTN for nucleotide sequences, BLASTP for protein sequences, and TBLASTN for searching a nucleotide database translated in six frames with a protein query (13). Multiple alignments of protein sequences were constructed by using the MACAW program (14). Multiple alignments of nucleotide sequences were aligned to correspond to the protein sequence alignments. Protein secondary structure was predicted by using the PHD program (http://cubic.bioc.columbia.edu/predict protein), with a multiple sequence alignment submitted as a query (15) (http://cubic.bioc.columbia.edu/predict protein). The phylogenetic tree was constructed by using the neighbor-joining method (16) as implemented in the MEGA2 program (17) with the maximum parsimony and maximum likelihood methods as implemented in the PAUP* program (http://paup.csit.fsu.edu). Evolutionary rates for synonymous and nonsynonymous positions in the coding sequences were calculated by using the modified Nei–Gojobori method (18). The evolutionary rates for noncoding sequences were calculated by using the two-parameter Kimura model (19).

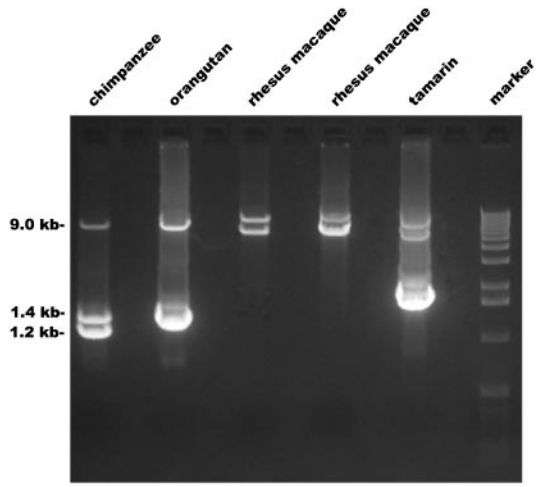

Fig. 1. PCR amplification of members of the *SPANX* family from chimpanzee (African great apes), orangutan (great apes), rhesus macaque (Old World monkey), and tamarin (New World monkey). Oligonucleotides were designed within the promoter and 3′ uncoding regions. The double upper bands for rhesus macaque and tamarin are presumably due to polymorphism in paralogs.

Results

The *SPANX* Family: Identification of a Second Subfamily in Primates and a Single *SPANX* Gene in Rodents. To shed light on the evolution of the *SPANX* family, we isolated and characterized primate genomic segments homologous to human *SPANX*. The corresponding regions from five species (chimpanzee, gorilla, orangutan, rhesus macaque, and tamarin) were amplified by using a set of primers developed from the conserved 5′ and 3′ flanking sequences of human *SPANX* genes (see *Materials and Methods*). PCR products with a size predicted for the *SPANX-A/D* genes (1.2 kb) were obtained only from African apes (Fig. 1). Sequence analysis of these fragments revealed genes with ≈65–90% nucleotide identity to the human *SPANX* genes and a similar organization, i.e., two exons and a 650-bp highly conserved intron that includes ≈400 bp of ERV LTR. In addition to or instead of the 1.2-kb fragment, we observed two other PCR products ≈1.4 and ≈9.0 kb in size. These bands were amplified from DNA samples of chimpanzee, gorilla, orangutan, and tamarin; for rhesus macaque, only the 9.0-kb fragment was detected (Fig. 1). Sequence analysis of the 1.4-kb product obtained from anthropoids revealed *SPANX*-like genes that were organized similarly to the known *SPANX-A/D* genes, i.e., contained two exons separated by a 650-bp intron with a solo LTR and shared ≈50% amino acid sequence identity with the SPANX-A/D proteins. A characteristic difference between the newly detected *SPANX*-related genes and human *SPANX-A/D* genes was found in exon 2. This exon is longer in primate genes than in human genes and is variable in size (280–320 bp vs. 219 bp in *SPANX-A/D*). Expansion and variability of exon 2 size are due to the presence of a 39-bp minisatellite sequence at its 5′ end. Similar amplification of minisatellites in exons without disruption of the ORF has been previously described for other genes (20, 21). End sequencing of 9.0-kb primate clones revealed

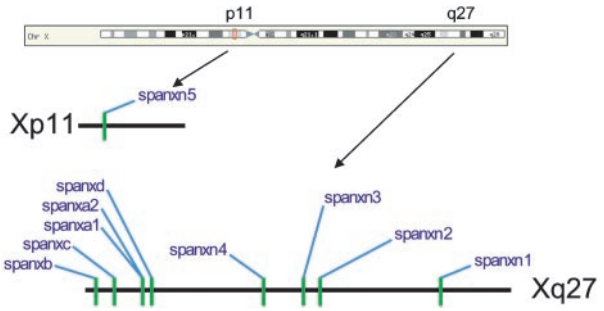

Fig. 2. Location of the *SPANX* family genes on human chromosome X. Five members of the *SPANX-A/D* subfamily, *SPANX-A1*, *-A2*, *-B*, *-C*, and *-D*, are clustered within an ≈800-kb region at Xq27.2. Four members of the *SPANX-N* subfamily, *SPANX–N1* (positions 142995930–143005820), *-N2* (positions 141495326–141490625), *-N3* (positions 141297635–141291834), and *-N4* (positions 140806882–140816198), are located ≈2 Mb apart from *SPANX-A/D* (University of California, Santa Cruz, July 2003, http://genome.ucsc.edu). *SPANX-N5* (positions 51791606–51793934) is located on the short arm of chromosome at Xp11. *SPANX* genes reside within ≈20-kb blocks, which are duplicated retaining a high homology across the chromosome.

significant sequence similarity to exon 1 and exon 2 of the 1.4-kb clones. The 1.4- and 9.0-kb clones differed in that the latter contained a second LTR upstream of exon 2. Sequencing of a 9.0-kb clone of rhesus macaque showed the presence of an intact ERV sequence. Presence of ERV in 9.0-kb clones of other species was confirmed by PCR (data not shown).

A search of the GenBank database detected five regions of significant similarity to the 1.4- and 9.0-kb primate sequences in the human genome. One of these regions, which is located at Xp11, produced a contiguous alignment with the 1.4-kb primate sequences. Analysis of this genomic sequence revealed a *SPANX*-like gene consisting of two exons separated by a 650-bp LTR-containing intron. Four other regions of similarity were identified at Xq27, ≈2 Mb away from the *SPANX-A/D* gene cluster (Fig. 2). In this region, the predicted exon sequences are separated by an intron containing a complete ERV, in an arrangement similar to the primate 9.0-kb clones. Like the *SPANX* homologs in primates, the five new members of the *SPANX* family in humans have a variable size of exon 2 due to the presence of a 39-bp minisatellite repeat at the 5′ end. The protein sequences encoded by the five human *SPANX*-like genes identified here share 50–80% identity with each other and 40–50% identity with the sequences of the *SPANX-A/D* proteins. In all cases, the exon boundaries and the splice sites were well conserved (data not shown), suggesting that these genes are expressed.

Based on the distinct gene organization and the relatively low sequence similarity to *SPANX-A/D* genes, the ERV-containing genes were classified as a second *SPANX* subfamily, which we named *SPANX-N*. In humans, the *SPANX-N* genes encode predicted proteins of 72 aa (*SPANX-N1*), 180 aa (*SPANX-N2*), 141 aa (*SPANX-N3*), 159 aa (*SPANX-N4*), and 72 aa (*SPANX-N1*). A search of the GenBank database revealed two regions of significant similarity to human *SPANX-N* in the mouse and rat genomes. Both mouse and rat *SPANX-N* homologs are previously unannotated genes; the expression of the mouse gene is supported by the detection of eight ESTs in Database of Expressed Sequence Tags (dbEST) (BU939216, CA463062, CA464820, CB273391, BX635129, BC048649, CB273391, and BU946237). The mouse and rat gene encode, respectively, 87- and 115-aa proteins with 28–36% amino acid identity to human *SPANX-N* genes. The mouse gene contains a 250-bp intron that shares ≈65% identity with the primate *SPANX* intron. The smaller size of the intron is due to the absence of the ERV sequence or the LTR, which is present in all primate *SPANX* genes. Because the murine *SPANX* homolog not only shows the closest similarity to the *SPANX-N* genes but also is located in the mouse chromosome X region syntenic to *SPANX-N1-N4*, we concluded that this gene is the single murine ortholog of the human *SPANX-N* genes. Thus, the *SPANX-N* subfamily is apparently represented not only in all primates but also in rodents, whereas the *SPANX-A/D* genes appear to be present exclusively in the African great apes and humans.

Identification of the *SPANX-N* family and, particularly, the rodent *SPANX* homologs, which showed limited sequence similarity to the primate *SPANX* sequences, provided an opportunity to gain some insight into the structure and putative functional motifs of the SPANX proteins. The most prominent conserved sequence feature of the *SPANX* family is the central hydrophobic patch ending with an arginine (Fig. 3). Secondary structure prediction suggested that the central conserved region formed a β-hairpin with a strongly hydrophobic proximal strand, followed by an α-helix. The rest of the protein seems to have a disordered structure with few residues conserved throughout the family but with considerable conservation within subfamilies and a marked preponderance of charged and polar residues. The bipartite nuclear localization signal that has been previously detected in the *SPANX-A/D* subfamily (8) is conserved in most of the *SPANX-N* proteins, with the exception of *SPANX-N2* and *-N4* but not in the rodent sequences; however, the latter contain a putative monopartite nuclear localization signal (Fig. 3). The presence of a small globular core embedded in apparently disordered structure suggests that SPANX protein monomers might be unstable and is compatible with the reported dimer formation (1, 10).

***SPANX-N* Genes Are Expressed in Normal Testis and in Melanoma Cell Lines.** Using primers specific to the *SPANX-N2* and *-N3* mRNA, we analyzed expression of these genes in a panel of normal tissues. A 264-bp band of expected size was detected only in testis (Fig. 4*a*). Sequencing of the RT-PCR products confirmed their identity to the *SPANX-N2* and *-N3* genes. Furthermore, the amplified sequences corresponded to two ESTs in dbEST (BU569937 and BF967778). Similar experiments with a panel of normal tissues from mice also detected expression of the mouse *SPANX* gene only in testis (Fig. 4*b*). The exclusive expression of these genes in normal testis correlated with the conservation of the promoter region, which contained two recognition sites for testis-specific transcription factors (Fig. 7, which is published as supporting information on the PNAS web site). *SPANX-N* expression in the NIH-60 panel of cancer cell lines that represent nine different types of cancers was also examined. RT-PCR products of *SPANX-N2* or *-N3* of the expected size were detected only in a melanoma cell line (Table 4, which is published as supporting information on the PNAS web site). Notably, the *SPANX-A/D* subfamily is also expressed in the same line. Coexpression of members of the two *SPANX* gene subfamilies is not surprising because of the remarkable conservation of the promoter sequences. Thus, expression profile analysis indicates that the *SPANX-N* subfamily, similar to the *SPANX-A/D* subfamily, consists of cancer/testis antigens (CTA) genes. Furthermore, the testis-specific pattern of expression of *SPANX* genes is conserved between primates and rodents.

Evolution of the *SPANX* Family: An Unusual Case of Apparent Positive Selection in both Nonsynonymous and Synonymous Positions. Reproduction-related genes, particularly those involved in spermatogenesis, tend to evolve rapidly. Moreover, many of them appear to be subject to positive selection, which is usually detected by measuring the d_a/d_s ratio, i.e., the ratio of the evolutionary rates in nonsynonymous and synonymous codon positions (22–25). For most genes, d_a/d_s is $\ll 1$, which is a sign

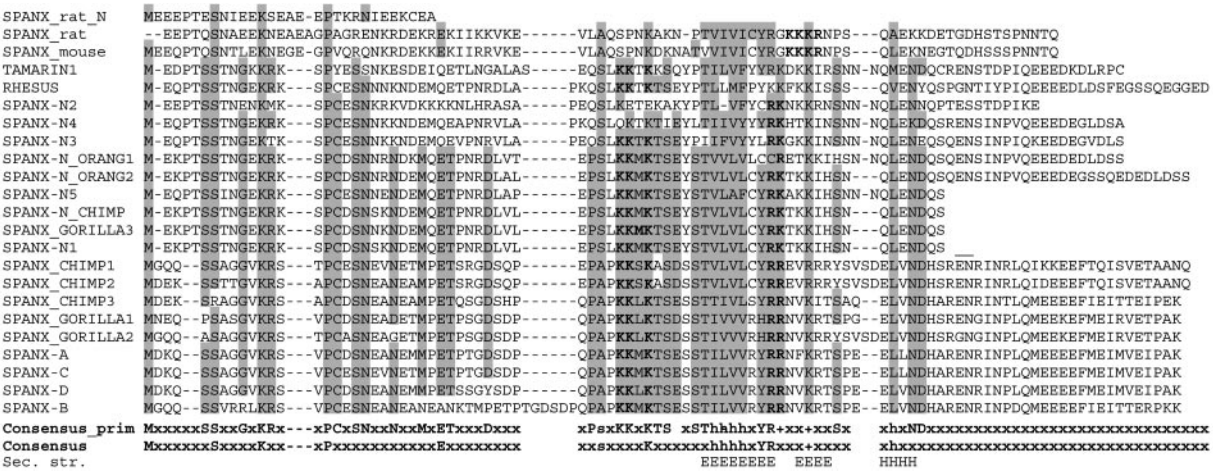

Fig. 3. Multiple alignment of the SPANX protein sequences. Consensus includes amino acid residues that are conserved in both rodent sequences and in the majority of sequences in each of the two primate subfamilies; consensus_prim consists of residues represented in the majority of the sequences in each primate family but not in rodents. h, hydrophobic residues; S, small residues; +, positively charged residues. Residues that conform to the consensus are shaded. In the predicted secondary structure: E, extended conformation (β-strand), and H, α-helix; the rest of the protein is predicted to consist of loops and unstructured regions. Residues that comprise predicted nuclear localization signal are show in bold type. The upper line corresponds to the N-terminal duplication of rat SPANX.

of evolution under purifying selection; in contrast, $d_a/d_s > 1$ is considered an indication of positive (diversifying) selection.

The rate of evolution of *SPANX* genes is outstanding even among reproductive proteins. The highest level of conservation between rodent SPANX proteins and human SPNAX-N family members is ≈36%, substantially less than the values observed for most spermatozoa- and testis-associated proteins and about the same as for transition protein 2, the most rapid evolving among analyzed human and mouse orthologs (22, 25). The d_a/d_s ratio for *SPANX* genes was typically close to 1 (Table 1), which normally would be interpreted as evolution under substantially relaxed purifying selection, perhaps near-neutral evolution. For several comparisons of closely related sequences, within the *SPANX-A/D* and *-N* subfamilies, d_a/d_s values close to 2 were observed; however, because of the small total number of nucleotide substitutions, these failed to pass statistical tests for positive selection (Table 5, which is published as supporting information on the PNAS web site). However, a highly unusual feature of the *SPANX* family is that both synonymous and nonsynonymous positions in the coding sequences of many *SPANX* genes evolved much faster than the noncoding sequences of the 5′ UTR and the intron (Tables 1 and 5). This anomalous mode of evolution was detected both among the closely related paralogs within the *SPANX-A/D* and *-N* subfamilies and in intersubfamily comparisons. Most of the intron sequences do not seem to contain specific functional signals and are believed to evolve (nearly) neutrally. Therefore, the almost 2-fold acceleration of evolution of all positions in the coding sequence compared to the intron (Table 1) seems to suggest that, in the *SPANX* family, positive selection acts with nearly equal strength on synonymous and nonsynonymous positions.

Given that the coding sequences of the *SPANX* genes appeared to have evolved under positive selection, they were not

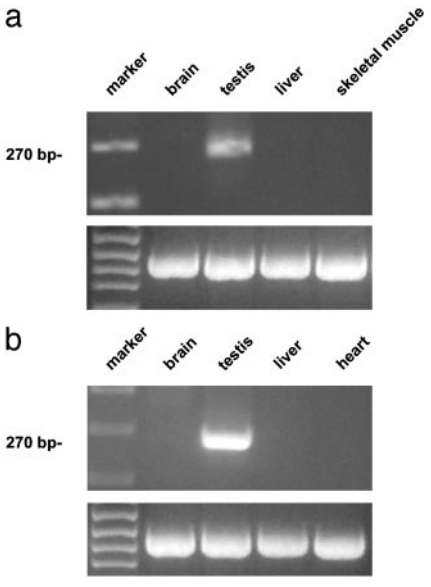

Fig. 4. RT-PCR analysis of the *SPANX-N* gene subfamily. (*a*) cDNA was prepared from a panel of human tissue mRNAs. Oligonucleotides were designed within exons 1 and 2 to amplify putative transcripts. A 264-bp band of the expected size was observed only in testis and the melanoma cell line LOX IMVI. Two members of the human *SPANX-N* subfamily, *SPANX-N2* and *-N3*, were revealed by cloning and sequencing of PCR products. (*b*) cDNA was prepared from a panel of mouse tissue mRNAs. Oligonucleotides were designed within exons 1 and 2 to amplify a putative transcript. A 264-bp band of the expected size was observed only in testis. Control PCRs were carried out with the same samples by using actin-specific primers.

Table 1. Mean evolutionary distances for the 5′ flanking regions, the intron, and the coding sequences of the *SPAN-X* genes

	All	*SPANX-N*	*SPANX-A/D*	*SPANX-N* vs. *SPANX-A/D*
5′	0.14	0.11	0.07	0.19
Intron	0.10	0.12	0.04	0.12
CDS_N	0.33	0.16	0.15	0.49
CDS_S	0.42	0.20	0.14	0.65

The modified Nei–Gojobori model was used for coding sequences, and the Kimura two-parameter model was used for noncoding sequences. CDS_N, nonsynonymous positions in the coding region; CDS_S, synonymous positions in the coding region. The probability that the rate difference between the coding and noncoding sequences was due to chance was $\ll 10^{-10}$ for both the intrasubfamily and intersubfamily comparisons according to the binomial test.

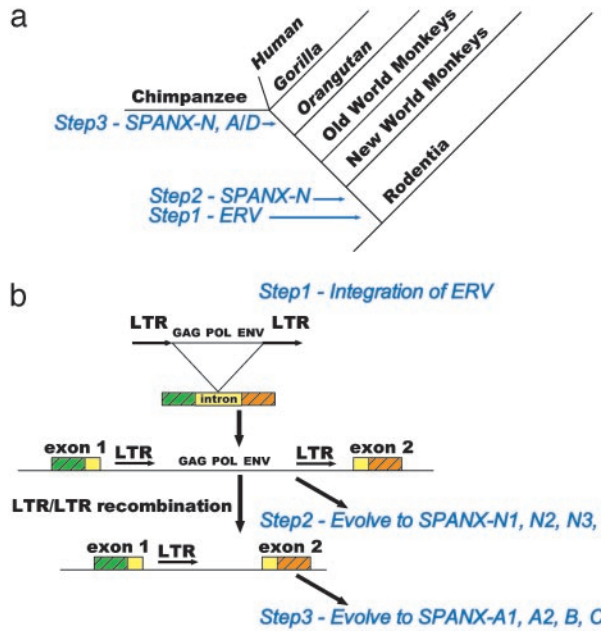

Fig. 5. A hypothetical evolutionary scenario for the *SPANX* gene family. The expansion of *SPANX* genes is superimposed on the tree of primate evolution.

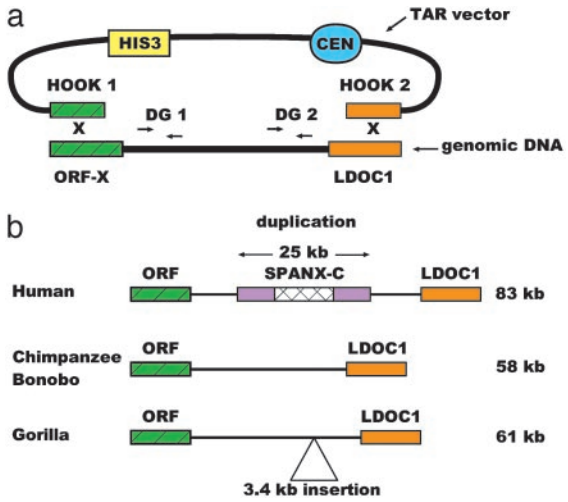

Fig. 6. Isolation of the syntenic genomic fragments containing the *LDOC1* gene from chimpanzee, bonobo, and gorilla by TAR cloning. (*a*) TAR vector containing a 3′ hook specific to a human unique sequence (ORF-X) that is ≈16 kb upstream of *SPANX-C* and a 5′ hook specific to the human *LDOC1* gene. Recombination between genomic DNA and the vector leads to cloning the syntenic regions of the species. *CEN* corresponds to the yeast chromosome VI centromere, *HIS3* is a yeast selectable marker, and DG1 and DG2 are diagnostic primers. (*b*) Comparison of sequences revealed that TAR clones from chimpanzee, bonobo, and gorilla do not contain the *SPANX-C* gene along with the duplication where it was inserted. The clone from gorilla contains a 3.4-kb insertion that corresponds to a hypothetical gene consisting of eight exons on human chromosome 3.

suitable for phylogenetic analysis. Phylogenetic tree constructed from alignments of the 5′-UTR and intron sequences clearly supported the fundamental split between the *SPANX-N* and *-A/D* subfamilies (Fig. 8, which is published as supporting information on the PNAS web site). Beyond that, however, the resolution of the phylogenetic trees was low, presumably because of the small number of informative positions, nonuniformity of evolutionary rates, or both. Nevertheless, comparison of the organization of *SPANX* genes in different mammalian species allowed a confident reconstruction of the main events in the evolution of this family (Fig. 5). The common ancestor of rodents and primates apparently had a single *SPANX-N* subfamily gene (extensive search for potential divergent members of the *SPANX* family in the mouse and rat genomes failed to detect any). Because the ERV-containing intron was detected in *SPANX-N* subfamily genes from all examined primates, including tamarin (a representative of early-branching New World monkeys), ERV insertion apparently occurred early in primate evolution, before the divergence of the major monkey lineages (≈50 million years ago). *SPANX-N* genes lacking the ERV and containing only a solo LTR in their intron apparently evolved independently in New World monkeys and great apes via duplications accompanied by homologous recombination between the ERV's LTRs. Other *SPANX-N* duplications left the ERV intact, as illustrated by the existence of four ERV-containing *SPANX-N* genes in humans (the exact number of such genes in apes and monkeys remains to be determined). The emergence of the *SPANX-A/D* gene subfamily appears to be a more recent event, subsequent to the separation of the hominoid lineage from orangutan. Apparently, this subfamily evolved via duplication of one of the *SPANX-N* genes accompanied by deletion of the distal part of exon 2 and rapid divergence (Fig. 5). Notably, the phylogenetic tree of the *SPANX-A/D* subfamily is best compatible with independent amplification of these genes in gorilla, chimpanzee, and humans (Fig. 8).

Lineage-Specific Amplification of *SPANX-A/D* Genes in Humans. Rapid evolution of the *SPANX-A/D* subfamily and location of these genes within segmental duplications impede a routine PCR analysis of syntenic chromosomal segments that is required to detect lineage-specific duplications. To overcome this problem, we selectively isolated syntenic genomic segments from African great ape genomes (chimpanzee, bonobo, and gorilla) by TAR cloning in yeast (12). Because up to 15% divergence in DNA sequences does not prevent selective gene isolation by TAR, the targeting hooks were developed by using the human genomic sequence.

The *SPANX-C* locus was chosen as a target. Because *SPANX-C* resides within a ≈20-kb segmental duplication, the targeting sequences in the TAR vector were designed from unique sequences flanking *SPANX-C* (Fig. 6). The vector efficiently clones an 83-kb human genomic segment containing *SPANX-C* and *LDOC1* genes (26). The size of the TAR yeast artificial chromosomes isolated from apes was different from that of human clones (≈58 kb for chimpanzee and bonobo and ≈61 kb for gorilla). To clarify the size differences between human and ape clones, the isolates from bonobo and gorilla were sequenced. All isolates shared a high level of sequence identity (≈95%) within the *SPANX-C* flanking sequences. The size difference is due to the absence of the 20-kb internal sequence containing the *SPANX-C* gene in African great apes. Partial sequencing of the chimpanzee clone revealed a similar organization of this syntenic region. Because this 20-kb sequence corresponds to a series of segmental duplications in chromosome X, we conclude that at least the duplication that yielded *SPANX-C* occurred only in the human lineage (the alternative would require independent deletion of the same region in the gorilla, bonobo, and chimpanzee lineages, a highly unlikely event). The *SPANX-C* duplication is likely not polymorphic, because we failed to detect the *SPANX-C* null allele in a human population analysis that involved 200 individuals by PCR by using specifically designed primers (Table 2). A detailed sequence analysis of the gorilla TAR clone showed that its greater length, compared with the bonobo and chimpanzee sequence, was caused by a 3.4-kb insertion. This insertion contains an ORF homologous to several human ESTs (AL136558). The human gene corresponding to the ESTs consists of eight exons and spans

≈30 kb on chromosome 3. Thus, the intronless insert in gorilla appears to represent a reverse-transcribed duplication of this gene, most likely a retropseudogene. Identification of the recent insertion of the apparent pseudogene in the syntenic region in gorilla suggests that the *SPANX-C*-harboring locus is a rearrangement hot spot. Thus, analysis of TAR clones revealed the lineage-specific amplification of *SPANX-A/D* genes in humans.

Discussion

In this study, we describe the previously unnoticed *SPANX-N* gene subfamily, which apparently predates the rodent–primate divergence and gave rise to the better characterized *SPANX-A/D* subfamily, with which it shares the testis-specific expression pattern. The presence of the ERV in the intron and the microsatellite at the end of exon 2 in most of the *SPANX-N* genes and the lack of both these elements in the *SPANX-A/D* subfamily clearly define the succession of events during the evolution of the *SPANX* family: *SPANX-N* is the ancestral form, which gave rise to the *SPANX-A/D* subfamily in the hominoid lineage (Fig. 5).

With the exception of the *SPANX-N5*, all nine *SPANX* genes are located in two gene clusters separated by ≈2 Mb at Xq27. All *SPANX* genes reside within recent segmental duplications (27). Recombination between the segmental duplications seems to be the mechanism for the rapid expansion of *SPANX* genes. It has been estimated that duplicated genomic segments, which represent ≈5% of the human genome, emerged during the past ≈40 million years of primate evolution (28–31). However, some of the duplications involved in the evolution of the *SPANX-A/D* subfamily appear to be even more recent, because the segment containing the *SPANX-C* gene is missing in the syntenic genomic regions of African great apes. The *SPANX* family expansion could still be ongoing in humans.

Similar to other reproduction-related genes, particularly those specifically expressed in spermatozoa and testis, the *SPANX* family appears to have evolved rapidly during the ≈90 million years separating primates and rodents from their common ancestor. Moreover, even among the reproductive genes, the *SPANX* genes are outstanding, being among the most rapidly changing ones. The major episode of accelerated evolution apparently occurred after the separation of the hominoid lineage from orangutan and led to the emergence of the *SPANX-A/D* subfamily. Most of the reproductive genes not only evolve fast but also appear to be subject to positive (diversifying) selection, which is typically manifest in an excess of amino acid-changing over synonymous substitutions, $d_a/d_s > 1$ (23, 25). The *SPANX* genes are highly unusual in showing dramatic acceleration of evolution not only in synonymous but also in nonsynonymous positions, as compared to the presumably (nearly) neutral rate of intron evolution. This results in $d_a/d_s \approx 1$, which, at face value, would suggest nearly neutral evolution of the coding sequences of the *SPANX* genes. The more likely interpretation, however, is that *SPANX* evolution is indeed driven by strong diversifying selection, which, in this case, affects both the amino acid sequences of the protein products and the codon choice.

The underlying mechanisms of the accelerated evolution of genes involved in spermatozoa and seminal fluid production are thought to involve spermatozoa competition and a specific form of sexual selection, cryptic female choice between spermatozoa (32, 33). Restricted expression of the *SPANX* genes during spermatogenesis and presence of the SPANX proteins in ejaculated sperms (3) are compatible with the notion that these proteins could contribute to a spermatozoa's fitness, although their molecular function remains unknown. The apparent positive selection acting on synonymous positions is much more puzzling. The only obvious explanation is that the codon choice of *SPANX* genes is rapidly adapted to achieve a high translation rate. Notably, the 5′-flanking sequences of *SPANX* genes appeared to evolve somewhat faster than the intron sequence, albeit much slower than the coding sequences; this seems to be compatible with positive selection acting at the level of expression regulation.

The present findings have potential implications for understanding the genetic differences between humans and other primates. *SPANX* genes are specifically expressed in testis, where the sudden emergence of a novel gene could potentially lead to reproductive barriers and thus play a role in speciation (23, 25, 34). Thus, the emergence of the *SPANX-A/D* family in the hominoids could be directly related to the speciation within this lineage.

Identification of the *SPANX* gene ortholog in mouse could be critical for elucidating the function(s) of this gene family and its possible role in malignancies. In addition, the fact that *SPANX-N* subfamily genes are expressed in melanoma cell lines suggests these genes could be potential targets for cancer immunotherapy.

1. Westbrook, V. A., Diekman, A. B., Naaby-Hansen, S., Coonrod, S. A., Klotz, K. L., Thomas, T. S., Norton, E. J., Flickinger, C. J. & Herr, J. C. (2001) *Biol. Reprod.* **64,** 345–358.
2. Zendman, A. J., Ruiter, D. J. & Van Muijen, G. N. J. (2003) *Cell Physiol.* **194,** 272–288.
3. Westbrook, V. A., Diekman, A. B., Klotz, K. L., Khole, V. V., von Kap-Herr, C., Golden, W. L., Eddy, R. L., Shows, T. B., Stoler, M. H., Lee, C. Y., *et al.* (2000) *Biol. Reprod.* **63,** 469–481.
4. Sousa, M. & Carvalheiro, J. (1994) *Anat. Embryol.* **190,** 479–487.
5. Larsen, R. E. & Chenoweth, P. J. (1990) *Mol. Reprod. Dev.* **25,** 87–96.
6. Baccetti, B., Burrini, A. G., Collodel, G., Magnano, A. R., Piomboni, P., Renieri, T. & Sensini, C. (1989) *Gamete Res.* **23,** 181–188.
7. Aktas, S., Ozkan, S. & Cisneros, P. L. (1996) *Int. Urol. Nephrol.* **28,** 819–829.
8. Zendman, A. J., Zschocke, J., van Kraats, A. A., de Wit, N. J, Kurpisz, M., Weidle, U. H., Ruiter, D. J., Weiss, E. H. & van Muijen, G. N. (2003) *Gene* **309,** 125–133.
9. Zendman, A. J., Cornelissen, I. M., Weidle, U. H., Ruiter, D. J. & van Muijen, G. N. (1999) *Cancer Res.* **59,** 6223–6229.
10. Westbrook, V. A., Scoppee, P. D., Diekman, A. B., Klotz, K. L., Allietta, M., Hogan, K., Slingluff, C., Patterson, J., Frierson, H., Irvin, W. P., Jr., *et al.* (2004) *Clin. Cancer Res.* **10,** 101–112.
11. Wang, Z., Zhang, Y., Liu, H., Salati, E., Chiriva-Internati, M. & Lim, S. H. (2003) *Blood* **101,** 955–960.
12. Kouprina, N. & Larionov, V. (2003) *FEMS Microbiol. Rev.* **27,** 1–21.
13. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
14. Schuler, G. D., Altschul, S. F. & Lipman, D. J. (1991) *Proteins* **9,** 180–190.
15. Rost, B., Sander, C. & Schneider, R. (1994) *Comput. Appl. Biosci.* **10,** 53–60.
16. Saitou, N. & Nei, M. (1987) *Mol. Biol. Evol.* **4,** 406–425.
17. Kumar, S., Tamura, K., Jakobsen, I. B. & Nei, M. (2001) *Bioinformatics* **17,** 1244–1245.
18. Zhang, J., Rosenberg, H. F. & Nei, M. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 3708–3713.
19. Kimura, M. (1980) *J. Mol. Evol.* **16,** 111–120.
20. Yang, F., Hanson, N. Q., Schwichtenberg, K. & Tsai, M. Y. (2000) *Am. J. Med. Genet.* **95,** 385–390.
21. Lievers, K. J., Kluijtmans, L. A., Heil, S. G., Boers, G. H., Verhoef, P., van Oppenraay-Emmerzaal, D., den Heijer, M., Trijbels, F. J. & Blom, H. J. (2001) *Eur. J. Hum. Genet.* **9,** 583–589.
22. Makalowski, W. & Boguski, M. S. (1998) *J. Mol. Evol.* **47,** 119–121.
23. Wyckoff, G. J., Wang, W. & Wu, C. I. (2000) *Nature* **403,** 304–309.
24. Swanson, W. J., Clark, A. G., Waldrip-Dail, H. M., Wolfner, M. F. & Aquadro, C. F. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 7375–7379.
25. Swanson, W. J. & Vacquier, V. D. (2002) *Nat. Rev. Genet.* **3,** 137–144.
26. Nagasaki, K., Schem, C., von Kaisenberg, C., Biallek, M., Rosel, F., Jonat, W. & Maass, N. (2003) *Int. J. Cancer* **105,** 454–458.
27. Bailey, J. A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V., Schwartz, S., Adams, M. D., Myers, E. W., Li, P. W. & Eichler, E. E. (2002) *Science* **297,** 1003–1007.
28. Eichler, E. E. (2001) *Trends Genet.* **17,** 661–669.
29. Samonte, R. V. & Eichler, E. E. (2002) *Nat. Rev. Genet.* **3,** 65–72.
30. Guy, J., Spalluto, C., McMurray, A., Hearn, T., Crosier, M., Viggiano, L., Miolla, V., Archidiacono, N., Rocchi, M., Scott, C., *et al.* (2000) *Hum. Mol. Genet.* **9,** 2029–2042.
31. Courseaux, A., Richard, F., Grosgeorge, J., Ortola, C., Viale, A., Turc-Carel, C., Dutrillaux, B., Gaudray, P. & Nahon, J. L. (2003) *Genome Res.* **13,** 369–381.
32. Birkhead, T. R. & Pizzari, T. (2002) *Nat. Rev. Genet.* **3,** 262–273.
33. Ball, M. A. & Parker, G. A. (2003) *J. Theor. Biol.* **224,** 27–42.
34. Singh, R. S. & Kulathinal, R. J. (2000) *Genes Genet. Syst.* **75,** 119–130.