12 Whitehurst, C. *et al.* (1992) Nucleotide sequence of the intron of the germline human kappa immunoglobulin gene connecting the J and C regions reveals a matrix association region (MAR) next to the enhancer. *Nucleic Acids Res.* 20, 4929–4930
13 Avramova, Z. *et al.* (1998) Matrix attachment regions and structural colinearity in the genomes of two grass species. *Nucleic Acids Res.* 26, 761–767
14 Greally, J.M. *et al.* (1999) Conserved characteristics of heterochromatin-forming DNA at the 15q11–q13 imprinting center. *Proc. Natl. Acad. Sci. U. S. A.* 96, 14430–14435
15 Shabalina, S.A. *et al.* (2001) Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet.* 17, 373–376
16 Kondrashov, A.S. and Shabalina, S.A. (2002) Classification of common conserved sequences in mammalian intergenic regions. *Hum. Mol. Genet.* 11, 669–674
17 Blasquez, V.C. *et al.* (1989) Immunoglobulin kappa gene expression after stable integration. I. Role of the intronic MAR and enhancer in plasmacytoma cells. *J. Biol. Chem.* 264, 21183–21189
18 Yi, M. *et al.* (1999) Evidence that the Igkappa gene MAR regulates the probability of premature V–J joining and somatic hypermutation. *J. Immunol.* 162, 6029–6039
19 Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16, 111–120
20 Liebich, I. *et al.* (2002) S/MARt DB: a database on scaffold/matrix attached regions. *Nucleic Acids Res.* 30, 372–374
21 Strissel, P.L. *et al.* (1998) Scaffold-associated regions in the human type I interferon gene cluster on the short arm of chromosome 9. *Genomics* 47, 217–229
22 Rollini, P. *et al.* (1999) Identification and characterization of nuclear matrix-attachment regions in the human serpin gene cluster at 14q32.1. *Nucleic Acids Res.* 27, 3779–3791
23 Jarman, A.P. and Higgs, D.R. (1988) Nuclear scaffold attachment sites in the human globin gene complexes. *EMBO J.* 7, 3337–3344
24 Bode, J. *et al.* (1995) Scaffold/matrix-attached regions: structural properties creating transcriptionally active loci. *Int. Rev. Cytol.* 162A, 389–454
25 Ogurtsov, A.Yu. *et al.* (2002) OWEN: aligning long collinear regions of genomes. *Bioinformatics* 18, 1703–1704

0168-9525/03/$ - see front matter. Published by Elsevier Science Ltd.
doi:10.1016/S0168-9525(03)00016-7

Increase of functional diversity by alternative splicing

Evgenia V. Kriventseva[1], Ina Koch[2], Rolf Apweiler[1], Martin Vingron[2], Peer Bork[3,4], Mikhail S. Gelfand[5] and Shamil Sunyaev[3,6]

[1]European Bioinformatics Institute (EMBL-EBI), Hinxton Wellcome Trust Genome Campus, Hinxton, Cambridge, UK CB10 1SD
[2]Max-Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany
[3]European Molecular Biology Laboratory (EMBL), Meyerhofstrasse 1, 69117 Heidelberg, Germany
[4]Max-Delbruck Centre for Molecular Medicine, Robert-Roessle-Strasse 10, 13122 Berlin, Germany
[5]State Scientific Centre GosNIIGenetika, Dorozhny pr. 1, 113545 Moscow, Russia
[6]Present address: Genetics Division, Dept. Medicine, Brigham & Women's Hospital, Harvard Medical School, 20 Shattuck St., Thorn 914, Boston, Massachusetts 02115, USA

A large-scale analysis of protein isoforms arising from alternative splicing shows that alternative splicing tends to insert or delete complete protein domains more frequently than expected by chance, whereas disruption of domains and other structural modules is less frequent. If domain regions are disrupted, the functional effect, as predicted from 3D structure, is frequently equivalent to removal of the entire domain. Also, short alternative splicing events within domains, which might preserve folded structure, target functional residues more frequently than expected. Thus, it seems that positive selection has had a major role in the evolution of alternative splicing.

Several recent large-scale computational studies were directed towards analysis of mRNA expression to quantify the extent of alternative splicing [1–7], to identify groups of genes with an increased level of alternative splicing [7], and to relate the level of alternative splicing to the organism complexity [8]. Experimental studies of individual genes [9–12] revealed cases of functional differences between splice forms, differences in the cellular localization, and even pathogenic consequences of a changed ratio of expression of alternative splice variants. With only rare exceptions, these observations were attributed to insertions, deletions or substitutions of protein domains or other protein structural modules. Therefore, the impact of alternative splicing on functional diversification can be analysed by quantifying the effect of alternative splicing on protein domains and mapping of alternative splicing regions on protein functional and structural units.

We extracted all alternatively spliced protein isoforms in higher organisms with fully sequenced genomes (*Homo sapiens*, *Mus musculus*, *Drosophila melanogaster* and *Caenorhabditis elegans*) from the SWISS-PROT database [13], producing a set of 4804 splicing variants of 1780 proteins. We then mapped the alternatively spliced regions onto protein domain annotations of the InterPro resource [14], which combines entries from several major protein domain databases. Our data are presented as supplementary information at http://archive.bmn.com/supp/tig/March_2003.pdf.

To interpret the mapping, we tested whether the distribution of alternative splicing is random with regard to the protein domain architecture. As we did not observe any bias of alternative splicing towards N-terminal, C-terminal or central parts of protein sequences

Corresponding author: Shamil Sunyaev (ssunyaev@rics.bwh.harvard.edu).

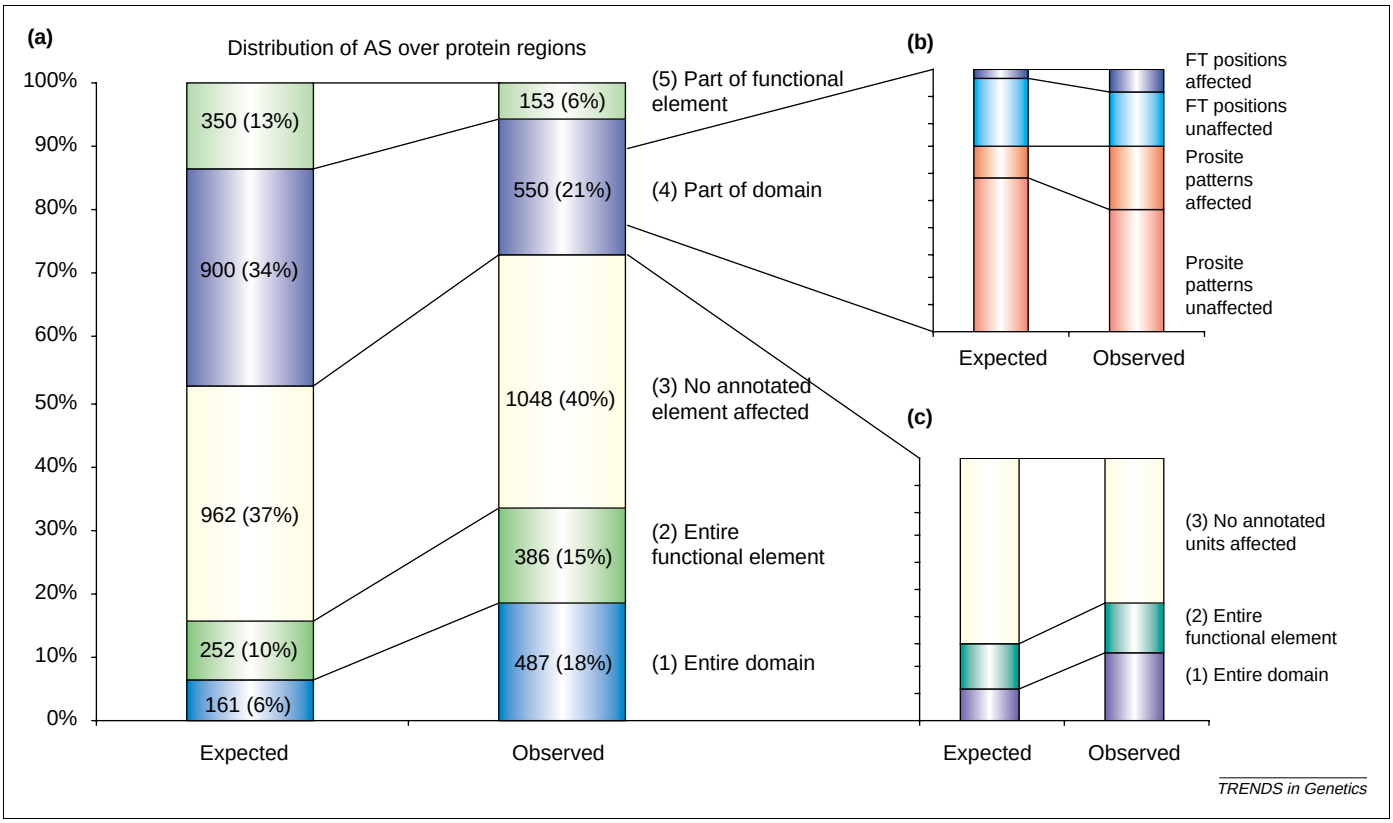

Fig. 1. The distribution of alternatively spliced regions in human with respect to the functional domains and structures in the proteins. (a) The observed distribution is compared with the distribution expected if alternative splicing were randomly located throughout the protein sequence. The events are divided into different classes: (1) alternative splicing inserts or deletes a complete protein domain (according to PROSITE profiles); (2) alternative splicing inserts or deletes entire functional elements that are not domains (i.e. transmembrane segments, signal peptides and coiled-coil regions); (3) alternative splicing does not affect any annotated or predicted protein functional elements; (4) alternative splicing inserts or deletes only part of a domain; (5) alternative splicing inserts or deletes only part of a non-domain functional element. The numbers of alternative splicing events in each class are shown and the percentage of the total for each class is in parentheses. (b) For cases where alternative splicing inserts or deletes a short fragment of a protein domain (less than 50 aa), the figure depicts the expected numbers of affected and unaffected functional sites. Only proteins with functional amino acids annotated in SWISS-PROT or in PROSITE were included in the analysis. The upper part of the figure corresponds to functional sites annotated in feature table (FT) in the SWISS-PROT entries. The lower part corresponds to sites annotated as PROSITE patterns. We focused solely on short alternative splicing within domains because long events are likely to destroy the domain structure. It is seen that short alternative splicing events within domains have a propensity to affect functional residues. (c) The expected and observed relative frequencies of alternative splicing events that either (1) insert or delete complete domains; (2) insert or delete complete non-domain functional elements or (3) have no effect on domains and other functional units. As opposed to (a), here partially affected elements are disregarded from the analysis.

(data not shown), our random model assumes a uniform placement of alternative splicing regions along the protein sequences while retaining the annotated domain architecture of proteins. The number of alternative splicing regions per protein and their length were assumed to match that of observed alternative splicing.

As shown in Fig. 1a, the proportion of alternative splicing boundaries within annotated domains is much lower than expected by chance; that is, alternative splicing tends not to occur within protein domains. This effect can be explained in two ways. First, the observation could simply reflect a correlation of exon and domain boundaries. Second, it could reflect the impact of negative natural selection, which would eliminate meaningless or even deleterious alternative splicing variants that might result from broken domains. To discriminate between these, we examined the boundaries of exons involved in constitutive splicing (i.e. not involved in alternative splicing), and we tested whether these constitutive splicing boundaries occur within domains more frequently than alternative splicing boundaries. The results clearly support the second possibility (Fig. 2).

The avoidance of disruption of structural elements is even stronger for elements that are not domains but that are structurally important, such as transmembrane regions, signal peptides and coiled-coil sequences (Fig. 1a).

The tendency of alternative splicing to occur outside of domains and other structural elements might, at a first glance, indicate that function is little affected by alternative splicing. However, here we show that alternative splicing that inserts or deletes whole domains affects function more than expected. To prove this, we modified our random expectation model and only considered a subset of alternative splicing variants that did not have split domains. Indeed, the observed number of insertions and deletions of complete domains in alternatively spliced isoforms is significantly higher than would be expected from random distribution of alternative splicing (Fig. 1c), even when disregarding partially affected domains. Alternative splicing variants that incorporate protein domains are over-represented, whereas alternative splicing affecting the inter-domain regions are under-represented. This suggests that alternative splicing that changes the domain architecture of proteins is favoured by selection.

Although the disruption of a domain sequence by alternative splicing is avoided in the statistical sense, a

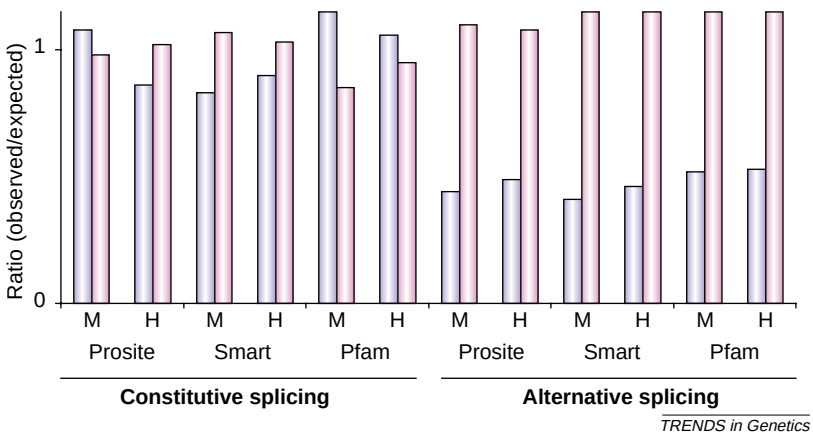

Fig. 2. Splicing events that occur within (blue) and outside (pink) of protein functional domains as annotated in the PROSITE Profiles, SMART and Pfam databases. The bars are ratios of observed numbers of splicing events to numbers expected under the assumption of the uniform random distribution of spliced regions. The graph shows ratios for constitutive splicing and alternative splicing for both *Homo sapiens* (H) and *Mus musculus* (M).

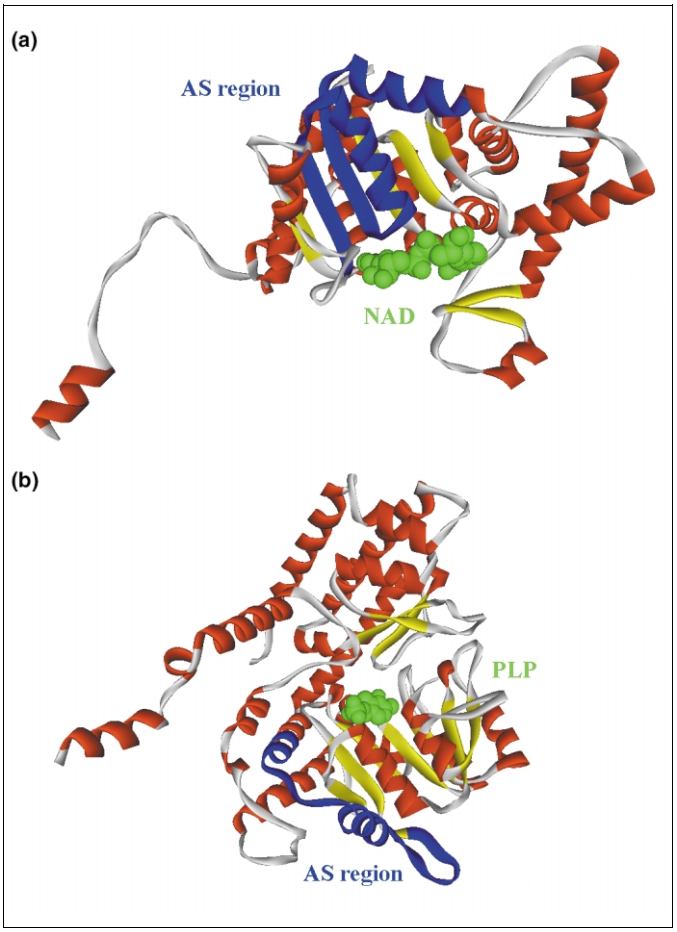

Fig. 3. Examples of structural location of alternatively spliced (AS) regions from the dataset of splicing variants annotated in SWISS-PROT. Longer isoforms are displayed, with the parts missing from the shorter isoforms coloured in dark blue. The ligand is coloured green. (a) Deoxynypusine synthase (PDB ID 1dhs, SWISS-PROT ID DHYS_HUMAN) catalyses the NAD-dependent oxidative cleavage of the spermidine. The region deleted by alternative splicing in the shorter isoform is relatively large and is located in the protein core closely to the ligand. The shorter isoform is known to be inactive. (b) Serine hydroxymethyltransferase (PDB ID 1bj4, SWISS-PROT ID GLYC_HUMAN) converts serine to glycine and is a key enzyme in the biosynthesis of purines, lipids, hormones, and other components. The region missing in the shorter isoform (isoform 2) is relatively short and is located on the protein surface.

considerable fraction (up to 28%) of alternative splicing variants do have split domains. The functionality of this type of alternative splicing depends on whether these isoforms can retain or even change their function.

The 3D structures of only 48 alternatively spliced proteins is known. Forty-three of the 71 alternative splicing variants that map to these proteins (60.5%) remove long essential parts of the domains, with a significant fraction of the hydrophobic core being deleted (e.g. Fig. 3a). We suggest that the resulting effect on function is probably equivalent to removal of the entire domain because only relatively short unfolded parts of the domain are retained in the shorter splicing isoforms. The remaining 28 alternative splicing variants do not affect most of the domain sequence and do not destroy the hydrophobic core (e.g. Fig. 3b). Of these, ten are sequence substitutions and 18 are short insertions or deletions. Two recent studies show that short alternative splicing in protein domains can leave the general structural scaffold unaffected and still be of functional importance [15,16].

To reveal whether insertions/deletions or substitutions of short regions of globular domains might constitute yet another mode of functional diversification by alternative splicing, we superimposed alternative splicing variants affecting functional residues in domains. From our set, we extracted alternative splicing variants that overlapped with a domain by less than 50 amino acids (e.g. in human, 106 such alternative splicing variants were identified). The annotation of functional sites was taken from the SWISS-PROT database [13] and from the database of PROSITE [17] patterns. Figure 1b shows that alternative splicing occurring inside protein domains preferentially targets functional amino acids (χ^2 test, P-value $<10^{-14}$). This suggests that alternative splicing frequently modulates function of protein domains either by inserting or deleting functional residues, and therefore possibly serving as a dominant–negative regulation mechanism [18], or by substituting the sequence that includes a functional site. Examples of the latter possibility include

the integrin α-7 (ITA7_HUMAN), where the fragment containing the GFFKR motif is substituted for a non-homologous fragment containing the same motif [19], and the Glandular Kallikrein 2 (KLK2_HUMAN), where sequence containing the serine residue involved in the charge relay system is substituted [20].

Under the natural assumption that mutational events leading to the emergence of multiple alternative splicing variants are randomly distributed, the observed non-random alternative splicing, tending to alter domain architecture and functional sites of proteins, suggests that positive Darwinian selection favours alternative splicing rather than splicing of unique variants. Such positive selection is strong evidence of the importance of alternative splicing in increasing functional diversity of proteomes.

Methods

The dataset of alternative splicing isoforms has been obtained by selecting all SWISS-PROT entries with the keyword 'alternative splicing' corresponding to proteins from *C. elegans*, *D. melanogaster*, *M. musculus* and *H. sapiens*. SWISS-PROT currently provides the largest set of experimentally identified splicing isoforms. Several organisms have been selected to verify that the observed effects are general and the organisms with fully sequenced genomes provide the most comprehensive set. VARSPLIC [21] routine was applied to these SWISS-PROT entries to form the complete collection of 4804 isoforms.

Sequence analysis of all splicing variants was performed by the InterProScan [22] tool. Protein domains were identified by characteristic signatures in SMART [23], Pfam [24], and PROSITE [17] Profiles. PROSITE [17] Patterns and SWISS-PROT [13] FT terms were used to select functional amino acids in protein domains.

The statistical analysis of the data included the following tests. First, the expected proportions of alternative splicing events, which insert or delete complete functional elements or affect these elements partially, were computed through direct enumeration via a sliding window. For each alternative splicing region from a given protein, a window of the length equal to the length of the alternative splicing region scanned the sequence of the protein and the counts of window positions, which affect functional elements as a whole or in part were enumerated. Then, the goodness-of-fit χ^2 test was used to compare the observed and expected proportions (Fig. 1a). Tests for all organisms and annotation databases considered resulted in highly significant *P*-values. In the test used to demonstrate the role of alternative splicing in functional diversification, we enumerated only those window positions, where both borders of the window were not inside a domain (Fig. 1c).

Second, we tested whether alternative splicing boundaries are underrepresented within domains and other structural modules. Under the assumption on the uniform random placement of alternative splicing regions on protein sequences, the expected number of alternative splicing boundaries inside protein domains would be proportional to the total number of amino acid residues in domains. We applied the χ^2 contingency test to demonstrate that alternative splicing boundaries in domains are under-represented. The test gave highly significant *P*-values in the range 10^{-5} to 10^{-148} for all considered organisms and all domain databases. The same test was applied to non-domain protein structural elements. We also repeated the same procedure for constitutive exons to verify that the observed effect is not due to positional correlation of exon and domain structures (Fig. 2).

Third, to demonstrate that short alternative splicing regions inside protein domain borders have a preference to affect functional amino acids, we applied the χ^2 contingency test. We tested whether the fraction of functional amino acids in alternative splicing regions is significantly higher than in constantly spliced regions (Fig. 1b). The test resulted in a highly significant *P*-value in range of 10^{-6} to 10^{-39} for SWISS-PROT annotated functional sites.

References

1 Hanke, J. *et al.* (1999) Alternative splicing of human genes: more the rule than the exception? *Trends Genet.* 15, 389–390
2 Mironov, A.A. *et al.* (1999) Frequent alternative splicing of human genes. *Genome Res.* 9, 1288–1293
3 Brett, D. *et al.* (2000) EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett.* 474, 83–86
4 Croft, L. *et al.* (2000) ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nat. Genet.* 24, 340–341
5 Kan, Z. *et al.* (2001) Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res.* 11, 889–900
6 Modrek, B. and Lee, C. (2002) A genomic view of alternative splicing. *Nat. Genet.* 30, 13–19
7 Modrek, B. *et al.* (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.* 29, 2850–2859
8 Brett, D. *et al.* (2002) Alternative splicing and genome complexity. *Nat. Genet.* 30, 29–30
9 Graveley, B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.* 17, 100–107
10 Lopez, A.J. (1998) Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. *Annu. Rev. Genet.* 32, 279–305
11 Roberts, G.C. and Smith, C.W. (2002) Alternative splicing: combinatorial output from the genome. *Curr. Opin. Chem. Biol.* 6, 375–383
12 Caceres, J.F. and Kornblihtt, A.R. (2002) Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends Genet.* 18, 186–193
13 Bairoch, A. and Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 28, 45–48
14 Apweiler, R. *et al.* (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* 29, 37–40
15 Peneff, C. *et al.* (2001) Crystal structures of two human pyrophosphorylase isoforms in complexes with UDPGlc(Gal)NAc: role of the alternatively spliced insert in the enzyme oligomeric assembly and active site architecture. *EMBO J.* 20, 6191–6202
16 Oakley, A.J. *et al.* (2001) The crystal structures of glutathione *S*-transferases isozymes 1-3 and 1-4 from *Anopheles dirus* species B. *Protein Sci.* 10, 2176–2185
17 Falquet, L. *et al.* (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.* 30, 235–238
18 Boise, L.H. *et al.* (1993) *bcl-x*, a *bcl-2*-related gene that functions as a dominant regulator of apoptotic cell death. *Cell* 74, 597–608
19 Leung, E. *et al.* (1998) A novel extracellular domain variant of the human integrin alpha 7 subunit generated by alternative intron splicing. *Biochem. Biophys. Res. Commun.* 243, 317–325
20 Liu, X.F. *et al.* (1999) Identification of three new alternate human

kallikrein 2 transcripts: evidence of long transcript and alternative splicing. *Biochem. Biophys. Res. Commun.* 264, 833–839
21 Kersey, P. *et al.* (2000) VARSPLIC: alternatively-spliced protein sequences derived from SWISS-PROT and TrEMBL. *Bioinformatics* 16, 1048–1049
22 Zdobnov, E.M. and Apweiler, R. (2001) InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17, 847–848
23 Letunic, I. *et al.* (2002) Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.* 30, 242–244
24 Bateman, A. *et al.* (2002) The Pfam protein families database. *Nucleic Acids Res.* 30, 276–280

0168-9525/03/$ - see front matter
doi:10.1016/S0168-9525(03)00023-4

Recombination explains isochores in mammalian genomes

Juan Ignacio Montoya-Burgos, Pierre Boursot and Nicolas Galtier

CNRS UMR 5000 – Génome, Populations, Interactions, Université Montpellier 2 – CC63, Place E. Bataillon, 34095 Montpellier Cedex, France

The mouse *Fxy* gene was translocated into the highly recombining pseudoautosomal region comparatively recently in evolutionary terms. This event resulted in a rapid increase of GC content. We investigated the consequences of the translocation further by sequencing exons and introns of *Fxy* in various rodent species. We found that the DNA fragment newly located in a highly recombining context has acquired every property of a GC-rich isochore, namely increased GC content (especially at the third codon positions of exons), shorter introns and high density of minisatellites. These results strongly suggest that recombination is the primary determinant of the isochore organization of mammalian genomes.

The pseudoautosomal region (PAR), a small region of homology between mammalian X and Y chromosomes, is intrinsically highly recombining because it is a short piece of DNA undergoing one obligatory cross-over per generation [1], like every bivalent. The *Fxy* gene is X specific in human and rat, but in the house mouse, *Mus musculus*, this gene was recently translocated and spans the pseudoautosomal boundary: its 3′ region (exons 4 to 10) now lies in the PAR [2]. Local recombination rate, therefore, has dramatically increased for the translocated 3′ *Fxy* in *M. musculus*. The detailed phylogenetic history of this gene was reconstructed by sequencing exons 8, 9 and 10 in nine rodent species (Fig. 1). Two copies of *Fxy* were found in *Mus spretus* (the previously known X-linked copy [2] and a new, presumably PAR-located copy). This indicates that a gene duplication occurred before the translocation event, 1–3 million years ago (that is, the date of the radiation between *M. spretus*, *M. musculus*, *Mus macedonicus* and *Mus spicilegus*) [3], followed by gene loss in *M. musculus*, in which only the PAR-specific copy could be amplified. Figure 1 confirms the previously reported increase in GC content at neutrally evolving third codon positions (GC3) in the PAR-located exons, together with an increase of evolutionary rate – a pattern not observed in exons 1–3, that have remained X specific [2]. This makes *Fxy* a unique 'natural laboratory' for investigating the relationship between recombination rate and isochore evolution [4] – an issue hotly debated during the past decade [5–7].

We contrasted the X-specific 5′ part and the PAR-located 3′ part of *Fxy* in *M. musculus* (data from ENSEMBL;

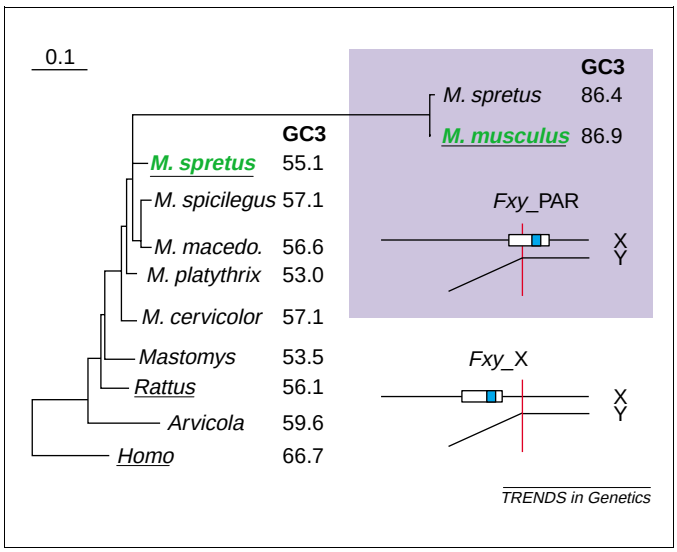

Fig. 1. Phylogenetic tree of *Fxy* in rodents. Exons 8, 9 and 10 (597 sites) were used. Nine rodent species were surveyed using PAR-specific primers (positive in *Mus musculus* and *Mus spretus*, top) and X-specific primers (positive in all but *M. musculus*, bottom). The GC content at redundant third codon positions (GC3) of each sequence (exons 8, 9 and 10, shown in cyan) is given. The topology was recovered using the maximum likelihood method [19]. Branch lengths were calculated from third codon positions using a model that accounts for variable GC content between sequences [20], removing a possible bias in the estimation of evolutionary rates. The two genes compared in further analyses are green. Genes whose chromosomal location has been determined experimentally [2] are underlined. Scale bar is in unit of average per site substitution rate. Diagrams show relationship of *Fxy* (white box) to the pseudoautosomal boundary (red line). GenBank accession numbers: *Homo*, AF035360; *Arvicola*, AY181220-22; *Rattus*, AF186461; *Mastomys*, AY181223-25; *Mus cervicolor*: AY181226-28; *Mus platythrix*: AY181232-34; *Mus macedonicus*, AY181229-31; *Mus spicilegus*, AY181235-37; *M. spretus Fxy*_X: AF186460; *M. spretus Fxy*_PAR: AY181238-40; *M. musculus*: AF026565.

Corresponding author: Nicolas Galtier (galtier@univ-montp2.fr).