Genet. Res., Camb. (2003), **82**, *pp.* 1–18. *With 5 figures.* © 2003 Cambridge University Press
DOI: 10.1017/S0016672303006268 Printed in the United Kingdom

Conserved fragments of transposable elements in intergenic regions: evidence for widespread recruitment of MIR- and L2-derived sequences within the mouse and human genomes

J. C. SILVA[1]*, S. A. SHABALINA[1], D. G. HARRIS[2], J. L. SPOUGE[1] AND A. S. KONDRASHOV[1]

[1] *National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20892, USA*
[2] *National Security Agency, Fort Meade, MD 20755, USA*

(*Received 30 April 2002 and in revised form 22 October 2002 and 29 January 2003*)

Summary

We analysed the distribution of transposable elements (TEs) in 100 aligned pairs of orthologous intergenic regions from the mouse and human genomes. Within these regions, conserved segments of high similarity between the two species alternate with segments of low similarity. Identifiable TEs comprise 40–60% of segments of low similarity. Within such segments, a particular copy of a TE found in one species has no orthologue in the other. Overall, TEs comprise only approximately 20% of conserved segments. However, TEs from two families, MIR and L2, are rather common within conserved segments. Statistical analysis of the distributions of TEs suggests that a majority of the MIR and L2 elements present in murine intergenic regions have human orthologues. These elements must have been present in the common ancestor of human and mouse and have remained under substantial negative selection that prevented their divergence beyond recognition. If so, recruitment of MIR- and L2-derived sequences to perform a function that increases host fitness is rather common, with at least two such events per host gene. The central part of the MIR consensus sequence is over-represented in conserved segments given its background frequency in the genome, suggesting that it is under the strongest selective constraint.

1. Introduction

Only a small proportion of the human genome, currently estimated to be less than 5%, consists of protein-coding DNA. In contrast, repeated DNA constitutes well over 45% of the human genome (International Human Genome Sequencing Consortium, 2001). Most of this repeated DNA is comprised of transposable elements (TEs) and their remnants. When their prevalence in genomes was first acknowledged, TEs were thought of as strictly selfish entities (Doolittle & Sapienza, 1980; Orgel & Crick, 1980). However, a growing number of cases are being reported in which TE sequences have been domesticated, i.e. recruited to perform some function that increases host fitness (Makalowski, 1995; Britten, 1997; Kidwell & Lisch, 1997; Miller *et al.*, 1997; Brosius, 1999; Smit, 1999; Makalowski, 2000; Nekrutenko & Li, 2001). Still, the prevalence of this phenomenon remains unknown.

Comparative genomics allows the detection of functionally important motifs by finding regions of similarity among otherwise highly divergent genomes. This makes it a powerful approach to studying different aspects of selection, such as the domestication of TEs, using distantly related species. Comparison of human and murine orthologous genome regions suggests mutational saturation between these species. The average number of substitutions per synonymous site (K_S) between the two species lies somewhere between 50% and 70% (Castresana, 2002; Mouse Genome Sequencing Consortium, 2002). Even if only a small fraction of all synonymous sites is under selective constraints (Bustamante *et al.*, 2002; Castresana, 2002), the average number of mutations per site between the two species will be considerably higher. In addition, the fraction of sites affected by insertions/deletions is very similar to that of sites affected by nucleotide substitutions (Silva & Kondrashov, 2002),

* Corresponding author. Current address: The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA. Tel: +1 (301) 8380200. Fax: +1 (301) 8380208. e-mail: jsilva@tigr.org

which further decreases any residual similarity between species. Finally, it has been shown experimentally that most non-coding sequences are not alignable between mouse and human (Jareborg *et al*., 1999; Shabalina *et al*., 2001). Indeed, this approach has been used to successfully identify conserved motifs in the mammalian genome (Hardison *et al*., 1997; Makalowski & Boguski, 1998; Hardison, 2000; Wasserman *et al*., 2000). These observations make this an ideal pair of species on which to use comparative genomics to study quantitatively the phenomenon of domestication of mammalian TEs.

Mammalian TE repeats can be split into four major classes (Smit, 1999): SINEs (short interspersed elements), LINEs (long interspersed elements), LTRs (long terminal repeat-containing retrotransposons) and DNA elements. The first three classes are composed of retrotransposons, i.e. TEs that transpose via an RNA intermediate. DNA elements transpose directly as DNA. Each of these classes is composed of several major families, which can differ in length, structure and gene composition and organization (Finnegan, 1992). Most families are in turn composed of subfamilies; these subfamilies correspond to different clades in the phylogeny of the family, and differ from each other by characteristic mutations, such as point mutations and/or insertions or deletions. Therefore, subfamilies of a single family have different DNA consensus sequences and may also differ in length. Within all four major mammalian TE classes there are old subfamilies whose age predates the split of the primate and murine lineages (Smit, 1993; Jurka *et al*., 1995; Smit *et al*., 1995; Smit & Riggs, 1996; Bénit *et al*., 1999; Malik *et al*., 1999). Barring horizontal TE transfer, only elements from such old subfamilies can be present in both human and mouse.

The life history of an old TE family present in two hosts can be described by one of two scenarios. First, elements from this family may have stopped transposing before the host species split from their common ancestor. If so, each element in one species must have an orthologue in the other, with the exception of cases where an element was lost later due to deletions or to the accumulation of point mutations. Alternatively, if transpositions continued after the split, a fraction of the elements in each species does not have, and has never had, an orthologue in the other species.

It has been argued that the transposition of MIR (a SINE family) and L2 (a LINE family) ceased before the primate–rodent split (Jurka *et al*., 1995; International Human Genome Sequencing Consortium, 2001) and so these families fit into the first scenario. If so, there should be no detectable similarity between selectively neutral orthologous elements in mouse and human. In addition, since the rate of molecular evolution in rodents has been higher than in primates (Li, 1997), intragenomic similarity between different, selectively neutral copies of MIR and L2 within the murine genome should also be negligible. Despite this, there are many murine fragments of MIR and L2 that diverge only by approximately 20% from their respective mammalian consensus sequences, and by approximately 30% from other elements of their family. This would imply that ubiquitous copies of murine MIRs and L2s, which are highly similar to many other murine (and human) elements, have all been under substantial negative selection.

We studied sequences of TE origin within 100 pairs of orthologous intergenic regions from the mouse and human genomes. These regions have been previously aligned (Shabalina *et al*., 2001). For each TE family, we compared its distribution in the two species in order to assess orthology between TE-derived sequences, and determined their densities within conserved (hits) and diverged (interhits) segments of intergenic regions. We have found that the locations of MIR and L2 elements are correlated in the two species, that the number of aligned elements between species is higher than expected by chance, and that these families are over-represented in conserved segments. These observations are best explained by the orthology of a substantial fraction of murine and human MIRs and L2s, implying their recruitment in the common ancestor and negative selection since then. A few other less common TE families show evidence of the same phenomenon, if to a lesser extent.

2. Materials and methods

(i) *Data*

Our sample of 100 pairs of complete, orthologous intergenic regions of a total length of approximately 10^6 nucleotides is located within 41 human and 38 murine continuous segments of the genome located on 12 different chromosomes in each of the two species (Shabalina *et al*., 2001). Each of these continuous fragments, which together form our genomic sample of approximately 6×10^6 nucleotides for each species, was analysed for their composition in TEs (Fig. 1*a*), using RepeatMasker versions 05051999 and 04042000 (A. F. Smit & P. Green, unpublished; available at http://repeatmasker.genome.washington.edu), and the rodent and human repeat databases in Repbase versions 05051999 and 04042000, available at http://www.girinst.org/server/Repbase (Jurka, 1998, 2000; Smit, 1999). RepeatMasker, in conjunction with Repbase, is currently the most powerful software to detect TE sequences (Rogozin *et al*., 2000). For methodological reasons we chose to follow strictly the nomenclature used by RepeatMasker. Even though this nomenclature is not standard (e.g. strictly speaking *mariner* elements are part of the Mer2_type family, etc.) and the identification of some elements is wrong (e.g. elements categorized as Alu in mouse are most

https://doi.org/10.1017/S0016672303006268 Published online by Cambridge University Press

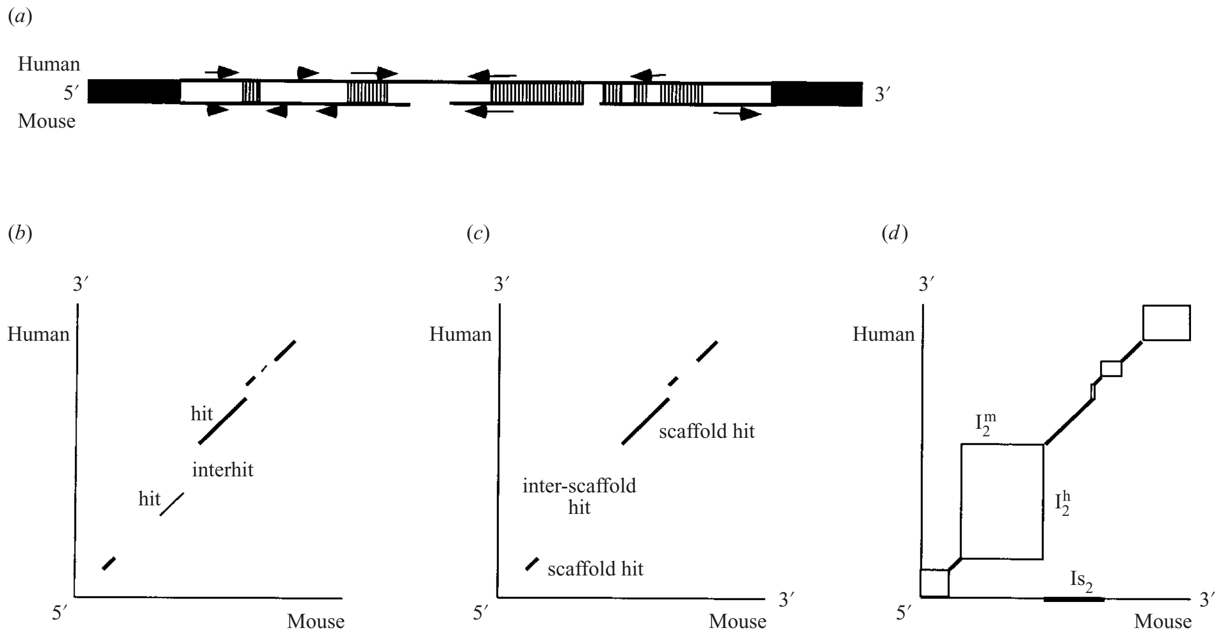

Fig. 1. Alignment of orthologous intergenic regions. (*a*) One-dimensional plot: top and bottom continuous lines represent the human and mouse sequences, respectively. Two insertion/deletion events are shown by an interruption in the mouse sequence. The genes that flank the spacer on its 5′ and 3′ ends (black boxes) and spacer regions where similarity exceeds 50 %, called hits, (hatched boxes) are shown. Arrows represent the location, length and orientation of transposable element (TE) sequences in the spacer. (*b*) Two-dimensional representation: vertical and horizontal axes represent human and mouse sequences, respectively. Hits are represented by diagonal lines; both hits formed mostly of TE sequences (thin lines) and those that are mostly TE-free (thick lines) are shown. Interhits are intergenic regions where similarity between the two sequences is lower than 50 %. (*c*) Similar to (*b*) but only hits that are mostly TE-free, called scaffold hits, are shown. Regions between scaffold hits are called inter-scaffold hits. (*d*) Scaffold hits are represented by diagonal lines. As an example, the length of the second scaffold hit in mouse (ls_2) is shown along the horizontal axis. Rectangles represent the area formed by inter-scaffolds, i.e. the product of inter-scaffold lengths in mouse and human (e.g. $l_2^m * l_2^h$, for the second inter-scaffold).

likely B1, etc.), the use of this classification is conservative for the purpose of determining the number of TE matches expected by chance (see below). For example, even though the mouse genome has no Alu elements, mouse sequences identified as such may be similar enough to human Alu elements to be aligned to them. This possibility is accounted for in the calculation of the expected number of matches. In this study we use mostly the element's family affiliation. When necessary, subfamily is also mentioned.

(ii) *Alignments*

All alignments were performed with Owen (ftp://ftp.ncbi.nih.gov/pub/kondrashov/owen), which uses a hierarchical procedure. First, all repeat sequences identified by RepeatMasker and by a histogram routine built into the software were masked. The remaining, unique sequences go through a first round of alignment (hash word size = 12), performed with stringent values for the alignment parameters (6 matches required within 8 consecutive nucleotides; minimum hit length = 40 nucleotides, $P < e^{-10}$). Hits are then resolved into a non-conflicting linear arrangement of alignments. Subsequent rounds of alignment of unique sequences are performed with progressively less stringent values for the parameters (down to 4 matches per 8 nucleotides and hit length ⩾25 nucleotides or 5 matches out of 8 and hit length ⩾15, and $P < e^{-5}$ in both cases). New hits are only allowed between the boundaries of others found in previous rounds. When no more hits of unique sequences are found, repeat sequences are unmasked and aligned as described. Note that only hits that do not conflict with existing ones are allowed and hence repeat sequences in either species will only be aligned to others that are located between the same pair of hits. In the end, a linear arrangement of hits (aligned segments within which between-species similarity exceeds 50 %) containing aligned unique and/or repeat sequences is obtained; hits are separated by regions of lower or no similarity that we call interhits (Fig. 1*b*). Hits are therefore conserved non-coding DNA segments, while interhits consist mostly of freely evolving, or unconstrained, DNA and TEs which jumped after the human–mouse split.

(iii) *Frequency of TEs in genome, intergenic regions and hits*

The number of elements of each TE family in the mouse and human genomic and intergenic region

https://doi.org/10.1017/S0016672303006268 Published online by Cambridge University Press

samples is defined as the number of fragments reported by RepeatMasker for that family. Note that if a particular element is split into multiple fragments by the insertion of repeat sequences within its boundaries, each of the resulting fragments is counted independently. The length of each TE family in the genomic and intergenic samples is defined as the cumulative count of nucleotide positions of all fragments of the family. The number and length of TE-derived sequences in hits are defined, respectively, as the number of elements that overlap hits in at least one nucleotide position, and the cumulative length of those overlaps. Counts requiring that the length of the overlap be at least 15 nucleotides did not differ much from these (results not shown).

The distribution of the number and length of TE sequences in hits given a random distribution of TEs within intergenic regions was estimated using Monte Carlo simulations. If the location of TEs is independent of that of hits, then the location of TEs should be uniformly distributed over the space of all possible arrangements of the TEs within an intergenic region, provided that no two TEs overlap.

We adopted the following procedure to randomly and uniformly choose arrangements from that space. For each intergenic region the following characteristics were recorded: length (l_{ig}), hit positions in the intergenic region, number of TEs, TE family and TE length (l_{TEi}). The cumulative length of all TE sequences ($\sum l_{TEi}$) was subtracted from the length l_{ig}, giving rise to a temporary intergenic region length tl_{ig}. Each TE was assigned a new insertion position, pushing all other positions (either TE-free or not) up by one and increasing tl_{ig} by one nucleotide. After all TEs had been inserted, their lengths were added following the insertion position, pushing all subsequent positions up by $l_{TEi}-1$. This brings tl_{ig} up to l_{ig}. The number and length of TEs that overlap hits were recorded as described previously. The process was repeated for all intergenic regions and results summed over all regions. The expected value of each of the two statistics (number and length of TEs in hits) was obtained by averaging over all replicates. The P value of the initial observation is approximately the fraction of replicates smaller than the observation (Efron & Tibshirani, 1993, p. 170). These simulations were performed both for all TEs as a whole and for individual TE families, and in each case consisted of 1000 replicates.

(iv) *Number of alignments,* x_A*, and the number of matches,* x_M

For each TE family, the number of alignments x_{A} corresponds to the number of times a pair of its elements is aligned (Fig. 2). This alignment must be reciprocal (i.e. same TE family in both species) and homologous (same region of the element in both species, and in the same orientation). Also, TE-derived sequences must be aligned for over 20 nucleotides. Finally, if multiple reciprocal TE alignments include the same TE, only one is counted (Fig. 2).

A realistic model to study the expected number of alignments is hard to design as the parameters of TE transposition (e.g. what determines the location and orientation of new insertions) and TE distribution (e.g. degree of clustering and fragmentation) are largely unknown. We devised a relatively simple statistic called the number of matches, x_{M}, which is a measure of the maximum number of ways in which m mouse elements can be aligned to h human elements present in orthologous genome regions, that is $m*h$. Because this is the maximum number of possible alignments, we will use the significance of x_{M} as a measure of how significant x_{A} is. While x_{M} is based on a simple biological model, it has the advantage of being mathematically tractable, unlike statistics based on more complex models.

In order to calculate x_{M}, a scaffold of unique hits was first created for each intergenic region. Unique hits are hits that have at least 50 aligned nucleotides that are repeat-free (i.e. not composed of TE sequences) in both species (Fig. 1). The scaffold is created by removing from the complete group of hits all those that are not unique (compare Fig. 1*b* and 1*c*). Scaffold hits are separated by inter-scaffold regions. x_{M} is the sum, over all intergenic regions, of the product, between species, of the number of fragments of a TE family in each inter-scaffold region.

(v) *The expected number of matches,* X_M

The number of matches expected by chance given the random location of TEs, X_{M}, corresponds to the sum, over all intergenic regions, of the product of the density of the TE family in mouse by its density on human sequences by each inter-scaffold area, i.e. the product of the length of the inter-scaffold in the two species (Fig. 1*d*).

X_{M} for each family was estimated according to four models. In models I and II, called 'global' models, the density of a TE family is averaged over all intergenic regions. Because the density of TE families can differ greatly among intergenic regions, we defined models III and IV. In these models, called 'local' models, X_{M} is calculated individually for each intergenic region (based on TE density in that region) and the values added over all regions. In models I and III, X_{M} is estimated for all inter-scaffold regions. Clearly, the contribution of each inter-scaffold region to X_{M} is roughly proportional to the square of its length, because it is dependent on the area (compare the area of the first and second inter-scaffold regions in Fig. 1*d*). Therefore, a few very large inter-scaffold regions can

https://doi.org/10.1017/S0016672303006268 Published online by Cambridge University Press

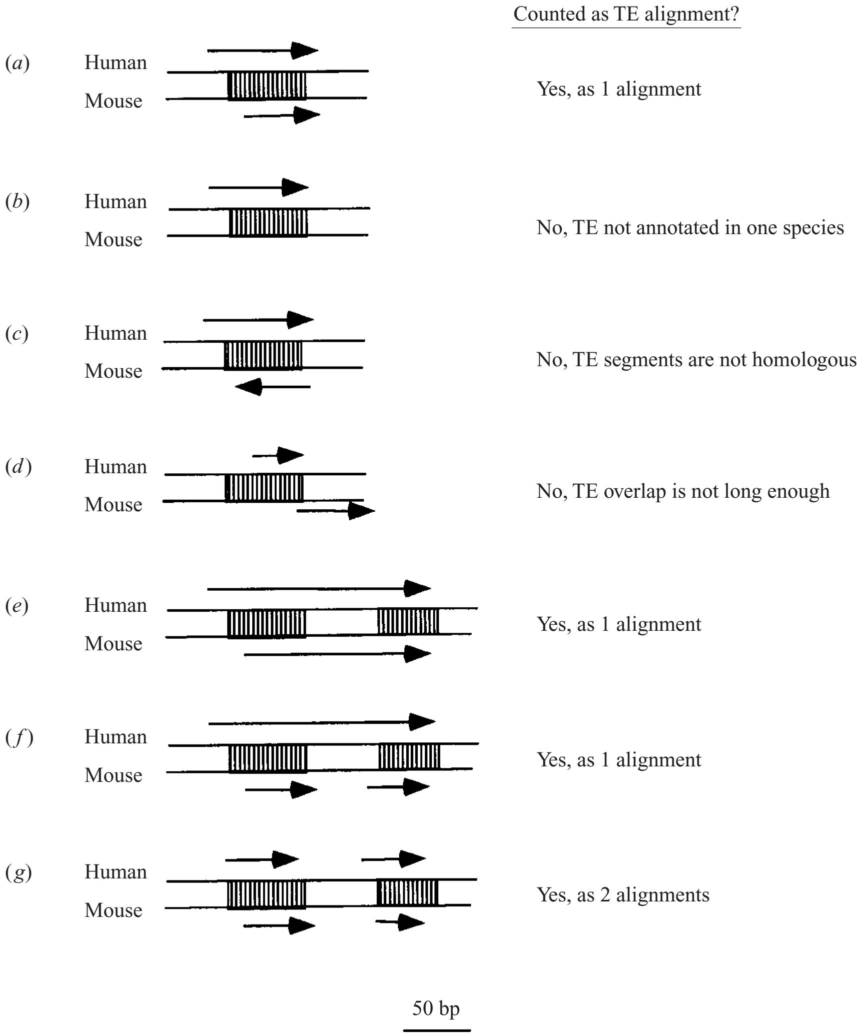

Fig. 2. Schematic of TEs in hits. Top and bottom horizontal lines represent the human and mouse sequences, respectively. Hatched boxes represent hits. Arrows represent TE sequences. (*a*) TE sequences in the two species are aligned to each other and >20 nucleotides of the alignment correspond to TE sequence. TEs have the same orientation and if the aligned region is homologous this arrangement will be scored as one alignment. (*b*) TEs are in opposite orientation. This is a chance event, not one due to conservation of orthologous sequences. (*c*) Elements do not overlap each other and a hit by ⩾20 nucleotides. (*d*) The mouse sequence was not annotated as TE. TEs in arrangements such as shown in (*b*), (*c*) and (*d*) will not be scored as alignments. Some TE sequences form two or more significant alignments with one (*e*) or more (*f*) elements in the other species. Because the same element is part of consecutive hits, these hits are not independent. In such cases, only one alignment is scored. (*g*) When two consecutive hits are formed by different TEs, they are all scored.

dramatically inflate X_{M} (the effect of these large regions disappears once inter-scaffold regions large than 10 kb are excluded). Models II and IV prevent this effect, by considering only inter-scaffold regions of length shorter than 1 kb in both species.

In models I and II, X_{M} is given by

$$\overline{D}_h \times \overline{D}_m \times \sum_{n=1}^{N} \sum_{i=1}^{k} (l_i^h \times l_i^m), \qquad (1)$$

where $\overline{D}$ is the density of a TE family, N is the total number of intergenic regions, k is the number of inter-scaffold regions in intergenic region n, l_i is the length of inter-scaffold region i (see Fig. 1*d*), and subscripts m and h stand for mouse and human, respectively. In these two models the density of a TE family, $\overline{D}$, is averaged over all intergenic regions as

$$\overline{D} = \frac{\sum_{n=1}^{N} m_n}{\sum_{n=1}^{N} L_n - \sum_{n=1}^{N} \sum_{i=1}^{s} ls_i}, \qquad (2)$$

where m is the number of fragments of the family in intergenic region n, L_n is the length of region n, s is the total number of scaffold hits in an intergenic region and ls_i is the length of scaffold hit i.

https://doi.org/10.1017/S0016672303006268 Published online by Cambridge University Press

X_{M} in models III and IV is then calculated independently for each intergenic region and added over all regions, as follows:

$$\sum_{n=1}^{N}\left(D_h \times D_m \times \sum_{i=1}^{k}(l_i^h \times l_i^m)\right),$$

$$D=\frac{m}{L-\sum_{i=1}^{s} ls_i}. \qquad (3)$$

The variation of X_{M}, under models III and IV, was calculated as

$$\sigma^2\left(\sum_{i=1}^{k} M_i H_i\right) = \mathrm{E}\left(\sum_{i=1}^{k} M_i H_i\right)^2 - \left(\mathrm{E}\left(\sum_{i=1}^{k} M_i H_i\right)\right)^2, \qquad (4)$$

where σ^2 stands for variance and E for expectation. M_i and H_i represent, for mouse and human respectively, the multinomial distributions $(M_i)_{i=1}^k$ and $(H_i)_{i=1}^k$ for each intergenic region, where k is the number of inter-scaffold regions. These distributions have parameters $(m; (p_i)_{i=1}^k)$ and $(h; (q_i)_{i=1}^k)$, where m and h are the number of TEs of the family in mouse and human intergenic regions, and p_i and q_i are the probabilities, in each species, of finding a TE in inter-scaffold i. This probability is proportional to the length of inter-scaffold i, l_i:

$$p_i=\frac{l_i}{L-\sum_{j=1}^{s} ls_j}. \qquad (5)$$

The first term in (4) is the expected value of the square of X_{M} and the second term is the square of X_{M}. The X_{M}, or mean value μ is given by expression (6) (which is equivalent to expression (3)).

$$\mathrm{E}\left(\sum_{i=1}^{k} M_i H_i\right) = \sum_{i=1}^{k}\mathrm{E}(M_i)\mathrm{E}(H_i) = m \times h \sum_{i=1}^{k} p_i q_i = \mu. \qquad (6)$$

The expected value of the square of x_{M} was calculated as follows:

$$\mathrm{E}\left(\sum_{i=1}^{k} M_i H_i\right)^2 = \mathrm{E}\left(\sum_{i=1}^{k}\sum_{j=1}^{k} M_i H_i M_j H_j\right) = \sum_{i=1}^{k}\sum_{j=1}^{k}\mathrm{E}(M_i M_j)\mathrm{E}(H_i H_j), \qquad (7)$$

$$\mathrm{E}(M_i M_j) = \left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\left(\sum_{i=1}^{k} p_i e^{\theta_i}\right)^m\right]_{\theta=0} = mp_i\left[\frac{\partial}{\partial\theta_j}\left(e^{\theta_i}\left(\sum_{i=1}^{k} p_i e^{\theta_i}\right)^{m-1}\right)\right]_{\theta=0} = mp_i((m-1)p_j+\delta_{ij}) = m^2 p_i p_j + mp_i(\delta_{ij}-p_j),$$

where $\delta_{ij}=1$ for $i=j$, and $\delta_{ij}=0$ for $i \neq j$.

$\mathrm{E}(N_i N_j)$ can be computed analogously.

For several intergenic regions $n=1,\ldots,N$, let (μ_n, σ_n^2) be the corresponding means and variances for the number of matches, and let X_n be the corresponding observation. Under the assumption of uncorrelated random placement of elements in the two species, the following expression is a standard normal variate, with mean 0 and variance 1:

$$\frac{\sum_{n=1}^{N} X_n - \sum_{n=1}^{N}\mu_n}{\left(\sum_{n=1}^{N}\sigma_n^2\right)^{1/2}}. \qquad (8)$$

Therefore, the critical value for each significance level can be readily obtained from the statistical table for the normal curve.

The rationale for this statistic can be understood intuitively as follows: the numerator contains several terms, each having the form

$$X-\mu=\sum_{i=1}^{k} M_i H_i - \sum_{i=1}^{k} mhp_i q_i = \sum_{i=1}^{k}(M_i - mp_i)(H_i - hq_i) \qquad (9)$$

of a covariance (the product of two deviations of the observations from their means, one for each species). Therefore the statistic will be large when mouse and human TEs tend to be present in the same inter-scaffold regions.

3. Results

(i) *Density of TEs*

We compared densities of DNA sequences of TE origin (whole TEs or their fragments) within three classes of DNA sequences: (i) approximately $\sim 6 \times 10^6$ nucleotides of genomic DNA in which our sample of 100 intergenic regions is located, (ii) the sample of 100 intergenic regions, and (iii) hits within this sample. Clearly, (iii) is a subset of (ii) is a subset of (i). The data are presented in Tables 1 and 2. Since, in both species, densities of TEs within (i) and (ii) are rather similar, our sample of intergenic regions is representative, in

https://doi.org/10.1017/S0016672303006268 Published online by Cambridge University Press

Table 1. *Distribution of TEs among the mouse genome sample, intergenic regions and hits*

	Genome sample			Intergenic regions			Hits		
TE family	No.[a]	Length[b] (bp)	%[c]	No.[a]	Length[b] (bp)	%[c]	No.[a]	Length[b] (bp)	%[c]
SINE									
Alu	3097	362 166	5·33	526	63 047	6·33	179	13 870	3·94
B2	1586	255 286	3·76	295	47 556	4·49	69	2747	0·78
B4	1011	133 154	1·96	142	18 295	1·73	28	1111	0·32
ID	357	25 780	0·38	49	3673	0·35	8	261	0·07
MIR	230	24 637	0·36	42	4033	0·38	35	2905	0·83
LINE									
L1	1141	476 158	7·01	168	59 934	5·66	62	9629	2·74
L2	146	20 256	0·30	17	2242	0·21	14	1652	0·47
CR1	6	636	0·01	2	351	0·03	2	271	0·08
Other	5	596	0·01	1	55	0·01	1	55	0·02
LTR									
Retroviral	113	55 419	0·82	15	19 870	1·87	4	509	0·14
MaLR	710	178 176	2·62	128	30 126	2·84	32	2562	0·73
ERV	4	819	0·01	0	0	–	0	0	–
ERV1	38	16 773	0·25	12	3217	0·30	2	130	0·04
ERVL	42	13 376	0·20	15	3807	0·36	6	620	0·18
ERVK	109	48 790	0·72	14	3441	0·32	1	47	0·01
HERVK	6	2343	0·03	0	0	–	0	0	–
MER21 group	7	1098	0·02	1	319	0·03	1	199	0·06
MER73 group	1	93	0·00	1	93	0·01	1	93	0·03
MER4I group	8	1498	0·02	1	390	0·04	0	0	0·00
MER4I group?	1	147	0·00	0	0	–	0	0	–
LTR	190	56 498	0·83	32	8429	0·80	7	409	0·12
LTR?	16	3837	0·06	7	1327	0·13	0	0	–
DNA									
Mer1_type	175	26 082	0·38	29	5010	0·47	12	801	0·23
Mer1_type?	3	251	0·00	1	114	0·01	0	0	–
Mer2_type	78	11 777	0·17	12	1596	0·15	1	106	0·03
Tc2	1	152	0·00	0	0	–	0	0	–
Mariner	4	1093	0·02	0	0	–	0	0	–
AcHobo	10	1463	0·02	0	0	–	0	0	–
DNA	6	656	0·01	1	157	0·01	1	157	0·04
Total (TEs)		1 719 010	25·30		277 082	26·52		38 134	10·84
Total data set[d]		6 794 015			996 283			351 667	

[a] Number of fragments of each family in the genome sample, in intergenic regions and in hits.
[b] Cumulative length of all the TE fragments in the three data sets (genome sample, intergenic regions and hits).
[c] Percentage of the length of the data set that is covered by each TE family.
[d] Total length of the genomic sample and cumulative length of all 100 intergenic regions and all mouse hits.

this respect, of the genome. Indeed, both the total TE density and that of individual families in (i) and (ii) are in good agreement with the genomic data for both species (Smit, 1999; Gu *et al.*, 2000; International Human Genome Sequencing Consortium, 2001). In particular, the fraction of (i) and (ii) in human DNA that is comprised of detectable TE sequences is approximately 45%. For murine DNA, this fraction is only about 26%. While, to some extent, this may be due to a real difference in TE frequencies between the human and murine genomes, it also appears that TEs are substantially underdetected in mouse. In particular, fractions of hits that are composed of TEs must be the same in both species (assuming that a sequence alignable with a TE sequence must also be TE-derived); yet, the apparent between-species difference that is observed at the genome level is also observed in hits (Tables 1, 2).

In contrast, TEs are substantially under-represented in hits. TE-derived sequences that can be recognized as such cover only 11% and 21% of the cumulative length of the hits in mouse and human, respectively (Tables 1, 2). On the other hand, inter-hits are enriched with TE-derived sequences which comprise approximately 37% and 57% of their cumulative length in mouse and human, respectively. Thus, TEs are 2 to 3 times rarer in hits than in inter-hits. Monte Carlo simulations show that, as a whole,

https://doi.org/10.1017/S0016672303006268 Published online by Cambridge University Press

Table 2. *Distribution of TE families among the human genome sample, intergenic regions and hits*

	Genome sample			Intergenic regions			Hits		
TE family	No.[a]	Length[b] (bp)	%[c]	No.[a]	Length[b] (bp)	%[c]	No.[a]	Length[b] (bp)	%[c]
SINE									
Alu	4518	1 145 034	18·03	1022	258 456	20·49	270	21 690	6·08
B4	3	225	0·00	1	85	0·01	0	0	–
MIR	897	113 746	1·79	192	23 441	1·86	108	8667	2·43
LINE									
L1	1744	640 179	10·08	362	146 889	11·64	142	21 328	5·98
L2	842	176 657	2·78	182	36 880	2·92	91	10 909	3·06
CR1	31	4939	0·08	9	773	0·06	1	192	0·05
Other	41	10 913	0·17	10	1808	0·14	4	387	0·11
LTR									
Retroviral	120	51 376	0·81	24	7971	0·63	1	220	0·06
MaLR	448	123 101	1·94	120	33 036	2·62	36	4754	1·33
ERV	0	0	–	0	0	–	0	0	–
ERV1	221	78 650	1·24	53	21 417	1·70	8	841	0·24
ERV2	0	0	–	0	0	–	0	0	–
ERVL	77	25 258	0·40	23	8692	0·69	9	1395	0·39
ERVK	19	14 911	0·23	2	1541	0·12	1	97	0·03
HERVK	0	0	–	0	0	–	0	0	–
MER21 group	40	11 132	0·18	16	3320	0·26	2	346	0·10
MER73 group	4	1202	0·02	1	550	0·04	1	204	0·06
MER4I group	94	24 772	0·39	51	11 583	0·92	6	304	0·09
MER4I group?	1	314	0·00	0	0	–	0	0	–
LTR	22	5871	0·09	6	1304	0·10	1	79	0·02
DNA									
MER1_type	367	65 086	1·03	61	11 655	0·92	25	1721	0·48
MER1_type?	28	3451	0·05	3	435	0·03	1	131	0·04
MER2_type	157	39 664	0·62	44	10 659	0·84	6	589	0·17
Tc2	2	218	0·00	1	86	0·01	0	0	–
Mariner	19	3622	0·06	4	233	0·02	2	115	0·03
AcHobo	18	2729	0·04	2	435	0·03	0	0	–
DNA	30	3512	0·06	5	535	0·04	3	269	0·08
T2_type	12	2586	0·04	4	900	0·07	0	0	–
Total (TEs)		2 549 148	40·15		582 684	46·19		74 238	20·82
Total data set[d]		6 349 108			1 261 561			356 488	

[a] Number of fragments of each family in the genome sample, in intergenic regions and in hits.
[b] Cumulative length of all the TE fragments in the three data sets (genome sample, intergenic regions and hits).
[c] Percentage of the length of the data set that is covered by each TE family.
[d] Total length of the genomic sample and cumulative length of all 100 intergenic regions and all human hits.

such TE-derived sequences are significantly excluded from hits in both human and mouse (Table 3).

The same pattern holds for most individual TE families (Tables 1, 2). However, two common TE families, MIR and L2, and a few rare families, are exceptions to this rule in that their density is higher in hits. Monte Carlo simulations of the most common families show that Alu and L1 elements are significantly excluded from hits in both species. Conversely, MIRs or L2s are overrepresented in hits, even though that tendency is only highly significant in mouse (Table 3).

(ii) *Distribution of TEs*

We studied the correlation between the location of TE-derived sequences in the murine and human genomes, taking advantage of the fact that hits involving unique (as opposed to repetitive) sequences create a scaffold that defines a linear arrangement of orthologous compartments (see Section 2 and Fig. 1). We used as a statistic the number of TE matches, x_{M} (see Section 2). Note that a match between TE fragments does not require that the two sequences are effectively aligned (only that they are in the same inter-scaffold region), but that an alignment always implies that the two sequences are matched. Therefore, $x_{\mathrm{M}} \geqslant x_{\mathrm{A}}$, and usually x_{M} largely overestimates x_{A} (see Section 4).

The number of matches expected if TEs are distributed randomly, X_{M}, for each TE family was estimated according to four models, as described in detail in Section 2. The values for this expectation are

Table 3. *Number and length of TEs in hits, in human and mouse intergenic regions*

	No. of TEs[a]			Length of overlap[b]		
	Observed	Expected[c]		Observed	Expected[c]	
		mean (SD)	range		mean (SD)	range
Human						
All TEs	718 ($P \ll 0\cdot001$)	982 (±17·9)	936–1044	74 238 ($P \ll 0\cdot001$)	153 020 (±24 920)	121 500–238 400
SINE						
Alu	270 ($P \ll 0\cdot001$)	446 (±12·7)	401–484	21 690 ($P \ll 0\cdot001$)	56 022 (±2276)	49 930–72 820
MIR	108 ($P \approx 0\cdot013$)	92 (±6·2)	72–111	8667 ($P \approx 0\cdot053$)	7646 (±703)	5500–13 160
LINE						
L1	142 ($P \approx 0\cdot002$)	168 (±7·9)	141–193	21 328 ($P < 0\cdot001$)	46 047 (±22 284)	24 950–137 600
L2	91 ($P \approx 0\cdot267$)	87 (±6·2)	66–108	10 909 ($P \approx 0\cdot236$)	10 250 (±980)	7208–13 520
Mouse						
All TEs	466 ($P \ll 0\cdot001$)	748 (±15·7)	699–794	38 134 ($P \ll 0\cdot001$)	103 270 (±30 620)	72 040–194 770
SINE						
Alu	179 ($P \ll 0\cdot001$)	245 (±9·8)	210–274	13 870 ($P \ll 0\cdot001$)	18 850 (±1186)	15 780–25 610
MIR	35 ($P < 0\cdot001$)	26 (±3·0)	17–34	2905 ($P \ll 0\cdot001$)	1891 (±262·0)	1221–2685
LINE						
L1	62 ($P \ll 0\cdot001$)	105 (±5·5)	89–124	9629 ($P \ll 0\cdot001$)	42 780 (±30 595)	14 070–134 630
L2	14 ($P \approx 0\cdot101$)	11 (±1·8)	5–16	1652 ($P < 0\cdot001$)	953·0 (±215·2)	331–1644

[a] Number of TE sequences that overlap a hit by at least one nucleotide.

[b] Cumulative length of overlaps between TE sequences and hits.

[c] The expected value was determined according to a model of random distribution of TE sequences in intergenic regions, and was calculated using 1000 Monte Carlo simulations as described in Section 2.

Note: No Bonferroni correction has been applied to the P values.

consistently higher under 'local' models III and IV, which take into account the distribution of TEs among intergenic regions, than under 'global' models I and II, in which TE density is averaged across all intergenic regions (Table 4). This indicates that most TE families have a clustered distribution, an observation reported for some TE families (Smit & Riggs, 1995). Therefore, we will focus on the estimates obtained using the more accurate 'local' models.

For most TE families, x_M does not differ significantly from X_M (Table 4). However, for a few families $x_M \gg X_M$, either for all inter-scaffold regions (model III) or, at least, when very large inter-scaffold regions (>1000 nucleotides) are eliminated (model IV) (Table 4). Therefore, while for most TE families independent location of TE-derived sequences between species cannot be ruled out, for those few others for which $x_M \gg X_M$ the location of such sequences in the two species is strongly correlated.

(iii) *Alignment of TEs in human and mouse*

The evolutionary distance between the two species precludes similarity between species of ancestral sequences devoid of selective constraints. Therefore, a likely explanation for the excessive x_M is the inheritance of the same TE-derived sequences from the common ancestor of human and mouse, followed by selection on some of those sequences. If this is the case, TE-derived sequences with matching locations should also be alignable and, therefore, x_A should be close to x_M. This is observed in MIR, L2, and several rare families, where $x_A \cong x_M \gg X_M$ (Table 4). The fact that a very large fraction (>80%) of the mouse MIR and L2 elements present in intergenic regions are aligned to a human sequence also points to this conclusion (Table 1).

(iv) *Representation of MIR (SINE) and L2 (LINE) sequences in alignments*

The relatively large number of alignments and the length homogeneity among elements of MIR and L2 families (at present only one predominant subfamily has been identified for each) allow us to compare the frequency of nucleotide positions in hits with their background frequency in intergenic regions and in the genome sample.

Thirty MIR elements in each species have an orthologue to which they are aligned. Four of these 30 pairs of TEs are aligned across two separate hits, leading to a total of 34 alignments that include MIR sequences. The MIR consensus sequence can be divided into three segments (Fig. 3*a*). The intermediate segment, or 'core', is the most abundant in the genome (Fig. 3*b*), an observation that has been reported before (Smit & Riggs, 1995) but whose cause is unknown. In addition, the relative frequency of the three

https://doi.org/10.1017/S0016672303006268 Published online by Cambridge University Press

Table 4. *Number of matches and alignments between species of TE-derived sequences of all TE families*

	All inter-scaffold regions				Inter-scaffold smaller than 1000 nucleotides			
	Expected number of matches[a]		Observed numbers (model III)[b]		Expected number of matches[a]		Observed numbers (model IV)[b]	
	Model I	Model III (SD)	Matches	Alignments	Model II	Model IV (SD)	Matches	Alignments
SINE								
Alu	1117·09	3563·6 (202·74)	3841 NS	137	24·39	45·88 (9·54)	38 NS	12
B4	0·30	0·93 (0·26)	1 NS		0·01	0·001 (0·04)	0 NS	0
MIR	16·76	22·15 (5·92)	31 NS	30	0·37	3·18 (1·84)	24***	**18** (+10)***
LINE								
L1	126·38	264·21 (33·30)	495***	45	2·75	7·38 (3·31)	6 NS	2
L2	6·43	20·34 (4·83)	23 NS	13	0·14	0·63 (0·85)	5***	**3** (+3)**
CR1	0·04	0·04 (0·18)	1***	**1*****	0·001	0·002 (0·05)	0 NS	0
Other (HAL1)	0·02	0·37 (0·64)	3***	1	<0·001	0·01 (0·09)	0 NS	0
LTR								
Retroviral	0·75	1·74 (1·77)	0 NS	1	0·02	0·01 (0·10)	0 NS	
MaLR	31·92	62·16 (12·14)	68 NS	10	0·70	1·36 (1·28)	5**	1 (+1)
ERVL	0·72	2·17 (1·66)	3 NS	3	0·02	0·07 (0·27)	1***	**1*****
ERVK	0·06	0	0	0	0·001	0	0	0
ERV1	1·32	0·67 (1·09)	2 NS	0	0·03	0·03 (0·18)	0 NS	0
LTR	0·40	0·77 (0·75)	0 NS	0	0·001	0·04 (0·19)	0 NS	0
MER21 group	0·03	1·03 (1·13)	1 NS	1	0·001	0·03 (0·18)	0 NS	0
MER73 group	0·002	0·13 (0·33)	1**	**1****	<0·001	0·003 (0·06)	0 NS	0
MER4I group	0·11	0	0	0	0·002	0	0	0
DNA								
MER2_type	1·10	2·08 (1·73)	1 NS	1	0·02	0·08 (0·29)	0 NS	0
MER1_type	3·68	10·01 (2·98)	25***	6	0·08	0·39 (0·70)	3***	**2** (+1)*
DNA (MER53)	0·01	0·08 (0·27)	1***	**1*****	<0·001	0·004 (0·06)	0 NS	0
MER1_type?	0·01	0·09 (0·29)	1***	0	<0·001	0·001 (0·03)	0 NS	0

[a] Expected number of matches estimated according to models I and II ('global models') and models III and IV ('local models'). Standard deviation was calculated for models III and IV (see Section 2). Significance levels: not significant (NS); $0{\cdot}01 < P \leqslant 0{\cdot}05$ (*); $0{\cdot}001 < P \leqslant 0{\cdot}01$ (**); $P \leqslant 0{\cdot}001$ (***).

[b] Observed number of matches and alignments, under models III and IV. In parentheses is the number of additional alignments that are contained entirely within scaffold hits.

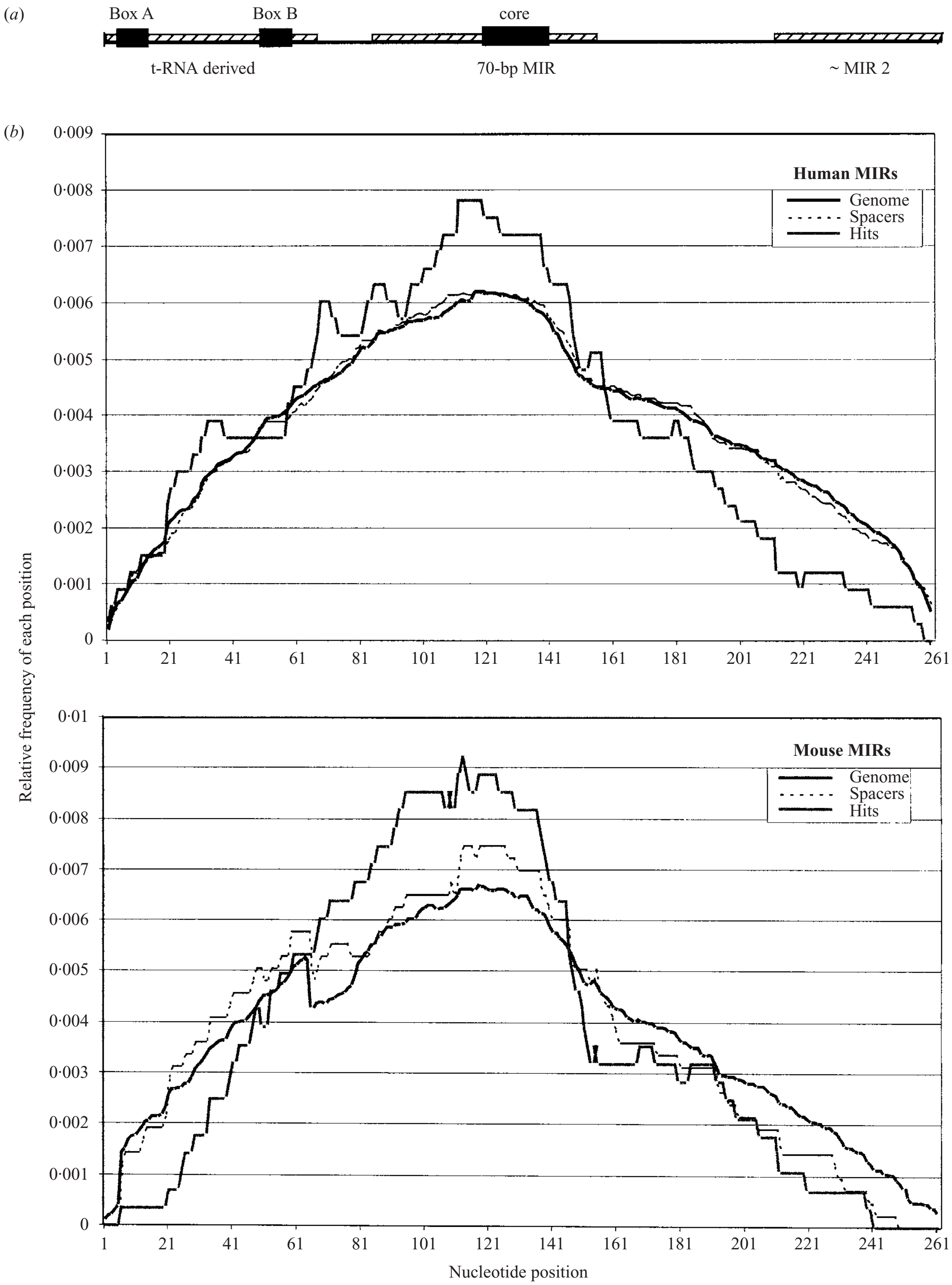

Fig. 3. The MIR subfamily. (*a*) Schematic of an element of the MIR subfamily. The first 70 nucleotides are derived from tRNA and contain an internal RNA polymerase III promoter (boxes A and B, located between nucleotides 5–15 and 50–60, respectively). The intermediate region (positions 85–155) is called the '70 bp MIR' and contains a conserved central part (positions 117–141) termed the 'core'. The 3′ end has similarity to MIR2 elements, which belong to the L2 family of LINEs (Smit & Riggs 1995). (*b*) Relative frequency of each position of the MIR consensus sequence in the genome (continuous line), intergenic regions (dotted line) and hits (grey lines) are shown for mouse ($n=221$, $n=40$ and $n=30$, respectively) and human ($n=820$, $n=177$ and $n=30$, respectively). The relative frequency of each position is the number of times that position is present in intergenic regions (or hits), divided by the frequency of all positions.

https://doi.org/10.1017/S0016672303006268 Published online by Cambridge University Press

MIR segments in hits differs from that in the genome. While the last 100 nucleotides of the MIR consensus are under-represented in hits relative to their genome frequency, the segment that corresponds to the '70 bp MIR' segment is present in hits in a relative frequency higher than that of the genome both species (Fig. 3).

There are thirteen L2 fragments aligned between the two species. One of the mouse fragments was aligned to two human L2s, for a total of 14 hits with L2 aligned fragments. The known consensus sequence of the L2 family is 3314 nucleotides long (Fig. 4). The background frequency distribution of the L2 consensus sequence reflects the outcome of its insertion mechanism. Integration of reverse-transcribed L2s often leads to a 5′-truncated element, the length of the truncated segment being variable. Consequently, the relative frequency of the different section of the element in the genome increases from 5′ to 3′. Two small sections approximately 300 nucleotides long that map to the last third of the L2-encoded ORF (positions 1750–2100 and 2300–2600) seem to be over-represented in hits relative to the genomic background frequency (Fig. 4). Because of the small number of L2 alignments used to draw the distribution of nucleotide frequency in hits, we constructed a similar plot using all 91 human L2 fragments that overlap hits, regardless of whether or not they are aligned to a mouse sequence annotated as being TE-derived. We found that now positions 2200 through 2500 are over-represented in hits (not shown). This region partially overlaps the second peak observed when only reciprocal TE alignments were used (Fig. 4). Conversely, the region between positions 2600 and 2900, which includes the last 100 nucleotides on the 3′ end of the open reading frame and some of the non-coding region of the element, is under-represented in hits relative to its background frequency.

This pattern cannot be explained by the presence of low-complexity regions in either MIR or L2 elements. There are no simple or tandem repeats in the aligned regions of these elements, and the MIR and L2 alignments are, on average, 80 and 120 nucleotides long, respectively, and approximately 55% AT-rich (only slightly less than the average AT composition of the human genome). Therefore, these are reliable and significant alignments.

4. Discussion

Our main findings are as follows. (1) Overall, TE-derived sequences are predominantly located in regions of low similarity between species, of which they can comprise over 50%; in contrast, they only cover approximately 20% of the total length of hits. (2) Contrary to what is observed for the other TE families, densities of MIR and L2 elements are 2 times higher in hits than in regions of low similarity. (3) The location of elements of the MIR, L2, and a few rare families is correlated between mouse and human. (4) A very large fraction (>75%) of MIR and L2 elements present in mouse intergenic regions have orthologous elements in the human genome. (5) For these families, the number of alignments is significantly higher than expected by chance, i.e. due to the independent insertion of TE into similar locations in the two species. (6) The central region of the MIR element, which corresponds to the '70-bp MIR', is over-represented in hits compared with its background frequency in the genome and in intergenic regions.

(i) *The exclusion of TE-derived sequences from hits*

The predominance of TE-derived sequences in interhits can be explained by the fact that a large fraction of the detectable mammalian TE sequences originated by transposition since the split of the murine and primate lineages (International Human Genome Sequencing Consortium, 2001). Usually, new TE copies will insert in a region that does not have a similar element in a related species and, therefore, this new TE will by definition be part of an interhit within the alignment of the genomes of the two species. The lack of a correlation between the location of most TE families in the two species supports this conclusion. In addition, the insertion of a TE in a conserved sequence, or hit, will either be eliminated by selection or cause the original conserved sequence to form two separate hits. This explains the exclusion of most TEs from hits.

(ii) *Correlation between the position of TEs in mouse and human*

We observed a correlation between species of the distribution of same-TE family sequences. The same conclusion has been reached, without quantitative analysis, by the Mouse Genome Sequencing Consortium (2002, pp. 533–534). This result indicates that the number of times that elements are found in the same inter-scaffold region in the two species (i.e. TEs are matched) exceeds the number expected by chance given the density of the TE family in the two host species. This correlation can arise in at least three ways.

First, elements could have a tendency to independently insert in the same genomic location in both host species. We are not aware of a biological reason why this should happen at the level of inter-scaffold regions. Second, if the original TE was very long, its subsequent fragmentation due to the insertion of other TEs within its boundaries will lead to a cluster of fragments. As the number of TE matches in an inter-scaffold region is the product of the numbers of

https://doi.org/10.1017/S0016672303006268 Published online by Cambridge University Press

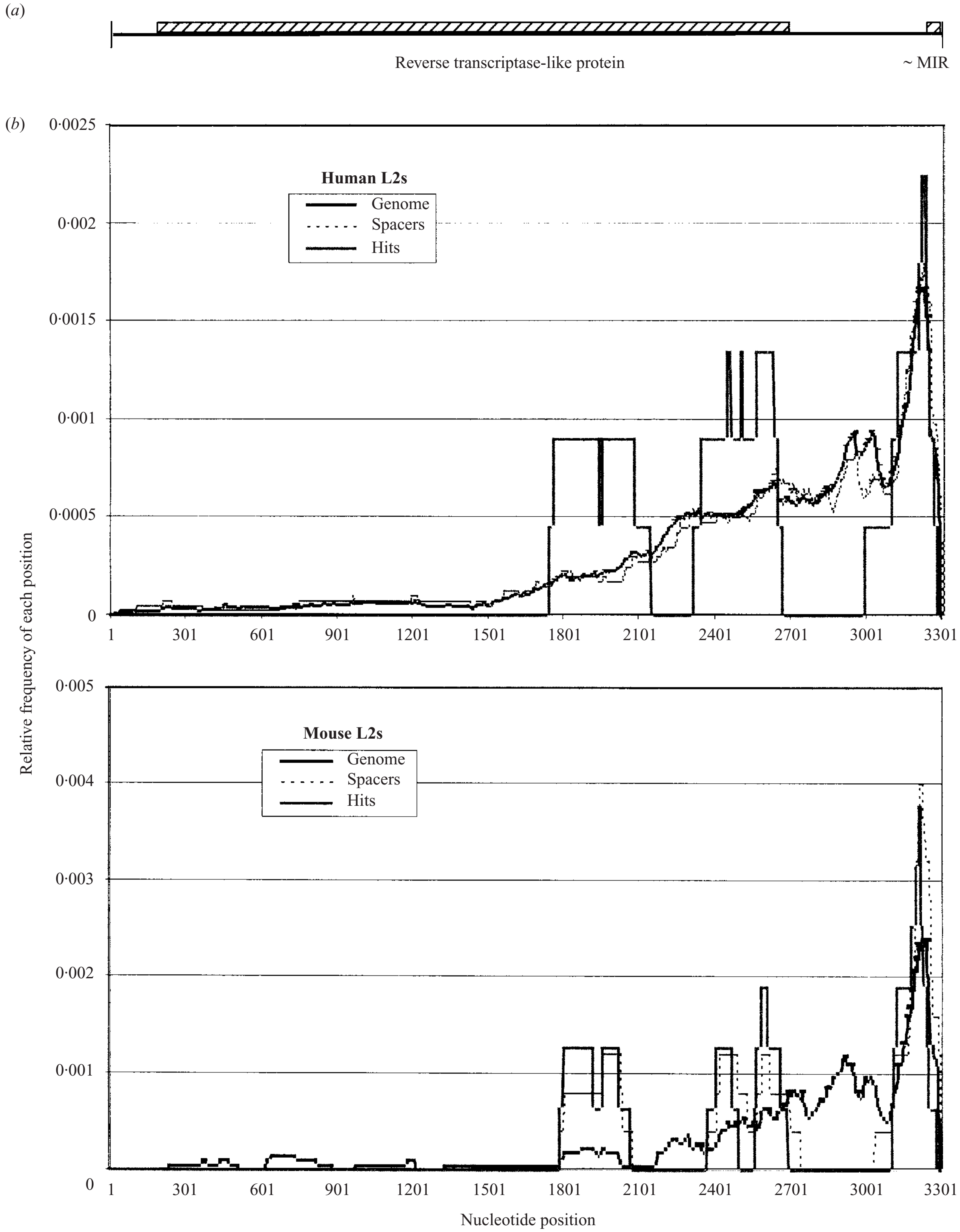

Fig. 4. The L2 subfamily. (*a*) Schematic of an L2 element (Smit, 1999, in Repbase). The 5′ end (188 nucleotides) is untranslated and has an internal RNA polymerase II promoter. Positions 189–2690 correspond to an open reading frame (ORF) with similarity to reverse transcriptase. The 3′ end contains a polyadenylation signal. (*b*) Relative frequency of each position of the element in the genomic sample (continuous line), intergenic regions (dotted line) and hits (grey lines) in mouse ($n=146$, $n=17$ and $n=13$, respectively) and in human ($n=840$, $n=182$ and $n=13$, respectively).

TE-derived fragments in that region in the two species, the fragmentation of one original element 'artificially' increases the number of matches. This seems to be, at least in part, the reason for the high number of matches observed for L1 elements. In fact, while the average length of an L1 insert is

https://doi.org/10.1017/S0016672303006268 Published online by Cambridge University Press

between 500 and 1000 nucleotides, the L1 fragments in human intergenic regions averaged only about 400 nucleotides in length, revealing some breakdown. In addition, when the length of the inter-scaffold regions considered is restricted to less than 1000 nucleotides (model IV), the number of L1 matches is no longer significantly higher than expected.

Finally, the correlation may be due to the presence of these TEs in the common ancestor of the host species, and their subsequent maintenance in both lineages. In this case TE matches correspond to TE pairs between species that are orthologous. This seems to explain the results obtained for some old TE families, such as MIRs, L2s and a few rare LINE, LTR and DNA TE families; the fact that, for these families, the number of alignments (which tend to reflect common ancestry) is also high provides strong support for this conclusion.

(iii) *Matches versus alignments*

x_M is a clear overestimator of x_A because there are several circumstances under which TE-derived sequences of the same family are present in the same inter-scaffold region in both species (and will, therefore, contribute to x_M) and yet can not be aligned. The following is a list of some of these circumstances:

(1) TEs are inserted in opposite orientations in the two species.

(2) Elements are located on opposite sides of a hit present within the inter-scaffold region.

(3) Elements of the same TE family are too different to align with each other. This can happen because the elements in the two species do not share a common ancestor for the entire length of their sequences (e.g. Alu subfamilies; Ullu *et al.*, 1982), and the fragments in each species happen to be non-homologous. Also, some elements are too old and, therefore, too divergent to be alignable (e.g. L1 subfamilies; Smit *et al.*, 1995).

(4) The TE-derived sequence in one species maps to a different part of the full-length element as the sequence found in the other species.

(5) The average number of elements per inter-scaffold region is high (>1) and differs between species. x_M in an inter-scaffold region is the product of the number of elements in human and in mouse for that region (see Section 2). However, as elements are arranged in a linear way, x_A can never be larger than the lowest of those two values. Therefore, x_M and x_A for an intergenic region may only be similar when the density of elements is low ($\leqslant 1$ per region).

(6) Only alignments for which the overlap of the two elements was >20 nucleotides were counted. No such restriction was imposed when calculating the number of matches.

Given all the above reasons, it is not surprising that, for most TE families, $x_A \ll x_M$. However, in the case of the MIR, L2 and a few other, less common, families, $x_A \cong x_M \gg X_M$. These represent the cases in which the correlation of TE location between species is due to an excessive number of orthologous pairs of elements.

(iv) *The time of the end of MIR and L2 expansions*

If a family of TEs stopped transposing before the human–mouse split, every copy in a mouse must have had a human orthologue and vice versa. Many copies of MIRs and L2 indeed come in such ortho-pairs (in some cases a member of a pair, overlooked by annotation, can be recognized from the alignment: see below). However, there are also many copies of human MIRs (at least 44%) and L2s (at least 50%) that do not have orthologues in mouse, and a few in mouse (16% of MIRs and 18% of L2s) with no discernible orthologue in human. It has been argued that most mouse elements may have become unrecognizable due to accumulation of neutral mutations (International Human Genome Sequencing Consortium, 2001), which, incidentally, would imply mutation saturation within the murine lineage alone. On the other hand, the loss of human orthologues for some of the extant mouse elements is not likely to be due to mutation, as the evolution rate along the human lineage has been twice as low as in the murine lineage (Li, 1997) and hence probably insufficient for mutation saturation between the human–mouse last common ancestor and modern humans.

Three other processes can explain the absence of human orthologues for some of the murine elements: (1) chromosomal rearrangements, (2) transpositions of murine (and, possibly, human) TEs after the human–mouse split and (3) deletions in the human chromosomes. The widespread synteny between the intergenic regions analysed precludes the first hypothesis. Circumstantial evidence against hypothesis (2) is the correlation between the location of elements between species. Had recent transposition been common, that correlation should have disappeared. This is corroborated by other studies that argue that transposition ceased before the split of the two lineages (Jurka *et al.*, 1995; International Human Genome Sequencing Consortium, 2001). Finally, analyses of the seven cases of murine intergenic MIR elements that are not part of hits (and, therefore, have no human orthologue) reveal that, in five of those cases, the length of the human inter-scaffold region in which they are located is only 3–30% of the size of the corresponding inter-scaffold region in mouse, favouring the deletion hypothesis. Thus, our analysis supports the view that transposition of MIR and L2 ceased before

https://doi.org/10.1017/S0016672303006268 Published online by Cambridge University Press

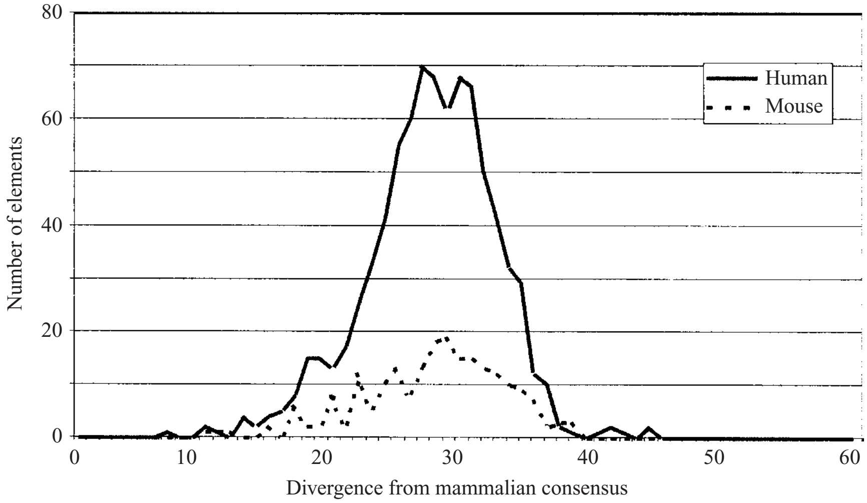

Fig. 5. Percentage divergence of mouse and human MIRs from the MIR mammalian consensus sequence. The distributions correspond to 820 and 221 MIR-derived fragments in the genomic sample of human and mouse, respectively.

the human–mouse split, although we cannot rule out some residual movements after it.

(v) *Recruitment of MIR and L2 elements*

At least two pieces of evidence from the MIR data are clearly inconsistent with unconstrained evolution of a majority of extant MIRs since the human–mouse split. First, the minimal divergence of individual elements from the mammalian consensus is approximately 20% in both mouse and human (a few less-diverged copies must be under selection; Fig. 5). This is too low to be consistent with the lack of selective constraint. In the murine lineage, introns are not alignable between *Mus–Oryctolagus* (rabbit), corresponding to a divergence well over 25% from their common ancestor to each of the terminal taxa (results not shown), while alignments of introns between *Mus* and *Cavia* (guinea pig) show approximately 35% divergence, or about 18% from their common ancestor. The value obtained for MIR divergence from the mammalian consensus falls in between these two. However, lagomorphs diverged from rodents after primates (Adkins *et al.*, 2001; Murphy *et al.*, 2001). As for humans, 20% divergence corresponds to the split of the human lineage from its common ancestor with the sister taxon of primates, the tree shrews Scandentia (Kupfermann *et al.*, 1999; Murphy *et al.*, 2001).

Second, the average divergence of MIR elements from its consensus is the same for both species (27·3% and 27·6% for human and mouse, respectively). The same is true for L2 elements, with an average divergence from the family consensus of 28·9% and 29·2% for human and mouse, respectively. However, selectively neutral TEs in the mouse genome should, on average, be at least 2 times more divergent from the mammalian consensus than are their human orthologues (Li, 1997).

Selection on TE sequences can act at two levels. Selection at the level of the host preserves individual TE-derived sequences. Alternatively, selection at the DNA level would favour only one (or a few) transposition-competent 'master copies' in each species, while all other copies remain free of selective constraint. As no transposition has occurred since the mouse–human split, and given the relatively low divergence of mouse elements from the mammalian consensus, selection on individual murine elements, and therefore recruitment of individual copies, is a more likely scenario.

MIR and L2 elements are proportionately more abundant in hits than in interhits in both species (Tables 1, 2); in addition, for these families, both x_M and x_A are significantly higher than X_M, and >75% of all elements in mouse intergenic regions have human orthologues. The MER73 group and a few DNA elements seem to share all these characteristics, suggesting domestication of members of these subfamilies. We argue that most (and, possibly, all) MIR and L2 elements for which an orthologue exists are indeed evolving under selection, and that the same is probably true for the MER73 group (LTR), MER53 and MER1-type (DNA) families, as well as CR1 (LINE) and ERVL (LTR) elements.

Additional evidence of selection acting on MIRs comes from the fact that the segment corresponding to the '70 bp MIR' is present in hits more often than expected given its background frequency in the genome. This indicates that chance is not the main factor determining which ancestral sequences are represented in hits, and that the '70 bp' segment is under the strongest selective constraint. These results provide strong support for previous studies that suggest

https://doi.org/10.1017/S0016672303006268 Published online by Cambridge University Press

that this MIR segment may play a functional role in the genome (Donehower *et al.*, 1988; Smit & Riggs, 1995), possibly by providing an alternative splicing sequence (Donehower *et al.*, 1988; Murnane & Morales, 1995) or as an additional promoter sequence (Thomas *et al.*, 1999). In addition, the histogram analysis of L2 elements suggests that a segment derived from the 3′ end of a reverse transcriptase-like ORF might also play a role in the mammalian genome that is position-dependent.

If a significant number of MIR- and L2-derived fragments have indeed evolved under selective constraints, care must be taken while calibrating the age of TE repeats. TE ages inferred under the assumption of lack of selective constraints (e.g. International Human Genome Sequencing Consortium, 2001) correspond to a lower boundary for the age which, depending on the strength of the selective constraints imposed on these sequences, may be considerably underestimated.

(vi) *Number of TE-derived recruited sequences and the detection of cryptic TE-derived sequences from non-reciprocal alignments*

TE-derived sequences present in hits in one species can be aligned either to sequences identified as TEs in the other species (reciprocal TE alignments), or to sequences that are not annotated as TEs (non-reciprocal TE alignments). TE families differ in their patterns of non-reciprocal alignments.

The annotated MIR or L2 element in non-reciprocal alignments is almost exclusively human. While most mouse MIR and L2 elements in hits correspond to reciprocal alignments (85% for MIRs, or 30 of 35, and 92% for L2s, or 13 of 14), the fraction is much smaller for aligned human elements (28% for MIRs and 14% for L2s). The lack of symmetry in this pattern is easily explained by the higher evolution rate of the murine lineage relative to the primate lineage (Li, 1997). The pattern observed for Alu elements is markedly different from that seen for MIR and L2 elements. Reciprocal Alu alignments are almost as common in human (~50% of the elements in hits) as they are in mouse (~70%).

The difference is probably due to the different history of these families in the host genome. As most (if not all) MIR and L2 elements were present in the common ancestor of the two species, non-reciprocal alignments are likely to result from old MIR and L2 elements, of which one of the two descendent copies (the murine one) has become unrecognizable. In contrast, as the Alu family diversified mostly after the two host lineages split, extant reciprocal alignments involving Alus are likely to correspond to elements that happened to insert by chance in similar inter-scaffold regions in both species. Because of the strong element of chance as opposed to common ancestry the pattern should be (close to) symmetrical. In addition, this explains why the proportion of Alus in intergenic regions that are present in hits is much more similar between mouse (34%) and human (26%), and much smaller overall compared with what was found for MIRs (83% vs 56%) and L2s (82% vs 50%). The pattern observed in other common families (e.g. L1, MaLR and the DNA family of MER1-type) is somewhat intermediate to that observed for MIRs and L2s versus Alus. This might result from the fact that these families are composed of multiple subfamilies, some of which pre-date the split of the murine and primate lineages while others post-date it.

Assuming that murine sequences in hits that are aligned to human MIR-derived sequences are also MIR-derived, then the total number of MIR-derived fragments is as high as 113, i.e. the number of reciprocal alignments (30) plus the MIR non-reciprocal alignments in human ($108-30=78$) and in mouse ($35-30=5$). This number drops to 104 if we exclude the cases where TEs overlap hits by fewer than 15 nucleotides in either species. Using the same rationale, there could be up to 92 domesticated L2-derived sequences (87 if only TEs that overlap hits by more than 15 nucleotides are counted). Thus, between these two families alone, the number of TE-derived fragments can be as high as 200 in our sample of 100 alignments of complete intergenic regions.

It should be stressed that these results do not necessarily imply that the primary sequence of the original TE is conserved, only that the conserved sequence is of TE origin. In addition, alignments of the murine and human genomes can be extremely useful in the annotation of these old TE mutation derivatives in the murine genome.

Finally, conservation in intergenic regions of mouse and human which leads to a pattern of hits and interhits could result from a short-scale alternation of hot and cold spots for mutation (Chiaromonte *et al.*, 2001). However, there is no evidence of variation in mutation rate at that scale (Silva & Kondrashov, 2002). This makes mutational cold spots a very unlikely explanation of the pattern observed.

(vii) *Conclusion*

The lack of evidence for the recruitment of sequences derived from most TE families does not show that they are non-existent. Our study can only detect the domestication of sequences belonging to TE families that pre-date the mouse–human split, and so those derived from young TE families will go unnoticed. For example, B2 elements, which are characteristic of the murine lineage, have been found to contain an active RNA polymerase II site which provides transcription promoters for host genes (Ferrigno *et al.*,

https://doi.org/10.1017/S0016672303006268 Published online by Cambridge University Press

2001). In addition, recruited TE-derived sequences may be present in coding regions (Makalowski, 1995; Brosius, 1999; Nekrutenko & Li, 2001), and we focused on intergenic regions only. Many other examples of recruitment of TE-derived sequences from young TE families can be found, for example in Brosius (1999). Finally, the function played by TEs may not be position-dependent, in which case the method used in this paper would not have detected domestication. Examples include a possible role of L1 elements in X chromosome inactivation (Bailey *et al*., 2000), and that of Alu elements in promoting protein translation during periods of stress (Chu *et al*., 1998; Schmid, 1998; but see Brookfield, 2001).

Among the ancestral TE families MIR and L2, and possibly CR1, MER53 (DNA) and MER73 (LTR), we were able to identify the potential recruitment of over 200 TE-derived sequences, i.e. DNA fragments of TE origin that may have been co-opted by the host genome for a function that increases host fitness. Since we analysed 100 intergenic regions we find, on average, at least two TE-derived such events per host gene. These results support the assertion that 'many of the coding, structural or regulatory sequences interspersed within the genome may have been derived from ancient transposable elements' (Murnane & Morales, 1995). We are just starting to compile evidence that reveals the extent to which this statement might be true.

We thank Suzanne Szak, King Jordan and Igor Rogozin for helpful discussions and suggestions on the manuscript.

References

Adkins, R. M., Gelke, E. L., Rowe, D. & Honeycutt, R. L. (2001). Molecular phylogeny and divergence time estimates for major rodent groups: evidence from multiple genes. *Molecular Biology and Evolution* **18**, 777–791.

Bailey, J. A., Carrel, L., Chakravarti, A. & Eichler, E. E. (2000). Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis. *Proceedings of the National Academy of Sciences of the USA* **97**, 6634–6639.

Bénit, L., Lalleman, J.-B., Casella, J.-F., Philippe, H. & Heidmann, T. (1999). ERV-L elements: a family of endogenous retrovirus-like elements active throughout the evolution of mammals. *Journal of Virology* **73**, 3301–3308.

Britten, R. J. (1997). Mobile elements inserted in the distant past have taken on important functions. *Gene* **205**, 177–182.

Brosius, J. (1999). Genomes were forged by massive bombardments with retroelements and retrosequences. *Genetica* **107**, 209–238.

Brookfield, J. F. Y. (2001). Selection on Alu sequences? *Current Biology* **11**, R900–R901.

Bustamante, C. D., Nielsen, R. & Hartl, D. L. (2002). A maximum likelihood method for analyzing pseudogene evolution: implications for silent site evolution in humans and rodents. *Molecular Biology and Evolution* **19**, 110–117.

Castresana, J. (2002). Genes on human chromosome 19 show extreme divergence from the mouse orthologs and a high GC content. *Nucleic Acids Research* **30**, 1751–1756.

Chiaromonte, F., Yang, S., Elnitski, L., Yap, V. B., Miller, W. & Hardison, R. C. (2001). Association between divergence and interspersed repeats in mammalian noncoding genomic DNA. *Proceedings of the National Academy of Sciences of the USA* **98**, 14503–14508.

Chu, W. M., Ballard, R., Carpick, B. W., Williams, B. R. & Schmid, C. W. (1998). Potential *Alu* function: regulation of the activity of double-stranded RNA-activated kinase PKR. *Molecular and Cellular Biology* **18**, 58–68.

Donehower, L. A., Slagle, B. L., Wilde, M., Darlington, G. & Butel, J. S. (1988). Identification of a conserved sequence in the non-coding regions of many human genes. *Nucleic Acids Research* **17**, 699–710.

Doolittle, W. F. & Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**, 601–603.

Efron, B. & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Ferrigno, O., Virolle, T., Djabari, Z., Ortonne, J.-P., White, R. J. & Aberdam, D. (2001). Transposable *B*2 SINE elements can provide mobile RNA polymerase II promoters. *Nature Genetics* **28**, 77–81.

Finnegan, D. J. (1992). Transposable elements. *Current Opinion in Genetics and Development* **2**, 861–867.

Gu, Z., Wang, H., Nekrutenko, A. & Li, W.-H. (2000). Densities, length proportions, and other distributional features of repetitive sequences in the human genome estimated from 430 megabases of genomic sequence. *Gene* **259**, 81–88.

Hardison, R. C. (2000). Conserved noncoding sequences are reliable guides to regulatory elements. *Trends in Genetics* **16**, 369–372.

Hardison, R. C., Oeltjen, J. & Miller, W. (1997). Long human–mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Research* **7**, 959–966.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.

Jareborg, N., Birney, E. & Durbin, R. (1999). Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Research* **9**, 815–824.

Jurka, J. (1998). Repeats in genomic DNA: mining and meaning. *Current Opinion in Structural Biology* **8**, 333–337.

Jurka, J. (2000). Repbase Update: a database and an electronic journal of repetitive elements. *Trends in Genetics* **9**, 418–420.

Jurka, J., Zietkiewicz, E. & Labuda, D. (1995). Ubiquitous mammalian-wide interspersed repeats (MIRs) are molecular fossils from the Mesozoic era. *Nucleic Acids Research* **23**, 170–175.

Kidwell, M. G. & Lisch, D. (1997). Transposable elements as sources of variation in animals and plants. *Proceedings of the National Academy of Sciences of the USA* **94**, 7704–7711.

Kupfermann, H., Satta, Y., Takahata, N., Tichy, H. & Klein, J. (1999). Evolution of *Mhc-DRB* introns: implications for the origin of primates. *Journal of Molecular Evolution* **48**, 663–674.

Li, W.-H. (1997). *Molecular Evolution*. Sunderland, MA: Sinauer Associates.

Makalowski, W. (1995). SINEs as a genomic scrap yard: an essay in genomic evolution. In *The Impact of Short*

https://doi.org/10.1017/S0016672303006268 Published online by Cambridge University Press

Interspersed Elements (*SINEs*) *on the Hpst Genome* (ed. R. J. Maraia), pp. 81–104. Austin, TX: R. G. Landes.

Makalowski, W. (2000). Genomic scrap yard: how genomes utilize all that junk. *Gene* **259**, 61–67.

Makalowski, W. & Boguski, M. (1998). Evolutionary parameters of the transcribed mammalian genome: an analysis of 2820 orthologous rodent and human sequences. *Proceedings of the National Academy of Sciences of the USA* **95**, 9407–9412.

Malik, H. S., Burke, W. D. & Eickbush, T. H. (1999). The age and evolution of non-LTR retrotransposable elements. *Molecular Biology and Evolution* **16**, 793–805.

Miller, J. M., McDonald, J. F. & Pinsker, W. (1997). Molecular domestication of mobile elements. *Genetica* **100**, 261–270.

Mouse Genome Sequencing Consortium (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562.

Murnane, J. P. & Morales, J. F. (1995). Use of a mammalian interspersed repetitive (MIR) element in the coding and processing sequences of mammalian genes. *Nucleic Acids Research* **23**, 2837–2839.

Murphy, W. J., Eizirik, E., Johnson, W. E., Zhang, Y. P., Ryder, O. A. & O'Brien, S. J. (2001). Molecular phylogenetics and the origins of placental mammals. *Nature* **409**, 614–618.

Nekrutenko, A. & Li, W.-H. (2001). Transposable elements are found in a large number of human protein-coding genes. *Trends in Genetics* **17**, 619–621.

Orgel, L. E. & Crick, F. H. C. (1980). Selfish DNA: the ultimate parasite. *Nature* **284**, 604–607.

Rogozin, I. B., Mayorov, V. I., Lavrentueva, M. V., Milanesi, L. & Adkison, L. R. (2000). Prediction and phylogenetic analysis of mammalian short interspersed elements (SINEs). *Briefings in Bioinformatics* **1**, 260–274.

Schmid, C. W. (1998). Does SINE evolution preclude *Alu* function? *Nucleic Acids Research* **26**, 4541–4550.

Shabalina, S. A., Ogurtsov, A. Y., Kondrashov, V. A. & Kondrashov, A. S. (2001). Selective constraint in intergenic regions of human and mouse genomes. *Trends in Genetics* **17**, 373–376.

Silva, J. C. & Kondrashov, A. S. (2002). Patterns in spontaneous mutation revealed by human–baboon sequence comparison. *Trends in Genetics* **18**, 544–547.

Smit, A. F. (1993). Identification of a new, abundant superfamily of mammalian LTR-transposons. *Nucleic Acids Research* **21**, 1863–1872.

Smit, A. F. (1999). Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Current Opinion in Genetics and Development* **9**, 657–663.

Smit, A. F. & Riggs, A. D. (1995). MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. *Nucleic Acids Research* **23**, 98–102.

Smit, A. F. & Riggs, A. D. (1996). *Tiggers* and DNA transposon fossils in the human genome. *Proceedings of the National Academy of Sciences of the USA* **93**, 1443–1448.

Smit, A. F., Toth, G., Riggs, A. D. & Jurka, J. (1995). Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *Journal of Molecular Biology* **246**, 401–417.

Thomas, C. P., Auerbach, S. D., Zhang, C. & Stokes, J. B. (1999). The structure of the rat amiloride-sensitive epithelial sodium channel gamma subunit gene and functional analysis of its promoter. *Gene* **228**, 111–122.

Ullu, E., Murphy, S. & Melli, M. (1982). Human 7S RNA consists of a 140 nucleotide middle repetitive sequence inserted in an *Alu* sequence. *Cell* **29**, 195–202.

Wasserman, W. W., Palumbo, M., Thompson, W., Fickett, J. W. & Lawrence, C. E. (2000). Human–mouse genome comparisons to locate regulatory sites. *Nature Genetics* **26**, 225–228.

https://doi.org/10.1017/S0016672303006268 Published online by Cambridge University Press