Do transposable elements really contribute to proteomes?

Valer Gotea and Wojciech Makałowski

Institute of Molecular Evolutionary Genetics and Department of Biology, Center for Comparative Genomics and Bioinformatics, The Pennsylvania State University, University Park, PA 16802, USA

Recent studies indicate that the initial classification of transposable elements (TEs) as 'useless', 'selfish' or 'junk' pieces of DNA is not an accurate one. TEs seem to have complex regulatory functions and contribute to the coding regions of many genes. Because this contribution had been documented only at transcript level, we searched for evidence that would also support the translation of TE cassettes. Our findings suggest that the proportion of proteins with TE-encoded fragments (~0.1%), although probably underestimated, is much less than what the data at transcript level suggest (~4%). In all cases, the TE cassettes are derived from old TEs, consistent with the idea that incorporation (exaptation) of TE fragments into functional proteins requires long evolutionary periods. We therefore argue that functional proteins are unlikely to contain TE cassettes derived from young TEs, the role of which is probably limited to regulatory functions.

Introduction

It is widely accepted that transposable elements (TEs; see Glossary) have had a major impact on the evolution of mammalian genomes. A well-documented example is that of the human genome, almost half of its sequence being derived from TEs [1]. TEs were initially regarded as 'junk' [2], 'selfish', and 'parasite' pieces of DNA [3–5]. Gradually, scientists realized that TEs should be regarded as 'seeds of evolution' [6] and 'genomic treasures' [7] because they seem to enhance the organisms' evolvability in many ways. TEs are active genomic components that can promote recombination [8,9] and provide ready-to-use motifs, such as transcriptional regulatory elements, polyadenylation and splicing signals, and even protein coding sequences [10–14].

The contribution of TEs to coding regions is of particular interest, because they can directly influence the phenotype by altering protein sequences. This aspect was documented, however, only at the transcript level, and the presence of TE-encoded fragments was not confirmed at the protein level [15]. Because of the important evolutionary implications, we attempted to clarify the issue of TE contribution to metazoan proteomes using computational methods and publicly available data by searching for TE cassettes in functionally well-characterized proteins. We found evidence indicating that functional proteins can indeed contain TE cassettes, but only those derived from old TEs. Those derived from young TEs, such as Alu short interspersed elements (SINEs) and L1 long interspersed elements (LINE1s), seem to disrupt the functionality of the proteins into which they are inserted.

TE fragments were found in coding regions of many transcripts but not in functional proteins

More than a decade ago, a few studies reported that some mRNAs contain TE cassettes in their coding regions [16–18] that sometimes resulted in disease phenotypes such as the gyrate atrophy of the choroid and retina [19]. These observations led to the hypothesis that, in other

Glossary

Transposable elements (TEs): all DNA segments that have the ability to move or multiply within genomes generating self-copies interspersed with non-repetitive DNA. The term is often used for referring to copies of such elements that lost the ability to move or multiply once integrated at a new genomic location because of either mutation or fragmentation. For those segments, 'TE-derived sequences' or 'transposed elements' would better describe the current status of the sequence. The more general term, 'interspersed repeats', could be used instead for referring to both active and non-active TE copies. TEs can be classified in two main classes [53]: DNA transposons and elements that move through an mRNA intermediate, which can be further divided into SINEs, LINEs and long terminal repeat (LTR) retrotransposons. Because multiple copies of the same DNA segment (e.g. >1 000 000 Alu copies in the human genome) can be found scattered throughout genomes, TEs are one of the most important categories of repetitive DNA. In humans, almost half (~45%) of the genome consists of TE-derived sequences and still active TEs [1].

TE cassette: a TE fragment inserted into an mRNA sequence [19,20]. In many cases, a TE cassette is generated after the activation of cryptic splice sites located in an intron-residing TE sequence. TE cassettes can be also generated by *de novo* insertions in exons.

Exaptation: a term introduced by Gould and Vrba [54] to describe the co-optation of different characters to new roles regardless of their original function. Those characters might have been shaped by natural selection for specific functions or might have had no function. The concept was applied to genomic level by Brosius and Gould [55] and fits perfectly in our case, because TE-derived sequences were originally part of mobile TEs. They were then fortuitously co-opted in the ORFs of different genes and now encode short stretches of amino acids that are completely unrelated to their original function within TEs. TE cassette could thus be called 'xaptonuons' according to the 'genomenclature' proposed by Brosius and Gould [55].

Nonaptation: according to Gould and Vrba's vision, this describes a character 'whose origin cannot be ascribed to the direct action of natural selection' [54]. Most of the TE fragments that were successfully exapted into protein coding regions can be thought of as nonaptations because after losing the ability to move, TEs were presumably subjected to neutral evolution in intronic or intergenic regions. If TE fragments are subjected to functional constraints even after loosing the ability to move, they should be called adaptations. Elements that can generate alternative messages capable of regulating the expression of cognate variants, such as Alu elements, could be considered as undergoing adaptive evolution, but it would need to be determined whether those sequences are under functional constraints.

Corresponding author: Makałowski, W. (wojtek@psu.edu).

Available online 29 March 2006

www.sciencedirect.com 0168-9525/$ - see front matter © 2006 Elsevier Ltd. All rights reserved. doi:10.1016/j.tig.2006.03.006

cases, TE exaptation could have neutral effects or even enhance fitness and, therefore, might increase protein variability with positive evolutionary consequences [20]. Since then, several studies discovered TE cassettes of many TE types in the coding region of many genes [21–23]. Despite a few reports of potentially functional proteins containing TE-encoded fragments [24–26], there is no strong evidence that supports the existence of such proteins *in vivo*. The presence of TE cassettes in transcripts does not guarantee their translation, because eukaryotic cells contain several quality control mechanisms that can initiate the degradation of the transcript and even of the protein product immediately after translation [27–29]. Moreover, even if translation occurs, the product might be non-functional and even 'mildly deleterious to the cell' [30]. Consequently, overstatements such as 'many translated repetitive elements are found in proteins' [31] and 'pieces of TEs found in exons are translated in functional proteins' [21] are misleading in the absence of evidence at the level of functional proteins. Evolutionary implications of TE exaptation would be more profound if TE cassettes were present not only at the transcript level but also at protein level, given that 'proteins, rather than genes or mRNA, represent the key players in the cell' [32] by determining the cellular phenotype, and thus directly affecting fitness. Pavlicek *et al.* have argued that TE contribution to proteins can be reliably studied only with directly sequenced proteins or with proteins that have their three dimensional structure determined [15]. Therefore, the contribution of TEs to the proteome needs to be confirmed [13,23].

Identifying proteins with TE-encoded fragments

We searched for TE cassettes only in functionally well-characterized proteins, to eliminate the uncertainty of translation associated with most transcripts (see the online supplementary material for more details). Among the 3764 Protein Databank (PDB; http://www.rcsb.org/pdb/) entries with non-redundant protein chains, we found only 24 proteins with fragments encoded by putative TE cassettes (Tables 1, S1 in the online supplementary material). No additional examples were identified in the Swiss-Prot collection of directly sequenced proteins (1765 non-redundant human protein sequences; http://www.expasy.ch/sprot). A common feature of all TE cassettes identified in the 24 proteins is their low RepeatMasker (RM) scores (http://www.repeatmasker.org/), which are all close to the empirically set thresholds for false positive matches (see online supplementary material for details on the RM scoring system). Therefore, we wanted to learn whether the exaptation of each TE cassette could be explained in the context of the evolutionary history of the host gene. Not surprisingly, phylogenies of 21 proteins (Table S1 in the online supplementary material) do not support the presence of a TE cassette as reported by RM. In 13 examples, where the putative TE cassette is located within one exon, the encoded fragment is conserved in invertebrate orthologs, which is inconsistent with the known activity times of the reported TEs. In eight examples, the putative TE cassette spans across multiple exons, which could be reasonably explained only by intron gain during vertebrate evolution. This would be a reasonable scenario [33], but one that we could not confirm for any of the eight cassettes, because the fragment and gene structure were either conserved in invertebrate orhtologs or the fragment was not well conserved in vertebrates. In either case, the data suggest that the putative TE cassettes are probably random matches to TE consensus sequences and not real TE cassettes.

Furthermore, we also found three proteins (Table 1), calpain 1 (CAPN1), granzyme A (GZMA) and protein tyrosine phosphatase, non-receptor type I (PTPN1), with phylogenies that provide support for the presence of the TE cassettes identified by RM (see online supplementary material for details on CAPN1 and GZMA).

PTPN1 (also known as PTP1B) is a 435-amino-acid protein that belongs to the large family of protein tyrosine phosphatases (PTPs), which catalyze protein dephosphorylation. Its sequence was initially determined by direct sequencing [34], its three dimensional structure was first determined by Barford *et al.* [35] and its functionality has been detailed by several consequent studies (links to all PDB structures can be accessed via the Swiss-Prot record of PTPN1, accession number P18031). RepeatMasker finds remnants of an L3 LINE in the coding region of the corresponding mRNA (NCBI gi:17390366) between coordinates 499 and 599 (Figure 1a). Although the length, divergence from consensus and RM score are similar to those of the TE cassettes that can be considered false positives, several arguments support the validity of the L3 cassette. The first argument is the origin of the L3 cassette in the second open reading frame of the L3 element (ORF2p). The L3 non-LTR retrotransposon is among the most ancient TEs reconstructed *in silico* and is characterized by the presence of two ORFs [36]. The second ORF is estimated to be ~902 residues long (Repbase v10.12, http://www.girinst.org) and contains a well-conserved reverse-transcriptase (RT) domain between coordinates 457 and 712, characteristic of retrotransposons and retroviruses. The L3 fragment found in the PTPN1

Table 1. Human proteins identified with TE-encoded fragments

Gene[a]	Corresponding PDB structure (PDB ID:chain)[b]	mRNA gi	CDS length (nt)	GC content of CDS (%)	TE type	Length of TE cassette (nt)[c]	Divergence from consensus sequence (%)[c]	RM score[c]	*P*-value[d]
CAPN1	2ARY:A	49900978	2145	59.53	MIRm	34	17.6	182	0.0003
GZMA	1OP8:A	184022	789	43.73	L3	85	34.1	183	0
PTPN1	1G7F:A	17390366	1308	50.69	L3	101	27.7	198	0.0013

[a] The gene name is given as the official NCBI gene symbol. [b] For every gene, only one structure and mRNA sequence is provided, even if many of these structures and sequences have been reported. [c] The length of the TE cassette, divergence from consensus sequence and scores are given as reported by RepeatMasker. [d] *P*-values represent the probability of a random match in the coding sequence (CDS) of the gene to the TE identified by RM as computed with the sequence randomization test (see online supplementary material).

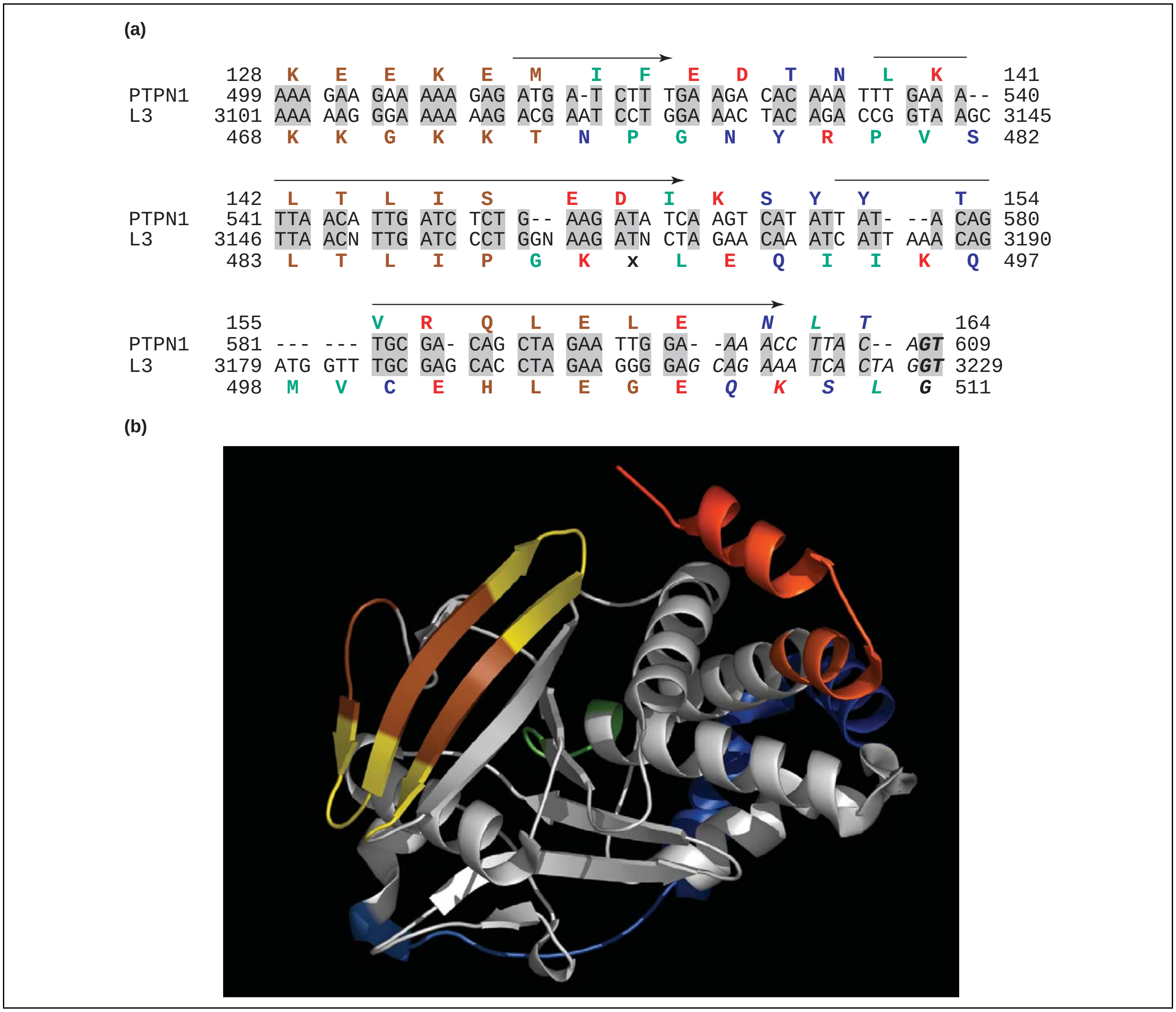
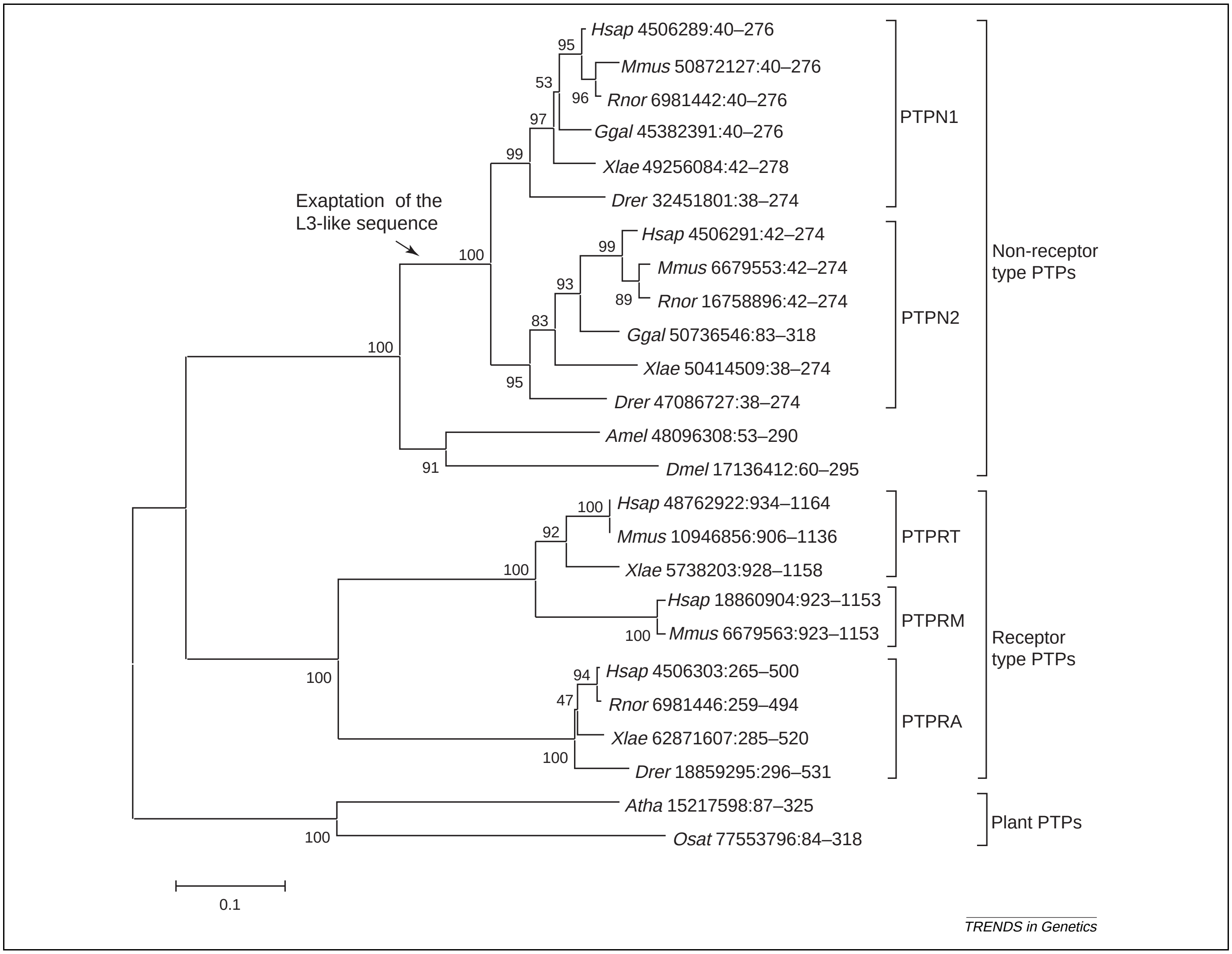

(a)
128 K E E K E M I F E D T N L K 141
PTPN1 499 AAA GAA GAA AAA GAG ATG A-T CTT TGA AGA CAC AAA TTT GAA A-- 540
L3 3101 AAA AAG GGA AAA AAG ACG AAT CCT GGA AAC TAC AGA CCG GTA AGC 3145
468 K K G K K T N P G N Y R P V S 482
142 L T L I S E D I K S Y Y T 154
PTPN1 541 TTA ACA TTG ATC TCT G-- AAG ATA TCA AGT CAT ATT AT- --A CAG 580
L3 3146 TTA ACN TTG ATC CCT GGN AAG ATN CTA GAA CAA ATC ATT AAA CAG 3190
483 L T L I P G K x L E Q I I K Q 497
155 V R Q L E L E N L T 164
PTPN1 581 --- --- TGC GA- CAG CTA GAA TTG GA- -AA ACC TTA C-- AGT 609
L3 3179 ATG GTT TGC GAG CAC CTA GAA GGG GAG CAG AAA TCA CTA GGT 3229
498 M V C E H L E G E Q K S L G 511
(b)

Figure 1. The contribution of L3 retrotransposon to PTPN1. (**a**) Nucleotide alignment of human PTPN1 mRNA (gi:17390366) with L3 consensus sequence, as determined by RepeatMasker. The alignment was extended manually (italics) to the end of the fifth human exon and into the donor splice of intron 5 (bold). This demonstrates that the L3 element carries the cryptic donor splice site. Conserved nucleotides are shaded. Amino acids are shown above and below the nucleotide in second codon position. β-strands occurring in human PTPN1 are represented by arrows above the alignment. Residues for which the reading frame was preserved between L3 ORF2p and PTPN1 are shown in brown. Residues from regions of non-preserved reading frame are shown in green (hydrophobic), red (charged) and blue (polar). (**b**) Three dimensional structure of human PTPN1 (PDB accession number 1G7F): N-terminus is shown in blue, the C-terminus in red, the PTP domain in white, the base of the active site cleft in green (Cys^{215} is the essential catalytic residue) and the L3-encoded fragment is shown in brown (preserved reading frame) and yellow (non-preserved reading frame).

mRNA corresponds to part of this RT domain (residues 468–506, Figure 1a). In fact, the L3 cassette donates almost unchanged the core residues of two β-strands that are part of a four-strand anti-parallel β-sheet (Figure 1a,b). It is difficult to imagine that such a complex structure could have been generated by a sequence that did not have previous coding capacity. We see a similar situation in GZMA (supplementary online material), consistent with the exaptation hypothesis, which implies the reuse of characters (i.e. protein coding sequences) for different functions.

TE exaptation: when did it happen?

A second argument supporting the validity of the L3 cassette is provided by the origin of PTPN1. It is known that PTP diversification occurred by a series of duplication events during early vertebrate evolution [37–39]. This can explain why PTPN1 is located ∼7.3-Mb apart from PTPRT on chromosome 20q, similar to their closest homologs, PTPN2 and PTPRM, respectively (Ref. [39]; Figure 2), which are located ∼4.4-Mb-apart on chromosome 18p. The most likely scenario is that an intrachromosomal duplication was followed by the exaptation of the L3-like sequence, followed by a larger interchromosomal duplication. This can explain why the L3 cassette is strongly conserved between PTPN1 and PTPN2, but seems strikingly non-conserved in the invertebrate non-receptor type PTPs (Figure 3). The average identity between vertebrate and invertebrate sequences for this segment (23.56 ± 10.67%)



TRENDS in Genetics

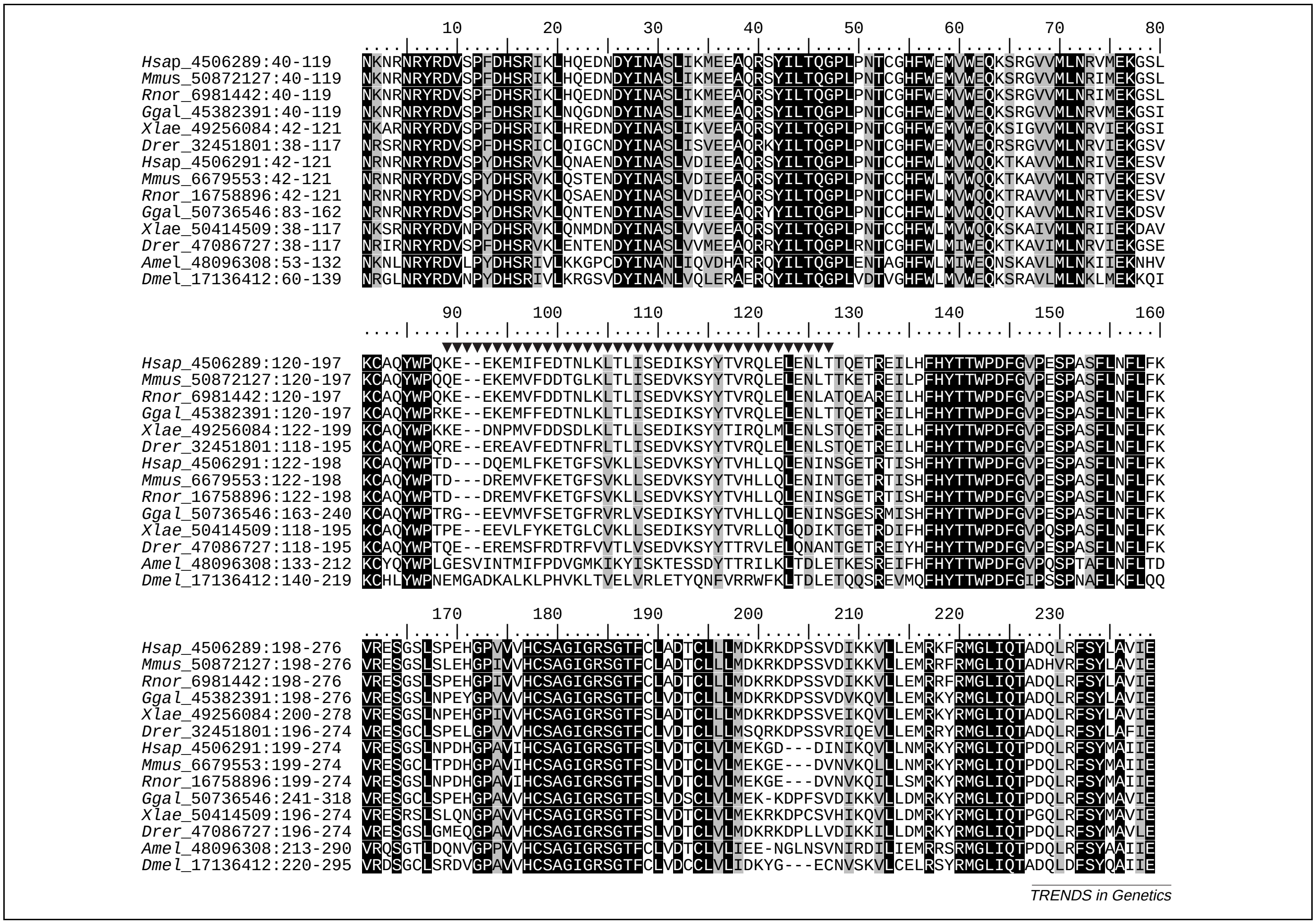
Figure 2. Phylogenetic history of human PTPN1. For constructing the tree, we used the neighbor-joining method [57], 188 sites after complete gap deletion, Poisson corrected distance and 1000 bootstrap replicates. The fragment corresponding to the L3 (coordinates 89–127 in Figure 3) was excluded for this step. Two plant PTPs (*Atha – Arabidopsis thaliana, Osat – Oryza sativa*) were added to root the tree of animal PTPs. The same tree topology was obtained by using maximum parsimony and maximum likelihood methods of phylogenetic reconstruction.

is significantly smaller (P-value $\ll 0.001$) than that of the rest of the PTP catalytic domain ($65.71 \pm 1.85\%$). In addition, the gene structure of invertebrate non-receptor type PTPs seems to have undergone major rearrangements (Figure 4a).

Although plausible, the timing of the events according to this scenario implies that L3 retrotransposons were active much earlier than the current estimate, which places L3 activity before the mammalian radiation [36]. Perhaps the L3 retrotransposon is much older than the current estimate of >200 million years because of the bias towards more-conserved copies used for the reconstruction of the consensus sequence [36]. An alternative explanation could be that the fragment identified as L3 originates in an older unknown RT-carrying retrotransposon. Support for this hypothesis is provided by the strong purifying selection observed in the vertebrate lineage (Table 2), which maintained the similarity to the original RT domain so that it now resembles the oldest known RT-carrying retrotransposon: the L3 LINE.

TE exaptation: how did it happen?

According to Ohno [40], gene duplications create the raw material for evolutionary ‘innovations’. He argues that newly duplicated genes are free of functional constraints and can undergo significant changes until they acquire new specific functions. Provided that duplications are the documented source for PTP diversification as discussed earlier, it is easy to imagine that the future PTPN1 could have easily acquired a TE fragment after the activation of a cryptic splice site in a manner similar to how genes currently acquire Alu fragments [20,41]. The position of the L3-like fragment at the end of exon 5 (Figure 4) supports this hypothesis (because non-active TE sequences are expected to mutate beyond recognition [42], it is not surprising that we cannot extend the alignment into the 3′ intronic region). We also observe that the ratio of dN to dS between the L3 consensus and the fragment identified in the human PTPN1 mRNA is almost one in either L3 or human PTPN1 reading frames (0.95 and 0.71, respectively – the assumption of neutrality cannot be rejected by a Z-test in either example). This



```
                              10        20        30        40        50        60        70        80
                      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
Hsap_4506289:40-119   NKNRNRYRDVSPFDHSRIKLHQEDNDYINASLIKMEEAQRSYILTQGPLPNTCGHFWEMVWEQKSRGVVMLNRVMEKGSL
Mmus_50872127:40-119  NKNRNRYRDVSPFDHSRIKLHQEDNDYINASLIKMEEAQRSYILTQGPLPNTCGHFWEMVWEQKSRGVVMLNRIMEKGSL
Rnor_6981442:40-119   NKNRNRYRDVSPFDHSRIKLHQEDNDYINASLIKMEEAQRSYILTQGPLPNTCGHFWEMVWEQKSRGVVMLNRIMEKGSL
Ggal_45382391:40-119  NKNRNRYRDVSPFDHSRIKLNQGDNDYINASLIKMEEAQRSYILTQGPLPNTCGHFWEMVWEQKSRGVVMLNRVMEKGSI
Xlae_49256084:42-121  NKARNRYRDVSPFDHSRIKLHREDNDYINASLIKVEEAQRSYILTQGPLPNTCGHFWEMVWEQKSIGVVMLNRVIEKGSI
Drer_32451801:38-117  NRSRNRYRDVSPFDHSRICLQIGCNDYINASLISVEEAQRKYILTQGPLPNTCGHFWEMVWEQRSRGVVMLNRVIEKGSV
Hsap_4506291:42-121   NRNRNRYRDVSPYDHSRVKLQNAENDYINASLVDIEEAQRSYILTQGPLPNTCCHFWLMVWQQKTKAVVMLNRIVEKESV
Mmus_6679553:42-121   NRNRNRYRDVSPYDHSRVKLQSTENDYINASLVDIEEAQRSYILTQGPLPNTCCHFWLMVWQQKTKAVVMLNRTVEKESV
Rnor_16758896:42-121  NRNRNRYRDVSPYDHSRVKLQSAENDYINASLVDIEEAQRSYILTQGPLPNTCCHFWLMVWQQKTRAVVMLNRTVEKESV
Ggal_50736546:83-162  NRNRNRYRDVSPYDHSRVKLQNTENDYINASLVVIEEAQRYYILTQGPLPNTCCHFWLMVWQQQTKAVVMLNRIVEKDSV
Xlae_50414509:38-117  NKSRNRYRDVNPYDHSRVKLQNMDNDYINASLVVVEEAQRSYILTQGPLPNTCCHFWLMVWQQKSKAIVMLNRIIEKDAV
Drer_47086727:38-117  NRIRNRYRDVSPFDHSRVKLENTENDYINASLVVMEEAQRRYILTQGPLRNTCGHFWLMIWEQKTKAVIMLNRVIEKGSE
Amel_48096308:53-132  NKNLNRYRDVLPYDHSRIVLKKGPCDYINANLIQVDHARRQYILTQGPLENTAGHFWLMIWEQNSKAVLMLNKIIEKNHV
Dmel_17136412:60-139  NRGLNRYRDVNPYDHSRIVLKRGSVDYINANLVQLERAERQYILTQGPLVDTVGHFWLMVWEQKSRAVLMLNKLMEKKQI

                              90       100       110       120       130       140       150       160
                      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
                              ▼▼▼▼▼▼▼▼▼▼▼▼▼▼▼▼▼▼▼▼▼▼▼▼▼▼▼▼▼▼▼▼▼▼▼▼▼▼▼
Hsap_4506289:120-197  KCAQYWPQKE--EKEMIFEDTNLKLTLISEDIKSYYTVRQLELENLTTQETREILHFHYTTWPDFGVPESPASFLNFLFK
Mmus_50872127:120-197 KCAQYWPQQE--EKEMVFDDTGLKLTLISEDVKSYYTVRQLELENLTTKETREILPFHYTTWPDFGVPESPASFLNFLFK
Rnor_6981442:120-197  KCAQYWPQKE--EKEMVFDDTNLKLTLISEDVKSYYTVRQLELENLATQEAREILHFHYTTWPDFGVPESPASFLNFLFK
Ggal_45382391:120-197 KCAQYWPRKE--EKEMFFEDTNLKLTLISEDIKSYYTVRQLELENLTTQETREILHFHYTTWPDFGVPESPASFLNFLFK
Xlae_49256084:122-199 KCAQYWPKKE--DNPMVFDDSDLKLTLLSEDIKSYYTIRQLMLENLSTQETREILHFHYTTWPDFGVPESPASFLNFLFK
Drer_32451801:118-195 KCAQYWPQRE--EREAVFEDTNFRLTLISEDVKSYYTVRQLELENLSTQETREILHFHYTTWPDFGVPESPASFLNFLFK
Hsap_4506291:122-198  KCAQYWPTD---DQEMLFKETGFSVKLLSEDVKSYYTVHLLQLENINSGETRTISHFHYTTWPDFGVPESPASFLNFLFK
Mmus_6679553:122-198  KCAQYWPTD---DREMVFKETGFSVKLLSEDVKSYYTVHLLQLENINTGETRTISHFHYTTWPDFGVPESPASFLNFLFK
Rnor_16758896:122-198 KCAQYWPTD---DREMVFKETGFSVKLLSEDVKSYYTVHLLQLENINSGETRTISHFHYTTWPDFGVPESPASFLNFLFK
Ggal_50736546:163-240 KCAQYWPTRG--EEVMVFSETGFRVRLVSEDIKSYYTVHLLQLENINSGESRMISHFHYTTWPDFGVPESPASFLNFLFK
Xlae_50414509:118-195 KCAQYWPTPE--EEVLFYKETGLCVKLLSEDIKSYYTVRLLQLQDIKTGETRDIFHFHYTTWPDFGVPQSPASFLNFLFK
Drer_47086727:118-195 KCAQYWPTQE--EREMSFRDTRFVVTLVSEDVKSYYTTRVLELQNANTGETREIYHFHYTTWPDFGVPESPASFLNFLFK
Amel_48096308:133-212 KCYQYWPLGESVINTMIFPDVGMKIKYISKTESSDYTTRILKLTDLETKESREIFHFHYTTWPDFGVPQSPTAFLNFLTD
Dmel_17136412:140-219 KCHLYWPNEMGADKALKLPHVKLTVELVRLETYQNFVRRWFKLTDLETQQSREVMQFHYTTWPDFGIPSSPNAFLKFLQQ

                             170       180       190       200       210       220       230
                      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....
Hsap_4506289:198-276  VRESGSLSPEHGPVVVHCSAGIGRSGTFCLADTCLLLMDKRKDPSSVDIKKVLLEMRKFRMGLIQTADQLRFSYLAVIE
Mmus_50872127:198-276 VRESGSLSLEHGPIVVHCSAGIGRSGTFCLADTCLLLMDKRKDPSSVDIKKVLLEMRRFRMGLIQTADHVRFSYLAVIE
Rnor_6981442:198-276  VRESGSLSPEHGPIVVHCSAGIGRSGTFCLADTCLLLMDKRKDPSSVDIKKVLLEMRRFRMGLIQTADQLRFSYLAVIE
Ggal_45382391:198-276 VRESGSLNPEYGPVVVHCSAGIGRSGTFCLVDTCLLLMDKRKDPSSVDVKQVLLEMRKYRMGLIQTADQLRFSYLAVIE
Xlae_49256084:200-278 VRESGSLNPEHGPIVVHCSAGIGRSGTFSLADTCLLLMDKRKDPSSVEIKQVLLEMRKYRMGLIQTADQLRFSYLAVIE
Drer_32451801:196-274 VRESGCLSPELGPVVVHCSAGIGRSGTFCLVDTCLLLMSQRKDPSSVRIQEVLLEMRRYRMGLIQTADQLRFSYLAFIE
Hsap_4506291:199-274  VRESGSLNPDHGPAVIHCSAGIGRSGTFSLVDTCLVLMEKGD---DINIKQVLLNMRKYRMGLIQTPDQLRFSYMAIIE
Mmus_6679553:199-274  VRESGCLTPDHGPAVIHCSAGIGRSGTFSLVDTCLVLMEKGE---DVNVKQLLLNMRKYRMGLIQTPDQLRFSYMAIIE
Rnor_16758896:199-274 VRESGSLNPDHGPAVIHCSAGIGRSGTFSLVDTCLVLMEKGE---DVNVKQILLSMRKYRMGLIQTPDQLRFSYMAIIE
Ggal_50736546:241-318 VRESGCLSPEHGPAVVHCSAGIGRSGTFSLVDSCLVLMEK-KDPFSVDIKKVLLDMRKYRMGLIQTPDQLRFSYMAVIE
Xlae_50414509:196-274 VRESRSLSLQNGPAVVHCSAGIGRSGTFSLVDTCLVLMEKRKDPCSVHIKQVLLDMRKYRMGLIQTPGQLRFSYMAVIE
Drer_47086727:196-274 VRESGSLGMEQGPAVVHCSAGIGRSGTFSLVDTCLVLMDKRKDPLLVDIKKILLDMRKYRMGLIQTPDQLRFSYMAVLE
Amel_48096308:213-290 VRQSGTLDQNVGPPVVHCSAGIGRSGTFCLVDTCLVLIEE-NGLNSVNIRDILIEMRRSRMGLIQTPDQLRFSYAAIIE
Dmel_17136412:220-295 VRDSGCLSRDVGPAVVHCSAGIGRSGTFCLVDCCLVLIDKYG---ECNVSKVLCELRSYRMGLIQTADQLDFSYQAIIE
```

TRENDS in Genetics

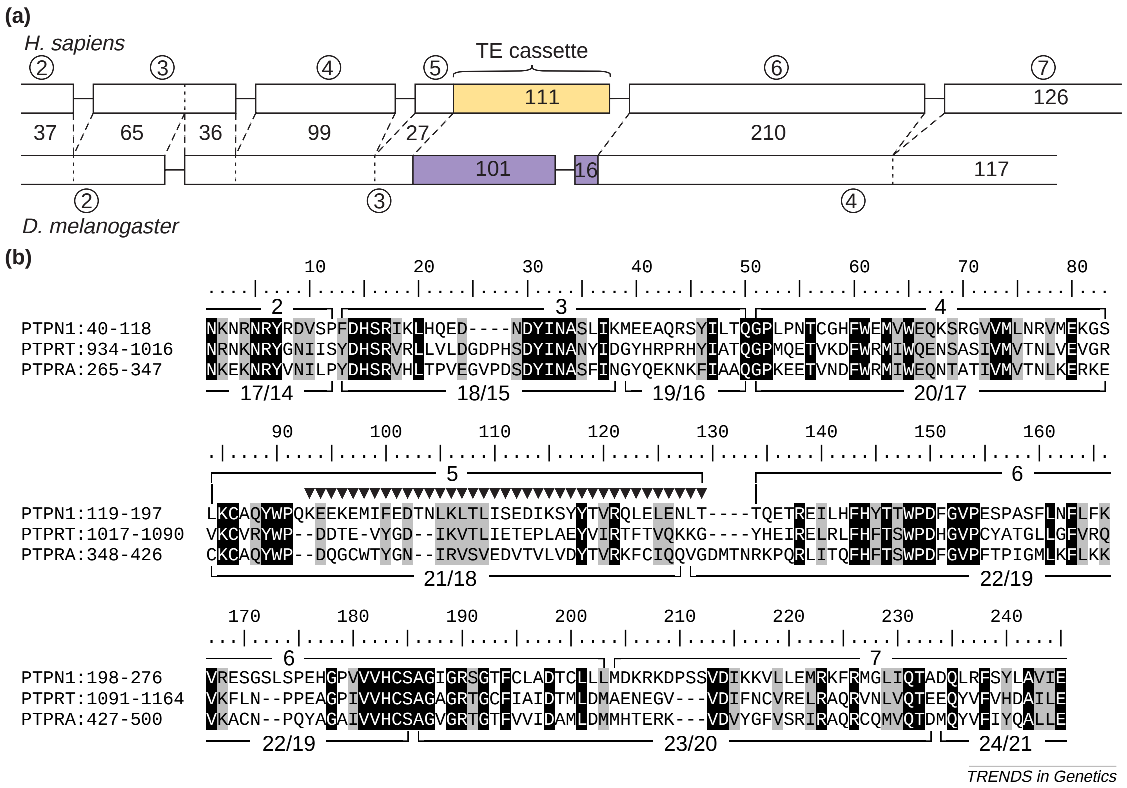
Figure 3. The amino-acid alignment of the PTP catalytic domain of animal non-receptor type PTPs. Species name, gi accession number and coordinates of residues included in alignment are indicated before every sequence. Identical and similar residues are highlighted in black and grey, respectively. The symbols (▼) above the alignment correspond to the L3-encoded fragment in PTPN1 (coordinates 89–127). Alignment coordinates 125–127 correspond to the manually extended alignment (shown in italics in Figure 1a).

is consistent with a period of neutral evolution that might have affected the TE in intron before exaptation. However, it is also consistent with a period of positive selection that the fragment could have experienced following exaptation, but before PTPN1 acquired a new specific function. This can explain why, despite the reasonably good nucleotide conservation, the coding function of the TE fragment has changed considerably (Figure 1). In addition to the

Table 2. Selection patterns in the catalytic domain of vertebrate and invertebrate non-receptor type PTPs, as determined by the ratio of dN to dS[a]

		PTPN1						PTPN2							
	Species	***Hsap*[b]**	***Mmus*[c]**	***Rnor*[d]**	***Ggal*[e]**	***Xlae*[f]**	***Drer*[g]**	***Hsap*[b]**	***Mmus*[c]**	***Rnor*[d]**	***Ggal*[e]**	***Xlae*[f]**	***Drer*[g]**	***Amel*[h]**	***Dmel*[i]**
PTPN1	***Hsap*[b]**		0.142	0.090	0.022	0.271	0.151	0.657	0.475	0.573	0.387	0.481	0.505	*0.737*	*0.964*
	***Mmus*[c]**	0.054		0.151	0.081	0.264	0.200	0.426	0.381	0.438	0.394	0.514	0.421	*0.650*	*0.981*
	***Rnor*[d]**	0.035	0.084		0.056	0.282	0.120	0.463	0.445	0.477	0.386	0.551	0.434	*0.602*	*0.892*
	***Ggal*[e]**	0.043	0.078	0.062		0.301	0.162	0.571	0.579	0.571	0.446	0.395	0.366	*0.789*	*0.764*
	***Xlae*[f]**	0.059	0.083	0.064	0.088		0.428	0.800	0.600	0.668	0.492	0.542	0.440	*0.555*	*1.048*
	***Drer*[g]**	0.109	0.143	0.128	0.136	0.139		0.468	0.350	0.436	0.438	0.503	0.341	*0.691*	*0.985*
PTPN2	***Hsap*[b]**	0.219	0.236	0.221	0.234	0.234	0.241		0.086	0.095	0.276	0.272	0.449	*0.956*	*1.096*
	***Mmus*[c]**	0.225	0.248	0.224	0.243	0.234	0.243	0.073		0.000	0.241	0.303	0.366	*1.309*	*1.029*
	***Rnor*[d]**	0.229	0.253	0.231	0.218	0.225	0.248	0.062	0.124		0.262	0.309	0.372	*1.069*	*1.088*
	***Ggal*[e]**	0.152	0.175	0.158	0.217	0.224	0.211	0.119	0.131	0.144		0.256	0.437	*0.969*	*0.983*
	***Xlae*[f]**	0.189	0.198	0.186	0.199	0.180	0.199	0.155	0.190	0.186	0.122		0.406	*1.019*	*0.874*
	***Drer*[g]**	0.199	0.225	0.213	0.207	0.191	0.227	0.213	0.178	0.191	0.150	0.156		*0.712*	*0.809*
	***Amel*[h]**	0.293	0.297	0.285	0.347	0.337	0.289	0.347	0.372	0.358	0.345	0.349	0.304		0.698
	***Dmel*[i]**	0.385	0.461	0.438	0.354	0.360	0.341	0.369	0.358	0.368	0.342	0.369	0.452	0.290	

[a] The order of sequences is the same as in Figure 2. Modified Nei-Gojobori method [56] using *P*-distance and complete gap deletion was used (Jukes-Cantor model of nucleotide substitution was not used because *P*-distance for many pairs is >0.75). The shaded grey cells correspond to pairs for which assumption of neutrality cannot be rejected by a Z-test. The variance was computed analytically for the L3 corresponding segments because of the few codons. The significance was tested with the variance computed both analytically and by bootstrap for the rest of the domain. The segment corresponding to the L3 consensus (89–127 in Figure 2) is not considered homologous between vertebrate and invertebrate sequences, and therefore dN/dS values are meaningless for those pairs (shown in italics). Values in the upper-right were computed for the segment matching the L3 consensus sequence (shown as triangles in Figure 2). Values in the lower-left side were computed for the rest of the PTP domain (195 codons).
[b] *Homo sapiens*, [c] *Mus musculus*, [d] *Rattus norvegicus*, [e] *Gallus gallus*, [f] *Xenopus laevis*, [g] *Danio rerio*, [h] *Apis mellifera*, [i] *Drosophila melanogaster*.


(a)
H. sapiens
②
③
④
⑤
TE cassette
⑥
⑦
111
126
37
65
36
99
27
210
101
16
117
②
③
④
D. melanogaster
(b)
10 20 30 40 50 60 70 80
2
3
4
PTPN1:40-118 NKNRNRYRDVSPFDHSRIKLHQED----NDYINASLIKMEEAQRSYILTQGPLPNTCGHFWEMVWEQKSRGVVMLNRVMEKGS
PTPRT:934-1016 NRNKNRYGNIISYDHSRVRLLVLDGDPHSDYINANYIDGYHRPRHYIATQGPMQETVKDFWRMIWQENSASIVMVTNLVEVGR
PTPRA:265-347 NKEKNRYVNILPYDHSRVHLTPVEGVPDSDYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWEQNTATIVMVTNLKERKE
17/14
18/15
19/16
20/17
90 100 110 120 130 140 150 160
5
6
PTPN1:119-197 LKCAQYWPQKEEKEMIFEDTNLKLTLISEDIKSYYTVRQLELENLT----TQETREILHFHYTTWPDFGVPESPASFLNFLFK
PTPRT:1017-1090 VKCVRYWP--DDTE-VYGD--IKVTLIETEPLAEYVIRTFTVQKKG----YHEIRELRLFHFTSWPDHGVPCYATGLLGFVRQ
PTPRA:348-426 CKCAQYWP--DQGCWTYGN--IRVSVEDVTVLVDYTVRKFCIQQVGDMTNRKPQRLITQFHFTSWPDFGVPFTPIGMLKFLKK
21/18
22/19
170 180 190 200 210 220 230 240
6
7
PTPN1:198-276 VRESGSLSPEHGPVVVHCSAGIGRSGTFCLADTCLLLMDKRKDPSSVDIKKVLLEMRKFRMGLIQTADQLRFSYLAVIE
PTPRT:1091-1164 VKFLN--PPEAGPIVVHCSAGAGRTGCFIAIDTMLDMAENEGV---VDIFNCVRELRAQRVNLVQTEEQYVFVHDAILE
PTPRA:427-500 VKACN--PQYAGAIVVHCSAGVGRTGTFVVIDAMLDMMHTERK---VDVYGFVSRIRAQRCQMVQTDMQYVFIYQALLE
22/19
23/20
24/21

TRENDS in Genetics

Figure 4. A comparison of PTPN1 gene structure with other invertebrate and human PTPs. **(a)** A comparison of the gene structure in the PTP domain area of human PTPN1 with that of *Drosophila melanogaster* PTP61F. Exon-intron boundaries were determined using Spidey (www.ncbi.nlm.nih.gov/spidey) and the following sequences (NCBI gi numbers): 17390366 (mRNA), 51511747 (genomic) for human and 24655162 (mRNA), 56411837 (genomic) for *D. melanogaster*. Sizes, in nucleotides, of corresponding exonic blocks are indicated between the two genes if identical in size, or in each exon if different between the two species. Exon numbers are shown (in circles) above and below the gene structure for human and *D. melanogaster*, respectively; the introns are not drawn to scale. (**b**) Alignment of the PTP catalytic domain of human PTPs located on chromosome 20. Exon boundaries and numbers are shown above the alignment for PTPN1, and below the alignment for PTPRT and PTPRA. For the latter, the first number corresponds to PTPRT and the second for PTPRA. The gene structures of PTPRT and PTPRA are identical, whereas that of PTPN1 is different after exon 5, which contains the L3-encoded fragment. However, it is difficult to determine the direction of causality, that is whether exaptation determined the change of gene structure or vice-versa.

expected nucleotide substitutions, deletions that were not multiple of three nucleotides determined the change of reading frame for two segments (Figure 1a). Consequently, the number of hydrophobic residues was reduced from nine to four in those regions, facilitating the formation of hairpin loops (Figure 1b). However, it is difficult to say whether most of these changes occurred after exaptation to increase protein functional efficiency, as can occur with any random sequence [43], or if they occurred in a fortuitous manner that actually facilitated the exaptation event. Whatever the case, it is remarkable that following the exaptation event and the subsequent intra-chromosomal duplication, both PTPN1 and PTPN2 acquired specific functions (Table 2) that probably do not exist in invertebrates, which have much fewer PTPs.

Concluding remarks and directions for further research

The confirmation that TEs are present at the protein level is by no means a surprise, and they are certainly not the only category of DNA sequence to be exapted successfully into functional proteins. Hayashi *et al.* showed that any random sequence could acquire biological functions if it had sufficient time to evolve [43]. It is, however, their prevalence and mobility within genomes that make TEs important players in molecular and genomic evolution.

Gene duplications – key events that favor exaptation

One common feature of the PTP, calpain and granzyme protein families is that they were all diversified by multiple gene duplication events. A newly duplicated gene is likely to be free of functional constraints, and therefore can more easily accommodate major changes [40], such as the exaptation of TE sequences. If those genes are preserved and acquire new specific functions, the influence of TEs is then directly reflected through the function of the host protein. This aspect can be further investigated in other protein families known to have been diversified through extensive gene duplications.

Phylogenies – the key for validation of low-scoring TE cassettes

Another common feature of the TE cassettes uncovered by us is that they all have lengths, divergence from TE consensus and RM scores similar to those of cases considered to be false positives (Tables 1 and S1). Therefore, an accurate distinction between random matches and real TE cassettes cannot be made based on any of those criteria. Moreover, not even the sequence randomization test (online supplementary material) can distinguish between the two because the *P*-values in all examples are small. In these conditions, it is only the phylogenetic history of a gene that can confirm the validity of an RM match, as shown for the TE cassettes in PTPN1, CAPN and GZMA. Interestingly, all three cassettes are derived from old repeats, which is consistent with the idea that a nonaptation period is usually required for the fortuitous shaping of such elements before successful exaptation into the ORF of a gene. In contrast to the two sequences derived from L3 ORF2p, which are both part of anti-parallel β-sheets in PTPN1 and GZMA, the sequence derived from the non-coding tRNA-like MIR region forms a simple loop region in CAPN1. The fate and importance of the exapted TE fragment therefore appears to be determined by its original role in the parent TE.

However, we cannot completely exclude the hypothesis of sequence convergence for any of the TE cassettes. This is because all real TE cassettes are likely to be old, and are therefore short and highly diverged from their original sequence, which means that random matches that would resemble such TE fragments are likely to occur (21/24 putative TE cassettes seem to be random matches). However, the exaptation scenario should be favored when support from phylogeny exists, because the probability of having both a random match and phylogenetic support for the same protein fragment is lower than having a random match alone. Therefore, we urge scientists to treat low scoring RepeatMasker matches with special attention because some might prove to be real 'treasures among the junk' [7].

TE cassettes – discrepancy between the frequency of occurrence in transcripts and functional proteins

We were surprised to find fewer TE cassettes (∼0.1%) in functional proteins than one would expect (∼4%) from the translation of TE-containing transcripts [22,23]. In contrast to our findings, most TE cassettes at the transcript level are derived from young TEs, and appear in a minor, alternatively spliced form of cognate mRNAs [22,23,44]. They can even persist as such over long evolutionary periods [45], indicating that they might represent neither successful exaptations for protein coding purposes nor the intermediate stages of such events. They must have a different important role or otherwise they would be lost. Two articles provide a clue as to what that role might be. First, Oh *et al.* showed that co-expression of the α, β and γ subunits of wild-type human epithelial sodium channel (hENaC) with an Alu-containing splicing variant of the α subunit (hαENaC+Alu) enhanced the expression of the amiloride-sensitive current in oocytes [46]. Second, Hirotsune *et al.* showed that an expressed pseudogene, *Makorin1-p1*, protected its cognate protein coding gene from mRNA decay, most likely by competition over a *trans*-acting RNA-destabilizing factor [47]. The expression of TE-containing transcript variants, or even of pseudogenes, can thus regulate the expression or enhance the function of the functional protein coding form. The significant number of TE-containing transcripts might indicate that the role of TEs in regulation of gene expression and function is more important that it is currently acknowledged, and requires further insight.

TE cassettes in functional proteins are currently underestimated

Despite the few real TE cassettes we found in functional proteins, we think that the real number is underestimated. One reason for this is that transmembrane-, signal-, disordered- and low-complexity protein regions are significantly under-represented in the PDB collection [48], because of the way targets are selected for structural genomics [49]. We can only hope that further studies will characterize more proteins from the under-represented classes. A second reason is that all TE-cassettes that we found are derived from old TEs (Table 1), which might cause the exaptation events to be obscured by long evolutionary periods. In addition, old TEs are usually difficult to identify because of their highly diverged and fragmented sequence. Similarity searching techniques currently employed for finding TEs are not optimized for these types of sequences; therefore, we would encourage the scientific community to implement better techniques for detecting fragments of older TEs. For example, the use of position weight matrices instead of consensus sequences for finding diverged MIR copies seemed to be a promising approach [50] and could be applied to other TEs. Despite these reasons, the real proportion of TE-containing proteins is probably closer to our estimate of ∼0.1% than to previous estimates of ∼4% [22,23] because we do not expect to find TE cassettes of young elements in functional proteins.

Young TEs: subject to future exaptation events

An important conclusion of our study is that functional proteins are unlikely to contain TE cassettes derived from young TEs, such as Alu and L1s. This is in contrast to previous reports [22,23], which estimated that they represent up to 60% of the human TE cassettes in ORFs. Even if that might be true at the transcript level, it seems unlikely that young TEs could be found in functional proteins because long evolutionary periods are needed for successful exaptation events. For example, Alu elements, which are found only in primate genomes, did not have enough time to evolve and adapt to new coding functions, but they seem to be currently undergoing that process [45]. As a result, they often cause problems when inserted into protein coding regions [51]. By contrast, there are examples of young repeats that contributed to the human proteome, but did not undergo exaptation. Elements such as the human endogenous retroviruses HERV-FRD were co-opted by the human genome, and their protein product has a function similar to that of the original retrovirus gene [52]. Nonetheless, we should not be surprised if exaptation of currently young TEs will eventually yield functional proteins – we just need to give nature enough time.

Acknowledgements

We thank Dimitra Chalkia for her contribution to the construction of MME and ARFIP2 phylogenies, and her constructive criticisms on the initial versions of the article. We are grateful to Vamsi Veeramachaneni for setting up the initial protein data set; to Jordi Bella and Paul McEwan for their help with protein structure visualization tools; to Nikolas Nikolaidis, Jongmin Nam and Arthur Lesk for critically reading the article; and to the reviewers for their useful comments and suggestions. V.G. was partially supported by Penn State's Worldwide Universities Network program.

Supplementary data

Supplementary data associated with this article can be found at doi:10.1016/j.tig.2006.03.006

References

1 Lander, E.S. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921

2 Ohno, S. (1972) So much 'junk' DNA in our genome. *Brookhaven Symp. Biol.* 23, 366–370

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

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

5 Hickey, D.A. (1982) Selfish DNA: a sexually-transmitted nuclear parasite. *Genetics* 101, 519–531

6 Brosius, J. (1991) Retroposons – seeds of evolution. *Science* 251, 753

7 Nowak, R. (1994) Mining treasures from 'junk DNA'. *Science* 263, 608–610

8 Chance, P.F. *et al*. (1994) Two autosomal dominant neuropathies result from reciprocal DNA duplication/deletion of a region on chromosome 17. *Hum. Mol. Genet.* 3, 223–228

9 Deininger, P.L. *et al*. (2003) Mobile elements and mammalian genome evolution. *Curr. Opin. Genet. Dev.* 13, 651–658

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

11 van de Lagemaat, L.N. *et al*. (2003) Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet.* 19, 530–536

12 Landry, J.R. *et al*. (2003) Complex controls: the role of alternative promoters in mammalian genome. *Trends Genet.* 19, 640–648

13 Kazazian, H.H., Jr. (2004) Mobile elements: drivers of genome evolution. *Science* 303, 1626–1632

14 Thornburg, B. *et al*. (2006) Transposable elements as a significant source of transcription regulating signals. *Gene* 365, 104–110

15 Pavlicek, A. *et al*. (2002) Transposable elements encoding functional proteins: pitfalls in unprocessed genomic data? *FEBS Lett.* 523, 252–253

16 Lundwall, A.B. *et al*. (1985) Isolation and sequence analysis of a cDNA clone encoding the fifth complement component. *J. Biol. Chem.* 260, 2108–2112

17 Caras, I.W. *et al*. (1987) Cloning of decay-accelerating factor suggests novel use of splicing to generate two proteins. *Nature* 325, 545–549

18 Brownell, E. *et al*. (1989) A human rel proto-oncogene cDNA containing an Alu fragment as a potential coding exon. *Oncogene* 4, 935–942

19 Mitchell, G.A. *et al*. (1991) Splice-mediated insertion of an Alu sequence inactivates ornithine delta-aminotransferase: a role for Alu elements in human mutation. *Proc. Natl. Acad. Sci. U. S. A.* 88, 815–819

20 Makałowski, W. *et al*. (1994) Alu sequences in the coding regions of mRNA: a source of protein variability. *Trends Genet.* 10, 188–193

21 Li, W.H. *et al*. (2001) Evolutionary analyses of the human genome. *Nature* 409, 847–849

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

23 Lorenc, A. and Makałowski, W. (2003) Transposable elements and vertebrate protein diversity. *Genetica* 118, 183–191

24 Gerber, A. *et al*. (1997) Two forms of human double-stranded RNA-specific editase 1 (hRED1) generated by the insertion of an Alu cassette. *RNA* 3, 453–463

25 Hilgard, P. *et al*. (2002) Translated Alu sequence determines nuclear localization of a novel catalytic subunit of casein kinase 2. *Am. J. Physiol. Cell Physiol.* 283, C472–C483

26 Hoenicka, J. *et al*. (2002) A two-hybrid screening of human Tau protein: interactions with Alu-derived domain. *Neuroreport* 13, 343–349

27 Jacobson, A. and Peltz, S.W. (1996) Interrelationships of the pathways of mRNA decay and translation in eukaryotic cells. *Annu. Rev. Biochem.* 65, 693–739

28 Hilleren, P. and Parker, R. (1999) Mechanisms of mRNA surveillance in eukaryotes. *Annu. Rev. Genet.* 33, 229–260

29 Wagner, E. and Lykke-Andersen, J. (2002) mRNA surveillance: the perfect persist. *J. Cell Sci.* 115, 3033–3038

30 Lovell, S.C. (2003) Are non-functional, unfolded proteins ('junk proteins') common in the genome? *FEBS Lett.* 554, 237–239

31 Deragon, J.M. and Capy, P. (2000) Impact of transposable elements on the human genome. *Ann. Med.* 32, 264–273

32 Pradet-Balade, B. *et al*. (2001) Translation control: bridging the gap between genomics and proteomics? *Trends Biochem. Sci.* 26, 225–229

33 Rogozin, I.B. *et al*. (2003) Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr. Biol.* 13, 1512–1517

34 Charbonneau, H. *et al*. (1989) Human placenta protein-tyrosine-phosphatase: amino acid sequence and relationship to a family of receptor-like proteins. *Proc. Natl. Acad. Sci. U. S. A.* 86, 5252–5256

35 Barford, D. *et al*. (1994) Crystal structure of human protein tyrosine phosphatase 1B. *Science* 263, 1397–1404

36 Kapitonov, V.V. and Jurka, J. (2003) The esterase and PHD domains in CR1-like non-LTR retrotransposons. *Mol. Biol. Evol.* 20, 38–46

37 Ono, K. *et al*. (1999) Multiple protein tyrosine phosphatases in sponges and explosive gene duplication in the early evolution of animals before the parazoan–eumetazoan split. *J. Mol. Evol.* 48, 654–662

38 Ono-Koyanagi, K. *et al*. (2000) Protein tyrosine phosphatases from amphioxus, hagfish, and ray: divergence of tissue-specific isoform genes in the early evolution of vertebrates. *J. Mol. Evol.* 50, 302–311

39 Andersen, J.N. *et al*. (2004) A genomic perspective on protein tyrosine phosphatases: gene structure, pseudogenes, and genetic disease linkage. *FASEB J.* 18, 8–30

40 Ohno, S. (1970) *Evolution by gene duplication*, Springer-Verlag

41 Lev-Maor, G. *et al*. (2003) The birth of an alternatively spliced exon: 3′ splice-site selection in Alu exons. *Science* 300, 1288–1291

42 Makałowski, W. (2001) The human genome structure and organization. *Acta Biochim. Pol.* 48, 587–598

43 Hayashi, Y. *et al*. (2003) Can an arbitrary sequence evolve towards acquiring a biological function? *J. Mol. Evol.* 56, 162–168

44 Sorek, R. *et al*. (2002) Alu-containing exons are alternatively spliced. *Genome Res.* 12, 1060–1067

45 Krull, M. *et al*. (2005) Alu-SINE exonization: *en route* to protein-coding function. *Mol. Biol. Evol.* 22, 1702–1711

46 Oh, Y.S. *et al*. (2001) An Alu cassette in the human epithelial sodium channel. *Biochim. Biophys. Acta* 1520, 94–98

47 Hirotsune, S. *et al*. (2003) An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature* 423, 91–96

48 Peng, K. *et al*. (2004) Exploring bias in the Protein Data Bank using contrast classifiers. *Pac. Symp. Biocomput.*, 435–446

49 Brenner, S.E. (2000) Target selection for structural genomics. *Nat. Struct. Biol.* 7(Suppl.), 967–969

50 Chaley, M.B. and Korotkov, E.V. (2001) Evolution of the MIR elements located in the coding regions of the human genome. *Mol. Biol.* 35, 874–882

51 Deininger, P.L. and Batzer, M.A. (1999) Alu repeats and human disease. *Mol. Genet. Metab.* 67, 183–193

52 Renard, M. *et al*. (2005) Crystal structure of a pivotal domain of human syncytin-2, a 40 million years old endogenous retrovirus fusogenic envelope gene captured by primates. *J. Mol. Biol.* 352, 1029–1034

53 Finnegan, D.J. (1989) Eukaryotic transposable elements and genome evolution. *Trends Genet.* 5, 103–107

54 Gould, S.J. and Vrba, E.S. (1982) Exaptation – a missing term in the science of form. *Paleobiology* 8, 4–15

55 Brosius, J. and Gould, S.J. (1992) On 'genomenclature': a comprehensive (and respectful) taxonomy for pseudogenes and other 'junk DNA'. *Proc. Natl. Acad. Sci. U. S. A.* 89, 10706–10710

56 Nei, M. and Gojobori, T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3, 418–426

57 Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425