Transposable elements have contributed to thousands of human proteins

Roy Britten*

California Institute of Technology, 101 Dahlia Avenue, Corona del Mar, CA 92625

Contributed by Roy Britten, November 18, 2005

This is a report of many distant but significant protein sequence relationships between human proteins and transposable elements (TEs). The libraries of human repeated sequences contain the DNA sequences of many TEs. These were translated in all reading frames, ignoring stop codons, and were used as amino acid sequence probes to search with BLASTP for similar sequences in a library of 25,193 human proteins. The probes show regions of significant amino acid sequence similarity to 1,950 different human genes, with an expectation of $<10^{-3}$. In comparison with previous REPEATMASKER (Institute for Systems Biology, Seattle) studies, these probes detect many more TE sequences in more human coding sequences with greater length than previous work using DNA sequences. If the criterion is opened, very many matches are found occurring on 4,653 different genes after correction for the number seen with random amino acid sequence probes. The processes that led to these extensive sets of sequence relationships between TEs and coding sequences of human genes have been a major source of variation and novel genes during evolution. This paper lists the number of sequence similarities seen by amino acid sequence comparison, which is surely an underestimate of the actual number of significant relationships. It appears that many of these are the result of past events of duplication of genes or gene regions, rather than a direct result of TE insertion. This report of observable relationships leaves to the future the functional implications as well as the detection of the events of TE insertion.

sequence similarity | mobile elements | genes | repeated sequences

The contribution of transposable elements (TEs) to eukaryotic genomes has been much studied and will not be reviewed here. I use the phrase TE to refer to any of the known repeated sequences, not including short tandem repeats and simple sequences. The presumption is that most TEs are the product of past insertions of active TE family members into the genomes of many eukaryote ancestors. No uncertainty in meaning results for the minority for which this may not have been demonstrated. Usage does not imply that the elements are currently active and transposable.

This article describes measurements of the relationships of TE sequences to human coding sequences, which probably represent much of the contribution TEs have made to the coding sequences of human proteins. The logic underlying this approach is that among the human "repeated sequences," there are many sequences that are related to TEs but are mostly incomplete and damaged, and the original ORFs of the once active TE cannot be recognized in many cases. Therefore, I have simply taken as probes the translation in each of the possible reading frames, ignoring stop codons, to make a set of six probes for each of the repeated sequences. The sequences available may well have undergone various mutations, including frame-shift events, and these will have broken up and limited the length of the TE probes. The relationship of the probes to the original TE is uncertain. The test is whether amino acid sequence similarities to human proteins are found. It turns out that these probes are more effective in finding TE sequence relationships to coding sequences than the previously used DNA sequence methods (1–5). The amino acid sequence probes identify about six times as many as the DNA sequence searches, and the regions of sequence similarity are much longer. The amino acid sequences of these probes are effective, in part, because the DNA sequences drift more rapidly during sequence evolution. In sum, these amino acid sequence probes work well, but the result is assuredly an incomplete view of the TE sequences in human proteins. It must be realized that the events responsible for the relationships occurred in the distant past, and thus the sequence relationships are very divergent and difficult to separate from similarities that may occur among random sequences. Among the reasons for the large number of observed relationships between the TEs and human proteins is that many gene regions have been duplicated in the past and are sequence-related to each other (6). The sequence regions that are similar to TEs are included among the regions that have many relationships to other proteins and thus have been duplicated many times in the past. Some genes include within their length many copies of local blocks of amino acid sequence as a result of internal regional events of duplication. In some cases, regions with similarity to TEs have been included in this process. As a result of these phenomena, the majority of similarities of amino acid sequences to TEs are the result of duplications rather than actual insertion events. Future work will be required to detect the actual insertion events.

In the past, it has been shown that TEs have supplied information for protein amino acid sequences (1–4) based on the DNA sequence relationship between the coding sequences and known TE sequences. Observations have identified 19 (2) and 47 (3) genes that appear to have coding sequences derived in part from TEs. There is also evidence that half a dozen protein coding sequences are derived almost entirely from TEs (5). Recent work, also based on DNA sequence relationships, reports 533 cases of TE sequences within human protein coding sequences (4), recognized by REPEATMASKER (RM) (Institute for Systems Biology, Seattle).

Results

RM Identification of TE Insertions in Coding Regions. RM was used to test for the presence of repeated sequences considered to be the product of insertion of TE. The results of the RM analysis of the library of 25,193 human protein coding sequences (see *Methods*) identify 934 examples of the presence of TE segments in these sequences. In our observations, some coding sequences contained more than one TE, and there were 814 different genes containing TE sequences. There is agreement with a previous observation (4), where 3.8% of 13,799 UniGene sequences were found to contain TEs according to RM. In comparison, we observed 937/25,193 or 3.7% of the total as containing TEs. Many TEs from this collection are identified as hypothetical genes or derived from ESTs. When the table of cases (4) in which TE were present was examined, 164 could be identified as

Conflict of interest statement: No conflicts declared.

Abbreviations: TE, transposable element; RM, REPEATMASKER; tv, transcript variant.

*E-mail: rbritten@caltech.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0510007103

Table 1. Comparison of methods to detect TE sequences among human genes

Method	No. of probes	TEs that match*	Different genes†	Average length‡
RM	?	263	814	42
DNA	835	283	537	31
Six frame	5,010	344	1,950	257

*The number of different TEs that match regardless of the reading frame.
†The number of different genes that are matched.
‡The longest match to each gene is averaged.

"known protein" genes. This is 1.1% of all 13,799 UniGene sequences, compared with the 262/25,193 or 1.04% we observed. Considering the statistical uncertainty and difficulty of identification of "known protein" genes from brief descriptions, this is an acceptable agreement. Both sets of observations show that a much larger fraction of putative or apparent proteins contain TE inserts. It is not certain whether this is an artifact.

Amino Acid Sequence Search for Ancient TE Insertions in Coding Sequences Using RM Finds. One might expect that insertions of TE into coding sequences has been a continuous process in the past, during the whole period when TEs were present in the genome. It is not known how long TEs have been active, but it probably extends far into the past. The data support such a view and indicate conservation of the resulting amino acid sequences in some cases; therefore, amino acid similarities permit looking into the distant past. We have used as probes the amino acid sequences of the regions of the genes in which the 934 examples of TE inserts in the coding regions of genes were recognized by RM. We compared them with the amino acid sequences of a library of 25,193 genes and found many thousands of proteins with apparent TE sequences. However, this work was terminated because of the superior quality of the amino acid sequence probes to be described.

Comparison of Three Methods of Detecting TE Sequences in Human Protein Coding Sequences. The three methods are: (*i*) RM, (*ii*) DNA sequence probes using BLAST, and (*iii*) six-frame translations of the DNA sequences as probes using BLASTP. For the first two methods, the target library was the mRNA sequences of 25,193 human genes. For the third method, the target was the protein sequences of the same library, using BLASTP. For the two BLAST runs, the criterion was that the expectation should be $\leq 10^{-3}$ (see *Methods*). The RM run was at the default criterion. Table 1 shows the comparison of the three methods. The number 344 is the number of different TEs that match regardless of the translation frame. It is obvious that the six-frame translation method is more sensitive, involves more TEs than the other methods, and makes much longer matches. The average length is in protein sequence equivalent or the DNA sequence length divided by three. For the RM length value in each case, RM chooses the best match by default criterion. For the other two methods, the longest match to each gene is averaged. For the average length of all matches, these numbers fall from 31 to 27 and from 257 to 210.

Search Using Amino Acid Sequence Probes. The first step was to translate, in all six reading frames, each of the collections of human repeated sequences. All of these may be considered potential TEs and, if not, may still be worth investigation if they find matches to human proteins. The resulting collection included 5,010 probes translated from 835 potential TEs. It is derived from the library used by RM (7). When this collection was compared by using blastp to a library of 25,193 human proteins, there were matches to 1,950 different human protein sequences with a limit that the expectation was $\leq 10^{-3}$.

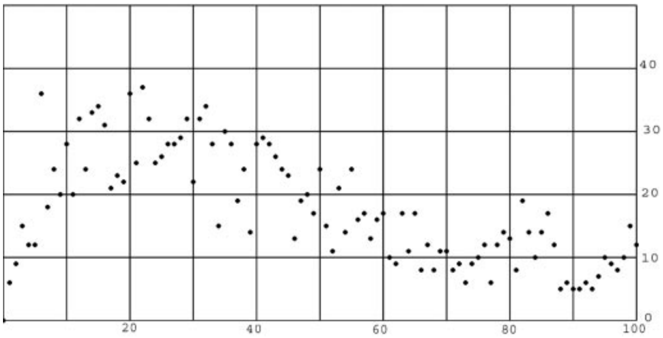

Fig. 1. The number of the 1,950 matches vs. coverage. Horizontally is the fraction of the length of the gene included in all of the matches to the amino acid sequence probes in 1% intervals. Vertically is the number of genes in each interval. Many different TEs contribute to the coverage of individual proteins.

The 1,950 different proteins matched have direct significance. The data in Table 1 more than triple the number of published relationships of TEs to human coding sequences from 533 (4) to 1,950. We have observed TEs in 814 different genes using RM to examine the library of 25,319 human genes. These data show the greater sensitivity to detect TE sequences using the six-frame translations of human repeated sequences, compared with RM. The 1,950 number is the largest estimate, up to this measurement of the TE sequences present in human proteins. Fig. 1 shows the fraction of the gene length that matches these amino acid sequence probes, also called the coverage. The number of genes is plotted against the percent of the length covered. At the right of the curve are a few hundred genes that have >80% of their length matching TEs.

The relationships observed are the result of the history of the proteins, including (for genes with repetitive structure) the results of unequal crossing over and slippage. There is also evidence (6) that the majority of genes are related to many other human genes, probably as the result of past duplications of regions. Studies (not shown) of the TE regions indicate they are often included in these gene-to-gene sequence similarities. The point is there is no one-to-one correspondence between the observation of a sequence relationship between a TE and any gene and the original insertion of the TE into a gene in the past. This is particularly obvious for genes that occur in large families and with large coverage, such as the collagens. Our studies

Table 2. Number of genes vs. percent coverage in 10% intervals

Percent*	Total†	KGP‡	Ratio
1–10	180	144	0.78
11–20	276	228	0.83
21–30	284	192	0.68
31–40	252	170	0.68
41–50	223	163	0.74
51–60	164	106	0.66
61–70	114	70	0.62
71–80	99	65	0.66
81–90	109	69	0.57
91–100	87	42	0.51

KGP, known gene product.
*The percent of the length of the genes included in matches.
†For all genes that show matches in the 25,193 member library.
‡For the genes that show matches to the 16,655 members known to produce known proteins.

Table 3. The functions of genes >80% covered by TE sequence matches

Ref Seq ID	Percent coverage	Description
NM_000501	98	Elastin (supravalvular aortic stenosis) (ELN)
	96	Collagen, type IX, α 1 (COL9A1), tv 1, 2
NM_174945	80	Zinc finger protein 575 (ZNF575) gi|28372566|
NM_004326	80	B-cell CLL/lymphoma 9 (BCL9) gi|72256199|
	95	Collagen, type IV, α 3 (Goodpasture antigen) (COL4A3), tv 1, 2, 3, 4, 5, 6
NM_000090	81	Collagen, type III, α 1 (Ehlers-Danlos syndrome type IV) (COL3A1)
NM_145056	86	Thymus expressed gene 3-like (MGC15476) gi|21450823|
NM_181950	95	HANP1 (H1T2) gi|32401436|
NM_005066	82	Splicing factor proline/glutamine-rich (SFPQ)
NM_012390	86	Submaxillary gland androgen regulated protein 3 homolog A (mouse) (SMR3A)
NM_005202	82	Collagen, type VIII, α 2 (COL8A2) gi|32964829|
NM_152263	85	Tropomyosin 3 (TPM3), tv 1 gi|22748618|
NM_003577	82	Undifferentiated embryonic cell transcription factor 1 (UTF1) gi|71043875|
NM_000089	80	Collagen, type I, α 2 (COL1A2) gi|48762933|
NM_004052	98	BCL2/adenovirus E1B 19 kDa interacting protein 3 (BNIP3)
NM_003407	82	Zinc finger protein 36, C3H type, homolog (mouse) (ZFP36) gi|4507960|
NM_006249	93	Proline-rich protein BstNI subfamily 3 (PRB3) gi|41349487|
NM_001948	96	dUTP pyrophosphatase (DUT) tv 2
	82	APEX nuclease (multifunctional DNA repair enzyme) 1 (APEX1), tv 1, 2, 3
NM_016426	81	G-2 and S-phase expressed 1 (GTSE1) gi|51317385|
NM_057160	86	Artemin (ARTN), tv 3 gi|16950644|
NM_003566	80	Early endosome antigen 1, 162 kDa (EEA1) gi|55770887|
NM_016333	92	Serine/arginine repetitive matrix 2 (SRRM2) gi|19923465|
NM_000493	82	Collagen, type X, α 1 (Schmid metaphyseal chondrodysplasia) (COL10A1)
NM_001950	83	Atrophin 1 (ATN1), tv 2 gi|55750040|
NM_002973	80	Ataxin 2 (ATXN2) gi|51479159|
NM_001852	85	Collagen, type IX, α 2 (COL9A2) gi|31083125|
	81	Collagen, type II, α 1 (primary osteoarthritis) (COL2A1), tv 1, 2
NM_002356	93	Myristoylated alanine-rich protein kinase C substrate (MARCKS)
NM_001845	85	Collagen, type IV, α 1 (COL4A1) gi|45580690|
NM_005839	85	Serine/arginine repetitive matrix 1 (SRRM1) gi|42542378|
NM_002687	85	Pinin, desmosome associated protein (PNN) gi|33356173|
NM_003724	99	Jerky homolog (mouse) (JRK) gi|22208998|
NM_000393	80	Collagen, type V, α 2 (COL5A2) gi|16554580|
	86	Collagen, type IV, α 5 (Alport syndrome) (COL4A5), tv 1, 2, 3
NM_004331	83	BCL2/adenovirus E1B 19 kDa interacting protein 3-like (BNIP3L) gi|47078259|

examine only the results of all of the underlying processes that lead to diverged copies of the original TE sequences.

The same set of probes was matched to a library of 16,655 genes known to produce proteins that have been studied (see *Methods*). For convenience, these are called known gene product genes. Table 2 lists the number of these genes as a function of the percent of their length matched by the probes (coverage) as well as the ratio between the number observed for the two libraries. The ratio of the sizes of the two libraries is 0.68, i.e., 16,655/25,193, whereas the ratio for 91–100% coverage is 0.51. These results suggest that many apparent genes in the library of 25,193 that show >90% coverage may be translatable or translated TE sequences in the genome rather than genes that produce useful proteins. A current search of their Web site (www.ncbi.nlm.nih.gov) shows that the National Center for Biotechnology Information has indeed removed some of these genes.

Table 2 shows the number of genes matched as a function of percent coverage in 10% intervals. The search of the known gene product gene library shows that >100 genes that produce studied proteins are matched by TEs over >80% of their length. Fifteen examples of apparent genes that are recognized as TEs have been removed from this list. Table 3 lists the genes that have been recognized as well as their functions. A good variety of different functional genes are present in Table 3, and many of them have multiple regional duplications, for example, collagen and elastin. The best explanation is that regions of this set of genes were originally derived in part from TE. For these examples, recognizable traces of their origins remain after a history of many duplications of both the whole gene and regions of it. In general, the amino acid sequence relationships are weak but significant. The average relationships show larger divergence than observed for the six genes previously identified that showed relatively precise relationships over most of their length (5). Because of the length of Table 3, the cases of genes with multiple transcript variants (tv) have been reduced to a single example. Thirty-one such cases, mostly collagen, have been removed, and the numbers of all of the tvs are listed for each example. In addition to these 96 examples, there are 241 examples with a coverage between 50% and 80% of the genes from the known gene product gene library that have significant similarity to TEs.

Relatives Among the Amino Acid Probes. In the comparison of the 5,010 probes with the human protein library, 17,190 total matches were reported by BLASTP. In most cases, these overlap with other matches and are of little individual consequence. On examination of the data, it was also observed that many of the genes matched a number of probes; this also is of little consequence but must be mentioned. The principal reason is that the probes are related to each other; for example, many are members of classes of repeats that are related to each other. Therefore, all

Table 3. (continued)

Ref Seq ID	Percent coverage	Description
	84	Collagen, type XIII, α 1 (COL13A1), tvs: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 19
NM_003609	80	HIRA interacting protein 3 (HIRIP3) gi|21396499|
NM_000427	93	Loricrin (LOR) gi|4557724|
NM_006501	84	Myelin-associated oligodendrocyte basic protein (MOBP), tv 2
NM_005416	89	Small proline-rich protein 3 (SPRR3) gi|4885606|
NM_152291	92	Mucin 7, salivary (MUC7) gi|22748664|
NM_033178	94	Double homeobox, 4 (DUX4) gi|15042962|
NM_000092	86	Collagen, type IV, α 4 (COL4A4) gi|62952498|
NM_014654	81	Syndecan 3 (N-syndecan) (SDC3) gi|57222246|
NM_000356	82	Treacher Collins–Franceschetti syndrome 1 (TCOF1), tv 2
NM_017804	99	Decreased expression in renal and prostate (DERPC), tv 1
NM_052897	91	Methyl-CpG-binding domain protein 6 (MBD6) gi|46852160|
NM_002723	89	Proline-rich protein BstNI subfamily 4 (PRB4) gi|41349489|
NM_005039	96	Proline-rich protein BstNI subfamily 1 (PRB1), tv 1
NM_153448	80	Extraembryonic, spermatogenesis, homeobox 1-like (ESX1L) gi|38455418|
NM_003387	99	Wiskott-Aldrich syndrome protein interacting protein (WASPIP) gi|38373694|
NM_005480	84	Trophinin associated protein (tastin) (TROAP) gi|33438581|
NM_152657	90	Gametogenetin (GGN), tv 1 gi|33286435|
	80	Collagen, type VIII, α 1 (COL8A1), tv 1, 2
NM_014805	99	EPM2A (laforin) interacting protein 1 (EPM2AIP1) gi|31982934|
NM_003472	85	DEK oncogene (DNA binding) (DEK) gi|31542502|
NM_173465	84	Collagen, type XXIII, α 1 (COL23A1) gi|29725623|
NM_003772	99	Jerky homolog-like (mouse) (JRKL) gi|22547223|
NM_005707	80	Programmed cell death 7 (PDCD7) gi|22027540|
NM_130761	87	Mucosal vascular addressin cell adhesion molecule 1 (MADCAM1), tv 1
	83	HLA-B associated transcript 2 (BAT2), tv 1, 2
	83	Collagen-like tail subunit of acetylcholinesterase (COLQ), tv 7, 8
NM_001846	87	Collagen, type IV, α 2 (COL4A2) gi|17986276|
NM_001853	94	Collagen, type IX, α 3 (COL9A3) gi|17921994|
	85	Collagen, type IV, α 6 (COL4A6), tv A, B
NM_022452	91	Fibrosin 1 (FBS1) gi|11967986|
NM_018942	87	Homeobox (H6 family) 1 (HMX1) gi|9506784|
NM_016423	84	Zinc finger protein 219 (ZNF219) gi|7705974|
NM_003125	82	Small proline-rich protein 1B (cornifin) (SPRR1B) gi|4507186|
NM_133264	99	WIRE protein (WIRE) gi|62414030|

847 six-frame translation amino acid sequence probes that find matches to human genes were compared with each other by using BLASTP. On average, these probes are significantly related to ≈30 other probes. The maximum number of relatives of any one probe is 84. Only 48 of these probes have no recognized relatives.

Cases in Which All Six Reading Frames Match Human Proteins. In some cases, all six reading frames matched human protein amino acid sequences with an expectation of $\leq 10^{-3}$. These are listed in Table 4, along with the number of times relationships are observed, regardless of frame, in the whole library. Many are translations of Alu repeated sequences, and many classes of Alus are included. Of course, there is no known functional Alu translation. These translations simply are similar in sequence to the various ways an Alu sequence can be incorporated into a gene-coding sequence and modified (by necessity for its survival) to achieve translatability. Many of the named classes of Alu sequences are very similar to each other, and the matching regions overlap severely. The largest number of TEs in this list are those found frequently in human proteins; thus, there has been an opportunity for matches to probes in all six reading frames.

Precision of Matches. The precision of the matches is of some value in attempting to understand the history of the sequence similarities to TEs. The interest is in the best matches of all the many matches that identify each of the 1,950 genes as including a TE sequence. Fig. 2 shows the distribution of the number of these

Table 4. TE probes for which all six frames make protein matches

Name	Matches
L1	45
TAR1	564
Alus	1,873
LTR12C	115
LTR12D	114
LTR12E	241
MER109	117
MER115	57
MER45C	162
MER52A	293
MER52C	97
MER52D	146
REP522	742
L1M2_5end	575
L1M3e_5end	3,589
L1PA13_5end	178

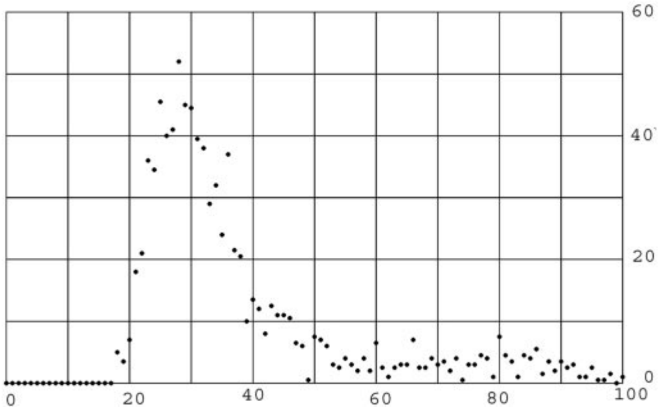

Fig. 2. The number of matches vs. maximum percent match. Vertically is the number of matches in each interval of 1%.

best matches vs. the percent match. The mode is at 28% match, and there are examples all the way up to one at 100%.

Random Sequence Comparison. The set of 5,010 probes (six-frame translations of potential TEs) were replaced by random amino acid sequences of the same set of lengths. The average amino acid composition was matched to that of the whole library of 25,193 genes. BLASTP was used to compare this random set of 5,010 sequences with the 25,193 human genes, with a requirement that the expectation be $\leq 10^{-3}$. This process was done repeatedly, and the number of matches found varied from two to six, with an average of four. Thus, almost all 1,950 examples of TE insertions in coding regions are likely to be statistically significant. The criterion that the expectation was limited to $\leq 10^{-3}$ was chosen for the first analysis to make sure the relationships were statistically significant and could be counted on as actual relationships between TEs and protein sequences.

Lower Criterion. The criterion of expectation of 10^{-3} is actually quite restrictive. For larger expectations, many more relationships are seen, rising to large values; however, at the same time, the number of background relationships rises so that it becomes impossible to tell which relationships are actual and which may just be noise. The measure chosen for background (i.e., not significant relationships) is the number of matches observed when the set of probes were replaced by random sequences of matching lengths and appropriate composition. Table 5 shows the actual effects of decreasing the lower limit of the BLASTP quoted score. Because the two measures of statistical expectation are inverse to each other, this is roughly equivalent to increasing the upper limit for the expectation, but the score is more convenient.

The third column of Table 5 lists the results of a search for matches with BLASTP using the 5,010 probes described earlier. Listed are the number of different proteins that included matches. The search was done at a very open criterion, and the highest-scoring match was selected for each individual protein. Matches were counted in each category of score limit quoted by the program. The fourth column lists the results where each probe was replaced by a random sequence of the same length and with composition equal on average to that of the library of 25,193 proteins. Table 5 shows a large increase in the number of matches for the lower score limits, whereas the background from random sequences also rises. The second column shows the difference between the number of proteins matched with the six-frame probes and the random probes and is labeled net as an estimate of the number of actual relationships between the TEs and the human proteins. The search with random sequence replacements for the 5,010 TE amino acid sequences was repeated 11 times, and the maximum net number of matches ranged from 4,491 to 6,153, with an average of 4,653 net matches. This large number of matches is apparently due to the presence of many more distant relationships to the TEs among the human proteins. Such a situation is expected if the process of TE insertions into proteins has a long history of insertions and duplications followed by drift of the sequences. This result leaves no doubt that the 1,950 identified in the first search with an expectation limit of 10^{-3} is an underestimate of the total number of distinct human proteins that include sequences related to TEs. The maximum net estimate in Table 5 is more than twice as large (after correction for background) at open criterion compared with an expectation of $\leq 10^{-3}$.

Table 5. Lower criterion comparison with random sequence probes

	Matches		
Score limit	Net	Six frame	Random
30	4,098	12,010	7,912
31	4,333	10,252	5,919
32	4,644	8,776	4,132
33	4,779	7,007	2,228
34	4,524	5,670	1,146
35	4,277	5,018	741
36	3,827	4,193	366
37	3,471	3,707	236
38	3,060	3,169	109
39	2,643	2,693	50
40	2,411	2,440	29
41	2,121	2,134	13
42	1,926	1,932	6
43	1,690	1,691	1
44	1,500	1,500	0

Number of different proteins that match with scores less than or equal to those in column 1.

Discussion

The process of choosing a sensitivity in measurements where background noise is detected and corrected for by subtraction is common in many fields but not in protein sequence comparison; I am not aware of previous publications. Nevertheless, it appears to be a sound procedure. The data of Table 5 suggest that there may be many TE sequences in human proteins that are the results of ancient events, and the sequences have drifted, making their recognition difficult. If this is so, the original estimate of 1,950 TEs in human proteins is a lower limit, because it was made with a rigid criterion. The much larger estimates of Table 5 are an improvement, but no upper limit of the number of TEs in human proteins is available. It seems likely that the processes responsible (TE insertion followed by gene duplication and regional copying) go back to the origin of eukaryotes or even earlier. Thus, the majority of such sequence relationships have diverged to such an extent that they are now unrecognizable.

Fig. 2 contributes to this argument, because it shows that there are examples with a precision of match over the range from the mode at 28% to 100%. The best explanation is that the most precise were recent events of inclusion of TEs in coding regions, and the others extend back in time to many ancient events. If this concept of a long history of such events is correct, which seems likely, then of course the more sensitive technique of the open criterion (Table 5) reaches back only part of the way. Both the included TEs in the coding regions and the TE examples used as probes have been subject to drift in their sequences, limiting search techniques based on amino acid sequence similarity.

Thus, the many thousands estimated are just the more recent examples, and the total could be very large. The inclusion of TEs in coding regions and subsequent multiplication of the gene regions could well be a nearly universal process. Alternative possibilities are that gene coding sequences have become part of TEs, or that many copies of genes have been made that have become nonfunctional and are classified as repeated DNA sequence in the human genome.

Previous studies (6) show that >80% of genes in the library of 25,193 human genes have regions of significant amino acid sequence relationships to other human genes, often very many others. In most cases, there are regions within the genes that show many relationships, and other regions that show few or none. There is a great variety of patterns resulting from many past events of multiplication and rearrangement. Previous studies (6) and the data reported in this paper give insights into the existence of long-term processes that lead to extensive sets of sequence relationships between genes and TEs and among the genes themselves. These insights are the result of the detection of very distant but significant relationships. The functional value of these processes during evolution presumably has been gene variation and the creation of novel genes and gene functions.

Methods

The set of 25,193 coding sequences was obtained by making use of the seq_gene.md file from the National Center for Biotechnology Information, build 34. RM, version 2002/05/05, was obtained from the Institute for Systems Biology, Seattle. A 24 processor Sun Microsystems (Mountain View, CA) E6500 computer was used for these studies. To prepare the file of "known protein" genes that produce the proteins studied, a list of the 25,193 genes with brief identifiers was alphabetized, and blocks were removed, for example, those identified as hypothetical or similar to other genes, with 16,655 remaining. BLAST (8) was used for DNA sequence comparisons and BLASTP (9) for amino acid sequence comparisons. The expectation reported by these programs is the result of a statistical calculation of the likelihood of a chance match with the length and precision reported for each find. As a check, the probes used in each BLASTP search were replaced by random amino acid sequences of the same lengths and of average composition equal to the whole protein library, and only an average of four matches were found with an expectation of $\leq 10^{-3}$.

John Williams carried out much of the data processing and wrote the necessary PERL programs, George Miklos made valuable suggestions, and Eric Davidson's laboratory supplied support.

1. Brownell, E., Mittereder, N. & Rice, N. R. (1989) *Oncogene* **4,** 935–942.
2. Smit, A. F. (1999) *Genet. Dev*. **9,** 657–663.
3. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al* (2001) *Nature* **409,** 860–921.
4. Nekrutenko, A. & Li, W. H. (2001) *Trends Genet.* **17,** 619–621.
5. Britten, R. J. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 16825–16830.
6. Britten, R. J. (2005) *Proc. Natl. Acad. Sci. USA* **102,** 5466–5470.
7. Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J. (2005) *Cytogenet. Genome Res.* **110,** 462–467.
8. Altschul, S. F., Gish, W., Miller. W., Myers, E. W, Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
9. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.