Coding sequences of functioning human genes derived entirely from mobile element sequences

Roy J. Britten*

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

Contributed by Roy J. Britten, September 20, 2004

Among all of the many examples of mobile elements or ''parasitic sequences'' that affect the function of the human genome, this paper describes several examples of functioning genes whose sequences have been almost completely derived from mobile elements. There are many examples where the synthetic coding sequences of observed mRNA sequences are made up of mobile element sequences, to an extent of 80% or more of the length of the coding sequences. In the examples described here, the genes have named functions, and some of these functions have been studied. It appears that each of the functioning genes was originally formed from mobile elements and that in some process of molecular evolution a coding sequence was derived that could be translated into a protein that is of some importance to human biology. In one case (AD7C), the coding sequence is 99% made up of a cluster of *Alu* sequences. In another example, the gene BNIP3 coding sequence is 97% made up of sequences from an apparent human endogenous retrovirus. The Syncytin gene coding sequence appears to be made from an endogenous retrovirus envelope gene.

Mobile elements form the majority of the human genome, but that is unimportant compared to all of the functional effects these ''parasites'' have had on our evolution. Insertions have influenced the regulation of transcription of some genes and the termination of transcription. Hundreds of examples have been recognized where individual exons have sequences that are similar or identical to fragments of mobile element (ME) sequences (1, 2). In many of these cases a single exon is involved, and its transcription yields a variant mRNA (3). The suggestion is that MEs are a source of variation as a result of the insertion of fragments of sequence into functioning genes. Here, I am using MEs (*sensu lento*) to represent any repeated sequence present in many copies in the genome. Smit (4) has made a list of 19 examples of human genes ''probably derived from transposable elements.''

Reported here are cases where almost the entire coding sequences (>89%) of functioning human genes are apparently derived from ME sequences. There are several examples of genes with named functions in which all or nearly all of coding sequences are quite similar to ME sequences as recognized by REPEATMASKER (www.repeatmasker.org). There are many other examples of observed mRNAs for which the coding sequences are defined by computer programs, and these sequences are identified by REPEATMASKER as MEs. However, in this subset of cases it is claimed that a functioning gene was derived entirely from ME sequences. There may be additional cases among a list of 49 unstudied examples derived by screening mRNA libraries to be described below.

These observations contribute an additional bit to the growing mass of evidence that indicates that mobile elements/repeats are not always junk and have made important contributions to the ''host'' (5–9). The MEs and DNA sequences derived from them have been a part of the eukaryotic ''genomic environment'' for a very long time. Thus, it is expected that they will have had important effects on gene function because it can be considered that living systems will sooner or later make use of whatever is available if it is at all possible, particularly in the genome. There have been theoretical proposals (10) of the evolutionary role of variety and change in these relationships, particularly in the control of gene expression. There is direct evidence for the evolutionarily significant role of mobile elements/repeats (11–13) and evidence for strong associations and functions including the regulation of transcription. The cases described in this paper add to this earlier evidence in that, in these cases, nearly the entire coding sequences of genes have apparently been derived from ME sequences.

A survey is in process to determine the fraction of the coding sequences recognized at present in available genomes that are derived from ME sequences. The early results turned up the AD7C or neural thread protein gene, which sparked interest because it is apparently derived entirely from a cluster of *Alu* repeated sequences. The investigators pointed out that the coding sequence contained regions of sequence similarity to four *Alu* sequences (14). Table 1 describes this and several other cases.

EVOLUTION

Methods

A collection of coding sequences was made from the NCBI file seq_gene.md. These were examined by REPEATMASKER, and those that were reported to be almost completely similar in sequence to mobile elements were set aside for further study. The examples examined in the first part of this paper were selected from this list on the basis of their known function. Some of the remainder of them are shown in Table 4.

Results and Discussion

AD7C. AD7C is a neuronal thread protein gene. It encodes a 41-kDa membrane spanning phosphoprotein that is useful in the diagnosis of early Alzheimer's disease (14, 15). The coding sequence is 1,128 nt long and REPEATMASKER shows that it consists of fragments of five (or four, see below) *Alu* sequences. All of the matches are with the reverse complements of the *Alu* repeats. The alignment is summarized in Table 2. Listed are the percent similarity and length of each of the regions from the best matching *Alu* sequences, which differ inconsequentially from those published in ref. 14.

First, an *Alu*Sp matches at 92% accuracy the first 281 nt of the coding sequence. After a gap of 3 nt, 141 nt of *Alu*Jo matches at 87% precision. Then, after 2 nt, an additional part of the *Alu*Jo sequence matches to 93% for 167 nt including a sizeable part of the poly(A) tail, modified by two substitutions that affect the translation. These two short fragments seem to represent one *Alu* sequence homolog in the coding sequence, but rearrangement has apparently occurred because there are overlapping regions of the *Alu*Jo. Next is a 92% match for 302 nt to an *Alu*Sc, including a sizeable part of the poly(A) tail that is modified. Finally, there is an 88% match for 239 nt to an *Alu*Sx, also including a sizeable region of the poly(A) tail that is modified. In the genome, this match continues after the end of the coding

Abbreviations: cds, coding sequence; ME, mobile element.

*To whom correspondence should be addressed. E-mail: rbritten@caltech.edu.

Table 1. Selected genes derived from ME sequences

Chr.	Name	cds, nt	% ME	% match	ME identifier	Accession no.
1?	AD7c	1,128	99.6	83–92	5 *Alu* segments	NM_014486
7	SYNCYTIN	1,615	100	97	HERV-W	AF072506
(7	GTF2IRD2*	1,607	97.7	80–88	Charlie8, DNA/MER1	NM_001003795)
8	HHCM	1,404	89.9	68–71	L1MD2, LINE/L1	NM_006543
10	BNIP3	585	97.1	84	HERV70, LTR/ERV1	NM_004052
13	LG30	216	100	74–76	MLT1E, MLT1G, LTR	AY138548

The first column is the chromosome (Chr.) number, which is not certain for AD7C.
*Exon 16 only, therefore in parentheses.

sequence region and there is another match to an *Alu* sequence (data not shown).

It appears that the whole gene coding region has been made from a cluster of *Alu* sequences. The gaps of a few nucleotides between the individual *Alu* sequence matches are probably just details of the REPEATMASKER alignment process and can be ignored. A matter of interest is how much change has occurred in the sequences to form a useful gene from the ME sequences. The *Alu* sequences summarized in Table 2 are simply the best matches from the REPEATMASKER collection and are not necessarily the *Alu* sequences that were present in the original *Alu* cluster, so that it is not possible in general to identify the sequence changes that have occurred. A sample can be estimated by examining the three poly(A) chains that are included. They total to 60 Ts in the complementary *Alu* sequences. In these poly(T) regions, eight changes have occurred, all leading to translatable codons for amino acids other than phenylalanine. They consist of six A substitutions and two insertions of two As each. This ≈17% change in this small sample suggests positive selection. Of course, there is only one possible silent substitution in a row of Ts, the transition from T to C in the third base. In addition, there are four cases of internal T-rich sequences in the five *Alu* sequences involved, and in one of those, such a silent substitution has occurred. In two of these cases, length differences have occurred resulting from a six-base deletion and a four-base insertion, leading, of course, to translatable codons. This is a clear case in which a cluster of *Alu* repeats has been converted into an active human gene. We do not yet know how the 5′ control region is organized. With that information we will someday be able to say more about the evolutionary process that created the gene. It was pointed out that an identifiable full-length representation in the human genome (build 34) is only 97% similar to the AD7C mRNA sequence (A. F. Smit, personal communication) (14). The differences are such that the genomic sequence is not translatable for a significant length. No better genomic copy of the mRNA has been found, but the gene could contain introns and might be hard to identify because of the *Alu* sequences.

Table 2. Alignment summary of AD7C

				Position in ME§	
%*	Start	End†	ME‡	End	Start
92	1	281	*Alu*Sp#SINE/*Alu*	280	1
87	284	411	*Alu*Jo#SINE/*Alu*	143	2
83	413	580	*Alu*Jo#SINE/*Alu*	301	134
92	581	884	*Alu*Sc#SINE/*Alu*	302	1
88	887	1128	*Alu*Sx#SINE/*Alu*	300	61

*Match between ME sequence and region of cds.
†Start and end positions in cds.
‡REPEATMASKER description of ME.
§End and start positions in reverse-oriented ME.

BNIP3. BNIP3 is the gene for a protein involved in controlling apoptosis through the interaction with other proteins (16–18). The heading for the entry in OMIM (Online Mendelian Inheritance in Man) is BCL2/ADENOVIRUS E1B 19KD PROTEIN-INTERACTING PROTEIN 3: BNIP3. Table 1 shows that 97% of the coding sequence is related closely to that of HERV70RM. HERV70RM is the name I am using for the version of HERV70 that is included in the REPEATMASKER library and it is named a human endogenous retrovirus, although it does not contain recognizable retroviral gene residues. It is more than 7 kb long, and the relationships to the BNIP3 coding sequence occurs after nucleotide 4641 of HERV70RM. The coding sequence of the BNIP3 mRNA aligns fully with the HERV70RM sequence even though the gene consists of 6 exons spread over almost 15 kb of DNA. To help resolve this relationship, REPEATMASKER was run against the whole gene, and the results are shown in Table 3. Most of these data are from REPEATMASKER output, and two columns are added to show the location of the exons in the gene. In most cases, the identification of an HERV70RM segment in the gene aligns closely with the exons. This agreement is so good that the history seems obvious. Likely, a part of the HERV70RM from about 4–7 kb was converted to a gene without introns, which must have evolved and become useful, and later the introns were inserted into it to lead to the modern BNIP3 gene. In fact, there is a BNIP3P sequence on chromosome 14 that is identified as a pseudogene because it lacks introns and gives a very good match in a search made with the BNIP3 mRNA by using BLAST the human genome. It is possibly a fossil of the early stage in this event or it may be an actual pseudogene made from the mRNA at a later stage.

To further explore this interpretation, the coding sequence was aligned with the HERV70RM sequence by using BLAST2 sequences. The result showed two copies of the almost complete cds region at locations 5507–6073 and 6732–7289 in the HERV70RM sequence, matching ≈80%. Thus, the locations shown in Table 3 in HERV70RM are simply the best fits of REPEATMASKER and do not necessarily show the actual sequence origins of the BNIP3 coding sequence. It seems likely that it originated as a copy of one of the regions in HERV70RM. Table 3 shows one example of a sequence similarity between HERV70RM and a region of the gene that is not an exon in BNIP3. The history of this region is unclear. In any case, it is clear that most of the exons of the BNIP3 gene derived from a continuous stretch of HERV70RM. This seems to be a good case of "introns late" because there is no other explanation that comes to mind for the presence of a series of connected pieces of HERV70RM spread widely in the BNIP3 gene.

An important issue is the nature of HERV70RM. The copy used in these studies is listed in the library of human repeated sequences listed in REPEATMASKER. It is incomplete and not a classical endogenous retrovirus. The HERVD database (http://herv.img.cas.cz) lists many regions in the human genome that are similar in sequence to what I call HERV70RM here, although none of them match a length of more than ≈1 kb. In fact, there

Table 3. MEs in the BNIP3 gene

Divergence			Distance from start of gene					ME identification		Location in ME	
%	Del	Ins	Exon		Start	End				Start	End
17.6	8.0	3.2	824	869	1	875	+	HERV70	LTR/ERV1	4641	5557
26.1	0.0	4.2			1241	1288	C	L2	LINE/L2	(86)	3227
28.3	16.3	0.0			1648	1739	C	MER5A	DNA/ MER1_type	(48)	141
9.0	4.1	0.0			2208	2473	+	*Alu*Sq	SINE/Alu	1	277
23.7	12.6	2.1			2753	2847	+	L1ME3A	LINE/L1	6021	6125
18.0	0.0	0.0	2938	3087	2937	3086	+	HERV70	LTR/ERV1	6776	6925
16.2	0.0	3.7	3164	3270	3169	3277	+	HERV70	LTR/ERV1	6933	7037
15.8	11.4	0.0			4574	4687	+	FLAM_C	SINE/Alu	1	127
13.8	0.0	0.0	5334	5418	5335	5421	+	HERV70	LTR/ERV1	7032	7118
13.6	1.3	0.0	6093	6243	6094	6247	+	HERV70	LTR/ERV1	5901	6056
19.3	2.8	0.0			6691	6980	C	*Alu*Jo	SINE/Alu	(14)	298
32.0	8.0	0.0			6997	7146	C	L1ME	LINE/L1	(734)	5436
7.0	1.1	1.1	None		7147	7233	+	HERV70	LTR/ERV1	7172	7258
27.5	5.6	1.4			7241	7384	C	L1ME	LINE/L1	(873)	5273
23.0	18.6	2.3			8613	8870	C	MER21C	LTR/ERV1	(88)	847
17.9	0.0	11.8			8909	8984	+	(CCCCAA) n	Simple_repeat	2	68
16.7	0.0	1.6			9224	9284	+	MER41B	LTR/ERV1	481	540
8.3	1.4	0.7			9297	9586	+	AluSq	SINE/Alu	6	297
24.1	3.7	3.7			9594	9675	C	MER21C	LTR/ERV1	(853)	82
23.7	17.2	0.0			9747	10036	C	MLT1A0	LTR/MaLR	(17)	348
34.9	1.8	3.6			11487	11596	+	MIR3	SINE/MIR	101	208
21.7	4.7	3.5			11902	11987	+	FRAM	SINE/Alu	75	161
4.7	3.0	3.0			12762	12892	+	*Alu*Jo/FLAM	SINE/Alu	1	131
			14061	14106			+	HERV70	LTR/ERV1	6053	6100

Del, deletion; Ins, insertion.

is a set of 63 sequences in this database that match the BNIP3 cds, although most of them show only a short matching region. The situation needs clarification because there are many entries in the HERVD database called HERV70 that show no sequence similarity to HERV70RM. There is no full-length copy of HERV70RM in the present version of the human genome, so its status as a human endogenous retrovirus sequence is doubtful. BLAST of the human genome (filter off) searching with HERV70RM finds many hits and graphs some examples as if they were full-length matches. They do not exist, and the program has assembled them from groups of nearby fragmentary matches.

When REPEATMASKER is run against HERV70RM, two small fragments of *Alu* sequences are found, as well as other MEs within it. There are regions that REPEATMASKER identifies as HERV70 (HERV70RM), and these include the region of the copies of the BNIP3 coding sequences. A warning is required here because BLAST of the human genome (filter off, default) finds only 3 matching sequences for the BNIP3 coding sequence of the 63 that exist in the HERVD database. I confirm the fact that there are many matching fragments to the coding sequence (cds), finding 120 in the human genome by using BLAST. This is an important point because these data, regardless of the interpretation of HERV70RM, show that the BNIP3 gene cds sequence is closely related *in toto* to sequences of a ME. We may not know exactly what this ME is, but there are many copies of this region of it in the human genome ranging from precise to quite divergent.

The BNIP3 gene occurs in the mouse genome [NM_009760], and the coding sequence matches the human with 89% accuracy. The protein sequences match to 90% accuracy except for a 5-aa gap and a 1-aa gap in the mouse protein. The gene arrangement is similar, with 6 exons extending over ≈15 kb. The exons are identical in length to the human exons except for the gaps of 15 and 3 nt corresponding to the protein differences. Because the cds match so closely in sequence, the mouse BNIP3 exons show the same relationship to the human HERV70RM as do the human BNIP3 exons. Interestingly, there is no sequence in the mouse genome, seen by BLAST of the mouse genome, that matches the human HERV70RM except for the BNIP3 exons. There is apparently no equivalent ERV in mouse genome, although, of course, many other HERVs and MERVs share sequence. REPEATMASKER may be used with either the human repeats or mouse repeats to examine the mouse BNIP3 gene region. With the human repeats, the mouse BNIP3 exons are recognized as HERV70RM sequences, but with the mouse repeats, no sequences match. The exons in the two genes are nearly identical. The nucleotide sequences of the mouse and human BNIP3 cds match closely (90%). K_s between the coding sequences of mouse and human are 0.41 and $K_a = 0.047$ (K_s is the divergence due to synonymous substitutions, and K_a is the divergence due to changes that cause amino acid replacement) (19). This similarity suggests that whatever the events were, they occurred far in the past.

The BNIP3 gene has also been sequenced from rat, and the cds is 95% similar to that of mouse BNIP3, so the same arguments apply. The K_s between the coding sequences of the rat and human is 0.37 and $K_a = 0.048$ (20). BLAST of the rat genome finds a BNIP3 exon and two other rat sequences similar to parts of human HERV70RM, whereas BLAST of the mouse genome finds only a BNIP3 exon with similarity to human HERV70RM. Based on a BLAST search of GenBank, chicken (*Gallus gallus*) has a similar mRNA sequence to the human BNIP3. There is a match of 367 of 453 nt, or 81%, in one large region and evidence of other smaller regions of similarity. It seems that a full examination of the evolution and relationships of BNIP3 and HERV70RM would be worthwhile in a number of species.

Table 4. Observed transcripts that match ME for >80% of their length

Chr.	% length*	Length of cds	ID	ID	% match†	Listing‡	Record§
1	98.25	171	NM_016646	LOC51336	82.7	L1M3	
1	96.00	150	XM_352936	PRO2012	78.1	L1ME	Record removed
2	100.00	384	NM_175853	LOC150759	78.7	L1P	
2	100.00	375	XM_208704	LOC283517	97.9	SVA	Record removed
2	100.00	723	XM_351431	LOC375197	95.8	L1PA4	Record removed
2	95.83	288	XM_173068	LOC253584	85.4	THE1C MLT1B	
2	92.33	300	XM_351509	LOC375299	90.0	L1P Tigger2	Record removed
2	86.38	279	XM_291017	LOC339793	89.8	L1P MLT1E2	
3	93.44	183	NM_018629	PRO2533	76.3	MLT1G3	Record removed
3	91.54	402	XM_353342	LOC375388	79.9	MLT2B4	Record removed
3	83.66	153	NM_014135	PRO0641	78.1	MLT1H	Record removed
4	99.95	1974	XM_209656	LOC285550	89.8	Charlie9	
4	97.40	1692	NM_024534	FLJ12684	72.6	MER34-int	
5	87.04	486	NM_173668	FLJ34836	89.1	LTR12C BaEV-int	MER50
5	86.45	369	XM_353366	LOC375433	81.2	L1MC/D LTR5B	Record removed
6	99.74	387	XM_291181	LOC340211	97.4	L1P	
6	99.60	249	NM_018572	PRO1051	87.9	L1P	Record removed
6	84.97	366	NM_178534	FLJ37940	85.7	HERVL18	Record removed
7	97.72	1491	NM_032203	GTF2IRD2	82.4	Charlie8	
7	94.02	1722	NM_145111	DKFZp727G1	86.4	Charlie9	
8	97.99	348	XM_351783	LOC375668	86.8	Tigger3 (Golem)	FLAM_C removed
8	85.35	273	XM_353456	LOC375664	70.3	L2	Record removed
9	100.00	375	XM_209180	LOC284397	97.1	SVA	
9	100.00	342	XM_351803	LOC375700	87.4	L1P4	Record removed
9	100.00	363	XM_353472	LOC375692	82.8	LTR1B	Record removed
9	92.10	291	XM_353479	LOC375732	73.1	SST1	Record removed
9	92.10	291	XM_353476	LOC375716	72.8	SST1	Record removed
9	91.75	291	XM_353477	LOC375726	72.6	SST1	Record removed
9	90.14	426	NM_030898	FLJ21673	84.1	*Alu*Sg/x L2	FLAM_A record removed
9	90.11	354	XM_353493	LOC375772	93.1	MLT2A1	Record removed
9	88.52	270	XM_353481	LOC375740	77.4	REP522	Record removed
9	88.52	270	XM_353480	LOC375738	77.8	REP522	Record removed
9	88.52	270	XM_353478	LOC375727	77.4	REP522	Record removed
10	99.67	600	NM_178512	FLJ37201	73.2	Tigger4 (Zombi)	
10	99.57	231	XM_352893	LOC374280	90.4	MER11B	Record removed
10	98.10	105	NM_173577	MGC45541	86.4	*Alu*Jo/FRAM	
1	85.21	2082	NM_021211	LOC58486	68.3	Charlie1	
12	85.48	303	XM_350891	LOC374483	93.8	HERVK22	Record removed
13	99.31	288	NM_138474	LOC144845	80.1	L1PA13	Record removed
13	97.80	501	NM_173604	FLJ25694	87.4	HERVE	
13	89.24	381	NM_153251	FLJ25952	85.5	*Alu*Sg/x AluSx	
13	80.13	297	XM_353050	LOC374511	77.0	MSTC MIR	Record removed
16	100.00	75	NM_030970	MGC3771	90.7	*Alu*Sp	
19	100.00	144	NM_178523	MGC45556	91.0	L1PA10	Record removed
19	88.17	372	XM_294914	LOC339358	81.2	MER41-int	MER41B MER77
19	84.00	225	NM_138781	LOC113386	77.8	HERVK3	
20	100.00	375	XM_209370	LOC284806	89.1	SVA	Record removed
21	88.89	378	XM_211658	LOC284837	75.3	L1MB8	Record removed
X	100.00	372	NM_153016	FLJ30672	87.8	THE1-int	

Chr., chromosome.
*Percent of length of cds that is ME.
†Percent sequence match of cds to ME.
‡REPEATMASKER listing.
§Added on March 2, 2004, when removals were found.

Syncytin. This example is listed by Smit (4) and is included here because recent evidence shows that Syncytin is a functioning gene in human placenta (21, 22). The mRNA is derived *in toto* from the endogenous retrovirus HERV-W, which is present in many copies in the human genome. The authors (21) identify ERVWE1 as the gene region that is the source of the transcript, although this may not be certain. ERVWE1 is 10.2 kb long and consists of the usual LTR–gag–pol–env–LTR arrangement. The Syncytin mRNA is 2.8 kb long and consists of the 5′ LTR, some additional sequence, the env gene, and the 3′ LTR. The cds of 1,617 nt includes just the env gene of the endogenous retrovirus. Within it, regions can be identified that are functionally significant to Syncytin. It is not clear how much evolutionary change occurred in the env gene to assume its present function. Entrez Gene lists what are termed GeneRIFs (www.ncbi.nlm.nih.gov/projects/GeneRIF/GeneRIFhelp.html):

1. Env HERV-W glycoprotein mediates cell–cell fusion upon interaction with the type D mammalian retrovirus receptor. Env protein was detected in the placental syncytiotrophoblast, suggesting a physiological role during pregnancy and placenta formation.
2. Contributor to normal placental architecture, especially in the fusion processes of cytotrophoblasts to syncytiotrophoblasts. The gene expression of Syncytin may be altered in cases with placental dysfunction such as preeclampsia or HELLP syndrome.
3. mRNA abundance for Syncytin showed stimulation by forskolin in BeWo cells.
4. Syncytin-mediated trophoblastic fusion in human cells is regulated by GCMa.
5. Syncytin gene activation is highest in term placenta.
6. HERV-W Env glycoprotein is directly involved in the differentiation of primary cultures of human villous cytotrophoblasts.
7. Hypoxia alters expression and function of Syncytin and its receptor during trophoblast cell fusion of human placental BeWo cells: Implications for impaired trophoblast syncytialization in preeclampsia.
8. Syncytin gene expression is down-regulated by hypoxia, which strengthens the hypothesis that Syncytin is reduced in disturbed pregnancies in the course of placental hypoxia.

HHCM. HHCM is identified as a human hepatocellular carcinoma 3.0-kb DNA sequence that encodes (in a 1,404-nt cds) a 52-kDa protein. It transforms both rat liver cells and NIH 3T3 fibroblasts.[†] Table 1 shows that it is almost 90% made up of L1 MEs. The sequence match is only ≈70%, so much sequence change has occurred because its origin from a part of the L1 sequence. It matches the regions 18–331 nt and 437-1470 nt of L1MD2. This is not apparently a beneficial contribution that L1 has made to our genome, although MEs act in strange ways. The record NM_006543 was ''temporarily removed by RefSeq staff for additional review'' and Smit (personal communication) did not find a closely matching genomic sequence. Thus, this example must be considered a candidate for future study.

LG30. LG30 is a gene of unknown function in the region G72/G30 of chromosome 13. Mutations in the region are connected to bipolar disorder (23, 24), but it appears that the G72 is more likely to be responsible (25). The LG30 coding region is only 216 nt long, and 100% of its length is related to LTR class ME (MLT1E, MLT1G).

GTF2IRD2. GTF2IRD2 was initially described as a transcription factor gene (26, 27), and the NCBI entry consisted of the fragment listed in Table 1. That is why it is included here. It has recently been studied in detail (28, 29), and it turns out that this fragment is actually exon 16, the 3′ exon and the only long exon, more than half the length of the whole coding sequence. This exon consists entirely of ME sequence Charlie8. What follows is a quotation from ref. 29. ''GTF2IRD2 is the third member of the novel TFII-I family of genes clustered on 7q11.23. The GTF2IRD2 protein contains two putative helix–loop–helix regions (I-repeats) and an unusual C-terminal CHARLIE8 transposon-like domain, thought to have arisen as a consequence of the random insertion of a transposable element generating a functional fusion gene. The retention of a number of conserved transposase-associated motifs within the protein suggests that the CHARLIE8-like region may still have some degree of transposase functionality that could influence the stability of the region in a mechanism similar to that proposed for Charcot–Marie–Tooth neuropathy type 1A. GTF2IRD2 is highly conserved in mammals and the mouse orthologue (Gtf2ird2) has also been isolated.''

Other Transcript Coding Sequences Apparently Derived from ME. Table 4 is a list of 49 examples of observed transcripts for which the coding sequences have been determined by computer programs, and these cds are made up from MEs at least to the extent of 80%. This collection was made by running REPEATMASKER against the NCBI collection of gene transcripts in February of 2004, but when checks were made in early March, all of the transcripts so marked had been removed from the collection. It seems likely that someone decided they were junk, which in a sense may be true, but from the point of view of this article they may be considered potentially useful and should be further examined. Some of them are likely to be examples of the transcription of fragments of ME, a process which occurs frequently. Regions of ME line 1 are expressed in mouse and rat and human RNA collections (unpublished data). Smit's table (4) has been extended (27) to include 47 potential genes derived at least in part from ME. However, the central issue for these two tables is whether these candidates are actually functioning genes. In fact, there is no evidence in the majority of cases that these mRNAs are produced by functioning genes. There are two examples in these tables where nearly the whole mRNA derives from an ME, and one of them is described above as Syncytin (21, 22). The other appears to be the transcription of a fragment of a sequence related fairly closely to HERV3, including the env gene and LTR, and the transcript is described as an env gene mRNA. The evidence of its function is transcription in placental trophoblast cells (28), reminiscent of intracysternal A-particles in mouse that are similar to ERVs and may be claimed to have an important role in placenta (29).

The cases described and possibly the example just mentioned (4, 27) show that parts of ME have been converted to form essentially complete gene coding sequences. There are probably more cases as indicated by Table 4. These observations add to the many known ways in which MEs have contributed to our evolution. This subject has been reviewed recently by Kazazian (30) who characterizes them as being in the driver's seat, rather than simply being useful to have around. Because of this review there is not reason for extensive discussion here.

I thank John Williams for assistance, Arian Smit and Mark Springer for criticism, and Eric H. Davidson's laboratory for support.

[†]Yang, S. S., Zhang, K., Vieira, W., Taub, J. V., Zeilstra-Ryalls, J. H. & Somerville, R. L., 14th International Symposium for Comparative Leukemia and Related Diseases, October 8–12, 1989, Vail, CO.

1. Nekrutenko, A. & Li, W. H. (2001) *Trends Genet.* **17,** 619–621.
2. Lorenc, A. & Makalowski, W. (2003) *Genetics* **118,** 183–191.
3. Sorek, R. R., Ast, G. & Graur, D. (2002) *Genome Res.* **12,** 1060–1067.
4. Smit, A. F. (1999) *Genet. Dev.* **9,** 657–663.
5. Brosius, J. (1999) *Gene* **228,** 115–134.
6. Makalowski, W. (2000) *Gene* **259,** 61–67.
7. Lagemaat, L., Landry, J. R., Mager, D. L. & Medstrand, P. (2003) *Trends Genet.* **19,** 530–536.
8. Jordan, K., Rogozin, I. B., Glazko, G. V. & Koonin, E. V. (2003) *Trends Genet.* **19,** 68–72.
9. Liu, G., Zhao, S., Bailey, J. A., Sahinalp, S. C., Alkan, C., Tuzun, E., Green, E, D. & Eichler, E. E. (2003) *Genome Res.* **13,** 358–368.
10. Britten, R. J. & Davidson, E. H. (1971) *Q. Rev. Biol.* **46,** 111–138.
11. Britten, R. J. (1996) *Mol. Phylogenet. Evol.* **5,** 13–17.
12. Britten, R. J. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 9374–9377.
13. Britten, R. J. (1997) *Gene* **205,** 177–182.
14. De la Monte, S. M. & Wands, J. R. (2002) *Front. Biosci.* **7,** 989–996.
15. De la Monte, S. M. & Wands, J. R. (2004) *J. Alzheimer's Dis.* **6,** 231–242.
16. Boyd, J. M., Malstrom, S., Subramanian, T., Venkatesh, L. K., Schaeper, U., Elangovan, B., D'Sa-Eipper, C. & Chinnadurai, G. (1994) *Cell* **79,** 341–351.

17. Kothari, S., Cizeau, J., Mcmillan-Ward, E., Israels, S. J., Bailes, M., Ens, K., Kirshenbaum, L. A. & Gibson, S. B. (2003) *Oncogene* **30,** 4734–4744.
18. Giatromanolaki, A., Koukourakis, M. I., Sowter, H. M., Sivridis, E., Gibson, S., Gatter, K. C. & Harris, A. L. (2004) *Clin. Cancer Res.* **10,** 5566–5571.
19. Comeron, J. M. (1995) *J. Mol. Evol.* **41,** 1152–1159.
20. Graur, D. & Li, W.-H. (2000) *Fundamentals of Molecular Evolution* (Sinauer, Sunderland, MA), pp. 362–363.
21. Mallet, F., Bouton, O., Prudhomme, S., Cheynet, V., Oriol, G., Bonnaud, B., Lucotte, G., Duret, L. & Mandrand, B. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 1731–1736.
22. Potgens, A. J., Drewlo, S., Kokozidou, M. & Kaufmann, P. (2004) *Hum. Reprod. Update*, in press.
23. Chen, Y.-S., Akula, N., Detera-Wadleigh, S. D., Schulze, T. G., Thomas, J., Potash, J. B., DePaulo, J. R., McInnis, M. G., Cox, N. J. & McMahon, F. J. (2004) *Mol. Psychiatry* **9,** 87–92.
24. Hattori, E., Liu, C., Badner, J. A., Bonner, T. I., Christian, S. L., Maheshwari, M., Detera-Wadleigh, S. D., Gibbs, R. A. & Gershon, E. S. (2003) *Am. J. Hum. Genet.* **72,** 1131–1140.
25. Chumakov, I., Blumenfeld, M., Guerassimenko, O., Cavarec, L., Palicio, M., Abderrahim, H., Bougueleret, L., Barry, C., Tanaka, H., La Rosa, P., *et al.* (2002) *Proc. Natl. Acad. Sci. USA* **99,** 13365–13367.
26. Strausberg, R. L., Feingold, E. A., Grouse, L. H., Derge, J. G., Klausner, R. D., Collins, F. S., Wagner, L., Shenmen, C. M., Schuler, G. D., Altschul, S. F., *et al.* (2002) *Proc. Natl. Acad. Sci. USA* **99,** 16899–16903.
27. 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.
28. Boyd, M. T., Bax, C. M., Bax, B. E., Bloxam, D. L. & Weiss, R. A. (1993) *Virology* **196,** 905–909.
29. Ball, M., McLellan, A., Collins, B., Coadwell, J., Stewart, F. & Moore, T. (2004) *Gene* **325,** 103–113
30. Kazazian, H. H., Jr. (2004) *Science* **303,** 1626–1632.