Ann. Rev. Biochem. 1986. 55:631–61

NONVIRAL RETROPOSONS: GENES, PSEUDOGENES, AND TRANSPOSABLE ELEMENTS GENERATED BY THE REVERSE FLOW OF GENETIC INFORMATION

Alan M. Weiner

Department of Molecular Biophysics and Biochemistry, Yale University School of Medicine, 333 Cedar Street, New Haven, Connecticut 06510

Prescott L. Deininger

Department of Biochemistry, Louisiana State University Medical Center, New Orleans, Louisiana 70112

Argiris Efstratiadis

Department of Human Genetics and Development, Columbia University, 701 West 168th Street, New York, NY 10032

CONTENTS

0066-4154/86/0701-0631$02.00

PERSPECTIVES AND SUMMARY

Movement of genetic information from one locus to another is known as transposition, and in principle, the information could be carried from the parental locus to the target locus as well by RNA as by DNA. Nonetheless, although DNA-mediated transposition has been documented in both prokaryotes and eukaryotes, RNA-mediated transposition appears to be restricted to eukaryotes. In order to distinguish RNA-mediated transpositions from DNA-mediated events, the "reverse" flow of genetic information from RNA back into DNA has been termed "retroposition," and the transposed information is therefore known as a "retroposon" (1, 2). The known retroposons can be divided into viral and nonviral superfamilies based on common structural features (Table 1). The viral superfamily, for which the avian and rodent retroviruses serve as a prototype, has been extensively reviewed elsewhere (3–7).

Over the past few years, retroposition of nonviral cellular RNA species has emerged as a major evolutionary force contributing to the continuous sequence duplication, dispersion, and rearrangement that maintain the remarkable fluidity of eukaryotic genomes. Because nonviral retroposition has given rise to many large families of pseudogenes and transposable elements that confer no obvious advantage on the organism, Orgel & Crick (8) and Doolittle & Sapienza (9) made the provocative proposal that nonviral retroposons could be thought of as molecular parasites that infest the genome but rarely confer a selective advantage (for a recent review, see Ref. 9a). This point of view is surely extreme, given the opportunistic ability of natural selection to find good use for even the most peculiar mutations. In fact, mobile elements in prokaryotes (bacteriophage lambda, mu, P1, and P2, as well as the insertion sequences IS10 and IS50) can confer a selective advantage on their hosts (reviewed in Ref. 10). Moreover, although the genome can be thought of at any instant as a mosaic of sequences that contribute to the organismal phenotype (genic sequences) and those that do not (nongenic sequences), the very fluidity of the genome guarantees a constant flow of sequences back and forth between these two abstract genetic compartments. Thus retroposition, which creates novel se-

Table 1 Retroposons[a]

Viral superfamily	Nonviral superfamily	
Retroviruses (all vertebrates examined)	RNA polymerase II transcripts	
Endogenous or exogenous	Functional semiprocessed retrogene[g]	
Nondefective or defective	Processed retropseudogenes (Table 2)	
Retrotransposons[b]	snRNA retropseudogenes (human and rodent)[h]	
Ty (yeast)	F family *(Drosophila)*[i]	LINES
Copia family *(Drosophila)*[c]	LINE1 family (human, monkey, mouse, rat)[i,j]	LINES
DIRS-1 *(Dictyostelium)*	RNA polymerase III transcripts	
Bs1 (maize)[d]	7SL RNA retropseudogenes	SINES
IAP (rodents)	Unprocessed (RNA sequence intact or truncated)[k]	SINES
THE1 repeats (human)[e]	"Processed" (internally deleted)[l]	SINES
VL30 (rat and mouse)	B1 superfamily (rodents)[m]	SINES
	Homologous composite of processed monomers	SINES
	Dimeric Type I Alu family (primates)	SINES
	Tetrameric Type I Alu family (human)	SINES
	7SK RNA retropseudogenes	SINES
	Unprocessed (RNA sequence intact or truncated)[k]	SINES
	tRNA retropseudogenes	SINES
	Unprocessed (3' truncated)[n]	SINES
	"Processed" (highly mutated and evolved)[o]	SINES
	Monomer family (galago)	SINES
	B2 superfamily (rodents)[p]	SINES
	ID repeats (rat and mouse[q])	SINES
	C repeats (rabbit)	SINES
	C repeats (artiodactyls[r])	SINES
	Heterologous composites[s]	SINES
	tRNA with 7SL retropseudogenes Type II Alu family (galago)[t]	SINES
	Type I Alu with LINE1 (human)	SINES
	7SK retropseudogene with Type I Alu family (human)	SINES
	C with BCS (artiodactyls)	SINES
	B2 with OBY (mouse)	SINES
	Polymerase unknown	
	BCS or A repeats (artiodactyls)	
	Type I and II rDNA insertions *(Drosophila* and *Bombyx)*[u]	
	RIME repeats (Trypanozoons)[t,v]	

Table 1 *(continued)*

Viral superfamily	Nonviral superfamily
Distinguishing Hallmarks	
Dispersed in genome	Dispersed in genome
Bounded by long terminal repeats (LTRs)	No terminal direct or inverted repeats
Transposition intermediate is RNA polymerase II transcript	Both RNA polymerase II and III transcripts serve as transposition intermediates
Active transposition (element presumed to encode reverse transcriptase and/or integrase)[f]	Passive transposition
Generate 4–6-bp target site duplications characteristic of the retroposon	Generate 7–21-bp target site duplications (occasionally shorter or longer)
No 3' terminal poly(A) tract	Often have 3' terminal poly(A) tract
May contain introns that are removed to generate subgenomic mRNAs	Intronless even when parental sequence contains introns (one known exception)

[a] References are given only for retroposons not discussed further in the text.
[b] Term proposed for virallike transposons with no obligatory extracellular phase (5).
[c] Includes copia, 297, 17.6, B104, gypsy, and HMS Beagle; copia itself appears to encode an *env* protein, and may be nondefective retrovirus (4).
[d] Tentatively classified as retrotransposon based on structural similarity to Ty and copia families (19).
[e] No homology detected with retroviral *pol* genes (20); could be progenitor or derivative of retrovirus.
[f] These activites could also be supplied in *trans* by another member of the viral superfamily.
[g] Rat and mouse preproinsulin I gene.
[h] Human U1 (21, 22), U2 (23, 24), U3 (25), U4 (26, 26a), U6 (27); mouse U6 (28); rat U3 (29, 30).
[i] Classified as RNA polymerase II because AATAAA polyadenylation signal precedes 3' poly(A) tract.
[j] Published experiments (31, 32) cannot distinguish RNA polymerase II readthrough transcription from transcripts initiating within the element.
[k] Included as SINEs because 7SL retropseudogenes with 5' truncations have given rise to "secondary" retropseudogenes with 5' substitutions (33), and 7SK retropseudogenes can generate mobile composites with the Type I Alu family (34).
[l] Provisionally classified as "processed," because the internal sequences may have been deleted at the RNA level (see text).
[m] Closely related to rodent 4.5S RNA; for a discussion see Ref. 2.
[n] See Ref. 35.
[o] Could be derived from tRNA gene or retrogene (see text).
[p] Closely related to rodent $4.5S_I$ RNA and rat R.dre.1 element; for a discussion see Ref. 2.
[q] J. G. Sutcliffe, personal communication.
[r] Goat and cow.
[s] Upstream element listed first.
[t] 5' Monomer family + 3' Alu right monomer.
[u] May not be retroposon; for discussion, see Refs. 22 and 22a.
[v] See Ref. 35a.

quence combinations through the duplicative dispersion of genetic information, can shape and reshape the eukaryotic genome in many different ways.

Nonviral retroposons have also been the subject of many previous reviews (1,2,11–18). We do not intend to belabor this information, but rather to update the earlier reviews. In particular, we discuss new data showing that many if not all SINEs [short mobile elements in the terminology of Singer, 1982 (13)] are retropseudogenes derived from known RNA polymerase III transcripts such as 7SL RNA and tRNA. We also review new data confirming that the major families of LINEs (long mobile elements) found in different mammals are closely related to each other. We then discuss the molecular mechanisms that

have been proposed for nonviral retroposition. Finally, we emphasize that the contemporary families of SINEs and LINEs arose very recently in evolution, so that youth itself can explain why the families are relatively homogeneous within a species, but exhibit characteristic differences between species.

THE AMAZING VARIETY OF RETROPOSONS

The Viral Superfamily

Table 1 lists representative retroposons of the viral superfamily and all the known nonviral retroposons. The str
vertebrate retroviruses have been reviewed elsewhere (3, 6, 7). While many members of the viral superfamily encode their own reverse transcriptase and integrase, other members may be defective in one or both of these activities, and require that the enzyme(s) be supplied in *trans* by a nondefective member of the viral superfamily residing in the same genome. For example, the human THE1 repeats are 2.3 kb long with 350-bp long terminal repeats (LTRs), are present in about 10,000 copies per genome, generate 5-bp duplications of the target sequence, and encode a 2.0-kb polyadenylated RNA. However, the DNA sequence of the retroposon appears to be unrelated to that of any known retrovirus (20). The intracisternal A-type particle (IAP) appears to be a retrovirus without an obligatory extracellular phase (36), and the viruslike 30S RNA sequence (VL30) may be similar (37).

Although it was immediately obvious that the Ty family of elements in yeast, and the copia family of elements in *Drosophila,* had most if not all of the distinguishing features of the viral superfamily, it was only recently that the extraordinary power of contemporary yeast genetics was harnessed to demonstrate that Ty elements must transpose through an RNA intermediate (5). A Ty element was constructed with a powerful inducible GAL1 promoter replacing part of the upstream LTR, and an intron derived from the rp51 ribosomal protein gene inserted at an innocuous site preceding the downstream LTR. Transposition of this element healed the upstream LTR and precisely eliminated the intron, thus verifying both transit through an RNA intermediate, and reverse transcription to generate a circular retrovirallike transposition intermediate with a solitary LTR that is a composite of the upstream and downstream LTRs (38). Interestingly, induction of the GAL1 promoter on the marked Ty element mobilized other unmarked Ty elements as well, implying that the reverse transcriptase can also work in *trans*. However, since yeast appears to lack retroposons of the nonviral superfamily, it may be that the Ty reverse transcriptase is specific for Ty transcripts, and cannot work efficiently on other cellular RNAs.

The complete DNA sequence of a representative copia element (4) revealed only weak homology with a number of retroviral proteins, including the reverse

transcriptase, but surprisingly good homology with the part of the *pol* gene encoding the retroviral integrase. Moreover, copia elements appear more similar to yeast Ty than to several vertebrate retroviruses or even the *Drosophila* copia-like element 17.6. In fact, the 17.6 element appears to encode an *env* product, but copia and Ty do not, suggesting that the retrovirallike particles containing copia RNA and a reverse transcriptase may be viral pseudotypes in which copia RNA is packaged by the *env* product of a nondefective retroposon of the viral superfamily.

New members of the viral superfamily will undoubtedly be found as additional species are scrutinized. For example, the DIRS-1 element of *Dictyostelium discoideum* may belong to this family, although it has inverted nonidentical LTRs (ITRs). One of the open reading frames in DIRS-1 encodes a protein with significant homology to regions of reverse transcriptase (39).

The Nonviral Superfamily

With the curious exception of the four ribosomal RNA species (18S, 28S, and 5.8S rRNAs transcribed by RNA polymerase I, and the 5S rRNA transcribed by RNA polymerase III), all the other major classes of cellular RNA species are known to have given rise to retroposons (Table 1). Despite the extraordinary variety of nonviral retroposons, these elements share certain features. All the nonviral retroposons correspond to a partial or complete DNA copy of a cellular RNA species. Some of the parental RNA species are cytoplasmic (tRNA, 7SL RNA, mRNA), while others are predominantly or exclusively nuclear (snRNAs, Alu, and Alu-equivalent sequences). With only two exceptions (the functional semiprocessed rat preproinsulin I gene and the incompletely processed U2 snRNA pseudogene discussed below), nonviral retroposons are derived from fully processed RNAs. For example, retropseudogenes derived from processed mRNAs always have 3' poly(A) tails and lack introns. The sequence of the mature RNA is often intact (particularly in the case of processed mRNAs) but can be truncated at either the 5' end, the 3' end, or rarely at both ends (33). [The RNA information present in LINE1 specimens can also be scrambled relative to the prototype LINE1 sequence, but it is not clear whether such scrambling takes place during or after insertion into a new chromosomal site (see below).] Insertion of a nonviral retroposon usually generates a target site duplication of 7–21 bp, but shorter and longer direct repeats (40) or none at all (22, 25) have occasionally been observed. Interestingly, the length of the target site duplication is not characteristic of the element. Alu sequences, for example, have been found to make target site duplications varying from 9 to 21 bp (41).

Perhaps most surprising of all, the nonviral retroposons exhibit no consistent structural similarities. Nonviral retroposons vary in length from as little as 33 bp (the U2.1 pseudogene) to over 6 kb (the LINE1 prototype). Although

nonviral retroposons usually have a 3' terminal poly(A) tract, most of the snRNA pseudogenes and all of the artiodactyl SINEs represent obvious exceptions. Finally, both RNA polymerase II and III transcripts can give rise to nonviral retroposons. Thus, neither the 5' terminal structure of the RNA (cap or triphosphate) nor the 3' terminal structure [usually poly(A) or oligo(U)] is a prerequisite for retroposition. As discussed in detail below, the amazing variety of nonviral retroposons strongly suggests that these elements transpose passively, i.e. they do not encode or even specify the enzymes responsible for their retroposition. However, the long open reading frame in the LINE1 prototype could be an exception to this generalization.

RETROPOSONS DERIVED FROM PROCESSED MESSAGES

Processed Retropseudogenes

Processed retropseudogenes (previously called "processed genes" in the terminology of Ref. 42) resemble a cDNA copy of a fully processed mRNA species (Tables 1 and 2). These retropseudogenes always include the 3' terminal poly(A) tract of the parental mRNA species, lack any introns present in the parental gene, and often extend to the normal 5' cap site (e.g. Refs. 63, 68, and 69); however, 5' truncations of processed mRNAs are not uncommon. In some cases, 5' truncation is caused by the insertion of another retroposon (53, 60), but in other cases the processed retropseudogene appears to be derived from an aberrant transcript generated by faulty splicing or by initiation downstream from the normal cap site. For example, the retropseudogene derived from the human lambda light chain is missing the V region (42), while the human epsilon heavy chain retropseudogene is missing the VDJ region (79, 80). Perhaps, genes that are subject to strict tissue-specific regulation in the soma usually give rise to retropseudogenes derived from aberrant germline transcripts. Of course we cannot exclude incomplete reverse transcription as the cause of 5' truncation; however, the abundance of full-length retropseudogenes (Table 2) makes incomplete reverse transcription a less likely explanation.

Since the essential promoter elements for RNA polymerase II lie upstream from the transcriptional initiation site, retroposition of a correctly initiated mRNA will almost always generate an inactive retropseudogene. For example, the processed human metallothionein II retropseudogene is inactive, despite the fact that the coding region remains intact (55–56a). Such inactive retropseudogenes degenerate by neutral drift, unless retroposition affected the function of adjacent DNA sequences. In contrast, when mRNA coding regions are duplicated by DNA-mediated events such as tandem duplication or translocation, the flanking regulatory elements (and introns) are preserved; thus the new locus is potentially active, and may be subject to positive or negative selection

Table 2 Retropseudogenes derived from processed messages

Species	Protein	Number of genes	Chromosomal location of gene(s)	Number of retropseudogenes	Chromosomal location of pseudogenes	mRNA sequence in pseudogene: intact	mRNA sequence in pseudogene: 5' truncated	mRNA sequence in pseudogene: Initiates upstream	Refs.
human	triosephosphate isomerase	1	12	5–6 [3]	≠12	+			43
human	argininosuccinate synthetase	1	9	14 [2]	6, 9, X, Y etc	+			44–46
human	phosphoglycerate kinase	1 somatic	X						
		1 testis-specific	?	1	6	+*			47
human	glyceraldehyde-3-phosphate dehydrogenase	1	12	~25	one on X	+			48–50
mouse	glyceraldehyde-3-phosphate dehydrogenase	1	6	200					48
human	lactate dehydrogenase	?		? [1]					51
human	dihydrofolate reductase	1	5	~5 [4]	≠5 (one on 3)		+ [2]		52, 52a, 53
human	metallothionein[a]	2 MT-I		~11 [1 MT-I] [1 MT-II]	1, 4, 18, 20 etc	+			54–58
		1 MT-II	16			+			
mouse	cytochrome *c*	1 somatic		20–30 [3]		+			59
		1 testis-specific							
rat	cytochrome *c*	1 somatic		20–30 [7]		+ [3]	+ [1][b]	+ [3]	60, 61
mouse	ribosomal protein L32	1*		16–20 [3][c]		+			62, 63
human	ribosomal protein L32	1*		~20 [1][d]		+			64
mouse	ribosomal protein L30	1*		≥15[4]		+			65
mouse	ribosomal protein L7	1–2*		≥20 [0]					66

mouse	ribosomal protein L18	?		≥8 [0]					67
rat	α-tubulin	2*		10–20 [1]		+			68
human	β-tubulin	2		15–20 [5]		+ [4]	+ [1]		69
human	β-actin	1*		~20 [2]		+			70, 71
mouse	cytoplasmic γ-actin	?		? [1][e]			+		72
mouse	myosin light chain	1		1 [1]			+		73
human	nonmuscle tropomyosin	1		≥3 [1]		+*			74
mouse	cytokeratin endo A	1		1 [0]					75
mouse	tumor antigen p53	1	11	1 [1]	14		+		76
mouse	α-globin	2	11	1 [1]	15			+	17
mouse	pro-opiomelanocortin	1	12	1 [1]	19*				77, 78
human	Ig ϵ heavy chain	1 Cϵ	14	1 [1][f]	9		+		79, 80
human	Ig λ light chain	6 Cλ	22	? [1][g]	≠22		+		42
human	c-Ha-ras	1	11	1 [1]	X		+*		81
human	c-Ki-ras	1	12	1 [1]		+*			82
human	c-raf	1	3	1 [0]	4				83

Numbers in brackets indicate number of retropseudogenes currently sequenced; * denotes uncertainty

[a] Protein sequencing data suggests the existence of a third MT-I gene.
[b] Rat LINE1 sequence lies upstream from point of truncation.
[c] One retropseudogene is in the second intron of the DHFR gene.
[d] The sequenced retropseudogene is in the first intron of HLA-SBβ gene.
[e] Inserted into a mouse LINE1 sequence.
[f] Missing VDJ sequences.
[g] Missing V sequence.

in addition to neutral drift. Occasionally, however, a processed mRNA may insert fortuitously downstream from a foreign promoter, or may acquire a promoter after retroposition. The intronless chicken calmodulin gene could have arisen in this way, although the other hallmarks of retroposition have yet to be demonstrated (84). Finally, in rare instances, retroposition of an aberrant mRNA initiating upstream from the normal cap site may move the normal promoter along with the mRNA sequence (Table 2), and thus potentially be able to generate a functional processed retrogene. The rat preproinsulin I gene discussed below is an example of such a rare event, and is presumably derived from an aberrant germline transcript of this otherwise tissue-specific mRNA. However, retropseudogenes derived from aberrant upstream transcripts are not necessarily subject to positive selection, as the mouse alpha-globin retropseudogene (reviewed in Ref. 17) and the rat cytochrome *c* retropseudogenes (60) demonstrate.

Curiously, nonviral retroposition is most commonly found in mammals, even when the same gene family has been examined in a variety of organisms. For example, the *Drosophila* and chicken alpha- and beta-tubulin genes are no less abundant than those of mammals, but mammals have 20–30 tubulin retropseudogenes while *Drosophila* and chicken have none. Similarly, all of actin and cytochrome *c* loci in the *Drosophila* and chicken genomes are true genes, while most of the human beta-actin and rodent cytochrome *c* loci are retropseudogenes. Most dramatically, retropseudogenes for the mammalian glyceraldehyde-3-phosphate dehydrogenase are very abundant (about 25 copies in man, rabbit, guinea pig, and hamster, but more than 200 copies in rat and mouse) while the same gene is single copy in the chicken. However, the existence of the F family in *Drosophila* and the putative calmodulin retrogene in the chicken (discussed above) suggest that retroposition can occur in other organisms, although it is much more frequent in mammals.

A Functional Semiprocessed Retrogene

The rat and mouse preproinsulin I gene is for the moment the sole example of a functional retrogene (40). Rats and mice (as well as three species of fish) have two nonallelic insulin genes, designated preproinsulin I and II, which are almost equally expressed. Gene I contains a single small intron in the 5' noncoding region, whereas gene II contains this same small intron as well as an additional larger intron within the coding region for the C peptide (85). Gene II most closely resembles the ancestral gene, because the unique chicken, dog, guinea pig, and human genes each have two introns. The precise deletion of the small intron from gene I suggested that this locus might be a semiprocessed retrogene derived from the parental gene II, but a deletion in the 5' flanking region of the rat gene I obscured the upstream direct repeat. Comparison of the mouse gene I with the rat genes I and II was therefore required to establish that

both the rat and mouse preproinsulin I genes are functional retrogenes derived from an aberrantly initiated and partially processed gene II transcript. The retroposed sequence is flanked by 41-nt direct repeats, and the polyadenylation signal is followed by an $ACCA_4$ tract in the rat and an $ACCA_8$ tract in the mouse. The aberrant transcript appears to have initiated at least 0.5 kb upstream from the normal cap site, so that the RNA intermediate for retroposition carried most if not all of the preproinsulin II promoter and regulatory sequences.

Although organisms can clearly survive with only one insulin gene, even when the product has diminished biological activity as is the case in guinea pig (86), fixation of the functional semiprocessed preproinsulin I gene appears to reflect positive selection. This conclusion is based on the independent preservation of two functional genes in two different species over sufficient evolutionary time for one of the genes to be inactivated by drift. Thus, replacement substitutions have accumulated in regions encoding the signal and C peptides (which are eliminated by protein processing), but such substitutions are completely absent from the regions encoding the A and B chains of the mature hormone. The existence of such strong negative selection on the A and B chains in gene I suggests that the two-copy state was initially subject to positive selection. This conclusion is strengthened by the observation that the sequences upstream from the promoter and within the single small intron (where rat gene I and II can be compared with mouse gene I) as well as the 3' flanking sequences (where the rat and mouse gene I can be compared) are evolving neutrally.

The murine preproinsulin gene I is unlikely to be the only functional mRNA retroposon. The intronless globin gene of Chironomus (87) might be another example. All but one of the 17–20 intronless actin genes in *Dictyostelium* are functional (88), and these might also be retrogenes; however, the extreme A-richness of the flanking sequences would obscure the hallmarks of retroposition. It is also possible that the overlapping subsets of introns in the actin multigene families of many species could be explained by invoking independent retropositions of different partially spliced transcripts.

LINES (LONG INTERSPERSED REPEATED SEQUENCES)

Mammalian genomes contain 20–50 thousand copies of a long (6–7-kb) interspersed element known as the LINE1 or L1 family (reviewed in Refs. 1, 2, 13–15). Many fragments of this family had been independently named and characterized in several mammalian species before they were recognized in 1983 as parts of a single family of retroposons. Now it is clear that each mammalian genome is inhabited by a related but generally species-specific L1 family. Each of these families can be thought of as an especially abundant and remarkably complex family of processed retropseudogenes, although we should keep in mind that some L1 elements may actually be functional ret-

rogenes. The structural features of typical L1 specimens support this conclusion. Most L1 elements terminate at the 3' end with an A-rich tract, sometimes preceded by a polyadenylation signal. L1 elements clearly contain one or more open reading frames (ORFs; Refs. 15, 89–92). L1 elements are mobile, and usually make target site duplications at the site of insertion (93–98). Moreover, a variety of polymorphisms caused by L1 insertions (or deletions) suggest that at least some contemporary L1 elements are capable of retroposition (99–101; A. V. Furano, personal communication). Finally, the putative retroposition intermediate, a homogeneous 6.5-kb polyadenylated transcript, can be detected in the cytoplasm of relatively undifferentiated human NTera2 teratocarcinoma cells (102).

However, this simple summary is misleadingly neat:

No complete L1 element has yet been sequenced; instead, when the sequenced fragments of the human, monkey, dog, rat, and mouse L1 families are aligned with the rat L1 family, for which over 6.5 kb of composite sequence is available, a "consensus" sequence can be derived for the rat, mouse, and human L1 families (92). An independently derived consensus for the primate (human and monkey) L1 sequence covering approximately 6 kb without interruption has also been compiled (sequence unpublished but available upon request; Ref. 15). Unfortunately, parts of these consensus sequences are represented by only a single specimen, and there is no guarantee that this specimen is free from mutations affecting the provisional ORFs. Thus, the three major ORFs in the primate L1 consensus may be fused or extended upstream by minor changes in the current consensus sequence (15).

Many L1 elements are severely truncated at the 5' end, so that the sequences from the 3' end are present at more than fivefold higher copy number than the complete elements (93–95, 103–106). Such 5' truncation has usually been attributed to incomplete reverse transcription, but other explanations are possible (see below).

Some L1 elements are internally scrambled or deleted, lack a 3' terminal A-rich tract, and have inserted cleanly into the target site without duplication (91). These L1 elements lack the obvious marks of retroposition, and were probably dispersed by a DNA-mediated event such as recombination between extrachromosomal L1-containing circles and the chromosome (107–109a).

Although human NTera2 cells contain a 6.5-kb L1 transcript, L1-related transcripts in more highly differentiated cells are heterogeneous, confined primarily to the nucleus, and may reflect readthrough transcription by RNA polymerase II from promoters outside the element (96, 104, 110–112). In addition, readthrough transcription from promoters for RNA polymerase III lying outside the element may also contribute to the heterogeneity of L1-related transcripts (113).

Finally, although the primate and rodent L1 families exhibit about 60%

homology over a region of about 1,500 bp that includes one of the ORFs, no significant homology can be detected in the first 2 kb from the 5' end of the L1 elements from the two species (T. N. H. Lee, M. F. Singer, T. Fanning, unpublished results quoted in Ref. 15), and the presumed 3' untranslated regions (200 bp in primates and about 700 bp in rats and mice) are completely different. Such species-specific differences at the 5' and 3' ends of mammalian L1 families may reflect the formation of composite elements derived from an ancestral L1 sequence which was present only in low copy number after the mammalian radiation (see below for further discussion of composite elements). Indeed, two human composite L1 elements, each starting with an Alu sequence, have been described (97).

The abundance of the mammalian L1 family raises several obvious questions:

Do the 6–7-kb elements represent the complete L1 sequence, or do these elements arise by retroposition of processed mRNAs derived from a larger functional gene that is present in much lower copy number? Such a hypothetical parental element could have introns, or even belong to the retroviral superfamily; in either case, regulation of the parental element, and of the processed retrogenes derived from it, might be very different.

Can the many truncated and scrambled L1 sequences themselves serve as templates for further retropositions, or are these structurally diverse genomic L1 sequences "frozen" in place like other processed mRNAs? The answer to this question presumably depends on the details of L1 transcription. Do L1 sequences contain internal promoters for RNA polymerase II, or could read-through from random RNA polymerase II promoters upstream [stimulated perhaps by one or more internal enhancers within the L1 sequence (114)] supply transcripts for retroposition? If a 3' terminal poly(A) tract is required for retroposition, can the AAUAAA polyadenylation signal remaining in a processed L1 retrogene function without the recently identified downstream components of the polyadenylation signal (115–117)?

Are L1 sequences abundant because they are abundantly transcribed during the time in development when germline retroposition occurs most efficiently, or because the L1 element itself carries sequences which make it a particularly efficient substrate for retroposition? For example, retroposition could occur in cleavage-stage embryos where the 6.5-kb transcript may be abundant (102), or in oocytes during the prolonged lampbrush stage characteristic of mammalian oogenesis (see below). Alternatively, L1 transcripts might encode a *cis*-acting reverse transcriptase, or a karyophilic protein that can transport the mRNA or cDNA into the nucleus (90). However, the 3' terminal ORF present in most mammalian L1 specimens does not appear to be homologous to the conserved domains of retroviral reverse transcriptases (118).

Why are the structures of genomic L1 sequences so much more complex than

those of other abundant retroposons such as the Alu family? The high frequency of 5' truncation might reflect blocks to reverse transcription caused by RNA secondary structure (90, 106, 119), or by RNA branch structures, nucleotide modification, or protein binding. Alternatively, the overrepresentation of sequences derived from the 3' end of L1 elements might reflect preferential nuclease attack at specific sites in genomic L1 sequences, followed by recombination and subsequent retroposition that is dependent on sequences in the 3' end of the element (92). This interpretation is consistent with the observation that some L1 specimens from different species share common sites of 5' (or even 3') truncation, regardless of whether they appear to have arisen by an RNA- or DNA-mediated process (92). Site-specific nuclease attack, followed by recombination, could also generate the discrete sizes of extrachromosomal L1-containing circles observed in monkey cells (107); subsequent integration of these circles followed by additional recombination events could then generate permuted and scrambled L1 sequences (91, 108). Still another possible explanation for scrambling of L1 sequences is the ability of retroviral reverse transcriptases to promote high levels of viral recombination, presumably because strand displacement is a fundamental property of the reverse transcriptase reaction (120); even a small amount of recombination between abundant L1 transcripts would quickly generate a diverse population of L1 sequences.

Although the LINE1 family is clearly most abundant in mammalian genomes, analogous LINEs are probably present in the genomes of lower eukaryotes. For example, the F elements of *Drosophila* are mobile, present in about 50 copies per genome, make 8–13-bp target site duplications upon insertion, and have a 3' poly(A) tract preceded by a AATAAA polyadenylation signal (121). Of five cloned F elements, three appear to be full-length 4.7-kb elements whereas the two others are truncated at the 5' end.

SINES (SHORT INTERSPERSED REPEATED SEQUENCES)

SINEs are short (approximately 70–300 bp) repetitive elements that are often present in over 100,000 copies per genome. Almost all known SINEs appear to be retroposons; only the bovine consensus sequence (BCS family; Refs. 122 and 123) and the mouse OBY family (2) are possible exceptions. Unlike members of LINE1 families, members of each SINE family generally have well-defined 5' and 3' ends, with variation between elements occurring primarily in the characteristic simple sequence [usually oligo(A)] found at the 3' end of the element. SINEs usually make target site duplications of 7–21 bp upon insertion. Most SINE families except for the BCS (123) and OBY families (2) have been shown to carry a functional internal promoter. As anticipated by

Jagadeeswaran et al (124) and first clearly documented by Rogers (2), SINEs with an internal RNA polymerase III promoter can cotranscribe adjacent chromosomal sequences and thereby mobilize them. In particular, the propensity of SINEs to insert into the 3' terminal simple sequence tail of other SINEs can generate mobile composite elements (Table 1; also, see below).

The Human Alu Family and the Rodent B1 Family

The human Alu family is the best studied of all SINEs, and will serve here as a paradigm for this class of retroposons. The 500,000 Alu elements in the human genome constitute a remarkable 5–6% of the genome by mass (125). The Alu, LINE1, and THE1 families, together with poly(CA), account for most of the highly repetitive DNA in the human genome (20, 126). Alu elements, which are approximately 300 bp long, are now known to be dimeric retropseudogenes, derived from 7SL RNA by one or more internal deletions of 7SL sequence followed by a dimerization (see below). The right monomer is 31 nucleotides longer than the left because the left monomer has sustained a more extensive internal deletion of 7SL sequence. Individual Alu elements diverge from the consensus by about 14% as a result of single base changes, insertions, and deletions (reviewed in Ref. 41).

The internal promoter defines the 5' end of the Alu element by directing the initiation of transcription by RNA polymerase III at a fixed distance upstream (131). This promoter exhibits a typical A and B block consensus (127). One or more A blocks lie close to the 5' end of the element (positions 5 and 31), and appear to increase both the strength (128, 129) and accuracy (129, 130) of initiation; however, only the B block between positions 70 and 100 appears to be essential for transcription (128, 129).

Curiously, no monomeric Alu elements have ever been found in the human genome, although the rodent equivalent of the Alu sequence (the B1 superfamily; Table 1) is almost exclusively monomeric. The absence of monomeric human Alu elements may not be difficult to explain if only the left Alu monomer has a functional promoter; the right monomer may be inactive (128–130). Inactivity of the right monomer promoter is puzzling, because the right monomer is more homologous to the 7SL sequence than the left (133) and might have been expected to better preserve the 7SL promoter structure; however, even single base changes could in principle affect the promoter activity of either monomer. Although it is tempting to attribute the transcriptional inactivity of the right monomer to the 31 additional nucleotides that lie within the right monomer B block, we would then be at a loss to explain what sequences function as the B block for intact 7SL genes (33).

The ability of a newly retroposed Alu element to serve as a template for further retropositions will depend on whether the Alu promoter can function

efficiently in the context of new 5' flanking sequences. However, the Alu promoter is derived from the 7SL promoter, which is known to require compatible upstream sequences for efficient transcription (134). Perhaps the many mutations in the left monomer relative to the right have rendered the left promoter less dependent on upstream sequences and therefore capable of more efficient retroposition.

A simple A-rich sequence defines the 3' end of Alu element and almost all other SINEs, except for artiodactyl SINEs, which lack an A-rich tail and usually have simple repeating sequences [e.g. $(AGC)_n$] instead (122, 123). The length of the A-rich sequence varies from 4 to more than 50 bp between individual Alu elements, and the A-rich sequence, though often a relatively pure homopolymer tract, is sometimes supplemented with or even replaced by simple sequences such as $(NA_x)_y$ (2, 12, 41).

Transcription by RNA polymerase III initiates at the 5' end of the element, transcribes through the entire element including the 3' terminal A-rich tract, and then terminates beyond the downstream direct repeat at one or more random oligo(dT) tracts in the adjacent chromosomal sequences (128–135). Reverse transcription is then primed on the 3' terminal A-rich tract, either by the 3' oligo(U) tract of the Alu transcript itself (124) or by another cellular RNA or DNA (21). In either case, the part of the Alu transcript derived from the 3' flanking chromosomal sequences will be lost upon reverse transcription, thus preserving the "anonymity" of the retroposed information. Although self-priming provides an attractive explanation for the high efficiency of Alu retroposition, one should keep in mind that processed mRNAs cannot possibly self-prime, and that artiodactyl SINEs do not terminate with self-complementary simple sequences (122, 123).

The abundance of Alu and other SINE families is usually attributed to the ability of many (and perhaps all) newly retroposed elements to serve as templates for further retroposition. But if reverse transcription is primed at random sites within the 3' A-rich tract, why doesn't the A-rich tract grow shorter with each successive round of retroposition? One possibility is that the first few bases of cDNA might slip back and prime again, thereby lengthening the tail and perhaps amplifying other simple sequences present in the parental A-rich tract (124). Another possibility is de novo addition of either a homopolymer or simple sequence at the staggered chromosomal break before insertion (2). A third very interesting possibility is that the A-rich region might be expanded at any time after retroposition by one or more of the mechanisms commonly invoked to explain the expansion of simple satellite sequences (1, 136–138). Perhaps the most compelling argument that a mechanism of this type may be at work on the 3' ends of Alu sequences is that similar expansions can also occur at the internal A-rich region following the left monomer (139, 140).

Alu and B1 Sequences are Processed 7SL Retropseudogenes

Human Alu sequences are dimeric, but the homologous rodent sequences (the B1 superfamily) are monomeric. How did these efficient retroposons arise? Today there is no doubt that both Alu and B1 sequences are derived from 7SL RNA, although the details of this process are still a matter of speculation (143).

The 7SL RNA is a component of the signal recognition particle, the cytoplasmic ribonucleoprotein particle required for cotranslational secretion of membrane and secretory proteins into the lumen of the rough endoplasmic reticulum (141). As expected for a component of the translational apparatus, 7SL RNA has been highly conserved throughout evolution. Thus, the fingerprints of chicken, mouse, and human 7SL RNA are similar. The sequences of rat and human 7SL RNA are identical, as are the 3' terminal sequences of dog and human 7SL RNA (literature reviewed in Ref. 142). Finally, the sequences of human and *Xenopus* 7SL RNA are very similar, and can be convincingly aligned with the *Drosophila* 7SL RNA sequence (143).

The sequence of the human Alu right monomer is almost identical to the 7SL RNA sequence, except for the deletion of 155 internal nucleotides from the 7SL sequence. The human Alu left monomer also lacks these 155 nt, as well as an adjacent 31 nt of internal 7SL sequence. The rodent B1 sequence resembles the human Alu right monomer, but the internal deletion of 7SL sequence extends 14 nt further downstream.

The belief that Alu and B1 sequences are derived from the 7SL sequence, rather than vice versa, rests primarily on the remarkable evolutionary conservation of 7SL RNA. In addition, the existence of abundant "unprocessed" 7SL retropseudogenes in the human genome (5' and/or 3' truncated; Ref. 33) demonstrates that mechanisms are available for generating 7SL retropseudogenes, some of which could be "processed" (internally deleted). The derivation of Alu and B1 sequences from 7SL is also consistent with the absence of Alu-like sequences (but the presence of 7SL genes) in the *Drosophila* genome (143, 144). Curiously, although both *Xenopus* and sea urchin contain sequences that cross-hybridize with Alu probes, 7SL probes fail to detect a 7SL RNA in the urchin (144).

How, then, did the 7SL sequence give rise to the monomeric B1 and dimeric Alu sequences, and why are the rodent and human repeats structurally different? For simplicity, we will assume that 7SL sequences originally gave rise to an ancestral retroposon closely resembling the contemporary human Alu right monomer, and that subsequent rearrangements of this ancestral monomeric Alu then generated the dimeric human Alu and the monomeric rodent B1 sequences; however, it is equally possible that these two families arose by independent deletion events directly from 7SL RNA. As we have seen above, both B1 and Alu sequences retropose using an RNA polymerase III transcript as the in-

termediate. Both these SINEs have an internal promoter for RNA polymerase III (which defines the 5' end of the element) and a 3' terminal A-rich tract that serves as the priming site for reverse transcription (thereby defining the 3' end of the element). Thus, generation of the ancestral Alu right monomer from the 7SL sequence (itself an RNA polymerase III transcription unit) would require at least two changes: acquisition of an A-rich tract immediately following the 3' end of the mature RNA sequence, and deletion of 155 internal nt. We will argue that both these initial events occurred at the RNA level.

Addition of an A-rich tract precisely at the 3' end of the 7SL RNA coding region is very unlikely to have taken place at the DNA level, i.e. by deletion or recombination. In contrast, it is easy to imagine addition of a poly(A) tail to 7SL RNA. Although this kind of polyadenylation might be considered aberrant, the 7SL7 and 7SL23 retropseudogenes provide a powerful precedent (33). In both these pseudogenes, the 3' end of the 7SL RNA sequence almost precisely abuts a poly(A) tail, and flanking direct repeats confirm that the pseudogenes were generated by retroposition. Moreover, there are other examples of retroposons with poly(A) tails that are derived from RNAs that are not normally polyadenylated (e.g. the human U1 snRNA pseudogene U1.101 of Ref. 21 and the rat U3 snRNA pseudogene H3.3 of Ref. 30). In fact, a significant fraction of 7SL RNA is retained on an oligo(dT) cellulose column (E. Ullu, personal communication). Other unusual polyadenylated RNA polymerase III transcripts have also been reported (145).

We speculate that formation of a 7SL retropseudogene with a poly(A) tail created a retroposon that could preserve the 3' end of the 7SL RNA sequence through successive retropositions. Although the subsequent deletion of 155 nt from this retroposon could have occurred at the DNA level, a case can be made that this event also took place at the RNA level (143). First, the major sites of micrococcal nuclease attack on 7SL RNA within the intact signal recognition particle (146) correlate well with the boundaries of the deletion found in the human Alu right monomer (133). Second, in the probable secondary structure of 7SL RNA (142) the resulting 5' and 3' halves of the nascent Alu right monomer would be held together by strong secondary structure. These two observations suggest that the human Alu right monomer might have resulted from nuclease attack around positions 100 and 250 of 7SL RNA in the intact signal recognition particle. Ligation of the 5' and 3' halves of the sequence, followed by retroposition of the internally deleted RNA sequence, would then generate the first human Alu right monomer.

The left human Alu monomer could have arisen independently from a polyadenylated 7SL RNA, following nuclease attack and deletion of 186 (155 + 31) nt. This would be consistent with the presence of a micrococcal nuclease-sensitive site at position 76. Alternatively, the human Alu left monomer could have arisen from the ancestral human Alu right monomer by a deletion at the DNA level between the hyphenated short direct repeats (TGCAgtgAGC and

TGCActccAGC), which are present at the appropriate positions; such deletions frequently remove all of the upstream direct repeat and portions of the downstream direct repeat (145a).

Finally, the human Alu right and left monomers were fused to form the contemporary dimeric element. This step is not difficult to imagine, since retroposons frequently insert into the A-rich tail of other retroposons (2). In fact, two dimeric Alu elements can occasionally fuse to form a tetrameric Alu element; the resulting tetramer is flanked by one set of direct repeats, indicating that it has transposed as a unit (145b, c, d). Similarly, although the Type I Alu sequence of the prosimian galago closely resembles the dimeric Alu of higher primates (147), the galago Type II Alu sequence appears to be an independent composite of an Alu right monomer and a tRNA-derived SINE designated the Monomer family (148, 149; see below).

As mentioned above, the absence of monomeric Alu elements from the human genome can be ascribed in part to the inactivation of the promoter in the right monomer (128–130). Further experiments will be required to determine whether this inactivation is due to single base changes, or to the 155-nucleotide deletion of 7SL sequence; the possibility remains that the deletion of 31 additional nt in the left monomer activates the left promoter. The failure of the left Alu monomer to spawn monomeric elements suggests that the internal oligo(A) tract following this monomer is too short to serve as an efficient template for reverse transcription.

Most Other SINEs are tRNA Retropseudogenes

Just as the Alu family arose from 7SL RNA, so many other families of SINEs are derived from tRNA or tRNA genes (2, 149–151). The homology between the internal RNA polymerase III promoters of SINEs and tRNAs was immediately obvious (123, 148, 152–154), but more extensive homologies were found only recently. Such homology was first noted between the goat C family and a cysteine tRNA (2). Subsequently, the rat ID, mouse B2, rabbit C, and bovine 73-bp repeats (149–151) were also found to display strong homologies to tRNA, although sequence divergence within the repetitive elements has led to some disagreement regarding the exact parental tRNA species. A particularly convincing homology can be found between the galago Monomer family and an initiator methionine tRNA (149). Structures resembling typical tRNA stems and loops can also be drawn for the various SINE sequences, although the base-pairing is much poorer than that in tRNA. However, the consensus sequences of several SINE families are more homologous to the presumed parental tRNA and exhibit much better base pairing. This suggests that individual SINE elements are subject to neutral drift from the parental tRNA sequence without strong selection for tRNA-like structures (149).

How were these SINEs derived from tRNA? In each case, the tRNA homolo-

gy lies at the 5' end of the SINE sequence as expected for an internal RNA polymerase III promoter; however, we should not discount the possibility that mutations have increased the efficiency of retroposition by rendering the promoter less dependent on compatible 5' flanking sequences (see above, and Ref. 134). Also, as for the ancestral Alu, addition of a 3' terminal oligo(A) or a simple repeating sequence was necessary to serve as an efficient priming site for reverse transcription. Since the 3' terminal poly(A) tract of the galago Monomer (149) and rat ID elements (151) very nearly abuts the tRNA sequence, just as the 3' terminal A-rich tract abuts the 7SL sequence in the Alu right monomer, these elements probably arose by aberrant polyadenylation of the parental tRNA. In contrast, the 3' terminal A-rich tract of the rodent B2 and rabbit C families lie far downstream from the tRNA homology (84 and 251 nt respectively; Ref. 151), suggesting that these SINEs arose by retroposition of a readthrough transcript derived from a tRNA gene, pseudogene, or retropseudogene.

We noted above that the tRNA homology within SINE sequences has diverged significantly from the parental tRNA sequence. Although much of this divergence can be attributed to neutral drift, another interesting interpretation is that the RNA sequence within SINEs must become nonfunctional before the element can evolve into an efficient retroposon. For the Alu family, the deletion of 155 nt from the 7SL sequence may have been sufficient in itself to alter function, for example by abolishing the ability of the RNA to bind the protein components of the signal recognition particle (141). For SINE families derived from tRNA sequences, an accumulation of point mutations may be necessary to abolish tRNA-like folding, RNA processing, or protein binding, before the retroposon can escape a strong negative selection. This may explain why the homologies between SINE families and tRNA seldom exceed 70%, while the homology between Alu and 7SL RNA exceeds 80% (133). Obviously, such mutations must occur prior to multiplication, and should therefore become obvious as the SINE consensus sequences improve; sequence divergence after multiplication to high copy number would be much less likely to affect the consensus sequence.

SINEs abound in all mammalian genomes. In addition to the human and rodent families already discussed, the rabbit C family (154) and the unrelated artiodactyl C family (122, 123, 155) display all the characteristics of short transposable elements. The rabbit C family terminates with the expected 3' terminal oligo(A) tract, but surprisingly, the artiodactyl C family terminates much more often with simple sequence repeats than with oligo(A). How reverse transcription could be primed on such a template is a mystery, although slippage of the initial cDNA and repriming provides an attractive possible mechanism which could also maintain the length of the 3' terminal A-rich tract in SINEs (124; also, see above).

SINEs may not be restricted to mammals. In vitro transcription of total genomic DNA from salmon, newt, and tortoise with a HeLa cell extract yields a small homogeneous RNA polymerase III transcript in each case (151, 156, 157). These transcripts are derived from highly repetitive sequences with tRNA-like structures, and may represent transposable elements (157). In addition, the mouse B1 sequence hybridizes well to DNA from a few species such as maize and chicken, but not to DNA from many other species (158). However, the abundant OAX transcript in *Xenopus* (159) and the CR1 sequence in chicken (160) now appear to lack the hallmarks of retroposition. Moreover, despite extensive studies on the structure and organization of genes encoding mRNA and snRNA in both *Xenopus* and chicken, no retropseudogenes have been definitively identified in either organism.

Can SINEs Serve as Tissue-Specific Markers?

Recently, the provocative proposal has twice been made that the presence of a particular mobile element within an mRNA transcript may serve to identify the tissue from which that mRNA is derived. The rat ID (or "identifier") sequence was proposed as a marker for brain-specific transcripts (153, 161), and the mouse "Set 1" (or B2) sequences were claimed to be a marker for mRNAs specific to both normal embryonic and oncogenically transformed cells (162, 163). How successfully have these claims been substantiated?

The 82-nucleotide rat ID sequence has all the distinguishing marks of a retroposon: flanking direct repeats and a 3' terminal A-rich region (164), an internal RNA polymerase III promoter (161), and high copy number in the genome of rats and mice (165). The element has been found in the 5' flanking region of the *v*-Ha-*ras* gene (166), in introns of mRNA precursors (164), and in the 3' untranslated regions of mature mRNA (153). The ID element is closely related to (and may in fact be a retroposon derived from) two small RNA polymerase III transcripts known as BC1 and BC2, which are found exclusively in neural tissue in the animal (161) but are present in many rat cell lines (J. G. Sutcliffe, personal communication). BC2 is about 160 nucleotides long, and may be a polyadenylated form of the 110-nucleotide BC1 RNA.

The bold hypothesis that ID elements could serve as a marker for neural-specific RNA polymerase II transcription units was based on two separate (and, in retrospect, unrelated) observations. First, the small BC1 and BC2 RNAs transcribed by RNA polymerase III are found in cytoplasmic RNA from brain, but not in RNA from kidney or liver (153). Second, the closely related ID retroposon sequences are found in introns of some brain-specific mRNAs (165). Thus it was conceivable that (*a*) the same factors that activate the absolutely neural-specific transcription of BC1 and BC2 by RNA polymerase III could also activate the transcription of ID sequences within a subset of neural-specific mRNA transcription units, and (*b*) the transcription of ID

sequences within neural-specific genes by RNA polymerase III could in turn activate tissue-specific transcription of these mRNAs by RNA polymerase II. The ability of the ID sequence to function as a transcriptional enhancer in certain transient expression assays (167) increased the credibility of this model for transcriptional regulation by ID sequences.

The most serious objection to the ID hypothesis is that the basic outlines of mammalian embryogenesis and development clearly antedate the relatively recent multiplication of species-specific SINEs in mammalian genomes. Thus retroposons such as ID cannot play an obligatory role in mammalian gene regulation. Although this objection rules out the ID hypothesis in its original form, the possibility remained that neural-specific transcription of ID sequences by RNA polymerase III could increase transcription from adjacent RNA polymerase II promoters. This more modest hypothesis predicts that ID sequences should be overrepresented in neural-specific genes: random insertion of the ID retroposon into nonneural genes would be expected to cause inappropriate expression of those genes in neural tissue, and thus be subject to negative selection. In contrast, ID sequences could accumulate in neural-specific genes because these insertions would simply "reinforce" normal tissue-specific regulation, and thus be selectively neutral. Consistent with prediction, Sutcliffe et al (161) found that readthrough transcription of ID sequences by RNA polymerase II in purified nuclei was greatly increased when the nuclei were isolated from brain tissue.

However, subsequent examination of total in vivo nuclear RNA by Owens et al (167a) demonstrated beyond a doubt that the earlier results (161) were misleading, and that ID sequences are present in similar abundance in the nuclear RNA polymerase II transcripts of other organs. Thus ID sequences are unlikely to function as transcriptional regulatory elements. The function of ID sequences, if any, remains to be determined. One interesting possibility is that RNA polymerase III transcripts of ID sequences might hybridize in vivo to RNA polymerase II transcripts containing ID sequences in inverted orientation; such hybrids could influence posttranscriptional regulation of neural-specific transcripts.

Scott et al (169) originally cloned cDNAs derived from mRNA species that are specifically activated by SV40 transformation of BALB/c 3T3 cells. These cDNA clones fell into a small number of cross-hybridizing sets, one of which (Set 1) turned out to share a small dispersed repetitive element. Murphy et al (162) then demonstrated by Northern blotting that polyadenylated mRNAs reacting with the Set 1 probe peak during embryonic organogenesis and decline thereafter; in addition, mRNAs reacting with the Set 1 probe are much more prevalent in undifferentiated EC embryonal carcinoma cells and F9 teratocarcinoma cells than after differentiation. Thus, it was proposed that the Set 1 repetitive element might be a "general onco-fetal marker." Surprisingly,

however, the Set 1 element was subsequently reported to be none other than a mouse B2 sequence, a typical SINE (163).

How could a mouse B2 sequence be mistaken for a general onco-fetal marker? Singh et al (170) have recently demonstrated that RNA polymerase III transcription of mouse B2 sequences is increased 5- to 20-fold by SV40 transformation of mouse NIH 3T3 cells; moreover, even in untransformed 3T3 cells, transcription of B2 sequences by RNA polymerase III is sensitive to growth conditions and virtually disappears at high cell densities. The careful observations of Singh et al (170) confirm previous observations that transcription of SINEs and LINEs is often derepressed in relatively undifferentiated tissues (102, 106, 162, 168). Although it is possible that B2 sequences might activate adjacent RNA polymerase II transcription units in undifferentiated cells (a version of the original ID hypothesis), the apparent enrichment for B2 sequences in mRNA from undifferentiated cells could be explained in many other ways. For example, SINEs might preferentially retropose into active chromatin, and the chromatin structure of germline cells might resemble that of undifferentiated somatic cells.

SPECULATIONS ON THE MECHANISM(S) OF RETROPOSITION

Despite considerable speculation, the mechanism of retroposition remains unknown. Here we attempt to clarify the major questions, and to classify the distinguishing features of each proposed mechanism as succinctly as possible. All authors agree that the structural characteristics of nonviral retroposons (Table 1) lead to the inescapable conclusion that RNA information has been inserted into a staggered chromosomal break; however, the agreement ends there. Given the extraordinary variety of RNA species that can serve as substrates for reverse transcription, as well as the variety of retrogene structures derived from them (Table 1), it is appropriate to ask whether one mechanism or many is responsible for retroposition. In the absence of proof to the contrary, we are tempted to believe that the known kinds of nonviral retroposition can be reconciled with a single mechanism, and our discussion reflects that prejudice.

Is the RNA copied into cDNA before insertion (21, 71, 124), or is the RNA inserted into the chromosome and then copied into cDNA in situ (2, 17, 26, 119)? Models of the second type are acceptable for retropseudogenes in which the 3' end of the RNA sequence overlaps the downstream direct repeat, but are inadequate for the many retropseudogenes where overlap is not observed. In addition, since incompletely processed nuclear RNA species might be expected to serve as substrates for reverse transcription in situ, models of the second type also suggest that unprocessed retropseudogenes should be relatively abundant. However, only two partially processed retrogenes are known out of the many

documented examples: the functional rat preproinsulin I retrogene which retains a single intron (40), and a human U2 snRNA pseudogene which retains the 10 extra nucleotides at the 3' end characteristic of preU2 snRNA (23, 171). In fact, the absence of introns from processed genes may imply that cytoplasmic but not nuclear RNA usually serves as the template for reverse transcription. To explain how small nuclear RNAs such as U1 and U2 could give rise to large families of processed retropseudogenes, we suggest that the snRNA may be reverse transcribed after export to the cytoplasm, but before assembly into a karyophilic snRNP particle (172). Similarly, transcripts of most SINEs, LINEs, and viral retroposons, although concentrated in the nucleus, may not be entirely excluded from the cytoplasm (4, 5, 96, 104, 172a).

What primes synthesis of the cDNA? If the RNA is copied into cDNA before insertion, reverse transcription might be self-primed for RNAs whose 3' end is complementary to an internal RNA sequence (proposed for SINEs in Ref. 124 and for certain snRNAs in Ref. 25); however, bimolecular priming by a second cellular RNA or by a DNA fragment must be invoked for all other RNA species (173). If the RNA is reverse transcribed in situ, the chromosome itself would presumably serve as the primer (2, 17, 23, 119).

What is the source of the reverse transcriptase activity? This vexing question remains unanswered (2, 17, 21, 23, 119, 124) and further speculation is unlikely to clarify the issue. We have already mentioned that retropseudogenes are relatively rare in *Drosophila* (121) and apparently absent in yeast (Table 1), despite the presence of endogenous viral retroposons that encode reverse transcriptases (4, 5). These reverse transcriptases may be specific for the viral templates. We also mentioned that retrogenes are rare or absent in birds, although endogenous and exogenous retroviruses abound in these species today. Thus the mere presence of reverse transcriptase activity in an organism is not sufficient to generate retrogenes; the right kind of reverse transcriptase must be present in the right part of the right cells at the right time in development.

We are intrigued by the possibility that differences in gametogenesis could account for the abundance of nonviral retrogenes in mammals compared to chickens, amphibia, or *Drosophila*. Spermatogenesis is very similar in all these species, and is therefore unlikely to make a significant contribution to the observed frequency of retroposition. In keeping with this conclusion, the three known mouse cytochrome *c* retrogenes are not derived from the testis-specific isozyme, but rather from the somatic isozyme which is presumably expressed in oocytes (59). In contrast, mammalian oogenesis differs from that in other organisms by prolonging the lampbrush stage (the diplotene of meiotic prophase) from birth to ovulation. This state of relatively suspended animation may last for as long as 40 years in humans, but for only several months in amphibians and for less than three weeks in birds; the meroistic oocytes of *Drosophila* skip the lampbrush stage altogether because nurse cells assume the

role of lampbrush chromosomes (174). If retroposition in mammals does occur predominantly (or even exclusively) in the female germline, retrogenes may be underrepresented on the Y chromosome.

Why does the staggered break usually vary in length from 7 to 21 bp, although the direct repeats can be as short as 0 bp (22, 25) and as long as 41 bp (40)? The variable length of the direct repeats confounds any simple argument that a single protein (or protein complex) could be responsible for integration (but see Ref. 24) as is thought to be the case for viral retroposons (7). However, although the direct repeats cannot be reconciled with a single consensus sequence, the strand corresponding to the RNA sequence is often rich in adenine (2, 26a) especially at the 5' end (17, 71, 175). Perhaps random nicking in A+T-rich regions produces staggered breaks with either 5' or 3' extended ends, the former serving as retroposon insertion sites, the latter as substrates for the DNA tailing reactions proposed by Rogers (1, 2) to account for "zero option" insertions. The preponderance of direct repeats with lengths between 7 and 21 bp could be explained if larger breaks were efficiently repaired, while smaller breaks were often fatal. Random nicking in A+T-rich regions would also explain why A-rich chromosomal sites often serve as hotspots for multiple insertions as first noted by Rogers (2). Thus the first retroposition event would not inactivate the target site and might in fact activate it, perhaps by introducing (or at least duplicating) oligo(dA) tracts. This, as Rogers (2) was the first to emphasize, must surely be the origin of the many mobile composite retroposons such as the human Alu, the galago Type II Alu (148), the artiodactyl C-BCS family (123), the mouse B2-OBY family (2), and the human 7SK-Type I Alu family (clone 11 of Ref. 34).

Why does the RNA sequence in many retrogenes exhibit significant overlap with the downstream direct repeat? This has been variously attributed to in situ reverse transcription of the RNA primed by an extended 3' end at the staggered break (2, 17, 23, 119), or to reverse transcription followed by hybridization of the 5' end of the cDNA to an extended 5' end at the staggered break (24, 71). However, the postulated hybridization cannot be obligatory in all cases, since the RNA sequence in many retrogenes fails to overlap the downstream direct repeat. Perhaps the 3' end of a cDNA (corresponding to the 5' end of the RNA sequence) is first attached to an extended 5' end on the upstream side of the staggered break; optional hybridization between the 5' end of the cDNA and the extended 5' end of the downstream side of the chromosomal break would then determine whether, and to what degree, the 3' end of the retrogene would be truncated (24). This would provide a natural explanation for the puzzling observation that retroposons with A-rich 3' tails (Alu elements and processed retropseudogenes derived from polyadenylated mRNAs) are rarely if ever truncated at the 3' end, whereas retropseudogenes derived from tRNA (35), 7SL RNA (33), 7SK RNA (34), and mature snRNAs (U1, U2, U3, U4, and U6)

are almost always truncated at the 3' end, except in those rare instances where the pseudogene appears to be derived from an aberrantly polyadenylated molecule of the RNA (21–23, 28, 33). In contrast, occasional examples of 5' truncation (29, 30, 33, 42, 79, 80) could reflect incomplete reverse transcription of a normally initiated RNA, complete reverse transcription of an aberrantly initiated RNA, or aberrant splicing. Given the large number of processed retropseudogenes spanning the complete mRNA sequence, aberrant initiation or splicing appear to be the more likely explanations for the 5' truncation of tissue-specific immunoglobulin mRNAs (42, 79, 80).

A WORD ABOUT THE EVOLUTION OF SINES AND LINES

Why are Alu and L1 sequences relatively homogeneous within each species, but often characteristically different between species? One possibility is that these repetitive sequences are subject to extensive gene conversion, so that each family is constrained to coevolve as a unit (see Ref. 176 for a general discussion of coevolution). However, recent work (138, 177) suggests that none of the Alu sequences in the chimpanzee beta globin cluster has been converted since the separation of the chimpanzee and human lineages. This result agrees well with the general observation that gene conversions are rare in mammals, except at hotspots such as certain immunoglobulin and histocompatibility loci (e.g. see Ref. 178). Another possibility is that Alu and L1 sequences might coevolve by frequent reciprocal recombination; however, we can rule this out because the observed stability of mammalian karyotypes is not compatible with even a low level of chromosomal translocation. A third possibility is that Alu and L1 sequences in the human and mouse genomes have reached a steady state, so that old sequences are continually replaced by new ones. However, a steady state would require a mechanism for specifically removing these sequences from the genome, and there is no real evidence for this. A steady state would also imply that some Alu and L1 sequences should be far older than average, while in fact divergence within the Alu and L1 families is relatively monodisperse. Most Alu elements diverge from the consensus by 8–20%, and only a very few of the more than 50 available sequences diverge by as much as 28% (41; P. L. Deininger, unpublished calculations). Similarly, many mouse L1 sequences differ from each other by less than 5% (90, 179).

If Alu and L1 sequences do not coevolve by gene conversion or recombination, and are not reliably eliminated from the genome before they degenerate, what accounts for the low level of intraspecific divergence and the existence of characteristic interspecific differences? We believe the evidence favors a fourth possibility, namely, that the expansion of these repetitive sequences within the genome began quite recently in evolutionary time, and probably continues

today. We therefore suggest that very few copies of the original Alu sequence were present in the ancestral primate some 55 Myr ago. The comparative youth of the family would then explain not only the relative lack of sequence divergence within it, but also the observed species-specific differences between the galago (prosimian) and human Alu families (147). Mutations occurring when there were few Alu sequences in the genome ("founder sequences") would be able to alter the consensus slightly in the two species. However, once the copy number of the Alu family increased, it would be difficult for subsequent mutations to alter the consensus. Thus, the Alu family consensus sequences for new world monkeys and humans do not differ significantly (147), suggesting that the copy number of the Alu family must have been high enough to stabilize the consensus before these species diverged. Similar arguments can be made for the mouse L1 family. Assuming that mouse L1 sequences were present at very low copy number 12 Myr ago, significant species-specific differences between the L1 families of *Mus caroli* and *Mus platythrix* (179) can be explained by expansion of different L1 "founder sequences" within each genome. The more subtle species-specific differences between the L1 families of *Mus domesticus* and *M. caroli* suggest that these closely related species diverged when the copy number of the L1 family was sufficiently high for the consensus to resist change.

Although the Alu and L1 families expanded quite recently in evolutionary time, we do not really know whether these families are continuing to expand today. We also do not know whether the Alu family has increased exponentially (as expected if each new Alu element can retropose as efficiently as its parent) or if only a small number of Alu elements can function as active retroposons (perhaps because 5' flanking regions strongly influence the efficiency of transcription; see Ref. 134). Thus the expansion of the Alu and L1 families may more nearly approximate a linear than an exponential function. Finally, we do not know whether there is a relatively sharp "saturation value," beyond which further expansion of a repetitive DNA sequence family would become subject to a disproportionately strong negative selection.

CONCLUDING REMARKS

Eukaryotic genomes are not as tidy as the genomes of prokaryotes. Introns, huge intergenic regions, satellite sequences, pseudogenes, and many families of transposable elements suggest that excess DNA is often not subject to a strong negative selection. Retroposition helps to maintain the complexity and fluidity of eukaryotic genomes by generating genes, pseudogenes, transposable elements, and novel combinations of DNA sequences. The resulting wealth of genetic variation serves as raw material for positive selection, as well as for negative selection and neutral drift. However, all retroposons are insertional

mutagens, and transposable elements in particular can increase rapidly if unchecked by strong negative selection (92). Thus retroposition, like all other forms of genetic variation, is both "good" and "bad" for the organism. Before we can reach a more sophisticated understanding of the role of retroposition in evolution, we will need to know more about the detailed molecular mechanism(s) of retroposition.

Literature Cited

1. Rogers, J. 1983. *Nature* 305:101–2
2. Rogers, J. 1985. *Int. Rev. Cytol.* 93:187–279
3. Weiss, R., Teich, N., Varmus, H., Coffin, J., eds. 1985. *RNA Tumor Viruses*. Cold Spring Harbor, NY: Cold Spring Harbor Lab. 2nd. ed.
4. Mount, S. M., Rubin, G. M. 1985. *Mol. Cell. Biol.* 5:1630–38
5. Boeke, J. D., Garfinkel, D. J., Styles, C. A., Fink, G. R. 1985. *Cell* 40:491–500
6. Baltimore, D. 1985. *Cell* 40:481–82
7. Panganiban, A. T. 1985. *Cell* 42:5–6
8. Orgel, L. E., Crick, F. H. C. 1980. *Nature* 284:604–7
9. Doolittle, W. F., Sapienza, C. 1980. *Nature* 284:601–3

9a. Wichman, H. A., Potter, S. S., Pine, D. S. 1985. *Nature* 317:77–81

10. Hartl, D. L., Dykhuizen, D. E., Miller, R. D., Green, L., de Framond, J. 1983. *Cell* 35:503–10
11. Jelinek, W. R., Schmid, C. W. 1982. *Ann. Rev. Biochem.* 51:813–44
12. Schmid, C. W., Jelinek, W. R. 1982. *Science* 216:1065–70
13. Singer, M. F. 1982. *Cell* 28:433–34
14. Singer, M. F. 1982. *Int. Rev. Cytol.* 76:67–112
15. Singer, M. F., Skowronski, J. 1985. *Trends Biochem. Sci.* 10:119–22
16. Sharp, P. A. 1983. *Nature* 301:471–72
17. Vanin, E. F. 1984. *Biochem. Biophys. Acta* 782:231–41
18. Jeffreys, A. J., Harris, S. 1984. *Bioessays* 1:253–58
19. Johns, M. A., Mottinger, J., Freeling, M. 1985. *EMBO J.* 4:1093–102
20. Paulson, K. E., Deka, N., Schmid, C. W., Misra, R., Schindler, C. W., et al. 1985. *Nature* 316:359–61
21. Van Arsdell, S. W., Denison, R. A., Bernstein, L. B., Weiner, A. M., Manser, T., Gesteland, R. F. 1981. *Cell* 26:11–17
22. Denison, R. A., Weiner, A. M. 1982. *Mol. Cell. Biol.* 2:815–28

22a. Eickbush, T. H., Robins, B. 1985. *EMBO J.* 4:2281–85

23. Hammarstrom, K., Westin, G., Bark, C., Zabielski, J., Pettersson, U. 1984. *J. Mol. Biol.* 179:157–69
24. Van Arsdell, S. W., Weiner, A. M. 1984. *Nucleic Acids Res.* 12:1463–71
25. Bernstein, L. B., Mount, S. M., Weiner, A. M. 1983. *Cell* 32:461–72
26. Hammarstrom, K., Westin, G., Pettersson, U. 1982. *EMBO J.* 1:737–39

26a. Bark, C., Hammarstrom, K., Westin, G., Pettersson, U. 1985. *Mol. Cell. Biol.* 5:943–48

27. Theissen, H., Rinke, J., Traver, C. N., Luhrmann, R., Appel, B. 1985. *Gene*. In press
28. Ohshima, Y., Okada, N., Tani, T., Itoh, Y., Itoh, M. 1981. *Nucleic Acids Res.* 9:5145–58
29. Reddy, R., Henning, D., Chirala, S., Rothblum, L., Wright, D., Busch, H. 1985. *J. Biol. Chem.* 260:5715–19
30. Stroke, I. L., Weiner, A. M. 1985. *J. Mol. Biol.* 184:183–93
31. Shafit-Zagardo, B., Brown, F. L., Zavodny, P. J., Maio, J. J. 1983. *Nature* 304:277–80
32. Heller, D., Jackson, M., Leinwand, L. 1984. *J. Mol. Biol.* 173:419–36
33. Ullu, E., Weiner, A. M. 1984. *EMBO J.* 3:3303–10
34. Murphy, S., Altruda, F., Ullu, E., Tripodi, M., Silengo, L., Melli, M. 1984. *J. Mol. Biol.* 177:575–90
35. Pratt, K., Eden, F. C., You, K. H., O'Neill, V. A., Hatfield, D. 1985. *Nucleic Acids Res.* 13:4765–75

35a. Hasan, G., Turner, M. J., Cordingley, J. S. 1984. *Cell* 37:333–41

36. Ono, M., Toh, H., Miyata, T., Awaya, T. 1985. *J. Virol.* 55:387–94
37. Stoye, J., Coffin, J. 1985. See Ref. 3, pp. 357–404
38. Varmus, H. E., Swanstrom, R. 1985. See Ref. 3, pp. 75–134
39. Cappello, J., Handelsman, K., Lodish, H. F. 1985. *Cell* 43:105–15
40. Soares, M. B., Schon, E., Henderson, A., Karathanasis, S. K., Cate, R., et al. 1985. *Mol. Cell. Biol.* 5:2090–103
41. Schmid, C. W., Shen, C.-K. J. 1986. *Molecular Evolutionary Genetics*, ed. R.

J. MacIntyre, pp. 323–58. New York: Plenum
42. Hollis, G. F., Hieter, P. A., McBride, O. W., Swan, D., Leder, P. 1982. *Nature* 296:321–25
43. Brown, J. R., Daar, I. O., Krug, J. R., Maquat, L. E. 1985. *Mol. Cell. Biol.* 5:1694–706
44. Freytag, S. O., Beaudet, A. L., Bock, H. G., O'Brien, W. E. 1984. *Mol. Cell. Biol.* 4:1978–84
45. Freytag, S. O., Bock, H. G., Beaudet, A. L., O'Brien, W. E. 1984. *J. Biol. Chem.* 259:3160–66
46. Su, T. S., Nussbaum, R. L., Airhart, S., Ledbetter, D. H., Mohandas, T., et al. 1984. *Am. J. Hum. Genet.* 36:954–64
47. Pani, K., Singer-Sam, J., Munns, M., Yoshida, A. 1985. *Gene* 35:11–18
48. Piechaczyk, M., Blanchard, J. M., Riaad-El-Sabouty, S., Dani, C., Marty, L., Jeanteur, P. 1984. *Nature* 312:469–71
49. Hanauer, A., Mandel, J. L. 1984. *EMBO J.* 3:2627–33
50. Benham, F. J., Hodgkinson, S., Davis, K. E. 1984. *EMBO J.* 3:2635–40
51. Tsujibo, H., Tiano, H. F., Li, S. S. 1985. *Eur. J. Biochem.* 147:9–15
52. Anagnou, N. P., O'Brien, S. J., Shimada, T., Nash, W. G., Chan, M. J., Nienhuis, A. W. 1984. *Proc. Natl. Acad. Sci. USA* 81:5170–74
52a. Maurer, B. J., Carlock, L., Wasmuth, J., Attardi, G. 1985. *Somatic Cell. Mol. Genet.* 11:79–85
53. Shimada, T., Chen, M. J., Nienhuis, A. W. 1984. *Gene* 31:1–8
54. LeBeau, M. M., Diaz, M. O., Karin, M., Rowley, J. D. 1985. *Nature* 313: 709–11
55. Karin, M., Eddy, R. L., Henry, W. M., Haley, L. L., Byers, M. G., Shows, T. B. 1984. *Proc. Natl. Acad. Sci. USA* 81:5494–98
56. Karin, M., Richards, R. I. 1982. *Nature* 299:797–802
56a. Karin, M., Richards, R. I. 1984. *Environ. Health Perspect.* 54:111–15
57. Varshney, U., Gedamu, L. 1984. *Gene* 31:135–45
58. Schmidt, C. J., Hamer, D. H., McBride, O. W. 1984. *Science* 224:1104–6
59. Linbach, K. J., Wu, R. 1985. *Nucleic Acids Res.* 13:617–30
60. Scarpulla, R. C. 1984. *Mol. Cell. Biol.* 4:2279–88
61. Scarpulla, R. C. 1985. *Nucleic Acids Res.* 13:763–75
62. Feagin, J. E., Setzer, D. R., Schimke, R. P. 1983. *J. Biol. Chem.* 258:2480–87
63. Dudov, K. P., Perry, R. P. 1984. *Cell* 37:457–68
64. Trowsdale, J., Kelly, A., Lee, J., Carson, S., Austin, P., Travers, P. 1984. *Cell* 38:241–49
65. Wiedemann, L. M., Perry, R. P. 1984. *Mol. Cell. Biol.* 4:2518–28
66. Klein, A., Mcyuhas, O. 1984. *Nucleic Acids Res.* 12:3763–76
67. Peled-Yalif, E., Cohen-Binder, I., Meyuhas, O. 1984. *Gene* 29:157–66
68. Lemischka, I., Sharp, P. A. 1982. *Nature* 300:330–35
69. Lee, M. G., Lewis, S. A., Wilde, C. D., Cowan, N. J. 1983. *Cell* 33:477–87
70. Leavitt, J., Gunning, P., Porreca, P., Ng, S.-Y., Lin, C.-S., Kedes, L. 1984. *Mol. Cell. Biol.* 4:1961–69
71. Moos, M., Gallwitz, D. 1983. *EMBO J.* 2:757–61
72. Tokunaga, K., Yoda, K., Sakiyama, S. 1985. *Nucleic Acids Res.* 13:3031–42
73. Robert, B., Daubas, P., Akimenko, M.-A., Cohen, A., Garner, I., et al. 1984. *Cell* 39:129–40
74. Maclead, A. R., Talbot, K. 1983. *J. Mol. Biol.* 167:523–37
75. Vasseur, M., Duprey, P., Brulet, P., Jacob, F. 1985. *Proc. Natl. Acad. Sci. USA* 82:1155–59
76. Zakut-Houri, R., Oren, M., Bienz, B., Lavie, V., Hazum, S., Givol, D. 1983. *Nature* 306:594–97
77. Takahashi, H., Mishina, M., Numa, S. 1983. *FEBS Lett.* 156:67–71
78. Uhler, M., Herbert, E., D'Eustachio, P., Ruddle, F. H. 1983. *J. Biol. Chem.* 258:9444–53
79. Battey, J., Max, E. E., McBride, W. O., Swan, D., Leder, P. 1982. *Proc. Natl. Acad. Sci. USA* 79:5956–60
80. Ueda, S., Nakai, S., Nishida, Y., Hisajima, H., Honjo, T. 1982. *EMBO J.* 1:1539–44
81. Miyoshi, J., Kagimoto, M., Soeda, E., Sakaki, Y. 1984. *Nucleic Acids Res.* 12:1821–28
82. McGrath, J. P., Capon, D. J., Smith, D. H., Chen, E. Y., Seeburg, P. H., et al. 1983. *Nature* 304:501–6
83. Bonner, T., O'Brien, S. J., Nash, W. G., Rapp, U. R., Morton, C. C., Leder, P. 1984. *Science* 223:71–74
84. Stein, J. P., Munjaal, R. P., Lagace, L., Lai, E. C., O'Malley, B. W., Means, A. R. 1983. *Proc. Natl. Acad. Sci. USA* 80:6485–89
85. Lomedico, P., Rosenthal, N., Efstratiadis, A., Gilbert, W., Kolodner, R., Tizard, R. 1979. *Cell* 18:545–58
86. Chan, S. J., Episkopou, V., Zeitlin, S., Karathanasis, S., MacKrell, A., et al. 1984. *Proc. Natl. Acad. Sci. USA* 81: 5046–50

87. Antoine, M., Niessing, J. 1985. *Nature* 310:795–98
88. Romans, P., Firtel, R. A. 1985. *J. Mol. Biol.* 183:311–26
89. Manuelidis, L. 1982. *Nucleic Acids Res.* 10:3211–19
90. Martin, S. L., Voliva, C. F., Burton, F. H., Edgell, M. H., Hutchison, C. A. III. 1984. *Proc. Natl. Acad. Sci. USA* 81: 2308–12
91. Potter, S. S. 1984. *Proc. Natl. Acad. Sci. USA* 81:1012–16
92. Soares, M. B., Schon, E., Efstratiadis, A. 1985. *J. Mol. Evol.* 22:117–33
93. Grimaldi, G., Skowronski, J., Singer, M. F. 1984. *EMBO J.* 3:1753–59
94. Lerman, M. I., Thayer, R. E., Singer, M. F. 1983. *Proc. Natl. Acad. Sci. USA* 80:3966–70
95. Thayer, R. E., Singer, M. F. 1983. *Mol. Cell. Biol.* 3:967–73
96. DiGiovanni, L., Haynes, S. R., Misra, R., Jelinek, W. R. 1983. *Proc. Natl. Acad. Sci. USA* 80:6533–37
97. Miyake, T., Migita, K., Sakaki, Y. 1983. *Nucleic Acids Res.* 11:6837–46
98. Nomiyama, H., Tsuzuki, T., Wasasugi, S., Fukado, M., Shumada, K. 1984. *Nucleic Acids Res.* 12:5225–34
99. Economou-Pachnis, A., Lohse, M. A., Furano, A. V., Tsichlis, P. N. 1985. *Proc. Natl. Acad. Sci. USA* 82:2857–61
100. Burton, F. H., Loeb, D. D., Chao, S. F., Hutchison, C. A. III, Edgell, M. H. 1985. *Nucleic Acids Res.* 13:5071–84
101. Shyman, S., Weaver, S. 1985. *Nucleic Acids Res.* 13:5085–93
102. Skowronski, J., Singer, M. F. 1985. *Proc. Natl. Acad. Sci. USA* 82:6050–54
103. Fanning, T. G. 1983. *Nucleic Acids Res.* 11:5073–91
104. Kole, L. B., Haynes, S. R., Jelinek, W. R. 1983. *J. Mol. Biol.* 165:257–86
105. Voliva, C. F., Jahn, C. L., Comer, M. B., Hutchison, C. A. III, Edgell, M. H. 1983. *Nucleic Acids Res.* 11:8847–59
106. Bennett, K. L., Hill, R. E., Pietras, D. F., Woodworth-Gutai, M., Kane-Haas, C., et al. 1984. *Mol. Cell. Biol.* 4:1561–71
107. Schindler, C. W., Rush, M. G. 1985. *J. Mol. Biol.* 131:161–73
108. Jones, R. S., Potter, S. S. 1985. *Proc. Natl. Acad. Sci. USA* 82:1989–93
109. Fujimoto, S., Tsuda, T., Toda, M., Yamagishi, H. 1985. *Proc. Natl. Acad. Sci. USA* 82:2072–76
109a. Riabowol, K., Shmookler Reis, R. J., Goldstein, S. 1985. *Nucleic Acids Res.* 13:5563–84
110. Whitney, F. R., Furano, A. V. 1984. *J. Biol. Chem.* 259:10:481–92
111. Schmeckpeper, B. J., Scott, A. F., Smith, K. D. 1984. *J. Biol. Chem.* 259:1218–25
112. Jackson, M., Heller, D., Leinwand, L. 1985. *Nucleic Acids Res.* 13:3389–403
113. Manley, J. L., Colozzo, M. T. 1982. *Nature* 300:376–79
114. Lueders, K. K., Fewell, J. W., Kuff, E. L., Koch, T. 1984. *Mol. Cell. Biol.* 4:2128–35
115. Gil, A., Proudfoot, N. J. 1984. *Nature* 312:473–74
116. McDevitt, M. A., Imperiale, M. J., Ali, H., Nevins, J. P. 1984. *Cell* 37:993–99
117. Sadofsky, M., Alwine, J. C. 1984. *Mol. Cell. Biol.* 4:1460–68
118. Patarca, R., Haseltine, W. A. 1984. *Nature* 309:728
119. Voliva, C. F., Martin, S. L., Hutchison, C. A., Edgell, M. H. 1984. *J. Mol. Biol.* 178:795–813
120. Junghans, R. P., Boone, L. R., Skalka, A. M. 1982. *Cell* 30:53–62
121. DiNocera, P. P., Digan, M. E., Dawid, I. B. 1983. *J. Mol. Biol.* 168:715–27
122. Watanabe, Y., Tsukada, T., Notake, M., Nakanishi, S., Numa, S. 1982. *Nucleic Acids Res.* 10:1459–92
123. Spence, S. E., Young, R. M., Garner, K. J., Lingrel, J. B. 1985. *Nucleic Acids Res.* 13:2171–86
124. Jagadeeswaran, P., Forget, B. G., Weissman, S. M. 1981. *Cell* 26:141–42
125. Rinehart, F. P., Ritch, T. G., Deininger, P. L., Schmid, C. W. 1981. *Biochemistry* 20:3003–10
126. Sun, L., Paulson, K. E., Schmid, C. W., Kadyk, L., Leinwand, L. 1984. *Nucleic Acids Res.* 12:2669–91
127. Fowlkes, D., Shenk, T. 1980. *Cell* 22:405–13
128. Fuhrman, S., Deininger, P., LaPorte, P., Friedmann, T., Geiduschek, E. P. 1981. *Nucleic Acids Res.* 9:6439–56
129. Perez-Stable, C., Ayres, T., Shen, C.-K. J. 1984. *Proc. Natl. Acad. Sci. USA* 81:5291–95
130. Paolella, G., Lucero, M. A., Murphy, M. H., Baralle, F. E. 1983. *EMBO J.* 2:691–96
131. Duncan, C. H., Jagadeeswaran, P., Wang, R. C., Weissman, S. M. 1981. *Gene* 13:185–96
132. Deleted in proof
133. Ullu, E., Murphy, S., Melli, M. 1982. *Cell* 29:195–202
134. Ullu, E., Weiner, A. M. 1985. *Nature* 318:371–74
135. Haynes, S. R., Jelinek, W. R. 1981. *Proc. Natl. Acad. Sci. USA* 78:6130–34
136. Smith, G. P. 1976. *Science* 191:528–35
137. Hamada, H., Petrino, M. G., Kakunga, T. 1982. *Proc. Natl. Acad. Sci. USA* 79:6465–69

138. Sawada, I., Willard, C., Shen, C.-K. J., Chapman, B., Wilson, A., Schmid, C. W. 1985. *J. Mol. Evol.* 22:316–22
139. Tsukada, T., Watanabe, Y., Nakai, Y., Imura, H., Nakanishi, S., Numa, S. 1982. *Nucleic Acids Res.* 10:1471–76
140. Saffer, J. D., Lerman, M. I. 1983. *Mol. Cell. Biol.* 3:960–64
141. Walter, P., Blobel, G. 1982. *Nature* 299:691–98
142. Gundelfinger, E. D., DiCarlo, M., Zopf, D., Melli, M. 1984. *EMBO J.* 3:2325–32
143. Ullu, E., Tschudi, C. 1984. *Nature* 312:171–72
144. Ullu, E., Esposito, V., Melli, M. 1982. *J. Mol. Biol.* 161:195–201
145. Carlson, D. P., Ross, J. 1983. *Cell* 34:857–64
145a. Efstratiadis, A., Posakony, J. W., Maniatis, T., Lawn, R. M., O'Connell, C., et al. 1980. *Cell* 21:653–68
145b. Degen, S. J. F., MacGillivray, R. T. A., Davie, E. W. 1983. *Biochemistry* 22:2087–97
145c. Lee, M. G.-S., Loomis, C., Cowan, N. 1984. *Nucleic Acids Res.* 12:5823–36
145d. Flemington, E., Traina-Dorge, V., Slagel, V., Bradshaw, H., Deininger, P. L. Submitted for publication
146. Gundelfinger, E. D., Krause, E., Melli, M., Dobberstein, B. 1983. *Nucleic Acids Res.* 11:7364–75
147. Daniels, G. R., Fox, G. M., Loewensteiner, D., Schmid, C. W., Deininger, P. L. 1983. *Nucleic Acids Res.* 11:7579–93
148. Daniels, G. R., Deininger, P. L. 1983. *Nucleic Acids Res.* 11:7595–10
149. Daniels, G. R., Deininger, P. L. 1985. *Nature* 317:819–22
150. Lawrence, C. B., McDonnell, D. P., Ramsey, W. J. 1985. *Nucleic Acids Res.* 13:4239–51
151. Sakamoto, K., Okada, N. 1985. *J. Mol. Evol.* 22:134–40
152. Krayev, A. S., Markusheva, T. V., Kramerov, D. A., Ryskov, A. P., Skryabin, K. G., et al. 1982. *Nucleic Acids Res.* 10:7461–75
153. Sutcliffe, J. G., Milner, R. J., Bloom, F. E. Lerner, R. A. 1982. *Proc. Natl. Acad. Sci. USA* 79:4942–46
154. Cheng, J.-F., Printz, R., Callaghan, T., Shuey, D., Hardison, R. C. 1984. *J. Mol. Biol.* 176:1–20
155. Schimenti, J. C., Duncan, C. H. 1984. *Nucleic Acids Res.* 12:1641–55
156. Matsumoto, K., Murakami, K., Okada, N. 1984. *Biochem. Biophys. Res. Commun.* 124:514–22
157. Endoh, H., Okada, N. 1986. *Proc. Natl. Acad. Sci. USA.* 83:251–55
158. Blin, N., Weber, T., Alonso, A. 1983. *Nucleic Acids Res.* 11:1375–88
159. Ackerman, E. J. 1983. *EMBO J.* 2:1417–22
160. Stumph, W. E., Baez, M., Beattie, W. G., Tsai, M.-J., O'Malley, B. W. 1983. *Biochemistry* 22:306–15
161. Sutcliffe, J. G., Milner, R. J., Gottesfeld, J. M., Lerner, R. A. 1984. *Nature* 308:237–41
162. Murphy, D., Brickell, P. M., Latchman, D. S., Willison, K., Rigby, P. W. J. 1983. *Cell* 35:865–71
163. Brickell, P. M., Latchman, D. S., Murphy, D., Willison, K., Rigby, P. W. J. 1983. *Nature* 306:756–60
164. Barta, A., Richards, R. I., Baxter, J. D., Shine, J. 1981. *Proc. Natl. Acad. Sci. USA* 78:4867–71
165. Milner, R. J., Bloom, F. E., Lai, C., Lerner, R. A., Sutcliffe, J. G. 1984. *Proc. Natl. Acad. Sci. USA* 81:713–17
166. Minarovits, J., Kovacs, Z., Foldes, I. 1984. *FEBS Lett.* 174:208–10
167. McKinnon, R. D., Shinnick, T. M., Sutcliffe, J. G. Submitted for publication
167a. Owens, G. P., Chaudhari, N., Hahn, W. E. 1985. *Science* 229:1263–65
168. Vasseur, M., Condamine, H., Duprey, P. 1985. *EMBO J.* 4:1749–53
169. Scott, M. R. D., Westphal, K.-H., Rigby, P. W. J. 1983. *Cell* 34:557–67
170. Singh, K., Carey, M., Saragosti, S., Botchan, M. 1985. *Nature* 314:553–56
171. Yuo, C.-Y., Ares, M. Jr., Weiner, A. M. 1985. *Cell* 42:193–202
172. Mattaj, I. W., DeRobertis, E. M. 1985. *Cell* 40:111–18
172a. Adeniyi-Jones, S., Zasloff, M. 1985. *Nature* 317:81–84
173. Chen, P.-J., Cywinski, A., Taylor, J. M. 1985. *J. Virol.* 54:278–84
174. Davidson, E. H. 1976. *Gene Activity in Early Development*, pp. 319–49. New York: Academic
175. Daniels, G. R., Deininger, P. L. 1985. *Nucleic Acids Res.* 13:8939–54
176. Dover, G. A., Flavell, R. B. 1984. *Cell* 38:622–23
177. Sawada, I., Beal, M. P., Shen, C.-K. J., Chapman, B., Wilson, A. C., Schmid, C. 1983. *Nucleic Acids Res.* 11:8087–101
178. Jeffreys, A. J., Wilson, V., Thein, S. L. 1985. *Nature* 316:76–79
179. Martin, S. L., Voliva, C. F., Hardies, S. C., Edgell, M. H., Hutchison, C. A. III. 1985. *Mol. Biol. Evol.* 2:127–40