HOW MANY PROCESSED PSEUDOGENES ARE ACCUMULATED IN A GENE FAMILY?

JAMES BRUCE WALSH[1]

Department of Molecular Genetics and Cell Biology, The University of Chicago, Chicago, Illinois 60637

Manuscript received October 30, 1984
Revised copy accepted February 19, 1985

ABSTRACT

A simple kinetic model is developed that describes the accumulation of processed pseudogenes in a functional gene family. Insertion of new pseudogenes occurs at rate ν per gene and is countered by spontaneous deletion (at rate δ per DNA segment) of segments containing processed pseudogenes. If there are k functional genes in a gene family, the equilibrium number of processed pseudogenes is $k(\nu/\delta)$, and the percentage of functional genes in the gene family at equilibrium is $1/[1 + (\nu/\delta)]$. ν/δ values estimated for five gene families ranged from 1.7 to 15. This fairly narrow range suggests that the rates of formation and deletion of processed pseudogenes may be positively correlated for these families. If δ is sufficiently large relative to the per nucleotide mutation rate μ ($\delta > 20\mu$), processed pseudogenes will show high homology with each other, even in the absence of gene conversion between pseudogenes. We argue that formation of processed pseudogenes may share common pathways with transposable elements and retroviruses, creating the potential for correlated responses in the evolution of processed pseudogenes due to direct selection for control of transposable elements and/or retroviruses. Finally, we discuss the nature of the selective forces that may act directly or indirectly to influence the evolution of processed pseudogenes.

Anything produced by evolution is bound to be a bit of a mess—S. Brenner

GENES that exist in multicopy families appear to be the rule rather than the exception for higher eukaryotes. Often many members of such families appear (from DNA sequence analysis) to be nonfunctional, these inactivated members being referred to as pseudogenes. Since their initial discovery (reviewed by PROUDFOOT 1980) subsequent molecular characterization has demonstrated that two distinct classes of pseudogenes exist: "traditional" pseudogenes and processed pseudogenes. Traditional pseudogenes, as exemplified in the globin gene families, appear to have arisen by gene duplication and been subsequently silenced by point mutations, small insertions and deletions (LITTLE 1982; PROUDFOOT 1980). Traditional pseudogenes are usually adjacent to functional copies and often show evidence of being under some form of selec-

[1] Address after September 15, 1985: Department of Genetics, University of California, Davis, California 95616.

tive constraint for several million years following their formation (PROUDFOOT and MANIATIS 1980; LI, GOJOBORI and NEI 1981; GOJOBORI, LI and GRAUR 1982). Processed pseudogenes differ from traditional pseudogenes in both their mode of formation and their evolutionary history following formation. Processed pseudogenes are characterized by a lack of introns and a remnant of a Poly-A tail, are often flanked by short direct repeats and are usually quite dispersed from functional copies, all of which suggests formation by the integration into the germline DNA of a reverse-transcribed processed RNA (SHARP 1983; MARX 1983; R. LEWIN 1983). We detail the molecular biology of processed pseudogenes in the next section. Since the focus of this paper is on processed pseudogenes, we will often simply use pseudogene or occasionally processed gene to denote this class and use traditional pseudogene when referring to the other class.

Processed pseudogenes present something of a dilemma for a cell in that, if a general reverse transcription and insertional mechanism exists, any gene that is active in the germline creates additional genomic DNA by generating the RNA for processed pseudogene formation. In this sense, processed pseudogenes share a copy-number control problem, similar to that for transposable elements (BROOKFIELD 1982; LANGLEY, BROOKFIELD and KAPLAN 1983; CHARLESWORTH and CHARLESWORTH 1983; DOOLITTLE, KIRKWOOD and DEMPSTER 1984). Here, we examine a simple nonselective model for the control of pseudogene copy number by the dynamic equilibrium between formation and spontaneous deletion of processed pseudogenes. We then discuss the possible selective forces that may influence the evolution of the genetic systems governing formation and deletion of processed pseudogenes.

MOLECULAR BIOLOGY OF PROCESSED PSEUDOGENES

Processed pseudogenes appear to be quite common in the mammalian genome. In humans, β-tubulin (WILDE, CROUTHER and COWAN 1982), β-actin (MOOS and GALLWITZ 1983), λ-immunoglobulin (HOLLIS *et al.* 1982), ϵ-immunoglobulin (BATTEY *et al.* 1982), C-Ha-*ras*2 proto-oncogene (MIYOSHI *et al.* 1984) and metallothionein-II (KARIN and RICHARDS 1982) all have processed pseudogenes, and processed pseudogenes are very likely associated with argininosuccinate synthetase (DAIGER, WILDIN and SU 1982). Human small nuclear RNAs (snRNA) U1, U2 and U3 also have numerous processed pseudogenes (DENISON *et al.* 1981; VAN ARSDELL *et al.* 1981; BERNSTEIN, MOUNT and WEINER 1983) as indicated by direct flanking repeats and (in some cases) truncated 3′ ends, presumably resulting from using that end for self-priming of reverse transcriptase (BERNSTEIN *et al.* 1983; VAN ARSDELL and WEINER 1984). Mouse α-globin (NISHIOKA, LEDER and LEDER 1980; VANIN *et al.* 1980, LUEDERS *et al.* 1982), myosin light chain (ROBERT *et al.* 1984), ribosomal proteins L7 (KLEIN and MEYUHAS 1984) and L32 (DUDOV and PERRY 1984); rat α-tubulin (LEMISCHKA and SHARP 1983) and cytochrome *c* (SCARPULLA and WU 1983) also have associated processed pseudogenes. Human HLA-SBβ and mouse dihydrofolate reductase share a highly conserved processed gene found within the introns of these genes (TROWSDALE *et al.* 1984). The high conservation

between the human and mouse pseudogene suggests that the source of the processed gene is an uncharacterized functional gene. Man, mouse, rat, hamster, guinea pig and hare all contain glyceraldehyde 3-phosphate dehydrogenase (GAPDH) processed pseudogenes (PIECHACZYK *et al.* 1984).

Nonmammalian examples of processed pseudogenes are less well documented, with chicken calmodulin (STEIN *et al.* 1983) being the only definitive example. However, the widespread occurrence of "orphons" suggests that processed pseudogenes may be a fairly common feature of eukaryotes. Orphons are "genes which have lost their families" (CHILDS *et al.* 1981) and are dispersed solitary copies of genes that otherwise exist in tandemly repeated clusters. Histone orphons have been found in yeast (*Saccharomyces cerevisiae*), sea urchin (*Lytechinus pictus*) and *Drosophila melanogaster* (CHILDS *et al.* 1981). Although the exact nature of orphons awaits DNA sequence analysis, their dispersed nature is at least consistent with origins as processed pseudogenes.

The formation of a pseudogene from a processed RNA requires two major events: reverse transcription of the RNA and the subsequent insertion of the DNA copy into the genome. We are quite ignorant about many features of this process. Major questions persist as to both the source of the reverse transcriptase and how the RNA is "primed" for reverse transcription. The only currently known source of reverse transcriptase is the *pol* gene of vertebrate retroviruses. For the *pol* gene product to generate a DNA copy from an RNA, it requires a "primer," a short RNA complementary to the 3′ end of the RNA to be reverse transcribed (KORNBERG 1980, pp. 223–225). This primer provides the 5′ end of the DNA copy and is subsequently degraded. Retroviruses solve the primer problem by using host tRNAs as primers and have special sites (primer-binding sites) on the viral RNA which are complementary to regions of the tRNA primer (reviewed by VARMUS 1982).

Where does the reverse transcriptase for processed pseudogenes come from? In mammalian and avian systems, endogenous retroviruses are potential sources (WEINBERG 1980; JAENISCH 1983). It is also likely that some form of reverse transcriptase exists in other eukaryotes as well, given the very striking resemblance between retorviruses and certain classes of transposable elements (TEMIN 1980). Both the *Ty* elements in yeast (ROEDER and FINK 1983) and the *copia*-like elements of Drosophila (FLAVELL and ISH-HOROWICZ 1983; FLAVELL 1984) show strong circumstantial evidence of requiring an RNA intermediate for transposition. If so, then, during the mobilization of these retrovirus-like transposable elements, processed pseudogenes may be created by the presence of reverse transcriptase activity. A very recent report (SAIGO *et al.* 1984) has identified open reading frames in the *copia*-like elements *17.6*, *297* and *412* which have amino acid sequence homology with the avian retrovirus *pol* gene. There is also evidence of potential reverse transcriptase activity in plants (VARMUS 1983), and a putative reverse transcriptase has been isolated and sequenced from cauliflower mosaic virus (GARDER *et al.* 1981). Finally, a formal possibility is that normal cellular polymerases may be subverted to reverse transcriptase activity by complexing with additional factors. The *Qβ*-replicase system of bacteriophage *Qβ* sets a precedent for this by altering part of its

host's translational machinery (elongation factors *Tu* and *Ts* and a ribosomal associated protein) which complexes with a phage-coded protein to form a polymerase that can replicate RNA from an RNA template (LEWIN 1977, pp. 819–824).

Whereas there are several potential sources for reverse transcriptase activity, the problem of a primer is more difficult to deal with. Certain small nuclear RNAs have complementary regions on their 3′ ends which allow them to self-prime, which would result in a truncated 3′ end as observed in some processed pseudogenes (BERNSTEIN, MOUNT and WEINER 1983; VAN ARSDELL and WEINER 1984). Given that many processed pseudogenes show a remnant of the Poly-A tail, priming must have occurred within the tail itself. RNAs that are very U-rich at their 5′ ends could potentially serve as primers, being able to base pair with the poly-A tail and create the necessary region of RNA duplex required for reverse transcriptase activity. Such RNAs would be able to prime any polyadenylated RNA. However, there is no evidence that such RNAs exist. If histone orphons are indeed processed pseudogenes, then an additional priming mechanism is required, as histone mRNAs lack a Poly-A tail (B. LEWIN 1983, p. 150).

We are even more ignorant about the mechanism of insertion of the DNA copy into the genome than we are about how such a copy is generated from an RNA. One characteristic feature of processed genes and transposons is that they are flanked by short direct repeats of genomic DNA. This observation suggests that a staggered double-strand break is created at the target site in the DNA, with subsequent replication creating duplicated repeats that flank the inserted DNA (SHAPIRO 1979). Mutational analysis has shown that the retrovirus *pol* gene region codes for an additional polypeptide in addition to reverse transcriptase. The 5′ region of *pol* codes for reverse transcriptase activity, whereas the 3′ region encodes a protein that is essential for insertion of the viral DNA into the genome (SCHWARTZBERG, COLICELLI and GOFF 1984; DONEHOWER and VARMUS 1984). Thus, the presence of reverse transcriptase is not sufficient for insertion. It is possible that normal cellular enzymes such as topoisomerase II can create pathways for the insertion of short DNA segments. Another possibility is that at least some of the enzymes required for insertion are supplied by endogenous transposable elements.

Finally, the type of RNA polymerase (Pol II *vs.* Pol III) which normally transcribes the gene that created the pseudogene is very important. Pol III processed genes have a completely different behavior than Pol II processed pseudogenes. Pol II (which transcribes mRNA and, hence, all protein-coding genes) has controlling regions and promoters that are 5′ (*i.e.*, outside) of the transcription unit (BREATNACH and CHAMBON 1981; MCKNIGHT and KINSBURY 1982). For these genes, a reverse-transcribed DNA copy contains the necessary translation information but is transcriptionally inactive, as the DNA copy lacks the necessary 5′ controlling regions. Thus, only under very unusual circumstances would we expect Pol II processed genes to create additional processed genes. Pol III genes, however, have their transcriptional control signals contained within their transcriptional unit (KORN 1982; HALL, CLARK and TOC-

CHINI-VALENTINI 1982), although silkworm 5S RNA genes appear to be an exception, requiring an additional element upstream from the Pol III transcriptional unit (MORTON and SPRAGUE 1984). Thus, DNA copies of the transcriptional unit potentially contain all of the information required to produce additional processed genes. ROGERS (1983a) has coined the term *retroposons* for such genes, as they can behave exactly like transposons in terms of generation of new copies which can themselves generate new copies. The *Alu* gene family in humans which makes up approximately 7% of total human DNA (RINEHART *et al* 1981) is a suspected retroposon (JAGADEESWARAN, FORGET and WEISSMAN 1981). Recent work confirms that human *Alu* sequences have functional Pol III promoters (PEREZ-STABLE, AYRES and SHEN 1984). Further work by ULLU and TSCHUDI (1984) strongly suggests that the *Alu* family may well have arisen from retroposons created from processed 7SL RNA genes, a small nuclear RNA essential for protein secretion in most eukaryotes (WALTER, GILMORE and BLOBEL 1984). In what follows, we exclude polymerase III genes, limiting our attention to those gene families whose reverse-transcribed copies cannot produce additional processed genes.

DYNAMICS OF PROCESSED PSEUDOGENE COPY NUMBER

SHARP (1983) has suggested that the creation of new processed pseudogenes would eventually be balanced by spontaneous deletion of DNA segments containing such pseudogenes. Homologous recombination between adjacent repeated sequences provides a mechanism for spontaneous deletions (SHARP 1983). Recombination between such elements would generate a circular DNA containing the sequences bounded by the repeats that would be lost from the cell and leave a chromosome lacking this internal DNA (Figure 1). *Alu* or *Alu*-related sequences occur on average about every 5 kb in mammalian genomes (SCHMID and JELINEK 1982), and regions flanked by *Alu* are unstable (CALABRETTA *et al.* (1982). The amplification of extrachromosomal circular copies of "inter-*Alu*" sequences during serial passage of human fibroblast cells provides further evidence for generation of spontaneous deletions by homologous recombination between dispersed repeated elements (SHMOOKLER REIS *et al.* 1983). Finally, a naturally occurring length variant of the human LDL receptor gene is a deletion created by recombination between adjacent *Alu* elements (LEHRMAN *et al.* 1984). Spontaneous deletions may still occur in the absence of dispersed-repetitive sequences such as *Alu*. The insertion of a processed gene creates direct repeats of flanking DNA, and these short regions of homology may be sufficient for recombination. Evidence from *E. coli* suggests that homologies on the order of 5–10 base pairs are sufficient for creating spontaneous deletions (ALBERTINI *et al.* 1982; EDLUND and NORMARK 1981), and this is about the size of the direct repeats flanking many processed pseudogenes.

Since only a very small fraction of the mammalian genome, perhaps as little as 1% or less (SHARP 1983), seems to participate directly in encoding proteins, we assume that both pseudogene insertions and deletions of DNA segments containing processed pseudogenes behave like neutral alleles. There is evidence

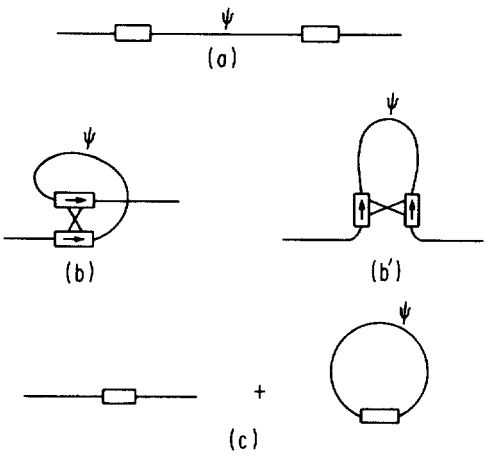

FIGURE 1.—Deletion of processed pseudogenes by recombination between flanking homologous regions. (a) Regions of homology (indicated by the open boxes) surround a DNA segment which contains an inserted processed pseudogene (indicated by ψ). (b) For repeats that are in direct orientation, the DNA loops over, and homologous recombination occurs between, the two repeated units. These units may simply be the direct repeats generated by the insertion of the processed pseudogene or they may be repetitive elements such as *Alu*. (b') For repeats in inverted orientation a slightly different intermediate structure forms to align the repeats for homologous recombination. (c) The resulting products of recombination are a chromosome deleted for the region between the repeats and a circular plasmid containing the processed pseudogene (which is subsequently lost or degraded).

that transposons and retroviruses may preferentially insert into regions of open chromatin, such as are found near active genes (EVGEN'EV, LEVINE and GUBENKO 1977; BREINDL, HARBERS and JAENISCH 1984; INOUYE, YUKI and SAIGO 1984; EIBEL and PHILIPPSEN 1984). The assumption of neutrality of insertion for most processed pseudogenes may, therefore, require additional modification as more data on sites of insertion and their effects on gene activity becomes available. Likewise, pseudogenes inserted near active genes may be protected from deletion, as only those deletions that do not affect the active genes are likely to be neutral. Thus, there may be certain "safe neighborhoods" that protect pseudogenes from deletion (GRANT 1981). In spite of these potential caveats the initial assumption of neutrality is not unreasonable for most processed genes.

Our model proceeds as follows: we count only those pseudogenes that are fixed in the population and assume for modeling ease that, when an insertion or deletion of a pseudogene appears, it is either lost or fixed in one generation. We will show later that exclusion of pseudogenes not destined to reach fixation has at most a negligible effect. Assume a fixed number, k, of active genes in the gene family under consideration, noting that $k = 1$ if we are considering a single-copy gene. This restriction to a fixed number of genes does not allow us to model retroposons, such as *Alu*, and is the key feature that separates our simple model from the more elaborate models required for transposable elements (*e.g.*, LANGLEY, BROOKFIELD and KAPLAN 1983; CHARLESWORTH and CHARLESWORTH 1983). Let v be the per generation per active gene copy rate

of formation of processed pseudogenes. In a population of N diploids, $2Nkv$ new pseudogenes are formed each generation, each with probability $1/(2N)$ of becoming fixed under the assumption of neutrality, so that on average $2Nkv[1/(2N)] = kv$ new pseudogenes destined to become fixed appear each generation. Likewise, if $n(t)$ is the number of pseudogenes at time t, and δ is the per generation per pseudogene deletion probability, then $\delta n(t)$ is the expected number of pseudogenes destined to be removed from the population that arises that generation. Putting these results together gives the differential equation describing the dynamics of the expected number of processed pseudogenes for a particular gene family:

$$dn(t)/dt = vk - \delta n(t) \tag{1}$$

With initial condition $n(0) = 0$ (*i.e.*, initially no pseudogenes), (1) has solution

$$n(t) = k(v/\delta)\,[1 - \exp(-\delta t)] \tag{2}$$

The approach to the steady-state equilibrium is monotonic increasing, with time scale set by the rate of spontaneous deletions. At $t = 1/\delta$, the number of processed genes is at 63% of its equilibrium value, and by $t = 5/\delta$ generations the population is essentially at equilibrium.

At the steady-state equilibrium, $n = k(v/\delta)$. If traditional pseudogenes are ignored, the gene family consists of $k + n$ members at equilibrium, giving the proportion of functional copies (π) within the gene family as

$$\pi = k/(k + n) = 1/[1 + v/\delta] \tag{3a}$$

or

$$v/\delta = 1/\pi - 1 \tag{3b}$$

From (3a), $\pi \approx \delta/v$ if $v \gg \delta$, whereas $\pi \approx 1 - v/\delta$ if $v \ll \delta$.

Our above analysis computes the expected number (*i.e.*, mean number) of processed pseudogenes associated with a particular gene family. It is of some interest to examine the complete process, and we do this by noting that our model is simply an immigration-death process, which is well characterized (COX and MILLER 1965, pp. 168–169). At equilibrium the number of processed pseudogenes associated with a particular gene family is simply a Poisson distribution with mean kv/δ (equation 58 of COX and MILLER 1965). Thus, the variance at equilibrium is simply the mean, and in addition we can compute p_0, the probability (at equilibrium) that a particular gene family has no associated processed pseudogenes. This is simply

$$p_0 = \exp\{-kv/\delta\} \tag{4}$$

Thus, if v/δ values are greater than 5, the probability that a single active gene ($k = 1$) has no associated processed pseudogenes is less than 0.006, whereas for $k = 2$, $p_0 < 5 \times 10^{-5}$. Finally, if we start the process with no processed pseudogenes, we see by inspection of the generating function given by equation 57 of COX and MILLER (1965) that the transient distribution is also Poisson, with mean for a particular time t given by (2).

TABLE 1

Expected percent divergence after t *generations*

μ/δ	$t = 1/\delta$	$t = 2/\delta$	$t = 5/\delta$
0.01	2	4	9
0.05	9	18	37
0.10	18	31	55
0.50	55	70	75
1.00	70	75	75

The amount of expected divergence after t generations was computed using $(3/4)[1 - \exp\{(8/3)\mu t\}]$, where μ is the per nucleotide mutation rate, and all mutants are selectively neutral (KIMURA and OHTA 1972). Because of the steady turnover of pseudogenes, we expect that 37, 14 and less than 1% of the pseudogenes are (respectively) older than $1/\delta$, $2/\delta$ and $5/\delta$ generations (δ is the per pseudogene rate of spontaneous deletion). Thus, if $\mu/\delta = 0.05$, 63% of the processed pseudogenes in a particular gene family should show less than 9% divergence and 86% should show less than 18% sequence divergence.

An additional feature of our model is that large amounts of sequence homology can be maintained between pseudogenes in the absence of gene conversion provided that the rate of spontaneous deletion is sufficiently high. The average lifetime of a processed pseudogene follows an exponential distribution with parameter δ, so the probability that the lifetime of a pseudogene exceeds τ generations is $1 - \exp(-\delta\tau)$. Only 37% of pseudogenes persist longer than $1/\delta$ generations, only 14% persist longer than $2/\delta$ generations and less than 1% persists for more than $5/\delta$ generations. Provided that δ is sufficiently large, most pseudogenes in a gene family will be of relatively recent origin and, hence, show high homology with one another. The critical quantity for estimating how much homology can be maintained by a given spontaneous deletion rate is the ratio μ/δ, where μ is the per nucleotide mutation rate. Table 1 lists the expected homology between two pseudogenes for various values of μ/δ at $1/\delta$, $2/\delta$ and $5/\delta$ generations. It is seen that high homology can be maintained, provided $\mu/\delta \leq 0.05$. Taking $\mu = 5 \times 10^{-9}$ (KIMURA 1983, pp. 172–175) implies that most pseudogenes in a gene family should differ by less than 20%, provided $\delta > 10^{-7}$. Of characterized processed pseudogenes, those for rat cytochrome *c* show 10–12% divergence (SCARPULLA and WU 1983), mouse L7 shows 5–15% divergence (KLEIN and MEYUHAS 1984) and human U1 snRNA shows 4–16% divergence (DENISON *et al.* 1981; VAN ARSDELL *et al.* 1981), all values consist with $\delta > 10^{-7}$, implying from (3b) and a conservative overestimate of $\pi = \frac{1}{3}$ (see below) that $\nu > 2 \times 10^{-7}$ for these gene families.

Our analysis of the expected amount of divergence between processed pseudogenes can be refined by using the method of identity coefficients. Such an analysis would proceed along the lines of recent models of the identity between a mixed number of transposable elements (which also allow for gene conversion) developed by OHTA (1985) and SLATKIN (1985), with the extension that

the active and processed pseudogenes need be followed separately. This can be simply done by adding two additional identity coefficients: the probability of identity between an active gene and a processed pseudogene and the probability of identity of two nonallelic pseudogenes (T. OHTA, personal communication).

Finally, some discussion is required of our decision to ignore those pseudogenes that are not destined to become fixed. We assumed newly created pseudogenes not destined to become fixed are lost instantaneously, but in fact they usually persist for a new generations before being lost. How many additional pseudogenes would be observed due to this transient polymorphism? In each generation $2Nvk$ new pseudogenes are created, only vk of which are destined to become fixed. For those not destined to become fixed, the conditional expected time until loss is the same as the standard result for neutral alleles: $4N_e[p/(1 - p)]\ln(1/p)$, where p is the initial frequency of the allele and N_e is the variance-effective population size (KIMURA and OHTA 1969), since we assume all sites of insertion are unique, $p = 1/2N$. Putting these results together gives the expected additional number of pseudogenes created by the transient segregation of sites that never reach fixation as

$$[2Nvk][1 - vk][\ln(2N)\ 2N_e/N] \tag{5a}$$

Under the assumption that $vk \ll 1$, (5a) simplifies further to

$$4vkN_e\ln(2N) \tag{5b}$$

Thus, at equilibrium the actual number of different sites in the population that have processed pseudogenes is

$$vk[1/\delta + 4N_e\ln(2N)] \tag{6}$$

Provided that $1/\delta \gg 4N_e\ln(2N)$, we can safely ignore sites not destined to reach fixation in our analysis. If the population is very large, care must be taken. In this case, however, most segregating sites not destined to reach fixation will have the processed gene at very low frequencies so that these sites are likely to be missed anyway from a small sample from the population.

ESTIMATED FORMATION/DELETION RATIOS FOR DIFFERENT FAMILIES

Equation (3b) indicates that the formation to deletion (v/δ) ratio for any particular gene family can be estimated simply from π, the proportion of functional genes (excluding traditional pseudogenes) for that family. Preliminary estimates of π are available for several gene families. For mouse ribosomal protein L7, $\pi = 1/6$, implying $v = 5\delta$ (KLEIN and MEYUHAS 1984); for murine ribosomal protein L32, $\pi = 1/16$, yielding $v = 15\delta$ (DUDOV and PERRY 1984); for human β-tubulin, $\pi = 3/8$, giving $v = 1.7\delta$ (CLEVELAND 1983); for rat cytochrome c, $\pi = 1/4$, giving $v = 3\delta$ (SCARPULLA and WU 1983); finally, for human U1 snRNA, π is about 10%, giving $v = 9\delta$ (DENISON and WEINER 1982; MANSER and GESTELAND 1982; BERNSTEIN, MOUNT and WEINER 1983; VAN ARSDELL and WEINER 1984). These estimates include only characterized clones and are best regarded as giving an upper bound for π. Additional processed

genes may be detected under conditions of lower stringency. Indeed, 13 additional regions hybridized under lower stringency for the mouse L7 gene but were not further characterized (KLEIN and MEYUHAS 1984). Furthermore, mutation experiments suggest that other mouse ribosomal protein gene families (which contain from eight to 20 members) may have only a single active gene (DUDOV and PERRY 1984), giving π values for these families potentially as low as 0.05.

The π values given above fall within a reasonably narrow range (0.05–0.375), implying that ν/δ values for the five families examined also have a fairly narrow range (1.7–15). If these π values do indeed represent true equilibrium values for these gene families, this narrow range is somewhat surprising. One might think that certain genes (*e.g.*, snRNAs that have the ability to self-prime *in vitro*) would have a much higher rate of pseudogene formation than other germline-active genes, but that all pseudogenes would have approximately the same rate of spontaneous deletion. If the tentative data presented here are correct, then these families either all have approximately the same rate of pseudogene formation (*e.g.*, within an order of magnitude) or else there is a strong positive correlation between the rates of formation and loss. It remains to be seen whether this preliminary conclusion will be substantiated as further sequence data become available.

If pseudogene deletion rates (δ) and formation rates (ν) are positively correlated, gene families with a high ν value should also show less divergence between pseudogenes (because of the correlated higher δ value) than gene families with lower ν values. Thus, it may be possible to distinguish between our two alternate explanations for similar ν/δ ratios by comparing the degree of homology between pseudogenes in different families, although gene conversion could obscure the results. Conversely, if the correlation between ν and δ is correct, gene families whose processed pseudogenes showed high homology would be also expected to have higher pseudogene formation rates.

If processed pseudogenes are a regular feature of any gene that is transcriptionally active in the germline, then how much excess DNA does this generate? The relative constancy of the ratio ν/δ for the five gene families examined here suggests that this ratio may be fairly constant for most genes that are active in the germline. Taking $\pi = 0.10$ for these families implies that, if those genes that produce polymerase II transcripts in the germline account for about 1% of total DNA, then approximately 9% of total DNA is processed pseudogenes. Of course such estimates are very crude and await subsequent refinement from additional estimates of π for other families that are active in the germline and some estimate of just what percentage of total DNA consists of such active genes. Potential complications can arise in that processed pseudogenes may be inserted into regions of DNA which are subsequently amplified, such as satellite DNA, where a repeat containing an insertion can be spread throughout an entire tandem array by unequal crossing over (SMITH 1976). The presence of a *Kpn*I element (see below) in African green monkey α-satellite DNA demonstrates that insertions into potentially amplifiable DNA do occur (THAYER and SINGER 1983). Insertions of pseudogenes into a subsequently amplified region of DNA has been suggested to account for the unu-

sual abundance of some vertebrate GAPDH pseudogenes, which are very abundant in mouse and rat (>200) but present in only ten to 30 copies of man, guinea pig and hampster (PIECHACZYK *et al.* 1984).

There is at least one Pol II gene family for which the ν/δ ratio is (apparently) much larger than the values we report here. This is the *Kpn* I family of repetitive DNA in humans which is equivalent to the mouse MIF-1 family (ROGERS 1983b; SUN *et al.* 1984; GRIMALDI, SKOWRONSKI and SINGER 1984). *Kpn* sequences make up of approximately 6% of the total human genome, and polymerase II transcripts of both strands of these sequences appear to be quite common, presumably resulting from read-through transcription from other nearby Pol II genes. However, some small fraction of *Kpn* sequences show discrete transcripts from one strand only, with a subset of these asymmetrically transcribed genes presumably being the active genes responsible for creating *Kpn* processed pseudogenes (SUN *et al.* 1984). Both the actual number of transcribing genes (k) and the function (if any) of these active genes are unknown, although MARTIN *et al.* (1984) report that a reading frame which is open for 326 amino acids is well conserved between mouse and primates, suggesting at least some nontrivial function for the active gene(s) generating *Kpn*.

Of the approximately 20% of human DNA that consists of repetitive sequences, very roughly 10% are unassigned interspersed repeats (SCHMID and JELINEK 1983; SUN *et al.* 1984), a figure close to our very crude estimate of the total percentage of genomic DNA consisting of processed pseudogenes produced by genes other than *Kpn*.

EVOLUTION OF THE GENETIC SYSTEMS RESPONSIBLE FOR PROCESSED PSEUDOGENES

So far we have treated processed pseudogenes as being selectively neutral and assumed that creation (ν) and deletion (δ) rates are evolutionary stable, although the actual values may vary from gene family to gene family. However, potential selection pressures exist both on particular pseudogenes (*i.e.*, those sites where pseudogene insertion has disrupted an essential gene) and more generally on the genetic system(s) associated with processed pseudogene formation. Thus, we have to ask about the evolution of the parameters ν and δ, in a similar manner to studies that have asked about the evolution of other genetic parameters, such as recombination or mutation rates (*e.g.*, FELSENSTEIN 1974; LEIGH 1973). Both processed pseudogenes and transposable elements share potential deleterious effects caused by increased genome size and by creation of deleterious mutations. Processed pseudogenes and transposons may also in some cases be advantageous by increasing the evolutionary flexibility of an organism or by generating advantageous mutations. In this section we examine these various potential selective forces and their evolutionary consequences. Our main theme is that, since transposable elements and processed genes may share common genetic systems, direct selection on the systems governing transposable elements can result in a correlated response in the systems for creating processed pseudogenes, and vice versa. It is our contention that

most evolution of ν and δ is due to correlated responses during the evolution of certain classes of transposons and retroviruses. Below, we first examine possible molecular mechanisms for altering ν and δ by either direct or correlated responses, and then we examine possible selection pressures on processed genes.

Insertion and deletion rates can be altered by either global changes or changes restricted to a particular gene family. Changes in the genetic system(s) producing processed genes, such as an alteration in the germline concentration of reverse transcriptase or enzymes required for insertion, affect most (if not all) gene families active in the germline. These global changes are the result of evolution of the entire system producing processed genes, contrasted with gene family-specific changes, which reflects evolution restricted to a particular gene family. Correlated responses are likely to produce global changes, whereas specific selection on a particular family is likely to produce gene family-specific changes. As an example of a possible correlated response, suppose that the frequency of dispersed repetitive elements (such as *Alu*) increases. This results in a global increase in δ due to increased chances for recombination between dispersed elements. Correlated responses are even more likely to alter ν. NELSON, LEVY and LEONG (1981) report the presence of a specific inhibitor of reverse transcriptase in human placentas. These authors speculate that such an inhibitor has evolved to reduce retroviral infections, but this would also result in a correlated reduction in ν. In a similar manner, evolution of copy number control in transposons may also result in a global increase or decrease in ν, due to changes in the levels of transposon-encoded proteins involved in the formation of processed pseudogenes (*i.e.*, insertional enzymes). Gene family-specific alterations in ν could occur by the evolution of secondary structures in the 3′ end of the transcription unit. Such structures might reduce primer binding or provide a block to reverse transcriptase. If direct selection has occurred for gene family-specific alterations, we would expect it in those families that produce major amounts of germline RNA, such as ribosomal genes.

One obvious type of selective pressure associated with processed pseudogene formation (and with transposable elements) is the increased cost in DNA replication associated with increased genome size. It is unclear just how important this increased cost might be for multicellular organisms (DOOLITTLE and SAPIENZA 1980; ORGEL and CRICK 1980; CAVALIER-SMITH 1980; DOVER 1980; SMITH 1980; ORGEL, CRICK and SAPIENZA 1980; DOVER and DOOLITTLE 1980; JAIN 1980). The actual time required for cell division may remain unchanged if additional origins of replication are added as the genome size increases. Thus, an increase in genomic size does not automatically imply an increase in generation time. CAVALIER-SMITH (1978) has even proposed that in some situations the only way to decrease cell division time is to increase the amount of genomic DNA. In his scheme, the excess DNA increases the volume of the nucleus, with a resulting increase in the surface area of the nuclear envelope, allowing for a faster rate of exchange with the cytoplasm, and hence, increased growth rate. Another subtle way for a change in total DNA content to have phenotypic effect is by altering the kinetics of gene regulation (LIN and RIGGS 1975; TRAVERS 1983), implying that there may exist a concentration of DNA

which optimizes the kinetics of DNA regulatory protein binding. Many other interesting arguments for and against a change in the amount of DNA in a cell can be found in the litany of papers cited at the beginning of this paragraph. The bottom line is that much speculation and little hard facts exist on the "cost" of excess DNA. The presence of widely differing amounts of DNA between even very closely related species is suggestive that organisms may be fairly robust with respect to the consequences of varying amounts of genomic DNA.

A more direct consequence of the insertion of processed pseudogenes is the creation of mutations, either by disrupting the coding region of a gene or by altering its regulation. Transposable elements are notorious for creating spontaneous mutations, especially in Drosophila (GREEN 1980; THOMPSON and WOODRUFF 1981; SPRADLING and RUBIN 1981). There are, however, fundamental differences between processed pseudogenes and transposons with respect to altering regulatory regions. Processed pseudogenes are transcriptionally inactive, whereas transposable elements can be quite active, being able to affect the expression of genes several kilobases away from the site of insertion (WILLIAMSON, YOUNG and CIRIACY 1981; MCGINNIS, SHERMOEN and BECKENDORF 1983; JACKSON 1984). Processed Pol II pseudogenes are unlikely to carry transcriptional promoters but may occasionally effect transcription from active promoters in other ways. Insertion of an *Alu* type 2 element in the 3′ region of the mouse class I histocompatibility *D* and *L* genes created new polyadenylation signals for these genes (KRESS *et al.* 1984). Overall, however, we would expect that the "target size" for mutations induced by transposable elements is likely to be much larger than that for processed genes, which most likely create mutations primilary by physically disrupting some important structural element in an active gene.

Processed pseudogenes may also have a deleterious effect by interfering with the active gene(s) that generated them. Insertion of a processed pseudogene near a strong promoter may create an antisense RNA, an RNA that has the opposite sense of the coding RNA for the active gene. Antisense RNAs have been shown to inhibit gene expression in both prokaryotes (MIZUNO, CHOU and INOUYE 1984; TRAVERS 1984) and eukaryotes (IZANT and WEINTRAUB 1984).

Thus, both transposable elements and processed pseudogenes can create potential problems by creating spontaneous mutations and increasing genomic DNA. For both of these events, the effects of transposable elements are likely to be more important than the effects of processed pseudogenes. Given that the formation of processed pseudogenes may require enzymes that are also involved in the mobilization of transposable elements, we suggest that the evolution of processed pseudogene copy number is more likely to be affected by correlated responses to selection for transposable element control rather than by direct selection on control of pseudogene copy number. Conversely, the spread of certain classes of transposable elements through a population may result in an increase in the equilibrium copy number of processed pseudogenes, again due to correlated responses.

Although we expect most features of processed pseudogenes and transposons

to be either neutral or deleterious, there are rare cases in which particular elements may be at a selective advantage. Certain transposons (IS*50*, IS*10*) confer a competitive advantage to *E. coli* strains that contain these elements over otherwise identical strains lacking them when both are grown in chemostats. The competitive advantage of IS*50* does not require transposition (HARTL *et al.* 1983) and is presumably due to the action of some IS*50*-encoded gene product, whereas IS*10* gives a selective advantage in certain cases by increasing the mutation rate of the host bacterium (CHAO *et al.* 1983). Under some situations then we expect the mutational effects associated with either transposons or processed genes to produce advantageous mutants. The evolutionary implications are, however, different for transposons and processed genes. A beneficial mutant created by the insertion of a processed gene will not affect the distribution of processed genes at other sites because no additional pseudogenes will be created from the copy that generated the beneficial mutant. A transposon that produces a beneficial mutant by an insertional event can become fixed at that particular site and subsequently generate elements for insertion elsewhere. Another difference between transposons and processed pseudogenes is that transposons are likely to have a much greater range of effects in producing regulatory mutants, which we might expect many beneficial mutants to be. Thus, although creation of beneficial mutants by a transposon may increase the number of other sites that contain this element, processed pseudogenes that create beneficial mutants are likely to have no effect on increasing the number of other sites containing pseudogenes.

Some processed pseudogenes may also have beneficial effects by becoming genes with new functions. OHNO (1970) has argued that gene duplication is the key to the creation of new gene functions. Direct duplication of the structural gene and part of its flanking 3′ and 5′ sequences is one option, and processed pseudogenes offer another. Evidence from traditional pseudogenes in the α-globin cluster (PROUDFOOT and MANIATIS 1980) suggests that directly duplicated copies of a gene may be under some selective constraint for several million years before diverging (in this case, being silenced). We interpret this constrained evolution as follows: duplicated genes that produce a slightly altered product may interfere with the action of the normal product. Thus, alterations of the coding sequence are likely to be deleterious for many duplicated genes, generating the observed constraint. On the other hand, complete inactivation of a duplicate gene is likely to be neutral, as the other unaltered copy is sufficient to carry out the required function. This suggests that a key step in the formation of genes with new functions may be mutational events that alter the regulatory pattern and then the subsequent divergence of the protein-coding regions. Processed pseudogenes could in rare cases acquire new regulatory regions, either by serendipitous insertional events (*i.e.*, near a retroviral LTR) or by conversion-like events. Such reactivated genes would be easily detectable, as they either lack the introns of the parental genes or (in those cases in which new introns might be inserted) have a different distribution of introns. There are examples of related genes with unrelated introns (CORNISH-BOWDEN 1982, 1984; ANTOINE and NIESSING 1984), but many other expla-

nations are possible in addition to the one suggested here. ANTOINE and NIESSING (1984) further suggest that gene conversion between an intron-containing active gene and either associated processed genes or processed gene precursors (*i.e.*, nonintegrated reverse-transcribed DNA) allows for the partial (or complete) removal of introns from active genes, a process that would reduce the amount of DNA associated with a particular gene.

We have offered the suggestion that processed pseudogenes may be involved in the creation of new gene functions primarily as a formal testable hypothesis. We believe it is quite unlikely. Although we have suggested some possibilities for which particular sites containing processed genes may be at a selective advantage, we do not believe that a general system for creating processed pseudogenes would be selected for directly. Evolution is very opportunistic, however, and may indeed occasionally come up with an advantageous use of particular processed pseudogenes, perhaps in some of the ways suggested here.

I thank HOLLY IRICK, WEN-HSIUNG LI, TOMOKO OHTA and NICK BARTON for helpful comments and criticisms. This work has supported by a National Institutes of Health postdoctoral fellowship.

LITERATURE CITED

ALBERTINI, A. M., M. HOFER, M. P. CALOS AND J. H. MILLER, 1982 On the formation of spontaneous deletions: the importance of short sequence homologies in the generation of large deletions. Cell **29:** 319–328.

ANTOINE, M. AND J. NIESSING, 1984 Intron-less genes in the insect *Chrionomus thummi thummi*. Nature **301:** 795–798.

BATTEY, J., E. E. MAX, W. O. MCBRIDE, D. WWAN AND P. LEDER, 1982 A processed human immunolgobulin ϵ gene has been moved to chromosome 9. *Proc. Natl. Acad. Sci. USA* **79:** 5956–5960.

BERNSTEIN, L. B., S. M. MOUNT AND A. M. WEINER, 1983 Pseudogenes for human small nuclear RNA U3 appear to arise by integration of self-primed reverse transcripts of the RNA into new chromosomal sites. Cell **32:** 461–472.

BREATNACH, R. AND P. CHAMBON, 1981 Organization and expression of eucaryotic split genes encoding for proteins. *Annu. Rev. Biochem.* **50:** 349–383.

BREINDL, M., K. HARBERS AND R. JAENISCH, 1984 Retrovirus-induced lethal mutation in collagen I gene of mice is associated with an altered chromatin structure. *Cell* **38:** 9–16.

BROOKFIELD, J. F. Y., 1982 Interspersed repetitive DNA sequences are unlikely to be parasitic. *J. Theor. Biol.* **93:** 281–299.

CALABRETTA, B., D. L. ROBBERSON, H. A. BARRERA-SALPANA, T. P. LAMBROU AND G. F. SAUNDERS, 1982 Genome instability in a region of human DNA enriched in *Alu* repeat sequences. *Nature* **296:** 219–225.

CAVALIER-SMITH, T., 1978 Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. *J. Cell Sci.* **34:** 247–278.

CAVALIER-SMITH, T., 1980 How selfish is DNA? Nature **285:** 617–618.

CHAO, L., C. VARGAS, B. B. SPEAR AND E. C. COX, 1983 Transposable elements as mutator genes in evolution. *Nature* **303:** 633–635.

CHARLESWORTH, B. AND D. CHARLESWORTH, 1983 The population dynamics of transposable elements. *Genet. Res. Camb.* **42:** 1–27.

CHILDS, G., R. MAXSON, R. H. COHN AND L. KEDES, 1981 Orphons: dispersed genetic elements derived from tandem repetitive genes of eucaryotes. *Cell* **23:** 651–663.

CLEVELAND, D. W., 1983 The tubulins: from DNA to RNA to protein and back again. *Cell* **34:** 330–332.

CORNISH-BOWDEN, A., 1982 Related genes can have unrelated introns. *Nature* **297:** 625–626.

CORNISH-BOWDEN, A., 1984 No introns in insect globin genes. *Nature* **301:** 724.

COX, D. R. AND H. D. MILLER, 1965 *The Theory of Stochastic Processes.* Chapman and Hall, London.

DAIGER, S. P., R. S. WILDIN AND T.-S. SU, 1982 Sequence on the human Y chromosome homologous to the autosomal gene for arginiosuccinate synthetase. *Nature* **298:** 682–684.

DENISON, R. A., S. W. VAN ARSDELL, L. B. BERNSTEIN AND A. M. WEINER, 1981 Abundant pseudogenes for small nuclear RNAs are dispersed in the human genome. *Proc. Natl. Acad. Sci. USA* **78:** 810–814.

DENISON, R. A. AND A. M. WEINER, 1982 Human U1 RNA pseudogenes may be generated by both DNA and RNA mediated mechanisms. *Mol. Cell. Biol.* **2:** 815–828.

DONEHOWER, L. A. AND H. E. VARMUS, 1984 A mutant murine leukemia virus with a single missense codon in *pol* is defective in a function affecting integration. *Proc. Natl. Acad. Sci. USA* **81:** 6461–6465.

DOOLITTLE, W. F., T. B. L. KIRKWOOD AND M. A. H. DEMPSTER, 1984 Selfish DNAs with self-restraint. *Nature* **307:** 501–502.

DOOLITTLE, W. F. AND C. SAPIENZA, 1980 Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284:** 601–603.

DOVER, G., 1980 Ignorant DNA? *Nature* **285:** 618–620.

DOVER, G., AND W. F. DOOLITTLE, 1980 Modes of genome evolution. *Nature* **288:** 646–647.

DUDOV, K. P. AND R. P. PERRY, 1984 The gene family encoding the mouse ribosomal protein L32 contains a uniquely expressed intron-containing gene and an unmutated processed gene. *Cell* **37:** 457–468.

EDLUND, T. AND S. NORMARK, 1981 Recombination between short DNA homologies causes tandem duplication. *Nature* **292:** 269–271.

EIBEL, H. AND P. PHILIPPSEN, 1984 Preferential integration of yeast transposable element TY into a promoter region. *Nature* **307:** 306–308.

EVGEN'EV, M., A. LEVINE AND I. GUBENKO, 1977 Are late replicating regions in polytene chromosomes of *Drosophila* enriched by repeated nucleotide sequences? *Nature* **268:** 766–767.

FELSENSTEIN, J., 1974 The evolutionary advantage of recombination. *Genetics* **78:** 737–756.

FLAVELL, A. J., 1984 Role of reverse transcription in the generation of extrachromosomal *copia* mobile genetic elements. *Nature* **310:** 514–516.

FLAVELL, A. J. AND D. ISH-HOROWICZ, 1983 The origin of extrachromosomal circular *copia* elements. *Cell* **34:** 415–419.

GARDER, R. C., A. J. HOWARTH, P. HAHN, M. BROWN-LUEDI, R. J. SHEPHERD AND J. MESSING, 1981 The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun sequencing. *Nucleic Acids Res.* **9:** 2871–2888.

GOJOBORI, T., W.-H. LI AND D. GRAUR, 1982 Patterns of nucleotide substitution in pseudogenes and functional genes. *J. Mol. Evol.* **18:** 360–369.

GRANT, B., 1981 The safe-neighborhood hypothesis of junk DNA. *J. Theor. Biol.* **90:** 149–150.

GREEN, M. M., 1980 Transposable elements in *Drosophilia* and other diptera. *Annu. Rev. Genet.* **14:** 109–120.

GRIMALDI, G., J. SKOWRONSKI AND M. F. SINGER, 1984 Defining the beginning and the end of *KpnI* family segments. *EMBO J.* **3:** 1753–1759.

HALL, B. D., S. G. CLARKSON AND G. TOCCHINI-VALENTINI, 1982 Transcription initiation of eucaryotic transfer RNA genes. *Cell* **29:** 3–5.

HARTL, D. L., D. E. DYKHUIZEN, R. D. MILLER, L. GREEN, AND J. DE FRAMOND, 1983 Transposable element IS50 improves growth rate of *E. coli* cells without transposition. *Cell* **35:** 503–510.

HOLLIS, G. F., P. A. HIETER, O., W. MCBRIDE, D. SWAN AND P. LEDER, 1982 Processed genes: a dispersed human immunoglobulin gene bearing evidence of RNA-type processing. *Nature* **296:** 321–325.

INOUYE, S., S. YUKI AND K. SAIGO, 1984 Sequence-specific insertion of the *Drosophila* transposable genetic element *17.6. Nature* **310:** 332–333.

IZANT, J. G. AND H. WEINTRAUB, 1984 Inhibition of thymidine kinase gene expression by antisense RNA: a molecular approach to genetic analysis. *Cell* **36:** 1007–1015.

JACKSON, I. J., 1984 Transposable elements and suppressor genes. *Nature* **309:** 751–752.

JAENISCH, R., 1983 Endogenous retroviruses. *Cell* **32:** 5–6.

JAGADEESWARAN, P., B. G. FORGET AND S. M. WEISSMAN, 1981 Short interspersed repetitive DNA elements in eucaryotes: transposable DNA elements generated by reverse transcription of RNA Pol III transcripts? *Cell* **26:** 141–142.

JAIN, H. K., 1980 Incidental DNA. *Nature* **288:** 647–648.

KARIN, M. AND R. I. RICHARDS, 1982 Human metallothionein genes—primary structure of the metallothionein-II gene and a related processed gene. *Nature* **299:** 797–802.

KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution.* Cambridge University Press, London.

KIMURA, M. AND T. OHTA, 1969 The average number of generations until extinction of an individual mutant gene in a finite population. *Genetics* **63:** 701–709.

KIMURA, M. AND T. OHTA, 1972 On the stochastic model for estimation of mutational distance between homologous proteins. *J. Mol. Evol.* **2:** 87–90.

KLEIN, A. AND O. MEYUHAS, 1984 A multigene family of intron lacking and containing genes, encoding for mouse ribosomal protein L7. *Nucleic Acids Res.* **12:** 3763–3776.

KORN, L. J., 1982 Transcription of *Xenopus* 5S ribosomal RNA genes. *Nature* **295:** 101–105.

KORNBERG, A., 1980 *DNA Replication.* W. H. Freeman, San Francisco.

KRESS, M., Y. BARRA, J. G. SEIDMAN, G. KHOURY AND G. JAY, 1984 Functional insertion of an Alu type 2 (B2 SINE) repetitive sequence in murine class I genes. *Nature* **226:** 974–977.

LANGLEY, C. H., J. F. Y. BROOKFIELD AND N. KAPLAN, 1983 Transposable elements in mendelian population. I. A theory. *Genetics* **104:** 457–471.

LEHRMAN, M. A., W. J. SCHNEIDER, T. C. SUDHOF, M. S. BROWN, J. L. GOLDSTEIN AND D. W. RUSSELL, 1984 Mutation in LDL receptor: *Alu-Alu* recombination deletes exons encoding transmembrane and cytoplasmic domains. *Science* **277:** 140–146.

LEIGH, E. G., 1973 The evolution of mutation rates. *Genetics* **73** (Suppl.): 1–18.

LEMISCHKA, I. AND P. A. SHARP, 1983 The sequences of an expressed rat α-tubulin gene and a pseudogene with an inserted repetitive element. *Nature* **300:** 330–335.

LEWIN, B., 1977 *Gene Expression 3: Plasmids and Phages.* John Wiley and Sons, New York.

LEWIN, B., 1983 *Genes.* John Wiley and Sons, New York.

LEWIN, R., 1983 How mammalian RNA returns to its genome. *Science* **219:** 1052–1054.

LI, W.-H., T. GOJOBORI AND M. NEI, 1981 Pseudogenes as a paradigm of neutral evolution. *Nature* **292:** 237–239.

LIN, S. AND A. D. RIGGS, 1975 The general affinity of *lac* repressor for *E. coli* DNA: implications for gene regulation in procaryotes and eucaryotes. *Cell* **4:** 107–111.

LITTLE, P. F. R., 1982 Globin pseudogenes. *Cell* **28:** 683–684.

LUEDERS, K., A. LEDER, P. LEDER AND E. KUFF, 1982 Association between a transposed α-globin pseudogene and retrovirus-like elements in the BALB/c mouse genome. *Nature* **295:** 426–428.

MANSER, T. AND R. F. GESTELAND, 1982 Human U1 loci: genes for human U1 RNA have dramatically similar genomic environments. *Cell* **29:** 257–264.

MARTIN, S. L., C. F. VOLIVA, F. H. BURTON, M. H. EDGELL AND C. H. HUTCHISON, 1984 A large interspersed repeat found in moue DNA contains a long open reading frame that evolves as if it encodes a protein. *Proc. Natl. Acad. Sci. USA* **81:** 2308–2312.

MARX, J. L., 1983 Is RNA copied into DNA by mammalian cells? *Science* **216:** 969–970.

MCGINNIS, W., A. W. SHERMOEN AND S. K. BECKENDORF, 1983 A transposable element inserted just 5′ to a Drosophila glue protein gene alters gene expression and chromatin structure. *Cell* **34:** 75–84.

MCKNIGHT, S. L. AND R. KINGSBURY, 1982 Transcriptional control signals of a eukaryotic protein-coding gene. *Science* **217:** 316–324.

MIYOSHI, J., M. KAGIMOTO, E. SOEDA AND Y. SAKAKI, 1984 The human *c*-HA-*ras*2 is a processed pseudogene inactivated by numerous base substitutions. *Nucleic Acids Res.* **12:** 1821–1828.

MIZUNO, T., M. CHOU AND M. INOUYE, 1984 A unique mechanism regulating gene expression: translational inhibition by a complementary RNA transcript (micRNA). *Proc. Natl. Acad. Sci. USA* **81:** 1966–1970.

MOOS, M. AND D. GALLWITZ, 1983 Structure of two human β-actin-related processed genes one of which is located next to a simple repetitive sequence. *EMBO J.* **2:** 757–761.

MORTON, D. G. AND K. U. SPRAGUE, 1984 *In vitro* transcription of a silkworm 5S RNA gene requires an upstream signal. *Proc. Natl. Acad. Sci. USA* **81:** 5519–5522.

NELSON, J. A., J. A. LEVY AND J. C. LEONG, 1981 Human placentas contain a specific inhibitor of RNA-directed DNA polymerase. *Proc. Natl. Acad. Sci. USA* **78:** 1670–1674.

NISHIOKA, P., A. LEDER AND P. LEDER, 1980 Unusual α-globin-like gene that has cleanly lost both globin intervening sequences. *Proc. Natl. Acad. Sci. USA* **77:** 2806–2809.

OHNO, S., 1970 *Evolution by Gene Duplication.* Springer-Verlag, Berlin.

ORGEL, L. E. AND F. H. C. CRICK, 1980 Selfish DNA: the ultimate parasite. *Nature* **284:** 604–607.

ORGEL, L. E., F. H. C. CRICK AND C. SAPIENZA, 1980 Selfish DNA. *Nature* **288:** 645–646.

OHTA, T., 1985 A model of duplicative transposition and gene conversion of repetitive DNA families. *Genetics* In press.

PEREZ-STABLE, C., T. M. AYRES AND C.-K. J. SHEN, 1984 Distinctive sequence organization and functional programming of an *Alu* repeat element. *Proc. Natl. Acad. Sci. USA* **81:** 5291–5295.

PIECHACZYK, M., J. M. BLANCHARD, S. RIAAD-EL SABOUTY, C. DANI, L. MARTY AND P. JEANTEUR, 1984 Unusual abundance of vertebrate glyceraldehyde 3-phosphate dehydrogenase pseudogenes. *Nature* **312:** 469–471.

PROUDFOOT, N., 1980 Pseudogenes. *Nature* **286:** 840–841.

PROUDFOOT, N. J. AND T. MANIATIS, 1980 The structure of a human α-globin pseudogene and its relationship to α-globin gene duplication. *Cell* **21:** 537–544.

RINEHART, F. P., T. G. RITCH, P. L. DEININGER AND C. M. SCHMID, 1981 Renaturation rate studies of a single family of interspersed repeated sequences in human deoxyribonucleic acid. *Biochemistry* **20:** 3003–3010.

ROBERT, B., P. DAUBAS, M. AKIMENKO, A. COHEN, I. GARNER, J. GUENET AND M. BUCKINGHAM, 1984 A single locus in the mouse encodes both myosin light chains 1 and 3, a second locus corresponds to a related pseudogene. *Cell* **39:** 129–140.

ROEDER, G. S. AND G. R. FINK, 1983 Transposable elements in yeast. pp. 298–328. In: *Mobile Genetic Elements,* Edited by J. A. SHAPIRO. Academic Press, New York.

ROGERS, J., 1983a Retroposons defined. *Nature* **301:** 460.

ROGERS, J., 1983b A straight LINE story. *Nature* **306:** 113–114.

SAIGO, K., W. KUGIMIYA, Y. MATSUO, S. INOUYE, K. YOSHIOKA AND S. YUKI, 1984 Identification of the coding sequence for a reverse transcriptase-like enzyme in a transposable genetic element in *Drosophila melanogaster. Nature* **312:** 659–661.

SCARPULLA, R. C. AND R. WU, 1983 Nonallelic members of the cytochriome c multigene family of the rat may arise through different messenger RNAs. *Cell* **32:** 473–482.

SCHMID, C. W. AND W. R. JELINEK, 1982 The *Alu* family of dispersed repetitive sequences. *Science* **216:** 1065–1070.

SCHWARTZBERG, P., J. COLICELLI AND S. P. GOFF, 1984 Construction and analysis of deletion mutations in the *pol* gene of moloney murine leukemis virus: a new viral function required for productive infection. *Cell* **37:** 1043–1052.

SHAPIRO, J. A., 1979 Molecular model for the transposition and replication of bacteriophage Mu and other transposable elements. *Proc. Natl. Acad. Sci. USA* **76:** 1933–1937.

SHARP, P. A., 1983 Conversion of RNA to DNA in mammals: *Alu*-like elements and pseudogenes. *Nature* **301:** 471–472.

SHMOOKLER REIS, R. J., C. K. LUMPKIN, J. R. MCGILL, K. J. RIABOWOL AND S. GOLDSTEIN, 1983 Extrachromosomal circular copies of an "*inter-Alu*" unstable sequence in human DNA are amplified during *in vitro* and *in vivo* ageing. *Nature* **301:** 394–398.

SLATKIN, M., 1985 Genetic differentiation of transposable elements under mutation and unbiased gene conversion. *Genetics* In press.

SMITH, G. P., 1976 Evolution of repeated DNA sequences by unequal crossover. *Science* **191:** 528–535.

SMITH, T., 1980 Occam's razor. *Nature* **285:** 620.

SPRADLING, A. C. AND G. M. RUBIN, 1981 *Drosophilia* genome organization: conserved and dynamic aspects. *Annu. Rev. Genet.* **15:** 219–264.

STEIN, J. P., R. P. MUNJAAL, L. LAGACE, E. C. LAI, B. W. O'MALLEY AND A. R. MEANS, 1983 Tissue-specific expression of a chicken calmodulin pseudogene lacking intervening sequences. *Proc. Natl. Acad. Sci. USA* **80:** 6485–6489.

SUN, L., K. E. PAULSON, C. W. SCHIMD, L. KADYK AND L. LEINWAND, 1984 Non-*Alu* family interspersed repeats in human DNA and their transcriptional activity. *Nucleic Acids Res.* **12:** 2669–2690.

TEMIN, H. M., 1980 Origin of retroviruses from cellular moveable genetic elements. *Cell* **21:** 599–600.

THAYER, R. E. AND M. F. SINGER, 1983 Interruption of an α-Satellite array by a short member of the *Kpn*I family of interspersed, highly repeated monkey DNA sequences. *Mol. Cell Biol.* **3:** 967–973.

THOMPSON, J. N., AND R. C. WOODRUFF, 1981 A model for spontaneous mutation in *Drosophila* caused by transposing elements. *Heredity* **47:** 327–335.

TRAVERS, A., 1983 Protein contacts for promoter location in eukaryotes. *Nature* **303:** 755.

TRAVERS, A., 1984 Regulation by anti-sense RNA. *Nature* **311:** 410.

TROWSDALE, J., A. KELLY, J. LEE, S. CARSON, P. AUSTIN AND P. TRAVERS, 1984 Linkage map of two HLA-SBβ and two HLA-SBα-related genes: an intron in one of the SBβ genes contains a processed pseudogene. *Cell* **38:** 421–429.

ULLU, E. AND C. TSCHUDI, 1984 *Alu* sequence are processed 7SL RNA genes. *Nature* **312:** 271–272.

VAN ARSDELL, S. W., R. A. DENISON, L. B. BERNSTEIN, A. M. WEINER, T. MANSER AND R. F. GESTELAND, 1981 Direct repeats flank three small nuclear RNA pseudogenes in the human genome. *Cell* **26:** 11–17.

VAN ARSDELL, S. W. AND A. M. WEINER, 1984 Pseudogenes for human U2 small nuclear RNA do not have a fixed site of 3′ truncation. *Nucleic Acids Res.* **12:** 1463–1471.

VANIN, E. F., G. I. GOLDBERG, P. W. TUCKER AND O. SMITHIES, 1980 A mouse α-globin-related pseudogene lacking intervening sequences. *Nature* **286:** 222–226.

VARMUS, H. E., 1982 Form and function of retroviral proviruses. *Science* **216:** 812–820.

VARMUS, H. E., 1983 Reverse transcription in plants? *Nature* **304:** 116–117.

WALTER, P., R. GLIMORE AND G. BLOBEL, 1984 Protein translocation across the endoplasmic reticulum. *Cell* **38:** 5–8.

WEINBERG, R. A., 1980 Origins and roles of endogenous retroviruses. *Cell* **22:** 643–644.

WILDE, C. D., C. E. CROWTHER AND N. J. COWAN, 1982 Diverse mechanisms in the generation of human β-tubulin pseudogenes. *Science* **217:** 549–552.

WILLIAMSON, V. M., E. T. YOUNG AND M. CIRIACY, 1981 Transposable elements associated with constitutive expression of yeast alcohol dehydrogenase II. *Cell* **23:** 605–614.

Communicating editor: B. W. WEIR