PSEUDOGENES

Author: **C. Deborah Wilde**
Department of Genetics
University of Cambridge
Cambridge, England

Referee: N. J. Proudfoot
Sir William Dunn School of Pathology
University of Oxford
Oxford, England

I. INTRODUCTION

With the advent of DNA sequencing and cloning techniques, it became possible to study the structure of a variety of vertebrate and invertebrate genes in detail at the molecular level. Such studies were initially confined to those genes for which RNA or cloned cDNA probes were readily available — 5S ribosomal RNA, histones, globins, and immunoglobulins. However, with more sophisticated cloning procedures, an increasing number of eukaryotic genes became accessible and amenable to study. The early investigations led to two intriguing discoveries that were quite unexpected: the presence of intervening sequences or introns within gene coding regions, and the presence within gene families of pseudogenes (linked sequences with considerable structural similarity to functional genes yet containing mutations that inactivated their transcription or the processing and translation of their transcripts). It soon became clear that pseudogenes and intervening sequences were not just peculiarities of the first gene families to be studied, as they were subsequently identified among a wide variety of eukaryotic genes. Pseudogenes rapidly lost their newsworthiness as it was no longer considered remarkable to discover them in each new gene family that was investigated.

While "first generation" pseudogenes had always been found linked to their functional counterparts, many of the "second generation" pseudogenes appeared to be unlinked to their parent genes and dispersed in the genome to different chromosomes. Once the DNA sequences of several of these dispersed pseudogenes became available, it was clear that they still held some surprises in store. Features of their structure — a lack of intervening sequences and oligoA tracts at their 3′ ends — strongly suggested that the pseudogenes arose through incorporation of mRNA reverse transcripts into the genome, probably at staggered breaks in the chromosome. Reverse transcription was clearly not a process confined to the retroviruses, but one that could occur in normal cells, though perhaps by accident rather than by design.

The description of this new class of pseudogenes and their probable origin shed new light on a mouse α-globin pseudogene that had been isolated some 2 years earlier. At that time, its structure was without precedent and was considered somewhat curious in that it had precisely lost both its intervening sequences. It now seemed plausible that it too had arisen through reverse transcription of a processed RNA intermediate. Furthermore, a similarity was noticed between these "intron-less" or "processed" pseudogenes and members of several families of interspersed repeated sequences — the human *Alu* and *Kpnl* elements and their rodent counterparts, and the F elements of *Drosophila*. All have 3′ oligoA tracts and are flanked by short direct repeats, suggesting that they have populated the genome through integration of copies derived from RNA intermediates. Thus the processed pseudogenes, comparatively small in number, seem to be part of a much larger class of sequences that have arisen through this common mechanism involving integration of RNA reverse transcripts into the genome.

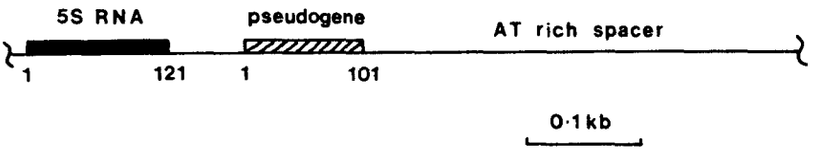

FIGURE 1. The 5S RNA repeat unit of *Xenopus laevis*, showing 5S RNA gene, the downstream pseudogene and AT-rich spacer region.[3]

Clearly, pseudogenes comprise a disparate group of structures and this presents some difficulties in finding an umbrella definition that would include all examples. Pseudogenes were intially defined as DNA sequences that have significant homology to functional genes, but which have sequence differences within their protein coding regions that would cause premature termination of translation or the formation of polypeptides with little homology to the functional gene product. Additionally, it was recognized that many pseudogenes also contained mutations that interfered with transcription initiation or the processing of RNA transcripts. While processed pseudogenes may have intact coding regions, they can still be classed as pseudogenes by virtue of their transcriptional silence, a consequence of their mRNA origins. Thus, either an aberrant coding region or transcriptional silence (or both) may be used to identify a pseudogene.

In this review, examples of different types of pseudogenes will be discussed, taking them broadly in their order of discovery. Perhaps inevitably a review of this type will become something of a catalog of the various pseudogenes and their deficiencies. However, it is hoped that this review will also show what we may learn from these genomic "mistakes" of the normal workings of the cell and of events that might relate to the evolution of eukaryotic genomes.

II. PSEUDOGENES LINKED TO FUNCTIONAL GENES

A. The *Xenopus* 5S Pseudogene

The first gene-like sequence to be dubbed a "pseudogene" was that for the *Xenopus laevis* 5S ribosomal RNA described by Jacq et al.[1] The pseudogene sequence occurs downstream of the functional 5S RNA gene and is part of the 700 nucleotide repeat unit that is amplified during oögenesis (Figure 1). The pseudogene is 20 nucleotides shorter at its 3′ end than its functional counterpart (101 instead of 121 nucleotides) and differs by only nine base changes.[2] No RNA corresponding to this pseudogene could be found in vivo, and thus it appeared to be an inert component of the genome. This raised questions as to why this pseudogene structure had been conserved; whether it served some function in processing of the mature 5S RNA or whether, being part of the duplicated repeat unit, it was just passively preserved along with the active gene. These questions remain largely unanswered, but the question of why no pseudogene transcripts are found in vivo has been addressed in further experiments involving micro injection of the isolated 5S gene and pseudogene into *Xenopus* öocytes.[3]

When the pseudogene is injected alone, it supports a rate of transcription of up to 85% of the level of normal 5S gene transcription. However, at least 75% of the pseudogene transcripts do not terminate correctly at the end of the gene (even though it contains a TTTT sequence thought to be important for correct termination), but read through into adjacent sequences. In vivo this would give rise to random termination in the downstream AT-rich spacer region, and hence no discretely sized transcripts would be formed; in addition, such randomly terminated transcripts might be somewhat unstable. Thus, the lack of pseudogene transcripts of defined length in vivo may be a reflection of inefficient transcription termination rather than a lack of transcriptional activity per se.

However, a further experiment[3] suggests that this may not be the whole explanation. If the 5S gene and the pseudogene are injected together, the rate of transcription from the pseudogene drops to one third of its level when injected alone. This indicates there is competition between the two promoters for RNA polymerase (or other transcription factors) and that the 5S gene has the more effective promoter. The two promoters only differ in sequence by four base changes; it is not clear whether this alone accounts for their differential activity or whether some other feature of the environment surrounding the two sequences is also important.

Thus, the apparent silence of the 5S pseudogene in vivo may in part be due to incorrect termination of transcription (and hence no stable or discretely sized transcripts), but also in part relate to its inefficient competition for RNA polymerase. This observation that pseudogenes may be transcriptionally active when assayed in vitro, while being silent in vivo, will be encountered again for several other pseudogenes. It appears to point towards a higher order of transcriptional control, above that of DNA sequences per se, which must be present in vivo and which is capable of rendering pseudogenes effectively transcriptionally silent.

B. Globin Pseudogenes — a Wealth of Examples

Historically, the next set of pseudogenes to be discovered were those within the α- and β-globin gene families of a variety of mammals.[4-14] Together, the mammalian globin gene families provide examples both of pseudogenes at different stages of their evolutionary decay and of the variety of processes whereby the different gene clusters have evolved.

1. Age and Origin of Pseudogenes

With the exception of two mouse α-globin pseudogenes that are dispersed to different chromosomes from the major α-globin gene cluster[15-18] (Section III.C below), all the globin pseudogenes are found linked to their functional counterparts. The most straightforward explanation for the origin of these pseudogenes is that they derive from duplicated genes formed within the gene clusters, which have diverged and become inactive, i.e., transcriptionally silent. Following inactivation, such genes would have been released from selection and would then rapidly accumulate mutations at a rate more characteristic of noncoding sequences.

Estimates of the evolutionary time spent by each present-day pseudogene, first under selection as an active gene and then without selection as a pseudogene, have been calculated from the percentage of silent and replacement base changes in the coding sequence compared to the active gene.[7,9,19] These estimates assumed that following inactivation, pseudogenes would accumulate mutations at the same rate as silent changes in active genes. However, it appears that there is some selective pressure against changes, even between synonymous codons in functional genes (presumably a reflection of bias in codon usage), and that the rate of nucleotide substitution in pseudogenes is approximately twice the rate of substitutions in the third codon position of active genes.[20-22] Many earlier estimates did not take this factor into account and thus will have tended to be overestimates of pseudogene age.

A further factor that has confounded these estimates is the realization that gene conversion events have played an important role in the evolution of globin gene clusters.[6,23-27] Gene conversion is the nonreciprocal copying of information from one gene to another homologous gene within a cluster, as the result of inter-[6] or intrachromosomal[23] exchange. A number of instances of gene conversion have been detected among α- and β-globin genes, and its effect has been to mask the true evolutionary age of genes, or pseudogenes, that have undergone conversion. Thus, two genes will appear to have arisen by duplication at the time of a conversion event, when in fact they may have a considerably older evolutionary history. For example, comparison of the protein coding regions of the two human adult globins δ and β suggests that they arose from a duplication event not more than 40 million years (MYr) ago.[28,29] However, various noncoding regions, the second intervening sequence, the

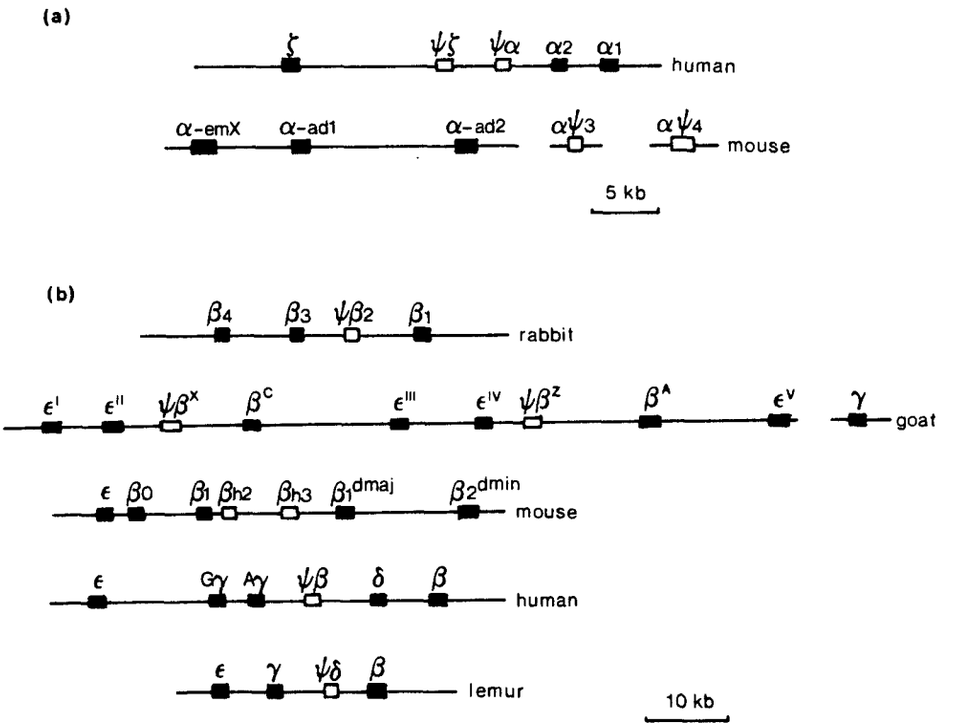

FIGURE 2. Mammalian globin gene families. (a) α-globin genes.[6,17] (b) β-globin genes.[14,29,41,43]

mRNA 3′ untranslated region and 5′ sequences upstream of the CCAAT box, appear to have diverged over a much longer period of time;[30,36] in addition, δ-like genes or pseudogenes are found in lower primates that diverged around 75 MYr ago.[14] Thus, the globin coding region appears to have undergone a recent conversion by the β-gene, which has covered the traces of its more ancient origin. Reliable estimates of evolutionary divergence times can, therefore, only be derived from those regions of the gene that have not been subject to gene conversion.

2. *Human α-Globin Pseudogenes*

In addition to the active embryonic (ζ) and adult (α1, α2) genes, the human α-globin gene cluster contains two pseudogenes, ψζ and ψα (Figure 2a).[6-8] Together, ψζ and ψα represent two extremes in the process of pseudogene formation and decay. Pseudogene ψζ shares >99.5% homology in its coding region to the functional ζ-globin gene and has a single deleterious mutation, a termination codon in its first exon;[8] presumably only very recently has it become a pseudogene. In contrast, ψα is only 75 to 80% homologous to the active α-globin genes and has a considerable array of mutations — base substitutions that introduce many missense codons and that affect the translation initiation codon and RNA splice donor and polyadenylation signal sequences, as well as deletions that cause frame shifts and termination codons in the coding sequence and alter the spacing between CCAAT and ATA boxes in the transcription promoter region.[7] Thus, ψα appears to be a relatively old pseudogene.

An interesting feature of these two pseudogenes is that they still retain unmutated transcription initiation signals, although their spacing in ψα has been altered by a deletion. The ψα promoter region is functional in vitro in HeLa cell extracts and in vivo following transvection into monkey Cos 7 cells, although at levels threefold (in vitro) or tenfold (in

vivo) lower than the α-globin gene promoter.[31] This lower transcription rate, together with the effect of the aberrant polyadenylation sequence (AATGAA) on RNA processing, is probably sufficient to explain the lack of ψα transcripts found in erythroid cells. The situation with regard to ψζ is somewhat different. Its promoter is identical to that of the active ζ-globin gene, and both are equally active following injection into *Xenopus* oöcytes.[32] However, ψζ transcripts cannot be detected in RNA from either human embryos or the K562 erythroid cell line, even though both these tissues contain transcripts of the active ζ-globin gene.[32a] Thus, some other feature of the chromosomal region around the ψζ pseudogene (possibly the presence of regulatory signals or its chromatin configuration) must be responsible for its silence in vivo.

The human α-globin cluster also provides insights into the evolutionary mechanisms that can give rise to pseudogenes. A comparison of the sequences surrounding the ψα pseudogene and the two active genes α1 and α2 suggest that they arose by gene duplication and subsequent unequal crossing over.[6,7] Such events still appear to be operating in present-day human populations, since chromosomes carrying either a single active α-globin gene (associated with the hemoglobin deficiency α-thalassemia),[33] or an α-globin gene triplication,[34,35] have been reported. Since the time the ψα, α1, α2 cluster was formed, the two active genes α1 and α2 have been maintained closely homologous by gene conversion events, while ψα has accumulated base changes to become a pseudogene. Sequences in the intergenic regions upstream of α1 and α2 show strong homology and have been implicated in gene conversion,[7] and their absence upstream of ψα may perhaps explain why it too has not been subject to conversion. Thus, gene duplication by itself may not be sufficient to set a gene on the path to becoming a pseudogene; a more crucial step may be the point at which a gene no longer becomes subject to conversion by neighboring genes and is free to diverge on its own.

3. β-Globin Pseudogenes

The β-globin gene clusters of a number of mammals show considerable variation in their complexity and organization (Figure 2b). However, using the DNA sequence information that is now available for a large majority of the β-globin genes, it has been possible to relate the different present day clusters back to a simple four (or five) gene cluster, which has evolved by various gene duplication and unequal cross-over events.[36,37,37a] As will be shown below, the different β-globin pseudogenes have often been important in establishing the evolutionary routes by which the various clusters have evolved.

a. Rabbit

The rabbit β-globin gene cluster comprises two embryonic β genes (β4, β3) and two adult genes (β2, β1).[38,39] The β2 gene is a pseudogene, which, like the human ψα gene, has many mutations that would affect its transcription, mRNA processing, and translation.[9] Naive estimates of its age, on the basis of the number of replacement and silent nucleotide substitutions in its coding region compared to that of the β1 gene, suggested that it had evolved for ~22 MYr as a functional gene and for ~33 MYr as a pseudogene.[9] However, a more detailed comparison of noncoding as well as coding regions of ψβ2 and β1, indicates that a gene conversion between their coding regions occurred in the past 50 MYr and suggests that the occurrence of two adult genes more probably predates the mammalian radiation, ~85 MYr ago.[40] The rabbit four gene cluster of two embryonic and two adult genes may be thought of as exemplifying the primordial β-globin gene family that was present in the common ancestor of present day mammals.[37]

b. Goat

A majority of the genes encoding β-like globin genes have now been isolated from the goat, and it is apparent that during the evolution of the goat lineage there has been at least a duplication and more probably a triplication of an initial four gene cluster.[41] The two

linked sets of genes each comprise two embryonic (ε) genes and two adult-like genes, one of which is a pseudogene. Further embryonic and adult-like genes, presumed to belong to a third cluster, have also been found (Figure 2b).

The two pseudogenes ψβ^x and ψβ^z are 90% homologous at the nucleotide level and also share several identical deleterious mutations (in phase terminators, missense codons, frameshifts, and altered ATA promoter box).[10,11] This strongly points to their both being derived from a common sequence that was already a defective pseudogene before the duplication of the gene cluster. Thus, like the *Xenopus* 5S pseudogene, the goat pseudogene has been a passive passenger in the amplification of a gene cluster of which it is part. Estimates of the age of the goat pseudogenes have again been complicated by the recognition that extensive gene conversion in all but the 5′ flanking region has occurred between the ancestral ψβ and β genes.[36,40] However, despite this conversion, their degree of sequence divergence suggests that the pseudogene has been nonfunctional for a major part of its subsequent evolutionary history.[11]

c. Mouse

While in the goat the repertoire of β-globin genes has been expanded by the triplication of a four gene set, the mouse β-globin genes have been expanded by gene duplications within the cluster. The mouse cluster comprises three embryonic genes (y, β0, β1), two adult genes β^{dmaj}, β^{dmin}) and two pseudogenes (βh2, βh3) lying between the adult and embryonic genes[12,36,42] (Figure 2b). Both pseudogenes appear to be related in part to the primordial adult-like gene from which the rabbit ψβ2 and goat ψβ^x and ψβ^z pseudogenes are also descended.[36]

Despite βh2 being a relatively divergent sequence that shares only 72% homology with the mouse adult β-globin genes, it has only two overtly deleterious mutations — a frameshift deletion in the first exon and an altered CCAAT box in the promoter region.[43] This suggests that βh2 diverged under selection as a functional gene for a large part of its history before becoming a pseudogene and acquiring mutations in conserved codons and its promoter.

The βh3 pseudogene shares even less homology than βh2 with the functional adult gene,[12,43] and it also seems to have suffered a large internal deletion that has removed ~150 nucleotides from the last ~20 nucleotides of the first intervening sequence to codon 75 in the middle of the second exon.[21] It has been suggested that βh3 is a recombinant between two different ancestral genes[36] — presumably the progenitors of βh2 and the adult β genes. Since βh3 lies between these genes in the cluster, it seems plausible that it was generated by an unequal cross-over event; such an event might also account for the internal deletion in this gene. Alternatively, the deletion may have occurred subsequently as part of the evolutionary decay of the βh3 pseudogene. Unequal cross-overs also appear to have been involved in the generation of other pseudogenes (see Sections II.B.3.d and II.G below) and recombinant Lepore δβ-globin chains,[44] and hence this is not an uncommon evolutionary mechanism.

d. Primates

The primate β-globins provide particularly clear examples of the roles of unequal cross-over events and gene conversion in the evolution of gene clusters. In addition, recent studies of the primate β-like pseudogene ψβ have clarified the evolutionary relationships between the various genes in the β-globin clusters of different mammalian species.[37a] The human β-globin gene cluster comprises an embryonic (ε) gene, two fetal (γ) genes, two adult genes (δ and β), and a pseudogene (ψβ) (Figure 2). A similar arrangement of genes is found in other primates, except that New World monkeys have a single γ-gene and the β-globin cluster in lemurs (lower primates) is smaller still by virtue of it containing an unusual hybrid ψβ-δ pseudogene.[14]

The nucleotide sequence of the hybrid ψβ-δ pseudogene from the brown lemur indicates that in its 5′ flanking region, first and second exons, and first intervening sequence, it is ~75% homologous to the human ψβ pseudogene,[45] while the 3′ part comprising second intervening sequence, third exon, and 3′ flanking sequence is more homologous to the human δ-globin gene.[14] The most plausible mechanism for its origin seems to be an unequal cross-over event between a pseudogene corresponding to the present day human ψβ and a δ-like gene, resulting in a fused ψβ-δ gene structure and a contracted β-globin gene cluster.

A detailed DNA sequence analysis of the β-like ψβ pseudogenes from man and a number of other primates (gorilla, chimpanzee, owl monkey) has recently become available.[45a,45b] These studies show that a ψβ gene is found in all primates and that this gene has probably been a pseudogene for the whole of primate evolution, suggesting that the ancestral primate β-globin gene cluster comprised a five gene set ε-γ-ψβ-δ-β. Comparisons of the ψβ sequences with those of other β-globin genes revealed that the primate ψβ gene is most closely related to the ε″ gene of goats.[37a] Both the primate ψβ pseudogene and the embryonically expressed goat ε″ gene appear to be derived from a common ancestral gene, named η, that is distinct from the ε, γ, δ, and β ancestral genes. The mouse and rabbit β-globin gene clusters lack η-like genes and are thus derived from an ε-γ-δ-β four gene set. The goat β-genes, however, lack descendants of the γ-type gene, and are derived from a triplicated ε-η-δ-β set of genes. Only in primates have descendants of all five types of ancestral genes been retained. Interestingly, the descendants of the ancestral δ-type gene, mouse βh2 and βh3, rabbit ψβ2, goat ψβ^{x}, and ψβ^{z}, and the minor adult δ-globin gene of primates, have all shown a tendency to become silent pseudogenes.

The evolutionary history of the primate δ-globin gene is particularly interesting. DNA sequence comparisons with other adult β-genes show that while its 5′ end has been subject to a relatively recent gene conversion by the β-gene, its 3′ end still bears significant homology to the pseudogenes of mouse (βh2) and rabbit (ψβ2).[30] It thus seems likely that the δ-gene was originally a pseudogene, akin to those of other mammals, and that it became rejuvenated in the early primate lineage by a gene conversion with the adult β-gene. Subsequently, the δ-gene has been poised on the edge between activity and becoming a silent gene, a potential pseudogene. In man and the great apes it is an active gene, albeit expressed at a very low level; however, in the Old World monkey lineage, it is transcriptionally silent[30] and will presumably now evolve into a pseudogene if it is not rescued by a further gene conversion event. Clearly, selection has been insufficient to maintain the activity of two so similar adult like genes in all primate lineages, and in the Old World monkeys we may be seeing the first step in the decay of the δ-gene once more into a pseudogene. Thus, the δ-gene illustrates that while pseudogenes may lie decaying among gene clusters for long periods in evolution, on occasion they can rise up phoenix-like from their ashes renewed by the process of gene conversion, and for a season, at least, regain their activity.

4. Pseudogene Position Within Gene Clusters

As has been seen, pseudogenes are common members of mammalian globin gene clusters, and it is intriguing that they all occupy homologous positions between the adult and embryonic or fetal genes of the cluster. It has been suggested that this might reflect some requirement for a pseudogene (a "spacer" gene) at this position for the correct functioning of the complex[31] — yet there is no evidence to support this and, indeed, both mouse α-globin and chicken globin clusters function quite satisfactorily without pseudogenes.[46-48] Alternatively, at least for the β-globin clusters, one can argue that the different pseudogenes occupy homologous positions by virtue of their evolutionary relatedness, all being descendants of an adult-like gene in the primordial cluster that was already a pseudogene. While present day β-pseudogenes have very different sets of inactivating mutations with no shared defects indicative of a common ancestral sequence, this may be more a reflection of the prevalance

of gene conversion events in the different lineages that would have masked any defects once held in common. What is certainly the case, however, is that there has been little evolutionary pressure to maintain two fully active adult β-genes. Despite rescue attempts by gene conversion (which on only one occasion restored activity in the primate δ-gene), there has been a repeated tendency for the second adult-like gene to become silenced and a pseudogene. What predisposes this second gene to inactivation is not clear, though it is feasible that some feature of its chromosomal environment or its relative position in the cluster may be important.[36]

A further explanation for the position of pseudogenes within gene clusters derives from the role of unequal crossing over in the evolution of these clusters. Cross-over events might generate aberrant genes through the recombination process itself — the mouse βh3 pseudogene is perhaps an example of such. Alternatively, potentially functional but hybrid genes, for example, between adult and embryonic globins, might be formed. Depending on where in the gene the cross-over occurred, this might result in either a hybrid polypeptide with both adult and embryonic characteristics or a promoter-gene combination that gives inappropriate expression of an adult gene in the embryo (or vice versa). Both conditions could well be deleterious and would favor selection to silence any such hybrid genes. This might in part explain the preponderance of pseudogenes in the middle of clusters. There might also be selection to maintain a pseudogene as a buffer in the middle of a cluster, so that recombination by unequal crossing over would tend to form pseudogene hybrids, like the lemur ψβ-δ hybrid pseudogene, rather than a potentially more deleterious embryonic-adult hybrid gene.

However, perhaps the simplest explanation for pseudogene position is just that it is much easier for pseudogenes to be "lost" from one end of a gene cluster by, for example, a deletion or DNA inversion. To remove a pseudogene from the center of a cluster by similar events would disrupt the integrity of a cluster and perhaps destroy the coordinate regulation of its component genes. Pseudogenes formed within gene clusters are perhaps, therefore, "trapped", and their only way out is to evolve away to nonhomology.

5. Pseudogene Fixation

A further question that arises from the study of globin pseudogenes in particular, is how pseudogenes have become fixed in the population. Exactly similar mutations in globin genes can, on the one hand, give rise to pseudogenes spread throughout all the human population and, on the other hand, result in β-thalassemias-deleterious conditions usually found only in certain ethnic groups.

The factors that influence gene fixation are complex; they depend on population size, the frequency of recurrent mutation, and on the degree of selection operating on the particular mutant allele.[48a] The main difference between whether a mutation leads to the development of a pseudogene or a thalassemic defect seems to lie in the realm of selection and genetic fitness. A mutation only has the potential for creating a pseudogene if it is either selectively advantageous or at least neutral in its effect on fitness. This condition is usually only fulfilled if the mutation occurs in a duplicated gene or a member of a multigene family, so that the genome retains at least one functional copy of the gene.

Thus, mutations that give rise to β-thalassemias occur in the β-globin gene, of which there is no duplicated counterpart; the δ-gene, although of essentially similar function, is expressed only at very low levels insufficient to compensate for the β-globin defect. The β-thalassemia mutations are, therefore, highly deleterious, particularly in the homozygous state, and they are only maintained in certain populations by the slight heterozygous advantage they afford in areas where malaria is endemic. In contrast, mutations in the minor δ-globin gene are probably not disadvantageous, and indeed such mutations have led to the δ-gene becoming a silent "pseudogene" in the Old World monkey lineage.[30] Thus, this is perhaps an example of an essentially neutral mutation resulting in the formation of a pseudogene.

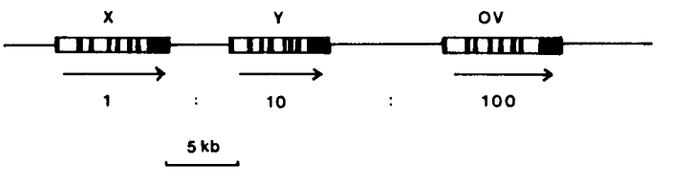

FIGURE 3. Structure of the chicken ovalbumin gene (OV) and pseudoovalbumin genes (X, Y). (■) Gene coding region; (□) intervening sequences; (→) direction of transcription. The ratio of transcription rates for the three genes is also shown.[51]

It is somewhat harder to envisage situations where there could have been a positive selection pressure for the silencing of a gene and the formation of a pseudogene. However, as the balance of α and β globin chain production is vital for the correct functioning of the red blood cell, it is possible that following a gene duplication or other mutational event that disrupted the normal balance of α and β globin expression, such positive selection for gene silencing could have operated. The resulting chromosome (carrying both the original duplication or mutation and the compensating pseudogene) would then have been selectively neutral and potentially capable of fixation in the population.

C. Ovalbumin — Pseudogenes in Embryo?

The chicken ovalbumin gene family provides an interesting comparison with the primate adult globin genes. As described above, the primate δ-gene might be thought of as an intermediate on the road to becoming a pseudogene, since it is either silent or expressed at much reduced levels. A similar situation seems to occur in the ovalbumin gene family. Upstream of the major ovalbumin gene there are two other closely similar ovalbumin-like genes, X and Y.[49,50] The three genes share DNA sequence homologies and have a common structure of eight exons spanning 6 to 8 kilobases (kb) of DNA, suggesting that the family evolved by duplication of a single ancestral gene (Figure 3). The two upstream genes X and Y are sometimes referred to as pseudogenes; however, they are expressed in the chick oviduct in response to estrogen, and, therefore, are not really pseudogenes in the strict sense. The minor ovalbumin-like genes X and Y might have a specific role in the developing or mature oviduct: alternatively, they may be semiredundant gene copies or presumptive pseudogenes that are diverging from the parental gene, but have yet to become completely silent transcriptionally. Genes X and Y are transcribed at significantly lower rates, and their transcription stimulation in response to estrogen is much less than the authentic ovalbumin gene. The ratio of hormone responsiveness of the three genes, ovalbumin:Y:X, is of the order 100:10:1.[51] The molecular basis for these very different transcription rates is not known; DNA sequencing of the 5′ flanking regions of the three genes has failed to reveal any obvious features that could relate to their differential hormonal responsiveness.[52] Presumably, some other higher order feature of chromatin structure must therefore be involved.

D. Interferon — Transcribed Pseudogenes?

Studies of the human leukocyte interferon gene family have shown that at least some members of this family comprise linked gene clusters and that the family contains pseudogenes.[53,54] Interferon pseudogenes were identified on the basis of their weak hybridization with a cDNA probe and their inability to form R loops with poly A^+ RNA from interferon producing leukocytes.[53] DNA sequence analysis confirmed that at least one of these weakly hybridizing regions was a pseudogene.

Sequence analysis of a further cluster of interferon genes revealed a pair of closely similar sequences with ~96% homology in both their coding and flanking regions.[54] However, one of the two sequences (LeIFN-L) had a termination codon in its signal peptide region indicating

that it was in fact a pseudogene. The 3′ end of a third interferon related sequence was also found in this cluster. It, too, was most likely a pseudogene since it showed weak homology to the other sequences and was interrupted by a number of insertions and deletions that destroyed any potential reading frame.[54]

Analysis of leukocyte interferon cDNA clones interestingly suggests that at least one interferon pseudogene is transcribed. One of eight cDNA clones isolated corresponded to a transcript that contained a frameshift insertion and several in-frame termination codons.[55] A sequencing error cannot be ruled out to account for this observation, but it may indeed be the case that a defective interferon "pseudogene" is transcribed. At first sight this is puzzling, since one might expect there to be strong selection to prevent expression of mutant transcript. However, since leukocyte interferons seem to be encoded by at least eight closely related genes,[55] a proportion of inactive transcripts might be tolerated, and consequently selection to silence any pseudogene might be reduced. Thus, in contrast to the primate δ-globin gene, where transcriptional silencing seems to precede the acquisition of coding region mutations, interferon pseudogenes may be evolving by the opposite route where a deleterious coding region mutation is the primary inactivating event.

E. Immunoglobin Pseudogenes — Casualties of Antibody Diversity?

The genes that encode the immunoglobulin (Ig) heavy and κ and λ light chains constitute three multigene families, each containing a large number of genes encoding the antigen binding variable (V) and joining (J) regions of the Ig chains and one or a few genes encoding the constant (C) regions that determine antibody class. Pseudogenes have been found in all these gene families, and while only two are C-region pseudogenes, among the V- and J-region genes pseudogenes occur with comparatively high frequency. This seems to reflect the different evolutionary constraints on these two sets of genes. In order to generate a wide diversity of antibodies, the immune system has evolved so as to contain a large repertoire of V-region genes, each gene differing by ~12% at the nucleotide level. This degree of diversity may unavoidably result in some genes acquiring deleterious mutations, and as selection operates on the V-gene repertoire as a whole, there can be no selective elimination of pseudogenes. Thus, they will persist in the genome until they either drift to complete nonhomology or until they are reactivated by gene conversion. In contrast, C-region genes are constrained to be nonvariant, since they encode the effector functions of the Ig molecule, and this seems to be reflected in the infrequent occurrence of pseudogenes among their number.

1. V- and J-Region Pseudogenes

V-Region pseudogenes have been found for mouse and human heavy chains[56-61] and human κ light chains.[62] Two of the pseudogenes studied, a human V_κ[62] and a mouse V_H[56], have diverged by ~25% from related active V genes and contain numerous small insertions and deletions, termination codons, and aberrant signal sequences for DNA joining with J or diversity (D) region segments. They are clearly examples of pseudogenes that have been inactive for some time. In contrast, all other heavy chain V-region pseudogenes have very minor defects, often only a single termination codon or a short deletion.[57-61] It has been estimated that up to 40% of the human and mouse V_H genes analyzed are pseudogenes by these criteria.[61] These pseudogenes seem to represent very recently diverged members of the V_H gene repertoire. This repertoire appears to change very rapidly since changes can be found even between two different mouse strains. The V_H genes of the C57BL/6 mouse that encode its restricted response to a simple hapten NP ([4-hydroxy-3-nitrophenyl] acetate) have homologues in the BALB/c mouse. The BALB/c V_H genes share >90% homology with the C57BL/6 genes, yet three of the five contain termination codons, and the remaining two have nucleotide changes in antigen binding regions and probably no longer contribute

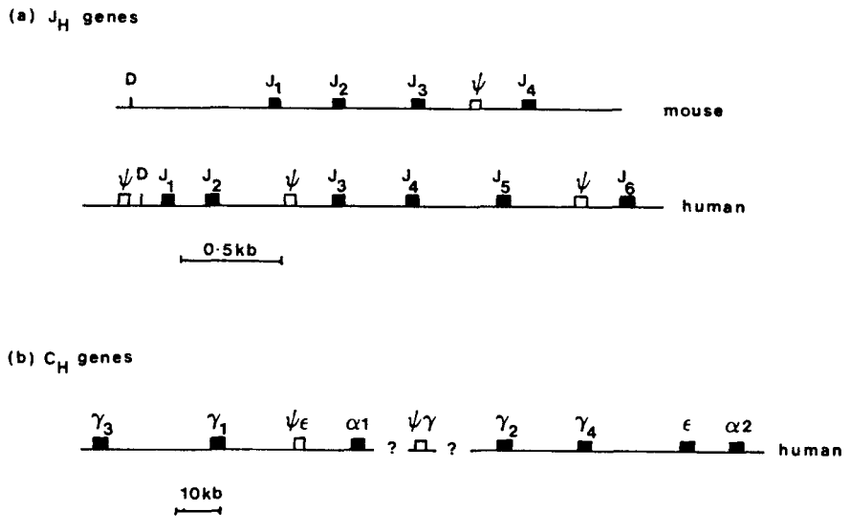

FIGURE 4. Arrangement of immunoglobulin heavy chain J region and C region genes.[69,74]

to an anti-NP response.[59,63] Thus, during V-region divergence, pseudogenes appear to arise with high frequency. While a proportion will probably continue to diverge away, it seems plausible that gene conversion may allow the less divergent pseudogenes to be re-recruited into the V-region repertoire.

The J-region genes of mouse κ and λ light chains[64-67] and mouse and human heavy chains[68,69] contain examples of pseudogenes that show a similar range of defects as V-region pseudogenes, including aberrant DNA joining and RNA splicing sequences (Figure 4). Since there are multiple J-region genes for each Ig chain, the presence of one (mouse J_H, J_κ, and J_λ) or a few (human J_H) pseudogenes in the cluster is unlikely to be unduly deleterious. This seems particularly the case in light of the high frequency of aberrant rearrangements associated with the V-J and V-D-J joining processes themselves, which the immune system already tolerates.[70]

2. C-Region Pseudogenes

In addition to two unusual processed pseudogenes derived from human λ light chain and C_ϵ region genes (Section III.E), two other C-region pseudogenes have been found in the human heavy chain cluster (Figure 4). The first, a C_γ pseudogene,[71,72] has a defective RNA splice signal and lacks a 5′ "switch" sequence required for the juxtaposition of a V-D-J region and hence for its expression. There is some evidence that it lies between the two linked sets of γ-γ-ε-α C-region genes in the heavy chain cluster,[73,74] though the mechanism of its origin remains obscure.

The second pseudogene, a truncated C_ϵ gene, lies in the upstream set of C-region genes and appears to be a casualty of an aberrant DNA rearrangement involving the heavy chain "switch" sequence.[74-76] It lacks its first two exons and apparently results from a recombination between its 5′ switch region and partially homologous sequences in its second intervening sequence. Since there is a duplicated set of γ-γ-ε-α C-region genes in man, the loss of one C_ϵ gene can clearly be tolerated. It is interesting that a mechanism of DNA rearrangement peculiar to Ig genes, the class "switch" sequences, has provided yet another route by which pseudogenes may be generated.

F. Major Histocompatibility Complex Pseudogenes — a Misnomer?

Pseudogenes have been found among the major histocompatibility complex (MHC) genes

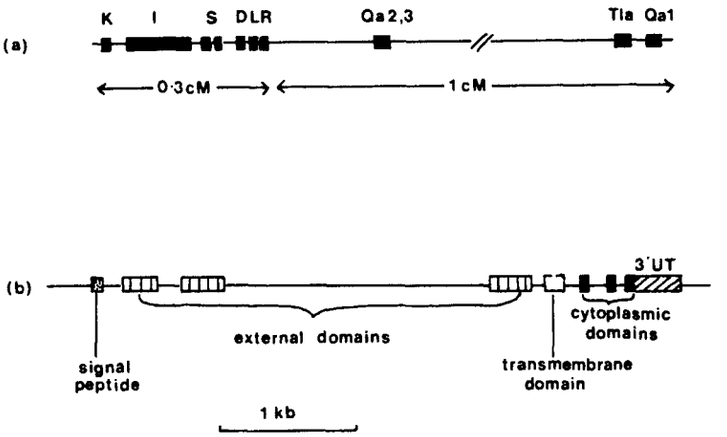

FIGURE 5. (a) Arrangement of genes in the mouse major histocompatibility complex. The K and D regions encode the major polymorphic antigens; the Qa 2,3 region encodes lymphoid differentiation antigens. (b) Organization of the mouse 27.1 gene. Exons are labeled according to their corresponding protein domains.[77]

that encode class I membrane antigens, of both mice[77,78] and man.[79] Like immunoglobulin genes, the MHC comprises a large gene family encoding proteins characterized by their high degree of sequence diversity. Thus, it might be expected that the MHC would also be a common source of pseudogenes; however, as will be seen below, it has not always been possible to define clearly what is and what is not a MHC pseudogene.

The human gene LN-11A, isolated by its homology to a class I cDNA probe, is clearly a pseudogene since it has a number of termination codons and deletions in all protein coding domains.[79] However, it retains extensive homology to a cDNA, corresponding to the HLA-B7 antigen, and has a similar exon structure to other class I genes.

Two presumptive mouse MHC pseudogenes have been isolated and both localized to the Qa2,3 region of the H-2 complex, which encodes lymphoid differentiation antigens. The 27.1 gene from BALB/c mice[77] shows striking homology to H-2 cDNA clones and to the sequence corresponding to an H-2K^b antigen, and has a characteristic class I gene structure (Figure 5). However, its sequence has two termination codons (in the transmembrane and second cytoplasmic exons), an aberrant splice acceptor for the second cytoplasmic exon, and nucleotide substitutions that introduce a charged aspartic acid residue into an otherwise hydrophobic transmembrane domain. This gene clearly could not encode a class I antigen that would insert into the cell membrane, and thus was judged to be a pseudogene.

Subsequent work suggested that this simple explanation was not the whole story. Analysis of cDNA clones corresponding to H-2 class I specific transcripts from the liver of SWR/J mice, revealed that ~20% of these transcripts had an unusual 3′ coding region with termination codons and substitutions that introduced charged residues in the transmembrane domain.[80] Thus, these unusual transcripts would encode a truncated class-I-like antigen lacking its transmembrane and cytoplasmic exons. It was suggested that such an antigen might be secreted rather than inserted into the cell membrane. These transcripts have been found in a number of different mouse strains, but they are expressed only in the liver[81] — and an antigen of the predicted structure has indeed been shown to be secreted by the liver and to be present in serum.[82] The gene Q10, which probably encodes this transcript, has been isolated from C57BL mice;[78] it has the same set of "mutations" found in the SWJ/R transcript and a structure that is closely similar to the BALB/c 27.1 "pseudogene". Thus, Q10 and 27.1 may not strictly be "pseudogenes" after all, since they can encode a secreted form of a class I antigen.

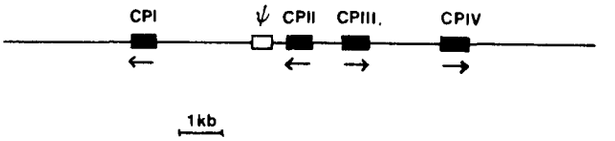

FIGURE 6. The cuticle protein gene cluster in *Drosophila*. Arrows indicate the direction of transcription of the four genes CPI, CPII, CPIII, and CPVI.[86]

A further twist to this comes from the observation that the Q10 gene is a potential donor for a gene conversion event that occurred at the H-2K locus in the C57BL mouse. The H-2K^{bm1} variant gene contains a clustered set of seven base changes, which corresponds exactly to a sequence found within the Q10 gene.[83-85] This supports the idea that gene conversion may play a significant role in generating the high degree of polymorphism found at the H-2K and H-2D loci.

Therefore, while genes such as Q10 and 27.1 may be classed as pseudogenes from the point of view of their ability to encode classical membrane bound class I antigens, they may in fact serve to encode liver specific secreted forms of these antigens. In addition, they may act as a source of diversity, through gene conversion, for the major polymorphic H-2K and H-2D loci. So, some pseudogenes at least are not purely useless components of the genome.

G. Cuticle Proteins — an Insect Pseudogene

The *Drosophila melanogaster* genome is characterized by its small size and a general absence of pseudogenes among those gene families studied to date. However, a single example of a pseudogene has been found in the cutical protein gene cluster, in band 44D of chromosome II.[86] The cluster spans 7.9 kb and comprises four closely spaced genes, each encoding a distinct cuticle protein (CPI-IV), arranged in two divergently transcribed sets of two genes (Figure 6). A fifth gene lies between the genes for CPI and CPII, but it is not transcribed in vivo and has a number of structural features that mark it out as a pseudogene — a deletion that removes the upstream promoter sequence, an aberrant splice acceptor site, and nucleotide substitutions, insertions, and deletions that cause frame shifts and premature termination. A clue to the origin of this pseudogene came from a comparison of its DNA sequence with those of the two flanking genes CPI and CPII. Upstream of nucleotide 468, the pseudogene most closely resembles gene CPI, while its 3′ end is more homologous to CPII. This suggests that an unequal cross-over event involving the two related genes CPI and CPII gave rise to a hybrid gene, which, either as a consequence of the cross-over event itself or subsequently, became a nontranscribed silent pseudogene and then gradually acquired its present array of deleterious mutations. This cuticle protein pseudogene, therefore, points once again to the involvement of unequal crossing over in the generation of pseudogenes.

H. A Mixed Bag of Further Examples

A number of further examples of pseudogenes have been reported; their main interest lies in their extending of the pseudogene repertoire to include RNA genes other than 5S RNA, and to organisms other than mammals. They, therefore, are an indication of the ubiquity of pseudogenes, and suggest that we can expect to find pseudogenes in many different gene families, from organisms throughout the plant and animal kingdoms.

1. Rat tRNA Pseudogenes

A set of clones corresponding to a repeating unit of at least 13.5 kb containing a tRNA gene cluster has been isolated from rat genomic libraries.[87,88] The cluster comprises genes

for $tRNA_{Asp}$ (GAU/C) and also $tRNA_{Gly}$ (GGA/G) and $tRNA_{Glu}$ (GAG) related sequences. In contrast to the $tRNA_{Asp}$ gene, which is highly conserved between different clusters, the $tRNA_{Gly}$ gene in all clusters and the $tRNA_{Glu}$ gene in three out of five clusters studied are divergent and contain substitutions and deletions that render them transcriptionally silent. Interestingly, in a different (Donryu) rat strain, these tRNA gene clusters appear not to contain pseudogenes,[89] suggesting that they have arisen in Sprague-Dawley rats relatively recently in evolution. It has been argued that the $tRNA_{Asp}$ gene encodes an essential abundant tRNA, while the defective tRNA pseudogenes are derived from minor tRNAs that have been passive passengers in the duplication of the cluster[87] — an analogous situation to the duplication of linked genes and pseudogenes in the goat β-globin gene cluster.

2. Pseudogene Fragments

An unusual H1A histone gene has been found in *Xenopus laevis*.[90] This gene lies adjacent to a functional H4 histone gene, and while it selects histone H1A mRNA in a hybrid selection-translation assay, it is not itself transcribed following injection into oöcytes. Nucleotide sequencing revealed that the 5′ third of this gene was absent, being replaced by a very AT-rich sequence; sequences corresponding to the 5′ region were not present within 1.8 kb of upstream sequence. The remainder of the histone pseudogene is highly conserved relative to other H1A sequences, suggesting that it is of recent evolutionary origin — possibly a DNA rearrangement during the amplification of histone genes in *Xenopus* resulted in loss of the 5′ portion of this gene.

Two other examples of truncated pseudogenes have been reported; a 5′ truncated human β-tubulin pseudogene,[91] and a 3′ truncated 5S RNA gene fragment from *Neurospora crassa*.[92] However, these two pseudogene fragments appear to be dispersed in the genome and not linked to any related genes. While their origin remains somewhat of a mystery, it seems simplest to explain them as casualties of DNA deletions or inversions or other genome rearrangements.

3. Plant and Slime-Mould Pseudogenes

A single example of a pseudogene in the slime mould *Dictyostelium discoideum* has been found among its actin genes.[93,94] This pseudogene (actin2-sub2) lies downstream of another actin related sequence; it is not expressed in vivo and encodes a protein that diverges considerably from the normal actin N-terminal sequence.

An unusual leghemoglobin gene has been found in soybean.[95] It is presumed to be a pseudogene since it is not transcribed in vivo, and its sequence does not correspond to any known leghemoglobin protein. Furthermore, it is about twice the size of other leghemoglobin genes, two of its intervening sequences being considerably longer than their counterparts in other genes. Mutations in its 5′ flanking sequence in regions associated with transcription initiation are though to be responsible for its inactivity in vivo.

A final example comes from the chloroplast genome of an unicellular flagellate *Euglena gracilis*. In the 16S-23S ribosomal RNA transcription unit, the spacer region between the 16S and 23S genes contains two tRNA genes, encoding $tRNA_{Ile}$ and $tRNA_{Ala}$. The leader sequence upstream of the 16S gene also bears significant homology with this spacer segment;[96] however, both of the two tRNA homologous sequences in the 16S leader have acquired mutations that render them nonfunctional. It appears that the 16S leader sequence originated as a partial duplication of the 16S-23S spacer, an event which perhaps also silenced the leader sequence transcriptionally.[96] Thus, freed from selection, it has diverged such that its two erstwhile tRNA genes have become pseudogenes. Thus, even the prokaryote-like genomes of chloroplasts are not immune to the occurrence of pseudogenes.

III. DISPERSED PSEUDOGENES

A. Introduction

In the previous section, we have seen that many pseudogenes are linked to their functional counterparts, lying silent within an active gene cluster. In addition to this type of pseudogene, a number of gene families (mouse globin, sea urchin, and *Drosophila* histones) revealed examples of presumptive pseudogenes that were dispersed in the genome, in some cases demonstrably to different chromosomes from their parent families. By analogy with bacterial transposons, their dispersion in the genome was explained by invoking some DNA transposition event or the possible involvement of retroviral long terminal repeat elements. However, studies on gene families encoding mammalian small nuclear RNAs (snRNAs) and the *Alu* interspersed middle repetitive elements[98-100] suggested an alternative model for the origin of many of these dispersed sequences — the incorporation of sequences contained in RNA into breaks in chromosomal DNA either via an RNA molecule itself or its reverse transcript. While reverse transcription had been known for some time to be involved in the retrovirus life cycle, its application to normal cellular RNAs was a new departure.

The proposal of such a mechanism set the scene for the subsequent discovery of other dispersed pseudogenes, which bore the hallmarks of being derived from mRNA molecules that had been incorporated into genomic DNA. It now appears that such RNA derived pseudogenes outnumber by far those pseudogenes that arose through DNA related gene duplication or transposition events, and in some cases they comprise the vast majority of members of a gene family. These pseudogenes are variously referred to as "processed" (pseudo)genes,[101] denoting the presence of features associated with RNA processing (the removal of intervening sequences and polyadenylation) or "retroposons",[102] denoting their transposition via RNA and its reverse transcription into DNA. This latter term also encompasses *inter alia* RNA derived snRNA pseudogenes and mammalian short and long middle repetitive elements.

B. Orphons

The name "orphon" was coined to describe single, dispersed members of the tandem multigene families that encode histones and ribosomal RNAs in yeast, *D. melanogaster*, and sea urchins.[103] Orphons were detected as variant sequences flanked by restriction enzyme sites not found in the major tandem repeat of the parent multigene family. A number of orphons encoding sea urchin histone genes were cloned and studied in more detail.

One histone H3 orphon from the sea urchin *Lytechinus pictus* retains greater than 98% homology in both its coding and 5′ and 3′ flanking regions with its counterpart in the parent cluster.[103] This is an indication that it was formed relatively recently in evolution. It was suggested that orphons arose from single elements of the tandem cluster that were excised during the unequal cross-over events that mediate sequence homogenization of the repeats. Misaligned tandem repeats might form looped out segments that would be potential targets for excision during repair processes. Reintegration of excised DNA segments at dispersed sites in the genome would create orphons.[103] The origin of the histone H3 orphon seems best explained by this type of DNA mediated mechanism. However, the structure of two other orphons, a *Drosophila* histone pseudogene,[104] and an early histone H2B orphon from the sea urchin *Strongylocentrotus purpuratus*,[105] suggests that they were more likely derived through reverse transcription of histone mRNAs. The *S. purpuratus* orphon has a 32-base-long polyA tract at its 3′ end (even through histone mRNAs are not normally polyadenylated) and is flanked by 6-nucleotide direct repeats indicative of its integration at a staggered chromosomal break. Thus, it appears that both DNA mediated and RNA mediated mechanisms may be responsible for the generation of "orphon" pseudogenes of tandem multigene families.

C. Mouse α-Globin Pseudogenes

The mouse α-globin gene family comprises two adult genes and one embryonic gene linked together on chromosome 11 and two pseudogenes α-ψ3 and α-ψ4 which, in contrast to all other globin pseudogenes, are dispersed to different chromosomal loci.[15-18] Using mouse x Chinese hamster hybrid cell lines, α-ψ3 and α-ψ4 have been mapped to chromosomes 15 and 17, respectively.[17,18]

The α-ψ4 pseudogene closely resembles other adult α-globin genes and retains both its intervening sequences.[18] The simplest explanation for its dispersion being one of chromosomal translocation or DNA transposition. This latter mechanism is perhaps made more plausible by the observation that goat, human, and mouse α-globin genes are flanked by vestigial direct repeat elements, possibly indicative of an ancient transposition event involved in the transfer of the ancestral α-globin to a new chromosomal location.[26] A similar event may therefore have been involved in the dispersal of the α-ψ4 pseudogene.

In contrast to α-ψ4, α-ψ3 has clearly lost both intervening sequences, and, in addition, it has a number of frame shift deletions and insertions in its erstwhile coding region.[15,16] The loss of both intervening sequences is precise, and it is, therefore, difficult to envisage how this could have occurred as the result of random deletions at the DNA level. This prompted the suggestion that an RNA intermediate was involved in its generation, either by gene conversion or recombination of an α-globin mRNA with a preexisting gene,[15] or through the involvement of a retroviral intermediate.[17,106] The observation that an intron-containing globin gene, incorporated into a Spleen Necrosis Virus derived retroviral vector, is converted into an intron-less gene after one round of productive infection through RNA intermediates[107] suggests that retroviruses could be involved in pseudogene formation. The α-ψ3 gene is flanked by mouse retroviral-like long terminal repeat (LTR) elements;[108] however, they are not orientated so as to promote transcription of the α-globin sequences. Furthermore, the flanking LTR elements are not found in all mouse stocks,[109] suggesting that their juxtaposition to α-ψ3 was fortuitous, rather than due to their involvement in the formation of this pseudogene. LTR-like elements are also found flanking an immunoglobulin C_ϵ processed gene,[110] though once again it is not clear whether this is fortuitous or a reflection of their role in pseudogene formation.

In the light of subsequent discoveries of intron-less processed pseudogenes, it is logical to propose that α-ψ3 was generated by the straightforward incorporation of an α-globin mRNA into chromosomal DNA. However, the extent of homology between α-ψ3 and α-globin genes does not coincide precisely with those sequences that are present in mRNA; at the 3′ side homology ends abruptly three nucleotides downstream of the polyadenylation signal, and upstream homology to the functional gene extends at least 350 nucleotides 5′ to the start of mRNA transcription. Clearly, a normal α-globin mRNA could not have been the substrate that gave rise to the α-ψ3 pseudogene. However, an aberrant transcript derived from a promoter upstream of the usual transcription start point, possibly an RNA polymerase III transcript,[109] could have been the source of this pseudogene. It has further been suggested that α-ψ4 may also derive from an RNA transcript that was reverse transcribed before removal of its intervening sequences.[109] Other examples of processed pseudogenes that appear to be the products of aberrant transcripts will be discussed below (Section III.E).

D. Small Nuclear RNA Pseudogenes

Small nuclear RNAs (snRNAs) are a family of abundant discrete RNAs found associated with proteins in ribonucleotprotein particles in the nuclei of eukaryotes. U3 RNA is restricted to the nucleolus and has been implicated in the processing of rRNA while U1, U2, U4, and U6 RNAs are nucleoplasmic and are thought to be involved in splicing and processing of mRNA precursors.[111] Each snRNA species is apparently encoded by ~100 to 2000 genes that are dispersed in the genome, these estimates being based on solution hybridization

experiments and on the frequency of clones in bacteriophage genomic libraries that hybridize to snRNAs.[112-117] However, sequence analysis of a number of cloned fragments hybridizing to U1, U2, U3, U4, or U6 snRNAs revealed that a vast majority contained snRNA pseudogenes;[112-122] perhaps as much as 80 to 90% of genomic sequences are pseudogenes.

These pseudogenes are of several different types, classified on the basis of their structural characteristics. Some encode full length snRNAs, but contain scattered base substitutions and insertions.[113,115,120,122,123] In view of the virtual invariance of snRNAs in evolution, it appears unlikely that these sequences encode functional snRNAs. These pseudogenes also show significant homology to bona fide snRNAs in their flanking regions, suggesting they were generated by divergence of duplicated snRNA genes. The significantly greater conservation of "coding" as opposed to flanking sequences even in the pseudogenes perhaps indicates that gene conversion has also been operating in this dispersed gene family.[123]

Other snRNA pseudogenes, in contrast, have characteristics that led to the suggestion that they were generated by the incorporation of reverse transcripts of snRNAs into the genome at either blunt or staggered chromosomal breaks.[124] A number of different mechanisms for the integration process have been elaborated to take into account the different flanking structures of these pseudogenes; these will be discussed more fully below (Section III.G). These pseudogenes are characterized by only containing sequences that are present in snRNA molecules themselves; their homology with snRNA genes begins precisely at the snRNA 5′ end and extends either to the 3′ end of the snRNA or shows a slight or more severe degree of 3′ truncation. Some, but not all, pseudogenes are flanked by short direct repeats of 16 to 21 nucleotides; the longest snRNA pseudogenes additionally have short 3′ A-rich segments at their ends or preceding a 3′ direct repeat sequence.[112,116,119,120,123,124] Since polyA is not normally present on snRNAs, such pseudogenes must have been derived from aberrantly polyadenylated molecules.

Many pseudogenes corresponding to U2, U3, and U4 snRNAs are severely truncated at their 3′ ends relative to the mature snRNA.[114,119,123-127] It appears that this truncation may be related to the way in which these pseudogenes were generated. The U3 pseudogenes are all truncated at nearly identical positions in the U3 sequence.[123-126] Interestingly, U3 snRNA acts as a self-priming template in vitro for AMV reverse transcriptase, generating a 74 nucleotide cDNA that is of closely similar length to the U3 pseudogenes.[126] This strongly suggests that these truncated pseudogenes might have been generated by the integration of similar self-primed cDNAs formed in vivo. The insertion could be at a staggered or blunt-ended chromosomal break since pseudogenes are found both with and without flanking direct repeats, though the absence of flanking repeats could also be due to subsequent sequence divergence.

The U2 pseudogenes, however, are truncated at a number of different sites in the snRNA sequence. All are flanked by direct repeats, 16 to 21 nucleotides in length, and in some cases the downstream repeat overlaps the 3′ end of the pseudogene by up to five or six bases.[126] It was suggested that the truncations in U2 pseudogenes reflect incomplete copying of an inserted cDNA during repair second strand synthesis.[127] If second strand synthesis terminated as a result of limited base pairing between the cDNA insert and downstream genomic sequences, this would explain both the 3′ truncation of pseudogenes and the overlap between pseudogene 3′ ends and their flanking direct repeats. Alternatively, premature termination of the second cDNA strand could have resulted from secondary structure within the cDNA or cDNA-snRNA hybrid insert that hindered the progress of the polymerase.[127] A further explanation for the 3′ truncation can be envisaged if limited homology between an snRNA molecule itself (rather than a cDNA) and a 3′ overhang of a staggered break in genomic cDNA were responsible for the priming of the first cDNA strand synthesis. This would give a mechanism for truncated pseudogene formation essentially similar to that for U3 pseudogenes, but with exogenous rather than internal self-priming of cDNA synthesis.

E. Processed Pseudogenes

Like some snRNA pseudogenes, processed pseudogenes have sequence characteristics that suggest they were derived from the incorporation of information contained in RNA transcripts into new chromosomal locations in the genome. All processed pseudogenes are related to protein coding genes, but lack the intervening sequences found in the functional parent gene. Most also have oligoA tracts correctly positioned relative to a polyA additional signal at their 3′ ends — a feature that further points to their mRNA origin. In addition, the dispersion of a number of processed pseudogenes to different chromosomes from their parent genes has been demonstrated using interspecies hybrid cell lines.[17,18,109,128-132]

1. Structure

Processed pseudogenes fall into two types. Some are colinear with normal cellular mRNAs, starting at the 5′ mRNA cap site and ending in an A-rich or oligoA stretch of 7 to 36 nucleotides and are flanked by direct repeat sequences of 9 to 25 bases. The first example of this type was a human β-tubulin pseudogene.[133] Subsequently, similar processed pseudogenes have been found in an ever increasing number of mammalian gene families — human metallothionein,[134] dihydrofolate reductase,[131,135,136] argino-succinate synthetase,[137] β-actin,[138,139] nonmuscle tropomyosin,[140,141] and Harvey and Kirsten *c-ras* oncogenes,[142,143] rat α-tubulin[144] and cytochrome c[145] and mouse tumor antigen p53,[146] and ribosomal proteins L7, L18, and L32.[147-149] Furthermore, the rat cytochrome c[145] and human β-tubulin[150] gene families demonstrate that where different mRNAs with 3′ untranslated regions of varying lengths are produced due to the use of alternative downstream polyadenylation sites, processed pseudogenes corresponding to each of the different sized mRNAs may be found.

Other processed pseudogenes are clearly also derived from RNA molecules, since they lack intervening sequences found in parent genes and end in oligoA or A-rich tracts; however, their structures do not correspond to the normal cellular mRNAs of their parent genes. There are four examples of this type (possibly five with the inclusion of the mouse α-ψ3 globin pseudogene): (1) a human immunoglobulin λ light chain pseudogene[101] containing spliced J and C regions, but no V region (which in immunoglobulin producing cells is normally joined directly onto the J region at the DNA level); (2) a human immunoglobulin ϵ heavy chain pseudogene,[110,128] comprising only the four spliced exons of the ϵ constant region, but no variable region coding elements (V, D, or J regions); (3) a mouse myosin light chain pseudogene,[132] consisting of the five terminal exons common to both myosin alkali light chains LC1 and LC3, and lacking either of the two combinations of N terminal exons normally present in the corresponding cellular mRNAs; and (4) a mouse pro-opiomelanocortin (POMC) pseudogene[129,151] that includes only those sequences downstream of codon 67 in the most 3′ exon of this gene.

The immunoglobulin J-C_λ and C_ϵ and POMC pseudogenes end in A-rich tracts of $(CA_x)_y$ or $(GA)_x$, while the myosin light chain pseudogene has a short oligoA tract preceding an A-rich sequence; all four are flanked by direct repeat sequences. All these pseudogenes are truncated at their 5′ ends relative to their parent genes and, with the exception of the POMC pseudogene, appear to have arisen from transcripts that initiated anomalously in the intervening sequence immediately upstream of those exons found in the pseuodgene. It is possible that these pseudogenes are derived from RNA polymerase III transcripts rather than the RNA polymerase II transcripts that are the precursors of normal cellular mRNAs.[109] The mouse globin α-ψ3 pseudogene also appears to be derived from an aberrant transcript, and while it is not 5′ truncated and lacks flanking repeats or A-rich tracts in the surrounding sequences thus far sequenced, it may be a further example in this class of pseudogene.[15,16]

2. Origins

Since processed pseudogenes are found in all, or most, individuals of a species and are

transmitted as inheritable components of the genome, they must have originally arisen in cells of the germ line. It follows from this that processed pseudogenes would be expected to be formed only from those genes that are expressed in germ line cells. Indeed, those processed pseudogenes that are essentially colinear with cellular mRNAs do seem to be derived either from "housekeeping" genes common to all cell types or from genes that might be preferentially expressed in the germ line (e.g., tumor antigen p 53, *c-ras* oncogenes).

In contrast, those processed pseudogenes that appear to be derived from aberrant transcripts originate from genes that are not normally expressed in the germ line since they encode products of highly differentiated somatic cells (i.e., lymphocyte immunoglobulin chains, erythrocyte α-globin, muscle myosin light chain, and pituitary hormone precursors). Presumably, the aberrant nature of the transcripts from which they appear to be derived is a reflection of their abnormal transcription in the germ line.

The human actin genes further exemplify very clearly this point that processed pseudogenes are usually only found in gene families that are expressed in the germ line. Processed pseudogenes appear to account for a large part of the multigene family encoding cytoskeletal β- and γ-actins, which are expressed in all cells;[138,139,152] in contrast, the α-cardiac and α-skeletal muscle actins, products of differentiated somatic tissues, are encoded by single copy genes with no related processed pseudogenes.[152] Several other gene families, including those for mouse ribosomal proteins L7, L18, and L32,[147-149] human nonmuscle tropomyosins,[141] a β-tubulin isotype,[150] and arginosuccinate synthetase,[137] comprise a single active gene and anything from 3 to 15 processed pseudogenes. The proportion of pseudogenes in any one family may be a reflection of the relative levels of transcription of the parent gene in the germ line.[150]

While the vast majority of processed pseudogenes have been found in mammalian species, a single calmodulin processed gene has been found in chickens,[153] and some at least of the histone orphons of sea urchins are derived from reverse transcribed mRNAs.[105] In addition, the F elements of *D. melanogaster* appear to be dispersed by the integration of polyadenylated RNA transcripts.[154] Therefore, the mechanisms responsible for the generation of processed pseudogenes are not exclusive to mammals, though some features of mammalian gamete production and germ line transcription may make them peculiarly susceptible to the formation of processed pseudogenes.

3. Age and Divergence

Unlike pseudogenes linked to their functional counterparts that may be as little as 75% homologous to their parent genes, processed pseudogenes seem to show strikingly high (90 to 99%) homology to the genes from which they derive. It suggests that they have arisen relatively recently in evolutionary history.

The myosin light chain pseudogene, for example, shares 99% nucleotide homology with the active gene and, furthermore, is found in *Mus musculus,* but not the related species *Mus spretus,* which diverged less than 7 MYr ago.[132] Similarly, a set of three human β-tubulin pseudogenes show homologies of 91, 92, and 97% with their parent gene, and it has been estimated that they have diverged around 13.4, 10.7, and 4.4 MYr ago, respectively.[150] A further indication of the relatively recent origin of some processed pseudogenes is the observation that a human dihydrofolate reductase pseudogene, hDHFR-ψ1, which has perfect homology to the functional gene, is only present in certain individuals of the species and shows an imbalance in its frequency in different racial groups.[131]

Thus, processed pseudogenes appear to be recent genomic acquisitions. However, because the examples of processed pseudogenes studied to date have been detected and isolated using DNA hybridization probes, the sample may be somewhat biased towards those that are little diverged from their parent genes. If probes were used at high stringency, more divergent processed pseudogenes may well have gone unnoticed. Indeed, when genomic blots are

performed at reduced stringency, additional genomic sequences with weaker homology to a probe can often be seen.[150] Also, an example of a highly divergent processed pseudogene with only 77 to 80% nucleotide homology to an active β-tubulin gene has been isolated from a human genomic library.[91,155] Therefore, genomes may in fact contain whole series of processed pseudogenes that have become progressively more and more divergent from their parental genes, gradually "fading out" into the genomic background. Those that have been detected using relatively stringent hybridization probes may be little more than the tip of an iceberg with respect to the total number of processed pseudogenes formed throughout evolutionary time.

4. Expression?

It has been assumed that processed pseudogenes will have been transcriptionally inactive since their time of formation. With the exception of the mouse α-ψ3 globin pseudogene, which retains upstream RNA polymerase II promoter sequences, all other processed pseudogenes are coterminal with their corresponding mRNAs and thus lack transcriptional promoters. While it is not impossible to envisage integration occurring correctly downstream of an RNA polymerase II (or III) promoter, it seems unlikely that this could occur without adversely affecting the activity of other genes. Thus, it is simplest to assume that, almost by definition, processed pseudogenes will have been incapable of expression from the time of their formation onwards, even though initially they will have had intact coding regions and only subsequently acquired the deleterious mutations characteristic of "classical" pseudogenes. Consistent with their transcriptional inertness, pseudogenes may show a higher degree of DNA methylation than their functional counterparts.[117,149] Also, unlike the parent gene, dihydrofolate reductase processed pseudogenes are not amplified during acquisition of methotrexate resistance in the human HeLa cell line.[131]

Given this, it is, therefore, somewhat surprising that a processed calmodulin "pseudogene" appears to be specifically expressed in chicken muscle.[153] However, clarification of this observation awaits the nucleotide sequence of regions flanking this processed gene and a more detailed structural analysis of the reported tissue specific transcript.

The vast majority of processed pseudogenes, however, seem to be transcriptionally silent "passenger" components of the genome, and being freed from any selection pressure, they must accumulate base changes until they bear very little relation to their parent genes. While theoretically the protein coding information they contain could be contributed back to the parent gene through the process of gene conversion, this would only be a viable event before the processed gene had acquired too many deleterious mutations. However, such events might explain how intervening sequences may be removed from genes during their evolution, as appears to have occurred in a rat preproinsulin gene.[156]

F. Other Retroposons

As has been mentioned in passing, both snRNA pseudogenes and processed pseudogenes share structural features with a number of dispersed repeated sequences in mammalian genomes[157] — the short interspersed primate *Alu* elements,[98,99] mouse B1 and B2 sequences,[158] rat R.dre.1 repeat elements,[159] the long interspersed *Kpn1* family of primates[160] and its rodent equivalent[161] as well as the F sequences in *D. melanogaster*.[154,162] All these elements appear to have been generated by reverse transcription of RNA intermediates, a feature which has earned them the name of retroposons.[102] They have been the subject of a recent detailed review[163] and will therefore not be discussed at length here.

G. Models and Mechanisms

In the preceding sections, the basic mechanism whereby snRNA and processed pseudogenes are formed has been taken as the insertion of an mRNA or its cDNA copy into a

staggered (or blunt) break in chromosomal DNA and subsequent repair of single stranded regions. While this outline mechanism has gained wide acceptance, it has been considerably more difficult to define in greater detail the precise series of molecular events that give rise to these pseudogenes. Since the only information concerning their mechanism of origin derives from their flanking sequence organization, perhaps the nearest it is possible to come to defining a mechanism will be to describe an outline "theme" and to elaborate a number of plausible "variations".

Any model for the formation of these pseudogenes must address a number of questions: what is the polymerase responsible for the reverse transcription? How is the reaction primed? Where and how do the insertions occur in the genome? Is the inserted molecule an RNA or a cDNA (or an RNA-cDNA heteroduplex)?

The reverse transcriptase activity responsible for the formation of these RNA derived pseudogenes could have come from an endogenous retrovirus or a transient germ line infection by a retrovirus.[126] However, given the prevalence of processed pseudogenes and other retroposon-like sequences in the genome, it seems equally possible that they were formed as the result of some secondary activity of a normal cellular DNA polymerase since human DNA polymerase β can copy synthetic RNA templates in vitro.[164] RNA is also intimately associated with DNA as a primer molecule in DNA replication, and it seems possible that DNA polymerases or repair enzymes may not distinguish between bona fide RNA primers and mRNA or snRNA molecules that become fortuitously associated with genomic DNA. Hence, such RNA "interlopers" may become converted into DNA in part by processes that normally repair away the RNA primers of DNA replication.

The sites into which processed pseudogenes and other retroposons have integrated are often found to comprise relatively AT-rich sequences. Since such sequences are more prone to local melting of DNA strands and hence strand breakage, they might be expected to be a common source of sites for pseudogene insertion. However, these sites also often have an asymmetric distribution of A and T residues, indicating that specifically A- or T-rich sequences might be important for the insertion process itself[127] (see below). It has also been suggested that DNA topoisomerases play an important role in generating the transient breaks in DNA between which the insertion may occur.[127] Type 1 topoisomerases preserve the energy of the phosphodiester bond in a phospho-protein linkage, which in eukaryotic topoisomerases is to the nucleotide at the 3′ side of the break in the DNA duplex.[165] It is also noteworthy that processed pseudogenes are formed in germ line cells that will be undergoing meiosis. The recombination and cross-over events of the first meiotic prophase involve considerable breakage and joining of DNA strands. Hence, at this stage, the genomic DNA may be particularly vulnerable and susceptible to the introduction of exogenous molecules and hence to the formation of processed pseudogenes and similar retroposons.

Questions concerning the primer for reverse transcription and the nature of the inserted molecule (RNA or cDNA) will be discussed together in comparing the different models proposed to account for pseudogene formation. The first model, that of van Arsdell et al. for snRNA pseudogenes,[124] suggested the following sequence of events: (1) synthesis of a cDNA copy of the snRNA; (2) covalent linkage of the cDNA 3′ end to a 5′ overhang of a staggered chromosomal break; (3) second strand cDNA synthesis primed from the recessed 3′ OH of the break; and (4) ligation and repair of the ends of the break, creating flanking direct repeats (Figure 7A). Although *a priori* the snRNA itself, rather than its reverse transcript, could be the inserted molecule, the authors preferred the cDNA alternative as it obviated the need to propose mechanisms for decapping the snRNA and for the ligation of RNA to DNA. Of itself, this model does not explain how synthesis of the first cDNA strand is primed. For some severely truncated snRNA pseudogenes, this presents no problem since the snRNAs from which they derive can act as self-priming templates for reverse transcriptase in vitro;[126] if similar cDNAs were formed in vivo, they could give rise to pseudogenes as indicated in the model. However, in extending this model to full-length snRNA pseudogenes

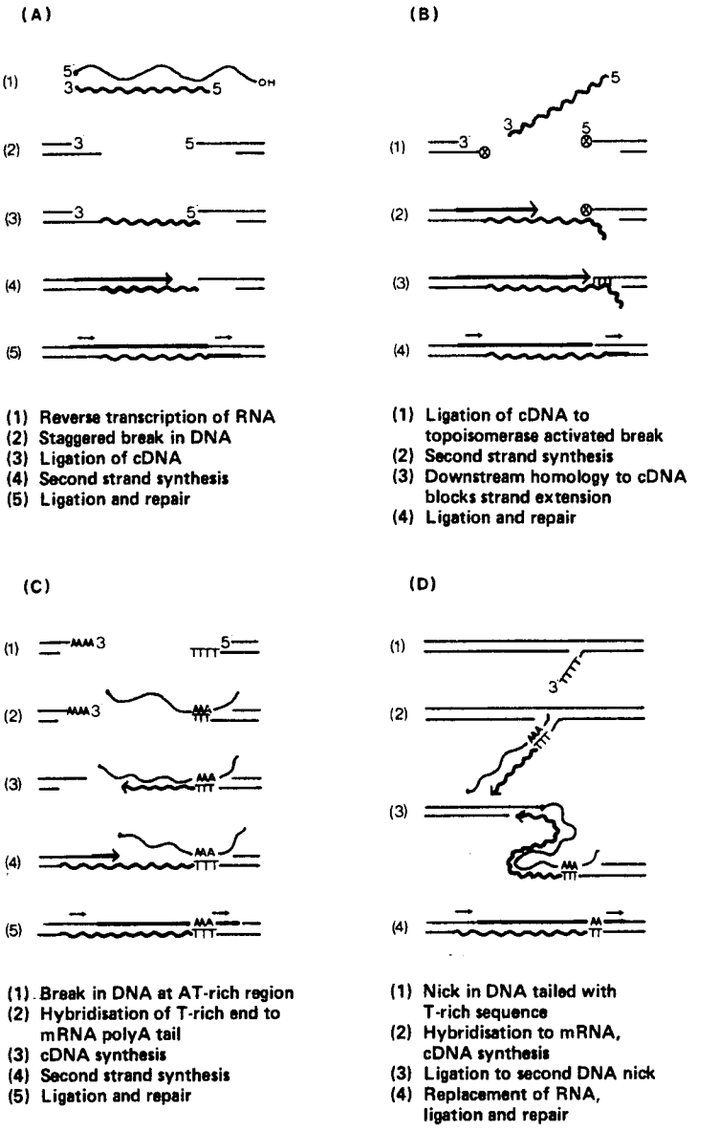

FIGURE 7. Models proposed for the generation of RNA-derived processed pseudogenes. Thin wavy lines represent RNA, thick wavy lines cDNA, and thick lines second strand or repair DNA synthesis. Flanking direct repeats resulting from the insertion are indicated by short arrows (→) and topoisomerase molecules by ⊗. (A), (B) "cDNA insertion" models for the generation of snRNA pseudogenes.[124,127] (C) "Primed insertion" model for mRNA derived pseudogenes.[109] (D) Retroposon insertion.[163]

and to processed pseudogenes that are full-length copies of mRNAs, it is presumably necessary to invoke some exogenous T-rich primer molecule for synthesis of the first cDNA strand — a somewhat less satisfactory situation.

This minimal "cDNA insertion" model has been elaborated to involve topoisomerases in the formation of staggered or blunt chromosomal breaks.[127] In addition, it was suggested that homology between the downstream direct repeat sequence and the incoming cDNA molecule might be instrumental in anchoring the cDNA relative to the staggered break[139]

Table 1
DIRECT REPEAT SEQUENCES FLANKING PROCESSED PSEUDOGENES[a]

Pseudogene	Upstream sequence	Ref.
Human immunoglobulin Cλ	AGAAGAGGATGTGAAT	101
Human metallothionein	GAGCAAAAGT-TTAAAAGGACAACAG	134
Human dihydrofolate-reductase ψ2	AAGCAAAAACCTTCCGGCC	135
$hTM_{nm}-2$	AGAAAAGAAAAACCC	140, 141
$hTM_{nm}-1$	CAAAAACTTTTGCC	140, 141
ψ1	AAAACTTATGTTT	139
ψ2	AAACCTCCTTACA	139
46 β	AAGAAGCTGAGGTGTC	133
11 β	AAAGAAATCAGAGA	91
7 β	ATACAATAAAATGCACAGGTCT	150
14 β	AAGAACAGAAAAGCTT	150
Rat α-tubulin	ATAAAAAGAGATTTTT	144
Mouse tumor antigen p53	AAAGAACTCAAGA	146

[a] The upstream flanking sequence is shown. That part which is repeated downstream of the pseudogene polyA tail is underlined. Note that A-rich sequences precede or form part of the direct repeats.

(and hence limiting second strand synthesis in the case of truncated pseudogenes)[127] (Figure 7B). This would account for the observation that flanking direct repeat sequences frequently overlap the 3′ end of truncated U2 snRNA pseudogenes or the 3′ oligoA or A-rich tails of full-length snRNA and processed pseudogenes (Table 1).

This latter observation also points to an alternative scenario for the formation of processed pseudogenes, which to a large extent overcomes the difficulty of "cDNA insertion" models in their requirement for the initial priming and synthesis of full-length cDNA copies of mRNA or snRNA molecules. The overlap between the 3′ ends of pseudogenes and their flanking direct repeats, suggests that 3′ overhangs at staggered chromosomal breaks might themselves act as the primers for the initial cDNA syntheses by virtue of their partial homology to an RNA. Thus, this scenario combines the two steps of cDNA synthesis and cDNA insertion. Since the cDNA molecule is primed by a single-stranded region of the genomic DNA itself, it is necessarily already linked into the chromosome. Subsequent steps would involve replacement of the RNA to generate a double stranded cDNA and repair and ligation of the ends (Figure 7C). A similar model to this has been independently proposed by Vanin.[109] This type of model also accounts for the frequent occurrence of processed pseudogenes among A-rich sequences in the genome, since the complementary T-rich strand could readily act as a primer for cDNA synthesis by base pairing with the poly A tail of an mRNA or aberrantly polyadenylated snRNA molecule.

A variation on this "primed insertion" theme has been suggested by Rogers in a general model for retroposon formation.[163] In this model, a nick in chromosomal DNA becomes tailed with T-rich sequences (in an analogous way to the amplification of telomeric sequences), which then act as a primer for cDNA synthesis. To ensure complete copy of the mRNA, the 5′ end of the inserted RNA is ligated to a second nick in the target DNA and repair synthesis completes the process to generate a retroposon flanked by direct repeats (Figure 7D).

Thus, a variety of models is available to account for the formation of processed pseudogenes, snRNA pseudogenes, and other retroposon like sequences. While a "primed insertion" type of model perhaps provides the simplest mechanism for a majority of these pseudogenes and retroposons, it is clear that at least for the truncated U3 snRNA pseudogenes, a "cDNA insertion" mechanism is favored. It appears that no one mechanism is likely to be universal, and the variety of pseudogene and retroposon structures and flanking "tail"

and repeat sequences probably reflects a variety of ways in which sequences contained in RNA may be reintroduced into the genome.

IV. CONCLUSION

In conclusion therefore, it can be said that while the genomes of many organisms do indeed contain all the genes required for their proper life and development, they also contain a surprising number of pseudogenes — dead and decaying erstwhile genes or "stillborn" retroposon-like processed pseudogenes. Clearly, the genome is able to tolerate this amount of "dead wood" since pseudogenes appear to have no function — except perhaps to intrigue and divert molecular biologists and thereby to throw new light on the workings of the genome.

ACKNOWLEDGMENTS

I am grateful to Nick Proudfoot, Sandy Macleod, and Tim Cripe for providing access to their unpublished data; to John Rogers and Margaret Buckingham for copies of their manuscripts prior to publication; to John Rogers for many helpful discussions; and to Rosemarie Baines for typing the manuscript.

REFERENCES

1. **Jacq, C., Miller, J. R., and Brownlee, G. G.,** A pseudogene structure in 5S DNA of *Xenopus laevis*, *Cell*, 12, 109, 1977.
2. **Miller, J. R., Cartwright, E. M., Brownlee, G. G., Fedoroff, N. V., and Brown, D.,** The nucleotide sequence of oöcyte 5S DNA in *Xenopus laevis*. II. The GC-rich region, *Cell*, 13, 717, 1978.
3. **Miller, J. R. and Melton, D. A.,** A transcriptionally active pseudogene in *Xenopus laevis* oöcyte 5S DNA, *Cell*, 24, 829, 1981.
4. **Proudfoot, N. J.,** Pseudogenes, *Nature (London)*, 286, 840, 1980.
5. **Little, P. F. R.,** Globin pseudogenes, *Cell*, 28, 683, 1982.
6. **Lauer, J., Shen, C-K. J., and Maniatis, T.,** The chromosomal arrangement of human α-like globin genes: sequence homology and α-globin gene deletion, *Cell*, 20, 119, 1980.
7. **Proudfoot, N. J. and Maniatis, T.,** The structure of a human α-globin pseudogene and its relationship to α-globin gene duplication, *Cell*, 21, 537, 1980.
8. **Proudfoot, N. J., Gil, A., and Maniatis, T.,** The structure of the human zeta globin gene and a closely linked nearly identical pseudogene, *Cell*, 31, 553, 1982.
9. **Lacy, E. and Maniatis, T.,** The nucleotide sequence of a rabbit β-globin pseudogene, *Cell*, 21, 545, 1980.
10. **Cleary, M. L., Haynes, J. R., Schon, E. A., and Lingrel, J. B.,** Identification by nucleotide sequence analysis of a goat pseudoglobin gene, *Nucl. Acids Res.*, 8, 4791, 1980.
11. **Cleary, M. L., Schon, E. A., and Lingrel, J. B.,** Two related pseudogenes are the result of a gene duplication in the goat β-globin locus, *Cell*, 26, 181, 1981.
12. **Jahn, C. L., Hutchison, C. A., III, Phillips, S. J., Weaver, S., Haigwood, N. L., Voliva, C. F., and Edgell, M. H.,** DNA sequence organization of the β-globin complex in the BALB/c mouse, *Cell*, 21, 159, 1980.
13. **Fritsch, E. F., Lawn, R. M., and Maniatis, T.,** Molecular cloning and characterization of the human β-like globin gene cluster, *Cell*, 19, 959, 1980.
14. **Jeffreys, A. J., Barrie, P. A., Harris, S., Fawcett, D. M., Nugent, Z. J., and Boyd, A. C.,** Isolation and sequence analysis of a hybrid δ-globin pseudogene from the brown lemur, *J. Mol. Biol.*, 156, 487, 1982.

15. **Vanin, E. F., Goldberg, G. I., Tucker, P. W., and Smithies, O.,** A mouse alpha globin-related pseudogene (ψα30.5) lacking intervening sequences, *Nature (London)*, 286, 222, 1980.
16. **Nishioka, Y., Leder, A., and Leder, P.,** An unusual alpha globin-like gene that has cleanly lost both globin intervening sequences, *Proc. Natl. Acad. Sci. U.S.A.*, 77, 2806, 1980.
17. **Leder, A., Swan, D., Ruddle, F., D'Eustachio, P., and Leder, P.,** Dispersion of α-like globin genes of the mouse to three different chromosomes, *Nature (London)*, 293, 196, 1981.
18. **Popp, R. A., Lalley, P. A., Whitney, J. B., and Anderson, W. F.,** Mouse α-globin genes and α-globin-like pseudogenes are not syntenic, *Proc. Natl. Acad. Sci. U.S.A.*, 78, 6362, 1981.
19. **Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodnev, R., and Dodgson, J.,** The evolution of genes: the chicken pre-pro-insulin gene, *Cell*, 20, 555, 1980.
20. **Miyata, T. and Yasunaga, T.,** Rapidly evolving mouse α-globin-related pseudogene and its evolutionary history, *Proc. Natl. Acad. Sci. U.S.A.*, 78, 450, 1981.
21. **Miyata, T. and Hayashida, H.,** Extraordinary high evolutionary rate of pseudogenes evidence for the presence of selective pressure against changes between synonymous codons, *Proc. Natl. Acad. Sci. U.S.A.*, 78, 5739, 1981.
22. **Li, W-H., Gojobori, T., and Nei, M.,** Pseudogenes as a paradigm of neutral evolution, *Nature (London)*, 292, 237, 1981.
23. **Slightom, J. L., Blechl, A. E., and Smithies, O.,** Human fetal $^{G}\gamma$ and $^{A}\gamma$-globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes, *Cell*, 21, 627, 1980.
24. **Shen, S., Slightom, J. L., and Smithies, O.,** A history of the human fetal globin gene duplication, *Cell*, 26, 191, 1980.
25. **Leibhaber, S. A., Gossens, M., and Kan, Y. W.,** Homology and concerted evolution at the α1 and α2 loci of human α-globin, *Nature (London)*, 290, 26, 1981.
26. **Schon, E. A., Wernke, S. M., and Lingrel, J. B.,** Gene conversion of two functional goat α-globin genes preserves only minimal flanking sequences, *J. Biol. Chem.*, 257, 6825, 1982.
27. **Weaver, S., Corner, M. B., Jahn, C. L., Hutchison, C. A., III, and Edgell, M. H.,** The adult β-globin genes of the "single" type mouse C57BL, *Cell*, 24, 403, 1981.
28. **Spritz, R. A., DeRiel, J. K., Forget, B. G., and Weissman, S. M.,** Complete nucleotide sequences of the human δ-globin gene, *Cell*, 21, 639, 1980.
29. **Efstratiadis, A., Posakony, J. W., Maniatis, T., Lawn, R. M., O'Connell, C., Spritz, R. A., DeRiel, J. K., Forget, B. G., Weissman, S. M., Slightom, J. L., Blechl, A. E., Smithies, O., Barralle, F. E., Shoulders, C. C., and Proudfoot, N. J.,** The structure and evolution of the human β-globin gene family, *Cell*, 21, 653, 1980.
30. **Martin, S. L., Vincent, K. A., and Wilson, A. L.,** Rise and fall of the delta globin gene, *J. Mol. Biol.*, 164, 513, 1983.
31. **Whitelaw, E. and Proudfoot, N.,** Transcriptional activity of the human pseudogene ψα globin compared with α globin, its functional gene counterpart, *Nucl. Acids Res.*, 11, 7717, 1983.
32. **Proudfoot, N. J., Rutherford, T. R., and Partington, G. A.,** Transcriptional analysis of human zeta globin genes, *EMBO J.*, 3, 1533, 1984.
32a. **Proudfoot, N. J.,** unpublished results.
33. **Orkin, S. H., Old, J., Lazarus, H., Altay, C., Gurgey, A., Weatherall, D. J., and Nathans, D. G.,** The molecular basis of α-thalassemias: frequent occurrence of dysfunctional α-loci among non Asians with HbH disease, *Cell*, 17, 33, 1979.
34. **Higgs, D. R., Old, J. M., Pressley, L., Clegg, J. B., and Weatherall, D.,** A novel α-globin gene arrangement in man, *Nature (London)*, 284, 632, 1980.
35. **Goosens, M., Dozy, A., Embury, S., Zacharides, Z., Hadjimas, M., Stamatoyannopoulos, G., and Kan, Y. W.,** Triplicated α-globin loci in humans, *Proc. Natl. Acad. Sci. U.S.A.*, 77, 518, 1980.
36. **Hardies, S. C., Edgell, M. H., and Hutchison, C. A., III,** Evolution of a mammalian β globin gene cluster, *J. Biol. Chem.*, 259, 3748, 1984.
37. **Hardison, R. C.,** Comparison of the β-like globin gene families of rabbits and humans indicates that the gene cluster 5′-ε-γ-δ-β-3′ predates the mammalian radiation, *Mol. Biol. Evol.*, 1, 390, 1984.
37a. **Goodman, M., Koop, B. F., Czelusniak, J., Weiss, M. L., and Slightom, J. L.,** The η-globin gene. Its long evolutionary history in the β-globin gene family of mammals, *J. Mol. Biol.*, 180, 803, 1984.
38. **Lacy, E., Hardison, R. C., Quon, D., and Maniatis, T.,** The linkage arrangement of four rabbit β-like globin genes, *Cell*, 18, 1273, 1979.
39. **Hardison, R. C., Butler, E. T., Lacy, E., Maniatis, T., Rosenthal, N., and Efstratiadis, A.,** The structure and transcription of four linked rabbit β-like globin genes, *Cell*, 18, 1285, 1979.
40. **Hardison, R. C. and Margot, J. B.,** Rabbit globin pseudogene ψβ2 is a hybrid of δ and β globin gene sequences, *Mol. Biol. Evol.*, 1, 302, 1984.
41. **Townes, T. M., Shapiro, S. G., Wernke, S. M., and Lingrel, J. B.,** Duplication of a four-gene set during the evolution of the goat β-globin locus produced genes now expressed differentially in development, *J. Biol. Chem.*, 259, 1896, 1984.

42. **Hill, A., Hardies, S. C., Phillips, S. J., Davies, M. G., Hutchison, C. A., III, and Edgell, M. H.,** Two mouse early embryonic β-globin gene sequences, *J. Biol. Chem.*, 259, 3739, 1984.
43. **Phillips, S. J., Hardies, S. C., Jahn, C. L., Edgell, M. H., and Hutchison, C. A., III,** The complete nucleotide sequence of a β-globin-like structure, βh2, from the [Hbb]^d mouse BALB/c, *J. Biol. Chem.*, 259, 7947, 1984.
44. **Baglioni, C.,** The fusion of two peptide chains in haemoglobin Lepore and its interpretation as a genetic deletion, *Proc. Natl. Acad. Sci. U.S.A.*, 48, 1880, 1962.
45. **Jagadeeswaran, P., Pan, J., Forget, B. G., and Weissman, S. M.,** Sequences of non-α-globin genes in man, *Cold Spring Harbor Symp. Quant. Biol.*, 47, 1079, 1982.
45a. **Chang, L-Y.E. and Slightom, J. L.,** Isolation and nucleotide sequence analysis of the β-type globin pseudogene from human, gorilla, and chimpanzee, *J. Mol. Biol.*, 180, 767, 1984.
45b. **Harris, S., Barrie, P. A., Weiss, M. L., and Jeffreys, A. J.,** The primate ψβ1 gene: an ancient β-globin pseudogene, *J. Mol. Biol.*, 180, 785, 1984.
46. **Leder, P., Hansen, J. N., Konkel, D., Leder, A., Nishioka, Y., and Talkington, C.,** Mouse globin system: a functional and evolutionary analysis, *Science*, 209, 1336, 1980.
47. **Engel, J. D. and Dodgson, J. B.,** Analysis of the closely linked adult chicken α-globin genes in recombinant DNAs, *Proc. Natl. Acad. Sci. U.S.A.*, 77, 2596, 1980.
48. **Dolan, M., Sugarman, B. J., Dodgson, J. B., and Engel, J. D.,** Chromosomal arrangement of the chicken β type globin genes, *Cell*, 24, 669, 1981.
48a. **Li, W-H.,** Evolution of duplicate genes and pseudogenes, in *Evolution of Genes and Proteins*, Nei, M. and Koehn, R. K., Eds., Sinauer, Sunderland, Mass., 1982, 14.
49. **Royal, A., Garapin, A., Cami, B., Perrin, F., Mandel, J. L., LeMeur, M., Brégégègre, F., Gannon, F., LePennec, J. P., Chambon, P., and Kouritsky, P.,** The ovalbumin gene region: common features in the organization of three genes expressed in chicken oviduct under hormonal control, *Nature (London)*, 279, 125, 1979.
50. **Heilig, R., Perrin, F., Gannon, F., Mandel, J. L., and Chambon, P.,** The ovalbumin gene family: structure of the X gene and evolution of duplicated split genes, *Cell*, 20, 625, 1980.
51. **Colbert, D. A., Knoll, B. J., Woo, S. L., Mace, M. L., Tsai, M. J., and O'Malley, B. W.,** Differential hormonal responsiveness of the ovalbumin gene and its pseudogenes in the chick oviduct, *Biochemistry*, 19, 5586, 1980.
52. **Knoll, B. J., Woo, S. L. C., Beattie, W., and O'Malley, B. W.,** Identification and sequence analysis of the 5′ domain of the X and Y pseudoovalbumin genes, *J. Biol. Chem.*, 256, 7949, 1981.
53. **Brack, C., Nogata, S., Mantel, N., and Weissmann, C.,** Molecular analysis of the human interferon α gene family, *Gene*, 15, 379, 1981.
54. **Ullrich, A., Gray, A., Goeddel, D. V., and Dull, T.,** Nucleotide sequences of a portion of a human chromosome nine containing a leukocyte interferon gene cluster, *J. Mol. Biol.*, 156, 467, 1982.
55. **Goeddel, D. V., Leung, D. W., Dull, T. J., Gross, M., Lawn, R. M., McCandliss, R., Seeburg, P. M., Ullrich, A., Yelverton, E., and Gray, P. W.,** The structure of eight district cloned human leukocyte interferon cDNAs, *Nature (London)*, 290, 20, 1981.
56. **Cohen, J. B. and Givol, D.,** Conservation and divergence of immunoglobulin V_H pseudogenes, *EMBO J.*, 2, 1795, 1983.
57. **Huang, H., Crews, S., and Hood, L.,** An immunoglobulin V_H pseudogene, *J. Mol. Appl. Genet.*, 1, 93, 1981.
58. **Givol, D., Zakut, R., Effron, K., Rechavi, G., Ram, D., and Cohen, J. B.,** Diversity of germ line immunoglublin V_H genes, *Nature (London)*, 292, 426, 1981.
59. **Loh, D. Y., Bothwell, A. L. M., White-Scharf, M. E., Imanishi-Kari, T., and Baltimore, D.,** Molecular basis of a mouse strain-specific anti-hapten response, *Cell*, 33, 85, 1983.
60. **Rechavi, G., Bienz, B., Ram, D., Ben-Neriah, Y., Cohen, J. B., Zakut, R., and Givol, D.,** Organization and evolution of immunoglobulin V_H gene subgroups, *Proc. Natl. Acad. Sci. U.S.A.*, 79, 4405, 1982.
61. **Rechavi, G., Ram, D., Glazer, L., Zakut, R., and Givol, D.,** Evolutionary aspects of immunoglobulin heavy chain variable region (V_H) gene subgroups, *Proc. Natl. Acad. Sci. U.S.A.*, 80, 855, 1983.
62. **Bentley, D. L. and Rabbitts, T. H.,** Human immunoglobulin variable region genes — DNA sequences of two V_κ genes and a pseudogene, *Nature (London)*, 288, 730, 1980.
63. **Bothwell, A. L. M., Paskind, M., Reth, M., Imanishi-Kari, T., Rajewsky, K., and Baltimore, D.,** Heavy chain variable region contribution to the NP^b family of antibodies: somatic mutation evident in a γ_{2a} variable region, *Cell*, 24, 625, 1981.
64. **Max, E. E., Seidman, J. G., and Leder, P.,** Sequences of five potential recombination sites encoded close to an immunoglobulin κ constant region gene, *Proc. Natl. Acad. Sci. U.S.A.*, 76, 3450, 1979.
65. **Sakano, H., Huppi, K., Heinrich, G., and Tonegawa, S.,** Sequences at the somatic recombination sites of immunoglobulin light chain genes, *Nature (London)*, 280, 288, 1979.
66. **Miller, J., Selsing, E., and Storb, U.,** Structural alterations in J regions of mouse immunoglobulin λ genes are associated with differential gene expression, *Nature (London)*, 295, 428, 1982.

67. **Blomberg, B. and Tonegawa, S.,** DNA sequences of the joining regions of mouse λ light chain immunoglobulin genes, *Proc. Natl. Acad. Sci. U.S.A.*, 79, 530, 1982.
68. **Gough, N. M. and Bernard, O.,** Sequences of the joining region genes for immunoglobulin heavy chains and their role in generation of antibody diversity, *Proc. Natl. Acad. Sci. U.S.A.*, 78, 509, 1981.
69. **Ravetch, J. V., Siebenlist, U., Kasmeyer, S., Walldmann, T., and Leder, P.,** Structure of the human immunoglobulin μ locus: characterization of embryonic and rearranged J and D genes, *Cell*, 27, 583, 1981.
70. **Coleclough, C., Perry, R. P., Karjalainen, K., and Weigert, M.,** Aberrant rearrangements contribute significantly to the allelic exclusion of immunoglobulin gene expression, *Nature (London)*, 290, 372, 1981.
71. **Krawinkel, U. and Rabbitts, T. H.,** Comparison of the hinge-coding segments in human immunoglobulin gamma heavy chain genes and the linkage of the gamma 2 and gamma 4 subclass genes, *EMBO J.*, 1, 403, 1982.
72. **Takahashi, N., Ueda, S., Obata, M., Nikaido, T., Nakai, S., and Honjo, T.,** Structure of human immunoglobulin gamma genes; implications for evolution of a gene family, *Cell*, 29, 671, 1982.
73. **Bech-Hansen, N. T., Linsley, P. S., and Cox, D. W.,** Restriction fragment length polymorphisms associated with immunoglobulin C_γ genes reveal linkage disequilibrium and genomic organization, *Proc. Natl. Acad. Sci. U.S.A.*, 80, 6952, 1983.
74. **Flanagan, J. G. and Rabbitts, T. H.,** Arrangement of human immunoglobulin heavy chain constant region genes implies evolutionary duplication of a segment containing γ, ε, and α genes, *Nature (London)*, 300, 709, 1982.
75. **Max, E., Battey, J., Ney, R., Kirsch, I. R., and Leder, P.,** Duplication and deletion in the human immunoglobulin ε genes, *Cell*, 29, 691, 1982.
76. **Hisajima, H., Nishida, Y., Nakai, S., Takahashi, N., Uedo, S., and Honjo, T.,** Structure of the human immunoglobulin $C_\gamma 2$ gene, a truncated pseudogene. Implications for its evolutionary origin, *Proc. Natl. Acad. Sci. U.S.A.*, 80, 2995, 1983.
77. **Steinmetz, M., Moore, K. W., Frelinger, J. G., Sher, B. T., Shen, F. W., Boyse, F. A., and Hood, L.,** A pseudogene homologous to mouse transplantation antigens: transplantation antigens are encoded by eight exons that correlate with protein domains, *Cell*, 25, 683, 1981.
78. **Mellor, A. L., Weiss, E. H., Kress, M., Jay, G., and Flavell, R. A.,** A nonpolymorphic class 1 gene in the mouse major histocompatibility complex, *Cell*, 36, 139, 1984.
79. **Biro, P. A., Pan, J., Sood, A. K., Kole, R., Reddy, V. B., and Weissman, S. M.,** Sequences of human major histocompatibility locus genes, *Cold Spring Harbor Symp. Quant. Biol.*, 47, 1079, 1982.
80. **Cosman, D., Khoury, G., and Jay, G.,** Three classes of mouse H-2 messenger RNA distinguished by analysis of cDNA clones, *Nature (London)*, 295, 73, 1982.
81. **Cosman, D., Kress, M., Khoury, G., and Jay, G.,** Tissue specific expression of an unusual H-2 (class 1)-related gene, *Proc. Natl. Acad. Sci. U.S.A.*, 79, 4947, 1982.
82. **Kress, M., Cosman, D., Khoury, G., and Jay, G.,** Secretion of a transplantation-related antigen, *Cell*, 34, 189, 1983.
83. **Schulze, D. H., Pease, L. R., Geier, S. S., Reyes, A. A., Sarmiento, L. A., Wallace, R. B., and Nathenson, S. G.,** Comparison of the cloned H-$2K^{bm1}$ variant gene with the H-$2K^b$ gene shows a cluster of seven nucleotide differences, *Proc. Natl. Acad. Sci. U.S.A.*, 80, 2007, 1983.
84. **Mellor, A. L., Weiss, E. H., Ramachandran, K., and Flavell, R. A.,** A potential donor gene for the bm1 gene conversion in the C57BL mouse, *Nature (London)*, 306, 792, 1983.
85. **Weiss, E. H., Mellor, A. L., Golden, L., Fahrner, K., Simpson, E., Hurst, J., and Flavell, R. A.,** The structure of a mutant H-2 gene suggests that the generation of polymorphism in H-2 genes may occur by gene conversion-like events, *Nature (London)*, 301, 671, 1983.
86. **Snyder, M., Hunkapiller, M., Yuen, D., Silvert, D., Fristrom, J., and Davidson, N.,** Cuticle protein genes of Drosophila: structure organization and evolution of four clustered genes, *Cell*, 29, 1027, 1982.
87. **Shibuya, K., Noguchi, S., Nishimura, S., and Sekiya, T.,** Characterization of a rat tRNA gene cluster containing the genes for $tRNA^{Asp}$, $tRNA^{Gly}$, and $tRNA^{Glu}$, and pseudogenes, *Nucl. Acids Res.*, 10, 4441, 1982.
88. **Makowski, D. R., Haas, R. A., Dolan, K. P., and Grunberger, D.,** Molecular cloning, sequence analysis and *in vitro* expression of a rat tRNA gene cluster, *Nucl. Acids Res.*, 11, 8609, 1983.
89. **Sekiya, T., Kuchino, Y., and Nishimura, S.,** Mammalian tRNA genes: nucleotide sequence of rat genes for $tRNA^{Asp}$, $tRNA^{Gly}$, and $tRNA^{Glu}$, *Nucl. Acids Res.*, 9, 2239, 1981.
90. **Turner, P. C., Aldridge, T. C., Woodland, H. R., and Old, R. W.,** Nucleotide sequence of H1 histone genes from *Xenopus laevis*. A recently diverged pair of H1 genes and an unusual H1 pseudogene, *Nucl. Acids Res.*, 11, 4093, 1983.
91. **Wilde, C. D., Crowther, C. E., and Cowan, N. J.,** Diverse mechanisms in the generation of human β-tubulin pseudogenes, *Science*, 217, 549, 1982.
92. **Selker, E. U., Free, S. J., Metzenberg, R. L., and Yanofsky, C.,** An isolated pseudogene related to the 5S RNA genes in *Neurospora crassa*, *Nature (London)*, 294, 576, 1981.

93. **Firtel, R. A., Timm, R., Kimmel, A. R., and McKeown, M.,** Unusual nucleotide sequence at the 5′ end of actin genes in *Dictyostelium discoideum*, *Proc. Natl. Acad. Sci. U.S.A.*, 76, 6206, 1979.
94. **McKeown, M. and Firtel, R. A.,** Differential expression and 5′ end mapping of actin genes in *Dictyostelium*, *Cell*, 24, 799, 1981.
95. **Wiborg, O., Hyldig-Nielsen, J., Ojensen, E., Paludan, K., and Marcker, K. A.,** The structure of an unusual leghaemoglobin gene from soybean, *EMBO J.*, 2, 449, 1983.
96. **Orozco, E. M., Rushlow, K. E., Dodd, J. R., and Hallick, R. B.,** *Euglena gracilis* chloroplast ribosomal RNA transcription units. II. Nucleotide sequence homology between the 16S-23S ribosomal RNA spacer and the 16S ribosomal RNA leader regions, *J. Biol. Chem.*, 255, 10997, 1980.
97. **Miyata, T., Kikuno, R., and Ohshima, Y.,** A pseudogene cluster in the leader region of the *Euglena* chloroplast 16S-23S rRNA genes, *Nucl. Acids Res.*, 10, 1771, 1982.
98. **Jelinek, W. R., Toomey, T. P., Leinwand, L., Duncan, G. H., Biro, P. A., Choudary, P. V., Weissman, S. M., Rubin, C. M., Houck, C. M., Deininger, P. L., and Schmid, C. W.,** Ubiquitous, interspersed repeated sequences in mammalian genomes, *Proc. Natl. Acad. Sci. U.S.A.*, 77, 1398, 1980.
99. **Schmid, C. W. and Jelinek, W. R.,** The Alu family of dispersed repetitive sequences, *Science*, 216, 1065, 1982.
100. **Jagadeeswaran, P., Forget, B. G., and Weissman, S. M.,** Short interspersed repetitive DNA elements in eucaryotes. Transposable DNA elements generated by reverse transcription of RNA pol III transcripts, *Cell*, 26, 141, 1981.
101. **Hollis, G. F., Hieter, P. A., McBride, O. W., Swan, D., and Leder, P.,** Processed genes: a dispersed human immunoglobulin gene bearing evidence of RNA type processing, *Nature (London)*, 296, 321, 1982.
102. **Rogers, J.,** Retroposons defined, *Nature (London)*, 301, 460, 1983.
103. **Childs, G., Maxson, R., Cohn, R. H., and Kedes, L.,** Orphons: dispersed genetic elements derived from tandem repetitive genes of eucaryotes, *Cell*, 23, 651, 1981.
104. **Maxson, R., Cohn, R., Kedes, L., and Mohun, T.,** Expression and organisation of histone genes, *Annu. Rev. Genet.*, 17, 239, 1983.
105. **Liebermann, D., Hoffman-Lieberman, B., Weinthal, J., Childs, G., Maxson, R., Mauron, A., Cohen, S. N., and Kedes, L.,** An unusual transposon with long terminal inverted repeats in the sea urchin *Strongylocentrotus purpuratus*, *Nature (London)*, 306, 342, 1983.
106. **Goff, S. P., Gilboa, E., Wilte, O. N., and Baltimore, D.,** Structure of the Abelson Murine Leukaemia virus genome and the homologous cellular gene: studies with cloned viral DNA, *Cell*, 22, 777, 1980.
107. **Shimotohno, K. and Temin, H. M.,** Loss of intervening sequences in genomic mouse α-globin DNA inserted in an infectious retrovirus vector, *Nature (London)*, 299, 265, 1982.
108. **Lueders, K., Leder, A., Leder, P., and Kuff, E.,** Association between a transposed α-globin pseudogene and retrovirus-like elements in the BALB/c mouse genome, *Nature (London)*, 295, 426, 1982.
109. **Vanin, E. F.,** Processed pseudogenes: characteristics and evolution, *Biochem. Biophys. Acta*, 782, 231, 1984.
110. **Ueda, S., Nakai, S., Nishida, Y., Hisajima, H., and Honjo, T.,** Long terminal repeat-like elements flank a human immunoglobulin epsilon pseudogene that lacks introns, *EMBO J.*, 1, 1539, 1982.
111. **Busch, H., Reddy, R., Rothblum, L., and Choi, Y. C.,** snRNA's, snRNP's, and RNA processing, *Annu. Rev. Biochem.*, 51, 617, 1982.
112. **Hayashi, K.,** Organization of sequences related to U6 RNA in the human genome, *Nucl. Acids Res.*, 9, 3379, 1981.
113. **Westin, G., Monstein, H-J., Zabielski, J., Philipson, L., and Pettersson, U.,** Human DNA sequences complementary to the small nuclear RNA U2, *Nucl. Acids Res.*, 9, 6323, 1981.
114. **Denison, R. A., Van Arsdell, S. W., Bernstein, L. B., and Weiner, A. W.,** Abundant pseudogenes for small nuclear RNAs are dispersed in the human genome, *Proc. Natl. Acad. Sci. U.S.A.*, 78, 810, 1981.
115. **Manser, T. and Gesteland, R. F.,** Characterization of small nuclear RNA U1 gene candidates and pseudogenes from the human genome, *J. Mol. Genet.*, 1, 117, 1981.
116. **Piechaczyk, M., Lelay-Taha, M. N., Sri-Widada, J., Brunel, C., Phiautard, J., and Jeanteur, P.,** Mouse DNA sequences complementary to small nuclear RNA U1, *Nucl. Acids Res.*, 10, 4627, 1982.
117. **Lund, E. and Dahlberg, J. E.,** True genes for human U1 small nuclear RNA, *J. Biol. Chem.*, 259, 2013, 1984.
118. **Ohshima, Y., Okada, N., Tani, T., Itoh, Y., and Itoh, M.,** Nucleotide sequences of mouse genomic loci including a gene or pseudogene for U6(4·8S) nuclear RNA, *Nucl. Acids Res.*, 9, 5145, 1981.
119. **Nojima, H. and Kornberg, R. D.,** Genes and pseudogenes for mouse U1 and U2 small nuclear RNAs, *J. Biol. Chem.*, 258, 8151, 1983.
120. **Watanabe-Nagasu, N., Itoh, Y., Tani, T., Okano, K., Koga, N., Okada, N., and Ohshima, Y.,** Structural analysis of gene loci for rat U1 small nuclear RNA, *Nucl. Acids Res.*, 11, 1791, 1983.
121. **Monstein, H-J., Westin, G., Philipson, L., and Pettersson, U.,** A candidate gene for human U-1 RNA, *EMBO J.*, 1, 133, 1982.

122. **Monstein, H-J., Hammarström, K., Westin, G., Zabielski, J., Philipson, L., and Pettersson, U.,** Loci for human U-1 RNA structural genes and evolutionary implications, *J. Mol. Biol.*, 167, 245, 1983.
123. **Denison, R. A. and Weiner, A. M.,** Human U-1RNA pseudogenes may be generated by both DNA mediated and RNA mediated mechanisms, *Mol. Cell. Biol.*, 2, 815, 1982.
124. **Van Arsdell, S. W., Denison, R. A., Bernstein, L. B., Weiner, A. M., Manser, T., and Gesteland, R. F.,** Direct repeats flank three small nuclear RNA pseudogenes in the human genome, *Cell*, 26, 11, 1981.
125. **Hammarström, K., Westin, G., and Pettersson, U.,** A pseudogene for human U4 RNA with a remarkable structure, *EMBO J.*, 1, 737, 1982.
126. **Bernstein, L. B., Mount, S. M., and Weiner, A. M.,** 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, 1983.
127. **Van Arsdell, S. W. and Weiner, A. M.,** Pseudogenes for human U2 small nuclear RNA do not have a fixed site of 3′ truncation, *Nucl. Acids Res.*, 12, 1463, 1984.
128. **Battey, J., Max, E., McBride, O., Swan, D., and Leder, P.,** A processed human immunoglobulin ϵ gene has moved to chromosome 9, *Proc. Natl. Acad. Sci. U.S.A.*, 79, 5956, 1982.
129. **Uhler, M., Herbert, E., D'Eustachio, P., and Ruddle, F. D.,** The mouse genome contains two non-allelic pro-opiomelanocortin genes, *J. Biol. Chem.*, 258, 9444, 1983.
130. **O'Brien, S. J., Nash, W. G., Goodwin, J. L., Lowy, D. R., and Chang, E. H.,** Dispersion of the *ras* family of transforming genes to four different chromosomes in man, *Nature (London)*, 302, 839, 1983.
131. **Anagnou, N. P., O'Brien, S. J., Shimada, T., Nash, W. G., Chen, M. Y., and Nienhuis, A. W.,** Chromosomal organization of the human dihydrofolate reductase genes: dispersion, selective amplification, and a novel form of polymorphism, *Proc. Natl. Acad. Sci. U.S.A.*, 81, 5170, 1984.
132. **Robert, B., Daubas, P., Akimenko, M. A., Cohen, A., Garner, I., Guenet, J. L., and Buckingham, M.,** A single locus in the mouse encodes both myosin light chains 1 and 3, a second locus corresponds to a related pseudogene, *Cell*, 38, 129, 1984.
133. **Wilde, C. D., Crowther, C. E., Cripe, T. P., Lee, M. G-S., and Cowan, N. J.,** Evidence that a human β-tubulin pseudogene is derived from its corresponding mRNA, *Nature (London)*, 297, 83, 1982.
134. **Karin, M. and Richards, R. I.,** Human methallothionein genes — primary structure of the metallolthionein-II gene and a related processed gene, *Nature (London)*, 299, 797, 1982.
135. **Chen, M-J., Shimada, T., Moulton, A. D., Harrison, M., and Nienhuis, A. W.,** Intronless dihydrofolate reductase genes are derived from processed RNA molecules, *Proc. Natl. Acad. Sci. U.S.A.*, 79, 7435, 1982.
136. **Masters, J. N., Yang, J. K., Cellini, A., and Attardi, G.,** A human dihydrofolate reductase pseudogene and its relationship to the multiple forms of specific messenger RNA, *J. Mol. Biol.*, 167, 23, 1983.
137. **Freytag, S. O., Bock, H-G. O., Beaudet, A. L., and O'Brien, W. E.,** Molecular structures of human arginosuccinate synthetase pseudogenes. Evolutionary and mechanistic implications, *J. Biol. Chem.*, 259, 3160, 1984.
138. **Moos, M. and Gallwitz, D.,** Structure of a human β actin-related pseudogene which lacks intervening sequences, *Nucl. Acids Res.*, 10, 7843, 1982.
139. **Moos, M. and Gallwitz, D.,** Structure of two human β actin-related processed genes one of which is located next to a simple repetitive sequence, *EMBO J.*, 2, 757, 1983.
140. **MacLeod, A. R. and Talbot, K.,** A processed gene defining a gene family encoding a human non-muscle tropomyosin, *J. Mol. Biol.*, 167, 523, 1983.
141. **MacLeod, A. R. and Talbot, K.,** unpublished results.
142. **McGrath, J. P., Capon, D. J., Smith, D. H., Chen, E. Y., Seeburg, P. H., Goeddel, D. V., and Levinson, A. D.,** Structure and organisation of the human Ki *ras* proto oncogene and a related processed pseudogene, *Nature (London)*, 304, 501, 1983.
143. **Miyoshi, J., Kagimoto, M., Soeda, E., and Sakaki, Y.,** The human c-Ha-ras 2 is a processed pseudogene inactivated by numerous base substitutions, *Nucl. Acids. Res.*, 12, 1821, 1984.
144. **Lemischka, I. and Sharp, P. A.,** The sequences of an expressed rat α-tubulin gene and a pseudogene with an inserted repetitive element, *Nature (London)*, 300, 330, 1982.
145. **Scarpulla, R. C. and Wu, R.,** Non-allelic members of the cytochrome c multigene family of the rat may arise through different messenger RNAs, *Cell*, 32, 473, 1983.
146. **Zakut-Houri, R., Oren, M., Bienz, B., Lavie, V., Hazum, S., and Givol, D.,** A single gene and a pseudogene for the cellular tumour antigen p53, *Nature (London)*, 306, 594, 1983.
147. **Klein, A. and Meynhas, O.,** A multigene family of intron-lacking and containing genes, encoding for mouse ribosomal protein L7, *Nucl. Acids Res.*, 12, 3763, 1984.
148. **Perled-Yalif, E., Cohen-Binder, I., and Meynhas, O.,** Isolation and characterization of four mouse ribosomal-protein-L18 genes that appear to be processed pseudogenes, *Gene*, 29, 157, 1984.
149. **Dudov, K. P. and Perry, R. P.,** The gene family encoding the mouse ribosomal protein L32 contains a uniquely expressed intron-containing gene and an unmutated processed gene, *Cell*, 37, 457, 1984.

150. **Lee, M. G-S., Lewis, S. A., Wilde, C. D., and Cowan, N. J.,** Evolutionary history of a multigene family: an expressed human β-tubulin gene and three processed genes, *Cell*, 33, 477, 1983.
151. **Notake, M., Tobimatsu, T., Watanabe, Y., Takahashi, H., Mishina, M., and Numa, S.,** Isolation and characterization of the mouse corticotropin-β-lipotropin precursor gene and a related pseudogene, *FEBS Lett.*, 156, 67, 1983.
152. **Ponte, P., Gunning, P., Blau, H., and Kedes, L.,** Human actin genes are single copy for α-skeletal and α-cardiac actin but multicopy for β and γ-cytoskeletal genes: 3′ untranslated regions are isotype specific but are conserved in evolution, *Mol. Cell Biol.*, 3, 1783, 1983.
153. **Stein, J. P., Munjaal, R. P., Lagace, L., Chai, E., O'Malley, B. W., and Means, A. R.,** Tissue specific expression of a chicken calmodulin pseudogene lacking intervening sequences, *Proc. Natl. Acad. Sci. U.S.A.*, 80, 6485, 1983.
154. **DiNocera, P. P., Digan, M. E., and Dawid, I. B.,** A family of oligo-adenylate-terminated transposable sequences in *Drosophila melanogaster*, *J. Mol. Biol.*, 168, 715, 1983.
155. **Cripe, T. P.,** unpublished results.
156. **Lomedico, P., Rosenthal, N., Efstratiadis, A., Gilbert, W., Kolodner, R., and Tizard, R.,** The structure and evolution of the two non-allelic rat preproinsulin genes, *Cell*, 18, 545, 1979.
157. **Sharp, P. A.,** Conversion of RNA to DNA in mammals: Alu-like elements and pseudogenes, *Nature (London)*, 301, 471, 1983.
158. **Krayer, A. S., Markusheva, T. V., Kramerov, D. A., Ryskov, A. P., Scryabin, K. G., Bayer, A. A., and Georgiev, G. P.,** Ubiquitous transposon like repeats B1 and B2 of the mouse genome: B2 sequencing, *Nucl. Acids Res.*, 10, 7461, 1982.
159. **Sutcliffe, J. G., Milner, R. J., Bloom, F. E., and Lerner, R. A.,** A common 82 nucleotide sequence unique to brain RNA, *Proc. Natl. Acad. Sci. U.S.A.*, 79, 4942, 1982.
160. **DiGiovanni, L., Haynes, S. R., Misra, R., and Jelinek, W. R.,** KpnI family of long-dispersed repeated DNA sequences of man: evidence for entry into genomic DNA of DNA copies of poly(A) terminated KpnI RNAs, *Proc. Natl. Acad. Sci. U.S.A.*, 80, 6533, 1983.
161. **Rogers, J.,** A straight LINE story, *Nature (London)*, 306, 113, 1983.
162. **Dawid, I., Long, E. O., DiNocera, P. P., and Pardue, M. L.,** Ribosomal insertion like elements in *Drosophila melanogaster* are interspersed with mobile sequences, *Cell*, 25, 399, 1981.
163. **Rogers, J. H.,** The origin and evolution of retroposons, *Int. Rev. Cytol.*, 93, 187, 1985.
164. **Weissbach, A.,** Eukaryotic DNA polymerases, *Annu. Rev. Biochem.*, 46, 25, 1977.
165. **Halligan, B. D., Davis, J. L., Edwards, K. A., and Lin, L. F.,** Intra and intermolecular transfer by HeLa DNA topoisomerase I, *J. Biol. Chem.*, 257, 3995, 1982.