PSEUDOGENES: Are They “Junk” or Functional DNA?

Evgeniy S. Balakirev1,2 and Francisco J. Ayala1
1Department of Ecology and Evolutionary Biology, University of California, Irvine, California 92697-2525; email: fjayala@uci.edu; 2Institute of Marine Biology, Vladivostok 690041, Russia and Academy of Ecology, Marine Biology, and Biotechnology, Far Eastern State University, Vladivostok 690600, Russia; email: esbalak@bio.dvgu.ru

Key Words neutral evolution, evolutionary rate, DNA polymorphism, Drosophila, potogene

Abstract Pseudogenes have been defined as nonfunctional sequences of genomic DNA originally derived from functional genes. It is therefore assumed that all pseudogene mutations are selectively neutral and have equal probability to become fixed in the population. Rather, pseudogenes that have been suitably investigated often exhibit functional roles, such as gene expression, gene regulation, generation of genetic (antibody, antigenic, and other) diversity. Pseudogenes are involved in gene conversion or recombination with functional genes. Pseudogenes exhibit evolutionary conservation of gene sequence, reduced nucleotide variability, excess synonymous over nonsynonymous nucleotide polymorphism, and other features that are expected in genes or DNA sequences that have functional roles. We first review the Drosophila literature and then extend the discussion to the various functional features identified in the pseudogenes of other organisms. A pseudogene that has arisen by duplication or retroposition may, at first, not be subject to natural selection if the source gene remains functional. Mutant alleles that incorporate new functions may, nevertheless, be favored by natural selection and will have enhanced probability of becoming fixed in the population. We agree with the proposal that pseudogenes be considered as potogene, i.e., DNA sequences with a potentiality for becoming new genes.

CONTENTS

INTRODUCTION .................................................... 124
PSEUDOGENES IN DROSOPHILA .......................... 125
  Evolutionary Conservation .................................. 125
  Pseudogene Expression .................................... 127
  Regulation of Gene Expression ............................ 127
  The β-esterase Cluster of D. melanogaster: Expression, Evolutionary
  Conservation, Recombination, and Gene Conversion ........ 128
EVOLUTIONARY CONSERVATION OF PSEUDOGENES ... 129
PSEUDOGENE TRANSCRIPTION AND EXPRESSION ........ 131
INTRODUCTION

Pseudogenes have been defined as nonfunctional sequences of genomic DNA that are originally derived from functional genes, but exhibit such degenerative features as premature stop codons and frameshift mutations that prevent their expression (97, 107, 132, 145, 164, 226, 231, 237). Pseudogenes are thought to arise by tandem duplication of genes, with ensuing loss of function as a result of gradual accumulation of disabling mutations (107, 132, 164, 237). “Processed” pseudogenes lack introns, and presumably arise by reverse transcription of processed mRNA, followed by integration into the genome (150, 226, 227, 231). Partially processed pseudogenes, generated by an RNA-mediated mechanism, containing or not the complete coding region, have been described in tomato (210) and human (60, 62, 194, 209). Chimeric nonfunctional genes combining parts of a functional gene and its pseudogene have been found in lemurs (108) and humans (129).

If the presence of one copy of the gene suffices for the needs of the organism, it is assumed that pseudogene mutations (disabling or not) will not be subject to purifying selection and will all have equal probability of becoming fixed in the population (85, 87, 116, 130, 131). It follows that pseudogenes will generally degenerate, owing to the rapid accumulation of recurrent mutations, and melt into the background of the surrounding DNA (e.g., 88). That process has indeed been detected in bacterial genomes (7). However, genomes of eukaryotes contain many pseudogenes that appear to have avoided full degeneration (96, 145). A large-scope analysis of pseudogene distribution based on complete genome draft sequences shows that there is less pressure to delete pseudogenes in eukaryotes than in prokaryotes (96).

We review a variety of pseudogene features in diverse organisms. We begin by reviewing the *Drosophila* literature. We then extend the discussion to other organisms, starting with cases of evolutionary conservation and proceeding to explore the various functions detected in putative pseudogenes, such as gene expression, regulation of gene expression, generation of genetic diversity (antibody, antigenic, and other), and impact on gene conversion and recombination.

We review those pseudogenes for which there is functional or evolutionary information, excluding from consideration some pseudogenes that have been characterized only by their sequence, such as tRNA (53, 197), U4 RNA (186), glutathione S-transferase (218, 219), and histone H2B (1).
PSEUDOGENES IN DROSOPHILA

Pseudogenes are rare in *Drosophila* (98, 161) relative to some other animals, especially vertebrates (145). DNA sequence evolution in many of the pseudogenes found in *Drosophila* manifests functional constraints, reflected in lower than expected (if the pseudogenes were not subject to selection) intraspecific variability and interspecific divergence; significant heterogeneity of nucleotide variability and divergence along the sequence; higher rate of substitution at synonymous than at nonsynonymous nucleotide positions; conservation of important functional regions; transcriptional activity; and codon bias (12, 16, 56, 110, 111, 162, 167, 208). Moreover, two *Adh* Drosophila genes originally identified as pseudogenes (70, 110) have later been considered to be novel functional genes (19, 136).

Evolutionary Conservation

Snyder et al. (203) studied a small multigene family encoding the larval cuticle proteins (LCP) of *Drosophila melanogaster*, located within a 9-kb region of the right arm of the second chromosome, at cytogenetic map position 44D. One of the five sequences was thought to be a pseudogene (ψLcp) because of several structural features and the absence of detectable transcripts. The disabling features included a 35-bp deletion eliminating the TATA box, two premature stop codons at positions 23 and 72 (162), a mutated splicing acceptor sequence, and several substitutions and indels that would prevent secretion of a functional protein (203). Another cuticle gene cluster located at 65A on the left arm of the third chromosome of *D. melanogaster* includes an intronless pseudogene, Lcp-aψ, with a 21-bp frame-shifting deletion in the signal-peptide coding region and without a consensus polyadenylation site (42).

The nucleotide polymorphism of ψLcp in *D. melanogaster* shows extensive sequence length variations that also result in premature stop codons in *D. simulans*, supporting the contention that ψLcp is not a functional gene (162). As expected in a pseudogene not subject to selection, the rates of synonymous and nonsynonymous substitutions are equal, and the overall nucleotide divergence between *D. melanogaster* and *D. simulans* is extremely high. However, the within-species nucleotide variation of ψLcp (π = 0.001 ± 0.001) is lower than for many functional genes. The HKA test (105) reveals that the ψLcp polymorphism in *D. melanogaster* is significantly lower than expected, given the amount of divergence between *D. melanogaster* and *D. simulans*. The low level of intraspecific polymorphism of ψLcp has been attributed to background selection (162).

Currie & Sullivan (56) have characterized an intronless phosphoglyceromutase pseudogene (*Pglym87*) located on the right arm of chromosome 3 at position 87B4.5, which most likely originated from the retroposition of a *Pglym78* transcript, a gene located on the left arm of chromosome 3 at bands 78A/B. RNase protection experiments and primer extension analyses have failed to detect any transcript from *Pglym87*, although the *Pglym87* open reading frame remains
intact except for deletion of the first two codons. The structural properties suggest that Pglym87 is a pseudogene, whereas the intact reading frame and codon bias suggest that it might be a functional gene (56).

A processed Adh pseudogene (ψAdh) has been found in D. yakuba and D. teissieri (melanogaster species subgroup) (110, 111). The pseudogene is located on the third chromosome, whereas the functional Adh gene is on the second chromosome of the species. ψAdh exhibits a peculiar pattern of evolution for a pseudogene, including retention of reading frame, codon bias, and a higher rate of substitution at synonymous than at nonsynonymous nucleotide positions. Long & Langley (136) have proposed that ψAdh is a functional gene, which they called jingwei. They obtained a ratio 27:4 silent:replacement polymorphism, a pattern typical of active genes; detected transcripts of the gene and suggested that it arose by retroposition from the alcohol dehydrogenase gene, followed by recruitment of additional 5' exons and introns from an unrelated gene.

An unprocessed Adh pseudogene (ψAdh) has been detected in the repleta group of Drosophila (70; reviewed in 208). Some evidence suggests that the ψAdh is a pseudogene, since mutations have rendered the gene incapable of being translated into a functional alcohol dehydrogenase (208). However, the molecular evolution of ψAdh exhibits characteristics that are atypical of pseudogenes: (a) the rate of evolution is substantially slower in the exons of ψAdh than in the intergenic region, and only slightly faster than the rate of exons of functional Adh genes; (b) codon bias is retained in most species studied; (c) silent substitutions in ψAdh significantly exceed replacement substitutions. The Ks/Ka ratio (silent:replacement substitutions) ranges from 10 to 14, in pairwise interspecific comparisons involving sequences from seven species. These ratios greatly depart from unity, which is the value expected for a pseudogene (assuming no selective constraints), and are only slightly lower than those obtained from equivalent comparisons of the functional Adh genes (208). The repleta group ψAdh may be a chimeric functional gene (19).

A pseudogene has been described in the α-esterase gene cluster, which appears to be a bona fide pseudogene, in D. melanogaster (DmaE4a-ψ), where it has multiple inactivating mutations that are fixed in natural populations (177, 178, 185). The α-cluster includes 11 carboxyl/cholinesterase genes (10 active genes and 1 pseudogene) and locates within cytological region 84D3-E2 on chromosome 3R of D. melanogaster. The α-esterases have diverged substantially from one another (amino acid similarity ranging from 37% to 66%). No gene conversion or intergenic recombination has been detected between the genes and pseudogene. However, the D. simulans and D. yakuba orthologs of DmaE4a-ψ do not have inactivating mutations and appear to be functional (178).

The Amylase multigene family of D. pseudoobscura is located within a series of highly polymorphic inversions on the third chromosome. Four pseudogenes have been reported in three gene arrangements (ST, SC, and TL): ST Amy3-ψ, with a premature stop codon shortening the protein to 31.6% of its normal length (36); and TL Amy2-ψ, TL Amy3-ψ, and SC Amy3-ψ, with large deletions in their coding
regions (160). The divergence among the four pseudogene sequences reveals a retardation of sequence evolution in the SC and ST arrangements, whereas the rates of substitution in the two TL pseudogenes do not depart from neutral expectations. Gene conversion is most likely responsible for slowing down the divergence of SC \textit{Amy3}-ψ, and ST \textit{Amy3}-ψ (160). In the closely related \textit{D. miranda}, an \textit{Amylase} pseudogene (\textit{Amy}3), located on the secondary sex chromosome X2, contains two large deletions (one 445 bp and the other 872 bp long) (205).

If pseudogenes are devoid of function, their pattern of nucleotide substitution is expected to reflect the pattern of spontaneous point mutations, which would make pseudogenes ideal models to study neutral evolution at the molecular level (131). Of the four \textit{D. melanogaster} \textit{Amy} pseudogenes, only two (\textit{TL Amy2-ψ} and \textit{TL Amy3-ψ}) seem to behave as bona fide pseudogenes, whereas in the other two, SC \textit{Amy3-ψ} and ST \textit{Amy3-ψ}, sequence evolution has been retarded, most likely by homogenization caused by gene conversion. Clearly, caution should be exercised when using pseudogenes as exemplars for determining patterns and rates of neutral nucleotide substitution.

### Pseudogene Expression

A genomic region encompassing the \textit{Cecropin} locus (\textit{Cec}, located at 99E on the right arm of the third chromosome of \textit{D. melanogaster}) includes three functional genes and two pseudogenes (\textit{Cecψ1} and \textit{Cecψ2}) (126). The pseudogenes have diverged considerably from the functional \textit{Cec} genes (only 50% of the residues remain identical), have no consensus splice signals, and contain multiple premature stop codons and deletions. However, both \textit{Cecψ1} and \textit{Cecψ2} have retained a promoter-like region with a TATA box and capping site homology (126). The two pseudogenes are also highly diverged between themselves. But nonsynonymous polymorphism is lower than synonymous polymorphism in the coding region of both pseudogenes, which suggests functional constraint on amino acid replacement changes, whereas the level of silent variation is the same (for \textit{Cecψ1}) or higher (for \textit{Cecψ2}) than in the functional genes. Both pseudogenes have conserved transcriptional signals and splice sites, and present an open reading frame; also, correctly spliced transcripts have been detected for both pseudogenes. \textit{Cecψ1} and \textit{Cecψ2} may be either active genes with some null alleles or young pseudogenes (167).

### Regulation of Gene Expression

The X chromosome \textit{Stellate} (\textit{Ste}) gene of \textit{D. melanogaster} (at 47.5 and band 12E) is homologous to a moderately repeated sequence that maps to the \textit{Suppressor of Stellate} locus (\textit{Su(Ste)}) (133) located on the long arm of the Y chromosome (93), which contains tandemly repeated pseudogene sequences with premature stop codons and frameshift mutations (17). Nevertheless, some of the tandem pseudogenes of \textit{Su(Ste)} are transcribed (113). The Y chromosome \textit{Su(Ste)} sequences regulate the activity of the \textit{Ste} locus on the X chromosome (94, 133, 134); they repress transcription and alter the splicing of \textit{Ste} transcripts (17, 134). The suggested
models for repression include competitive interaction between Su(Ste) and Ste for positively acting transcription factor(s) (113, 134); Su(Ste) antisense transcription (see below) (57); and DNA-protein or protein-protein interaction affecting Ste transcription (113). However, no regulatory interactions have been demonstrated for another D. melanogaster gene–pseudogene pair, respectively located on the X and Y chromosomes (mst77F gene family) (184).

The β-esterase Cluster of D. melanogaster: Expression, Evolutionary Conservation, Recombination, and Gene Conversion

Balakirev & Ayala (12) have described a “cryptic” pseudogene within the β-esterase gene cluster of D. melanogaster. The β-esterase gene cluster is on the left arm of chromosome 3 of D. melanogaster, at 68F7-69A1 in the cytogenetic map (but see (163)). The cluster comprises two tandemly duplicated genes, first described as Est-6 and Est-P (50). The coding regions are 1686 and 1691 bp long, respectively, and consist of two exons (1387 bp and 248 bp) and a small (51 bp in Est-6 and 56 bp in Est-P) intron (152). The Est-6 gene is well characterized [reviewed in (153, 175)]. The gene encodes the major β-carboxylesterase (EST-6), which is transferred by D. melanogaster males to females in the seminal fluid during copulation (174) and affects the female’s consequent behavior and mating proclivity (90). Less information is available for Est-P. It was first described as a functional gene, based on several lines of evidence: transcriptional activity, intact splicing sites, no premature termination codons, and presence of initiation and termination codons (50). However, Balakirev & Ayala (12, 16) found premature stop codons within the Est-P coding region and some other indications suggesting that Est-P might be in fact a pseudogene, which they designated as ψEst-6. Dumancic et al. (65) showed that some alleles of Est-P produce a catalytically active esterase, corresponding to the previously identified EST-7 isozyme (101) and renamed the gene Est-7.

The comparative analysis of ψEst-6 and Est-6 has recently been extended to 28 strains and has included (13–16) entropy analysis, using spectral methods (43–47; reviewed in 135). Some features of ψEst-6 indicate that it could be a pseudogene: 11 premature stop codons among 28 sequences are hardly compatible with the functionality of the encoded protein; the number of amino acid replacements is 2.9 times higher in ψEst-6 than in Est-6, and some of them are drastic; nucleotide polymorphism is 2.1 times higher in ψEst-6; structural entropy analysis reveals significantly lower structural regularity and higher structural divergence for ψEst-6, in accordance with the expectations if it is a pseudogene or nonfunctional gene (16). However, as noted above, the gene can be expressed (50), and some alleles of ψEst-6 produce a catalytically active esterase (65), although this is detected in late larvae and adults of both sexes, whereas the functional Est-6 genes transcripts are found in all life stages but predominantly in adult males (50, 65), consistent with the significant role of EST-6 in male mating (90, 174). Moreover, the rate of synonymous substitutions is higher than the rate of nonsynonymous substitutions.
and neutrality tests are significant (16). However, and perhaps unexpectedly, the population recombination rate is 2.6 times lower in $\psi$Est-6 than in Est-6, so that linkage disequilibrium is more pronounced in $\psi$Est-6 than in Est-6 (81.2% versus 50.3% statistically significant associations between sites). Gene conversion occurs within both Est-6 and $\psi$Est-6, but rarely between the two genes, and has only been detected by using the inferred protein alignment, on the basis of mutations at silent sites that occurred after the gene conversion event. The haplotype structure is dimorphic in both $\psi$Est-6 and Est-6, each consisting of two very divergent sets of very similar alleles, but the divergent sequences of the two genes are not congruent; those of Est-6, but not those of $\psi$Est-6, correspond to a Slow/Fast allozyme dimorphism (16).

In conclusion, it seems that $\psi$Est-6, like other, if not most, Drosophila putative pseudogenes, exhibits features of both functional and nonfunctional genes (16).

**Evoluutionary Conservation of Pseudogenes**

Patterns of pseudogene evolution that are inconsistent with the neutral patterns expected if the pseudogenes lacked function have been found in other organisms besides Drosophila. In chicken, IgIV and IgHV pseudogenes have been identified that include crippling mutations (182). However, only a few IgIV and IgHV pseudogenes contain more than one crippling mutation and, notably, very few are crippled because of stop codons or frameshift mutations. The IgIV and IgHV pseudogene sequences contain significantly fewer stop codons generated by point mutation than would be expected under a model of random nucleotide change. Most nucleotide substitution-generated stop codons are rescued by additional nucleotide substitutions within the same codon. The majority of insertions and deletions are in frame (182).

Furthermore, the majority of IgV pseudogenes have not diverged significantly from functional genes; rather, they exhibit a nonrandom distribution of variability along the complementarity-determining regions (CDR), with a peak of nucleotide and amino acid variability at the 3′ end of the pseudogenes. The framework regions (FR) appear to have been highly conserved. Thus, the chicken IgIV and IgHV pseudogenes exhibit evidence of selection for protein function and display all the hallmarks of expressed IgV genes: selection for open reading frames, diversification of the CDRs, and conservation of the FRs (182). One plausible hypothesis is that selection for open reading frames acts on these genes because they donate DNA sequence during somatic gene conversion. However, an analysis of 217 gene conversion events at the chicken IgIVI locus reveals that pseudogenes are rarely utilized as gene conversion sequence donors; three of the pseudogenes were never used (144). Furthermore, most gene conversion events involve only a small segment of the pseudogene [mean size approximately 27 bp; (144)].

A similar state of affairs may obtain in the mouse, where a comparison of seven functional $V_H$ genes with nine $V_H$ pseudogenes leads to the conclusion that the IgV
pseudogenes may be subject to the same, or nearly the same, selection pressure as the functional genes (190). The pseudogenes have generally relatively minor defects, most frequently single nucleotide substitutions. Indels of single nucleotides leading to frameshifts or stop codons, or the loss of an initiation codon are not common. The relatively “intact” status of most immunoglobulin pseudogenes in the mouse has been noticed (22, 69, 81, 190). A single point mutation is responsible for the nonfunctionality of the human TCRG-V10 gene (246).

An interesting example of a “preserved” pseudogene, ψPgiC, has been described in the wildflower genus Clarkia (86). Relative-rate tests show that exon nucleotides have not diverged faster in the pseudogene than in the homologous functional gene. The ratio of synonymous to replacement substitutions in the pseudogene is low, as expected if there is selection. The observed pattern cannot be accounted for by gene conversion from PgiC1 to ψPgiC2. A possible explanation for the conservation of ψPgiC2 is the occasional formation of PGIC1-PGIC2 heterodimers, so that the pseudogene may be subject to selection against mutants causing defective heterodimers (86). Experimental evidence of such deleterious subunit interactions has recently been described for Cu/Zn superoxide dismutase in Drosophila (158).

The conservation of pseudogene sequences retaining 90% and more homology with their functional counterparts has been reported in many organisms, such as human [steroid 21-hydroxylase pseudogene CYP21P (103); folate binding protein (187); laminin receptor pseudogene, LAMRIL5 (173); macrophage stimulating protein, MSP (225)]; mouse [X-linked lymphocyte-regulated, KLR, unprocessed pseudogenes (79); heat-shock, hsp25, pseudogene (74)]; rat [U6snRNA pseudogene (200)]; snake [phospholipase A2-like pseudogene (112)]; maize [Gpa pseudogenes (165)]; Aspergillus nidulans [5SrRNA pseudogene (25)]; and Chlamydomonas reinhardtii [cysteine-rich pseudogene (140)]. For additional examples, see Vanin (226), Wilde (237), and Mighell et al. (145).

The extent of similarity between the pseudogenes and their homologous functional genes is often not uniform along the sequences. In the human ferrochelatase gene and pseudogene, nucleotide identity is 82–93% in exons 2–11, but only 63% in the 5′-flanking region (234). Between a human lactate dehydrogenase gene and its pseudogene, the 5′-flanking region has twice as many differences as the coding and 3′-flanking regions (207). In salmon, there is high conservation (up to 80–90%) between the intron sequences of the growth hormone pseudogene and the functional gene, but much less homology in the coding regions (114). Similarly, in a phospholipase pseudogene from the Mojave rattlesnake, there is high conservation of the intronic regions but much less in the protein-coding regions (112). On the contrary, in the green alga C. reinhardtii, a cysteine-rich protein pseudogene and the functional gene are similar within coding regions, but the similarity drops dramatically in the intron and in the 5′- and 3′-noncoding regions. The first and second exon sequences show a 91% and 74% similarity, respectively, whereas the 5′- and 3′-flanking regions are only 51% and 41% similar. Similarity between the intron sequences is even lower, 40%, and is clustered toward the exon-intron boundaries (140).
PSEUDOGENE TRANSCRIPTION AND EXPRESSION

The original definition of pseudogenes implied that they should be transcriptionally and translationally silent and hence not subject to selection (131). However, non-functional as well as functional transcripts of pseudogenes have been described in many organisms. The human examples are numerous and include interferon pseudogene $\psi$LeIFNE [(84); IFNAP22 (63)]; glyceraldehydes-3-phosphate dehydrogenase pseudogene $\psi$Gapdh (9, 224); glucocerebrosidase pseudogene psGBA (204); dopamine D5 pseudogene $\psi$DRD5-1 (148); dopamine $D_1$ pseudogene (232); complement factor H-related protein 1 pseudogene $H36-2$ (201); DNA topoisomerase 1 pseudogene TOP1 (248); serotonin 5-hydroxytryptamine receptor $\psi$5HT1D (18); glutamine synthetase $\psi$GS (39); neuropeptide Y1-like receptor $\psi$NPY Y1-like (181); steroid 21-hydrolase pseudogene CYP21P or CYP21A (67); type I hair keratin $\psi$hHaA, which is transcribed in human (180), whereas its orthologs are also expressed in chimpanzee and gorilla (238); tumor suppressor $\psi$PTEN (75); 5-HT7 receptor $\psi$5-HT7 (154); translationally controlled tumor protein pseudogene $\psi$TPT1, transcribed in all tissues investigated, but with widely variable expression in different human tissues, whereas its orthologs are fully expressed in the rabbit (215); and myosin pseudogene $\psi$MYO15BP (24).

Pseudogene transcripts have been detected in other mammals, including mouse glyceraldehyde-3-phosphate dehydrogenase $\psi$Gapdh (77, 224); bovine aromatase pseudogene Cyp19 (76) and cytochrome b$_5$ pseudogene (55); and ribonuclease pseudogene $\psi$BS RNAase in sheep (52). Other animal examples include the silk moth chorion locus, high-cysteine protein pseudogene $\psi$HcB.12/13 (72). Plant examples are rps19 pseudogene in Oenothera berteriana (191), rps14 pseudogene (10) and myrosinase pseudogene $\psi$TGG3 (245) in Arabidopsis thaliana, potato ribosomal protein $\psi$rps14 (166), and liverwort NADH dehydrogenase pseudogene (211). Other examples are erythrocyte binding protein $\psi$EBA165 in Plasmodium falciparum (222) and yeast adenine phosphoribosyltransferase pseudogene $\psi$APT2 (2).

Human pseudogene translation activity has been shown in vivo (33, 142) as well as in vitro (146; for mouse, see 27) or, simply, suggested (18, 232). The human glucocerebrosidase pseudogene psGBA promoter has some activity when attached to a reporter gene (169). This pseudogene is transcriptionally active in vivo and the amount of pseudogene-derived mRNA is sometimes comparable to the amount of mRNA derived from the active gene, indicating that pseudogene transcription occurs at a high level (204). However, there is considerable subject-to-subject variation in the amount of pseudogene transcript relative to the functional gene, which could be due to polymorphism in the promoter regions of the pseudogene or to polymorphisms affecting the stability of pseudogene-derived mRNA (204). Interestingly, the pseudogene transcripts can be detected in tissues not known to express the functional counterpart gene (154), or in only some of the tissues where the functional homolog is specifically expressed (245). This situation is reminiscent of $\psi$Est-6 in Drosophila melanogaster, which is transcribed mostly
in late larvae and adults of each sex, whereas the functional Est-6 gene transcripts are detected in all life stages but predominate in adult males (50, 65).

REGULATION OF GENE EXPRESSION

McCarrey & Riggs (141) have suggested a possible role for pseudogenes in development as a source of intracellular inhibitors. They hypothesized that pseudogenes could be the source of the antisense RNA that hybridizes with the sense RNA from the determinator genes, thereby blocking their expression. Pseudogenes, or portions thereof, may be transcribed from the opposite strand relative to their functional counterparts, which would make them a source of antisense RNA. If the inhibitors act at the protein level, then they should be identifiable, for example, as proteins that modify the electrophoretic migration of determinator proteins. One prediction derived from this hypothesis has been confirmed, namely that the sequence of duplicated genes that are transcribed in opposite directions should be conserved whenever the sense-antisense interaction is advantageous (141).

There are numerous examples in various organisms of unusually high level of pseudogene conservation, exceeding 90% of the sequence, as pointed out above. Moreover, pseudogenes are prevalent in the genomes of multicellular organisms and scarce in those of unicellular organisms, such as bacteria and yeast (97, 128), which is consistent with the hypothesis of McCarrey & Riggs (141). Korneev et al. (122) have shown that a nitric oxide synthase pseudogene (pseudo-NOS) and its paralogous functional gene (nNOS) are co-expressed in identifiable neurons of the mollusk Lymnaea stagnalis. The pseudo-NOS transcript includes a region with significant antisense homology to the nNOS mRNA. The antisense region of the pseudo-NOS RNA specifically suppresses the synthesis of the nNOS protein (122). Thus the pseudo-NOS transcript acts as an antisense regulator of nNOS protein synthesis.

There has been much speculation and some evidence that pseudogenes may have regulatory roles for the genes from which they have been derived (72, 73, 106, 248). Healy et al. (100) have shown that 3′ sequences that lie within the ψEst-6 pseudogene transcription unit of D. melanogaster contain elements that modulate the expression of Est-6, which obviously implies some regulatory function for ψEst-6. Sequence similarity exists in a 390-bp block within the 609-bp 3′-flanking region of Est-P (D. melanogaster) and the 3′-flanking sequence of Est-5A (D. pseudoobscura) (28, 50). In particular, a segment of 110 bp within this region shows 76% sequence similarity between the Est-P and Est-5A, which sharply contrasts with 20% or lower similarity in the 5′ flanking region between the same two genes or between Est-6 and Est-5B (D. melanogaster versus D. pseudoobscura) (28). Obvious reductions in intraspecific variation as well as interspecific divergence also occur in this region between D. melanogaster and D. simulans (16).

Troyanovsky & Leube (223) have described an interesting example of gene/pseudogene cooperation in human cytokeratin 17 expression. A detailed examination of cytokeratin transcription regulation using gene/pseudogene chimeric
constructs has identified specific promoter/enhancer elements that are inactive by themselves but can interact to induce strong transcriptional activity of reporter genes. The process includes the interaction between the proximal region of the inactive cytokeratin pseudogene promoter and the distal upstream region of the actively transcribed cytokeratin gene. Troyanovsky & Leube (223) conclude that cis elements in the proximal 5′-upstream region of the pseudogene promoter can cooperate with distal enhancer elements of the functional gene to induce strong transcriptional activity in transfected HeLa cells. In mice, the Makorin1-p1 pseudogene regulates the messenger-RNA stability of its homologous coding gene (Makorin1) by competitive interaction, either at the RNA or DNA level (104a). Similar mechanisms have been previously suggested for regulatory gene-pseudogene interaction (113, 134).

**PSEUDOGENES AS RESERVOIRS FOR GENERATING GENETIC DIVERSITY**

**Antibody Diversity**

A functional role for pseudogenes in the immune response has been proposed for human (4, 155, 228, 242), chicken (20, 143, 144, 171, 172, 189), rabbit (71, 118), and other vertebrates. Immunoglobulin gene diversity is generated by somatic gene conversion events in which sequences derived from alleles or paralogous genes generate new gene sequences. Gene conversion may also occur between genes and pseudogenes, so that selection may contribute to conserving the sequence and other functional characteristics of pseudogenes. In chicken, the process has been shown to occur in the germ line as well as somatically during antibody diversification (20). Together with the observations that the immunoglobulin pseudogenes are highly conserved, have open reading frames, and retain canonical structure sequence patterns, these observations support the hypothesis that the immunoglobulin pseudogenes resemble a functional multigene family maintained by selection for its functional role in generating somatic antibody diversity (143).

**Antigenic Variation**

Pseudogenes act as a reservoir of sequences that recombine with their functional paralogous genes and thus generate antigenic diversity. Gene conversion events have been reported in several bacterial pathogens as a mechanism for generating sequence diversity in expressed antigenic genes. In the bacterium *Borrelia hermsii*, a vmp pseudogene serves as a source of antigenic variation during relapsing fever (170). Bacteria of the genus *Borrelia* generate antigenic diversity of the vmp/vls coat proteins through recombination from a tandem array of silent partial pseudogene cassettes into a telomeric expression site on a linear plasmid (244). Antigenic variation is generated in *Anaplasma marginale* (a member of ehrlichial genogroup II) by recombination of pseudogenes into the functional expression site (30, 31).
Pseudogenes are also a source of genetic diversity in gonococci. These bacteria generate pilin variants that allow evasion of the host immune system. In this process, a silent, partial pilin gene sequence is introduced into the expression locus to form a complete pilin gene (91). Pilin phase variation can also occur by deletions between direct repeats internal to the expressed pilin gene or between a pilin gene and an adjacent silent pseudogene (104).

Antigenic variation of surface proteins in Mycoplasma (151) is achieved by recombination between multiple pseudogenes and the expressed vlbA gene; the multiple pseudogenes serve as a repertoire for the creation of vlbA sequence diversity, probably by site-specific recombination, with recombination between the single complete vlbA gene and one of the multiple partial copies creating new vlbA gene variants (151). Subsequent gene conversion, drawing on the shorter pseudogenes, may introduce further variability into this new variant. Thus, the expressed genes are likely to be chimeric, with sequences derived from several overlapping pseudogenes (151). The discrete boundaries of sequence similarity observed for the pseudogenes strongly suggest a site-specific mechanism of gene conversion (151).

In Trypanosoma cruzi antigenic diversity is generated by segmental gene conversion (3). Recombination between vsg genes and pseudogenes has been suggested as a mechanism by which antigenic diversity is generated in Trypanosoma equiperdum (217), as well as in the T. cruzi genome, by recombination between surface protein pseudogenes (212, 213).

Other Gene Systems

The human olfactory receptor (OR) pseudogenes may be important for the generation and maintenance of receptor diversity (82). Intensive intergenic gene conversion has been revealed for this multigene family that leads to segment shuffling in the odorant binding site, an evolutionary process reminiscent of somatic combinatorial diversification in the immune system (196). Although OR pseudogenes have lost full coding function, they are apparently under new evolutionary constraints: OR pseudogenes adopt noncoding functions as CpG islands (82), enhancers (37), and matrix attachment regions (80).

RECOMBINATION AND SEQUENCE POLYMORPHISM

Recombination is a key enabling mechanism in the continuing evolution of life on earth. It is widely believed to have originated in bacteria or their progenitors as a DNA repair process. The biochemistry of recombination has reinforced a long-established link between recombination and repair (54, 125, 179). Actively transcribed yeast genes show increased recombination rates when compared to identical, nontranscribed genes (216).

Disruption of homologous sequences by indels or nucleotide polymorphism can significantly reduce recombination frequencies (32, 83, 230). Recombination
within the maize \(bz1\) gene is inhibited when one allele with a transposon insertion is paired with a second allele lacking the insertion (64). A similar outcome was observed at the maize \(a1\) locus (241). In \(D.\ melanogaster\) the observed reduction of recombination in \(\psi\) Est-6 and the absence of gene conversion between Est-6 and \(\psi\) Est-6 (16) may be due in some strains to indels, such as the insertion of the \(mdg-3\) retrotransposon within the intron of \(\psi\) Est-6 detected in some strains (16, 78).

The recombination machinery may be sensitive even to a single nucleotide mismatch. Individual nucleotide substitutions have been shown to affect recombination in yeast (26, 195), bacteria (198, 199), and mammalian cells (137). Strong inhibition of intrachromosomal recombination, sometimes by as much as 100- to 1000-fold, due to small sequence divergence, has been observed in the mouse (60, 229), mammalian cells (66, 243), yeast (11, 48, 58, 59, 95, 195), and bacteria (49, 139, 168). Inhibition is largely mediated by the mismatch repair system, which is thought to scan hybrid DNA and abort recombination when too many mismatches are detected (41, 58, 59, 61, 147, 195). Conversion tract lengths for spontaneous ectopic events are apparently reduced by sequence divergence, an effect that can also be explained by hybrid DNA rejection (48, 95).

As noted earlier, the recombination rate is lower in \(\psi\) Est-6 of \(D.\ melanogaster\) than in the Est-6 gene. Moreover, no (or little) gene conversion occurs between Est-6 and \(\psi\) Est-6 (16). The nucleotide sequences of the coding regions of Est-6 and \(\psi\) Est-6 show 64% similarity (50). This level of nucleotide divergence probably does not satisfy the homology requirements for efficient intergenic conversion and represents sufficient sequence divergence to disrupt conversion between the genes. However, interlocus gene conversion occurs between Est-5A and Est-5B of \(D.\ pseudoobscura\) that are orthologous to Est-6 and \(\psi\) Est-6, respectively, of \(D.\ melanogaster\), but exhibit much higher similarity, 82.5%, between the coding regions (28, 117).

Recombination is not random within the human \(\beta\)-globin gene cluster (115, 233). There is a “hot spot” of recombination between the \(\delta\) and \(\beta\) genes (115). This phenomenon could be caused by specific sequences promoting recombination. The 5′-flanking region of the gene is known to have a chi-like sequence (5′-GTCGGTG-3′), which is presumably a promoter of recombination (40). Chi-like sequences are involved in recombination in bacteriophage \(\lambda\) (202), murine immunoglobulin genes (124), human and murine class I and class II genes in the major histocompatibility complex (239), and in human \(CYP21\) genes (6, 129). The sequence GGXGGX has been implicated to function as a hot spot for gene conversion events in an early chorion multigene family (102, 123). The conversion events between \(C\beta1\) and \(C\beta2\) genes of the T-cell antigen receptor \(\beta\)-chain locus in murine (183) and rabbit (138) are limited to exon 1. Specific recognition sequences that promote conversion were suggested to explain the conversion pattern observed (183).

Other examples of specific sequences promoting recombination are suggested in the vertebrate \(Major\ Histocompatibility\ Complex\ (MHC-)\ DRB\) genes. Every second intron of all investigated \(MHC-DRB\) genes contains a defined stretch of tandemly repeated simple DNA sequences (5, 176, 192), the basic structure of
which is \((gt)\_m(ga)\_m\). These simple sequences have been found in proximity to presumed hot spots of recombination in the mouse \textit{MHC} (119) and the human \textit{HLA-DQ} region (188). This fact has led to the suggestion that a causal relation exists between the simple repeated DNA structures and the recombination hot spots in the \textit{MHC} [reviewed in (188, 206)]. It is interesting that unprecedented evolutionary stability of simple repeats in the expressed mammalian \textit{MHC-DRB} genes occurs in some specific genome locations (5, 8, 29, 89, 176).

In mammalian \textit{DRB} pseudogenes, the simple repeat stretch seems to gradually reduce its characteristic pattern in the evolutionary course. Particularly, the \textit{HLA-DRBX} pseudogene (127) harbors only a short \((ga)\_m(ca)\_m\) stretch, hence lacking the basic \((gt)\_m(ga)\_m\) structure. The simple repeat structure of the sheep \textit{Ovar-DRB3} pseudogene is highly disintegrated and in the \textit{Ovar-DRB4} pseudogene only three copies of each dinucleotide remained detectable. Simple repeat sequences in the human pseudogene \textit{DRB2} also show signs of shortening and disintegration (176).

Defective DNA recombination signals in the T-cell receptor A joining-region pseudogenes have been identified in mice (121), primates (214), and human (120). Absence of recombination signal sequences or crippling mutations in these signal sequences have been revealed in chicken \textit{IglV} and \textit{IghV} pseudogenes (171, 172). Bliskovskii et al. (23) have described the structure of the human \textit{son} processed pseudogene, which has a 96% homology with the \textit{son} functional gene. Despite the high sequence homology, the \textit{son} pseudogene lacks five monomers of the perfect tandem repeat area, which are present in the homologous \textit{son} gene region and which, as suggested, are associated with the initiation of recombination processes (51, 109, 221).

In several yeast genes, the sequence signals initiating recombination often occur within the promoter but not within the gene itself (38, 68, 157, 236, 240). Disruption of the \textit{HIS4}, \textit{GAL10}, and \textit{ARG4} promoter regions inhibit recombination and gene conversion within these loci (149, 157, 216, 235). Even partial deletions of the yeast \textit{ARG4} and \textit{HIS4} promoter regions inhibit recombination within these genes (149, 235). A role for promoter sequences in initiating recombination has been identified in maize \textit{a1} (241) and \textit{b1} (156) genes.

In \textit{D. melanogaster}, the promoter of \(\psi\)\textit{Est-6} is limited to the 193 bp encompassing the intergenic region between \textit{Est-6} and \(\psi\)\textit{Est-6} (50), and it may be that the observed reduction in the rate of recombination of \(\psi\)\textit{Est-6} is associated with the promoter’s truncation. The sequences promoting recombination could also be eroded within \(\psi\)\textit{Est-6} owing to stochastic accumulation of mutations, as in cases of the \textit{HLA-DRB} (127, 176) and \textit{son} (23) pseudogenes.

**CONCLUSIONS**

We have described in \textit{Drosophila} and other organisms, including humans, pseudogene features that would be unexpected if pseudogenes were nonfunctional sequences of genomic DNA (“junk” DNA). An extensive and fast-increasing
literature does not justify a sharp division between genes and pseudogenes that would place pseudogenes in the class of genomic “junk” DNA that lacks function and is not subject to natural selection. Pseudogenes are often extremely conserved and transcriptionally active. Moreover, there are direct indications that pseudogenes can be functional by being involved in the regulation of gene expression and in generating genetic diversity. Eukaryote genomes contain many pseudogenes that have avoided strong degeneration because they are important components of the genome (34, 35, 96). Pseudogenes may lose some specific functions but retain others, and even acquire new ones, which may not be simply recognizable.

Pseudogenes may recover the full original function of the genes from which they derive. Pseudogene function has been restored in vitro by mutagenesis (247), transfection (193), or in vivo by site-specific (or intermolecular) recombination (21, 170). Possible restoration of function by gene conversion has been proposed for the ribonuclease pseudogene in artiodactyl evolution (220). In the chlorella virus PBCV-1, a nonfunctional, nontranscribed cytosine DNA methyltransferase pseudogene differs by eight amino acids from a functional counterpart, but a single amino acid change or deletion by site-directed mutagenesis restores the pseudogene’s activity (247). Restoration by site directed mutagenesis has been obtained in the yeast Schizosaccharomyces pombe (159). A strain of E. coli deleted for the lacZ gene has been shown to give rise to spontaneous lactose-utilizing mutants as a result of the activation of the ebgp pseudogene [reviewed in (92)].

How pervasive are “functional” pseudogenes? Many pseudogenes have been identified in all sorts of organisms on the grounds that they are duplicated genes that exhibit stop codons or other disabling mutations in their DNA sequences, so that they cannot have the full function of the original genes from which derived. In many of these cases, however, it remains unknown, because it has not been investigated, whether the pseudogenes, described only on the basis of DNA sequences, may have acquired regulatory or other functions, or play a role in generating genetic variability. There seems to be the case that some functionality has been discovered in all cases, or nearly, whenever this possibility has been pursued with suitable investigations. One may well conclude that most pseudogenes retain or acquire some functionality and, thus, that it may not be appropriate to define pseudogenes as nonfunctional sequences of genomic DNA originally derived from functional genes, or as “genes that are no longer expressed but bear sequence similarity to active genes” (99, p. 114). Rather, pseudogenes might be defined as DNA sequences derived by duplication or retroposition from functional genes that are often subject to natural selection and therefore retain much of the original sequence and structure because they have acquired new regulatory or other functions, or may serve as reservoirs of genetic variability.

Much remains to be learned about the population dynamics of pseudogenes. This much is clear: It is unwarranted to assume that “all mutations occurring in pseudogenes are selectively neutral and become fixed in the population with equal probability” (87, p. 124). Once a pseudogene appears in a population, presumably in a single genome, by duplication or retroposition, the pseudogene may at first
follow neutral population dynamics and become extinct or not. A pseudogene may rapidly become fixed if it is closely linked to a gene experiencing a selective sweep, or slowly increase in frequency by neutral drift. However, a mutant that provides a functional role for a pseudogene may be favored by natural selection and, thus, have a higher probability of becoming fixed than its nonfunctional alleles (137a,b). It may be appropriate to include pseudogenes within the class named by Brosius & Gould (34) as “potogenes.” These authors pointed out that the products of gene duplication, including those that become pseudogenes, may eventually acquire distinctive functions, and thus might be called potogenes to call attention to their potentiality for becoming new genes.

ACKNOWLEDGMENTS

This work is supported by NIH grant GM42397 to F. J. Ayala.

The Annual Review of Genetics is online at http://genet.annualreviews.org

LITERATURE CITED

in mismatch repair in *Saccharomyces cerevisiae* stimulates ectopic recombination between homeologous genes by an excision repair dependent process. Genetics 126:535–47


48. Chen W, Jinks-Robertson S. 1999. The role of mismatch repair machinery in
regulating mitotic and meiotic recombination between diverged sequences in yeast. *Genetics* 151:1299–313


75. Fuji GH, Morimoto AM, Berson AE, Bolen JB. 1990. Transcriptional analysis of the PTEN/MMAC1 pseudogene, ΨPTEN. *Oncogene* 18:1765–69


112. John TR, Smith JJ, Kaiser II. 1996. A phospholipase A 2 -like pseudogene retaining the highly conserved introns of Mojave toxin and other snake venom group II PLA 2s, but having different exons. *DNA Cell Biol.* 15:661–68


143. McCormack WH, Thompson CB. 1990. IgL variable gene conversion display pseudogene donor preference and 5′ to 3′ polarity. Genes Dev. 4:548–58
144. McCormack WH, Thompson CB. 1990. IgL variable gene conversion display pseudogene donor preference and 5′ to 3′ polarity. Genes Dev. 4:548–58
146  BALAKIREV □ AYALA


gene arrangement in *Solanum tuberosum*: *rps14* is a transcribed and unedited pseudogene. *Plant Mol. Biol.* 31:937–43


190. Schiff C, Mili M, Fougereau M. 1985. Functional and pseudogenes are similarly organized and may equally contribute to the extensive antibody diversity of the IgVHI family. *EMBO J.* 4:1225–30


four clustered genes. Cell 29:1027–40
222. Triglia T, Thompson JK, Cowman AF. 2001. An EBA175 homologue which is transcribed but not translated in


CONTENTS

IRA HERSKOWITZ (1946–2003), The Editors

TRANSPOSON-BASED STRATEGIES FOR MICROBIAL FUNCTIONAL
GENOMICS AND PROTEOMICS, Finbarr Hayes

ERROR-PRONE DNA POLYMERASES: WHEN MAKING A MISTAKE IS
THE ONLY WAY TO GET AHEAD, Alison J. Rattray and
Jeffrey N. Strathern

GENETICS OF HAIR AND SKIN COLOR, Jonathan L. Rees

THIOL-BASED REGULATORY SWITCHES, Mark S.B. Paget and
Mark J. Buttner

PSEUODGENES: ARE THEY “JUNK” OR FUNCTIONAL DNA?
Evgeniy S. Balakirev and Francisco J. Ayala

UNUSUAL LIFE STYLE OF GIANT CHLORELLA VIRUSES,
James L. Van Etten

GENETICS OF LACTASE PERSISTENCE AND LACTOSE INTOLERANCE,
Dallas M. Swallow

CELL POLARITY AND THE CYTOSKELETON IN THE CAENORHABDIS
ELEGANS ZYGOTE, Stephan Q. Schneider and Bruce Bowerman

THE SPINDLE ASSEMBLY AND SPINDLE POSITION CHECKPOINTS,
Daniel J. Lew and Daniel J. Burke

LATERAL GENE TRANSFER AND THE ORIGINS OF PROKARYOTIC
GROUPS, Yan Boucher, Christophe J. Douady, R. Thane Papke,
David A. Walsh, Mary Ellen R. Boudreau, Camilla L. Nesbo,
Rebecca J. Case, and W. Ford Doolittle

GENETICS OF AGING IN THE FRUIT FLY, DROSOPHILA MELANOGASTER,
Stephen L. Helfand and Blanka Rogina

NATURAL SELECTION AND THE EVOLUTION OF GENOME IMPRINTING,
Elena de la Casa-Esperón and Carmen Sapienza

THE NEED FOR WINTER IN THE SWITCH TO FLOWERING,
Ian R. Henderson, Chikako Shindo, and Caroline Dean

TRANSMISSION RATIO DISTORTION IN MICE, Mary F. Lyon
CONTENTS

STRUCTURE, DIVERSITY, AND EVOLUTION OF PROTEIN TOXINS FROM SPORE-FORMING ENTOMOPATHOGENIC BACTERIA, Ruud A. de Maagd, Alejandro Bravo, Colin Berry, Neil Crickmore, and H. Ernst Schneef 409

YEAST VACUOLE INHERITANCE AND DYNAMICS, Lois S. Weisman 435

HEPARAN SULFATE CORE PROTEINS IN CELL-CELL SIGNALING, Kenneth L. Kramer and H. Joseph Yost 461

RETOGRANPOSONS PROVIDE AN EVOLUTIONARILY ROBUST NON-TELOMERASE MECHANISM TO MAINTAIN TELOMERES, Mary-Lou Pardue and P. G. DeBaryshe 485

A CYANOBACTERIAL CIRCADIAN TIMING MECHANISM, S. S. Golden, J. L. Ditty, and S. B. Williams 513

REGULATION OF CELL CYCLES IN DROSOPHILA DEVELOPMENT: INTRINSIC AND EXTRINSIC CUES, Laura A. Lee and Terry L. Orr-Weaver 545

RECOGNITION AND RESPONSE IN THE PLANT IMMUNE SYSTEM, Zachary Nimchuk, Thomas Eulgem, Ben F. Holt III, and Jeffrey L. Dangl 579

RECA-DEPENDENT RECOVERY OF ARRESTED DNA REPLICATION FORKS, Justin Courcelle and Phillip C. Hanawalt 611

INDEXES
Subject Index 647
Cumulative Index of Contributing Authors, Volumes 33–37 679
Cumulative Index of Chapter Titles, Volumes 33–37 682

ERRATA
An online log of corrections to Annual Review of Genetics chapters may be found at http://genet.annualreviews.org/errata.shtml