MOLECULAR AND CELLULAR BIOLOGY, Aug. 1985, p. 2090–2103 Vol. 5, No. 8
0270-7306/85/082090-14$02.00/0
Copyright © 1985, American Society for Microbiology

RNA-Mediated Gene Duplication: the Rat Preproinsulin I Gene Is a Functional Retroposon

MARCELO BENTO SOARES,[1] ERIC SCHON,[1†] ANN HENDERSON,[1‡] SOTIRIOS K. KARATHANASIS,[2] RICHARD CATE,[3§] SCOTT ZEITLIN,[1] JOHN CHIRGWIN,[4] AND ARGIRIS EFSTRATIADIS[1*]

Department of Human Genetics and Development, Columbia University, New York, New York 10032[1]; Department of Cardiology, Harvard Medical School, Children's Hospital Medical Center, Boston, Massachusetts 02115[2]; The Biological Laboratories, Harvard University, Cambridge, Massachusetts 02138[3]; and Department of Anatomy and Neurobiology, Washington University School of Medicine, St. Louis, Missouri 63110[4]

Received 27 March 1985/Accepted 13 May 1985

Rats and mice have two, equally expressed, nonallelic genes encoding preproinsulin (genes I and II). Cytological hybridization with metaphase chromosomes indicated that both genes reside on rat chromosome 1 but are approximately 100,000 kilobases apart. In mice the two genes reside on two different chromosomes. DNA sequence comparisons of the gene-flanking regions in rats and mice indicated that the preproinsulin gene I has lost one of the two introns present in gene II, is flanked by a long (41-base) direct repeat, and has a remnant of a polydeoxyadenylate acid tract preceding the downstream direct repeat. These structural features indicated that gene I was generated by an RNA-mediated duplication-transposition event involving a transcript of gene II which was initiated upstream from the normal capping site. Sequence divergence analysis indicated that the pair of the original gene and its retroposed, but functional, counterpart (which appeared about 35 million years ago) is maintained by strong negative selection operating primarily on the segments encoding the chains of the mature hormone, whereas the segments encoding the parts of the polypeptide that are eliminated during processing and also the introns and the flanking regions are evolving neutrally.

Rats (as well as mice and three fish species) have two insulins instead of one, in contrast to other organisms (13, 23). These rat hormones are the products of two nonallelic preproinsulin genes (57) that are almost equally expressed (11). Characterization of the duplicated genes (36), isolated from a rat chromosomal DNA library, indicated a significant structural difference between them. The gene for preproinsulin I has a single 119-base-pair (bp) intron interrupting the segment corresponding to the 5′ noncoding region of the mRNA, whereas the other gene, encoding preproinsulin II, contains in addition to this small intron a second 499-bp intron interrupting the segment encoding the C-peptide. Since the structures of the unique chicken (46) and human (2, 62) genes are similar to that of the rat gene II (two introns at corresponding positions) we concluded that the two-intron organization corresponds to that of the common ancestor and that introns can be lost during evolution (46). This conclusion was strengthened by the subsequent determination of the structures of the unique dog (30) and guinea pig (7) preproinsulin genes, which are of the two-intron type.

Intron loss was subsequently documented in a mouse α-globin pseudogene (45, 69) and later in other pseudogenes that exhibit clearly the features of processed genes (22; see references 34, 56, and 68 for reviews). These observations raised the possibility that the rat preproinsulin I gene was a retroposon (51) that had been generated by an RNA-mediated duplication-transposition event but which for some reason remained functional. To examine this possibility we mapped the chromosomal location of the two genes and characterized their flanking regions by DNA sequencing to define the break points of the duplication unit. Here we show that the rat (and also the mouse) preproinsulin gene I is a functional semi-processed gene, generated by reinsertion into the genome (retroposition) of the structural information of a partially spliced transcript initiated upstream from the bona fide capping site.

MATERIALS AND METHODS

Enzymes and DNAs. Restriction enzymes were from New England Biolabs, Bethesda Research Laboratories, and Boehringer-Mannheim; DNA polymerase, DNA ligase, and RNA ligase were from New England Biolabs; T4 polynucleotide kinase was from Bethesda Research Laboratories; reverse transcriptase was from Life Sciences; S1 nuclease was from Boehringer-Mannheim and Bethesda Research Laboratories; phytohemagglutinin was from Wellcome Reagents; colcemid was from Sigma Chemical Co.; trypsin was from GIBCO Laboratories; [α-^{32}P]dNTPs (600 or 3,200 Ci/mmol), [γ-^{32}P]ATP (7,000 Ci/mmol), and [^{125}I]dCTP (2,000 Ci/mmol) were from New England Nuclear Corp.

DNA fragments were subcloned into pUC9 or M13 vectors (70). DNA sequencing was performed by the enzymatic method (54) and occasionally by the chemical method (39) for verification. Chromosomal DNA was analyzed by the method of Southern (58). Northern analysis was performed as described previously (76).

Cytological hybridization. Rat chromosomes were obtained from the established rat cell lines BRL and E-11, and also from peripheral lymphocytes. The chromosome com-

[*] Corresponding author.

[†] Present address: Department of Neurology, College of Physicians and Surgeons, Columbia University, New York, NY 10032.

[‡] Present address: Department of Biological Sciences, Hunter College, New York, NY 10021.

[§] Present address: Biogen Research Corp., Cambridge, MA 02142.

plement of BRL cells is pseudodiploid, and deviations from the normal karyotype, none of them involving chromosome 1, have been described (48). E-11 cells (16), provided by P. Fisher, are also pseudodiploid and occasionally trisomic for chromosome 1. They also contain a $7q^+$ chromosome and an $8q^+$ chromosome and a marker chromosome of unknown origin. Chromosomes from peripheral lymphocytes were obtained by phytohemagglutinin stimulation, as described previously (M. B. Soares and A. Henderson, submitted for publication). Cytological hybridizations were carried out as described (48), using DNA probes nick translated with $[^{125}I]$dCTP to a specific activity of 0.5×10^9 to 2×10^9 dpm/μg. Autoradiographic exposure times were between 14 and 21 days depending on the experiment.

In vitro transcription and S1 mapping. In vitro transcription was carried out by using a HeLa cell extract (38). RNA was synthesized in a 20-μl reaction containing 12 mM HEPES (*N*-2-hydroxyethylpiperazine-*N*-2′-ethanesulfonic acid) (pH 7.9), 7.5 mM $MgCl_2$, 90 mM KCl, 60 μM EDTA, 1.2 mM dithiothreitol, 10% glycerol, 500 μM each ATP, GTP, and CTP, 50 μM UTP, 4 mM creatine phosphate, and 75 μg of double-stranded DNA template per ml (gel-purified restriction fragment). After incubation for 60 min at 30°C, the reaction was terminated by the addition of sodium acetate (pH 5.2) to a final concentration of 50 mM, Sarkosyl to 0.5%, and EDTA to 10 mM (final volume of 300 μl). A 40-μg amount of tRNA was added, and the reaction was phenol extracted and ethanol precipitated.

Single-stranded, end-labeled probes for S1 mapping were generated by strand separation (39) with 6% nondenaturing acrylamide gels (59:1, acrylamide-bis acrylamide).

S1 mapping was performed as described previously (4), using for each assay RNA from one-half of a transcription reaction and about 5 μg of end-labeled probe. Protected fragments were analyzed on denaturing urea-polyacrylamide gels.

Sequence divergence analysis. Alignments were derived according to the arbitrary, but consistent, rules described previously (46). Each gap, regardless of length, was scored as one site and one mismatch. The alignments used for the calculations shown in Table 2 will be made available upon request. The sequences used for the alignments were retrieved from the GenBank and the EMBL data libraries, except for the 5′ flanking region of the rat POMC gene (J. Roberts, unpublished data) and the 3′ flanking region of the mouse serum albumin gene (R. W. Scott and S. Tilghman, unpublished data).

Uncorrected percent divergences (percent substitution values) were calculated as 100 × substitutions/sites. A percent substitution value (λ) for a noncoding region can be corrected for back mutations by using the formula $-3/4 \ln (1 - 4/3\lambda)$. Uncorrected divergences for coding regions were calculated as described previously (7). Because the percent substitution values were derived by using different methods for coding and noncoding regions, there is no one-to-one correspondence between such values, and they cannot be compared without special corrections (an issue outside the scope of this article). Since our rat-to-mouse or rat (mouse)-to-human comparisons involved species which diverged at the same evolutionary time, we used mostly uncorrected percent divergence values to avoid making the (possibly unfounded) assumptions that are necessary for corrections.

It should be emphasized that the numbers we have derived from the calculations are only indicative, because the statistical significance of the comparisons is low (small samples and, on occasion, short lengths of compared sequences). Moreover, the calculations depend on the quality of the alignments, for which unique solutions do not necessarily exist. This problem (proper alignment of only a few divergent sequences) is mainly due to the fact that the observed differences in a pairwise comparison are the combined result of the fixation of two kinds of evolutionary events, point mutations and deletions-additions of nucleotide blocks.

To calculate averages we assumed that all comparisons have equal weight, which is an oversimplification. For example, the mouse-to-rat comparisons for the IVS2 of the POMC gene and for the intragenic region between the 18S and 5.8S rRNA sequences (shown in Table 2) yielded the same divergence (26%), but the latter figure (derived from the ratio 251:972) is statistically much more significant than the former number (9:35).

RESULTS

Chromosomal localization of the two rat preproinsulin genes. The restriction maps of the previously described chromosomal DNA clones rI1 and rI2 (36) carrying, respectively, the rat preproinsulin genes I and II are shown in Fig. 1. Since the two clones did not contain overlapping restriction fragments, we did not know whether the two genes were linked. Attempts to resolve this issue by Southern analysis, using restriction enzymes with very rare recognition sequences, produced negative results. Thus, we decided to establish the chromosomal location of the two genes by cytological hybridization. For this purpose we first identified by Southern analysis (data not shown) suitable hybridization probes from the flanking regions of the two genes that did not contain repetitive sequences. The map location of the selected probes is shown in Fig. 1.

We then labeled these probes to high specific activity by nick translation with $[^{125}I]$dCTP and hybridized them to rat metaphase chromosomes obtained from either rat cell lines (BRL or E-11) or rat peripheral lymphocytes. We used these different chromosome sources to examine the consistency of our results. Results of six different experiments supported the interpretation that both preproinsulin genes reside on rat chromosome 1 (the longest chromosome of the complement, by definition) (Table 1). Other chromosomal regions with sequence homology to the probes were not detected. In most of these experiments we identified all rat chromosomes on the basis of their characteristic G-banding patterns. However, it is known that the G-banding procedure applied before hybridization can lower the hybridization efficiency (21). For this reason, we also analyzed size groups of the chromosomes hybridized to the probes without prior G-banding (see footnote *c*, Table 1). The results remained the same. Examples of the hybridization of chromosome 1 regions to preproinsulin gene I and II probes are shown in Fig. 2a and b.

If chromosome 1 is divided into six arbitrary segments of approximately equal length, it can be seen that gene I is located in segment 3 and gene II is located in segment 5, both on the long chromosome arm (Fig. 2d). We can roughly estimate the distance between the two genes as follows. Rat chromosome 1 is by length approximately 10% of the entire haploid chromosome complement (43). Assuming uniform DNA packaging for all chromosomes, this size would correspond to 0.3 pg of DNA (10% of the mammalian C-value of 3 pg per haploid genome) or 300,000 kilobases (kb). Thus, the two genes, which are separated by approximately two of the six arbitrary segments of chromosome 1, are not closely linked, since they are located at a distance of 100,000 kb from each other. From this mapping we conclude that the

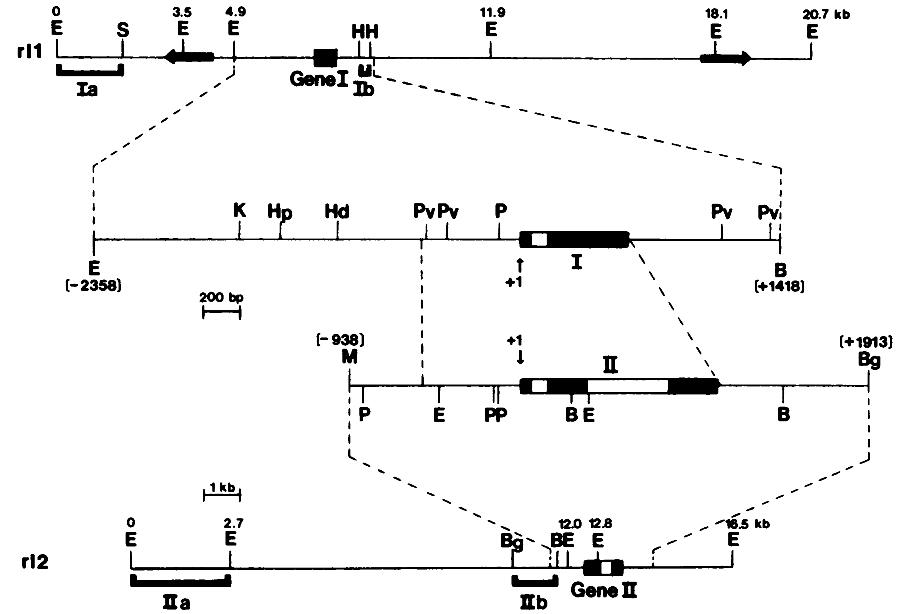

FIG. 1. Restriction maps of the inserts of clones rI1 and rI2 (36) containing the rat preproinsulin genes I and II, respectively. The 5′ to 3′ transcriptional orientation of the genes is from left to right. The distances of the *Eco*RI sites (in kb) from the leftmost *Eco*RI site (arbitrary starting point) are indicated. The relative positions of the 6.2-kb (11.9 to 18.1) and 2.6-kb (18.1 to 20.7) *Eco*RI fragments of rI1 were erroneously shown in inverted order previously (36). The thick arrows flanking gene I denote a 1.2-kb palindromic DNA element (see text). The thick brackets below the maps (marked Ia, Ib, IIa, and IIb) indicate the DNA segments that were used as probes for cytological hybridization. The blowups in the center of the diagram indicate the regions sequenced (Fig. 3). Numbering is relative to the capping site at +1. Gene regions are denoted by shaded (exon) and unshaded (intron) boxes. The dotted lines connecting the two blowups indicate the boundaries of the homologous DNA segments. Restriction sites are: B, *Bam*HI; Bg, *Bgl*II; E, *Eco*RI; H, *Hin*fI; Hd, *Hin*dIII; Hp, *Hpa*II; K, *Kpn*I; M, *Mst*II; P, *Pst*I; Pv, *Pvu*II; and S, *Sac*I.

appearance of two preproinsulin genes in rats is not the result of a simple duplication event generating linked copies (as in the case of globin genes, for example). Rather, during or after duplication, one (or both) of the genes moved to a new chromosomal location. This conclusion is consistent with the fact that the two genes reside on different chromosomes in the mouse; gene I is on mouse chromosome 6, while gene II is on chromosome 7, which also carries the β-globin gene (32).

In addition to insulin gene duplication-transposition

TABLE 1. Chromosomal localization of rat preproinsulin genes by cytological hybridization[a]

Chromosome source	Probe type	Probe size (kb)	No. of metaphase plates analyzed	χ^2 values		Expt no.
				Chromosome 1 (segment 3)	Chromosome 1 (segment 5)	
Peripheral lymphocytes	Ia	1.8	27[b]	75		1
	IIa	2.7	46[b]		70	2
E-11 cells	IIa	2.7	35[b]		127	3
BRL cells	Ib	0.45	122[b]	104		4
	IIb	1.5	24[b]		74	5a
			62[c]		185	5b
	Mixture of Ib, IIb, and cDNA	~2.3	17[b]		113	6a[d]
			27[c]	409	423	6b[d]
	β-Globin[e]	~0.54	15[b]	152		7a
			51[c]	94		7b

[a] Chromosomes were hybridized to the indicated probes either after G-banding and photography or without prior treatment. In the series of experiments with G-banding, each chromosome was identified according to its banding pattern and divided into segments (96 segments in the entire complement). Grains in each segment were counted. The χ^2 value is used as an arbitrary index of significance. It was derived in this case by comparing the observed number of grains to the expected value (assuming that grains over all of the plates were equally distributed throughout the complement). Values less than 10 were not considered significant. In the series of experiments without G-banding, chromosome 1 (the largest in the complement) was identified by size. In this case, grains were counted by dividing the chromosome complement into size groups. With the exception of segments 3 and 5 of chromosome 1, no other regions were consistently labeled (although occasionally segments of other chromosomes exhibited χ^2 values larger than 10). The probes that are specific for preproinsulin gene I and II are shown in Fig. 1.

[b] G-banding before hybridization.

[c] No prior treatment.

[d] The coding region probe used in experiment 6 was the cDNA clone pI47 (71). In experiment 6a hybridization to segment 3 of chromosome 1 did not have a significant χ^2 value, possibly because the primary component in the mixture (with respect to size) was a gene II-associated sequence.

[e] The mouse β-globin probe was the cDNA clone pCR1β (53).

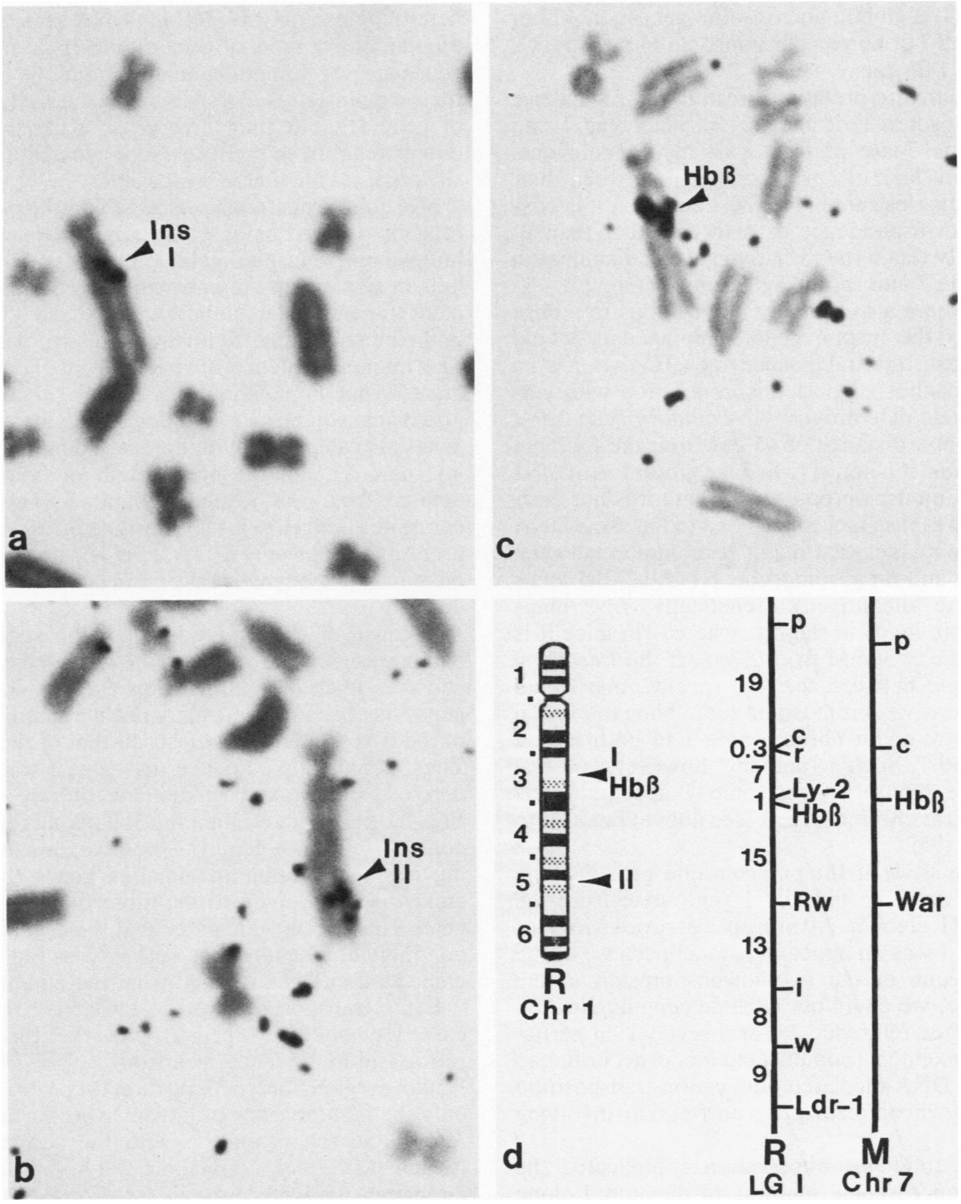

FIG. 2. Chromosome assignment of rat preproinsulin and β-globin genes. Areas of metaphase plates of rat chromosomes containing chromosome 1 are shown. (a, b, and c) Hybridization experiments 1, 2, and 7a, respectively (Table 1). (d; left) Diagram of G-banded rat chromosome 1 (R Chr 1 [17]), divided into six arbitrary segments. The approximate positions of the preproinsulin I–β-globin and preproinsulin II genes (in segments 3 and 5, respectively) are indicated by arrowheads. The percent cross-over values between the markers in rat linkage group I (R Lg I) are shown (right) and are correlated to a portion of the genetic map of mouse chromosome 7 (M Chr 7). The genetic markers are: p, pink eye; c, coat color (albino); r, ruby (or red) eye; *Ly-2* (or PtaA or AgF), peripheral T-lymphocyte antigen A; Hbβ, β-globin; Rw (rat) or War (mouse), warfarin resistance; i, ichthyosis; w, waltzing; *Ldr-1*, lactate dehydrogenase regulator.

(which obviously preceded the separation of rats and mice), the rat-to-mouse comparison also points to one or more chromosome fusion or translocation event or both that must have occurred after speciation. Our mapping data, however, cannot define the details of the chromosomal rearrangement(s).

Our results localizing the rat preproinsulin genes to chromosome 1 are among the very few assignments of genes to specific chromosomes in this species (see reference 74 and 75 for other examples). Thus, we took this opportunity to define the chromosomal location of an entire rat linkage group. In comparison with mouse genetics (18, 50), rat genetics is at a very primitive stage and only a few markers, belonging to nine linkage groups (I to IX), have been studied (49). The rat β-globin gene locus has been assigned to linkage group I. If the insulin and β-globin genes are linked, as they are on mouse chromosome 7 and human chromosome 11, then rat linkage group I is on chromosome 1. Hybridization of a mouse β-globin probe to rat metaphase chromosomes (Table 1 and Fig. 2) showed that the rat β-globin gene is indeed on chromosome 1, and (within the limits of resolution by this type of mapping) it seems to be located closer to the preproinsulin retrogene I than to the ancestral gene II. However, the lack of information about the physical distance between the β-globin and insulin genes on either mouse chromosome 7 or human chromosome 11 precludes a direct comparison with rats.

A corollary of our interpretation (positioning rat linkage group I on chromosome 1) is that rat chromosome 1 and mouse chromosome 7 are at least partially homologous, despite the apparent lack of correspondence between their characteristic banding patterns (44). This conclusion is consistent with the correspondence of markers other than β-globin which closely retain their relative genetic distances in mouse chromosome 7 and rat linkage group I (Fig. 2d).

Is rat chromosome 1 also partially homologous to mouse chromosome 6, as the insulin gene mapping data would suggest? In this case, the independent genetic data are so limited that they cannot support this conclusion with certainty. The rat lactate dehydrogenase regulatory gene *Ldr-1* has been mapped at a distance of 45 cM from the β-globin gene locus (60). Thus, it belongs to linkage group I. An *Ldr-1* locus is present on mouse chromosome 6, but it is not clear whether the rat and mouse loci correspond to the same DNA sequence. Another marker that might be common between rat linkage group I and rat chromosome 6 is *Ly-2* (the locus for a T-lymphocyte alloantigen). Genetically, *Ly-2* maps close to the β-globin locus in the rat, whereas in mice it is located at a distance of 38 cM from *Ldr-1*. If the *Ldr-1* and *Ly-2* loci correspond between the two species, one would have to invoke extensive chromosome scrambling to explain the partial homology of rat chromosome 1 to both mouse chromosomes 6 and 7. Such scrambling, however, is a real possibility because mouse chromosome 7 appears to be related not only to rat chromosome 1 (see above) but also to rat chromosome 7 (75).

DNA sequence analysis of the preproinsulin gene flanking regions. How did the preproinsulin I gene arise from the ancestral type II gene? Although retroposition of preproinsulin gene I was an appealing model because of the clean excision of one of the two introns present in the ancestral gene type, we could not exclude gene duplication at the DNA level (see reference 25 for a review). In particular, we could not exclude a complicated model according to which there was a DNA-mediated duplication-transposition event, followed by incomplete gene conversion involving cDNA (34).

In addition, the following observation complicated the picture. Electron microscopic analysis of denatured clone rI1 had indicated that gene I resides in the loop of a stem-loop palindromic structure. Since certain *Drosophila* palindromes are transposons (see reference 47 and other references therein), we could not exclude the possibility of indirect gene I transposition mediated by the flanking inverted repeats. However, sequence analysis of the palindrome (M. B. Soares, E. Schon, and A. Efstratiadis, manuscript in preparation) showed that this element consists of two members of a family of long interspersed DNA repeats (LINES), known as the Kpn family in the human genome and as the MIF-Bam-R family in the mouse genome (see reference 52 for a review). The repeat upstream of the insulin gene (in regard to the transcriptional orientation of gene I) is longer than the downstream repeat and is transcribed in the opposite direction from the insulin gene. For this reason, we excluded involvement of these repeats in gene transposition, although transcripts from such LINE elements are known to retropose (52).

To define the break points of the duplication unit, we characterized the flanking regions of both genes by DNA sequencing (Fig. 3). In exact agreement with previous heteroduplex analysis (36), the DNA sequence alignments of the flanking regions of genes I and II (Fig. 4) show that the genes are very homologous for about 500 bp upstream from their capping sites (540 bases of gene I align with 526 bases of gene II), but they diverge so extensively immediately downstream from their polyadenylic acid [poly(A)] addition sites that no alignment is possible.

Most of the upstream region of homology between the two rat genes is also homologous to the upstream sequence of the human preproinsulin gene (positions −1 to −360 of the human sequence). An internally repetitive polymorphic element (ranging in length between 100 and 3,000 bp) is present in the human sequence upstream from position −360 (3, 63). The human sequence upstream from the polymorphic element is not homologous to the rat sequence. The known upstream sequences of the dog and guinea pig preproinsulin genes also align well with the rat and human sequences (Fig. 4A, part A), but the break point of homology cannot be defined from the available data. In contrast, the DNA sequence of the 5′ flanking region of the chicken preproinsulin gene (Fig. 4A, part B) is so divergent from the mammalian sequences that no meaningful alignment can be derived.

Because of their relationships, the sequences of the 5′ flanking regions of mammalian preproinsulin genes do not allow us to draw a conclusion as to which of the two rat genes has been transposed, since it cannot be defined which of the two sequences is closer to that of the ancestral DNA. Thus, convincing evidence that gene I was transposed was derived from sequence alignments of the 3′ flanking regions. Fig. 4B, part A shows that the 3′ flanking region of gene II (in contrast to that of gene I) is homologous to the corresponding regions of other mammalian genes (human, dog, and guinea pig) and even to the more distantly related chicken gene. This strongly indicated that the original gene is gene II, not only because it retains both introns but also because it is embedded in DNA derived from the common ancestor.

Is the transposed gene I a retrogene? Had that been the case we would expect to find the three hallmarks of retroposition, absence of introns, a poly(A) tract at the 3′ flanking region, and flanking direct repeats. Of these criteria, only the first (absence of introns) was partially fulfilled. The 3′ flanking region started with the sequence ACCAAAA [which was not a convincing poly(A) tract], and flanking direct repeats were not evident. For these reasons (and to avoid the ad hoc argument that the hallmarks of retroposition had been erased during evolution), we decided to sequence the corresponding regions of the mouse preproinsulin I gene for comparison. Details of the mouse chromosomal DNA clones will be published elsewhere (B. Wentworth, L. Villa-Komaroff, and J. Chirgwin, manuscript in preparation).

Alignment of the rat and mouse gene I 3′ flanking regions showed a significant degree of homology (Fig. 4B, part B). Moreover, the beginning of the mouse 3′ flanking region (ACCAAAAAAAA) constituted a convincing poly(A) tract. Alignment of the 5′ flanking sequences indicated that the break point between the two rat genes does not correspond to the upstream end of the duplication unit, because the mouse I and the rat II sequences can be aligned for 57 to 60 additional bases further upstream, whereas a deletion of 116 bases is evident in the rat gene I sequence. After this deletion, the alignment of the mouse I and rat I genes resumes and continues to the end of the available sequence (despite the detection of additional deletions in both sequences).

RAT PREPROINSULIN GENE II

```
(-938)CCTGAGGAGAACCTCTCCACACTGCCCTGGTCTTCCCACCCTGGTGTCCCAGATACCCGGAGTGTGAGTGGCTGCAGCACTTTCTGGGGGACAAGAAGTAGGGAGCAAGGGGCT
CACAGTTCAAGTCTGGTGGCTATAAAGCCCTGCATAGGGTAGAGTTCTCGCTCATGCAACGACACCAAGGGTTTTTGCTGTCTGCTCGGGGAACAGGGCAGTACCAAATCAGGAACAGAA
AGAGTCAAGGATCCCCCAACCACTCCAAGTGGAGGCTGAGAAAGGTTTTGTAGCTGGGTAGAGTATGTACTAAGAGATGGAGACAGCTGGCTCTGAGCTCTGAAGCAAGCACCTCTTATG
GAGAGTTGCTGACCTTCAGGTGCAAATCTAAGATACTACAGGAGAATACACCATGGGCTTCAGCCCAGTTGACTCCCGAGTGGGCTATGGGTTTGTGGAAGGAGAGATAGAAGAGAAGGG
ACCTTTCTTCTTGAATTCTGCTTTCCTTCTACCTCTGAGGGTGAGCTGGGGTCTCAGCTGAGGTGAGGACACAGCTATCAGTGGGAACTGTGAAACAACAGTTCAAGGGACAAAGTTACT
AGGTCCCCCAACAACTGCAGCCTCCTGGGGAATGATGTGGAAAAATGCTCAGCCAAGGACAAAGAAGGCCTCACCCTCTCTGAGACAATGTCCCCTGCTGTGAACTGGTTCATCAGGCCA
CCCAGGAGCCCCTCTTAAGACTCTAATTACCCTAAGGCTAAGTAGAGGTGTTGTTGTCCAATGAGCACTTTCTGCAGACCTAGCACCAGGCAAGTGTTTGGAAACTGCAGCTTCAGCCCC
TCTGGCCATCTGCTGATCCACCCTTAATGGGACAAACAGCAAAGTCCAGGGGTCAGGGGGGGGGTGCTTTGGACTATAAAGCTAGTGGGGATTCAGTAACCCCCAGCCCTAAGTGACCAG
CTACAGTCGGAAACCATCAGCAAGCAGGTATGTACTCTCCAGGGTGGGCCTGGCTTCCCCAGTCAAGACTCCAGGGATTTGAGGGACGCTGTGGGCTCTTCTCTTACATGTACCTTTTGC
TAGCCTCAACCCTGACTATCTTCCAGGTCATTGTTCCAACATGGCCCTGTGGATCCGCTTCCTGCCCCTGCTGGCCCTGCTCATCCTCTGGGAGCCCCGCCCTGCCCAGGCTTTTGTCAA
ACAGCACCTTTGTGGTTCTCACTTGGTGGAAGCTCTCTACCTGGTGTGTGGGGAGCGTGGATTCTTCTACACACCCATGTCCCGCCGCGAAGTGGAGGACCCACAAGGTAAGCTCTGCTC
CTGAATTCTATCCCAAGTGCTAACTACCCTGTTTGTCTTTCACCCTTGAGACCTTGTAAATTGTGCCCTAGGTGTGGAGGGTCTCAGGCTAACCAGTGGGGGGCACATTTCTGTGGGCAG
CTAGACATATGTAAACATGGTAGCTGCCAAGAAGGAGTGAGAATCCTTCCTTAAGTCTCCTAGGTGGTGACGGGTGGCCAGGCCCCAGGATAGGTACCCATTTGGGGACCCCATAGAGCA
CCGCACCGACCGAGGGATGGTAACAGGATGTGTAGGTTTTGGAGGCCCATATGTCCATTCATGACCAGTGACTTGTCTCACAGCCATGCAACCCTTGCCTCCTGTGCTGACTTAGCAGGG
GATAAAGTGAGAGAAAGCCTGGGCTAATCGGGGGGTCGCTCGGCTCCTCCTAACTGGATTGTCCTATGTGTCTTTGCTTCTGTGCTGCTGATGCTCTGCCCTGTGCTGACATGACCTCCC
TGGCAGTGGCACAACTGGAGCTGGGTGGAGGCCCGGGGGCCGGTGACCTTCAGACCTTGGCACTGGAGGTGGCCCGGCAGAAGCGCGGCATCGTGGATCAGTGCTGCACCAGCATCTGCT
CTCTCTACCAACTGGAGAACTACTGCAACTAGGCCCACCACTACCCTGTCCACCCCTCTGCAATGAATAAAACCTTTGAAAGAGCACTACAAGTTGTGTGTACATGCGTGCATGTGCATA
TGTGGTGCGGGGGGAACATGAGTGGGGTCGGCTGGAGTGGTCGCGGCTTAATCTATCTGTCCAGCAGCAATATCCCTAAAGGGACAGCCCGAGACTTCTTTGGCTTTCTCTGAATCCCCG
GTTTTCTCCTGACTCCCTGGCACCTCTCACTGGTGCCTGTTAGACCGAACCTGAGCAAAGGCAGTTCCTTCCCCCGAGATAGCTAACGCCTTCGGTTGTCTGGGAGGACCGTGACCCCTA
CCCCTGATGCTTCTAGCCGGCTAGAGAAGAGTTAGAGGTCTTTGGAATGCATGGGTGATGGCATCTAGGAACTAATAATTCGTTGCTCTATAGCCCTGGAGGATCCTCAAGGGCCCTTAT
TACTTTTATAAGGAAGAAGACCAAAATATCCCAACCACAGCTTTCACCTAGCCCTCCTTAGTTCTTTTTCCAGAGCTATTTTAGAGGAGTATCTGACTGGGAAGAAATTGGGCTTGGTAC
CTTGAGCTGGAAGGCCATGGAGTCATTCTTAAAGAGCTTATCCNAGCTCTGGGAGGACAGAGAACAGCCCCAACTGCTCTCAGACTATCCAATGACCTTTAGTGCTCTGAACAGGACAGG
CACCCCACACAGGTGAATAACATACTGAAGGACTGGGCAGGAGCAGAACTCCACTTCTCAGGAATGCCAGTTGCAAGTTCTAAGCAAGGTAGCCATGAGAGAGGTCCAGGGCTGGGGTCT
CCTATTACTTCTCAAGTCCCTCCTCCCAAGACAGGTAGGGCCTCTCATCAAGATTTTCTTGAAACTTAAACTGAAAAGCCACATAACATCTAAGATCT(+1913)
```

RAT PREPROINSULIN GENE I

```
(-2358)GAATTCCTATTCTTTCCAGGACTTTTATCATGAAGGGGTGTTAAAACACGATAATTTAATGGAGATGTTTATGTTTGGGAGGACCTTGAGAGAAAGGCACTGGGTCCAAGTCA
TTATAGAATCATAGTCTAAGCCAAAAATACTCATATATGTTCAGTTTCACTACCCTCCTTCTCCTCGAGACGTGTGACCTAAGTAGCTGTCTGTATACGGTAATATACTTGCCAGTTCCT
GAGATAAGCAGGCGCACAGTGCCCCATTCCGAAATAAGTGTGCAGCCAGTGTTCTAGCTTTAGTCTCCGATGGAATCCCCTGGTTCAAGTATGATCTCATGTTGTCGCATCCCATTTCCT
TCCAGGGAGCTCCCTCAGGAGAACGACTGCTCACCCGCGACCTTTCCACAGGATTGAAGCTCAGGGAAGGAAACAATATTTGAAGTTTCAGTCCAGTGGTGCATATGAACTTTAGAATCA
CTCCACCACTATCTACAACGACAAAAGTCACATGGCTAAAGAAGAGTGTCAGAAAAGAGCTCCTTAACAAATAAAATTCAATTCAATTCAATTCTTGAGCAGTCCTAGTCTCGTCCATAA
GCCTTCTCTCTGGTCCCTCTAATTCTCTGACTCACTCTCTGACCAGTGTAGCCAGTAAAGGGGACATATTCTCTAAATAAATGTTTTCTTCCAACTTGAAACCTAATGTTTACAAGCCTT
TATATTACAGTACCTTGGTTTCTCCACTCTGATGAAATACAAACATCTTACCTATCATTCAGTCATGGGTTTTCTTTCTTTGTTCTCAGCTCTCTGGGTACCTGTCATGGACTGCCACAT
GATGATTGGCTGAGCAGTGGTGAAGACACTAACCTCACTCGGAATCATAAAGTAGATTTCTAATTAAGAAAAAGGAGTGCCACCCATTGTTCAAAGGAACACTGTGTTAATTTAGTTAGA
GTGGAGGGTGCAAATTCTAAGAGGAAAGCTAATTTATAAACACGGTGCCAGAACCTGTTTCCCAAAGTTCTGTTGTGCCGGGCATGAACCACACATTTGAGGTTGTAGTCAATGGCTGAG
AACAGGAGAGATCAAGCACCCCCAGAAAGGTCCTCTTTTTGAACAAAGATTATCAGTGCATCCAGTTAAATAAGAAAGTTGACAGTGGTTGAGAGAAGCAACCAGGTGCTTGCCAGCCCC
ACATAACTCCAGGCTAGCAGCAAAGATGACCTTACTCTCCCGATACAAGCTGGGTCACTTACAGTCATATCTTCGCTATTTTCCTCTACTCATACCCTTTAGAAAACTGCCAAATTGATA
TGTTTTCCATACCTATAATCAGACTAAGAAAAAGAGGCAGAAACAAGTTTTAAAGCTTTCATGGATGGCACTGGAGAAGTTAAATTTTTTTGCTGTATAACCTATGTTTATTTTATACCA
TAAGGACTTCACAAACTAACTTAACGTTATTGGACCAGTACAAACCTACTTTCTTTAATGGCTTCTGAGTTACTTGAAGACAAGTAACGTGTCTACTTTATTATGGCGCCTACATTGGAA
CCCACTGCACATCACCAGAGATTTTTCTCAGCAAATGTTTGTTGACTTAATTGAATGCTGTCATTGACAGAAAGTCTGGTTTGCATTTCCTCTCAACTCCTTGAAAATAGCTACCTTTCC
TAATTGACCCTGTTGGCAGATGTTTAAACTGGGGTGAACGCTGTGCTACTGAGGCCTGATGCTTTGTGGAAAATCAATTAGAATGAGACCAAATGTTCCACCTAATTCAAATGGTTGTCA
AAAAATAGAATTTGAGTATCTATATTTCGTCCCAGCTGACCCCTGAGTGGGCTATGGGTTTGTGGAAGTAGAGATAGAGGAGAAGGGACCATTACATGTCCTGCTGCCTGAGTTCTGCTT
TCCTTCTCCCTTTGAAGGTGAGCTGGGGTCTCAGCTGAGCTAAGAATCCAGCTATCAATAGAAACTATGAAACAGTTCCAGGGACAAAGATACCAGGTCCCCAACAACTGCAACTTTCTG
GGAAATGAGGTGGAAAATGCTCAGCCAAGGAAAAAGAGGGCCTTACCCTCTCTGGGACAATGATTGTGCTGTGAACTGCTTCATCAGGCCATCTGGCCCCTTGTTAATAATCTAATTACC
CTAGGTCTAAGTAGAGTTGTTGACGTCCAATGAGCGCTTTCTGCAGACTTAGCACTAGGCAAGTGTTTGGAAATTACAGCTTCAGCCCCTCTCGCCATCTGCCTACCTACCCCTCCTAGA
GCCCTTAATGGGCCAAACGGCAAAGTCCAGGGGGCAGAGAGGAGGTGCTTTGGACTATAAAGCTAGTGGAGACCCAGTAACTCCCAACCCTAAGTGACCAGCTACAATCATAGACCATCA
GCAAGCAGGTATGTACTCTCCTGGGTGAGCCCGGTTCCCCCAGCCAAAACTCTAGGGACTTTAGGAAGGATGTGGGTTCCTCTCTTACATGGACCTTTTCCTAGCCTCAACCCTGCCTAT
CTTCCAGGTCATTGTTCCAACATGGCCCTGTGGATGCGCTTCCTGCCCCTGCTGGCCCTGCTCGTCCTCTGGGAGCCCAAGCCTGCCCAGGCTTTTGTCAAACAGCACCTTTGTGGTCCT
CACCTGGTGGAGGCTCTGTACCTGGTGTGTGGGGAACGTGGTTTCTTCTACACACCCAAGTCCCGTCGTGAAGTGGAGGACCCGCAAGTGCCACAACTGGAGCTGGGTGGAGGCCCGGAG
GCCGGGGATCTTCAGACCTTGGCACTGGAGGTTGCCCGGCAGAAGCGTGGCATTGTGGATCAGTGCTGCACCAGCATCTGCTCCCTCTACCAACTGGAGAACTACTGCAACTGAGTCCAC
CACTCCCCGCCCACCCCTCTGCAATGAATAAAGCCTTTGAATGAGCACCAAAATGAGAGAGTTTTTATGAATACAAAGGGATTGTGTGAACGGGAATCTTTTTCTCTGTCATTTAGTATC
GTGCTAGCGTATTACTAAGCAGTTGTTAAAACTGCATGATTGTGTAACCATTTAAGAAGCTCATGATAAAACAGACATAATTCAAAGTATCCAGAATTTGGATAATAAGAGAGTCAAAGG
GTAAATGCCACTGAAGTTACTAGTCCAAAAGTGATTTGCTTCTGTCTTGGCTCACCCCCCTGGTACTATTTAGCTCACAATTCCAGGTTACAGTTCATCCACAGAAGGGGAGTGGAATAG
AGACGGATAAATAAATGTACCCATACATCAACCTATCTAGATAGTGTCCCACTGAGAATCCCCTTCCTAGATTATCCTAGACTATCAAGTAGACAGTTAAAACATCACAATCACTTTGCA
CATCCCACACCATCCTGCAATAATTCATTCACTTTCTCATATACATCAGAGAGATCACTCCTTCCACAGATAACTACAGTGGTAAAGGTCAGTGTGAGTTCCTATACAGCTGTGGAGTGA
TACAGTCTGTGCCCTTAACACTTTGGTTCCAGACAGAGGAAGGAGGAATAAACACGAGAACACAGGTATGCGGTTTGTAAACCAGTACACGAATGTTTAGATTTTTAAAAAGATTTTTGT
GGGCCAACTAAGGAGTGAGTGTCTGTTGCCAGAAAATGTCTTTCCAACCTTGGTGTGAACACTGACCCATCAACACTAGAATACAAGGTTCTGACTTCCAGGCTTTCTAAGTCAGAACTC
CTGTTGTCACAGCTGTAGCTATGAATCTAACAAAAGCTCTTCAGTCTTCGTATGAATAGGATCC(+1418)
```

FIG. 3. DNA sequence of the rat preproinsulin I and II genes and their flanking regions. The sequence of the mRNA-like strand of the gene is presented from the 5′ to the 3′ direction. Numbering is relative to the capping site at +1. Intron regions are in small capital letters. Dots at the top of each sequence are spaced at 10-bp intervals. Gene landmarks [ATA box, ATG initiator, termination codons, and poly(A) addition signal] are underlined. The poly(A) addition site is marked by an asterisk. The arrowheads denote the breakpoint of upstream homology between the two rat genes. However, there is a deletion in gene I at this point (see Fig. 4). Thus, the true homology break (comparison between rat gene II and mouse gene I, Fig. 4) is at the position of the dot. The direct repeat flanking gene I (target site duplication during retroposition) is boxed. Dots underline a putative upstream ATA box in gene II, dotted arrows underline the location of potential ATA and CCAAT boxes located on the bottom strand in gene I. Nucleotides underlined in the gene II sequence are corrections of the sequence previously presented (36).

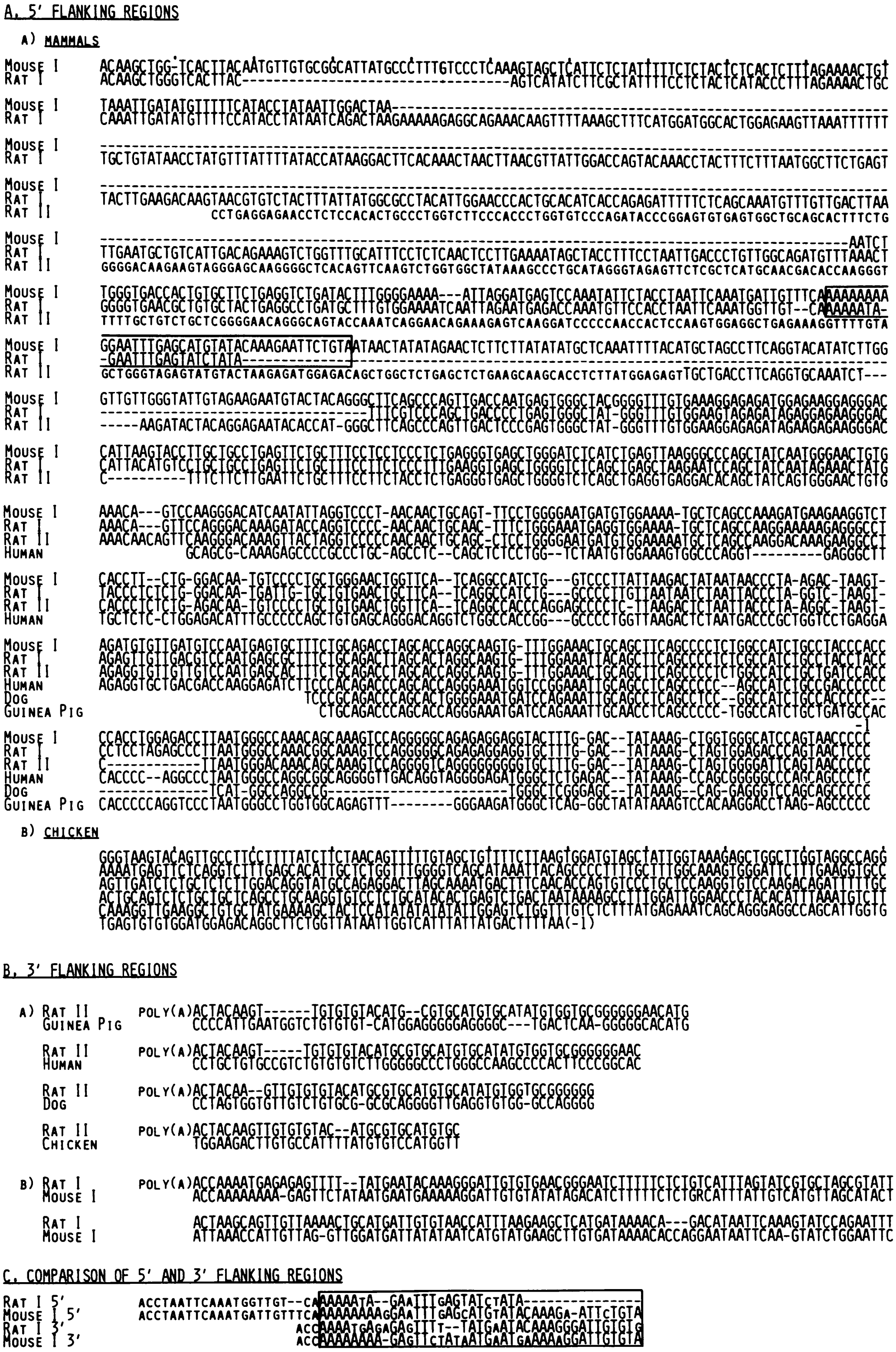

FIG. 4. DNA sequence alignments of preproinsulin gene flanking sequences. (A) Alignment of the 5′ flanking regions. In Fig. 4A, part A the sequences of the indicated mammalian species extend leftward from the first base before the capping site at −1. Gaps were introduced to maximize homology. The lower case letters in the rat gene II sequence denote divergence from the mouse gene I sequence. The apparent break in homology with the rat I gene sequence at gene II position −527 is due to the presence of a 116-bp deletion in rat DNA relative to mouse DNA. (Fig. 4A, part B) The 5′ flanking region of the chicken preproinsulin gene, which cannot be aligned to the mammalian sequences. (B) Alignment of the rat gene II 3′ flanking sequence to the corresponding sequences of other mammalian preproinsulin genes (part A). The sequences begin with the first nucleotide following the poly(A) addition site and extend downstream. Because of sequence divergence it is difficult to align these sequences as a group, but homology is evident in pairwise comparisons. A similar alignment in Fig. 4B, part B compares the rat and mouse gene I sequences. (C) Comparison of rat and mouse gene I 5′ and 3′ flanking region sequences at the breakpoints of homology between the mouse I and rat II genes. The 5′ sequences begin just before the boxed region in Fig. 3 and 4A, and the 3′ sequences begin at the first base of the 3′ flanking region. Mismatches are indicated by small capital letters. Dashes indicate gaps introduced to maximize homology. The duplicated sequence of the target site is boxed.

A sequence of 41 bases present in the mouse gene I at a distance of 42 bases upstream from the mouse I-rat II break point is repeated in both the mouse I and rat I 3' flanking region. The remnant of the poly(A) stretch can be considered as part of the direct repeat, as in other cases of retroposition (see, for example, references 8, 37, and 61), which might have implications for the formulation of certain retroposition models (see references 19 and 72). Because of the 116-base deletion in the rat I sequence only a portion of the upstream member of the direct repeat has survived. Despite this and despite some degree of divergence a convincing alignment of the mouse and rat direct repeats can be derived (Fig. 4C). Except for its length (which is approximately twice as long as that in other retroposons), the direct repeat seems typical in the sense that its 5' side is A-rich (42).

The origin of the 42 bases which are present in the mouse gene I sequence between the upstream member of the direct repeat and the mouse I-rat II break point is unknown. A similar situation (presence of 15 or 16 bases of unknown origin in an equivalent position) has been described in two human 7SL RNA retropseudogenes (65). We note that the sequence of unknown origin seems totally unrelated to the ancestral DNA in the region, because the rat II sequence upstream from the break point can be aligned with the mouse gene II sequence without rearrangements or deletions (B. Wentworth and L. Villa-Kamaroff, unpublished data). From these data we conclude that all three of the retroposition criteria are fulfilled and that the rat and mouse preproinsulin gene I is a functional, semiprocessed retrogene.

Retroposition of upstream transcripts. Because of the sequence relationships between the original insulin gene and its retroposed counterpart we concluded that the reinserted genetic information was derived from a transcript initiating upstream from the normal capping site. However, examination of the DNA sequence of gene II upstream from the break point of strong homology did not reveal the presence of an ATA box around position −30 from this point. Nevertheless, we did identify an ATA box at position −804 (approximately 220 bases upstream from the break point), which is homologous to the bona fide ATA box of the gene. Assuming that 5' truncation might occur during retroposition, this ATA box could have been part of the promoter for the expression of the upstream transcript. The performance of this upstream ATA box of gene II was examined both in vitro and in vivo. In parallel, we examined the performance in vitro of a putative ATA box present on the antimessage strand (in regard to insulin gene transcription) in the upstream nonhomologous region of gene I (position −930).

We first examined by S1 nuclease analysis, using single-stranded, end-labeled probes, the initiation sites of RNA synthesized in an in vitro transcription system (38) from truncated double-stranded DNA templates. Figure 5 (bottom) shows the templates and probes used for each gene. For gene II we used the following templates: N (containing only the natural ATA box), U (containing only the upstream ATA box), and B (containing both boxes). Probes n and u were used to assay transcription from the natural or upstream box, respectively. RNA transcribed from templates N and B protected 190 nucleotides (nt) of probe n, exactly as expected for RNA initiating at the natural capping site (Fig. 5, part II, lanes 2 and 3). RNA from templates U and B protected 80 to 85 nt of probe u (Fig. 5, part II, lanes 5 and 6). Because of the position of the label in probe u, the start sites of the protecting transcripts are 28 to 33 nt downstream from the second ATA box, which clearly functions in vitro. The controls (Fig. 5, part II, lanes 1 and 4) showed that RNA from template U did not protect probe n, and RNA from template N did not protect probe u. The larger-than-expected fragments (Fig. 5, part II, lanes 2, 3, 5, and 6, dots) correspond to probe protection by end-to-end transcripts from the in vitro templates. The origin of the smaller-than-expected fragments (Fig. 5, part II, lanes 2 and 3) is unknown.

What are the relative strengths of the natural and upstream promoters in vitro? Since an equivalent amount of RNA from template B was used for protection of probes n and u (Fig. 5, part II, lanes 3 and 6), the ratio of signal in the 190-nt and 80- to 85-nt fragments will yield an estimate of the relative promoter strengths, provided that a correction is applied for the differences in specific activities of the two probes. For such a normalization we used as internal standards the 300- and 190-nt fragments (dots in lanes 3 and 6, respectively) protected by end-to-end transcripts (assuming that this type of transcription is not differential). Thus, we first calculated (from densitometric analysis) the ratios 190:300 nt (lane 3) and 80 to 85:190 nt (lane 6) and then divided these two numbers to normalize the data. This final ratio (4:1) suggests that the natural promoter is four times stronger in vitro than the upstream one.

For gene I (by analogy to the gene II experiments) we used templates N', U', and B', and probes n' and u' (Fig. 5, part I, bottom). RNA initiating at the natural capping site of gene I should protect 125 nt of probe n', which is exactly the size of the doublet of fragments protected by RNA synthesized from templates N' and B' (Fig. 5, part I, lanes 2 and 3). RNA initiating approximately 30 nt downstream from the second (inverted) putative ATA box of gene I should protect about 200 nt of probe u'. However, RNA from templates U and B (Fig. 5, part I, lanes 5 and 6) did not protect such a fragment, which means either that the second ATA box is not functional in vitro or that the transcripts are very rare and below the sensitivity limits of this assay. Probe protection by end-to-end transcripts was again observed (Fig. 5, part I, dots).

Does the upstream promoter of gene II function also in vivo? We attempted to answer this question by Northern analysis and S1 protection experiments with upstream probes and RNA from an insulin-expressing cell line, derived from a transplantable rat insulinoma (38 cells; see reference 15). The results were negative. From the amounts of polyadenylated and nonpolyadenylated RNA we used (30 μg each), the results of parallel experiments probing mature insulin mRNA, and the known specific activities of the probes (approximately 10^9 cpm/μg), we calculate that if any upstream transcripts are present in this cell line, their concentration is at least three orders of magnitude lower than that of the transcripts initiating at the natural capping site. This is a dramatic difference from the results of the in vitro experiments, which simply assay for the recognition of a DNA signal by the general transcriptional machinery, in the absence of tissue-specific controls. Nevertheless, we consider the in vitro results (suggesting that the upstream ATA box has potential for function) as consistent with the hypothesis that an upstream promoter of this kind was responsible for the (possibly aberrant) transcriptional event leading to gene I retroposition, which must have occurred in the germ line of an ancestral murid. We note that upstream ATA boxes present in the 5' flanking regions of the ovomucoid gene (31) and the ovalbumin-like X gene (20) produce rare transcripts in vivo. Alternatively, promiscuous transcription in the germ line, leading to the generation of

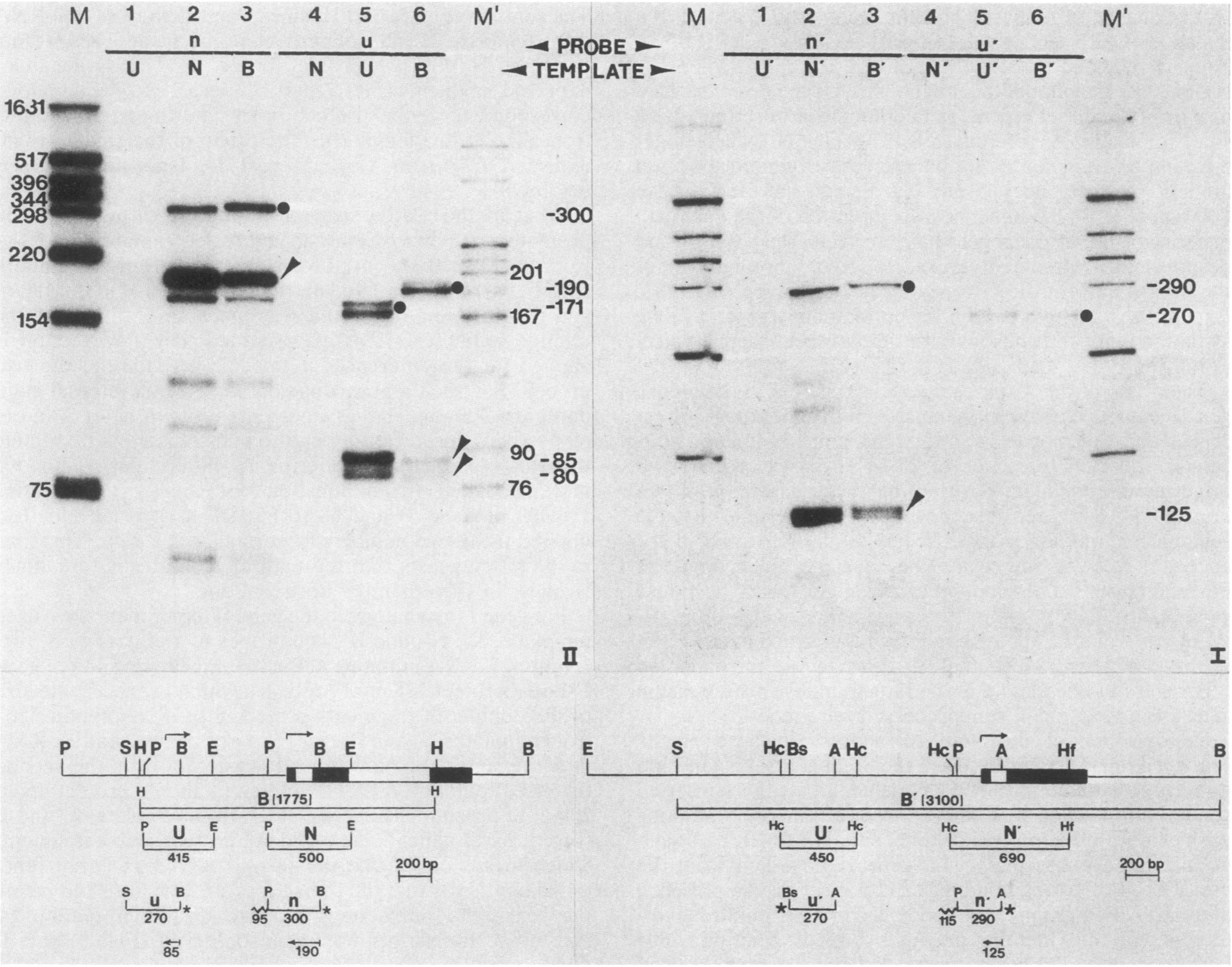

FIG. 5. S1 nuclease mapping of transcripts synthesized in vitro from rat gene II and I templates (left and right autoradiogram, respectively, of a 6% denaturing urea-polyacrylamide gel). Maps of the gene regions and also the templates used for in vitro transcription, the probes used for the analysis, and the sizes (in nt) of the protected fragments are shown at the bottom of the figure. Introns and exons are indicated by shaded and open boxes, respectively. The templates and probes (described in the text) are denoted by capital and small letters, respectively. Asterisks indicate the position of the 5' end label. Numbers below or next to the templates denote fragment sizes in bp. The wavy lines on probes n and n' denote 95 and 115 bp of vector sequences, respectively. Arrows above the maps indicate positions of transcription initiation in vitro. Arrows and numbers below each probe denote the length of the S1-protected fragments for that probe. S1-protected fragments are indicated by arrowheads on each autoradiogram. End-to-end transcripts are indicated by dots. Sizes of protected fragments are indicated at the right of each autoradiogram. The size markers (lanes M and M') are the *Hinf* or *Hpa*II fragments of pBR322, respectively. Restriction sites are as in Fig. 1 except for: A, *Ava*II; Bs, *Bst*NI; H, *Hpa*II; Hc, *Hinc*II; Hf, *Hinf*I.

retroposons, might initiate downstream from sequences resembling the ATA$^-$ promoters of housekeeping genes (see reference 66 and other references therein). In other cases, upstream transcripts may be RNA polymerase III products (6). Our data cannot exclude the possibility that the insulin gene retroposition was the consequence of polymerase III action. However, polymerase III internal promoter sequences (10) are not recognizable at the beginning of the homologous regions between gene II and the rat and mouse retroposons.

The rat (mouse) preproinsulin I gene is not the only retroposon derived from an upstream transcript. Some of the rat cytochrome *c* retropseudogenes also belong to this category (55). We note that an analogous retroposon is the mouse α-globin processed pseudogene (45, 69). This retropseudogene has been considered as atypical because of the apparent absence of poly(A) and direct repeats. However, this conclusion was based on a misalignment of the 5' flanking regions of the pseudogene and the wild-type sequences (45, 69). When the sequences are properly aligned (68) to the upstream end of the available sequence, it becomes evident that DNA sequencing upstream from position −350 of the pseudogene and the wild-type gene is necessary to identify a sequence break point that would locate the putative upstream member of the direct repeat. A candidate for a short poly(A) remnant is present at position +19 of the 3' flanking region.

DISCUSSION

The most interesting conclusion from our data is that the duplication-transposition of a functional preproinsulin gene is a case of nonviral retroposition. The virus-type retroposons (which include not only retroviruses but also *Drosophila* elements, such as copia, yeast Tyl elements, etc.) encode their own, known or presumed, reverse transcriptases, and presumed integrases. The nonviral retroposons, such as Alu repeats and cDNA-like pseudogenes (or retropseudogenes) including those of small nuclear RNAs, seem to transpose in a passive way. It has been postulated that a cDNA copy of a transcript is inserted into the site of a staggered chromosomal break (for models see reference 67 and also references 24 and 42). However, other models postulate hybridization of a transcript to the chromosomal break, which precedes copying of RNA into DNA (19, 72).

Why are retropseudogenes fixed in the population? The most likely explanation is that their fixation is due to random drift. Since the cDNA-like copies do not carry regulatory sequences with them, they are inactive from the beginning and they accumulate nucleotide substitutions rapidly, because negative selection is not in operation. Occasionally, a retrogene might become functional by association with foreign regulatory sequences (either preexisting at the site of insertion or more likely generated by mutation). Such a case might be represented by an intronless chicken gene, encoding a calmodulin-like polypeptide, which appears to be functional in vivo (59) (although other hallmarks of retroposition have not yet been documented). In contrast, the insulin I retroposon was active from the beginning because it carried its own promoter with it.

Is the fixation of the functional retroposon we describe a neutral event? This is plausible, since an organism can easily survive with only one insulin gene (even if its product is of diminished biological activity, as in the guinea pig). However, the fact remains that both genes are maintained as functional entities in two different species, although enough time was available for one of the copies to drift (it is almost certain that the two genes were never corrected against each other). Thus, we believe that the two genes provide a selective advantage for rats and mice. What this advantage might be is unclear, because specific roles must be assigned to both genes or their products. Although the new gene (I) might respond better to certain stimuli (27), there is no evidence to suggest a reason for the preservation of the original gene. Nevertheless, sequence divergence analysis (Table 2) strongly supports our interpretation that the two-copy state is positively selected.

Positive (Darwinian) natural selection is the differential reproduction of variants with increased fitness. If a new gene appearing in the population enhances the fitness of the individual that carries it its frequency in the population will increase, and the gene will eventually be fixed (more rapidly than by random drift). The same will happen with preexisting genes that are neutral or even slightly deleterious if they become advantageous by a change in the environment. In contrast, negative selection is the elimination from the population of functionally deleterious mutant genes which reduce the fitness of the individuals that carry them (see reference 29 for an extensive discussion). In pairwise sequence comparisons, the operation of negative selection is revealed indirectly when sequence preservation is documented in genes that did not diverge recently.

As we have discussed previously (7), replacement substitutions not only in the C-peptide region but also in the signal peptide sequence of the preproinsulin gene are neutral. Five such substitutions (9%), leading to five neutral (i.e., not harmful) amino acid replacements, can be seen in the pairwise comparison of the preregion of the mouse I and rat I genes. The corresponding silent substitutions are 15%. For the C-peptide region, the percent substitution values for replacement and silent sites are 4 and 21%, respectively. In contrast, there is a complete absence of replacement substitutions in the sequence encoding the B and A insulin chains, whereas some silent substitutions (11%) do appear. Only the operation of strong negative selection can account for this differential preservation of sequence in the gene regions encoding the mature hormone relative to the segments encoding the parts that are eliminated during proteolytic processing. Moreover, this differential negative selection, operating independently for millions of years on the same gene in two different species, strongly suggests that the two-copy state was originally fixed by positive selection. The equivalent argument that the first fixation was neutral, followed by positive selection of both copies because of a shift in the environment, is less likely because some minimal divergence in the replacement sites of the B and A chains should have appeared.

The degree of negative selection maintaining the two rat-mouse insulins can be appreciated further from the fact that the genes are embedded in DNA which, as we show below, is evolving neutrally.

To examine neutrality, we decided to compare any known gene sequences in rat, mouse, and human (or any combination of these three species). We also decided to focus primarily on introns and flanking sequences, because by this approach we could test one of the strongest predictions of the neutral theory (see below).

The only gene-associated DNA segment that can be considered as functionless a priori is the 3′ flanking region, probably excluding the segment in the immediate vicinity of the poly(A) addition site, which might be involved in processing (40). Introns (at least for most of the length of their primary structure) can also be considered as functionless, since the length of the same intron can vary significantly between organisms. Thus, if these gene-associated segments (3′ flanks and introns) are evolving neutrally, their divergence should be similar regardless of gene type. Such a prediction cannot be made for coding regions of different genes, because of differential negative selection possibly constraining even silent sites.

The fulfillment of the prediction (Table 2) constitutes clear-cut evidence in support of the neutral mutation-random drift hypothesis (28, 29), within the limitations of the analysis (see above for details) and for these particular DNA regions.

The analysis indicates that in the rat (or mouse) to human comparisons (85 million years [MY] of divergence) the 3′ flanking regions (average of 37 ± 3% substitutions in six different genes) are diverging at the same rate as the introns (average of 38 ± 4% substitutions calculated from ten introns belonging to six different genes). Most interestingly, the percent substitution is the same in the 5′ flanking region distal to the gene (average of 37 ± 3% in four genes). Thus, this latter gene-associated segment (for which a prediction of neutrality could not be made a priori) is also evolving neutrally. Surprisingly, the segment of the 5′ flanking region which is proximal to the gene (and which is expected to be under negative selection because of promoter functions) is not significantly constrained (with one exception among six examples). Exactly the same conclusions can be drawn from the rat to mouse comparisons (including comparisons of preproinsulin genes), although in this case the data are even more limited. Nevertheless, a neutral divergence value of 19 ± 4% can be assigned from the numbers derived from four

TABLE 2. Percent substitution values[a]

Gene[b]	Comparisons			
		R-to-M comparisons		
		Coding regions		
		Preregion	*C-peptide*	*B + A chains*
Insulin				
RI/MI				
Replacement sites		9 (5/53)	4 (3/70)	0
Silent sites		15 (4/26)	21 (7/34)	11 (6/54)
		Noncoding regions		
	5′ Flanking		*Introns*	*3′ Flanking*
	Distal	*Proximal*		
Insulin				
RI/MI	19 (95/511)	11 (29/257)	IVS1 14 (16/115)	26 (46/175)
RI/RII	16 (45/279)	14 (35/247)	IVS1 16 (18/115)	
MI/RII	17 (58/334)	11 (28/249)	IVS1 15 (17/115)	
POMC	17 (15/89)	17 (12/70)	IVS2 26 (9/35)	
Serum albumin		8 (7/90)		22 (20/89)
Cytochrome *c*			IVS2 19 (67/361)	10 (63/611)
18S-5.8S rRNA intergenic region			26 (251/972)	
		R or M-to-H comparisons		
		Coding regions		
Insulin		*Preregion*	*C-Peptide*	*B + A chains*
RII/H				
Replacement sites		17 (9/52)	14 (10/73)	3 (4/131)
Silent sites		27 (7/26)	38 (14/37)	31 (17/54)
		Noncoding regions		
	5′ Flanking		*Introns*	*3′ Flanking*
	Distal	*Proximal*		
Insulin RII/H	38 (41/107)	30 (76/252)	IVS1 38 (46/120)	
POMC				
M/H		34 (31/90)	IVS2 41 (29/71)	34 (21/61)
R/H		38 (23/61)	IVS2 46 (53/115)	
Immunoglobulin kappa-chain, M/H				37 (125/334)
Growth hormone, R/H	37 (64/173)	21 (29/140)	IVS1 36 (71/195) IVS2 41 (81/196) IVS3 32 (31/97) IVS4 38 (79/207)	36 (16/44)
Parathyroid hormone, R/H	32 (56/175)	27 (28/103)	IVS1 40 (103/257)	41 (32/78)
α-Globin, M/H		38 (30/78)	IVS2 37 (47/126)	39 (46/119)
β-Globin, M/H	40 (93/230)	28 (28/100)	IVS1 32 (37/117)	34 (50/149)

[a] Numbers indicate percent substitutions calculated from pairwise comparisons, without correction for back mutations (see text). The numbers in parentheses are the ratios of the counted nucleotide substitutions to the total number of sites in each case. In the calculation of an average divergence from the rat-to-mouse comparisons, the 10% substitution value (3′ flanking region of the cytochrome *c* gene) was excluded, because this gene-associated segment is presumably constrained (multiple polyadenylation sites, generating different lengths of 3′ noncoding regions in different transcripts [35]).

[b] R, Rat; M, mouse; H, human.

[c] The division of the 5′ flanking region of insulin genes into distal and proximal segments was based on the position of the border of a putative tissue-specific enhancer element, defined by in vivo assays of deletion mutant templates (73). A similar border was assigned for globin genes from the results of transcription studies using deletion mutants (14, 41). For the rest of the genes, the border was positioned arbitrarily (point of an abrupt change in the density of substitutions in the alignment). Absence of a calculated value from a column is due to either unavailability of data or to inability to derive a convincing alignment.

different genes, considering altogether introns, 3′ flanks, and distal 5′ flanks.

From the value of 38% for the rat (or mouse)-human divergence, corrected to 53% for back mutations, we can assign a neutral unit evolutionary period of 1.6. (The unit evolutionary period is the time in MY required for the fixation of 1% divergence between two initially identical sequences.) Accordingly, the retroposition event occurred approximately 35 MY ago (1.6 × 22%, 22% being the corrected value for 19% divergence).

How frequent are the fixations of retroposons? A rough estimate can be made using as an example the human Alu repetitive family (26). Alu repeats are almost certainly retropseudogenes of 7SL RNA transcripts (5, 64). An equivalent family exists in rodents. However, rodents have a second such family (type 2 Alu), which is not present in humans, and rabbits have exclusively a third family of repetitive retroposons (C repeats) that are absent in other mammals (9). Thus, we can assume that the independent formation of these families began around the time of the mammalian radiation, 85 MY ago. We can further assume that such families are expanding (or at least expanded initially) in a geometric fashion, because of the presence of internal polymerase III promoters in the retroposons (24). Since the presumed geometric progression has a common ratio of 2 and it took 85 MY to establish from the initial gene a family of approximately 300,000 to 500,000 members in humans or 170,000 members in rabbits, we estimate that about 17 to 19 rounds of fixation should have occurred or about one fixation event every 5 MY. Thus crude (but indicative) estimate will not change substantially even if we assume that only half of the members of each family are transcriptionally active.

Although retropseudogenes lacking internal promoters differ from the Alu-type repeats because they are frozen after retroposition, their rate of fixation should have been the same, because any neutral retroposon has the same probability to be fixed in the population by random drift as any other (assuming, of course, that the amounts of the two types of transcripts in the germ line are not vastly different). This rate is in excellent agreement with the observed frequency of retroposons (12, 33, 35) in certain mammalian gene families (15 to 30 retropseudogenes in the tubulin, actin, and cytochrome *c* families). However, the rate of appearance of functional retroposons, like the rat-mouse preproinsulin I gene, is expected to be lower because of the rarity of upstream transcripts. The frequency of appearance of cDNA retroposons that become functional by association with promoters (as the putative chicken calmodulin-like retrogene possibly did) cannot be assessed. In any event, insulin gene I is unlikely to be the only example of functional retroposition. In this regard we think that the intronless *Chironomus* globin gene (1) is another good candidate. In addition, the 17 to 20 intronless *Dictyostelium* actin genes (P. Romans and R. A. Firtel, J. Mol. Biol., in press), which (with one exception) are functional, might also be retrogenes. Unfortunately, the extreme A-richness of their flanking regions precludes a firm conclusion based on sequence comparisons. Nevertheless, the possibility that at least some actin genes are functional retroposons could explain the appearance of different (occasionally overlapping) subsets of actin gene introns in different species. This gene family (which in mammals includes retropseudogenes) might have been formed by independent retropositions of differentially and partially spliced transcripts.

ACKNOWLEDGMENTS

We thank B. Wentworth, L. Villa-Komaroff, J. Roberts, R. W. Scott, S. Tilghman, R. Firtel, and A. Furano for making available unpublished data, P. Fisher and M. Chao for materials, S. Ripley, S. Cotter, and L. Mills for help with some of the experiments, and one of our reviewers for several constructive comments for improvement of the manuscript.

This work was supported by Public Health Service grants to A.E. and J.C. from the National Institutes of Health, by a grant to A.H. from the National Science Foundation and by a gift to the laboratory of A.E. from the Bristol-Myers Co. A.H. is a Scholar of the Leukemia Society of America. M.B.S. was supported by a fellowship from the Brazilian Conselho Nacional do Desenvolvimento Cientifico e Tecnologico (CNPQ).

LITERATURE CITED

1. **Antoine, M., and J. Niessing.** 1984. Intron-less globin genes in the insect *Chironomus thummi thummi*. Nature (London) **310**:795–798.
2. **Bell, G. I., R. L. Pictet, W. J. Rutter, B. Cordell, E. Tischer, and E. M. Goodman.** 1980. Sequence of the human insulin gene. Nature (London) **284**:26–32.
3. **Bell, G. I., M. J. Selby, and W. J. Rutter.** 1982. The highly polymorphic region near the human insulin gene is composed of simple tandemly repeating sequences. Nature (London) **295**: 31–35.
4. **Brosius, J., R. Cate, and A. P. Perlmutter.** 1982. Precise location of the two promoters for the β-lactamase gene of pBR322. J. Biol. Chem. **257**:9205–9210.
5. **Brown, A. L.** 1984. On the origin of the Alu family of repeated sequences. Nature (London) **312**:106.
6. **Carlson, D. P., and J. Ross.** 1984. α-Amanitin-insensitive transcription of mouse β^{major}-globin 5′-flanking and structural gene sequences correlates with mRNA expression. Proc. Natl. Acad. Sci. U.S.A. **81**:7782–7786.
7. **Chan, S. J., V. Episkopou, S. Zeitlin, S. K. Karathanasis, A. MacKrell, D. F. Steiner, and A. Efstratiadis.** 1984. Guinea pig preproinsulin gene: an evolutionary compromise? Proc. Natl. Acad. Sci. U.S.A. **81**:5046–5050.
8. **Chen, M.-J., T. Shimada, A. Moulton-Davis, M. Harrison, and A. W. Nienhuis.** 1982. Intronless human dihydrofolate reductase genes are derived from processed RNA molecules. Proc. Natl. Acad. Sci. U.S.A. **79**:7435–7439.
9. **Cheng, J.-F., R. Printz, T. Callaghan, D. Shuey, and R. C. Hardison.** 1984. The rabbit C family of short interspersed repeats. J. Mol. Biol. **176**:1–20.
10. **Ciliberto, G., G. Raugei, F. Costanzo, L. Dente, and R. Cortese.** 1983. Common and interchangeable elements in the promoters of genes transcribed by RNA polymerase III. Cell **32**:725–733.
11. **Clark, J. L., and D. F. Steiner.** 1969. Insulin biosynthesis in the rat: demonstration of two proinsulins. Proc. Natl. Acad. Sci. U.S.A. **62**:278–285.
12. **Cleveland, D. W.** 1983. The tubulins: from DNA to RNA to protein and back again. Cell **34**:330–332.
13. **Dayhoff, M. O.** 1978. Atlas of protein sequence, vol. 5, supplement 3, p. 150–151. National Biomedical Research Foundation, Washington, D.C.
14. **Dierks, P., A. van Ooyen, M. D. Cochran, C. Dobkin, J. Reiser, and C. Weissman.** 1983. Three regions upstream from the cap site are required for efficient and accurate transcription of the rabbit β-globin gene in mouse 3T6 cells. Cell **32**:695–706.
15. **Episkopou, V., A. J. M. Murphy, and A. Efstratiadis.** 1984. Cell-specified expression of a selectable hybrid gene. Proc. Natl. Acad. Sci. U.S.A. **81**:4657–4661.
16. **Fisher, P. B., I. B. Weinstein, D. Eisenberg, and H. S. Ginsberg.** 1978. Interactions between adenovirus, a tumor promoter and chemical carcinogens in the transformation of rat embryo cells. Proc. Natl. Acad. Sci. U.S.A. **75**:2311–2314.
17. **Gallimore, P. H., and C. R. Richardson.** 1973. An improved banding technique exemplified in the karyotype analysis of two strains of rat. Chromosoma **41**:259–263.
18. **Green, M. C.** 1981. Genetic variants and strains of the laboratory mouse, Gustav Fischer Verlag, Stuttgart.
19. **Hammarstrom, K., G. Westin, C. Bark, J. Zabielsky, and U. Petterson.** 1984. Genes and pseudogenes for human U2 RNA. J. Mol. Biol. **179**:157–169.
20. **Heilig, R., R. Muraskowsky, and J.-L. Mandel.** 1982. The ovalbumin gene family. The 5′ end region of the X and Y genes. J. Mol. Biol. **156**:1–19.
21. **Henderson, A.** 1982. Cytological hybridization to mammalian chromosomes. Int. Rev. Cytol. **76**:1–46.
22. **Hollis, G. F., P. A. Hieter, O. W. McBride, D. Swan, and P. Leder.** 1982. Processed genes: a dispersed human immunoglobulin gene bearing evidence of RNA-type processing. Nature

(London) **296**:321–325.

23. **Humbel, R. E., H. R. Bosshard, and H. Zahn.** 1972. Chemistry of insulin, p. 111–132. *In* D. F. Steiner and N. Freinkel (ed.), Handbook of physiology, endocrinology. Williams & Wilkins, Baltimore.
24. **Jagadeeswaran, P., B. G. Forget, and S. M. Weissman.** 1981. Short interspersed repetitive DNA elements in eucaryotes: transposable DNA elements generated by reverse transcription of RNA Pol III transcripts? Cell **26**:141–142.
25. **Jeffreys, A. J., and S. Harris.** 1982. Processes of gene duplication. Nature (London) **296**:9–10.
26. **Jelinek, W. R., and C. W. Schmid.** 1982. Repetitive sequences in eukaryotic DNA and their expression. Annu. Rev. Biochem. **51**:813–844.
27. **Kakita, K., S. Giddings, and M. A. Permutt.** 1982. Biosynthesis of rat insulins I and II: evidence for differential expression of the two genes. Proc. Natl. Acad. Sci. U.S.A. **79**:2803–2807.
28. **Kimura, M.** 1982. The neutral theory as a basis for understanding the mechanism of evolution and variation at the molecular level, p. 3–56. *In* M. Kimura (ed.), Molecular evolution, protein polymorphism and the neutral theory, Springer-Verlag, Berlin.
29. **Kimura, M.** 1983. The neutral theory of molecular evolution, p. 117–148. Cambridge University Press, Cambridge.
30. **Kwok, S. C. M., S. J. Chan, and D. F. Steiner.** 1983. Cloning and nucleotide sequence analysis of the dog insulin gene. J. Biol. Chem. **258**:2357–2363.
31. **Lai, E. C., D. R. Roop, M.-J. Tsai, S. L. C. Woo, and B. W. O'Malley.** 1982. Heterogeneous initiation for transcription of the chicken ovomucoid gene. Nucleic Acids Res. **10**:5553–5567.
32. **Lalley, P. A., and J. M. Chirgwin.** 1984. Mapping of mouse insulin genes. Cytogenet. Cell Genet. **37**:515.
33. **Leavitt, J., P. Gunning, P. Porreca, S.-Y. Ng, C.-S. Lin, and L. Kedes.** 1984. Molecular cloning and characterization of mutant and wild-type human β-actin genes. Mol. Cell. Biol. **4**: 1961–1969.
34. **Lewin, R.** 1983. How mammalian RNA returns to its genome. Science **219**:1052–1054.
35. **Limbach, K. J., and R. Wu.** 1985. Characterization of a mouse somatic cytochrome c gene and three cytochrome c pseudogenes. Nucleic Acids Res. **13**:617–630.
36. **Lomedico, P., N. Rosenthal, A. Efstratiadis, W. Gilbert, R. Kolodner, and R. Tizard.** 1979. The structure and evolution of the two nonallelic rat preproinsulin genes. Cell **18**:545–558.
37. **MacLeod, A. R., and K. Talbot.** 1983. A processed gene defining a gene family encoding a human non-muscle tropomyosin. J. Mol. Biol. **167**:523–537.
38. **Manley, J. L., A. Fire, A. Cano, P. A. Sharp, and M. L. Gefter.** 1980. DNA-dependent transcription of adenovirus genes in a soluble whole-cell extract. Proc. Natl. Acad. Sci. U.S.A. **77**: 3855–3859.
39. **Maxam, A. M., and W. Gilbert.** 1980. Sequencing end-labeled DNA with base-specific chemical cleavages. Methods Enzymol. **65**:499–560.
40. **McLauchlan, J., D. Gaffney, J. L. Whitton, and J. B. Clements.** 1985. The consensus sequence YGTGTTYY located downstream from the AATAAA signal is required for efficient formation of mRNA 3' termini. Nucleic Acids Res. **13**:1347–1368.
41. **Mellon, P., V. Parker, Y. Gluzman, and T. Maniatis.** 1981. Identification of DNA sequences required for transcription of the human α1-globin gene in a new SV40 host-vector system. Cell **27**:279–288.
42. **Moos, M., and D. Gallwitz.** 1983. Structure of two human β-actin-related processed genes one of which is located next to a simple repetitive sequence. EMBO J. **2**:757–761.
43. **Mori, M., and M. Sasaki.** 1973. Fluorescence banding patterns of the rat chromosomes. Chromosoma **40**:173–182.
44. **Nesbitt, M. N.** 1974. Evolutionary relationships between rat and mouse chromosomes. Chromosoma **46**:217–224.
45. **Nishioka, Y., A. Leder, and P. Leder.** 1980. Unusual α-globin-like gene that has cleanly lost both globin intervening sequences. Proc. Natl. Acad. Sci. U.S.A. **77**:2806–2809.
46. **Perler, F., A. Efstratiadis, P. Lomedico, W. Gilbert, R. Kolodner, and J. Dodgson.** 1980. The evolution of genes: the chicken preproinsulin gene. Cell **20**:555–566.
47. **Potter, S. S.** 1982. DNA sequence analysis of a *Drosophila* foldback transposable element rearrangement. Mol. Gen. Genet. **188**:107–110.
48. **Robins, D., S. Ripley, A. Henderson, and R. Axel.** 1981. Transforming DNA integrates into the host chromosome. Cell **23**:29–39.
49. **Robinson, R.** 1982. Linkage in the Norway rat. Genet. Maps **2**:299–301.
50. **Roderick, T. H., and M. T. Davisson.** 1982. Linkage map of the mouse. Genet. Maps **2**:277–286.
51. **Rogers, J.** 1983. Retroposons defined. Nature (London) **301**: 460.
52. **Rogers, J.** 1983. A straight LINE story. Nature (London) **306**:113–114.
53. **Rougeon, F., and B. Mach.** 1977. Cloning and amplification of α and β mouse globin gene sequences synthesized in vitro. Gene **1**:229–239.
54. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A. **74**:5463–5467.
55. **Scarpulla, R. C.** 1984. Processed pseudogenes for rat cytochrome c are preferentially derived from one of three alternate mRNAs. Mol. Cell. Biol. **4**:2279–2288.
56. **Sharp, P. A.** 1983. Conversion of RNA to DNA in mammals: Alu-like elements and pseudogenes. Nature (London) **301**: 471–472.
57. **Smith, L. F.** 1966. Species variation in the amino acid sequence of insulin. Am. J. Med. **40**:662–666.
58. **Southern, E. M.** 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. **98**:503–517.
59. **Stein, J. P., R. P. Munjaal, L. Lagace, E. C. Lai, B. W. O'Malley, and A. R. Means.** 1983. Tissue-specific expression of a chicken calmodulin pseudogene lacking intervening sequences. Proc. Natl. Acad. Sci. U.S.A. **80**:6485–6489.
60. **Stolc, V., and T. J. Gill.** 1983. Linkage and polymorphism of a gene controlling lactate dehydrogenase in the rat. Biochem. Genet. **21**:933–941.
61. **Ueda, S., S. Nakai, Y. Nishida, H. Hisajima, and T. Honjo.** 1982. Long terminal repeat-like elements flank a human immunoglobulin epsilon pseudogene that lacks introns. EMBO J. **1**:1539–1544.
62. **Ullrich, A., T. J. Dull, A. Gray, J. Brosius, and I. Sures.** 1980. Genetic variation in the human insulin gene. Science **209**:612–615.
63. **Ullrich, A., T. J. Dull, A. Gray, J. A. Philips, and S. Peter.** 1982. Variation in the sequence and modification state of the human insulin gene flanking regions. Nucleic Acids Res. **10**:2225–2240.
64. **Ullu, E., and C. Tschudi.** 1984. Alu sequences are processed 7SL RNA genes. Nature (London) **312**:171–172.
65. **Ullu, E., and A. M. Weiner.** 1984. Human genes and pseudogenes for the 7SL RNA component of signal recognition particle. EMBO J. **3**:3303–3310.
66. **Valerio, D., M. G. C. Duyvesteyn, B. M. M. Dekker, G. Weeda, T. M. Berkvens, L. van der Voorn, H. van Ormondt, and A. J. van der Eb.** 1985. Adenosine deaminase: characterization and expression of a gene with a remarkable promoter. EMBO J. **4**:437–443.
67. **Van Arsdell, S. W., and A. M. Weiner.** 1984. Pseudogenes for human U2 small nuclear RNA do not have a fixed site of 3' truncation. Nucleic Acids Res. **12**:1463–1471.
68. **Vanin, E. F.** 1984. Processed pseudogenes. Biochem. Biophys. Acta **782**:231–241.
69. **Vanin, E. F., G. I. Goldberg, P. W. Tucker, and O. Smithies.** 1980. A mouse α-globin-related pseudogene lacking intervening sequences. Nature (London) **286**:222–226.
70. **Vieira, J., and J. Messing.** 1982. The pUC plasmids, an M13mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. Gene **19**:259–268.
71. **Villa-Komaroff, L., A. Efstratiadis, S. Broome, P. Lomedico, R. Tizard, S. Naber, W. Chick, and W. Gilbert.** 1978. A bacterial clone synthesizing proinsulin. Proc. Natl. Acad. Sci. U.S.A.

75:3727–3731.

72. **Voliva, C. F., S. L. Martin, C. A. Hutchison, III, and M. H. Edgell.** 1984. Dispersal process associated with the L1 family of interspersed repetitive DNA sequences. J. Mol. Biol. **178:** 795–813.

73. **Walker, M. D., T. Edlund, A. M. Boulet, and W. J. Rutter.** 1983. Cell-specific expression controlled by the 5′ flanking region of insulin and chymotrypsin genes. Nature (London) **306:**557–561.

74. **Yoshida, M. C.** 1978. Rat gene mapping by rat-mouse somatic cell hybridization and a comparative Q-banding analysis between rat and mouse chromosomes. Cytogenet. Cell Genet. **22:**606–609.

75. **Yoshida, M. C.** 1984. GP1, LDHA and PEPD are syntenic and assigned to rat chromosome 7. Cytogenet. Cell Genet. **37:**613.

76. **Zeitlin, S., and A. Efstratiadis.** 1984. In vivo splicing products of the rabbit β-globin pre-mRNA. Cell **39:**589–602.