Volume 8 Number 23 1980 Nucleic Acids Research

General occurrence and transcription of intervening sequences in mouse genes expressed via polyadenylated mRNA

Ian H.Maxwell, Françoise Maxwell and William E.Hahn

Department of Anatomy, University of Colorado, School of Medicine, Denver, CO 80262, USA

Received 10 September 1980

ABSTRACT

cDNA of modal size ~1600 nucleotides, transcribed from mouse brain polyadenylated mRNA, was annealed with excess of high molecular weight (~20 kb) genomic DNA. The S_1 nuclease method was then applied to determine possible sequence discontinuity between the cDNA and genomic DNA. A substantial reduction in the average size of the annealed cDNA was observed following S_1 nuclease treatment. Large single copy genomic DNA, annealed with excess high molecular weight DNA, and cDNA, hybridized with its template mRNA, were resistant to cleavage by S_1 nuclease. We interpret these results to indicate a high frequency of discontinuous coding sequences in the genomic DNA that annealed with the cDNA. The same result was obtained using fractionated cDNA, enriched in transcripts of relatively infrequent or abundant mRNA species. The result obtained with the infrequent sequence class cDNA indicates that tens of thousands of split genes exist in the mouse genome. Extensive cleavage of the cDNA by S_1 nuclease was also observed after hybridization with >30S nuclear RNA, indicating that intervening sequences are generally transcribed.

INTRODUCTION

Recently it has been shown that some gene products in eukaryotes are encoded discontinuously in the genomic DNA. Since the initial demonstration that some 28S rRNA genes in Drosophila are split [1], other examples of discontinuous coding sequences for rRNA and tRNA have been reported (e.g. [2,3]). The first evidence that mRNA species were not encoded in a continuous colinear sequence in DNA came from studies of adenovirus infection: certain sequences that are widely separated in adenovirus DNA were shown to be juxtaposed in viral mRNAs (reviewed in [4]). Subsequently, restriction mapping, electron microscopy, and sequence analysis of specific eukaryotic genes, in total DNA or in cloned fragments, led to the detection within these genes of sequences absent from the corresponding mRNAs. The presence of such intervening sequences within polypeptide coding regions was first reported for globin [5], ovalbumin [6] and immunoglobulin genes [7,8] and

has since been recognized in other specific eukaryotic genes (e.g., [9-12]).

These observations have led to the concept that the presence of intervening sequences may be a general feature of the genes of eukaryotes [13-16]. However, most of the genes which have been shown to be split code for highly abundant polypeptides in specialized cells (although, recently, discontinuity of the mRNA coding sequences has been shown for two genes which are usually represented by infrequent mRNAs [12,17]). In this report, we present evidence for the existence of intervening sequences in a substantial proportion of the genes expressed via abundant and infrequent species of polyadenylated mRNA in the mouse brain. Our results apply to a much larger number of genes than could be studied individually, thus strongly supporting the concept of the generality of split genes in higher eukaryotes, at least for those genes expressed via polyadenylated mRNA. Evidence is also presented that intervening sequences are generally transcribed, as has been demonstrated to be the case for certain specific genes (e.g., [18-20]).

MATERIALS AND METHODS

Synthesis and Fractionation of cDNA.

RNA was extracted from mouse brain polysomes pelleted through 0.6 M sucrose [21] or from polysomes excluded from Sepharose 4B [22]. Polyadenylated RNA was isolated from the polysomal RNA by binding twice to oligo(dT)-cellulose (Collaborative Research, grade T3) with the inclusion of denaturation steps [23] and was used as the template for cDNA synthesis [24], using dT_{10} primer, under conditions based on those described by Kacian and Myers [25], including Na pyrophosphate. After incubation for 60 minutes at 37°C, excess EDTA was added and extraction with phenol + chloroform and then chloroform was performed, in the presence of carrier *E. coli* RNA and phage fd DNA. After ethanol precipitation, RNA was hydrolysed in 0.1 M NaOH, 45 mM Na_4EDTA at 37°C overnight. cDNA was purified from the neutralized solution by exclusion from Sephadex G100 in the presence of fd DNA (1 µg/ml), included in order to reduce loss of cDNA by adsorption on surfaces. The specific activity of the cDNA was about 5×10^6 cpm/µg and its size ranged from about 200 to 5000 nucleotides, the number average size being about 1000 nucleotides (i.e., approximately 2/3 that of the template mRNA population [21]).

For removal of lower molecular weight cDNA, labeled cDNA was centrifuged (23 hr, 37,000 rpm, Beckman SW41 rotor, 4°C) in a 5-20% (w/v) sucrose gradient containing 0.9 M NaCl, 0.1 M NaOH and 1 µg/ml fd DNA. Markers of

SV40 ^{3}H-DNA Hae III fragments were centrifuged in a parallel gradient. cDNA fractions corresponding to >1600 nucleotides (~40% of the mass) were pooled, neutralized, diluted to 0.6 M Na^{+} and ethanol-precipitated, together with carrier nucleic acid (70 µg E. coli RNA and 5 µg fd DNA). The precipitated cDNA (recovery ~20% of that applied to the gradient) had a modal size of 1600 nucleotides and a number average size of 1100 nucleotides as determined from alkaline agarose gel electrophoresis.

For kinetic fractionation, to obtain an infrequent sequence class, cDNA was hybridized with a 200 fold excess of poly(A)$^{+}$mRNA to a Cot of 20. Infrequent sequences were isolated by hydroxyapatite (HAP) chromatography as the single strand (unbound) fraction. Application to HAP was in 0.12 M phosphate buffer at 70°C (in order to destabilize A:T duplexes formed between the T and A tails of the cDNA and mRNA [26]); the DNA:RNA hybrids were efficiently retained. cDNA was recovered from the single strand fraction by hollow fiber (F-36, Biomed Instrument, Chicago) concentration and ethanol precipitation with carrier nucleic acids.

Preparation of High Molecular Weight Genomic DNA.

DNA was prepared from crude nuclear pellets, obtained in the course of mouse brain polysome preparation [21], by phenol extraction [27]. The ethanol-precipitated DNA was spooled and redissolved and high molecular weight RNA was removed by precipitation with 3 M NaCl at 0°C and centrifugation [27]. The DNA was purified by treatment with Proteinase K, further phenol extractions and repeated ethanol precipitation and spooling.

The DNA was then treated with S_1 nuclease with the object of lowering the fragment size [28] and facilitating handling of concentrated solutions. The DNA (110 µg/ml) was incubated for 30 min at 37°C with S_1 nuclease (4900 units/ml; Miles) in 0.25 M NaCl, 1 mM Zn acetate, 30 mM Na acetate (pH 4.5). The nuclease was then removed by Proteinase K digestion and phenol extraction.

The average single strand fragment size of the final DNA preparation was about 20 kb (determined by alkaline agarose gel electrophoresis, with reference to markers of EcoRl digested λ DNA [29]). No fragments smaller than 10 - 12 kb were detected.

Isolation of Plasmids pBR321 and pBR322.

Strains of E. coli harboring pBR321 (RRl) and pBR322 (GM8) were subjected to plasmid amplification and supercoiled plasmid DNA was purified from cleared lysates [30]. ^{3}H-labeled pBR322 was prepared from cells labeled with ^{3}H-thymidine during the plasmid amplification.

Treatment with EcoRl (new England Biolabs; 80 units/ml) was performed at a DNA concentration of 250 µg/ml at 37°C for 17 hr (ionic conditions as specified by the supplier).

Preparation and Fractionation of Nuclear RNA.

Mouse brain nuclei were purified and RNA was extracted by a hot phenol method as described [21], except that the RNA was not treated with DNase or pronase. The RNA was precipitated with 3 M NaCl (0°C, overnight) and pelleted by centrifugation through 6 M NaBr, 10 mM EDTA [27] before being redissolved and precipitated with ethanol. The RNA was free from significant levels of DNA as shown by the fact that no detectable labeled DNA:DNA duplex (i.e. resistant to low salt RNase digestion [21,31]) was formed during hybridization with labeled cDNA. (This was verified for both poly(A)$^+$ and poly(A)$^-$ fractions of the nuclear RNA.)

Fractionation on oligo(dT)-cellulose was as described for polysomal RNA, above. Size fractionation of nuclear RNA was performed by centrifugation in sucrose gradients containing 50% (v/v) dimethyl sulfoxide as described previously [23], except that SDS was omitted. RNA sedimenting faster than ~30S (as determined by reference to rRNA markers) was recovered by ethanol precipitation.

Nucleic Acid Annealing and S_1 Nuclease Treatment.

Mixtures for annealing (10-20 µl) contained the quantities of driver and tracer nucleic acids indicated in the Figure legends, in 0.4 M or 0.24 M phosphate buffer (for DNA- or RNA-driven reactions, respectively), 1-2 mM EDTA and 0.1% SDS, sealed in glass capillary tubes. The mixtures were heated 2 min at 102°C (for DNA drivers) or 0.5 min at 80°C (for RNA drivers) and were then incubated at 64°C for the times indicated. We did not attempt to drive the reactions to completion and relatively short incubation periods were used to minimize possible nucleic acid degradation. The mixtures were diluted and the nucleic acids were excluded from Sephadex G100 and precipitated with ethanol. After redissolving, a sample (usually 0.1 - 0.2 of the total) was removed for electrophoretic analysis without S_1 nuclease treatment. The remainder was incubated with S_1 nuclease (Miles; final concentration 4900 units/ml) in a volume of 80 µl, containing 0.25 M NaCl, 1 mM Zn acetate, 30 mM Na acetate (pH 4.5), for 30 min at 37°C. Samples (1-5 µl were applied to DE-81 discs (Whatman) before and after the incubation to determine the extent of digestion of the tracer [32]. EDTA was added to a concentration of 10 mM and the mixture was applied to a column (6x0.5 cm) of Sephadex G50, equilibrated with 0.1 M NaCl, 10 mM Tris-HCl (pH 7.4).,

1 mM EDTA. The S_1 nuclease-resistant nucleic acids in the excluded fraction were then precipitated with ethanol. In most experiments it was observed that the proportion of the radioactivity excluded from G50 was similar to the proportion that bound to DE-81 discs.

In control experiments, it was established that, after digestion with S_1 nuclease under the conditions given above, less than 1% of labeled cDNA bound to DE-81 discs. This extent of digestion was not affected by the presence of RNA or of native DNA. However, when mouse denatured DNA (1 µg/µl) was present during the S_1 digestion, about 6% of labeled cDNA bound to DE-81 discs and was also excluded from Sephadex G50. When this cDNA was analyzed by alkaline agarose gel electrophoresis, significant radioactivity was only observed very near the bottom of the gel, representing chain lengths less than 200 nucleotides. This material therefore did not contribute significantly to the gel profiles presented in this report.

Alkaline Agarose Gel Electrophoresis.

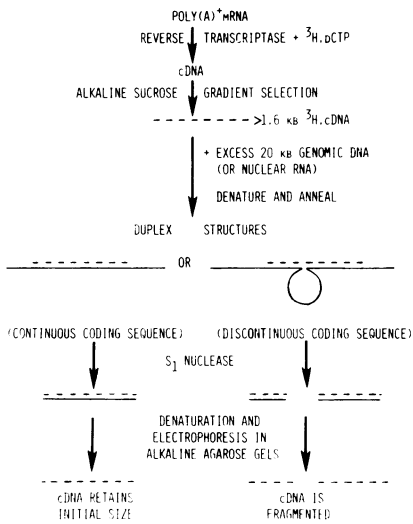
Cylindrical gels (87x6 mm) of 1% agarose (Seakem HGT(P)), 30 mM NaOH, 2 mM Na_4EDTA were electrophoresed in a vertical apparatus containing 30 mM NaOH, 2 mM Na_4 EDTA in anode and cathode compartments [33]. Samples (30 µl per gel; 1600-10,000 cpm) were applied in 0.1 M NaOH, 1 mM EDTA, 8% glycerol, ~0.02% bromocresol green. Electrophoresis was at 80 volts for 5 min and then at 30 volts for 4-4.5 hr, until the dye was ~25 mm from the bottom of the gel. The gels were cut into 2 mm slices which stood in toluene-Triton scintillation fluid for 24 hr before determination of radioactivity. Recovery was 70-90% of the radioactivity applied to the gels.

In all experiments, a parallel gel was included, containing markers consisting of HaeIII fragments of SV40 ^{3}H-DNA [34,35] together with (in some experiments) full-length linear SV40 ^{3}H-DNA (produced by EcoRl digestion). These markers gave prominent peaks of radioactivity at 5200 (SV40 linear DNA), 1660, 750 and 540 nucleotides (Hae III fragments).

RESULTS

Experimental Design

The experimental approach for detecting the presence of discontinuous coding sequences is shown in the diagram below. A labeled cDNA probe, consisting of transcripts of an entire mRNA population, is annealed with excess genomic DNA or nuclear RNA or, in a control experiment, with the



corresponding mRNA. Any intervening sequences (by definition absent from the mRNA and hence the cDNA molecules) will form single strand loops within the heteroduplexes resulting from annealing with the cDNA. These loops would be expected to be susceptible to cleavage by the single strand specific nuclease S_1, as would the cDNA strand at the corresponding position ([36]; see below for discussion of the efficiency of such cleavage). The presence of intervening sequences should therefore be indicated by a size reduction of the annealed cDNA following exposure to S_1 nuclease.

Effect of S_1 Nuclease on cDNA-Genomic DNA Duplex.

cDNA was annealed with a large excess of DNA (average size ~20 kb)

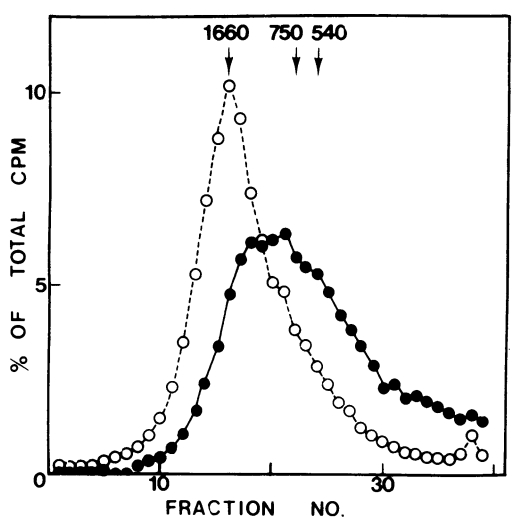


Figure 1. Effect of S_1 nuclease treatment on the size distribution in alkaline agarose gels of the large cDNA probe, after annealing with genomic DNA. cDNA (35,000 cpm; ~7 ng; modal size 1.6 kb) was annealed with mouse brain ~20 kb DNA (120 µg) for 12.5 hr (equivalent Cot 3300). A sample was removed and analyzed by electrophoresis without prior S_1 nuclease treatment. The remainder was subjected to electrophoresis after digestion with S_1 nuclease. 29% of the total cDNA, after annealing, was resistant to S_1 nuclease digestion as determined by binding to DEAE paper [32] or by exclusion from Sephadex G50. The arrows show the position of markers containing the indicated number of nucleotides, electrophoresed in a parallel gel. The ordinate is expressed in terms of % total cpm recovered from each gel. o----o, without S_1 nuclease; ●——●, with S_1 nuclease.

from mouse brain nuclei to a Cot at which ~30% of the cDNA was resistant to S_1 nuclease (as determined by binding to DE-81 paper [32]; synthesis and size selection of cDNA, transcribed from mouse brain polyadenylated mRNA, is described in Materials and Methods). After annealing, samples were subjected to alkaline agarose gel electrophoresis, with or without prior S_1 nuclease digestion, to determine the size distribution of the labeled cDNA. As shown in Fig. 1, the undigested cDNA showed a peak of radioactivity at approximately 1600 nucleotides as well as a heterogeneous distribution of smaller molecules, some of which resulted from partial degradation during the annealing incubation. Digestion with S_1 nuclease eliminated the 1600 nucleotide peak and increased the relative proportion of smaller molecules (Fig. 1). The proportion of the cDNA smaller than 750 nucleotides was increased from 23% to 52% by the S_1 nuclease treatment. The experiment was repeated twice, each time with a different preparation of cDNA, and essentially the same result was obtained.

The above results are consistent with the idea that many of the genomic sequences that annealed with cDNA species in the probe were interrupted by intervening sequences, although other interpretations are possible. Therefore, alternative explanations of the results were investigated in control experiments presented below.

S_1 Nuclease Does Not Significantly Cleave Double-Strand DNA.

First, the possible spurious cleavage of double strand DNA by our S_1 nuclease was tested using plasmid pBR322 (4.36 kbp [37]), labeled *in vivo* with ^{3}H-thymidine, and made linear with EcoRl. The heat denatured plasmid DNA was renatured and treated with S_1 nuclease under conditions as used for the mouse DNA above and was then electrophoresed in an alkaline agarose gel. A sharp peak of radioactivity, including most of the labeled DNA, was observed in the position expected for pBR322 intact single strands and a negligible amount of labeled DNA was present in the region of the gel corresponding to less than 1000 nucleotides (results not shown). This result ruled out the possibility that the cleavage by S_1 nuclease shown in Fig. 1 might be due to a spurious attack on DNA duplex or to the generation of spurious S_1 nuclease-sensitive sites during the incubation period required for annealing.

Effect of S_1 Nuclease on cDNA Hybridized to Template RNA.

Further assurance against an artifactual explanation for the S_1 nuclease cleavage shown in Fig. 1 would be provided if the cDNA were shown to be

resistant to cleavage after hybridization with its template RNA. The cDNA was therefore hybridized with excess poly(A)$^+$mRNA and was then subjected to alkaline agarose gel electrophoresis, with or without prior exposure to S_1 nuclease. As shown in Fig. 2, the hybridized cDNA showed very little cleavage (see also Fig. 5C). This result excluded the possibility that the cDNA was, for any reason, intrinsically incapable of forming duplexes resistant to cleavage by S_1 nuclease (e.g. owing to copying errors by reverse transcriptase).

Size Reduction of Annealed cDNA Is Not Due to End Overlap.

The high molecular weight genomic DNA used as driver (Fig. 1) was randomly cleaved to a limited extent during its preparation (see Materials and Methods). Although the average size of this DNA was more than 10 times that of the cDNA, a small fraction of the cDNA could have formed partial duplexes by annealing in a position overlapping the end of a driver DNA fragment. To assess the possible contribution of this effect to the subsequent cleavage of the annealed cDNA by S_1 nuclease, labeled, randomly cleaved single copy DNA of similar size to the cDNA was prepared from mouse brain nuclear DNA, labeled with ^{3}H-dCTP by nick translation (38) to 5.5×10^6 cpm/µg. The same experiment as in Fig. 1 was performed using this single copy probe in place of the cDNA. Since this probe was prepared from genomic DNA, it contained the same distribution of intervening and coding sequences as the driver DNA. Therefore, in this experiment, an annealed tracer molecule should be subject to cleavage by S_1 nuclease only in the

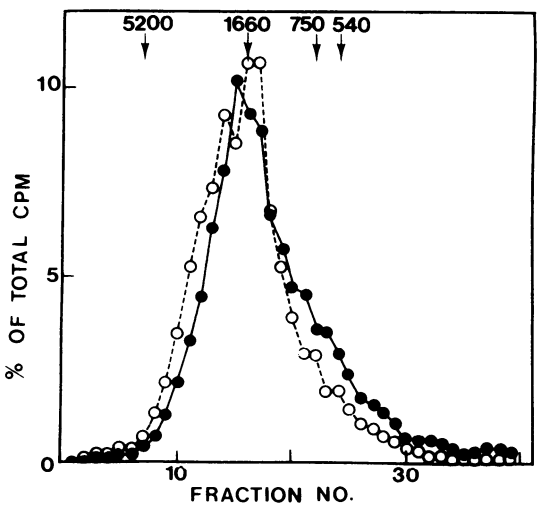


Figure 2. Effect of S_1 nuclease on the size distribution of the large cDNA probe, after hybridization with its template mRNA. cDNA (35,000 cpm; ~7 ng) was hybridized with mouse brain poly(A)$^+$mRNA (~38 µg) for 3 hr (Cot ~150) and subjected to alkaline agarose gel electrophoresis, with or without S_1 nuclease treatment under the same conditions as in Fig. 1. About 65% of the probe was resistant to S_1 nuclease after the hybridization. o----o, without S_1 nuclease; ●——●, with S_1 nuclease.

event that it overlapped the end of a driver DNA fragment. The frequency of such occurrence should be the same as in the case where cDNA of the same average size was annealed with the same driver DNA (Fig. 1). We observed that exposure to S_1 nuclease resulted in only slight cleavage of the annealed single copy tracer and the ~1600 nucleotide peak of radioactivity was preserved. The proportion of the tracer smaller than 750 nucleotides was increased from 22% to 29% by S_1 nuclease treatment (cf. 23% to 52% for the cDNA in Fig. 1). Therefore, the effect of tracer-driver end-overlap was much too small to account for the observed S_1 nuclease cleavage of the annealed cDNA shown in Fig. 1. (Direct comparison of the percentages stated gives a maximum estimate of the contribution of end-overlap to the cDNA cleavage. This is because such direct comparison involves the assumption that both tracers were uniformly labeled. In fact, the single copy tracer may have been, to some extent, deficient in label in its 5' proximal sequences owing to our selection of large molecules after nick translation. This would result in an increased proportion of radioactivity in smaller fragments derived from end-overlapping molecules than in the case of a uniformly labeled tracer.)

cDNA Is Largely Complementary to Single Copy Genomic DNA.

Another explanation of the results shown in Fig. 1 might be in terms of mismatched duplexes formed between repetitive transcripts in the cDNA and related repetitive sequences in the genomic DNA. This explanation is unlikely because we have observed that total cDNA, transcribed from mouse brain poly(A)$^+$mRNA, anneals with excess mouse DNA fragments very largely at the rate characteristic of single copy DNA (unpublished observations); less than 5% of this cDNA was found to anneal with repetitive sequences. Nevertheless, the following experiment was performed in order to determine whether the cDNA that annealed with genomic DNA under our conditions was enriched in repetitive sequence transcripts. High molecular weight cDNA was annealed with excess genomic DNA and treated with S_1 nuclease as above. 40% of the cDNA was not digested by the nuclease. A further excess of mouse genomic DNA was added and the mixture was sheared with a pressure cell under conditions producing DNA fragments of ~400 nucleotides (modal size). After denaturation, the kinetics of annealing of the recovered cDNA under standard conditions (0.4 M phosphate buffer, 65°C) were determined using hydroxyapatite chromatography (results not shown). Most of the observed annealing (maximum 44%, at equivalent Cot 5×10^4) occurred at a rate characteristic of single copy DNA. The kinetics observed at low

Cot (<100) indicated that no more than 8-12% of the recovered cDNA annealed with repetitive sequences in the genomic DNA. This observation sets an upper limit to the contribution of mismatched repetitive sequences to the observed S_1 nuclease cleavage in Fig. 1. In fact, any such contribution is probably much less than this limit since many of the repetitive sequence duplexes formed during reassociation of eukaryotic DNA fragments are not cleaved under conditions of S_1 nuclease digestion similar to those we have used [39,40].

Even if any duplexes formed by the cDNA with genomic repetitive sequences were efficiently cleaved by S_1 nuclease the maximum cumulative effect of such cleavage and of end-overlap (discussed above) could account for only about half the extent of cleavage observed in Fig. 1. Therefore it is probable that most of the cleavage by S_1 nuclease that was observed in this experiment was indeed due to the frequent occurrence of intervening sequences in the genomic DNA (see also Discussion).

Efficiency of Cleavage by S_1 Nuclease.

While much of the S_1 nuclease-resistant cDNA was smaller than 500-800 nucleotides, a significant fraction was at least 1600 nucleotides in size (Fig. 1), suggesting the possible existence of some long uninterrupted coding sequences in the genome. However, the size distribution of the S_1 nuclease-resistant cDNA would reflect that of the genomic coding sequences accurately only if all potential cleavage sites were actually cleaved by the nuclease under the conditions used. (By potential cleavage sites, we are referring to the cDNA strand at the position opposite a loop in the type of duplex structure illustrated in the diagram). In an attempt to determine whether this was the case, a model DNA heteroduplex containing a single strand loop was constructed from plasmids pBR322 and pBR321, as shown in Fig. 3 (inset).

The electrophoretic profiles in Fig. 3 show the effect of S_1 nuclease treatment on the single strand size of the shorter strand (pBR322, present as labeled tracer: see legend to Fig. 3) of the heteroduplex. The three peaks of radioactivity correspond to intact strands of pBR322 (4.36 kb) and to the arms of the heteroduplex (about 2.5 kb and 1.8 kb). The most slowly migrating species (i.e. intact strands) contained about 60% of the total radioactivity in all three peaks, indicating incomplete cleavage of the pBR322 strand of the heteroduplexes by S_1 nuclease. The result was essentially the same whether or not an excess of mouse denatured DNA was present during the S_1 nuclease digestion. If this experiment represents a

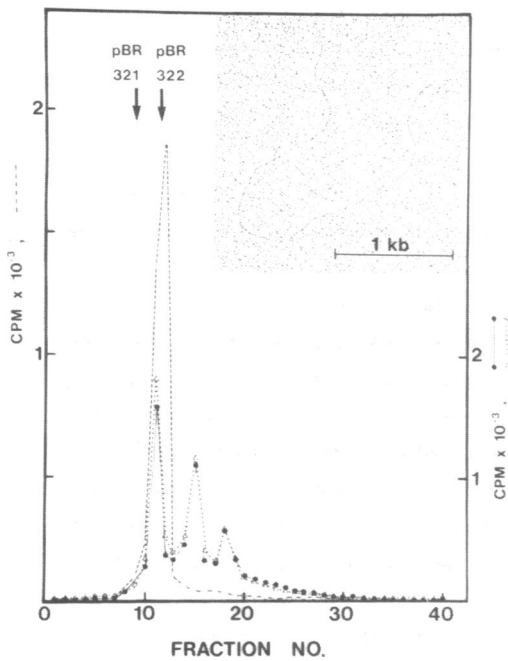


Figure 3. Effect of S_1 nuclease on a plasmid heteroduplex DNA used as a model for estimating the efficiency of S_1 nuclease cleavage in the experiment illustrated in Fig. 1. The inset shows an electron micrograph of a heteroduplex molecule that was formed in an annealing mixture containing equal amounts of pBR322 and pBR321 (both cleaved by EcoRl). Spreading was from 50% formamide, 0.1 M Tris-HCl (pH 8.5), 1 mM EDTA, 30 µg/ml cytochrome c on water as hypophase. The single strand loop is in the pBR321 strand and represents a region deleted in pBR322[41].

The electrophoretic profiles show the effect of S_1 nuclease treatment on the size of labeled pBR322 after annealing with excess unlabeled pBR321. A mixture of ^{3}H-labeled pBR322 (25,500 cpm, 1 µg) and unlabeled pBR321 (60 µg) DNA (both plasmids cleaved with EcoRl) was denatured with NaOH and then neutralized and allowed to anneal at 60°C for 30 min. Samples were analyzed by alkaline agarose gel electrophoresis, with or without S_1 nuclease treatment under the same conditions as in Fig. 1. About 90% of the pBR322 DNA was resistant to S_1 nuclease after annealing, as determined by exclusion from Sephadex G50. Excess mouse denatured DNA (50 µg) was included in one digestion to determine whether its presence might lower the efficiency of S_1 nuclease cleavage of the heteroduplexes. The position of the pBR321 driver DNA (about 5.2 kb, arrowed) was determined by ethidium staining of the gels, before slicing. o----o, without S_1; ●——●, with S_1 in absence of mouse DNA; Δ....Δ, with S_1 in presence of mouse DNA.

valid model for duplex structures formed between cDNA and genomic DNA, it can be concluded that the frequency of occurrence of intervening sequences would be underestimated from the extent of cleavage by S_1 nuclease. Some, or all, of the largest S_1 nuclease-resistant cDNA shown in Fig. 1 may, therefore, reflect inefficiency of cleavage rather than the existence of long continuous coding sequences.

The above result was not entirely surprising since it has been observed that a substantial fraction of nicked circular SV40 DNA is resistant to S_1 nuclease cleavage under high salt conditions [42].

Complex Class mRNAs Are Encoded Discontinuously in the Genome.

Kinetics of mRNA-driven cDNA hybridization indicate that individual poly(A)$^+$mRNA species occur in widely different frequencies in the mouse brain [26,43-45], as in other cells and tissues [44, 46]. The experiment shown in Fig. 1 did not establish whether mRNAs derived from split genes were represented in each frequency class in the mRNA population. Because the infrequent (complex) class of mRNA represents by far the largest number of individual species it was important to determine whether this class is encoded discontinuously, in attempting to assess the generality of split genes. An infrequent class large cDNA probe was therefore prepared. Total cDNA was fractionated kinetically by template RNA-driven hybridization to appropriate Cot and hydroxyapatite chromatography and then unhybridized >1.6 kb cDNA was isolated by alkaline sucrose gradient centrifugation (see Materials and Methods). As shown in Fig. 4A, this cDNA fraction hybridized with its template RNA about 10 fold more slowly than did total >1.6 kb cDNA, indicating substantial enrichment in transcripts of infrequent mRNA species.

The infrequent class cDNA was annealed with excess genomic DNA (~20 kb) and its size distribution, with or without treatment with S_1 nuclease was determined. As shown in Fig. 4B, S_1 nuclease reduced the size of the annealed cDNA to an extent similar to that observed when total large cDNA was used as the probe (Fig. 1). As will be discussed later, this result implies the existence of many thousand genes containing intervening sequences.

A cDNA probe representing the more abundant poly(A)$^+$mRNA species was also isolated and annealed with genomic DNA. Subsequent exposure to S_1 nuclease resulted in substantial size reduction of the annealed probe (results not shown), indicating that many of the more abundant poly(A)$^+$mRNA species in the mouse brain are also transcribed from split genes.

Sequences Homologous with Poly(A)$^+$mRNA Are Discontinuous in High Molecular Weight Nuclear RNA.

The same procedure was employed to determine whether intervening sequences are represented in various fractions of nuclear RNA. If intervening sequences are generally transcribed and are then removed from the RNA by a splicing process it would be expected that the larger nuclear RNA molecules would contain these transcripts. This possibility was tested employing high molecular weight fractions of both polyadenylated and non-polyadenylated RNA prepared from mouse brain nuclei. RNA sedimenting

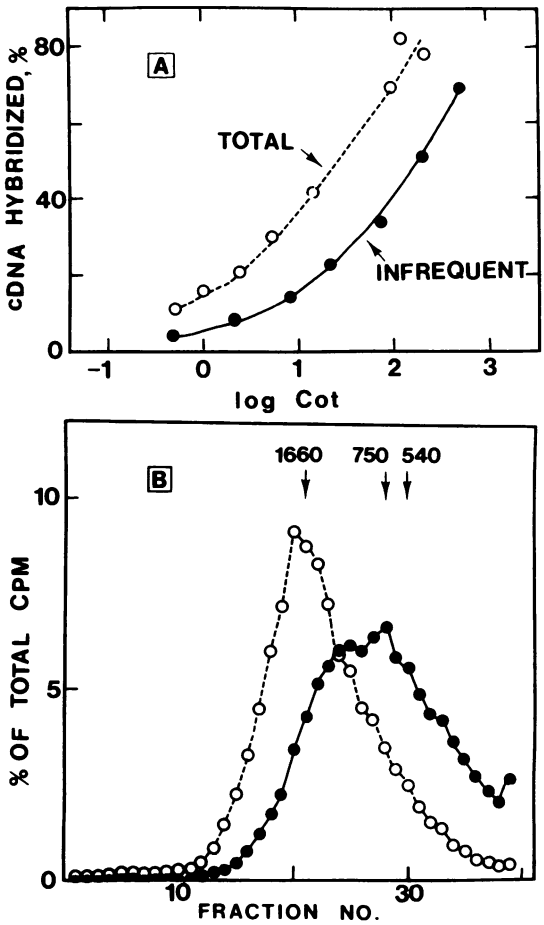


Figure 4. A. Poly(A)$^+$ mRNA driven hybridization kinetics of large (>1.6 kb) cDNA, isolated either from total cDNA or from a cDNA fraction comprised of infrequent sequences. cDNA (500-3500 cpm/µl) was hybridized with poly(A)$^+$mRNA (0.2-0.5 µg/µl) in phosphate buffer (0.36 M Na^+, pH 6.8) containing 1.2 mM EDTA, 0.1% SDS, at 64°C. After various times of incubation 1 µl samples were assayed for S_1 nuclease resistance as described (32). B. Effect of S_1 nuclease on the size distribution of the cDNA fraction comprised of infrequent sequences, after annealing with genomic DNA. The experiment was performed in the same way as for the unfractionated probe (see Fig. 1). The cDNA (30,000 cpm; ~6 ng) was annealed with mouse brain ~20 kb DNA (100 µg) for 18 hr (equivalent Cot 4500), after which 30% of the cDNA was resistant to S_1 nuclease digestion. o---o, without S_1 nuclease; ●——●, with S_1 nuclease.

faster than ~30S in sucrose gradients, (under conditions minimizing aggregation; see Materials and Methods) was recovered and hybridized with large cDNA, transcribed from poly(A)$^+$mRNA. The size distribution of the hybridized cDNA was then determined, with or without S_1 nuclease treatment.

The results obtained after hybridization of the cDNA with large poly(A)$^+$ nuclear RNA are shown in Fig. 5A and with large poly(A)$^-$ nuclear RNA in Fig. 5B and D (after different incubation periods). Although some cleavage by S_1 nuclease is apparent in Fig. 5A, much of the hybridized cDNA appeared to retain its initial size. The results obtained with large poly(A)$^-$ nuclear RNA as driver (Fig. 5B and D) were strikingly different. In this case, marked cleavage of the annealed cDNA by S_1 nuclease was

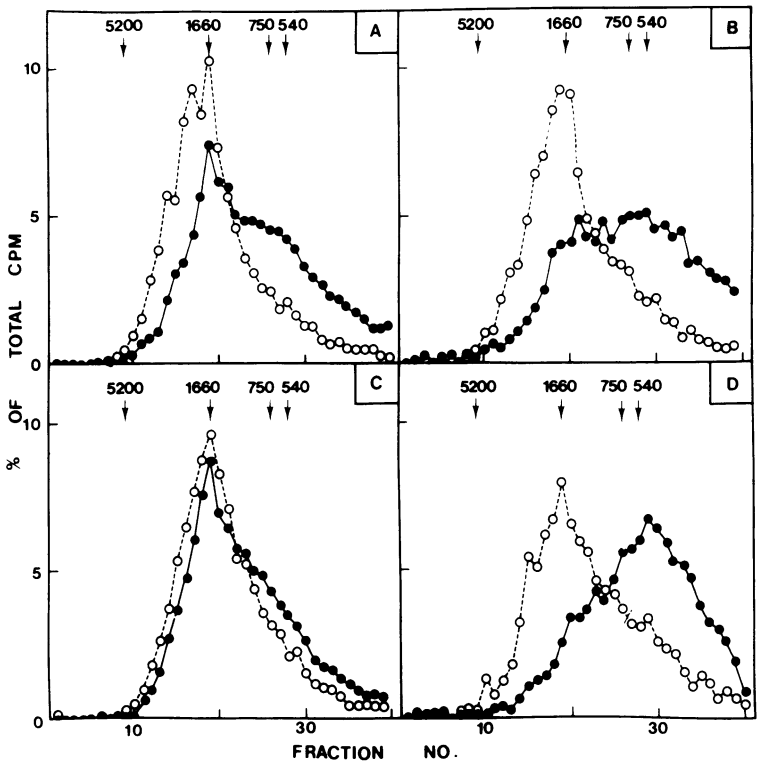


Figure 5. Effect of S_1 nuclease on the size distribution of the large cDNA probe after hybridization with high molecular weight fractions of polyadenylated or nonpolyadenylated nuclear RNA.
A. cDNA (47,000 cpm; ~9 ng) was hybridized with >30S poly(A)$^+$RNA (~80 µg) for 12.5 hr (Cot ~850). About 45% of the cDNA became resistant to S_1 nuclease digestion.
B., D. cDNA (23,000 cpm; ~5 ng, each reaction) was hybridized with >30S poly(A)$^-$RNA (~70 µg each reaction) for 16.5 hr (B) or 50 hr (D)(Cot ~1150 and 2800, respectively. About 12% and 20% of the cDNA became resistant to S_1 nuclease digestion in B. and D. respectively.
C. Control experiment showing that the cDNA, hybridized with mRNA, remained resistant to S_1 nuclease cleavage even when the hybridization occurred during a 40 hr incubation period. cDNA (23,000 cpm; ~5 ng) was hybridized with mouse brain poly(A)$^+$mRNA (~1 µg), in the presence of E. coli RNA (50 µg), for 40 hr (Cot ~40). About 60% of the cDNA became resistant to S_1 nuclease digestion. o----o, without S_1 nuclease; ●——●, with S_1 nuclease.

observed, to an extent at least as great as when genomic DNA was used to drive the reaction (Fig. 1). The control experiment shown in Fig. 5C shows that these results were not an artifact of the prolonged incubation

required to obtain sufficient cDNA hybridization; cDNA that was hybridized with poly(A)$^+$mRNA during a 40 hr incubation was largely resistant to cleavage by S_1 nuclease (Fig. 5C; compare with Fig. 2). We interpret the results to indicate that intervening sequences in genomic DNA generally are transcribed. The apparently higher frequency of these transcripts (relative to coding sequences) in poly(A)$^-$ than in poly(A)$^+$ large nuclear RNA suggests that polyadenylation and splicing of poly(A)$^+$mRNA precursors are events that occur in rapid succession (see Discussion).

DISCUSSION

We have presented evidence for the frequent occurrence and transcription of intervening sequences within coding sequences in mouse DNA. Our results add support to the idea of the generality of split genes (at least for genes expressed via poly(A)$^+$mRNA) which has been suggested from studies on a few specific genes. Our evidence is based on the observation of extensive S_1 nuclease cleavage of a complex cDNA probe, after annealing with high molecular weight DNA or nuclear RNA. The principle of this type of experiment has previously been employed for mapping transcripts of viral DNA on their respective genomes [47,48] and in the demonstration of discontinuous genes coding for mouse immunoglobulins [8,49].

Measurement of the complexity of mouse brain poly(A)$^+$mRNA by saturation hybridization with single copy DNA [21,24] has shown that this population is comprised of about 90,000 different species of average size (~1.5 kb). We have estimated that ~70,000 species are present in the lowest frequency class of poly(A)$^+$mRNA whose transcripts contribute significantly to the mass of cDNA [45]; about 40% of the total cDNA mass was representative of this complex mRNA class. Assuming that the cDNA was representative of its template mRNA in terms of complexity (see below for discussion of this point), the observation (Fig. 1) that a substantial proportion of cDNA, annealed with genomic DNA, was cleaved by S_1 nuclease suggests that the coding sequences of many genes are interrupted. This conclusion was substantiated by the observation that the extent of S_1 nuclease-cleavage of infrequent class cDNA, annealed with genomic DNA (Fig. 4), was similar to that observed with a total cDNA probe (Fig. 1). This result implies that tens of thousands of split genes are present in the mouse genome.

Intervening sequences have previously been demonstrated in numerous individual genes in eukaryotes (see Introduction). However, apart from

viral genes, the genes studied have, with few exceptions [12,17], been those coding for highly abundant polypeptides produced in specialized cell types. To our knowledge, the experiments presented here provide the first evidence that a highly complex mRNA population is encoded by split genes. As already noted (results not shown), we also observed substantial S_1 nuclease cleavage of fractionated cDNA representing more abundant poly(A)$^+$ mRNA species, after annealing with genomic DNA. It therefore appears that mechanisms determining the abundance of particular mRNA species are not related in any simple manner to the presence or absence of intervening sequences in the corresponding genes.

The conclusions stated above imply the assumption that our cDNA probes were at least approximately representative of the sequence composition of the poly(A)$^+$mRNA from which they were copied. It has recently been found that the complexity of cDNA transcribed from mouse liver poly(A)$^+$mRNA, determined by saturation hybridization with labeled single copy DNA, is essentially the same as that of the template RNA (J. Van Ness and W. E. Hahn, submitted for publication), indicating that the vast majority of individual species in this mRNA population is copied by reverse transcriptase. It is therefore likely that the cDNA used in the experiments described here, which was synthesized under the same conditions, was also largely representative of the sequence composition of the poly(A)$^+$ mRNA template. In any case, the conclusion that we have detected some tens of thousands of split genes could only be seriously in error if any preferential copying of mRNA species by reverse transcriptase were positively correlated with the derivation of those species from split genes. There is no reason to expect such a correlation to exist.

Our conclusions regarding the frequency of split genes, obtained using a large cDNA probe, are clearly not necessarily applicable to genes specifying mRNAs much smaller than average. Meyuhas and Perry [50] have reported that, in mouse L cells, there exists a strong bias towards more abundant species among the smaller poly(A)$^+$mRNA molecules. It is not known whether this is also true for mouse brain mRNA. If this is the case, any preferential removal of abundant species in selecting the large cDNA must presumably have been balanced by the removal of incomplete cDNA transcripts of less frequent mRNA species, since it was observed that the poly(A)$^+$mRNA driven hybridization kinetics of the large cDNA fraction and of the unfractionated cDNA were similar (results not shown).

It is unlikely that the cleavage of cDNA shown in Fig. 1 is due, to

any marked extent, to base pair mismatching owing to the annealing of cDNA with closely related but not perfectly homologous sequences in the genome. Although there are gene families comprised of sequences sufficiently homologous for cross annealing [9,51,52] it is generally thought that most different mRNA specifying genes are unique (i.e. present once per haploid genome, reviewed by Davidson [53]). Recent studies on restriction digests of genomic DNA have shown some genes to be unique (e.g. conalbumin [54]). The cDNA used in our experiments annealed with genomic DNA with kinetics characteristic of single copy sequences. However, the accuracy of this measurement does not exclude the possibility that some of the cDNA may have annealed with sequences repeated a few times per genome. Such cross-annealing would only contribute to the S_1 nuclease cleavage observed in our experiments in cases where the resulting duplexes were sufficiently well-matched to be stable under stringent annealing conditions but contained enough mismatching to be susceptible to S_1 nuclease attack. Under the digestion conditions we used, cleavage of the DNA strand opposite a deletion loop in a heteroduplex was ~50% efficient (see Results) and cleavage opposite a single strand nick was only ~10% efficient (unpublished observation). Therefore it seems unlikely that efficient S_1 nuclease cleavage would occur at the sites of single mismatched base pairs. This conclusion is supported by evidence that repetitive DNA duplexes that are substantially mismatched can be isolated following S_1 nuclease digestion under high salt conditions similar to those we employed [39,40]. From these considerations, we think it unlikely that cross-annealing of cDNA to closely related genes could have made a major contribution to the S_1 nuclease cleavage of cDNA annealed to genomic DNA. Moreover, the absence of cleavage of cDNA hybridized to mRNA places a further constraint on the extent to which cross-annealing to related genes may have contributed to our results; such putative related genes would have to remain unexpressed (at least as abundant mRNA species) in brain tissue.

In the experiment using genomic DNA driver (Fig. 1), about 34% of the annealed cDNA remained larger than 1000 nucleotides after S_1 nuclease treatment. This cDNA does not necessarily represent uninterrupted coding sequences but, instead, may have resulted from inefficient S_1 nuclease cleavage, as suggested by the observed extent of cleavage of a model heteroduplex constructed from plasmid DNAs (see Fig. 3). In all experiments, conditions of S_1 nuclease treatment were such that almost complete digestion of single strand DNA was obtained (see Materials and Methods)

and it is not known why only a fraction of the heteroduplex was cleaved (Fig. 3). It is possible that efficient cleavage of the unlooped strand of a heteroduplex depends on local destabilization of base pairing resulting from the presence of the loop. Thus, the probability of S_1 nuclease cleaving the unlooped strand before substantially digesting the single strand loop would determine the extent of cleavage observed. Alternatively it is possible that S_1 nuclease might show some sequence preference for cleavage and that the result shown in Fig. 3 was due to efficient cleavage of one of the two kinds of heteroduplex present (i.e. with the loop in either the + strand or the - strand of pBR321).

The expression of split genes is generally believed to involve the removal of intervening sequences by a process of splicing of nuclear RNA molecules (for review, see [15]). Transcripts of intervening sequences have been demonstrated in nuclear RNA in a limited number of specific cases, e.g. mouse β-globin [18], chicken ovalbumin [19] and ovomucoid [55] and numerous viral genes [48, 56-58]. Indirect evidence from S_1 nuclease cleavage experiments has been reported for the transcription of an intervening sequence of a mouse light chain immunoglobulin gene [20] Our results showing cleavage of cDNA hybridized to high molecular weight fractions of nuclear RNA indicate the transcription of intervening sequences. However, the extent of S_1 nuclease cleavage of the cDNA seen after hybridization with >30S poly(A)$^+$RNA was relatively small compared with that seen after hybridization with the >30S poly(A)$^-$ fraction (cf. Figs. 5A, B and D). It therefore appears that only the latter fraction contained a high frequency of interrupted coding sequences. These observations suggest that polyadenylation, cutting and splicing of mRNA precursor molecules generally occur in rapid succession. Evidence that polyadenylation precedes splicing has been reported for viral transcripts in SV40 [48] and adenovirus infected cells [56].

Finally, it should be emphasized that the conclusions drawn here apply only to genes expressed via poly(A)$^+$mRNA. Histone genes in sea urchin [59] and Drosophila [60] lack intervening sequences; conceivably this might generally be true for other genes expressed via nonpolyadenylated mRNA. We have recently provided evidence for the existence of a complex class of nonpolyadenylated mRNA in the mouse brain [24]. Work is in progress to determine whether this class of mRNA is encoded discontinuously.

ACKNOWLEDGEMENTS

This work was supported by grants from the National Institutes of Health. We thank Marlene Lauth for electron microscopy.

REFERENCES

1 Glover, D. M. and Hogness, D. S. (1977) Cell 10, 167-176
2 Allet, B. and Rochaix, J.-D. (1979) Cell 18, 55-60
3 Etcheverry, T., Colby, D. and Guthrie, C. (1979) Cell 18, 11-26
4 Sambrook, J. (1977) Nature 268, 101-104
5 Jeffreys, A. J. and Flavell, R. A. (1977) Cell 12, 1097-1108
6 Breathnach, R., Mandel, J. L. and Chambon, P. (1977) Nature 270, 314-319
7 Hozumi, N. and Tonegawa, S. (1976) Proc. Nat. Acad. Sci., USA, 73, 3628-3632
8 Rabbitts, T. H. and Forster, A. (1978) Cell 13, 319-327
9 Fiddes, J. C., Seeburg, P. H., DeNoto, F. M., Hallewell, R. A., Baxter, J. D. and Goodman, H. M. (1979) Proc. Nat. Acad. Sci., USA, 76, 4294-4298
10 Sargent, T. D., Wu, J.-R., Sala-Trepat, J. M., Wallace, R. B., Reyes, A. A. and Bonner, J. (1979) Proc. Nat. Acad. Sci., USA, 76, 3256-3260
11 Cochet, M., Gannon, F., Hen, R., Maroteaux, L., Perrin, F. and Chambon, P. (1979) Nature 282, 567-574
12 Nunberg, J. H., Kaufman, R. J., Chang, A. C. Y., Cohen, S. N. and Schimke, R. T. (1980) Cell 19, 355-364
13 Doolittle, W. F. (1978) Nature 272, 581
14 Gilbert, W. (1978) Nature 271, 501
15 Darnell, J. E., Jr. (1978) Science 202, 1257-1260
16 Crick, F. (1979) Science 204, 264-271
17 Wahl, G. M., Padgett, R. A. and Stark, G. R. (1979) J. Biol. Chem. 254, 8679-8689
18 Tilghman, S. M., Curtis, P. J., Tiemeier, D. C., Leder, P. and Weissman, C. (1978) Proc. Nat. Acad. Sci., USA, 75, 1309-1313
19 Roop, D. R., Nordstrom, J. L., Tsai, S. Y., Tsai, M.-J. and O'Malley, B. W. (1978) Cell 15, 671-685
20 Rabbitts, T. H. (1978) Nature 275, 291-296
21 Bantle, J. A. and Hahn, W. E. (1976) Cell 8, 139-150
22 Eschenfeldt, W. H. and Patterson, R. J. (1975) Preparative Biochemistry 5, 247-255
23 Bantle, J. A., Maxwell, I. H. and Hahn, W. E. (1976) Anal. Biochem. 72, 413-427
24 Van Ness, J., Maxwell, I. H., and Hahn, W. E. (1979) Cell 18, 1341-1349
25 Kacian, D. L. and Myers, J. C. (1976) Proc. Nat. Acad. Sci., USA, 73, 2191-2195
26 Ryffel, G. U. and McCarthy, B. J. (1976) Biochemistry 14, 1379-1384
27 Kirby, K. S. and Cook, E. A. (1967) Biochem. J. 104, 254-257
28 Perlman, S., Phillips, C. and Bishop, J. O. (1976) Cell 8, 33-42
29 Helling, R. B., Goodman, H. M. and Boyer, H. W. (1974) J. Virol. 14, 1235-1244
30 Clewell, D. B. (1972) J. Bact. 110, 667-676
31 Galau, G. A., Britten, R. J. and Davidson, E. H. (1974) Cell 2, 9-20
32 Maxwell, I. H., Van Ness, J. and Hahn, W. E. (1978) Nucl. Acids. Res. 5, 2033-2038
33 McDonell, M. W., Simon, M. N. and Studier, F. W. (1977) J. Mol. Biol. 110, 119-146
34 Yang, R. C.-A., Van de Voorde, A., and Fiers, W. (1976) Eur. J. Biochem. 61, 101-117

35 Reddy, V. B., Thimmappaya, B., Dhar, R., Subramanian, K. N., Zain, B. S., Pan, J., Ghosh, P. K., Celma, M. L. and Weissman, S. M. (1978) Science 200, 494-502
36 Shenk, T. E., Rhodes, C., Rigby, P. W. J. and Berg, P. (1975) Proc. Nat. Acad. Sci., USA, 72, 989-993
37 Sutcliffe, J. G. (1978) Nucl. Acids Res. 5, 2721-2728
38 Rigby, P. W. J., Dieckmann, M., Rhodes, C. and Berg, P. (1977) J. Mol. Biol. 113, 237-251
39 Britten, R. J., Graham, D. E., Eden, F. C., Painchaud, D. M. and Davidson, E. H. (1976) J. Mol. Evol. 9, 1-23
40 Houck, C.-M., Rinehart, F. P. and Schmid, C. W. (1978) Biochim. Biophys. Acta 518, 37-52
41 Bolivar, F., Rodriguez, R. L., Greene, P. J., Betlach, M. C., Heyneker, H. L., Boyer, H. W., Crosa, J. H. and Falkow, S. (1977) Gene 2, 95-113
42 Beard, P., Morrow, J. F. and Berg, P. (1973) J. Virol. 12, 1303-1313
43 Young, B. D., Birnie, G. D. and Paul, J. (1976) Biochemistry 15, 2823-2829
44 Hastie, N. D. and Bishop, J. O. (1976) Cell 9, 761-774
45 Hahn, W. E., Van Ness, J. and Maxwell, I. H. (1978) Proc. Nat. Acad. Sci., USA, 75, 5544-5547
46 Bishop, J. O., Morton, J. G., Rosbash, M. and Richardson, M. (1974) Nature 250, 199-204
47 Berk, A. J. and Sharp, P. A. (1977) Cell 12, 721-732
48 Lai, C.-J., Dhar, R. and Khoury, G. (1978) Cell 14, 971-982
49 Matthyssens, G. and Tonegawa, S. (1978) Nature 273, 763-765
50 Meyuhas, O. and Perry, R. P. (1979) Cell 16, 139-148
51 Kindle, K. L. and Firtel, R. A. (1978) Cell 15, 763-778
52 Tobin, S. L., Zulauf, E., Sanchez, F., Craig, E. A. and McCarthy, B. J. (1980) Cell 19, 121-131
53 Davidson, E. H. (1976) in Gene Expression in Early Development, Chapter 6, Academic Press
54 Perrin, F., Cochet, M., Gerlinger, P., Cami, B., LePennec, J. P. and Chambon, P. (1979) Nucl. Acids Res. 6, 2731-2748
55 Nordstorm, J. L., Roop, D. R., Tsai, M.-J. and O'Malley, B. W. (1979) Nature 278, 328-331
56 Nevins, J. R. and Darnell, J. E., Jr. (1978) Cell 15, 1477-1493
57 Tal, J., Ron, D., Tattersall, P., Bratosin, S. and Aloni, Y. (1979) Nature, 279, 649-651
58 Green, M. R., Lebovitz, R. M. and Roeder, R. G. (1979) Cell 17, 967-977
59 Schaffner, W., Kunz, G., Daetwyler, H., Telford, J., Smith, H. O. and Birnstiel, M. L. (1978) Cell 14, 655-671
60 Lifton, R. P., Goldberg, M. L., Karp, R. W., and Hogness, D. S. (1977) Cold Spring Harbor Symp. Quant. Biol. 42, 1047-1051