
No more than seven interruptions in the ovalbumin gene: comparison of genomic and double-stranded cDNA sequences

K.O'Hare, R.Breathnach, C.Benoist and P.Chambon

Laboratoire de Génétique Moléculaire des Eucaryotes du CNRS, Unité 184 de Biologie Moléculaire et de Génie Génétique de l'INSERM, Institut de Chimie Biologique, Faculté de Médecine, Strasbourg, France

Received 10 July 1979

ABSTRACT

We have determined the sequence of ovalbumin RNA (ov-mRNA) using a double-stranded cDNA (dscDNA) plasmid. We have also determined the sequence of the previously characterized exonic regions of the chicken ovalbumin gene. The comparison of these various sequences has shown that there are no additional interruptions in the mRNA-coding sequences above those 7 already characterized. There is only one single base discrepancy between the two mRNA sequences determined using the dscDNA or the genomic clones. This demonstrates the accuracy and reproducibility of the cloning and sequencing techniques. The ovalbumin mRNA sequence was found to be 1872 nucleotides in length, 13 nucleotides larger than the previous value reported by McReynolds et al. [Nature 273, 723-728 (1978)].

INTRODUCTION

The cloning of extensive sequences of ovalbumin double-stranded cDNA (dscDNA) has been reported by Humphries et al. (1). We have used this plasmid to show that the ovalbumin gene is split in chicken DNA (2) and to isolate genomic clones containing the 8 identified ovalbumin exons (3-6). These exons were mapped on the cloned fragments by electron microscopy and use of restriction enzymes. These techniques are not sufficiently sensitive to detect intervening sequences less than about 50 bp, however, and so the possibility remained that some or all of the exons might contain additional hitherto undetected interruptions. To investigate this possibility, we have sequenced the cloned double-stranded cDNA of Humphries et al. (1) and the regions of our ovalbumin genomic clones known to contain messenger-coding sequences. Comparison of these dscDNA and genomic clone sequences has shown that the 8 ovalbumin

exons defined by our previous work (6, 7, 8) are not further interrupted in the chicken genome. This allows us to present a definitive map of the chicken ovalbumin gene [Fig. 4(b)].

While this work was being completed, McReynolds et al. (9) reported the sequence of an independently constructed ovalbumin dscDNA plasmid. The sequences are in very good agreement - we detect, however, a sequence of 13 nucleotides in the 3' non-translated region of the messenger missed by McReynolds et al.

MATERIALS AND METHODS.

DNA fragments

Cellular DNA fragments shown in Fig. 4 were prepared as described in Breathnach et al. (8). The ovalbumin dscDNA insert in pCR1ov2.1 was prepared as an HhaI or HpaII ovalbumin-containing fragment (Fig. 2) by digestion with these enzymes and separation of the digest on 5-20% sucrose gradients (Hhaov or Hpaov).

DNA sequencing

The strategies used for DNA sequencing are shown in figures 2 and 4. End-labelling and sequencing using the Maxam-Gilbert (10) technique was as described as in Breathnach et al. (8). Electrophoresis of cleavage products was on 90 cm long, 2 mm thick, 8 or 20% polyacrylamide gels or on "thin" gels as described by Sanger and Coulson (11). Sources of material have already been described (8). Further details on the sequencing strategies and autoradiographs of the sequencing gels are available on request.

mRNA sequencing

Direct sequencing on ovalbumin mRNA was performed by elongation of a primer with reverse transcriptase in the presence of chain-terminating inhibitors (12). The RNA used as template was an oligo-dT cellulose purified oviduct RNA preparation kindly provided by A. Krust. Reverse transcriptase from avian myeloblastosis virus was from Dr. J. Beard.

To overlap the HgaI site at position 1125 (Figs 1 and 2) an MboII fragment defined by the sites at positions 1197 and

1229 was used as a primer (P in Fig. 2). (Here and throughout this paper the position given for a restriction enzyme site is the first (5') nucleotide of the recognition sequence and not the point of cutting). The primer was prepared from the cloned 2.6 kb Eco2-Hind1 fragment which contains the entirety of exon 7 (corresponding to nucleotides 830 to 1872 of the messenger) as this produced fewer small fragments than did Hhaov on digestion with MboII (our unpublished observations). The DNA was digested with MboII and electrophoresed on a non-denaturing, 10% polyacrylamide 0,35 mM thick gel (A.J.H. Smith, personal communication). The DNA was recovered by diffusion into 50 mM Tris HCl pH 8, 0.2 M NaCl, 10 mM EDTA overnight at 37°C and precipitated with ethanol in the presence of 50 µg of tRNA as carrier. An amount equivalent to 2.5 µg of the 2.6 kb fragment was used with 8 µg of RNA. Sequencing was exactly as described by Zimmern and Kaesberg (13) using their 2:2 mixtures of chain-terminating inhibitors.

Biohazards associated with the experiments described in this publication were examined previously by the French National Committee. The experiments were carried out under L3-B1 conditions (Le Progrès Scientifique N° 191, Nov/Dec 1977).

RESULTS AND DISCUSSION.

Ovalbumin mRNA sequence deduced from a cloned double-stranded cDNA.

The hybrid plasmid pCR1ov2.1 containing a double-stranded cDNA (dscDNA) copy of ovalbumin mRNA was constructed by Humphries et al. (1). Briefly, an oligo-dT primer hybridised to ovalbumin mRNA was extended by reverse transcription, the resulting cDNA made double-stranded, treated with S1, and the S1-treated double-stranded cDNA then tailed with dG residues before cloning into pCRI plasmid previously cut at the EcoRI site and tailed with dC residues. The resulting dscDNA clone pCR1ov2.1 contains the cDNA insert flanked by poly [dG:dC] tracts. We have determined the sequence of this insert using the technique of Maxam and Gilbert (10). The sequence obtained is shown in Fig. 1. The sites used for 5' end-labelling and the direction and extent of the sequences determined are shown in Fig. 2. Where possible,

Fig. 1 : Sequence of ovalbumin mRNA. The data shown is the result of sequencing of both ovalbumin dscDNA (see Fig. 2) and cellular exons (see Fig. 4). Nucleotides are numbered from the 5' end. The termination codon (UAA) and initiation codon (AUG) are boxed. Bracketed nucleotides were not present in the dscDNA, but were present in ovalbumin exons and are taken for the 5' end from McReynolds et al.(9), and Gannon et al.(6), and for the 3' end from Cheng et al.(14) and our sequence of the ovalbumin gene exon 7. Vertical arrows define the messenger nucleotides encoded by the leader-coding sequence and each of the ovalbumin gene exons. Two *HinfI* sites located around position 1470 and discussed in the text are identified. The ovalbumin dscDNA has a dA residue at position 626. The corresponding residue in the ovalbumin gene exon 5 is a dG. This is indicated on the figure by a G residue below the A residue at position 626.

the sequence was determined on both strands; otherwise several sequence determinations were made on the one strand. In this way more than 90% of the sequence we present has been confirmed either on the complementary strand or on the same strand from a different site. In addition, we have used the chain-terminator method of Sanger et al. (12) with reverse transcriptase for direct sequencing on the messenger RNA in order to overlap the

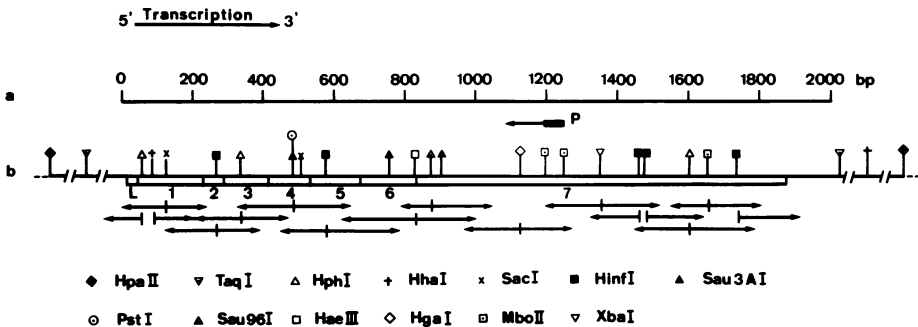


Fig. 2 : Strategy for sequencing ovalbumin dscDNA. a) scale, b) restriction map of an *HpaII* fragment containing the ovalbumin dscDNA insert in pCR1ov2.1. The restriction enzyme sites used for 5' end-labelling and the direction and extent of sequences read are indicated by arrows. Double lines correspond to the inserted sequence, and the light lines to the flanking plasmid sequences. L and 1-7 indicate areas of the dscDNA corresponding to the leader-coding sequence and the various exons of the cellular gene [see Fig. 4(b)]. P indicates an *MboII* primer used to overlap the *HgaI* site using direct priming on the messenger (see Materials and Methods).

HgaI site at position 1125 (Figs. 1, 2, and Materials and Methods), as this proved difficult to accomplish using the chemical degradation technique. Thus all restriction sites used for sequencing were overlapped. The accuracy of the sequence presented in Fig. 1 was further cross-checked by sequencing the ovalbumin exons on cloned genomic fragments (see below).

Extent of ovalbumin sequences in pCR1ov2.1

If the insert in pCR1ov2.1 contained a complete copy of the sequences at the 3' end of the messenger, there would have been a poly [dA:dT] tract (from the oligo-dT primer, see above) in the insert immediately before the poly [dG:dC] tract marking the end of the insert. As we did not find such a tract, we compared our sequence with that deduced by Cheng et al. (14) for the extreme 3' end of the ovalbumin messenger. In this way we could show that the insert in pCR1ov2.1 lacks only one coded nucleotide at the 3' end (shown bracketed in Fig. 1). This nucleotide, and the poly [dA:dT] tract, were presumably removed during the S1 nuclease treatment of the dscDNA prior to tailing and cloning (see above).

The extent of the 5' terminal messenger sequences missing in pCR1ov2.1 was similarly determined by comparison of the insert sequence with the sequence of the 5' end of the messenger (Fig. 3a). The 5' end of the messenger has been sequenced by reverse transcription of ovmRNA using a restriction enzyme fragment as primer by McReynolds et al. (9), and we have verified the sequence they obtained by sequencing on the corresponding genomic clone (6) (Fig.3b). It is clear from Fig. 3c that the first 14 nucleotides of the messenger are not represented in pCR1ov2.1, and they are therefore shown bracketed in Fig. 1.

We conclude that pCR1ov2.1. contains an ovalbumin dscDNA insert lacking the first 14 nucleotides from the 5' end of the messenger, and the last nucleotide at the 3' end of the messenger.

A peculiar sequence arrangement in the ovalbumin dscDNA plasmid, pCR1ov2.1

As shown in Fig. 3c, the insert in pCR1ov2.1 contains an extra 48 nucleotides between the end of the sequences correspond-

favoured in this instance by the conformation of the mRNA or cDNA. McReynolds et al. (9) have pointed out that the sequence at the 5' end of ovmRNA may be folded into a hairpin loop and it has not proved possible to 5' end-label ovmRNA after "decapping" under conditions where globin mRNA was efficiently labelled (16). It is also possible that events during the cloning and propagation of pCR1ov2.1 have contributed to the generation of this sequence arrangement.

Correspondence between mRNA sequence and ovalbumin exons.

Our previous studies (2-8) have shown that the sequences coding for ovalbumin mRNA are interrupted at least seven times in the chicken genome so that the gene consists of at least eight exons and seven introns [Fig. 4 (b)]. We have sequenced

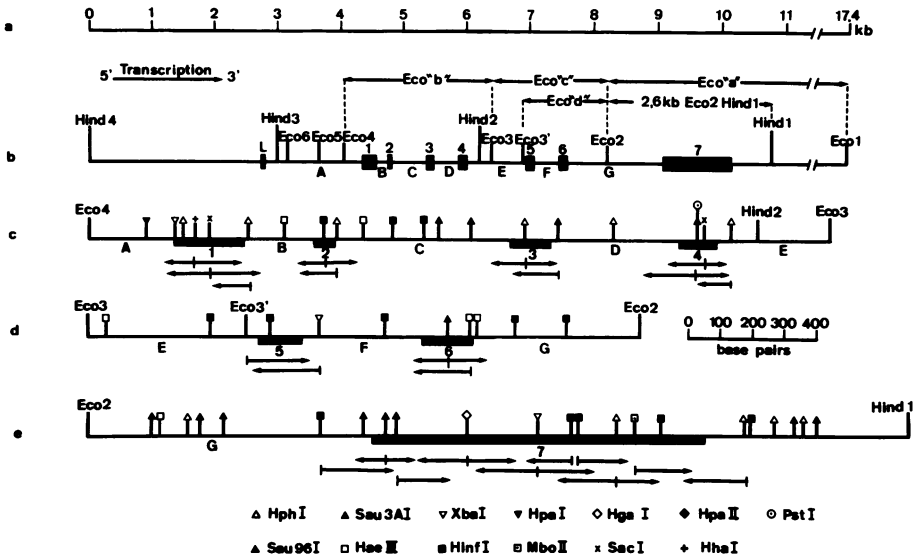


Fig. 4 : Strategy for sequencing the ovalbumin gene exons. (b): Restriction map of the cellular ovalbumin gene. The eight exons (L and 1-7) are shown by heavy lines (2-8). The location of EcoRI fragments "b" and "c" and the 2.6 kb EcoRI-HindIII fragment (Eco site 2-Hind site 1) used for sequencing are shown. (c), (d), and (e): Restriction maps of EcoRI fragments "b", "c", and the 2.6 kb EcoRI-HindIII fragment, respectively. Arrows indicate the restriction enzyme sites used for 5' end-labelling and the direction and extent of sequences read. Numbers identify ovalbumin exons (heavy lines) as in (b).

the limits of these exons previously, to determine the corresponding exon-intron junctions (8). It is, however, necessary to sequence the exons in their entirety to rule out the possibility of further small intervening sequences within them which might have been beyond the limits of detection of electron microscopy or restriction enzyme mapping. We have done this using the technique of Maxam and Gilbert (see Fig. 4, and Materials and Methods for sequencing strategy) on the appropriate genomic clones. The full sequence of the exon L was reported in ref. 6. The sequences obtained for the previously identified exons are exactly those predicted from the mRNA sequence presented in Fig. 1 (with one single nucleotide exception at position 626, see below). The RNA nucleotides encoded by the various exons are indicated on Fig. 1 by the vertical arrows, assuming that the GT-AG rule (8) for splicing applies to all of them. From this result we conclude that there are no additional interruptions in the hitherto characterised exons of the ovalbumin gene, and that the ovalbumin gene map shown in Fig. 4 (a) accounts for all the nucleotides of the messenger and is definitive.

Sequence heterogeneity in the ovalbumin gene and fidelity of the dscDNA synthesis/cloning procedure.

Three sequences are now available for ovalbumin mRNA : the one we have deduced from the cloned dscDNA present in pCR1ov2.1, that deduced by McReynolds et al. (9) from an independently cloned dscDNA, called pov230 (17) and the one we can deduce from the ovalbumin exon sequences. As each of these sequences represents messenger from different individual chickens (and for the data of McReynolds et al., a different breed) any differences could be due to sequence heterogeneity in the gene or to sequencing errors or could have arisen during the construction and propagation of the clones. There are in fact very few differences between the two dscDNA sequences and only one between our dscDNA and exon sequences. These are shown in Table 1.

Of the single base differences, that at position 79 creates an HhaI site in pCR1ov2.1 which is present in the appropriate genomic clone constructed by Garapin et al. (4) but is absent both from a genomic clone (our unpublished observations) from a bank constructed by Dodgson, Strommer and Engel (18) and from

Nucleotide number (Fig. 1)	Source of sequence		
	pCRlov2.1	Ovalbumin exons	pov230
34	C	C	G
43	A	A	G
79	C	C	T
223	G	G	A
626	A	G	G
1307	C	C	A
1456-1477	TGACTCAGTACTEAGTCAAT	TGACTCAGTACTEAGTCAAT	TGACTCAGTCAAT

Table 1 : Differences between 3 available sequences for ovalbumin mRNA. Data for pCRlov2.1 and ovalbumin exons are those presented in this paper. Data for pov230 are taken from McReynolds et al. (9). Nucleotides are numbered as in Fig. 1. The *Hinf*I sites at 1458 and 1471 boxed in this figure are shown in Fig. 1.

pov230 (9). Similarly the difference at position 43 creates a *Taq*I site in pov230 which is absent from pCRlov2.1 and our genomic clones (6). The remaining single base differences do not affect the pattern of restriction enzyme sites so that the only evidence for their existence is the sequencing data. However we are confident that our identification of nucleotides at positions 34, 223 and 1307 is correct for our messenger since we found the same nucleotide in the corresponding positions of the genomic exons. We have in addition confirmed our identification for nucleotides 34 and 43 by reverse transcription on ovmRNA using the chain terminator technique and an *Hha*I-*Sac*I restriction enzyme fragment as primer (see Fig. 1, data not shown). These experiments confirmed that there are no differences between our sequence for ovmRNA and that of McReynolds et al. (9) in the first 14 nucleotides (ie, those not present in pCRlov2.1) and also produced no evidence for heterogeneity at the 5' end of the messenger as has been found for SV40 mRNAs (19,20).

The differences at positions 79 and 223 are in the third base "wobble" position and do not change the amino-acid encoded.

The difference at positions 626 changes an alanine codon in the mRNA sequences derived from our genomic clones and from pov230 to a threonine codon in the mRNA sequence from pCR1ov2.1. We do not believe that this is a sequencing error but rather that in addition to the known variants of ovalbumin protein (21) there is one (as yet undescribed) with such an amino-acid substitution. It is also possible that this difference is due to an error by reverse transcriptase or DNA polymerase. Indeed recent estimates on the frequency of mis-matching by reverse transcriptase indicate that an error rate of one in 1872 nucleotides is not to be unexpected (22).

From the nature of the discrepancy between our sequences and that of McReynolds et al. around the *Hinf*I site at position 1458 it is clear that the difference is due to sequencing error rather than variance. McReynolds et al. did not overlap this *Hinf*I site and therefore did not realise that a second close *Hinf*I site exists at position 1471 (see Fig. 1 for the two sites and Fig. 5 for sequence). By sequencing outward in both directions from what they assumed to be a single site, they missed the 13 nucleotides between the two sites. In our experience, it is not possible to identify with certainty nucleotides very near the site of 5' end-labelling so that it becomes necessary to determine the sequence around each site by reading the sequence from another distant site.

Two of the single base differences between our ov-mRNA sequence and that of McReynolds et al. are in the 5' non-translated region, one in the 3' non-translated region and the remaining three in the coding region. It is perhaps surprising that there are two differences in the 64 nucleotide long 5' non-translated region which is thought to have a role in ribosome binding and only one in the 650 nucleotide long 3' non-translated region whose only role may be to link the stop codon with a poly A addition site. However, it is clear that there is not much more sequence heterogeneity in the non-translated regions than in the coding region.

In general our results testify to the remarkable fidelity of the techniques used to clone and propagate cDNA copies of RNA and eukaryotic DNA fragments in prokaryotes. However, it

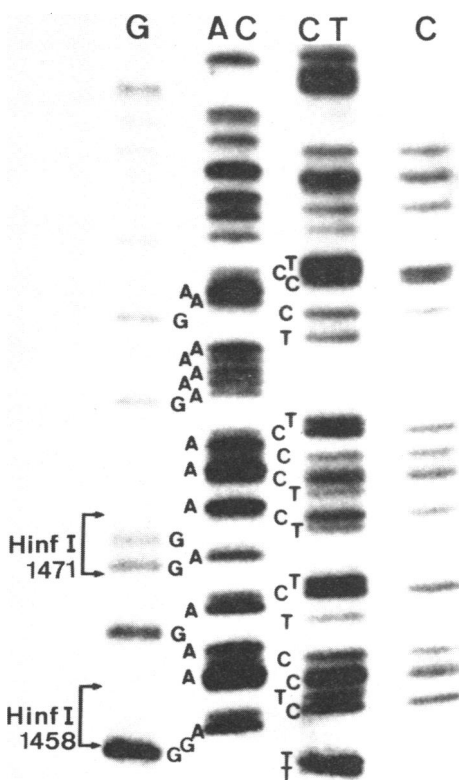


Fig. 5 : Sequence between the two Hinfl sites. The autoradiograph shows the sequence between the two Hinfl sites at position 1458 and 1471 as determined on an XbaI end-labelled fragment cut with TaqI (see Fig. 2).

is equally clear that peculiar sequence arrangements (5' end of the insert in pCRlov2.1), single base deletions (23) and perhaps single base changes (position 626 in pCRov2.1) may be generated. Our results also show how reproducible the modern sequencing techniques are.

ACKNOWLEDGEMENTS

We thank the Viral Cancer Program, National Cancer Institute (Dr. Beard) for gifts of avian myeloblastosis virus reverse transcriptase and A. Krust for a gift of ovalbumin mRNA. The technical assistance of M.C. Gesnel, A. King, J.M. Garnier, E. Taubert, A. Landmann is gratefully acknowledged. This work was supported by grants from the INSERM (CRT 76.5462 and 76.5.468), the CNRS (ATP 3558) and the Fondation pour la Recherche Médicale Française. K. O'Hare was supported by an EMBO long-term fellowship.

In accordance with the current policy of this journal concerning sequence papers, our complete data was made available to the Editors and reviewers, but is not presented.

REFERENCES

1. Humphries, P., Cochet, M., Krust, A., Gerlinger, P., Kourilsky, P., and Chambon, P. (1977) *Nucleic Acids Res.* **4**, 2389-2406.
2. Breathnach, R., Mandel, J.L., and Chambon, P. (1977) *Nature* **270**, 314-319.
3. Garapin, A.C., LePenec, J.P., Roskam, W., Perrin, F., Cami, B., Krust, A., Breathnach, R., Chambon, P., and Kourilsky, P. (1978) *Nature* **273**, 349-354.
4. Garapin, A.C., Cami, B., Roskam, W., Kourilsky, P., LePenec, J.P., Perrin, F., Gerlinger, P., Cochet, M., and Chambon, P. (1978) *Cell* **14**, 629-638.
5. LePenec, J.P., Baldacci, P., Perrin, F., Cami, B., Gerlinger, P., Krust, A., Kourilsky, P., and Chambon, P. (1978) *Nucleic Acids Res.* **5**, 4547-4562.
6. Gannon, F., O'Hare K., Perrin, F., LePenec, J.P., Benoist, C., Cochet, M., Breathnach, R., Royal, A., Garapin, A., Cami, B., and Chambon, P. (1979) *Nature* **278**, 428-434.
7. Mandel, J.L., Breathnach, R., Gerlinger, P., LeMour, M., Gannon, F., and Chambon, P. (1978) *Cell* **14**, 641-653.
8. Breathnach, R., Benoist, C., O'Hare, K., Gannon, F., and Chambon, P. (1978) *Proc. Natl. Acad. Sci.* **75**, 4853-4857.
9. McReynolds, L., O'Malley, B.W., Nisbet, A.D., Fothergill, J.E., Gibvol, D., Fields, S., Robertson, M., and Brownlee, G.G. (1978) *Nature* **273**, 723-728.
10. Maxam, A., and Gilbert, W. (1977) *Proc. Natl. Acad. Sci.* **74**, 560-564.
11. Sanger, F., and Coulson, A.R. (1977) *FEBS Lett.* **87**, 107-110.
12. Sanger, F., Nicklen, S., and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci.* **74**, 5463-5467.
13. Zimmern, D., and Kaesberg, P. (1978) *Proc. Natl. Acad. Sci.* **75**, 4257-4261.
14. Cheng, C.C., Brownlee, G.G., Carey, N.H., Doel, M.T., Gillam, S., and Smith, M. (1976). *J. Mol. Biol.* **107**, 527-547.
15. Ghosal, D., and Saedler, H. (1978) *Nature* **275**, 611-617.
16. Chu, L.Y., Lockard, R.E., RajBhandary, U.L., and Rhoads, R.E. (1978) *J. Biol. Chem.* **253**, 5228-5231.
17. McReynolds, L.A., Catterall, J.F., and O'Malley, B.W. (1977) *Gene* **2**, 217-231.
18. Dodgson, J.B., Strommer, J., and Engel, J.D. (1979) *Cell*, in press.
19. Ghosh, P.K., Reddy, V.B., Swinscoe, J., Lebowitz, P., and Weissman, S.M. (1978) *J. Mol. Biol.* **126**, 813-846.
20. Reddy, V.B., Ghosh, P.K., Lebowitz, P., and Weissman, S.M. (1978) *Nucleic Acids Res.* **5**, 4195-4213.
21. Wiseman, R.L., Fothergill, J.E., and Fothergill, L.A. (1972) *Biochem. J.* **127**, 775-780.
22. Gropinathan, K.P., Weymouth, L.A., Kunkel, T.A., and Loeb, L.A. (1979) *Nature* **278**, 857-859.

23. Browne, J., Paddock, G.V., Liu, A., Clarke, P., Heindell, H.C., and Salser, W. (1976). *Science* 195, 389-391.