Volume 10 Number 19 1982 Nucleic Acids Research

Organisation of feather keratin genes in the chick genome

P.L.Molloy*, B.C.Powell, K.Gregg, E.D.Barone and G.E.Rogers

Department of Biochemistry, University of Adelaide, Adelaide, South Australia, Australia 5001

Received 28 May 1982; Revised and Accepted 8 September 1982

SUMMARY

A genomic clone containing sequences of five feather keratin genes has been isolated using cDNA to chicken embryonic feather keratin mRNA as a probe. The clone probably represents part of a longer cluster of tandemly spaced genes; the genes are evenly spaced with a centre to centre separations of 3.3 kb and are transcribed from the same DNA strand, suggesting that the cluster has arisen by a series of tandem duplications. The organisation and complete sequence of the central gene has been determined. The protein encoded by the gene contains 97 amino acids and its sequence is typical of proteins of the embryonic and adult feather family. The transcript from the gene has been deduced to contain a long 3' non-coding region of 435 nucleotides and a 58 base 5' non-coding region interrupted in the gene 21 bases prior to the initiation codon by the gene's only intron of 324 bases.

INTRODUCTION

The major protein synthesised in feather and other epidermal cells of birds is the intracellular disulphide-linked protein, keratin. In the chick embryo, at about day 13, a rapid increase in keratin synthesis begins, being maximal at day 14 to 15; at day 21 keratin is essentially the only protein found in the feather (1). Analysis of the proteins of the embryonic feather has resolved the keratins into a family of at least 20 homologous proteins of molecular weight about 10,000. Amino and carboxy-terminal sequences have been determined for some of these species and are consistent with the presence of multiple keratin proteins differing from each other by one or a few amino acid substitutions (2).

As a gene family the keratin system has a number of features of particular interest. It comprises a relatively large number of genes which are both related in an evolutionary sense and co-ordinately expressed. The conservation through evolution of non-coding sequences which may be important in control of gene expression or involved, for example, in protein-RNA and protein-DNA interaction can be examined in a larger sample

0305-1048/82/1019-6007$ 2.00/0

of genes than is possible with most gene families. An additional major advantage is that the embryonic feather family is one of a number of related gene families, e.g., the adult feather keratins and scale, beak and claw keratins, all of which are believed to share common ancestral genes but which are expressed in different tissues and at different times during development.

A fraction of mRNA has been isolated, from the polysomes of 14-day embryonic chick feather, which upon translation in cell free systems yields the keratin family of proteins as the predominant product (3). The mRNA is about 800 nucleotides long and studies on kinetics of hybridization and reannealing indicate that there are about 30 different keratin mRNA species present. The hybridization studies indicated that each mRNA contains a repetitive region (presumably the coding region) and unique sequences (4). Complementary DNA to this feather keratin mRNA has been used to construct an mRNA library (5) and here we report the use of this mRNA fraction to isolate embryonic feather keratin genes from a chick genomic library.

EXPERIMENTAL METHODS

Isolation of Genomic Clones

Keratin genomic clones were selected from a λ Charon 4A chicken library (prepared and kindly donated by J. Dodgson, J. Engel and R. Axel) using embryonic feather keratin cDNA under conditions of moderate stringency (0.5 x SSC, 0.1% SDS at 65°). Approximately 100,000 phage recombinants were screened by the technique of Benton and Davis (6). Prehybridization and hybridization conditions were as described by Wahl et al. (7). About 100 plaques bearing keratin sequences were detected and 8 were plaque-purified. Phage, prepared either by a modification of the PDS method of Blattner et al. (8) or from plate lysates, were banded in CsCl, dialysed extensively against 10 mM Tris-HCl (pH 7.5), 1 mM EDTA and repeatedly extracted with phenol-chloroform. The phage DNA was finally purified from a 10-40% sucrose gradient.

Subcloning into Plasmid Vectors

Chimeric plasmids containing regions of λCFK1 were obtained by ligation of a total HindIII digest of λCFK1 into pBR322 to generate the subclones pλCFK1-4, 5A, 5B, 6, 7 and 8 containing insert fragments of size 2.94, 2.15, 1.84, 1.64 and 1.28 kb respectively. The 3.2 kb HindIII/EcoRI fragment was eluted from a low melting point agarose gel and ligated into

pBR322 to generate the subclone pλCFK1-2. Plasmid DNAs were isolated by a procedure involving alkali-SDS lysis and selective precipitation with polyethylene glycol 6000, essentially similar to that described by Birnboim and Doly (9).

R-loop Mapping

Keratin mRNA was isolated as described by Kemp et al. (1). Plasmid DNAs were linearised with an appropriate restriction enzyme and ethanol precipitated prior to hybridization. Hybridization was performed in a volume of 10 µl of buffer containing 60% formamide, 80 mM PIPES (pH 8.5), 400 mM NaCl, 5 mM EDTA, 1.1 µg/ml of DNA and 20 µg/ml of feather keratin mRNA. Incubation in a sealed capillary was for 3 minutes at 80°, 12 h at 60° and finally 30 minutes at 30°. Three µl of the hybridization mix was added to 12 µl of a stock hyperphase solution consisting of 72.5% formamide, 120 mM Tris-HCl (pH 8.5), 12 mM EDTA, 10 mM NaCl, 62.5 µg/ml cytochrome C, 0.2 µg/ml pBR322 DNA and 0.1 µg/ml M13 DNA. This was then spread on a hypophase of 40% formamide in water. Spreads were transferred to parlodion-coated grids, stained with uranyl acetate and shadowed with Pt/Pd prior to visualisation in a Siemens Elmiscop 102 electron microscope. Length measurements were standardised using pBR322 and M13 for the double and single stranded regions respectively. Measurements from 37 molecules were considered in determining the length and spacing of λCFK1 genes.

Restriction Enzymes and Digestions

Restriction enzyme (New England Biolabs and Boehringer-Mannheim) digests of DNA were performed under standard conditions. Mapping of restriction sites was done by a combination of single and double enzyme digests and partial enzyme digests.

Labelling of Nucleic Acids

[^{32}P]cDNA to embryonic feather mRNA was prepared as previously described (4). Nick-translated DNA was prepared as described by Rigby et al. (10) using α-[^{32}P]dGTP. End-labelling of DNA restriction fragments was done by adding 1 unit of the Klenow fragment of DNA polymerase I (Boehringer) to restriction digests of 5 µg DNA in 300 µl containing 30-50 µCi of [^{32}P]-dNTP's and incubating at 37° for 30 minutes followed by a chase for 10 minutes with 50 µM unlabelled nucleotides. For 5' end-labelling DNA restriction digests were treated with calf intestinal phosphatase (Boehringer) and individual fragments eluted from gels; labelling was as described by Maxam and Gilbert (11). γ-[^{32}P]ATP,

α-[^{32}P]dCTP and α-[^{32}P]dGTP were prepared (12) at specific activities of 2,000, 500 and 500 Ci/mmol respectively.

DNA Sequencing

DNA sequencing of end-labelled DNA fragments was as described by Maxam and Gilbert (11) using the G, G+A, T+C and C reactions and in some cases the A > C reaction. In most cases the T specific $KMnO_4$ reaction of Rubin and Schmid (13) was also used. Most of the sequence data relies on more than one sequencing of one or both strands of a region and predicted restriction sites have been checked by digestion with an appropriate enzyme. Sequence of two regions (see Fig. 4) was determined by cloning of fragments into M13mp83 (14; J. Messing, unpublished) and sequencing by the chain termination procedure using a synthetic 17 base primer (Collaborative Research).

Gel Transfers and Hybridizations

Transfers from agarose gels and hybridization with cDNA or nick-translated probes were as described by Wahl et al. (7). Transfer of DNA from acrylamide gels was done as described by Smith and Summers (15).

Primer Extension on RNA

End-labelled DNA fragments were hybridized with 0.5 μg of keratin mRNA in 15 μl of 1 M NaCl, 0.5 mM EDTA, at 65° for 3 h. This hybridization mix was diluted into 200 μl of a solution of 20 mM Tris-HCl (pH 8), 8 mM $MgCl_2$, 10 mM dithiothreitol and 400 μM of each deoxynucleotide, 5 units of reverse transcriptase was added and the reaction incubated at 37° for 45 min. For sequencing reactions 7 μl of an annealing mix was diluted into 100 μl of a solution containing 20 mM Tris-HCl (pH 8), 8 mM $MgCl_2$, 10 mM dithiothreitol and 20 μM of each deoxynucleotide triphosphate. Of this mix 25 μl was added to each dideoxynucleotide triphosphate to give final concentrations of 50 μM ddGTP, 250 μM ddATP, 80 μM ddTTP and 40 μM ddCTP, 1 unit of reverse transcriptase was added to each tube and the reaction was incubated at 37° for 30 min. After ethanol precipitation, products were electrophoresed on 5% polyacrylamide gels in the presence of 7 M urea.

RESULTS

Isolation of Genomic Clones

Screening of a portion of a λ Charon 4A chicken library with embryonic feather keratin cDNA yielded about 100 positive recombinant phage. However, multiple isolates of some clones were obtained, and of 8 which were plaque-purified only 3 were different. These designated λCFK1, 5 and

7 (λ Charon 4A-Chicken-Feather-Keratin) contained inserts of 15.5 kb, 12.5 kb and 17 kb respectively. Restriction enzyme mapping and heteroduplex analyses (data not shown) suggested that λCFK5 may have derived from λCFK1 through an internal recombination event, during phage propagation, leading to deletion of part of the DNA. This type of event has been observed in mouse immunoglobulin genomic clones (16) and in α-globin genomic clones (17) in which there are extensive regions of homology. The present report details the characterization of λCFK1.

Identification of Coding Regions

In order to identify the regions specifying keratin sequences, λCFK1 was digested with HindIII and Southern blot hybridization carried out using cDNA to embryonic feather mRNA (Fig. 1). The cDNA hybridized to six of the seven HindIII fragments containing cloned sequences (see map in Fig. 3), indicating that the coding regions are distributed widely throughout the clone. That the genes are indeed keratin-coding was confirmed by the hybridization to the HindIII digest of λCFK1 of a nick-translated fragment of the sequenced cDNA clone pCFK23 (5) containing only protein-coding sequences. Three of the HindIII bands, which correspond to 4 fragments since the 2.15 kb band is a doublet, hybridized to this probe (Fig. 1).

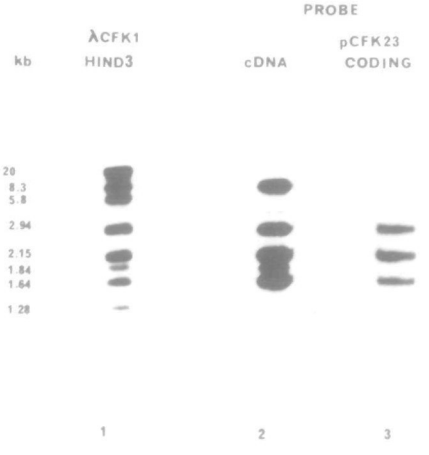


Fig. 1. Hybridization of keratin sequences to a digest of λCFK1. λCFK1 was digested with HindIII, the fragments separated on a 1% agarose gel and transferred to nitrocellulose. Track 1 contains nick-translated λCFK1 digested with HindIII as markers; track 2 was hybridized with feather keratin cDNA and track 3 with a nick-translated segment of the keratin cDNA clone, pCFK23, containing only keratin-coding sequences.

Clearer identification of the number of keratin genes and their location was obtained by R-loop mapping. Keratin mRNA annealed with λCFK1 was visualized by electron microscopy and in a high proportion of molecules five R-loops were visible (Fig. 2). The average size of the hybrid region for genes A to D was 600 bp, somewhat shorter than the observed message size. It is not clear whether this is an artifact of the spreading or is due to different extents of homology with the non-coding parts of the genes. The R-loop for the gene labelled E was consistently shorter than for the other genes; sequencing of this region indicated that the 3' end only of the gene is present and that it abuts the end of the cloned DNA fragment (K. Gregg, unpublished results). The size of the other four R-loops was consistent with each representing a separate gene; also noteworthy is the fact that the hybrid regions did not appear to contain any looped-out single-stranded DNA indicating an absence of intervening sequences within the annealed region.

Complete or partial sequence analysis of the five genes, A, B, C, D and E (K. Gregg and P. Molloy, unpublished data) showed that all five share

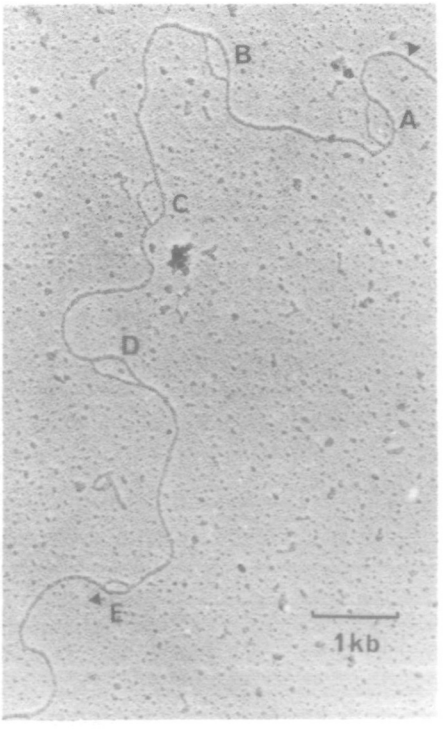


Fig. 2. R-loop formation between λCFK1 and embryonic feather mRNA. The boundaries between the vector arms and the inserted DNA are indicated by the arrows, and the R-loops of the five genes by A, B, C, D and E. The scale represents 1 kb.

the same orientation and are transcribed as indicated in Fig. 3. The clone λCFK1, therefore, appears to represent a segment of a tandemly repeating cluster of feather keratin genes transcribed from the same DNA strand and spaced about 3.3 kb centre to centre.

Gene C Codes for a Protein of 97 Amino Acids

A detailed analysis has been made of gene C, the central gene in the clone, including its complete sequence. The sequencing strategy and the gene sequence are shown in Fig. 4 and the different regions of the gene are considered below.

The sequence of the gene C region of λCFK1 revealed a reading frame encoding a protein of 97 amino acids (plus the initiating methionine) unbroken by any intervening sequences. The size of the predicted protein is in close agreement with the observed size (M.W. 10,000) of embryonic feather keratin and in Fig. 5 its sequence is compared with previously determined sequences of amino and carboxyl terminal peptides from different embryonic feather keratin proteins (2). In the amino terminal region only a single amino acid substitution, Asn → Asp is seen when this terminal region is compared with the sequence of an N-terminal peptide of 36 residues (Fig. 5). More differences are found near the carboxyl terminus, in agreement with the greater level of substitution found among the sequenced peptides in this region. The sequence of the gene C product is not identical to any of the sequenced peptides but it is clearly part of a closely related family of sequences. For example, comparison with the entire length of a mRNA-derived embryonic feather keratin clone, pCFK23, (5) shows a total of seven amino acid substitutions in 77 residues, six of which reflect a single nucleotide change (Fig. 5). It is quite possible that differences near the end of the cDNA clone, especially the Gln → Lys,

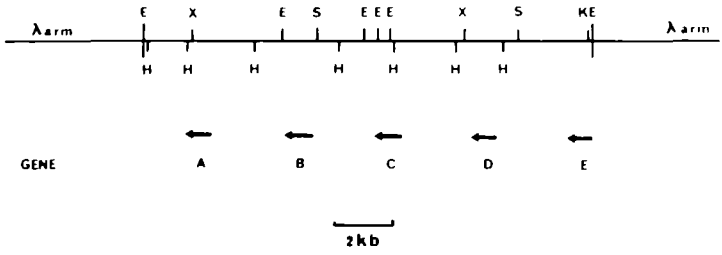
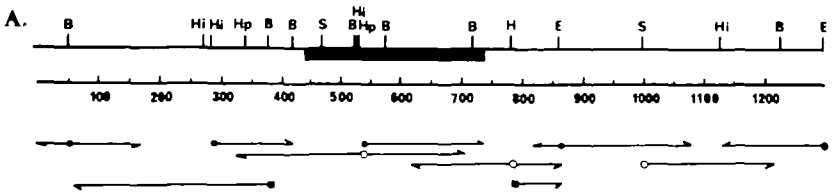


Fig. 3. Map of the chicken genomic clone λCFK1. The restriction sites indicated are E, EcoRI; H, HindIII; X, XbaI; S, SalI; K, KpnI. Below the map is shown the location of each of the five genes, the arrows indicating their direction of transcription, 5' to 3'. The scale represents 1 kb.



B.

50 100
GCTTGCCGCT TGCCTACTAT AGTTACATAT GCATAAATTA ACTCTAAACC AGGCTCCCTC ATCCACTTCT CTTGCCTTCT CCTCCTTGGT GAACAAGGTG

150 200
AGCTGCTGTG GCTTTCTTCT CACTCTTGCT CTGCTTCTCC TCTTGCTCTT TGTGCCTCCA TTAAGTCTTT GCCTTTGTGC AGACCTTGGT ATGAGCCTTG

250 300
CTCCTTGTTG TGCTCCCACC AGCTCTGTTC TGCTTGCCAG TGTGAGGTGT GGTGGGAGAA GGCCTTTGGC TGACTCTACT GGTGATTCTG GGGATCTGGG

350 400
CTCGCCAAGC CTCTGTCCCT GCTCCGCCTG TCCCTGATCC CCGGCCTGTG CCTCCTGTCT TGGCAATCAG GCCACCAGGC TACTTCTCAC GCAGCCTGTG

450 500
CTTTTCTTGT CCTCTCTCCA GGTCTACTCC CATCCTACAG CCATGTCCTG CTTCGATCTG TGCCGTCCCT GTGGCCCGAC CCCACTGGCC AACAGCTGCA

550 600
ACGAGCCCTG TGTGCGCCAG TGCCAGGACT CCCGGGTGGT GATCCAGCCC TCTCCCGTGG TGGTGACCCT GCCTGGACCC ATCCTCAGCT CCTTCCCCCA

650 700
GAACACCGCT GCGGGCTCCA GCACCTCTGC TGCTGTTGGC AGTATCCTGA GTGAGGAGGG AGTGCCCATC TCCTCTGGTG GCTTTGGCAT CTCTGGCCTG

750 800
GGTAGCCGCT TCTCTGGCAG GAGGTGTCTG CCCTGCTAAG GACGAGGTGT TCATCCCATG GATGCATCCT CAGGAAACCC AAAGCTTGGT GCTGGACTGC

850 900
TGACTGAGCT TCTGAGCAGG ATCCACTGAG CACCCTCCTG CTCTCCTGCA AAGCAAAGAG GGAATTCAAG TTGCCAGCCT GTGCTGCCTG TAGACACAGA

950 1000
CAGCAGCTGT CTTCTTCTTT TCCTTCTTTC TTTCATCATT AGGGGTTCTT GTTGTGTCCT TTGTCCTGTG CCCTGGGTTC ATCCTGAAGC AAGTTGAGAT

1050 1100
GGCCCTGCTT CTTCTTCCAC TTGTCTTGTG ATGGGGAAGA CATGCATCCC ATCTTCCTGT AGTTTCCCTC CTTATGGCCA ATATGTTTGC CAGCTGTATT

1150 1200
TGTAGCAGCC TTTAACTTTC ACTTTTTTGA TTCCTTCTTT GAGCTCAATA AAATTTATGC TGCATTGTAA TCTCAGTCTC CTCATGTTTC TTGCCTCATA

Fig. 4. Nucleotide sequence of gene C.
(a) Restriction map and sequencing strategy. The map shows the cleavage sites of those enzymes which were used in sequencing the region. The region to the left of the HindIII site is contained in the subclone pλCFK1-5B, and that to the right in pλCFK1-6. B, BstNI; Hi, HinfI; Hp, HpaII; S, Sau96I; H, HindIII; E, EcoRI. The blocked region is that coding for the keratin protein. The arrows indicate the direction and length of sequence obtained from a particular restriction site; those beginning with represent fragments 5' end-labelled, 3' end-labelled and fragments sequenced by subcloning into M13mp83.
(b) Sequence of gene C. The bases in large type are those which appear in the final mRNA, while those in small type correspond to the flanking and intervening sequences. Sequences shown boxed are the promoter sequence (32-38), the mRNA start site (61), the 5' and 3' splice sites (95-100 and 419-423 respectively), the initiation codon (443), the termination codon (736) and the polyadenylation signal (1147-1152).

are caused by cloning aberrations because lysine is an amino acid not detected in feather keratins. The codon usage of gene C shows a strong bias toward the use of C (50 times) over G and U (21 and 25) and particularly A (4) as the third base of a codon. This bias, which also occurs in the mRNA clone, pCFK23, resembles that seen for the chicken globin gene, though differing in some details, but differs dramatically from codon usage in another chicken gene, ovalbumin, in which all four bases are used with equal frequency (18). Codon usage in various mRNAs has been observed to differ widely and is apparently specific to genome or cell type (18), these differences possibly reflect tissue or species specific differences in tRNA populations or requirements for mRNA secondary structure and may be important in the regulation of translation.

3' Non-coding Region of Gene C

The 3' end of the mRNA region of gene C can be identified by

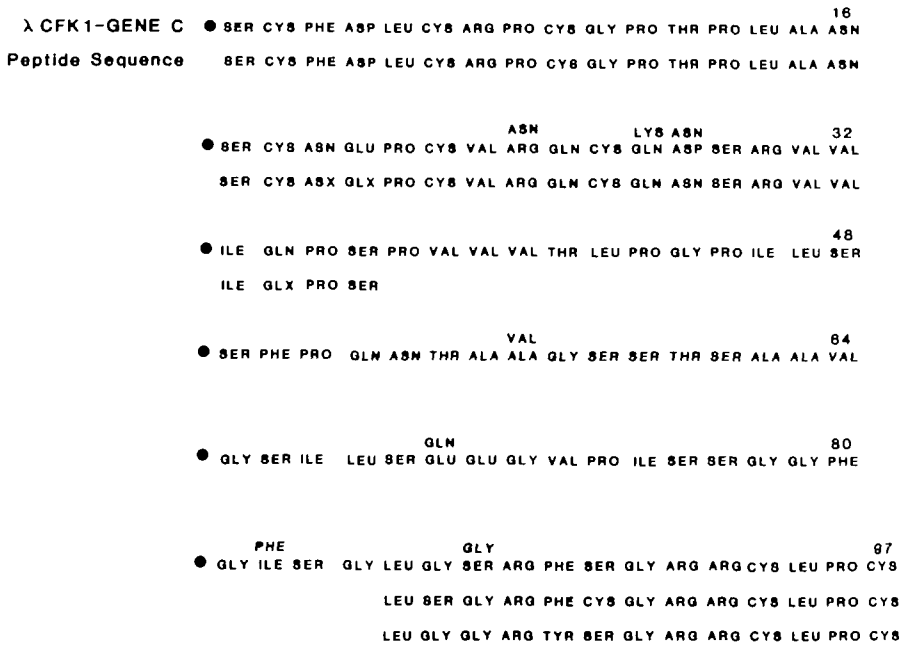


Fig. 5. Protein coded by gene C. The amino acid sequence of the protein coded by gene C corresponds to those lines indicated by . The amino acids shown above this line are the differences seen in the protein coded by the mRNA clone pCFK23; the sequences shown below the line are those of the sequenced amino terminal peptide T1b, and two carboxyl terminal peptides a and b of feather keratins (2). Glx and Asx are residues in which differentiation between Glu and Gln and Asp and Asn respectively was not made.

comparison of its sequence with that of the mRNA clone, pCFK23, (5). Both contain the AAUAAA motif required for polyadenylation (19) and the sequence of 22 nucleotides between this and the polyadenylate tail of pCFK23 is exactly reproduced in the gene C sequence. The 3' non-coding region of gene C is therefore 435 nucleotides long, again very similar to pCFK23, and sequence comparison over the entire length of this region shows an average of about 60% homology. This includes two blocks of about 70 nucleotides, one adjacent to the 3' end of the message, which are highly conserved and another block of sequence which exhibits no significant homology. The highly conserved nature of these two blocks of sequence indicates a functional selection for their retention equivalent to the selective pressure on the protein coding region.

The 5' Non-coding Region of Gene C Contains an Intron

The length of keratin mRNA has been previously estimated to be about 800 nucleotides (1), so given the length of the coding and 3' non-coding sequences of gene C (altogether 732 nucleotides) a short 5' non-coding region could be expected. Examination of the sequence 5' to the coding region of gene C did not reveal any sequence comparable to the "TATA box" sequence normally found prior to the site of transcription initiation and implicated in the binding of RNA polymerase (20). To determine which regions 5' to the coding region are present in the keratin mRNA, cDNA to keratin mRNA was hybridized to digests of the subclone, pλCFK1-5B, which had been transferred to nitrocellulose (data not shown). This indicated that sequences transcribed into mRNA are found in the region between 180 and 390 bases 5' to the initiation codon. Thus the 5' non-coding region must be longer than previously believed or must be interrupted by an intervening sequence.

To determine the length of the 5' non-coding region and to test for the presence of an intron, the 243 bp, HinfI fragment which overlaps the AUG initiation codon was isolated, recut with the enzymes BstNI and Sau3AI, and the kinased subfragments used to prime cDNA synthesis on keratin mRNA. The BstNI site partially overlaps the HinfI site and is poorly cleaved in the secondary digestion. Thus these subfragments share the same HinfI end. The products of priming on mRNA with the 72 bp Sau3A fragment and the 110 bp BstNI fragment (respectively bases 459-530 and 421-530 of the priming strand in Fig. 4) are shown in Fig. 6. The run-off cDNA products are in both cases 148 bases long, indicating a 5' non-coding region of about 58 bases. In addition none of the HaeIII, BstNI or HinfI

subfragments of the 243 bp HinfI fragment which extend beyond (to the 5' side) the 110 bp BstNI fragment were able to prime cDNA synthesis on the keratin mRNA (data not shown). Taken together with the results of hybridization to cDNA, these results suggest the presence of an intron in the 5' sequence. This is confirmed by comparison of the DNA sequence of gene C with the sequence derived when dideoxynucleotides are included in the primer extension of the 110 bp BstNI fragment (Fig. 6). The sequence derived from the mRNA differs from the genomic DNA sequence found adjacent to the 3' end of the primer but is identical to that of the DNA sequence between residues 61 and 97 (Fig. 4), though the level of dideoxynucleotide has prevented the reading of the last 5 U residues of the sequence and the strong primer band masks the first few residues. This defines the structure of the gene as consisting of a 5' leader sequence of 37 bases which is spliced to a point separated from it by 324 bases and preceding the initiation codon by 21 bases. The following features evident in the sequence (Fig. 4) are consistent with this interpretation:

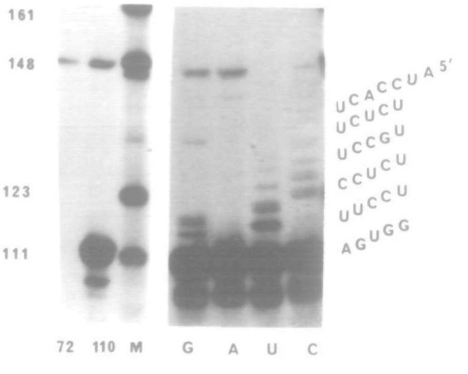


Fig. 6. The 5' end of feather keratin mRNA. Restriction fragments of pλCFK1-5B extending from the HinfI site in the coding region of the gene (cleavage site between 530 and 531 on opposite strand of that shown in Fig. 4) to either the Sau3AI site (cleavage site between bases 458-459), giving a 72 base primer, or the BstNI site (cleavage between bases 420 and 421), giving a 110 base primer, were kinased and annealed to embryonic feather mRNA. Extension of the primers was carried out in the absence of dideoxynucleotides, tracks 72 and 110 respectively for the 72 and 110 base primers. Track M, end-labelled HpaII fragments of pBR322 as markers, the base numbers on the left refer to the lengths of the end-filled markers. Tracks G, A, U and C are extensions of the 110 base primer in the presence of ddCTP, ddTTP, ddATP and ddGTP yielding the complementary base sequences for the RNA template. The sequence of the RNA is shown to the right, the presence of the last 5 U's being inferred from the spacing of the bands and comparison with the DNA sequence. Products were analysed on a 5% urea-acrylamide gel.

1. The site of initiation of mRNA synthesis is an A residue in a pyrimidine-rich region, TCCCTC'ATCC, as seen for a number of other genes (21).
2. The 5' and 3' splice sites AAG/GTGAG and CAG/GT respectively are in close accord with the consensus sequences derived from a number of genes (21) and the position of the splice sites is that predicted by the primer extension experiments.
3. The transcription start site is preceded 29 bp 5' by the sequence CATAAAT, which differs from the consensus promoter sequence only in that the first base is a C rather than a T (21). This same change has been observed in other genes, e.g., rabbit globin (22). It is of interest that other sequences similar to the promoter sequence, TATAGTT and TACATAT, are found nearby at positions 43 and 37 bases prior to the mRNA start site.

Thus the overall gene structure for gene C, shown in Fig. 7, consists of a 58 base 5' non-coding region which is interrupted by an intervening sequence of 324 bases, a continuous protein-coding region of 297 bases and a 3' non-coding region of 435 bases giving the gene a total length of about 1100 bases. Failure to detect the presence of the intervening sequence in R-loop experiments is probably due to the short length, 37 bases, of the 5' exon which was probably not capable of forming a stable hybrid under the spreading conditions.

DISCUSSION

Using cDNA to embryonic feather mRNA as a probe, a genomic clone containing five keratin genes has been isolated and characterised. The protein sequence coded by gene C of λCFK1 shows strong homology to sequenced carboxyl and amino terminal peptides of keratins isolated from embryonic feather, while lacking features evident in scale-specific keratins (23). The gene thus belongs to the homologous family of feather keratins, but lack of protein sequence data for feather keratins expressed

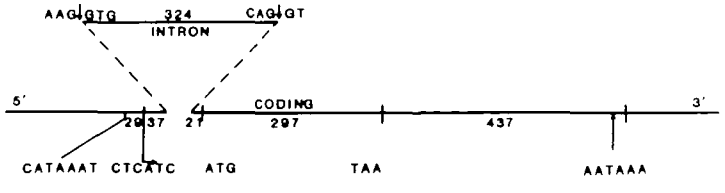


Fig. 7. The structure of gene C. The overall structure of gene C, with the lengths of the various segments and recognised signal sequences is shown.

at different stages of development does not permit us to assign the protein to a particular developmental period.

At the nucleic acid sequence level, the gene C sequence can be compared to that of two cDNA clones, pCFK17 and pCFK23 (5), obtained from embryonic feather mRNA. The homology observed in the 3' non-coding between gene C and either cDNA clone is as great as between the two cDNA clones, with regions of strong conservation of sequence being the same for all three genes. The use of a restriction fragment from gene C to prime cDNA synthesis on embryonic feather mRNA demonstrated that mRNAs with 5' terminal sequences identical to that of gene C are present in the embryonic feather, and this makes it probable that gene C is expressed in this tissue.

The dominant feature of the cluster of genes isolated in λCFK1 is their regular, tandem spacing with a centre to centre separation of the genes of 3.3 kb. Although a regular spacing of the genes is maintained, preliminary hybridization analysis suggests there is little nucleotide sequence homology within the spacer regions (data not shown). Given the estimated number of about 30 embryonic feather keratin genes it is probable that λCFK1 represents a segment of a longer tandem array of genes, though current data do not exclude the existence of more, smaller clusters of genes.

The arrangement of the cluster suggests that it evolved through a series of tandem duplications. It will be interesting to examine the detailed pattern of nucleotide sequence and evolution in view of the models suggested by studies on globin genes (22). The relationship of these genes to the others of the embryonic family and to the ancestrally related adult feather, scale, beak, claw and epidermis families is of great interest as an extreme example of an apparent primordial gene which has undergone multiple duplications.

The structure of gene C is particularly interesting as it has a single intron which is present in the 5' non-coding region, 21 bases prior to the AUG codon. Preliminary results (K. Gregg and P. Molloy, unpublished data) suggest that all genes of the cluster share the feature. The occurrence of the intron is therefore not related to the separation of functional protein domains brought together during evolution to form a single protein, a suggested rationale for the positions of a number of introns in eukaryotic genes (24). It is possible that the concept could be expanded to include the bringing together of functional blocks of DNA sequences, coding or non-

coding, into composite genes during evolution. Subsequent loss of introns from non-coding regions should occur more easily as small alterations in the non-coding region are likely to be less deleterious than in the coding region.

A more likely explanation for the presence of the intron is that the same gene may be expressed in a different tissue (e.g., scale or claw) or at a different stage of development (adult feather) using a different promotor region and 5' leader sequence. This would be analogous to the expression of the mouse α-amylase gene in the liver and salivary gland where different primary transcripts of the same gene are spliced to give mRNAs differing only in their 5' leader sequences (25). Numerous examples of differential splicing of 5' ends of viral mRNAs have been demonstrated (26) and it is probable that similar strategies are used as the control of eukaryote genes. If this is the case for feather keratin genes we would expect to find within the intergenic spacer regions alternative promoters and 5' leader sequences.

ACKNOWLEDGMENTS

We wish to thank Dr. Trevor Lockett for his help in the initial steps of the study and Dr. R.H. Symons for ^{32}P-nucleotides. Able technical assistance was given by Mrs. Lesley Crocker and Michael Calder. All cloning work was carried out under C_3 containment conditions according to rules laid down by the Australian Scientific Committee on Recombinant DNA. We are grateful to the Australian Research Grants Committee for the financial support of this project. One of us (P.L. Molloy) was on secondment from the Molecular and Cellular Biology Unit, C.S.I.R.O., Sydney, during the completion of the study.

*Present address. Molecular and Cellular Biology Unit, CSIRO, North Ryde, NSW, Australia

REFERENCES

1. Kemp, D.J., Partington, G.A. and Rogers, G.E. (1974). Biochim. Biophys. Res. Commun. 60, 1006-1014.
2. Walker, I.D. and Rogers, G.E. (1976). Eur. J. Biochem. 69, 341-350.
3. Partington, G.A., Kemp, D.J. and Rogers, G.E. (1973). Nature, New Biology 240, 33-36.
4. Kemp, D.J. (1975). Nature 254, 573-577.
5. Saint, R.B., Morris, C.P., Crowe, J.S., Kuczek, E.S.A. and Rogers, G.E. (1982). Paper in preparation.
6. Benton, W.D. and Davis, R.W. (1977). Science 196, 180-182.
7. Wahl, G.M., Stern, M. and Stark, G.R. (1979). Proc. Natl. Acad. Sci.

USA 76, 3683-3687.
8. Blattner, F.R., Williams, W.G., Blechl, A.E., Denniston-Thompson, K., Faber, H.E., Furlong, L-A., Grunwald, D.J., Kiefer, D.O., Moore, D.D., Schumm, J.W., Sheldon, E.L., Smithies, O.L. (1977). Science 196, 161-169.
9. Birnboim, H.C. and Doly, J. (1980). Nucleic Acids Res. 7, 1513-1523.
10. Rigby, P.J.W., Dieckmann, M., Rhodes, C. and Berg, P. (1977). J. Molec. Biol. 113, 237-251.
11. Maxam, A.M. and Gilbert, W. (1980). Methods Enzymol. 65, 499-560.
12. Symons, R.H. (1977). Nucleic Acids Res. 4, 4347-4355.
13. Rubin, C.M. and Schmid, C.W. (1981). Nucleic Acids Res. 8, 4613-4619.
14. Messing, J., Crea, R. and Seeburg, R.H. (1981). Nucleic Acids Res. 9, 309-321.
15. Smith, G.E. and Summers, M.D. (1980). Analyt. Biochem. 109, 123-129.
16. Adams, J.M., Webb, E., Gerondakis, S. and Cory, S. (1980). Nucleic Acids Res. 8, 6019-6032.
17. Lauer, J., Shen, C-H.J. and Maniatis, T. (1980). Cell 20, 119-130.
18. Grantham, R., Gautier, C. and Guoy, M. (1980). Nucleic Acids Res. 8, 1893-1912.
19. Fitzgerald, M. and Shenk, T. (1981). Cell 24, 251-260.
20. Wasylyk, B. and Chambon, P. (1981). Nucleic Acids Res. 9, 1813-1824.
21. Breathnach, R. and Chambon, P. (1981). Ann. Rev. Biochem. 50, 349-383.
22. Efstratiadis, A., Posakony, J.W., Maniatis, T., Lawn, R.M., O'Connell, C.,Spritz, R.A., DeRiel, J.K., Forget, B.G., Weissman, S.M., Slightom, J.L., Blechl, A.E., Smithies, O.L., Baralle, F.E., Shoulders, C.C. and Proudfoot, N.J. (1980). Cell 21, 653-668.
23. Walker, I.D. and Bridgen, J. (1976). Eur. J. Biochem. 67, 283-293.
24. Gilbert, W. (1979). In Eukaryote Gene Regulation, ICN-UCLA Symposium on Molecular and Cellular Biology, XIV. R. Axel, T. Maniatis and C.F. Fox, eds. (New York: Academic Press) pp 1-12.
25. Young, R.A., Hagenbuchle, O. and Schibler, U. (1981). Cell 23, 451-458.
26. Reddy, U.B., Ghosh, P.K., Lebowitz, P. and Weissman, S.M. (1978). Nucleic Acids Res. 5, 4195-4213.