Volume 14 Number 2 1986 Nucleic Acids Research

Structure of gene and pseudogenes of human apoferritin H

Francesco Costanzo[1], Maurizio Colombo[1], Susanne Staempfli[1], Claudio Santoro[3], Maria Marone[3], Rainer Frank[1], Hajo Delius[1] and Riccardo Cortese[1,2]

[1]EMBL, Meyerhofstrasse 1, Heidelberg, FRG, [2]Istituto di Scienze Biochimiche, University of Naples, Naples and [3]Istituto di Biologia Generale, University of Turin, Turin, Italy

Received 30 October 1985; Revised and Accepted 17 December 1985

ABSTRACT

Ferritin is composed of two subunits, H and L. cDNA's coding for these proteins from human liver (1,2,3), lymphocytes (4) and from the monocyte-like cell line U937 (5) have been cloned and sequenced. Southern blot analysis on total human DNA reveals that there are many DNA segments hybridizing to the apoferritin H and L cDNA probes (1,2,4,6). In view of the tissue heterogeneity of ferritin molecules (7,8), it appeared possible that apoferritin molecules could be coded by a family of genes differentially expressed in various tissues (1,2). In this paper we describe the cloning and sequencing of the gene coding for human apoferritin H . This gene has three introns; the exon sequence is identical to that of cDNA's isolated from human liver, lymphocytes, HeLa cells and endothelial cells. In addition we show that at least 15 intronless pseudogenes exist, with features suggesting that they were originated by reverse transcription and insertion. On the basis of these results we conclude that only one gene is responsible for the synthesis of the majority of apoferritin H mRNA in various tissues examined, and that probably all the other DNA segments hybridizing with apoferritin cDNA are pseudogenes.

INTRODUCTION

Ferritin is an ubiquitous protein, present in every cell of all known organisms: it is composed of 24 subunits arranged so as to form a hollow sphere, containing large amount of iron. It is postulated that its main role is that of iron storage even though many aspects of its function are not clear (for reviews see 9 and 10).

An intriguing biochemical property is the extreme degree of heterogeneity of ferritin molecules extracted from various tissues, when analysed by electrofocusing: apparently each cell contains a mixed population of ferritin molecules ranging from more basic to more acidic species (7,8). Each cell however, has

a more or less constant and characteristic isoelectric focusing pattern, indicating that the particular combination of ferritins present in a cell type depends on regulatory mechanisms specific for that cell type. The molecular basis for this heterogeneity is still unknown. Two main mechanisms have been proposed: 1) that there are many different apoferritin molecules, either coded by many, cell specifically expressed genes, or generated by cell specific post-translational modifications (11,12) or 2) that there are only two different apoferritin subunits, differing in the isoelectrofocusing behaviour. The tissue specific heterogeneity of ferritin could be explained assuming that in each cell the ferritin is composed of a characteristic proportion of acidic and basic subunits (8). Biochemical analysis has, in our opinion, led more support to the second hypothesis, because it was possible to show that there are indeed two types of apoferritin subunits, called H and L (from Heart and Liver type respectively), and that their proportion in the ferritin molecule is different in different tissues, in agreement with the isoelectrofocusing behaviour (8,13). On the other hand, it could be and was argued that there are not in fact two subunits, and that the observations leading to this conclusion could be artificial or non conclusive (14,15).

More recently the purification of cDNA's coding for human and rat apoferritins has added new elements to this issue: first of all it has been possible to show beyond any doubt, that there are two distinct gene products, corresponding to the apoferritin H and L (1,3,4,5,16). At the same time it was also clear that in humans and rats there are many DNA sequences homologous to the apoferritin cDNA's, suggesting the possibility of a multigenic family. In addition to that, the sequences of human apoferritin H cDNA from liver and lymphocytes appeared to be different in one crucial point (4). We have reinvestigated this last point and reached the conclusion that the difference was due to a sequence mistake in the cDNA from liver, regrettably discovered only after publication (1).

In this paper we describe the isolation and characterization of several genomic segments containing

sequences homologous to apoferritin H cDNA, and show that all but one, are most probably processed pseudogenes. The real gene has three introns, and the exon sequence is identical to that of cDNA's from liver, lymphocytes, and HeLa cells.

MATERIALS AND METHODS

Bacterial strains, plasmids and phage vectors

Escherichia coli K12 (strains 71/18 and TG1) was used for transformation (17). The M13 derivatives mp18 and mp19 (18), tg130 and tg131 (19) were used as phage vectors for subcloning and sequencing. Phage EMBL3 (20) and the cosmid vector pcos2EMBL (21) were used in the construction of the human genomic libraries. Transformation and preparation of single- and double-stranded DNA were as described (22,23). Large-scale preparation of EMBL3 and pcos2EMBL recombinants was done according to (20) and (21).

Enzymes and chemicals

Restriction endonucleases, T4 polynucleotide kinase, T4 DNA ligase, DNA polymerase I holoenzyme and Klenow fragment were purchased from B.R.L. and Biolabs. ^{32}P-labelled compounds were purchased from Amersham. AMV reverse transcriptase was purchased from Boehringer.

Screening of the human libraries

Human cDNA libraries from HeLa cells (gift of G. Persico), lymphocytes (kindly provided by T. Baralle) and from umbilical cord endothelial cells (Costanzo, F. and Santoro, C., manuscript in preparation) were screened using as a probe the nick-translated 528 bp Pst-Pst fragment from the human liver H chain cDNA (1). The same probe was used in screening the λ genomic library provided by G. Bensi (24) and the cosmid library provided by A.M. Frischauf. The screening was performed according to (20) and Crkvenjakov, R. (personal communication).

Electron microscopy analysis

For the heteroduplex and hybrid analysis separated strands of the recombinant λ DNAs and of the EMBL3 vector were isolated (25). Supercoiled DNA of the cosmid 25 was cut with ClaI. Full length single strands were isolated from agarose gels.

Separated single strands of the H cDNA plasmids were

prepared by incorporation of biotinylated dUTP into the termini of the EcoRI-linearized plasmid DNA, followed by a second cut by SalI restriction enzyme, and the fractionation of the complementary strands of the two fragments on an avidin agarose column (26). For the heteroduplex preparation equal amounts of the separated strands of the different recombinants were incubated in 50% formamide, 10 mM Tris-HCl, pH 7.4, 1 mM EDTA, 0.2 M CsCl for 45 min at 37° C. Samples were spread with cytochrome (Type VI, Sigma) from 30% formamide, 0.1 M Tris-HCl pH 8.5 onto 0.005% octyl glucopyranoside. The preparations were stained with uranyl acetate, rotary shadowed with platinum.

Hybrids were prepared by passing a mixture of complementary strands of the recombinants and of vector through a Sephadex G-50 column (2x45 mm) equilibrated with 80% formamide, 0.1 M HEPES, pH 8.3, 0.4 M NaCl, 0.01 M EDTA (27). One µl of total liver RNA (8.5 mg/ml) was added and the sample incubated in a Teflon tubing sealed with needles for 15 h at 45°C. Aliquots of 5 µl were spread after 10-fold dilution in 30% formamide spreading solution as described above.

Oligodeoxynucleotide synthesis

Oligodeoxynucleotides were synthesized following the phosphite-amidite method (28).

DNA sequencing

Sequence analysis was done using the dideoxy chain termination method (29).

Primer elongation

The oligodeoxynucleotide primer was labelled at the 5' end with [γ-^{32}P] ATP and T4 polynucleotide kinase. 1 or 3 µg of polyadenylated RNA were mixed with 9 volumes of DMSO, heated at 45° C for 20 min, ethanol precipitated and resuspended in 50mM Tris-Cl pH 8.3, 5mM $MgCl_2$, 50mM KCl, 1mM DTT, 0.4mM dNTPs with 0.2 pmol of kinased primer and 20 Units of AMV reverse transcriptase. Incubation was carried out at 42° C for 2 h. After phenol extraction the samples were ethanol precipitated, washed with 80% ethanol and loaded on a 6% denaturing polyacrylamide gel.

RESULTS

cDNA sequences from various tissues

We have originally reported the sequence of a cDNA clone, from a liver library, coding for human apoferritin H subunit (1). More recently we have determined the sequence of apoferritin H cDNA isolated from HeLa cells, lymphocytes and endothelial cells. All these sequences are identical and differ from that originally reported by us at position 734 (position 617 of the Fig. 2 in [1]), 777 and 800 for the insertion of a C residue. These differences are artifacts: reinvestigation of the liver cDNA clones (6 independent isolated) shows that our previous sequence was mistaken in these positions. The conclusion of this set of experiments is that cDNA's from various cell types have an identical sequence (shown in Fig. 5, in comparison with the genomic clone sequence), in agreement with what has been independently shown by other authors (3).

Apoferritin H mRNA from various tissues has approximately the same length, as measured by Northern analysis (data not shown). A precise measurement, by primer elongation, shows that mRNA's from human liver, Daudi, HeLa cells and human placenta, initiate in the same position, 208 nucleotides upstream from the initiating ATG (Fig. 1). It is therefore probable that one gene is responsible for the synthesis of the apoferritin H mRNA in the various tissues and cells examined.

Isolation of several genomic clones carrying apoferritin H sequences

We screened two human genomic libraries, one constructed in the lambda phage vector EMBL3 and the other in the cosmid pcos2EMBL. About $5x10^5$ plaques and $3x10^5$ colonies were screened using as probe a 528 bp long cDNA fragment spanning the cDNA sequence from the internal Pst site towards the 3' end (1). DNA extracted from several positive plaques and colonies was analysed by restriction mapping and was used in a series of heteroduplex experiments. Heteroduplex analysis was done on 15 of the recombinant λ clones. A representative electron micrograph is shown in Fig. 2a. The conclusion from these experiments is that all clones studied contain an uninterrupted segment, about 1000 base pairs long, hybridizing to apoferritin

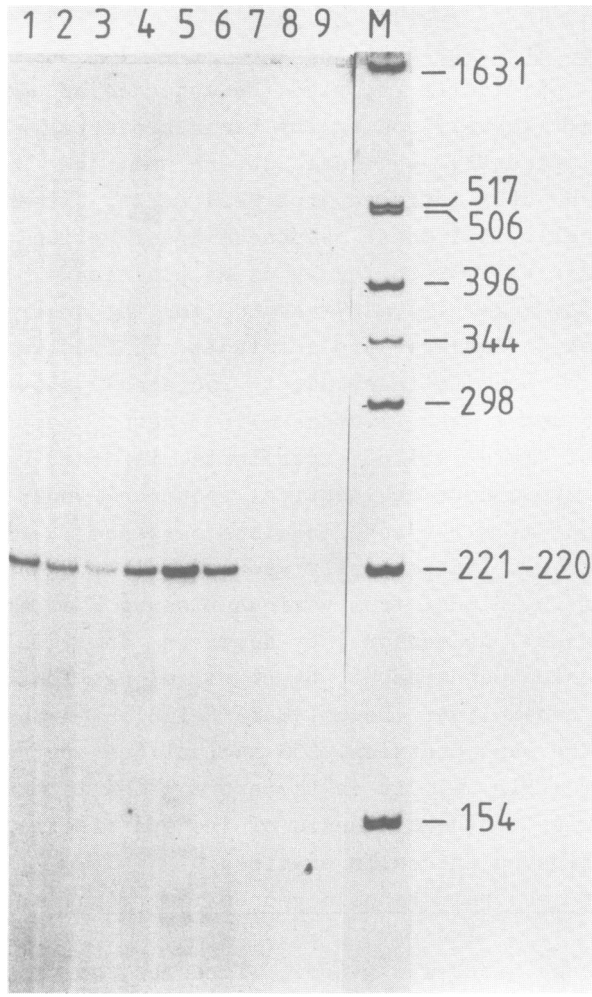

Fig. 1. Primer elongation analysis. An oligonucleotide complementary to the region from position +209 to +226 (see fig. 3) of the cDNA was used as a primer in the experiment and annealed to 1 or 3 µg of poly(A)+ RNA from human liver (lanes 1-2), Daudi cell line (3-4), HeLa cells (5-6) and human placenta (7-8). Lane 9 shows the negative control (3µg of tRNA as template) and lane 10 the molecular weight marker pBR322 HinfI cut.

H cDNA. It was also possible to confirm that the majority of clones derived from different regions of the human genome as recently reported (30). EM analysis of hybrids between the λ

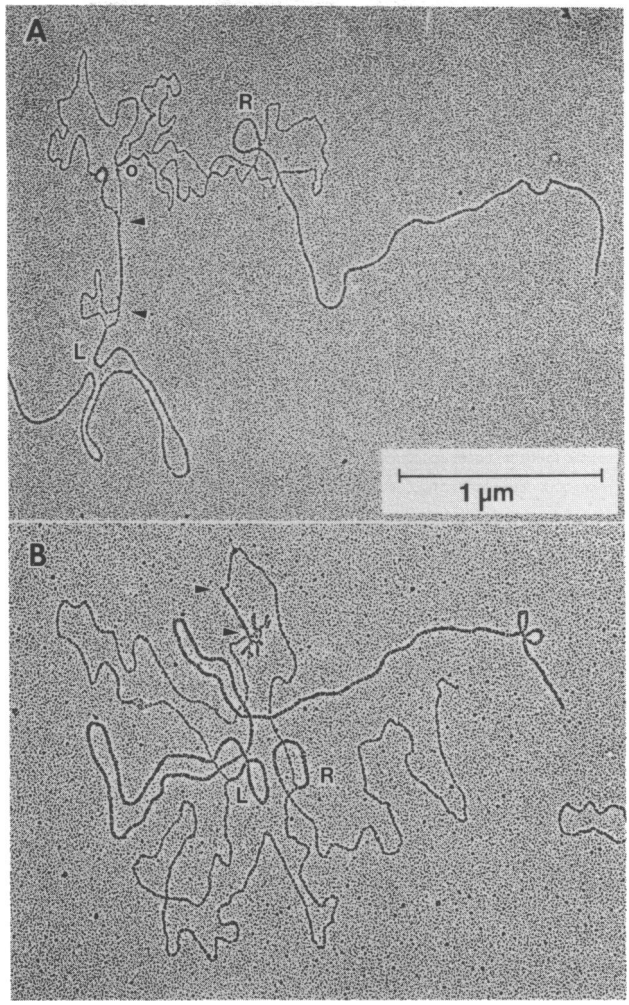

Fig. 2. a) Heteroduplex between single stranded DNAs from the recombinants 133 and 156. The homologous region marked by arrowheads corresponds to the regions of homology with the cDNA probes. Short homologies as the one marked by the circle are probably due to repetitive sequences. L and R mark the ends of the left and right λ arms.
b) mRNA was hybridized to the slow strand of the recombinant 112. The hybrid region between the two arrowheads does not display intron loops. The 5' end of the mRNA frequently appears to be aggregated with other RNA.

recombinants and mRNA extracted from human liver (Fig. 2b) confirmed that the hybridizing region corresponds to an uninterrupted stretch of about 1000 bases.

A) 5' and 3' Flanking regions of 123 and 133

```
-89
aacacagtatcgctagcagaggagaggtagaaaaggcgagggtactggacaggcatgattttggcacagtgagaaagtt
ataccatgacatcaccacagtcttaaaacttgtactgttacgtagcaacttccatttctgcaagttgcaaaattaataa

         -1                  +921
gtgggtagca..coding sequence..aaaaaaaaaaagaagagaaagttgtcggtagcaaagctattacaggacaga
tataaatata..coding sequence..aaaaataatcaatataaatatatgtattag
```

B) 5' "untranslated" region of cDNA , 123 and 133

```
+1

CAGACGTTCTTCGCCGAGAGTCGTCGGGGTTTCCTGCTTCAACAGTGCTTGGACGGAACCC  - 1=cDNA
GAGACATTCTTCGCCGAGAGTCGTCGGGGTTTCCTGCTTCAACAGTGCTTGGACGGAACCC  - 2=123
GAGACGTTCTTTGCCGAGAGTCGTCGGGGTTTCCTGCTTCAACAGTGCTTGGACGGAACCC  - 3=133
1    2     3


GGCGCTCGTTCCCCACCCCGGCCGGCCGCCCATAGCCAGCCCTCCGTCGACCTCTTCACCG
AGCGCTCGTCCCCCACCCCGGCCGGCCGCCCGTAGCCAGCCCTCTGTCCC CTCTTCACTG
GGTGCTCGTCCCCCACCCCAGCCGGCCGCCCATAGCCAGCCCTCCCTC ACCTCTTCACGG
2 3      1          3            2             23  122         1
                                                   2           2
                                                   3           3

CACCCTCGGACTGCCCCAAGGCCCCCGCCGCCGCTCCAGCGCCGCGCAGCCACCGCCGCCG
CACCCTCGGACCATCCCAAGGCCCCCGCCGCCGCTCCAGCGCCGCGCAGCCACCGCCGCCA
CACCCTCAGACCGCCCCAAGGCCCCTGCCGCCGCCCCAGCGCCACGCAGCCACTG   CCA
       3   122           3         3         3          3 333  1
                         ----------------------------------


                        +208
CCGCCG   CCTCTCCTTAGTCGCCGCC
CCTC         TCCTTAGTTGCTGCC
CCGCCGCCACCTCTCCTTAGTCGCCGCC
  2 222222222        2  2
      333
```

C) "coding" region of cDNA, 123 and 133

```
+209
ATG ACG ACC GCG TCC ACC TCG CAG GTG CGC CAC AAC TAC CAC CAG GAC TCA GAG
ATG ACA ACT GCG TCC ACC TCG CAG GTG CGC CAG AAC TAC CAC CAG GAC TCA GAG
ATG AGG ACC ACG TCC ACC TCA CAG GTG CGC CAG AAC TAC CAC CAG GAC TCA GAG
     32   2 3            3

GCC GCC ATC AAC CGC CAG ATC AAC CTG GAG CTC TAC GCC TCC TAC GTT TAC CTG
GCC GCC ATC AAC GGC CAG ATC AAC CTG GAG CTC TAC GCC TCC TAC GTT TAC CTG
GCC GCC ATC AAC CGC CAG ATC AAC CTA GAG CTC TGT GCC TCC TAC GTT TAC CTG
                  2                    3          33

TCC ATG TCT TAC TAC TTT GAC CGC GAT GAT GTG GCT TTG AAG AAC TTT GCC AAA
TCC ATG TCT TAC AAC TTT GAC CGC GAT GAT GTG GCT TTG AGG AAC TTT GCC ACA
TCC ATG TCT TAC TGC TTT GAC CGT GAT GAT GTG GCT TTG AAG AAC TTT GCC AAA
                 23            3                    2               2


TAC TTT CTT CAC CAA TCT CAT GAG GAG AGG GAA CAT GCT GAG AAA CTG ATG AAG
TAC TTT CTT CAC CAA TCT CAT GAG GAG AGG GAA CAT GCC GAG AAA CTG ATG AAG
TAC TTT CTT CAC CAA TCT CAT GAG GAG AGG GAG CAT GCT GAG AAA CTG ATG AAG
                                          3        2
                                                   ----------------------

CTG CAG AAC CAA CGA GGT GGC CGA ATC TTC CTT CAG GAT ATC AAG AAA CCA GAC
CTG CAA AAC TAT CGT GGT GGC CAA ATC TTC CTT CAG GAT ATC AAG AAA CCA GTC
CTG CAG AAC CAA CGA GGT GGC CGA ATC TTC CTT CAG GAT ATC AAA AAA CCA GAC
      2     2 2   2          2                                3          2
------------------------------
TGT GAT GAC TGG GAG AGC GGG CTG AAT GCA ATG GAG TGT GCA TTA CAT TTG GAA
TGT GAT GAC TGG GAG AGT GGG CTG AAT GCA ATG GAG TGT GCA TCA CAT TTG GAA
TGT GAT GAC TGG GAG AGC GGG CTG AAT GTG ATG GAG TGT GCA TTA CAT TTG GAA
                       2              33                 2

AAA AAT GTG AAT CAG TCA CTA CTG GAA CTG CAC AAA CTG GCC ACT GAC AAA AAT
AAA AAT GTG AAT CAG TCA CTA TTG GAA CTG CAC AAA CTA GCC ACT GAC AAA AAT
AAA AAT GTG AAT CAG TCA CTA CTG GAA CTG CAC AAA TTG GCC ACT GAC AAA AAT
                             2                   3 2

GAC CCC CAT TTG TGT GAC TTC ATT GAG ACA CAT TAC CTG AAT GAG CAG GTG AAA
GAC CCC CAT TTG TGT GAC TTC ATT GAG ACA TGT TAC CTG AAT GAG CAG GTG AAA
GAC CCC CAT TTG TGT GAC TTC ATT GAG ACA CAT TAC CTG AAT GAG CAG GTG AAA
                                       22
```

```
GCC ATC AAA GAA TTG GGT GAC CAC GTG ACC AAC TTG CGC AAG ATG GGA GCG CCC
GCC ATC AAA GAA TTG GGT GAC CAC GTG ACC AAC TTG CAC AAG ATG GGA GCC CCC
GCC ATC AAA GAA TTG GAT GAC CAT GTG ACC AAC TTG CAC AAG ATG GGA GCA CTC
                      3        3                    1                1  3
                                                                     2
                                                                     3

GAA TCT GGC TTG GCG GAA TAT CTC TTT GAC AAG CAC ACC CTG GGA GAC AGT GAT
GAA TCT GGC TTG GCA GAA TAT ATC TTT AAC AAG AAC ACC CTG GGA GAC AGT GAT
GAA TCT GGC TTG GCA GAA TAT CTC TTT GAC AAG CAC ACC CTG GGA GAC AGT GAT
                   1          2       2       2
                              ------------------------------------------

            +760
AAT GAA AGC TAA
AAT GAA AAC TAA
AAT GAA AGC TAA
        2
----------
```

D) 3' "untranslated" region of cDNA,123,133

```
+761
GCCTCGGGCTAATTTCCCCATAGCCGTGGGGTGACTTCCCTGGTCACCAAGGCAGTGCATGCATGTTGGGGTTTCCTTTA
GCCTCAGGCTAATTTCCCCATAGCCATGGGGCGAATTCCCTGGTCACTAAGGCAGTGCATGCATGTTGGGATTTCCTTTA
GCCTCG GCTAATTTCCCCGTAGCCGTGGGGTGACTTCCCTGGTCACCAAGGCAGTGCATGCATGTTGGGGTTTCCTTTA
    23            3     2     2  2          2                2
------            ------------------------------------------------------------

                                                                        +920
CCTTTTCTATAAGTTGTACCAAAACATCCACTTAAGTTCTTTGATTTGTACCATTCCTTCAAATAAAGAAATTTGGTACC
CCTTTTCTATAAGTTGTGCCAAAACATCCACTTAAGTTCTTCGATTTGTACCATTCCTTCAAATAAAGAAATTTGGTACC
CCTTTTCTATAAGTTGTACCAAAACATCCACCTAAGTTCTTTGATTTGTGCCATTCCTTCAGATGAAGAAATTTGGTACC
                2             3         2       3           3  3
------------------
```

Fig. 3. Sequence comparison of the clones 123 and 133. The comparison is extended to the cDNA in the region corresponding to the putative retrotranscripts. The direct repeats are boxed. See text for further details.

Intronless genes have been described in a variety of multigene families from different organisms; in many cases the intronless genes proved to be pseudogenes, with in frame stop codons or deletions and absence of sequences resembling promoters (31-37). On the other hand there are known cases of active nuclear genes without introns (38-41). To gain more information we determined the DNA sequence of two apoferritin H genomic clones.

DNA sequence analysis of 123 and 133

Two genomic clones, 123 and 133, have been sequenced making use of the Sac I site present in the coding region to construct subclones in the M13 vector. The complete sequence is shown in Fig. 3. In panel A) in small cases letters are the 5' and 3' flanking regions to the putative 123 and 133 apoferritin processed retrotranscript; there is no homology between them. In both cases the coding region is flanked by direct repeats (boxed in the figure); in panel B) in continuous capital case letters is the 5' untranslated region of the putative retrotranscribed sequences (123 and 133 are the second and third line

respectively, the first line is the revised version of the cDNA sequence); in panel C) capital letters in groups of three correspond to the coding region; in panel D) in continuous capital letters are the 3' untranslated flanking regions. The flanking regions of the two clones (panel A) show features characteristic of pseudogenes derived from processed RNA intermediates. 133 coding sequence is flanked in the 5' and 3' region by the direct repeat TAATAATATAAATATA suggesting DNA insertion. In 123 we found an imperfect direct repeat flanking the coding region, and a polyadenylation site followed by a stretch of A's.

The homology between the two genes in the 5' flanking region spans for about 200 nucleotides, corresponding exactly to the 5' flanking untranslated region, as determined by primer elongation (Fig. 1) and diverges after this point. Similarly the homology in the 3' flanking untranslated region extends only up to the polyA (traces of which are present in both 123 and 133).

Unlike most pseudogenes, however, 123 and 133 have an uninterrupted reading frame as long as the expressed cDNA segment. Even though 123 and 133 have been obviously generated by retrotranscription and insertion, they can potentially code for a protein similar to apoferritin H. In the alignment shown in Fig. 3, we have indicated the sequence differences with the following notation: 1 denotes a sequence position in which the nucleotide is identical in the 123 and 133 and is different in the cDNA; 2 denotes a sequence position in which the same nucleotide is present in the cDNA and in 133 but is different in 123; 3 denotes a position in which cDNA and 123 have the same nucleotide and 133 has a different one.

The distribution of the differences is interesting. We assume that the retrotranscribed DNA derived from the same ancestral gene coding for the cDNA: all nucleotides identical in all three sequences or common to two of them are likely to be conserved from the original gene prior to the retrotranscriptional event (or are due to subsequent gene conversion). All differences present only in one of the sequences must correspond to mutations occurred after the

separation. The higher number of 2's and 3's means therefore that the cDNA sequence has diverged much less than the others from the original gene, as one would expect if the cDNA sequence, being responsible for the synthesis of the functional protein, has been under some structural constraint, whereas 123 and 133 are unselected pseudogenes.

The lower number of 3's might indicate that 133 has been generated by a more recent retrotranscriptional event, or that it was changed by gene conversion; the apparent clustering of 2's in at least two positions along the sequence (see underlined regions in Fig. 3) might in fact indicate gene conversion zones (presumably between 133 and the cDNA coding gene) (42,43).

From these results it was reasonable to expect that also the other intronless genes were pseudogenes and we did not further analyse them.

Isolation and identification of an expressed gene

We wished to identify the genomic clone containing the gene expressed in human liver. We chose a strategy aimed at the identification of the clones with the highest degree of sequence homology with the cDNA. For this purpose we synthesized a set of oligonucleotides spanning both the coding and the 3' flanking region to be used as probes on all the genomic clones identified, spotted in equimolar amounts onto a nitrocellulose filter. By using different washing temperatures, calculated on the basis of the T_m of the different oligos, it was possible to identify two clones with a hybridization pattern identical to that of the cDNA used as control. These two clones, named c18 and c25, contained the same genomic region and therefore only c25 was further analysed. Sequences hybridizing to apoferritin H cDNA are contained within a HindIII-HindIII fragment of about 6.5 kb. An electron micrograph of the heteroduplex between the cosmid clone and the cDNA segment isolated from a liver cDNA library is shown in Fig. 4a. There appear to be three introns of the approximate sizes of 1.6, 0.2 and 0.04 kb. A map of the extensions of exon and intron sizes derived from the EM measurements is given in Fig. 4b. Further experiments showed that the coding region is comprised between two BamHI restriction fragments of about 4kb (the 5' part) and of 2.5 kb

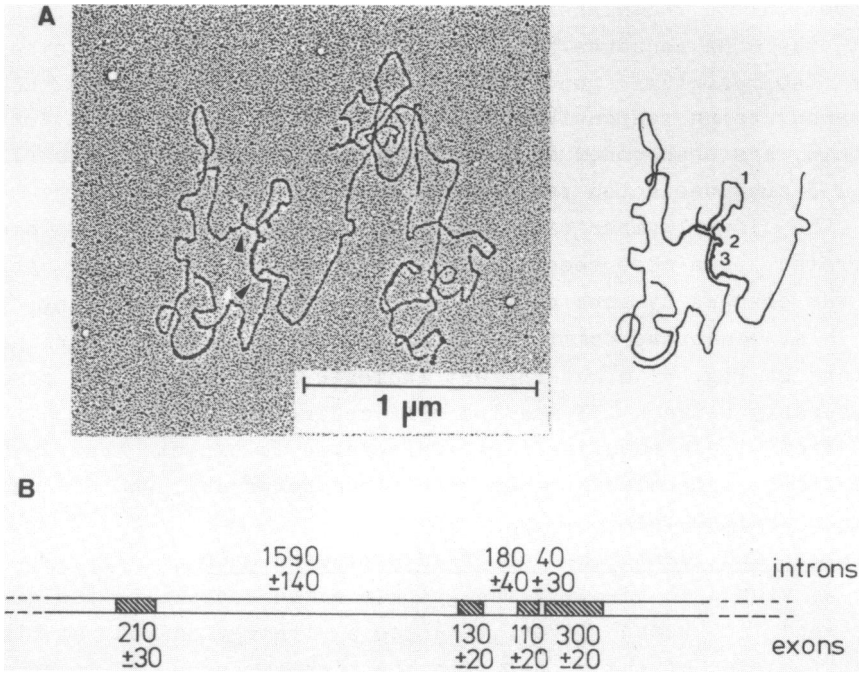

Fig. 4. a) Heteroduplex between a single strand of cosmid 25 cut by restriction enzyme ClaI and a separated strand from the plasmid containing the complete cDNA of the apoferritin H gene. The heteroduplex is interrupted by three intron regions numbered in the tracing to the right (the third intron is so small that it cannot be seen in all heteroduplexes). A homology of pBR322 sequences present in the plasmid and cosmid vectors leads to the formation of a circular structure.
b) Map of the sizes of exon and intron regions as determined by electron microscopy. The average lengths and standard deviations are given in nucleotides.

(the 3' part). These two fragments were subcloned and further analysed by restriction mapping and DNA sequencing. The results are shown in Fig. 5. The coding region is identical to that of the cDNA's and is interrupted by three introns, as indicated by the heteroduplex analysis. The first exon comprises 208 bases of untranslated region and a segment coding for the first 38 aminoacids. The second exon codes for aa 39 to 87, the third exon from aa 88 to 129 and the fourth from aa 130 to the stop codon and 160 bases of 3' untranslated region. Upstream from the point of initiation of transcription there is a canonical

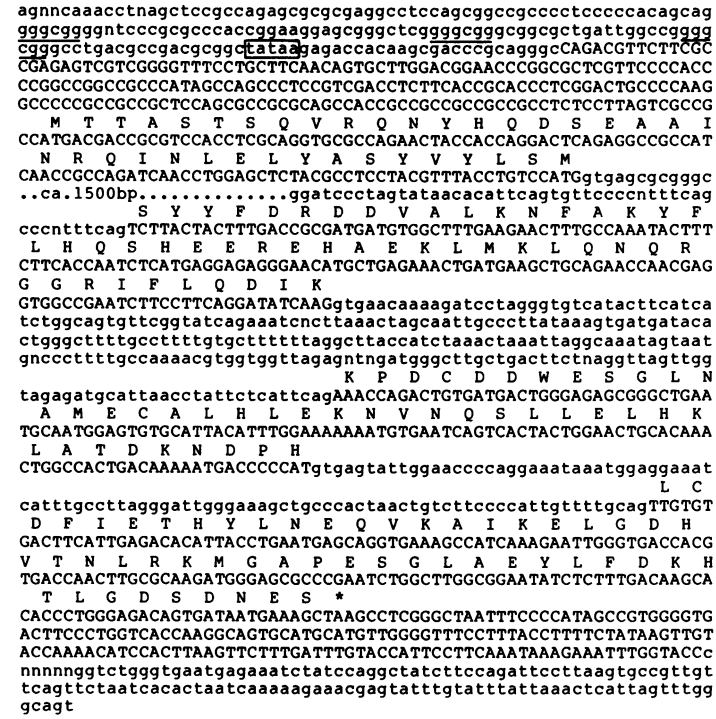

Fig. 5. Nucleotide sequence of the human H chain apoferritin gene. Capital letters indicate the regions present in the mature mRNA. The presumptive TATA box is indicated as well as the putative Sp1-binding sites.

TATA box; further upstream there are three ggcggg boxes, which are frequently found in eukaryotic promoters and which, in some cases, have been shown to interact with transcriptional factor Sp1 (44). This is in agreement with the proposal that such factor is important for efficient expression of "house-keeping" proteins.

The fact that cDNA's and c25 have identical sequence indicates that c25 codes for the apoferritin H subunit in a variety of cells. Further support to this conclusion is given by the observation that the apoferritin gene contained in clone c25 is present in one copy per haploid genome, as shown by Southern blot on total human DNA. A segment of DNA from the 5' flanking region of the c25 clone (see Fig. 6, for experimental

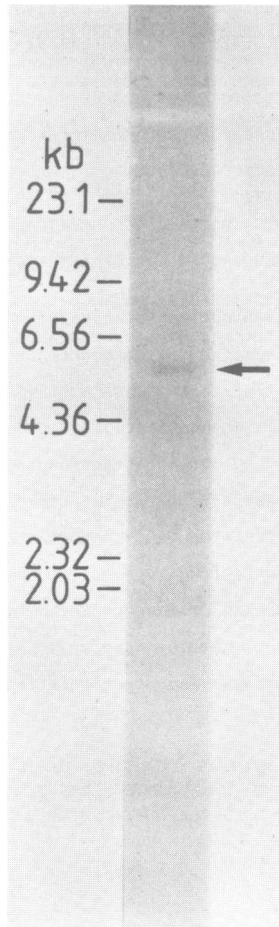

Fig. 6. Southern blot analysis on 10 µg of EcoRI restricted human genomic DNA. An EcoRI-MaeII fragment, about 200 bp in size, including the TATA box and the first few nucleotides of the transcribed region was nick-translated and used as a probe. This promoter-specific probe detects only one band of 6 kb in contrast with the high number of bands hybridizing to the cDNA probes.

details), was used as a probe for hybridization to total human DNA restricted with EcoRI. A single band is revealed, approximately 6 kb long, as predicted from the analogous experiment on the c25 cloned DNA. The intensity of the signal corresponds to that expected from single copy genes. It is very probable therefore that there is only one gene coding for human

apoferritin H subunit (and only one coding for apoferritin L subunit; manuscript in preparation). The complex pattern seen on the Southern blots is a consequence of a high number of processed genes (pseudogenes).

ACKNOWLEDGMENTS

We thank H. Seifert for typing this manuscript and P. Stevenson for technical assistance. F.C. was the recipient of a training contract GBI-047-D from the "Biomolecular Engineering Programme" of the Commission of the European Communities. M.C. was supported by a fellowship from Deutscher Akademischer Austauschdienst (DAAD). The work done in Italy was supported by P.F. Ingegneria Genetica e Basi Molecolari delle malattie ereditarie and by P.F. Oncologia, CNR, Rome.

REFERENCES

1) Costanzo, F., Santoro, C., Colantuoni, V., Bensi, G., Raugei, G., Romano, V. and Cortese, R. (1984) EMBO J. 3, 23-27.
2) Costanzo, F., Santoro, C. and Cortese, R. (1984) in Ferritins and Isoferritins as Biochemical Markers (Albertini, A. et al., eds), pp. 79-85, Elsevier Science Publishers B.V.
3) Boyd, D., Vecoli, C., Belcher, D.M., Jain, S.K. and Drysdale, J.W. (1985) J. Biol. Chem. 260, 11755-11761.
4) Boyd, D., Jain, S.K., Crampton, J., Barrett, K.J. and Drysdale, J. (1984) Proc. Natl. Acad. Sci. USA 81, 4751-4755.
5) Dörner, M.H., Salfeld, J., Will, H., Leibold, E.A., Vass, J.K. and Munro, H.N. (1985) Proc. Natl. Acad. Sci. USA 82, 3139-3143.
6) Jain, S.K., Barrett, K.J., Boyd, D., Favreau, M.F., Crampton, J. and Drysdale, J.W. (1985) J. Biol. Chem. 260, 11762-11768.
7) Drysdale, J.W. (1970) Biochim. Biophys. Acta 207, 256-258.
8) Arosio, P., Adelman, T.G. and Drysdale, J.W. (1978) J. Biol. Chem. 253, 4451-4458.
9) Munro, H.N. and Lindner, M.C. (1978) Physiol. Rev. 58, 317-396.
10) Aisen, P. and Listowsky, I. (1980) Ann. Rev. Biochem. 49, 357-393.
11) Crichton, R.R., Millar, J.A., Cumming, R.L.C. and Bryce, C.F.A. (1973) Biochem. J. 131, 51-59.
12) Harrison, P.M., Banyard, S.J., Hoare, R.J., Russell, S.M. and Treffry, A. (1977) Ciba Found. Symp. 51, 19-40.
13) Adelman, T.G., Yokota, M. and Drysdale, J.W. (1977) in Proteins of Iron Metabolism (Brown, E.B. et al., eds), pp.49-55, Grune and Stratton, New York.
14) Fish, W.W. (1976) J. Theor. Biol. 60, 385-392.
15) Bryce, C.F.A., Magnusson, C.G.M. and Crichton, R.R. (1978) FEBS Lett. 96, 257-262.
16) Brown, A.J.P., Leibold, E.A. and Munro, H.N. (1983) Proc. Natl. Acad. Sci. USA 80, 1265-1269.

17) Gronenborn, B. and Messing, J. (1978) Nature 272, 375-377.
18) Messing, J. (1983) Methods Enzymol. 101, 28-78.
19) Kieny, M.P., Lathe, R. and Lecocq, J.P. (1983) Gene 26, 91-99.
20) Frischauf, A.M., Lehrach, H., Poustka, A.M. and Murray, N. (1983) J. Mol. Biol. 170, 827-842.
21) Poustka, A., Rackwitz, H.R., Frischauf, A.M., Hohn, B. and Lehrach, H. (1984) Proc. Natl. Acad. Sci. USA 81, 4129-4133.
22) Cortese, R., Melton, D.A., Tranquilla, T. and Smith, J.D. (1978) Nucl. Acids Res. 5, 4593-4611.
23) Cortese, R., Herland, R. and Melton, D.A. (1980) Proc. Natl. Acad. Sci. USA 77, 4147-4151
24) Bensi, G., Raugei, G., Klefenz, H. and Cortese, R. (1985) EMBO J. 4, 119-126.
25) Koller, B., Delius, H., Buenemann, H. and Mueller, W. (1978) Gene 4, 227-239.
26) Delius, H., van Heerikhuizen, H., Clarke, J. and Koller, B. (1985) Nucl. Acids Res. 13, 5457-5469.
27) Chow, L.T., Roberts, J.M., Lewis, B. and Broker, T.R. (1977) Cell 11, 819-836.
28) Winnacker, E.L. and Dorper, T. (1982) in Chemical and Enzymatic Synthesis of Gene Fragments (Gassen, H.G. and Lang,A. eds), Verlag Chemie, pp. 97-102.
29) Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463-5467.
30) Cragg, S.J., Drysdale, J. and Worwood, M. (1985) Hum. Genet. 71, 108-112.
31) Vanin, E.F., Goldberg, G.I., Tucker, P.W. and Smithies, O. (1980) Nature 286, 222-226.
32) Van Arsdell, S.W., Denison, R.A., Bernstein, L.B., Weiner, A.M., Manser, T. and Gesteland, R.F. (1981) Cell 26, 11-17.
33) Lemischka, I. and Sharp, P.A. (1982) Nature 300, 330-335.
34) Hollis, G.F., Hieter, P.A., McBride, O.W., Swan, D. and Leder, P. (1982) Nature 296, 321-325.
35) Karin, M. and Richards, R. (1982) Nature 299, 797-802.
36) Wilde, C.D., Crowther, C.E., Cripe, T.P., Gwo-Shu, M. and Cowan, N.J. (1982) Nature 297, 83-84.
37) Benham, F.J., Hodgkinson, S. and Davies, K.E. (1984) EMBO J. 3, 2635-2640.
38) Schaffner, W., Kunz, G., Daetwyler, H., Telford, J., Smith, H.O. and Birnstiel, M.L. (1978) Cell 14, 655-671.
39) Nagata, S., Mantei, N. and Weissmann, C. (1980) Nature 287, 401-408.
40) Houghton, M., Jackson, I.J., Porter, A.G., Doel, S.M., Catlin, G.H., Barber, C. and Carey, N.H. (1981) Nucl. Acids Res. 9, 247-266.
41) Stein, J.P., Munjaal, R.P., Lagace, L., Lai, E.C., O'Malley, B.W. and Means, A.R. (1983) Proc. Natl. Acad. Sci. USA 80, 6485-6489.
42) Ohta, T. (1984) Genetics 106, 517-528.
43) Nagylaki, T. (1984) Proc. Natl. Acad. Sci. USA 81, 3796-3800.
44) Gidoni, D., Dynan, W.S. and Tjian, R. (1984) Nature 312, 409-413.