A gene encoding 22 highly related zinc fingers is expressed in lymphoid cell lines

Ruth Lovering and John Trowsdale
Imperial Cancer Research Fund Laboratories, PO Box 123, Lincoln's Inn Fields, London WC2A 3PX, UK

Received March 7, 1991; Revised and Accepted May 3, 1991 EMBL accession no. X59244

ABSTRACT

A cDNA was isolated from a T cell library using an oligonucleotide probe corresponding to a sequence conserved in proteins with multiple zinc fingers of the C_2H_2 type. The predicted protein structure of this cDNA (ZNF43) showed that it contained 22 of the *Krüppel* type of zinc finger motifs in tandem. The amino acid sequence was strongly conserved between each of the finger domains of this cDNA, except for variable residue positions within the putative DNA binding site. Within the zinc finger domain the amino acid sequence of the four zinc fingers 6 to 9 was very similar to the amino acid sequence of fingers 10 to 13, the DNA sequence bound by this group of eight fingers may include a short repeat. Southern blotting showed that ZNF43 was one of a closely related family of proteins with 10 to 20 members. The members of the ZNF43 family did not appear to be clustered at the chromosomal level. The transcription of many members of this gene family was increased in lymphoid cell lines. After *in vitro* induced terminal differentiation of the human HL60 cell line the expression of the ZNF43 family was reduced. The expression of the ZNF43 gene was mainly limited to T and B cell lines. The gene was differentially spliced and different cell lines expressed different combinations of transcripts.

INTRODUCTION

The regulation of transcription by DNA binding proteins is critical for the development and differentiation of all organisms. Several distinct motifs have been identified which are capable of binding DNA in a sequence specific manner and these have been shown to be present in a variety of transcription factors which play a major role in development (1). One of these motifs, the zinc finger, is present in several characterised transcription factors, eg. in the *Drosophila Krüppel* and *hunchback* genes, which are zygotic segmentation genes (2, 3) and in the glucocorticoid receptor protein, which is required for transcriptional regulation by steroid hormones (4, 5). The zinc finger family of genes has been subdivided into two classes according to the zinc chelating amino acids present. The *Krüppel* or TFIIIA-like proteins are characterised by two cysteine and two histidine residues (C_2H_2) which chelate the zinc ion whereas the steriod hormone receptors usually have two pairs of cysteines (10, 11, 5). In most of the genes that have been studied there is more than one zinc finger motif forming the DNA binding region of the protein (10). The zinc finger region of a protein is usually composed of tandemly repeated zinc finger motifs, however, occasionally one group of zinc fingers are separated from another by a non-finger region eg. Xfin 37 (12) and PRDII-BFI (13).

There are over 300 genes in the human genome containing zinc finger domains (6). Most of the characterised zinc finger proteins specifically bind DNA and are transcription regulators (7, 8); however, the TFIIIA zinc finger motif binds both DNA and RNA (9) and it is possible that some zinc finger gene products have functions other than transcription factors.

Lymphoid cells provide a good system for the study of cell differentiation as the stages of T-cell and B-cell differentiation have been well characterised using reagents and antibodies which recognise stage specific intracellular and cell surface markers (14). It should therefore be possible to identify transcription regulators which are specific for particular stages of cell differentiation and are involved in the maturation pathway. This study identified a zinc finger gene family which is mainly expressed in lymphoid cell lines.

MATERIALS AND METHODS

Library screening

A λgt11 cDNA library derived from a B-cell line, ROF-NL was screened with the oligonucleotide ACTCACACTGGGGAGAA-GCCCTACGAGTGCACCGAGTGTGGG end labeled with γ^{32}P-ATP (15). This encodes the amino acid sequence THTGEKPYECTECG which is highly conserved in zinc finger sequences of many species. 110 positive clones were obtained from 200 000 recombinant clones after washing to a stringency of 6×SSC, at room temperature for 30 minutes, then at 42°C for 2 minutes. All clones of interest were subcloned into the plasmid Bluescript (Stratagene). One cDNA clone, 2w6 (an 840 bp *Eco* RI fragment), was of particular interest as it was expressed predominantly in lymphoid cell lines. A cDNA library, in the expression vector CDM8 (20), with mRNA from a T-cell line CEM was screened twice for longer cDNA clones of 2w6. First it was screened with the 840 bp *Eco* RI fragment (2w6) resulting in over 800 positive clones from 1×10^6 recombinant clones. One of these cDNA clones of 3.05 kb (2A1.2) was identified

4B1-5-PRIME

```
                                                                CTTCGTTCTTCTGTGTCCTCTGCTGCTAGAGGTCCA   36
GCCTCTGTGGCTCTG TGA CCT GCG GGT ATT GGG GGA TCC ACA GCT AAG ACG CCA GGA CCC CCC GGA AGC CTA GAA ATG GGA CCA TTG ACA TTT  129
                .  Pro Ala Gly Ile Gly Gly Ser Thr Ala Lys Thr Pro Gly Pro Pro Gly Ser Leu Glu MET Gly Pro Leu Thr Phe
```

2A1.2-5-PRIME

```
      GTGATCTGCAAGTCCTGGGAGACGCACAGCTAAGATGCCCGGACATCCTGGAAGCTGGGAAATGGTGATGTGACTCTTCTTCAGCCTGCACCCTCCAAAAGAGGATTGTGATG  113
TATCACTGGACCCAGCACCTAGATGACGTGTGATTGTGACACATACCTCTGCTCAGAATGTGAGCGATTTGACTCTCCTGCCTGGGCCCAGTACACAGATGGGATTGTGACATATCGCT  232
GGACCCAGCATGTAGATCATGTGACCCTATACTCTTGCCTTGGTGCTGTCGCAGAGGGCATTG TGA CAT ATC ACT GGT CCC TGC ACC CAG GGA CCA TTG ACA TTT  337
                                                          .  His Ile Thr Gly Pro Cys Thr Gln Gly Pro Leu Thr Phe
```

```
ATG GAT GTG GCC ATA GAA TTC TGT CTG GAG GAG TGG CAA TGC CTG GAC ATT GCA CAG CAG AAT TTA TAT AGG AAT GTG ATG TTA GAG AAC  427
MET Asp Val Ala Ile Glu Phe Cys Leu Glu Glu Trp Gln Cys Leu Asp Ile Ala Gln Gln Asn Leu Tyr Arg Asn Val MET Leu Glu Asn

TAC AGA AAC CTG GTC TTC CTG GGT ATT GCT GTC TCT AAG CCA GAC CTG ATC ACC TGT CTG GAG CAA GAA AAA GAG CCT TGG GAG CCT ATG  517
Tyr Arg Asn Leu Val Phe Leu Gly Ile Ala Val Ser Lys Pro Asp Leu Ile Thr Cys Leu Glu Gln Glu Lys Glu Pro Trp Glu Pro MET

AGG AGA CAT GAA ATG GTA GCC AAA CCC CCA GTT ATG TGT TCT CAT TTT ACC CAA GAC TTT TGG CCA GAG CAG CAT ATA AAA GAT CCT TTC  607
Arg Arg His Glu MET Val Ala Lys Pro Pro Val MET Cys Ser His Phe Thr Gln Asp Phe Trp Pro Glu Gln His Ile Lys Asp Pro Phe

CAA AAA GCG ACA CTG AGA AGA TAT AAA AAC TGT GAA CAT AAA AAT GTA CAT TTA AAA AAA GAC CAT AAA AGT GTG GAT GAG TGT AAG GTG  697
Gln Lys Ala Thr Leu Arg Arg Tyr Lys Asn Cys Glu His Lys Asn Val His Leu Lys Lys Asp His Lys Ser Val Asp Glu Cys Lys Val

CAC AGA GGA GGT TAT AAT GGA TTT AAC CAA TGT TTG CCA GCT ACC CAG AGC AAA ATA TTT CTA TTT GAT AAA TGT GTG AAA GCC TTT CAT  787
His Arg Gly Gly Tyr Asn Gly Phe Asn Gln Cys Leu Pro Ala Thr Gln Ser Lys Ile Phe Leu Phe Asp Lys Cys Val Lys Ala Phe His

AAA TTT TCA AAT TCA AAC AGA CAT AAG ATA AGC CAT ACT GAA AAA AAA CTT TTC AAA TGC AAA GAA TGT GGC AAA TCA TTT TGC ATG CTT  877
Lys Phe Ser Asn Ser Asn Arg His Lys Ile Ser His Thr Glu Lys Lys Leu Phe Lys Cys Lys Glu Cys Gly Lys Ser Phe Cys MET Leu
                                                                                1

CCA CAT CTA GCT CAA CAT AAA ATA ATT CAT ACC AGA GTG AAT TTC TGC AAA TGT GAA AAA TGT GGA AAA GCT TTT AAC TGC CCT TCA ATC  987
Pro His Leu Ala Gln His Lys Ile Ile His Thr Arg Val Asn Phe Cys Lys Cys Glu Lys Cys Gly Lys Ala Phe Asn Cys Pro Ser Ile
                                                                    2

ATC ACT AAA CAT AAG AGA ATT AAT ACT GGA GAG AAA CCC TAC ACA TGT GAA GAA TGT GGC AAA GTC TTT AAT TGG TCC TCA CGC CTT ACT 1057
Ile Thr Lys His Lys Arg Ile Asn Thr Gly Glu Lys Pro Tyr Thr Cys Glu Glu Cys Gly Lys Val Phe Asn Trp Ser Ser Arg Leu Thr
                                                        3

ACA CAT AAA AAA AAT TAT ACT AGA TAC AAA CTC TAC AAA TGT GAA GAA TGT GGC AAA GCT TTT AAC AAG TCC TCA ATC CTT ACT ACC CAT 1147
Thr His Lys Lys Asn Tyr Thr Arg Tyr Lys Leu Tyr Lys Cys Glu Glu Cys Gly Lys Ala Phe Asn Lys Ser Ser Ile Leu Thr Thr His
                                                            4

AAG ATA ATT CGC ACT GGA GAG AAA TTC TAC AAA TGT AAA GAA TGT GCC AAA GCT TTT AAC CAA TCC TCA AAC CTT ACT GAA CAT AAG AAA 1237
Lys Ile Ile Arg Thr Gly Glu Lys Phe Tyr Lys Cys Lys Glu Cys Ala Lys Ala Phe Asn Gln Ser Ser Asn Leu Thr Glu His Lys Lys
                                            5

ATT CAT CCT GGA GAG AAA CCT TAC AAA TGT GAA GAA TGT GGC AAA GCC TTT AAC TGG CCC TCA ACT CTT ACT AAA CAT AAG AGA ATT CAT 1327
Ile His Pro Gly Glu Lys Pro Tyr Lys Cys Glu Glu Cys Gly Lys Ala Phe Asn Trp Pro Ser Thr Leu Thr Lys His Lys Arg Ile His
                                        6

ACT GGA GAG AAA CCC TAC ACA TGT GAA GAA TGT GGC AAA GCT TTT AAC CAG TTC TCA AAC CTT ACT ACA CAT AAG AGA ATC CAT ACT GCA 1417
Thr Gly Glu Lys Pro Tyr Thr Cys Glu Glu Cys Gly Lys Ala Phe Asn Gln Phe Ser Asn Leu Thr Thr His Lys Arg Ile His Thr Ala
                                    7

GAG AAA TTC TAT AAA TGT ACA GAA TGT GGT GAA GCT TTT AGC CGG TCC TCA AAC CTT ACT AAA CAT AAG AAA ATT CAT ACT GAA AAG AAA 1507
Glu Lys Phe Tyr Lys Cys Thr Glu Cys Gly Glu Ala Phe Ser Arg Ser Ser Asn Leu Thr Lys His Lys Lys Ile His Thr Glu Lys Lys
                8

CCC TAC AAA TGT GAA GAA TGT GGC AAA GCT TTT AAG TGG TCC TCA AAG CTT ACT GAA CAT AAG TTA ACT CAT ACT GGA GAG AAA CCC TAC 1597
Pro Tyr Lys Cys Glu Glu Cys Gly Lys Ala Phe Lys Trp Ser Ser Lys Leu Thr Glu His Lys Leu Thr His Thr Gly Glu Lys Pro Tyr
            9

AAA TGT GAA GAA TGT GGC AAA GCC TTT AAC TGG CCC TCA ACC CTT ACT AAA CAT AAC AGA ATT CAT ACT GGA GAG AAA CCC TAC AAA TGT 1687
Lys Cys Glu Glu Cys Gly Lys Ala Phe Asn Trp Pro Ser Thr Leu Thr Lys His Asn Arg Ile His Thr Gly Glu Lys Pro Tyr Lys Cys
    10                                                                                                          11

GAA GTA TGT GGC AAA GCT TTT AAC CAG TTC TCA AAC CTT ACT ACA CAT AAG AGA ATT CAT ACT GCA GAA AAA CCG TAC AAA TGT GAA GAA 1777
Glu Val Cys Gly Lys Ala Phe Asn Gln Phe Ser Asn Leu Thr Thr His Lys Arg Ile His Thr Ala Glu Lys Pro Tyr Lys Cys Glu Glu
                                                                                                    12

TGT GGC AAA GCT TTT AGC CGG TCC TCA AAC CTT ACT AAA CAT AAG AAA ATT CAC ATT GAA AAG AAA CCC TAC AAA TGT GAA GAA TGT GGC 1867
Cys Gly Lys Ala Phe Ser Arg Ser Ser Asn Leu Thr Lys His Lys Lys Ile His Ile Glu Lys Lys Pro Tyr Lys Cys Glu Glu Cys Gly
                                                                                13

AAA GCT TTT AAG TGG TCC TCA AAG CTT ACT GAA CAT AAG ATA ACT CAT ACT GGA GAG AAA CCC TAC AAA TGT GAA GAA TGT GGC AAA GCT 1957
Lys Ala Phe Lys Trp Ser Ser Lys Leu Thr Glu His Lys Ile Thr His Thr Gly Glu Lys Pro Tyr Lys Cys Glu Glu Cys Gly Lys Ala
                                                                            14

TTT AAC CAT TTC TCA ATC CTT ACC AAA CAT AAG AGG ATT CAT ACT GGA GAG AAA CCC TAC AAG TGT GAA GAA TGT GGC AAA GCT TTT ACC 2047
Phe Asn His Phe Ser Ile Leu Thr Lys His Lys Arg Ile His Thr Gly Glu Lys Pro Tyr Lys Cys Glu Glu Cys Gly Lys Ala Phe Thr
                                                                        15

CAA TCC TCA AAC CTT ACT ACA CAT AAG AAA ATT CAT ACT GGA GAG AAA TTC TAC AAA TGT GAA GAA TGT GGC AAA GCT TTT ACC CAA TCT 2137
Gln Ser Ser Asn Leu Thr Thr His Lys Lys Ile His Thr Gly Glu Lys Phe Tyr Lys Cys Glu Glu Cys Gly Lys Ala Phe Thr Gln Ser
                                                                16

TCA AAC CTT ACT ACA CAT AAA AAA ATT CAT ACT GGA GGA AAA CCC TAC AAA TGT GAA GAA TGT GGC AAA GCT TTT AAC CAG TTC TCA ACT 2227
Ser Asn Leu Thr Thr His Lys Lys Ile His Thr Gly Gly Lys Pro Tyr Lys Cys Glu Glu Cys Gly Lys Ala Phe Asn Gln Phe Ser Thr
                                                    17

CTT ACT AAA CAT AAG ATA ATT CAC ACT GAG GAG AAA CCC TAC AAA TGT GAA GAA TGT GGC AAA GCC TTT AAG TGG TCC TCA ACC CTT ACT 2317
Leu Thr Lys His Lys Ile Ile His Thr Glu Glu Lys Pro Tyr Lys Cys Glu Glu Cys Gly Lys Ala Phe Lys Trp Ser Ser Thr Leu Thr
                                                        18

AAA CAT AAG ATA ATT CAT ACT GGA GAG AAA CCC TAC AAA TGT GAA GAA TGT GGC AAA GCT TTT AAA CTG TCC TCA ACC CTT TCT ACA CAT 2407
Lys His Lys Ile Ile His Thr Gly Glu Lys Pro Tyr Lys Cys Glu Glu Cys Gly Lys Ala Phe Lys Leu Ser Ser Thr Leu Ser Thr His
                                                    19
```

```
AAG ATT ATT CAT ACT GGA GAG AAA CCC TAC AAA TGT GAA AAA TGT GGC AAA GCT TTT AAC CGA CCC TCA AAC CTT ATT GAA CAT AAG AAA 2497
Lys Ile Ile His Thr Gly Glu Lys Pro Tyr Lys Cys Glu Lys Cys Gly Lys Ala Phe Asn Arg Pro Ser Asn Leu Ile Glu His Lys Lys
                                           20
ATT CAT ACT GGA GAG CAA CCC TAC AAA TGT GAA GAA TGT GGC AAA GCA TTT AAC TAT TCC TCA CAC CTT AAT ACA CAT AAG AGA ATT CAT 2587
Ile His Thr Gly Glu Gln Pro Tyr Lys Cys Glu Glu Cys Gly Lys Ala Phe Asn Tyr Ser Ser His Leu Asn Thr His Lys Arg Ile His
                                        21
ACT AAA GAG CAA CCC TAC AAA TGT AAA GAA TGT GGC AAA GCT TTC AAC CAA TAT TCA AAC CTT ACT ACA CAT AAC AAA ATT CAT ACT GGA 2677
Thr Lys Glu Gln Pro Tyr Lys Cys Lys Glu Cys Gly Lys Ala Phe Asn Gln Tyr Ser Asn Leu Thr Thr His Asn Lys Ile His Thr Gly
                                        22
GAG AAA CTC TAC AAA CCT GAA GAT GTG ACA GTG ATT TTG ACA ACA CCT CAA ACT TTT TCA AAC ATA AAA TAA ATTATACTGGTGAGAAATTCTAG 2772
Glu Lys Leu Tyr Lys Pro Glu Asp Val Thr Val Ile Leu Thr Thr Pro Gln Thr Phe Ser Asn Ile Lys  .

AAATGTAAAGAATGTGATAAAGGCTTTACATGGTTGTCACACTTGATTGTAGGTAAGATAATTTACATTGGAGTAAACTTCTACAAGTGTGAAGAATGTGGCAAAACTTTTAATTAATG 2891
CTCATACCTTATTGCACAGGAAAGAATTTTTACTTGGAAAAAGGTATATACACAAAGAATGTGGAAAAGCCATTAATATGTGCTCATATCTTACTCAACATCAGAGAGTCTG        3003
```

4B1-3-PRIME

```
TACTTAATAAAACCATTATAGATGCAACTAGTGTCAAACGATCTTTCAGAAAATAAAA  2753 . . . . . 1 kb . . . . . .

. . . . . TCGTCGCCAAAGATATGAGAGATTCTTTTATTAGGTGGGCATTATTTAAAACATTTTTTATGGAACAGTAAGGATATTAAAGTGTAAGATGCAGCCAGGCATGGTGGCT
CATGCCTATAATCCCAGCACTTTGGGAGGCTGAGGCCGGTGGATCACCTGAGGTCAGGAGTTTGAGACCAGCCTGACCAACATGGTGAAACCCTGTCTTTACTAAAAATACAAAAATTT
ACCAGGCACCTATAATCCCAGCTACTTCAAAGGCTGAGGCAGGAGAATCACTTGAACCTGGAAAGTGGAGGTCTCAGTGAGCCGAAATCATACCATTGCACTCCAACCTGGGCAACAAA
AGTCAAACTCCATCTCAAAAAGAAAAAAAAAGAAA
```

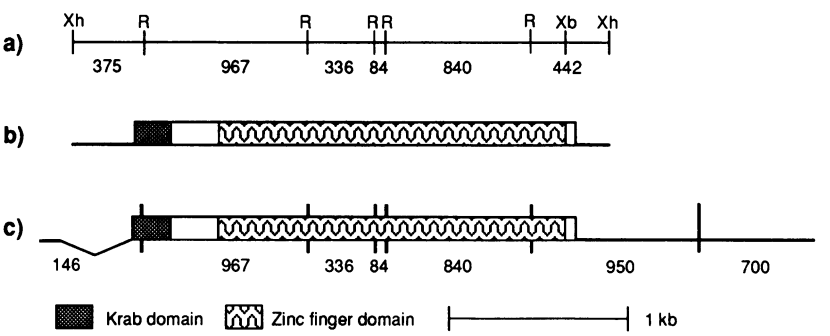

Figure 1. The DNA sequence and predicted amino acid sequence of the 2A1.2 cDNA (EMBL Accession number X59244). The 2A1.2 clone is 3.05 kb long and contains an open reading frame between nucleotides 299 and 2746, whereas the 4B1 cDNA is approximately 4 kb with an open reading frame between nucleotides 55 and 2502. The first cysteine of each of the 22 finger domains are numbered and underlined along with the next three amino acids. The 5′ sequence of the 2A1.2 and 4B1 cDNAs are in boxes above the coding region and these both continue directly into the open reading frame below. These two cDNAs are homologous 3′ of nucleotide 322 of 2A1.2 and 3′ of nucleotide 114 of 4B1, but share no homology 5′ of these positions. The consensus 3′ splice junction sequence py_nNCAGG (19) is present at the putative splice site of 2A1.2 (base 322). The open reading frame in both clones terminates at the same stop codon, at position 2747 in 2A1.2 and 2503 in 4B1. In 2A1.2 the 3′ untranslated region is 257 nucleotides long, whereas in 4B1 it is 1.3 kb. The partial sequence of the 3′ untranslated region of 4B1 is shown below the 2A1.2 sequence, directly continuing from where it overlaps with the sequence of 2A1.2. The 4B1 cDNA terminates in a poly A^+ tail, 22 nucleotides from the poly A^+ tail is a modified poly-adenylation signal (35), which is underlined. The poly A^+ tail in 4B1 is unusual in that it has two G residues. Conversely, 2A1.2 does not have a poly A^+ tail and there are no AATAAA poly-adenylation signals within the last 50 nucleotides, suggesting that this cDNA is not complete at its 3′ end. At the bottom of the figure a schematic diagram of the two cDNAs. **a)** the restriction map of the cDNA 2A1.2 inserted in the CDM8 vector, *Eco* RI (R), *Xba* I (Xb) and *Xho* I (Xh), the sizes of the *Eco* RI fragments are indicated (in base pairs) below the map (there are no *Bam* HI sites and 18 sites for *Hind* III). **b)** and **c)** show the organisation of the 2A1.2 and 4B1 cDNAs respectively. In **c)** the *Eco* RI sites are shown as vertical lines and the *Eco* RI fragment sizes are described below the figure (in base pairs).

as a longer copy of 2w6. The CEM cDNA library was then screened with a 3′ unique probe (250 bp *Xba* I-*Xho* I fragment, see figure 1) from the 2A1.2 clone which produced 51 positive clones. In both screenings of the CEM cDNA library the hybridization conditions were 6×SSC, 5×Denhardts solution, 10% dextran sulphate, 0.5% SDS, 100μg/ml salmon sperm DNA at 65°C, washed at 65°C to the stringency of 0.1×SSC.

DNA sequencing

Sequencing reactions were performed by the primed synthesis chain termination method using the Sequenase 2.0 enzyme (U.S.B.) directly from plasmid DNA. *Eco* R1 fragments of the clone 2A1.2 were subcloned into Bluescript and sequenced using the Bluescript primers KS and SK and the M13 reverse and −20 primers. The sequence was confirmed and extended using oligonucleotides homologous to internal 2A1.2 sequences. The 4B1 clone was partially sequenced using primers from within the 2A1.2 clone and from the CDM8 vector. Sequences were compiled using the Intelligenetics GEL and SEQ programs.

Cells

All cell lines were obtained from the cell production unit at the ICRF, and grown to a density of 4×10^5/ml before being harvested for RNA isolation. Non-adherent cell lines were cultured in RPMI with 10% FCS, at 37°C and 5% CO_2. Adherent cell lines were cultured in E9 with 10% FCS, at 37°C and 5% CO_2. The following human cell lines were used: ICRF-23 (embryonic lung); HeLa (cervical carcinoma); HFF (foreskin fibroblast); HL60 (acute promyelocytic leukaemia); U937 (macrophage); T-cells, HSB.2 (peripheral blood acute lymphoblastic leukaemia), Molt-4 and CEM (acute lymphoblastic leukaemia), J6 (Jurkat derivative, lymphoma); B-cells, Namalva (Burkitt's lymphoma), Mann and ROF-NL (transformed resting B lymphocyte), IM9 (multiple myeloma).

HL60 cells were grown in RPMI with 15% FCS to a density of 2×10^5 before being induced to differentiate. Morphological changes were observed within 24 hours of addition of 16nM TPA. After 3 days of treatment 85% of the cells became adherent, stopped dividing and underwent morphological changes indicative of their differentiation to a cell type with macrophage characteristics (16). During the induction of the HL60 cells to differentiate with 1μM retinoic acid all manipulations were performed in subdued light. From the increase in cell number during the retinoic acid induction it was apparent that over half of the cells were still actively dividing, however, after 3 days 30% of the cells had differentiated to metamyelocytes and neutrophiles, as observed by staining the nuclei with Geimsa (17).

DNA isolation and Southern blot analysis

DNA was isolated from U937 cells by the CsCl gradient method (18). 10μg of DNA was digested with restriction enzymes and the DNA fragments separated on 0.7% agarose gels in 1×TBE overnight. The DNA was transfered to Hybond N+ (Amersham) by blotting in 20×SSC overnight and was fixed to the membrane with NaOH. Hybridization conditions used were as for screening the T-cell library.

RNA isolation and Northern analysis

PolyA$^+$ RNA was isolated from human cell lines and human thymic tissue using Fast-track (Invitrogen), whereas the guanidine isothiocyanate method was used to obtain total RNA (18). 1% agarose-formaldehyde gels were run overnight with either 3μg of polyA$^+$ RNA or 20μg total RNA per lane and the RNA was then transfered to Hybond N (Amersham) by blotting in 20×SSC overnight. The filters were baked at 80°C for 2 hours and then U.V. illuminated for 1 min. The northern blots were hybridized in 50% formamide, 5×SSPE, 5×Denhardts solution, 0.1% SDS, 100μg/ml salmon sperm DNA at 42°C and washed at 65°C to a stringency of 0.1×SSC. The DNA probes were labeled with the multiprime labelling kit (Amersham). Filters were exposed to autoradiographic film at −70°C with intensifying screens.

```
residue
no:    2   4   6   8  10  12  14  16  18  20  22  24  26  28

                                                - - I S - - E
K - L F - - K - - - - S - C M L P H - A Q - - I - - - R  1
V N F C - - - K - - - - - - C P - I I - K - - R - N - -  2
- - - - T - - - - - - - V - - W - - R - - T - - K N Y - R  3
Y - L - - - - - - - - - - - K - - I - - T - - I - R - -  4
- - F - - - K - - A - - - - Q - - N - - E - - K - - P -  5
- - - - - - - - - - - - - - W P - N - - K - - R - - - -  6
- - - - T - - - - - - - - - Q F - N - - T - - R - - - A  7
- - F - - - T - - - E - - S R - - N - - K - - K - - - E  8
K - - - - - - - - - - - - - K W - - K - - E - - L T - - -  9
- - - - - - - - - - - - - - W P - T - - K - N R - - - -  10
- - - - - - - V - - - - - - Q F - N - - T - - R - - - A  11
- - - - - - - - - - - - - - S R - - N - - K - - K - - I E  12
K - - - - - - - - - - - - - K W - - K - - E - - I T - - -  13
- - - - - - - - - - - - - - H F - I - - K - - R - - - -  14
- - - - - - - - - - - - - - T Q - - N - - T - - K - - - -  15
- - F - - - - - - - - - - - T Q - - N - - T - - K - - - -  16
G - - - - - - - - - - - - - - Q F - T - - K - - I - - - E  17
- - - - - - - - - - - - - - K W - - T - - K - - I - - - -  18
- - - - - - - - - - - - - - K L - - T - S T - - I - - - -  19
- - - - - - - K - - - - - - - R P - N - I E - - K - - - -  20
- Q - - - - - - - - - - - - Y - - H - N T - - R - - - K  21
- Q - - - - K - - - - - - - - Q Y - N - - T - N K - - - -  22
- - L - -

con: e K p Y K C E E C G K A F n x s S x L T x H K x I H T g
```

Figure 2. Amino acid comparison of the zinc finger motifs within 2A1.2 and its consensus sequence. In the consensus sequence at the bottom of the figure: capital letters show the amino acids present in over 80% of the residue positions; small characters are used to indicate that the amino acid is present in at least 50% of the sequences. Dashes are used in the sequence to indicate the consensus sequence. The fingers are numbered 1−22 on the right and the residue positions are numbered 1−28 above the sequence (10). The four reiterated fingers are boxed.

RESULTS

Isolation of zinc finger cDNAs

The sequence TGEKPYE is usually found within zinc finger domains of the C_2H_2 type and probes from this region have been used to isolate many zinc finger genes (6). In order to identify transcription factors involved in the differentiation of haematopoietic lineages a B-cell line cDNA library was screened with a 42 base oligonucleotide probe encoding this sequence as described in Materials and Methods. To identify the cDNAs of interest the positive clones obtained were characterised by northern blot and Southern blot analysis. One clone (2w6, an 840 bp *Eco* RI fragment) was analysed further as it encoded a sequence which was strongly expressed in T and B cell lines and at a low level in macrophage and epithelial cell lines. As 2w6 was much shorter than the mRNA transcripts it identified it was used to isolate longer cDNA clones from a T-cell library. One 3.05 kb clone, 2A1.2, included all of the 2w6 sequence and was used for further studies. In order to obtain other longer cDNA clones the T-cell cDNA library was rescreened with a unique 3′ probe from 2A1.2 and a third clone of 4.0 kb (4B1) was isolated.

Sequence analysis of 2A1.2 and 4B1 cDNAs

The complete sequence of clone 2A1.2 and the partial sequence and restriction enzyme map of 4B1 confirmed that these two cDNAs were transcripts from the same gene, named ZNF43 (HGM Nomenclature). Although the two cDNAs 2A1.2 and 4B1 differed in size by a kilobase they both contained an open reading

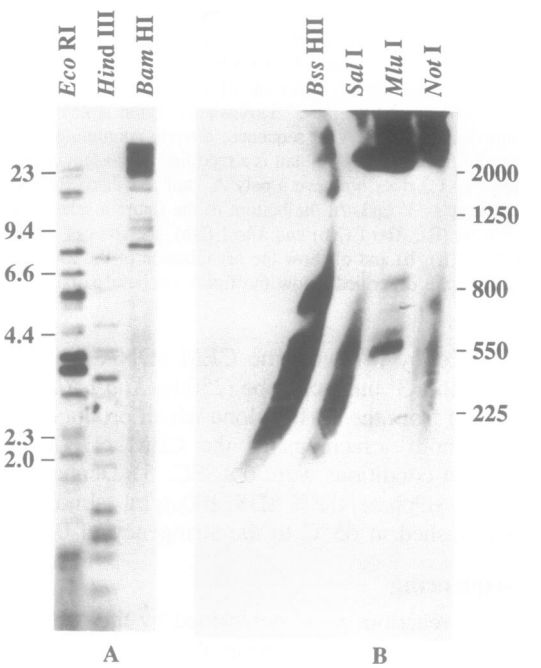

Figure 3. Genomic analysis of ZNF43. The 840 bp *Eco* RI fragment which includes fingers 12 to 21 (figure 1) was used as a probe and hybridized to genomic DNA digested with the enzymes shown on a Southern blot (**A**) and a PFGE blot (**B**) and washed to a stringency of 0.1×SSC at 65°C. The migration of the size markers is indicated.

frame of similar length (2.5 kb). The predicted protein structure of the ZNF43 gene suggested that it encoded 22 contiguous zinc finger domains of the C_2H_2 type with a short C terminal region of 18 amino acids. The cDNA 2A1.2 had a 157 amino acid N terminal region in contrast to that of the cDNA 4B1 which was 163 amino acids long (Figure 1).

When the finger sequences of the ZNF43 gene were aligned (Figure 2) several interesting characteristics were apparent. In the second, third and fourth fingers the last zinc binding histidines (residue 26) were replaced by alternative amino acids, asparagine, tyrosine and arginine respectively. These amino acid substitutions are likely to have destroyed or at least reduced the DNA binding capacity of these three fingers.

Another interesting feature was the strong homology between each of the finger domains. This conservation of the finger sequence (about 74%) was more apparent than that within most of the other sequenced finger proteins, for example 63% within the *Krüppel* protein and 52% in *Xfin* (2, 13). One other exception is Hf.12 (77%, 21). 19 amino acids were conserved in at least 18 of the 22 fingers and 5 more amino acid positions were conserved in 50% of the fingers (Figure 2). There were only four positions in which a conserved residue was not present in the majority of the fingers (residues 15, 18, 21, 24) and all of these were within the putative DNA binding region, thought to be within the helical region of the motif (residues 15–27; 10, 22). Residues 15, 17, 18 and 21 are likely to be involved in determining DNA sequence specificity according to Gibson's model of the zinc finger structure (31, personal communication, T.Gibson)

Finally when the amino acid sequence of the four zinc fingers 6, 7, 8 and 9 were aligned as a group against fingers 10, 11, 12 and 13 a 91% sequence homology was observed (boxed in Figure 2). Of the probable DNA binding residues of these repeated zinc fingers there were only two amino acid substitutions and both of these were relatively conservative changes, leucine to isoleucine and asparagine to threonine. This implies that the block of eight zinc fingers may bind to a tandemly-repeated nucleic acid sequence.

The sequences of the two cDNA clones 2A1.2 and 4B1 overlapped entirely except at the most 5′ and 3′ regions. The difference between the two clones in the 5′ region was compatible with alternatively spliced transcripts of the ZNF43 gene. The sequence preceding nucleotide 114 of 4B1 shared no homology with that 5′ of nucleotide 322 of 2A1.2. The conserved intron acceptor sequence (py_nNCAGG, 19, 23) is present at the point of divergence of 2A1.2 and 4B1, suggesting that DNA 5′ to nucleotide 322 may have been removed from 4B1. This idea is supported by the sequence of a shorter cDNA clone (HTF.6, 3 kb) independently isolated by Bellefroid (24, E.Bellefroid, personal communication). This cDNA was spliced differently to both 2A1.2 and 4B1 and consequently had a much shorter N terminal domain. However the 19 amino acids 5′ of the first methionine of 4B1 were identical to those 5′ of the first methionine of HTF.6 suggesting that both of these cDNAs contained the same putative initiation codon, which had presumably been spliced into an active position.

The sequences of 2A1.2 and 4B1 were homologous after the first in-frame initiation codon of 4B1 (nucleotide 114) such that the second in-frame initiation codon of 4B1 (nucleotide 130) was equivalent to the first in-frame initiation codon of 2A1.2 at position 338. When the sequence around the initiation codon shared by both of the cDNA clones 2A1.2 and 4B1 was compared

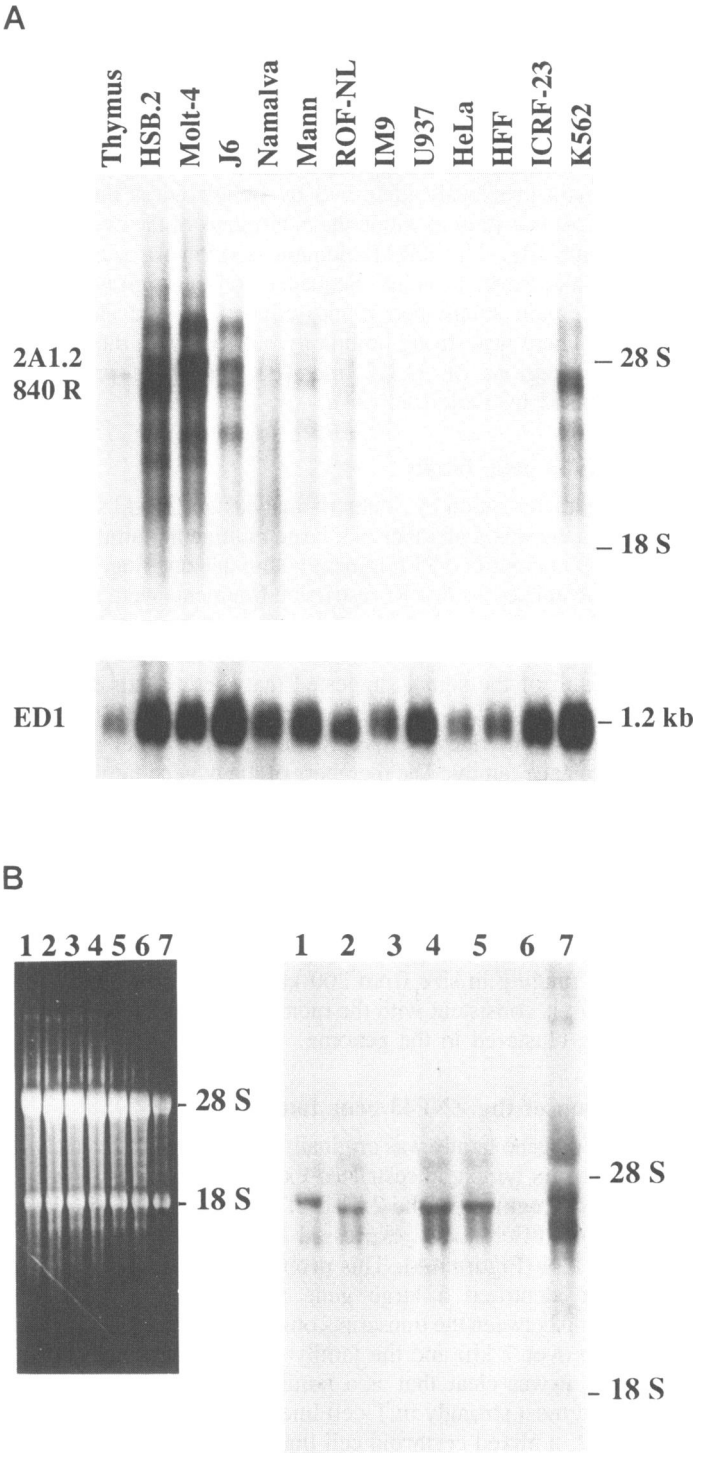

Figure 4. Expression of ZNF43. The 840 bp *Eco* RI fragment of 2A1.2 covering 10 zinc fingers was labeled and hybridized to two northern blots. In **A** Poly A^+ RNA was obtained from the cell lines indicated: T-cell lines J6, Molt-4 and HSB.2; B-cell lines IM9, ROF-NL, Mann and Namalva; K562 a mixed erythroid cell line; non-lymphoid cell lines U937, (macrophage), ICRF-23 (embryonic lung), HeLa (cervical carcinoma), HFF (foreskin fibroblast); and from thymic tissue. In **B** total RNA was extracted from the acute promyelocytic leukaemia cell line HL60 (lanes 1 and 4) and also from HL60 cells which had been induced to differentiate with retinoic acid for 1 day (lane 2) or 3 days (lane 3) or induced with TPA for 1 day (lane 5) or 3 days (lane 6) RNA from the B cell line Namalva was run in lane 7. A probe from the esterase D gene (ED1) was hybridized to the northern shown in **A** as a control for the quantity of RNA in each track (28). The RNA was uniformly loaded in the HL60 northern (**B**) as determined by both spectrophotometry and ethidium bromide staining (see figure on left).

to that of the consensus for eukaryotic translation initiation (G/A-CCAUGG, 25) only the G at position +4 was present in addition to the AUG. In contrast the first intiation codon in 4B1 included the G at position −3 as well as the G at position +4, which suggested that translation initiation of 4B1 was more likely to occur at this first initiation codon.

A domain previously identified by Bellefroid as the KRAB domain (24) was present within the N terminus of the two cDNAs 2A1.2 and 4B1. The KRAB domain is often associated with *Krüppel*-like finger proteins. Sequence conservation within the KRAB domain subdivided it into element A and element B regions. There was strong homology between both the element A and B regions of 2A1.2 and 4B1 and those previously characterised by Bellefroid (24).

The ZNF43 gene family

Initial characterisation by Southern analysis of ZNF43 suggested that this gene was a member of a large multigenic family. Using an 840 bp internal *Eco* RI fragment within the zinc finger domain as a probe at least 20 *Eco* RI restriction fragments were identified in the human genome with stringent washing conditions of 0.1×SSC at 65°C (Figure 3A). The small size of the probe and the strength of the signal suggested that these bands were not due to introns within the area covered by the probe but were most likely to be due to cross hybridization to other members of a large gene family. The members of many gene families have been shown to be clustered (ANT-C, 26; *Zfp-1* and *Zfp-4*, 27; HLA,14) and therefore the ZNF43 gene family was analysed by pulsed field gel electrophoresis (PFGE) to determine whether there was any evidence for the linkage of the members of this family. After PFGE analysis, the same 840 bp probe identified at least 5 restriction fragments in genomic DNA digested with *Bss* HII, ranging in size from 200 kb to 1000 kb (Figure 3B). This result is consistent with the members of the ZNF43 family not being clustered in the genome.

Expression of the ZNF43 gene family

The ZNF43 gene family was originally chosen for further analysis because of its lymphoid restricted expression pattern. The 840 bp *Eco* RI fragment of the 2A1.2 cDNA hybridised to a large number of differentially expressed transcripts on a poly A^+ northern blot (Figure 4A). This probe was not specific for one gene but identified a large gene family and therefore the relationship between the transcripts observed (ranging in size from 2.5 kb to over 7 kb) and the family members was not known. However it was clear that as a family the ZNF43 genes were expressed most strongly in T-cell lines as well as in B-cell lines and K562, a mixed erythroid cell line. The differentiation state of T-cell lines used was known; HSB.2 being the most immature, Molt-4 intermediate, and J6 the most mature cell line. The maturity of the B-cell lines increased from Namalva to Mann and Rof-NL and then to IM9. There was no obvious pattern of increase or diminution of bands as these cells progressed through stages of lymphoid differentiation. However the transcription of the ZNF43 gene family was too complex to establish such a relationship on a limited panel of cell lines. In thymic tissue the expression was relatively low, but this was expected as only a small proportion of these cell were T-cells, the majority being terminally differentiated thymic cells. The expression of the ZNF43 family was barely detectable in the non-lymphoid cell lines U937, HeLa, HFF and ICRF-23.

The expression of a gene involved in differentiation is expected to change as a cell proceeds along a differentiation pathway (2, 21, 29, 30), therefore we examined the effect of *in vitro* induced differentiation on the expression of the ZNF43 gene family. The induction of HL60 cells to differentiate using TPA resulted in a down regulation of the ZNF43 gene family mRNA within 24 hours (Figure 4B). No change in the expression of this gene family was observed after the induction of these cells with retinoic acid. However, in this experiment only 30% of the retinoic acid induced cells were terminally differentiated and at least half were still actively dividing, therefore any effect of retinoic acid may have been masked by the uninduced cells. Terminal differentiation of HL60 cells appears to result in a down regulation of almost every zinc finger gene that has been studied in this way (21, 29). These observations suggest that the transcription of many transcription factors is stopped following the terminal differentiation of HL60 cells but that these genes may not be directly involved in this differentiation.

Since the data showed that the ZNF43 family contained many members specific probes were necessary to obtain precise information on expression. A unique probe was isolated from the 3′ end of the 2A1.2 cDNA. This probe hybridized to only one fragment on a Southern blot (Figure 5B) and identified a subset of the transcripts previously observed by northern blot analysis (Figure 5A). These transcripts were expressed at a low level. Transcripts of around 7.0 kb were identified in the T-cell line only. The 3′ probe also hybridized to a broad band of transcripts ranging in size from 3.0 to 4.8 kb in T-cell and B-cell lines. The 4 kb 4B1 and the 3.05 kb 2A1.2 cDNA transcripts could be encompassed within this range.

We conclude from these data that the family of genes represented by the cDNA probes described has a complex pattern

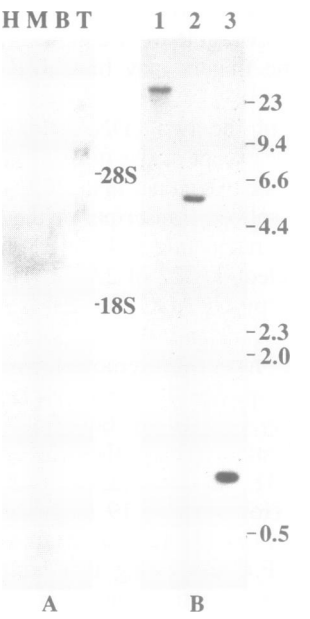

Figure 5. Southern and northern blot analysis using a unique 3′ probe. The 2A1.2 *Xba* I-*Xho* I fragment (250 bp) was hybridized to northern (**A**) and Southern (**B**) blots as described in Materials and Methods. The northern blot in **A** was of poly A^+ RNA from the cell lines; HeLa (H), epithilial; U937 (M), macrophage; Namalva (B), B-cell; Molt-4 (T), T-cell; the position of the ribosomal RNA is indicated. In **B** genomic DNA was digested with *Bam* HI (1), *Hind* III (2) and *Eco* RI (3); the migration of the λ *Hind* III markers is indicated.

of expression of extremely large transcripts. The expression of ZNF43 was shown to be almost exclusively confined to T-cell and B-cell lines.

DISCUSSION

The gene ZNF43 potentially encodes 22 contiguous zinc finger domains of the C_2H_2 type. The amino acid sequence of the finger domains is strongly conserved between the 22 fingers and a major consensus sequence could be obtained for all but four residue positions. The first finger of ZNF43 has a typical zinc finger structure but is separated from the other 18 zinc finger domains by three degenerate finger motifs. From the DNA sequence it is apparent that two other degenerate finger motifs are also present in ZNF43, one 5′ and one 3′ to the zinc finger domain. The ZNF43 gene may therefore encode a protein with only 19 functional zinc fingers, 18 in one domain, as a tandem repeat, and one zinc finger in a separate domain. The position of the non-conserved residues (15, 18, 21 and 24) within the putative DNA binding region of the zinc finger motif (10, 22) suggests that these variable amino acids are involved in sequence specific recognition of the DNA. Indeed two of these variable residues (18 and 21) have been shown in Krox-20 to combine together to identify a specific base in the DNA recognition sequence (31). Furthermore, Gibson's model of the structure of the zinc finger motif predicts that residues 15 and 17, as well as 18 and 21, are likely to be involved in determining the DNA sequence bound by a zinc finger (31). However, the conserved serine residue, position 17, is in a restricted space in Gibson's model which may limit its variability and in ZNF43 this residue is conserved in 21 of the 22 zinc fingers. Gibson has also suggested that the variable residue 24 in ZNF43 has the potential to modulate the finger position with respect to the DNA (personal communications, T.Gibson).

The expected length of DNA which could be bound by 22 zinc fingers falls in the range of 45-140 bp (10, 32, 33). The structure of the ZNF43 protein suggests that within this 45-140 bp of DNA would be an internal repeat of 8-28 bp bound by the four repeated finger domains. However it is possible that the specific DNA sequence bound by this protein maybe determined by only a few of the zinc fingers and that the rest of the finger region binds non-specifically to strengthen the DNA-protein interaction (34).

As the majority of the ZNF43 protein is composed of tandemly repeated zinc finger domains the unique N-terminal domain is the most likely region for any protein-protein interactions. Half of the N-terminal region is strongly homologous to the KRAB domain (24, personal communications, E.Bellefroid). The amino acids of the KRAB domain and the remainder of the ZNF43 N-terminal have the potential capacity to form helixes which could easily be involved in protein-protein interactions (personal communications, E.Bellefroid, T.Gibson).

From the sequence data presented here and from others (HTF.6, personal communications, E.Bellefroid, 24) three different transcription species of the ZNF43 gene have been identified, and many more may exist. The predicted protein products of the cDNAs 2A1.2 and 4B1 differ by only 5 amino acids in the N-terminal region and the effects on the protein may be minimal, however, translation of the HTF.6 RNA species would delete one of the KRAB domains, element A. It is interesting that one of the regions possibly involved in protein-protein interactions was included in only a subset of the ZNF43 transcripts (4B1 and 2A1.2) and deleted in at least one of the other transcript species (HTF.6).

The ZNF43 gene is a member of a large family of related zinc finger proteins. The cross hybridization of the 2A1.2 cDNA with a large number of human genes is characteristic of this cDNA and is not observed with all the zinc finger cDNAs studied (unpublished data, 15). This cross hybridization shows that the members of this zinc finger gene family are very closely related in sequence. Many of the members of the ZNF43 gene family are expressed specifically in lymphoid cell lines. Between the different cell lines the expression of the ZNF43 gene family is variable in both RNA transcript size and transcription level suggesting that these genes may play a role during T-cell and B-cell differentiation. The expression of one member of this family, the ZNF43 gene, is low and limited to T-cell and B-cell lines.

In conclusion, we have identified a member of a highly complex subset of the superfamily of multiple zinc finger proteins. Having characterised the genes, it will now be of interest to determine the nucleic acid motif to which the proteins bind, in order to probe their function.

ACKNOWLEDGEMENTS

We would like to thank Toby Gibson for his advice and suggestions regarding the protein structure of ZNF43, and Eric Bellefroid for his cooperation. We thank Pat Miller for providing excellent laboratory services.

REFERENCES

1. Mitchell, P.J. and Tjian, R. (1989) *Science*, **245**, 371–378.
2. Rosenberg, U.B., Schroeder, C., Preiss, A.,Cote, S., Riede, I. and Jäckle, H. (1986) *Nature*, **319**, 336–339.
3. Tautz, D., Lehmann, R., Schnürch, H., Schuh, R., Siefert, E., Kienlin, A., Jones, K. and Jäckle, H. (1987) *Nature*, **327**, 383–389.
4. Yamamoto, K.R. (1985) *Ann. Rev. Genet.*, **19**, 209–52.
5. Evans, R.M. (1988) *Science*, **242**, 889–895.
6. Bellefroid, E. J., Lecocq, P.J., Benhida, A., Poncelet, D.A., Belayew, A. and Martial, J.A. (1989) *DNA*, **8**, 377–387.
7. Green, S. and Chambon, P. (1987) *Nature*, **325**, 75–78.
8. Blumberg, H., Eisen, A., Sledziewski, A., Bader, D. and Young, E.T. (1987) *Nature*, **328**, 443–445.
9. Miller, J., McLachlan, A.D. and Klug, A. (1985) *EMBO J.*, **4**, 1609–1614.
10. Gibson, T.J., Postma, J.P.M., Brown, R.S. and Argos, P. (1988) *Protein Engineering*, **2**, 209–218.
11. Berg, J.M. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 99–102.
12. Ruiz i Altaba, A., Perry-O'Kleefe, H. and Melton, D.A. (1987) *EMBO J.*, **6**, 3065–3070.
13. Fan, C-M. and Maniatis, T. (1990) *Genes and Development*, **4**, 29–42.
14. Klein, J. (1986) Natural History of Major Histocompatibility Complex. John Wiley and Sons, Inc., USA.
15. Cunliffe, V., Koopman, P., McLaren, A. and Trowsdale, J. (1990) *EMBO J.*, **9**, 197–205.
16. Rovera, G., Santoli, D. and Damsky, C. (1979) *Proc. Natl. Acad. Sci. USA.*, **76**, 2779–2783.
17. Breitman, T.R., Selonick, S.E. and Collins, S.J. (1980) *Proc. Natl. Acad. Sci. USA.*, **77**, 2936–2940.
18. Maniatis, T., Fritsch,E.F. and Sambrook,J. (1989) Molecular cloning: A Laboratory Manual. Cold Spring Harbor University Press, Cold Spring Harbor.
19. Mount, S.M. (1982) *Nucleic Acids Res.*, **10**, 459–472.
20. Seed, B. (1987) *Nature*, **329**, 840–842.
21. Pannuti, A., Lanfrancone, L., Pascucci, A., Pelicci, P-G., La Mantia, G. and Lania, L. (1988) *Nucleic Acids Res.*, **16**, 4227–4237.
22. Lee, M.S., Gippert, G.P., Soman, K.V., Case, D.A. and Wright, P.E. (1989) *Science*, **245**, 635–637.
23. Csank, C., Taylor, F.M. and Martindale, D.W. (1990) *Nucleic Acids Res.*, **18**, 5133–5141.

24. Bellefroid, E.J., Poncelet, D.A., Lecocq, P.J., Revelant, O. and Martial, J.A. (1991) *Proc. Natl. Acad. Sci. USA.*, in Press.
25. Kozak, M. (1984) *Nature*, **308**, 241–246.
26. Lewis, E.B. (1978) *Nature*, **276**, 565–570.
27. Nadeau, J.H., Birkenmeier, C.S., Chowdhury, K., Crosby, J.L. and Lalley, P.A. (1990) *Genomics*, **8**, 469–476.
28. Squire, J., Dryja, T.P., Dunn, J., Goddard, A., Hofmann, T., Musarella, M., Willard, H.F., Becker, A.J., Gallie, B.L. and Phillips, R.A. (1986) *Proc. Natl. Acad. Sci. USA.*, **83**, 6573–6577.
29. Lania, L., Donti, E., Pannuti, A., Pascucci, A., Pengue, G., Feliciello, I., La Mantia, G., Lanfrancone, L. and Pelicci, P-G. (1990) *Genomics*, **6**, 333–340.
30. Chowdhury, K., Deutsch, U. and Gruss, P. (1987) *Cell*, **48**, 771–778.
31. Nardelli, J., Gibson, T.J., Vesque, C. and Charnay, P. (1991) *Nature*, **349**, 175–178.
32. Berg, J.M. (1990) *J. Ann. Rev. Biophys. Chem.*, **20**, 405–421.
33. Rhodes, D. and Klug, A. (1986) *Cell*, **46**, 123–132.
34. Green, S., Kumar, V., Theulaz, I., Wahli, W. and Chambon, 20. (1988) *EMBO J.*, **7**, 3037–3044.
35 Sheets, M.D., Ogg, S.C. and Wickens, M.P. (1990) *Nucleic Acids Res.*, **18**, 5799–5805.