Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 11371–11375, December 1993
Genetics

Cloning of Nrf1, an NF-E2-related transcription factor, by genetic selection in yeast

JEFFERSON Y. CHAN*, XIAO-LIANG HAN†, AND YUET WAI KAN*†

*Department of Laboratory Medicine and †Howard Hughes Medical Institute, University of California, San Francisco, CA 94143-0724

Contributed by Yuet Wai Kan, August 11, 1993

ABSTRACT **We have devised a complementation assay in yeast to clone mammalian transcriptional activators and have used it to identify a human basic leucine-zipper transcription factor that we have designated Nrf1 for NF-E2-related factor 1. Nrf1 potentially encodes a 742-aa protein and displays marked homology to the mouse and human NF-E2 transcription factors. Nrf1 activates transcription via NF-E2 binding sites in yeast cells. The ubiquitous expression pattern of Nrf1 and the range of promoters containing the NF-E2 binding motif suggest that this gene may play a role in the regulation of heme synthesis and ferritin genes.**

The human β-globin gene cluster consists of the embryonic ε-, the fetal $^{G}\gamma$- and $^{A}\gamma$-, and adult δ- and β-globin genes. Upstream from the ε-globin gene are four erythroid-specific DNase I hypersensitive sites (HS1–4) collectively termed the locus control region (1–3). Genetic results suggest that the locus control region functions as a powerful enhancer and is essential for the erythroid-specific expression of the globin genes (4–9). Among the HSs, enhancer activity appears to reside predominantly within HS2 and HS3 (10–15). Studies in transgenic mouse and transient transfection assays have localized the erythroid-specific enhancer activity within the 732-bp *Hin*dIII–*Bgl* II fragment of HS2. This fragment includes a direct sequence repeat (GCTGAGTCATGATGAGTCA) identified by *in vitro* and *in vivo* footprinting experiments (11, 16, 17). The AP1 consensus sequence (TGAGTCA) plus the bases G and C 5′ to it (GCTGAGTCA) have been shown to be an activator for the porphobilinogen deaminase gene and bind an erythroid-specific factor termed NF-E2 (18, 19). This dimeric motif is necessary for high-level expression of erythroid-specific genes in transgenic mice and transient transfection experiments (10–12).

By using a yeast expression system, we set out to clone cDNAs that encode proteins that bind to the tandem NF-E2/AP1 site and activate transcription. The cloning by complementation of genes from heterologous species in yeast has been described (20–22); however, this approach is limited by the availability of corresponding yeast mutants to isolate the gene of interest. To overcome this problem, we developed a yeast expression system in which the neomycin-resistance gene was placed under the control of the tandem NF-E2/AP1 motif. Expression of the neomycin-resistance gene is in turn dependent on the expression of transfected cDNAs that encode the appropriate activating proteins via the linked NF-E2/AP1 enhancer sequence. Using this yeast genetic selection strategy, we screened a human erythroleukemia cell line (K562) cDNA library constructed in a yeast expression vector and isolated a cDNA‡ that encodes a member of the basic leucine-zipper (bZIP) protein family that activates transcription via the NF-E2/AP1 binding site. The deduced amino acid sequence displays remarkable similarity to the transcription factors mouse NF-E2 (23) and human NF-E2 (41) and *Drosophila* CNC protein (24). We have designated the protein Nrf1 for NF-E2-related factor 1, based on its homology to the NF-E2 protein.

MATERIALS AND METHODS

Nucleic Acid Cloning and Analysis. cDNA libraries were constructed (25) from mRNA prepared from hemin-treated K562 cells (40 μM for 24 h). A yeast expression library was generated (22) in plasmid pDB20. λ phage libraries were generated in λGT10 and screened (25). DNA sequencing was performed by a modified dideoxynucleotide chain-termination method using *Taq* DNA polymerase (Promega fmol DNA sequencing system). Sequence analyses were done using the GCG sequence analysis (26). Total RNA from various cell lines were isolated, size-fractionated, and blotted on to nylon membranes (25). Northern blots containing multiple human tissue RNA samples were purchased from Clontech. Hybridizations and washings were done by standard procedures (25).

Yeast Strains, Transformations, Growth Conditions, and β-Galactosidase Assay. Standard yeast genetic methods were used (27). JCN1 strain was constructed by transformation of *Saccharomyces cerevisiae* YPH303a (*MATa leu2Δ1 his3Δ200 trp1Δ63 lys2-801 ura3-52 Gal3*$^{+}$) by electroporation with plasmid yCPN$_3$Neo (see below) (27). Transformants were grown on Ura^{-}Trp^{-} SD medium and then replated on YPD medium supplemented with G418 antibiotic (GIBCO/BRL; 50 μg/ml). Surviving colonies were grown on YPD and plasmids were extracted for a second round of transformation in yeast and also for transformation into *Escherichia coli* for further analysis. Assay for β-galactosidase activity was done by the method of Miller (28).

Plasmid Construction. The plasmid yCPN$_3$Neo, for generating the JCN1 yeast strain, was made as follows: the β-galactosidase gene in the vector pLGΔSS (29) was deleted by *Bam*HI/*Tth*1III digestion and replaced in-frame with the neomycin-resistance gene obtained by PCR amplification of the plasmid pNEO (data not shown). The *Ura3* gene and 2-μm sequences contained within the *Sma* I–*Aat* II fragment in the resulting plasmid were replaced with a *Sma* I–*Aat* II fragment containing the *Trp1* gene and *CEN–ARS* sequences from the plasmid YCplac22 (30). yCPN$_3$Neo was then obtained by insertion of the oligomer (CTCGAGCAATGCTGAGTCATGATGAGTCATGCCATTGCTGAGTCATGATGAGTCATGCCAATGCTGAGTCATGATGAGTCATGCCTCGAG) containing three tandem NF-E2/AP1 motifs (underlined) into the *Xho* I site upstream of the *CYC* minimal promoter.

Constructs for activation assay in yeast are as follows. Yeast reporter plasmid yCPLΔSS, containing the β-galactosidase gene, was obtained by replacing the *Ura3* gene and

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: bZIP, basic leucine zipper; HS, hypersensitive site; ORF, open reading frame.

‡The sequence reported in this paper has been deposited in the GenBank data base (accession no. L24123).

2-μm sequences contained within the *Sma* I–*Aat* II fragment of pLGΔSS with the *Sma* I–*Aat* II fragment containing the *Trp1* gene and *CEN–ARS* sequences from plasmid YCplac22. Plasmid yCPLN$_3$ was obtained by cloning the oligomer containing the three tandem NF-E2/AP1 binding sites into the *Xho* I site upstream of the *CYC* minimal promoter of yCPLΔSS.

***In Vitro* Translation and Immunoblots.** *In vitro* expression of Nrf1 was done using TnT reticulocyte system (Promega). Inserts containing Nrf1 sequences were cloned downstream of the T7 promoter in the plasmid pBluescript SK+ (Stratagene).

Immunoblots were done with rabbit anti-Zip and anti-Term antibodies prepared by Caltag (South San Francisco, CA). Antibodies are against synthetic peptides corresponding to residues 652–666 and 728–742, respectively, of the Nrf1 protein sequence. Horseradish peroxidase-conjugated goat anti-rabbit antibody was used as a secondary antibody and developed with diaminobenzidine (25).

RESULTS

Cloning of Human Nrf1 cDNA. Yeast strain JCN1 was transformed with a cDNA library constructed in the yeast expression vector pDB20 with cDNA derived from hemin-induced K562 cell line. Double transformants were selected first for uracil and tryptophan prototrophy on tryptophan- and uracil-deficient synthetic dextrose medium. The transformants were recovered and subsequently replated as pools to screen for neomycin-resistant colonies on a rich medium containing the antibiotic G418. We screened 100,000 double transformants, and several independent clones resistant to the antibiotic G418 were identified. DNA from these clones was isolated and used for a second round of transfection in JCN1 to verify that the neomycin-resistant phenotype was indeed the result of expression from the transfected cDNA and not due to a random mutational event in the yeast cell allowing it to propagate in the presence of G418. Two clones were isolated and found to be identical.

Nrf1 Is a bZIP Protein. A 2.1-kb cDNA clone, designated E517, contained a long open reading frame (ORF). As Northern blot analysis (see below) revealed two large transcripts of ≈5 kb, we screened a K562 cDNA library in λGT10 using 5′ and 3′ E517 cDNA subprobes and isolated three overlapping clones totaling 5 kb of cDNA. Inspection of the entire sequence revealed a single long ORF encoding potentially 742 aa, beginning at the first ATG at nt 929 and terminating at nt 3157 (Fig. 1). The ORF is preceded by 11 in-frame stop codons and is followed by two overlapping polyadenylylation signals at extreme 3′ end of the sequence. Thus, these findings suggest that the entire length of the cDNA has been obtained.

The predicted amino acid sequence contained a region near the C terminus with marked similarity to the bZIP family of transcription factors. This region is characterized by heptad repeats of leucine and hydrophobic residues within a putative amphipathic helical domain of 40 aa and is preceded by a 30-aa domain rich in arginine and lysine residues (Fig. 2). Protein sequence comparison found similarity closest to a *Drosophila* bZIP protein termed "cap and collar" (CNC) (24) and the mouse (23) and human (41) NF-E2 protein. On the basis of its homology to NF-E2, we named the protein Nrf1 for NF-E2 related factor 1 (see below).

Alignment of amino acids in the basic DNA binding domain of Nrf1 with several other bZIP proteins showed 85% (22/26)

```
   1 GGTATCTGTACGTCTTCGAACCTCCGACTTTCGTTCTTGATTAATGAAAACATTCTTGGCAAATGCTTTCGCTCTGGTCCGTCTTGCGCCGGTCCAAGAATTTCACCTCTAGCGGCGCAATACGAATGCCCCCGGCCGTCCCTCTTAATC
 151 ATGGCCTCAGTTCCGAAAACCAACAAAATAGAACCGCGGTCCTATTCCATTATTCCTAGCTGCGGTATCCAGGCGGCTCGGGCCTGCTTTGAACACTCTAATTTTTTCAAAGTAAACGTTCGGGCCCCGCGGGACACTCAGCTAAGAGCA
 301 TCGAGGGGGCGCCGAGAGGCAAGGCCGAGCTCTAGGCCGGCCGGCGGTGGCGGCGGCGAGGCCGGGACTCGGGCTTAGGGCCTGCTGTGGAGGCAGCGGCGGACGCCGAGCTAAGCAGTTTCTCTGGAAACCCCCCTGGTAAGTGTGGAG
 451 GAGGCGGGACACTCTGACCCAAGACGAAAGGCCTGTAGCTCCAGCCAAAGAAAATAAACCTTAGGAGGGAGAAGGAAAAAAAAAATCCATCAGCTGTTCCTGAGAACAGCCTGCATTGGAATCTACAGAGAGGACAACTAATGTGAGTGAG
 601 GAAGTGACTGTATGTGGACTGTGGAGAAAGTAAGTCACGTGGGCCCTTGAGGACCTGGACTGGGTTAGGAACAGTTGTACTTTCAGAGGTGAGGTGTCGAGAAGGGAAAGTGAATGTGGTCTGGAGTGTGTCCTTGGCCTTGGCTCCACA
 751 GGGTGTGCTTTCCTCTGGGGCCGTCAGGGAGCTCATCCCTTGTGTTCTGCCAGGGTGGGGTACGGGGTTTGACACTGAGGAGGGTAACCTGCTGGCTGGAGCGGCAGAGCAGTGGCCTTGATTTGTCTTTTGGAAGATTTTAAAAACCAA
 901 AAAGCATAAACATTCTGGTCCTTCAGCAATGCTTTCTCTGAAGAAATACTTAACGGAAGGACTTCTCCAGTTCACCATTCTGCTGAGTTTGATTGGGGTACGGGTGGACGTGGATACTTACCTGACCTCACAGCTTCCCCCACTCCGGGA
                                   M  L  S  L  K  K  Y  L  T  E  G  L  L  Q  F  T  I  L  L  S  L  I  G  V  R  V  D  V  D  T  Y  L  T  S  Q  L  P  P  L  R  E
1051 GATCATCCTGGGGCCCAGTTCTGCCTATACTCAGACCCAGTTCCACAACCTGAGGAATACCTTGGATGGCTATGGTATCCACCCCAAGAGCATAGACCTGGACAATTACTTCACTGCCCGGCGGCTCCTCAGTCAGGTGAGGGCCCTGGA
      I  I  L  G  P  S  S  A  Y  T  Q  T  Q  F  H  N  L  R  N  T  L  D  G  Y  G  I  H  P  K  S  I  D  L  D  N  Y  F  T  A  R  R  L  L  S  Q  V  R  A  L  D
1201 CAGGTTCCAGGTGCCAACCACTGAGGTAAATGCCTGGCTGGTTCACCGAGACCCAGAGGGGTCTGTCTCTGGCAGTCAGCCCAACTCAGGCCTCGCCCTCGAGAGTTCCAGTGGCCTCCAAGATGTGACAGGCCCAGACAACGGGGTGCG
      R  F  Q  V  P  T  T  E  V  N  A  W  L  V  H  R  D  P  E  G  S  V  S  G  S  Q  P  N  S  G  L  A  L  E  S  S  S  G  L  Q  D  V  T  G  P  D  N  G  V  R
1351 AGAAAGCGAAACGGAGCAGGGATTCGGTGAAGATTTGGAGGATTTGGGGGCTGTAGCCCCCCCAGTCAGTGGAGACTTAACCAAAGAGGACATAGATCTGATTGACATCCTTTGGCGACAGGATATTGATCTGGGGGCTGGGCGTGAGGT
      E  S  E  T  E  Q  G  F  G  E  D  L  E  D  L  G  A  V  A  P  P  V  S  G  D  L  T  K  E  D  I  D  L  I  D  I  L  W  R  Q  D  I  D  L  G  A  G  R  E  V
1501 TTTGACTATAGTCACCGCCAGAAGGAGCAGGATGTGGAGAAGGAGCTGCGAGATGGAGGCGAGCAGGACACCTGGGCAGGCGAGGGCGCGGAAGCTCTGGCACGGAACCTGCTAGTGGATGGAGAGACTGGGGAGAGCTTCCCTGCACA
      F  D  Y  S  H  R  Q  K  E  Q  D  V  E  K  E  L  R  D  G  G  E  Q  D  T  W  A  G  E  G  A  E  A  L  A  R  N  L  L  V  D  G  E  T  G  E  S  F  P  A  Q
1651 GTTTCCAGCAGACATTTCCAGCATAACAGAAGCAGTGCCTAGTGAGAGTGAGCCCCCTGCTCTTCAAAACAACCTCTTGTCTCCTCTTCTGACCGGGACAGAGTCACCATTTGATTTGGAACAGCAGTGGCAAGATCTCATGTCCATCAT
      F  P  A  D  I  S  S  I  T  E  A  V  P  S  E  S  E  P  P  A  L  Q  N  N  L  L  S  P  L  L  T  G  T  E  S  P  F  D  L  E  Q  Q  W  Q  D  L  M  S  I  M
1801 GGAAATGCAGGCCATGGAAGTGAACACATCAGCAAGTGAAATCCTGTACAGTGCCCCTCCTGGAGACCCACTGAGCACCAACTACAGCCTTGCCCCCAACACTCCCATCAATCAGAATGTCAGCCTGCATCAGGCGTCCCTGGGGGGCTG
      E  M  Q  A  M  E  V  N  T  S  A  S  E  I  L  Y  S  A  P  P  G  D  P  L  S  T  N  Y  S  L  A  P  N  T  P  I  N  Q  N  V  S  L  H  Q  A  S  L  G  G  C
1951 CAGCCAGGACTTCTTACTCTTCAGCCCCGAGGTGGAAAGCCTGCCTGTGGCCAGTAGCTCCACGCTGCTCCCGTTGGCCCCCAGCAATTCTACCAGCCTCAACTCCACCTTCGGCTCCACCAACCTGACAGGGCTCTTCTTTCCACCCCA
      S  Q  D  F  L  L  F  S  P  E  V  E  S  L  P  V  A  S  S  S  T  L  L  P  L  A  P  S  N  S  T  S  L  N  S  T  F  G  S  T  N  L  T  G  L  F  F  P  P  Q
2101 GCTCAATGGCACAGCCAATGACACAGCAGGCCCAGAGCTGCCTGACCCTTTGGGGGGTCTGTTAGATGAAGCTATGTTGGATGAGATCAGCCTTATGGACCTGGCCATTGAAGAAGGCTTTAACCCTGTGCAGGCCTCCCAGCTGGAGGA
      L  N  G  T  A  N  D  T  A  G  P  E  L  P  D  P  L  G  G  L  L [D  E  A  M  L  D  E  I  S  L  M  D  L  A  I  E  E  G  F  N  P  V  Q  A  S  Q  L  E  E
2251 GGAATTTGACTCTGACTCAGGCCTTTCCTTAGACTCGAGCCATAGCCCTTCTTCCCTAAGCAGCTCTGAAGGCAGTTCTTCCTCTTCTTCCTCCTCCTCTTCCTCTTCTTCCTCTGCTTCTTCCTCTGCCTCTTCCTCCTTTTCTGAGGA
     E  F  D  S  D] S  G  L  S  L  D  S  S  H  S  P  S  S  L  S  S  S  E  G  S  S  S  S  S  S  S  S  S  S  S  S  S  A  S  S  S  A  S  S  S  F  S  E  E
2401 AGGTGCGGTTGGCTACAGCTCTGACTCTGAGACCCTGGATCTGGAAGAGGCCGAGGGTGCTGTGGGCTACCAGCCTGAGTATTCCAAGTTCTGCCGCATGAGCTACCAGGATCCAGCTCAGCTCTCATGCCTGCCCTACCTGGAGCACGT
      G  A  V  G  Y  S  S  D  S  E  T  L  D  L  E  E  A  E  G  A  V  G  Y  Q  P  E  Y  S  K  F  C  R  M  S  Y  Q  D  P  A  Q  L  S  C  L  P  Y  L  E  H  V
2551 GGGCCACAACCACACATACAACATGGCACCCAGTGCCCTGGACTCAGCCGACCTGCCACCACCCAGTGCCCTCAAGAAAGGCAGCAAGGAGAAGCAGGCTGACTTCCTGGACAAGCAGATGAGCCGGGATGAGCACCGAGCCCGAGCCAT
      G  H  N  H  T  Y  N  M  A  P  S  A  L  D  S  A  D  L  P  P  P  S  A  L  K  K  G  S  K  E  K  Q  A  D  F  L  D  K  Q  M  S  R  D  E  H  R  A  R  A  M
2701 GAAGATCCCTTTCACCAATGACAAAATCATCAACCTGCCTGTGGAGGAGTTCAATGAACTGCTGTCCAAATACCAGTTGAGTGAAGCCCAGCTGAGCCTCATCCGAGACATCCGGCGCCGGGGCAAGAACAAGATGGCGGCGCAGAACTG
      K  I  P  F  T  N  D  K  I  I  N  L  P  V  E  E  F  N  E  L  L  S  K  Y  Q  L  S  E  A  Q  L  S  L  I [R  D  I  R  R  R  G  K  N  K  M  A  A  Q  N  C
2851 CCGCAAGCGCAAGCTGGACACCATCCTGAATCTGGAGCGTGATGTGGAGGACCTGCAGCGTGACAAAGCCCGGCTGCTGCGGGAGAAAGTGGAGTTCCTGCGCTCCCTGCGACAGATGAAGCAGAAGGTCCAGAGCCTGTACCAGGAGGT
     R  K  R  K  L  D  T  I  L  N](L) E  R  D  V  E  D (L) Q  R  D  K  A  R (L) L  R  E  K  V  E (F) L  R  S  L  R  Q (M) K  Q  K  V  Q  S (L) Y  Q  E  V
3001 GTTTGGGCGGCTGCGAGATGAGAACGGACGACCCTACTCGCCCAGTCAGTATGCGCTCCAGTACGCCGGGGACGGCAGTGTCCTCCTCATCCCCCGCACGATGGCCGACCAGCAGGCCCGGCGGCAGGAGAGGAAGCCAAAGGACCGGAG
      F  G  R  L  R  D  E  N  G  R  P  Y  S  P  S  Q  Y  A  L  Q  Y  A  G  D  G  S  V  L  L  I  P  R  T  M  A  D  Q  Q  A  R  R  Q  E  R  K  P  K  D  R  R
3151 AAAGTGAGCCTGGGGAAGAAGGGGGTTTGAAGCCCACCAAGACCGAAACTGGAGAAGGGCTGGACCTGGACCTGGACCTGGACCTACAGCGGGGACTTAAATGCCTTCTTATCCAATATATCTTCTCAGATGGGATGACTGCGGGTCAGT
      K  *
3301 GTACAGGAAGAGGCAGGCACTGGCTGGCTCAGCTCCACTCGGGTGGAGTGGAAGTGGCCAGACCATTTAGACGGACAGGGTCCTCACCCTACCCCTTTCCTGTGAGGCAGGGGTGGTGGAGTTGCTGGAGGTAGAGGAGCTATGTGG
3451 AGCAAAGGCCGACAGAGGGGAAGGAATGGACCTGTGAGAGGAAGGGAAGGTGGCAGAAAGTCTCATTTCAGGAAGGAGGGATAGAAGGAAGGAAGGAAGGAACCCCCCCCCCCCCGAAAAAAAAAATCAAAGCGGGAAGAAAATCAGAGGG
3601 AAGGTTAAGGTTGGCTCTGGCCAGGATTCCAGGCAGCAGGTTGGAGTGACTGGTGGGCCTAGATCACTGGTGTGATAAACCCCATTTCACCCCGGGGGGGGTGGGGTACACAGACACAGGGTGGGGGTGGGGAGGGGCGGTGTTAACTCT
3751 TTCTGCTCCTTGCATTTTGACATCCCTGAAGGGGAGCTCTTGGATATCATTGGCCATGTTTCAATCGAATGGAGCCACTGGGCCCCAACACTGGCTTTGAGATTTAGAGTCAAAGGGTAGAGTGAACAGGAAAGGGTCACGTGGTCCCAT
3901 GTTGCAACAGCCCCAACATACGCATGTCATTCACTGCCTTGCCACTCCATCTCCCTCCGTGCTCCAGCCACCCCTGAGCTGAGGCTCCCATTGTCTCCATCAGAGCCTGCATGTGTATGCCGTCCTCCCCTGGTCCGGTGTTTGTGTTCC
4051 CCACCCCTCACAGACTGCCTGAGCTCTTCTGTAAGCTGGGGTAGGGTGATGGCAGTGCTCCGGGAACTGGGCCTGCAGCCTTCCTCTTCTGGGACTGCTGTGAGGCAGAGGAATGATGGAGAATCTAGTGTAGCAGCCTCCAGGCAGGAT
4201 TCAGCACAACACTGGGGAGTCACCCTTCCCTCGGGCCTCTGCCTACCAACAACTGGGCTTATCACTGGGAAAACACAAAAAATTACACAACCCAGCAACAACAAAAGAACTAGTCCTCTTAGAATTTCTTGCGCTTTGATTTTTTAGGG
4351 CTTGTGCCCTGTTTCACTTATAGGGTCTAGAATGCTTGTGTTGAGTAAAAAGGAGATGCCCAATATTCAAAGCTGCTAAATGTTCTCTTTGCCATAAAGACTCCGTGTAACTGTGTGAACACTTGGGATTTTTCTCCTCTGTCCCGAGGT
4501 CGTCGTCTGCTTTCTTTTTGGGTTTCTTTCTAGAAGATTGAGAAGTGCATATGACAGGCTGAGAGCACCTCCCCAAACACACAAGCTCTCAGCCACAGGCAGCTTCTCCACAGCCCCAGCTTCGCACAGGCTCCTGGAGGGCTGCCTGG
4651 GGGAGGCAGACATGGGAGTGCCAAGGTGGCCAGATGGTTCCAGGACTACAATGTCTTTATTTTTAACTGTTTGCCACTGCTGCCCTCACCCCTGCCCGGCTCTGGAGTACCGTCTGCCCCAGACAAGTGGGAGTGAAATGGGGGTGGGGG
4801 GAAGCACTGATTCCCAGTTAGGGGGTGCCTAACTGAGCAGTAGGGATAGAAGGTGTGAACCTGGGAGTGCTTTTATAAATTATTTCCTTGTAGATTTTATTTTTAATTTATCTCTGTGACCTGCCAGGGAGAGGGGAGAGAGAGAGAGA
4951 TGCTGTTGAGCACATGACAAAATAAAATAAAATGGATGATTCAAAAAAAAAA   5002
```

FIG. 1. Nucleotide sequence and the deduced amino acid sequence of human Nrf1 cDNA. The basic region is boxed and the leucine and hydrophobic heptad repeats are circled. The acidic domain is boxed and shaded. Nucleotides surrounding AUGs that most closely agree with Kozak's consensus sequence are underlined (see text). The two polyadenylylation signals are double underlined.

```
                |----------Basic region----------||----------Leucine Zipper region----------|
h-NFE2    R D I R R R G K N K V A A Q N C R K R K L E T I V Q L E R E L E R L T N E R E R L L R A R G E A D A T L E V M R Q Q L T E L
m-NFE2    - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - S S - - - - - - - - - - - - - r - - - - - - - - - a - -
h-Nrf1    - - - - - - - - - - M - - - - - - - - - - - D - - L N - - - D V - d - Q r D K a - - - - e K v - F l r S - r q - K - k V Q s -
CNC       - - L - - - - - - - - - - - - - - - - - - - D Q - L T - - d - V n a V v k r K t q - n q d R D h L E s e r k r I s N k F a m -
JUN B     K v e - K - l R - R L - - T k - - - - - - - - r - A r - - d k V k t L k a - n a g - s s - a - l L r e Q V a q L K - k V m t H
cJUN      K a e - K - m R - R I - - S k - - - - - - - - r - A r - - e k V k t - k a q n s e - A s t a n m L r e Q V a q L K - k V m n H
JUN D     K a e - K - l R - R I - - S k - - - - - - - - r - s r - - e k V k t - k S q n t e - A s t a s l L r e Q V a q L K - k V l s H
Fos       K r r i - - e R - - M - - a k - - n - R r - l - d T - q a - t D q - e d - K s a - q t e i a n L l k e k - k L e f i - a a H
```

FIG. 2. Amino acid sequence alignment of the basic and Zip regions of Nrf1 and several members of the bZIP family. The first line shows the bZIP region of human NF-E2 (J.Y.C. and Y.W.K., unpublished data). The leucine and hydrophobic heptad repeats are indicated in boldface type, uppercase type denotes conserved amino acid changes, lowercase type denotes nonconservative changes, and dashes indicate identity. Conserved amino acid groups are (E,D), (L,M,F,A,V,I), (R,K,H), and (S,T,Q,N). h, Human; m, murine.

identity and 100% (26/26) similarity to the human NF-E2 and essentially equivalent levels of identity and similarity to CNC (Fig. 2). It is less similar to cJUN with 50% (13/26) identity and 77% (20/26) similarity and to Fos with 50% (13/26) identity and 62% (16/26) similarity. In contrast, the Zipper region showed smaller degrees of identity and similarity. The highest homology was found with NF-E2 showing 39% (14/36) identity and 72% (26/36) similarity, and the lowest was found with Fos showing only 14% (5/36) identity and 33% (12/36) similarity. Whereas homology among the different bZIP proteins is apparent in the basic and Zip regions, the similarity between Nrf1 and NF-E2 is remarkable especially in the putative DNA binding region–basic domain. Hence, these proteins are closely related to one another and probably represent a distinct subfamily of bZIP proteins.

The Nrf1 sequence includes a 34-aa block (nt 2164–2265) that is rich in acidic residues (35%) and is bracketed by a stretch of serine/threonine residues at the N terminus and by serine repeats at the C terminus. Acidic-residue-rich domains have been suggested to be important in activation of RNA polymerase II transcription factors, and a serine/threonine-rich domain has been suggested to function as a surface for protein–protein interaction for transcriptional coactivators (31, 32).

Characterization of the Protein Product Encoded by Nrf1. Conceptual translation of the ORF from the first methionine residue at nt 929 predicts a protein of 81 kDa. However, only the in-frame ATGs at nt 1799 and 1814 are in a good context for initiation based on Kozak's rules (33). If translation does indeed begin from these internal AUG codons, the proteins will have a predicted size of 50 kDa. Interestingly, *in vitro* transcription and translation of the entire coding region of Nrf1 revealed two products—a major product of 110 kDa and a minor product of 65 kDa (Fig. 3*a*, lane 1). Translation of a plasmid containing the partial clone, E517, lacking nt 1–1780 that contains the first ATG codon of Nrf1, resulted in a shorter peptide of 65 kDa (Fig. 3*a*, lane 2). We presume that the minor 65-kDa product detected from translation of the full-length clone was derived from one of the internal methionines at nt 1799 or 1814. The disparity between the predicted sizes of 81 kDa and 50 kDa and the actual sizes of 110 kDa and 65 kDa, respectively, detected by SDS/PAGE may be due to the clustering of charged amino acids and serine residues in the protein that may cause them to migrate aberrantly.

Immunoblot experiments using rabbit polyclonal antibodies against two synthetic peptides (see Fig. 3*b*) were carried out to detect the endogenous Nrf1 protein in K562. A major protein with molecular mass of ≈65 kDa was detected migrating with the shorter ^{35}S-labeled *in vitro*-transcribed and translated product (Fig. 3*b*, lane 2). No specific polypeptide corresponding to the longer 110-kDa product derived from the *in vitro* transcription and translation of the plasmid containing the entire coding region was detected (Fig. 3*b*, lane 1). Absence of the larger protein may be due to differential usage of translational initiation or post-translational processing of a larger protein precursor. It is possible that the larger polypeptide predicted from the utilization of the first methionine triplet at nt 929 is by-passed in the K562 cell but served efficiently as the translational start site in an *in vitro* system or other cell types. We infer from these results that the endogenous Nrf1 protein in K562 can be derived entirely from sequence information starting from nt 1799, by-passing the first AUG codon and potential coding information for an additional 286 aa, and that the encoded 50-kDa molecule migrates at 65 kDa.

Expression of Nrf1 mRNA in Human Tissues and Cell Lines. Hybridization of RNA blots with DNA probes derived from either the 5′ or 3′ end of the Nrf1 cDNA detected two transcripts of ≈5 kb in the erythroid (K562, KU24410, and HEL) and nonerythroid (RAJI, HPBALL, 293, HELA, and Plc/Prf/5) cell lines tested (Fig. 4*a*). The molecular basis of these two transcripts may be due to alternate usage of polyadenylylation signals. In addition to the canonical

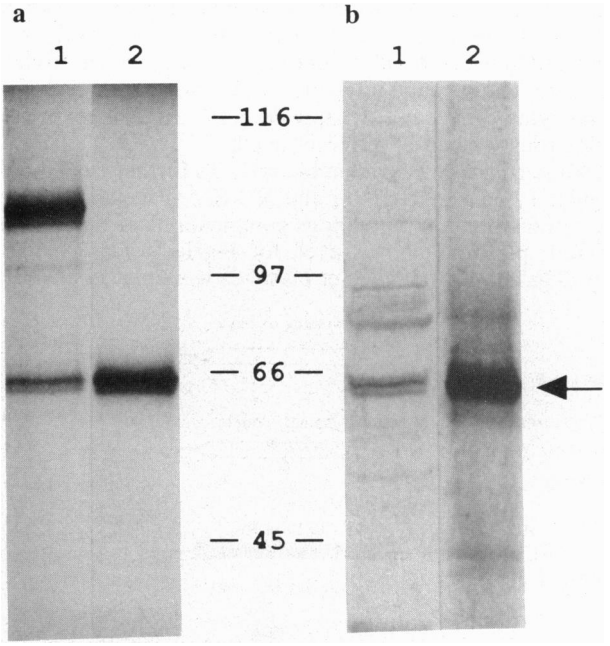


FIG. 3. (*a*) *In vitro* translation of recombinant Nrf1 in a reticulocyte lysate system. Lane 1 shows a major product and a minor product of 110 and 65 kDa, respectively, derived from expression of a plasmid containing the entire coding region of Nrf1. Lane 2 shows the 65-kDa product derived from expression of a plasmid containing the partial Nrf1 clone (see text). Sizes of the molecular mass standards are indicated (in kDa). (*b*) Immunoblot analysis of K562 whole cell extract. Cell extract (10–20 μg) was analyzed by SDS/PAGE on an 8% gel and transferred on to nitrocellulose filters for probing with anti-Zip antibody. Lane 1 shows the endogenous Nrf1 protein detected by the antibody as indicated by the arrow. Cross-reacting proteins in lane 1 are also detected by preimmune serum (data not shown). A similar result using anti-Term antibody was obtained (data not shown). Lane 2 shows comigration of the *in vitro*-expressed product derived from the plasmid as in *a*, lane 1, with the endogenous protein in K562 cells.

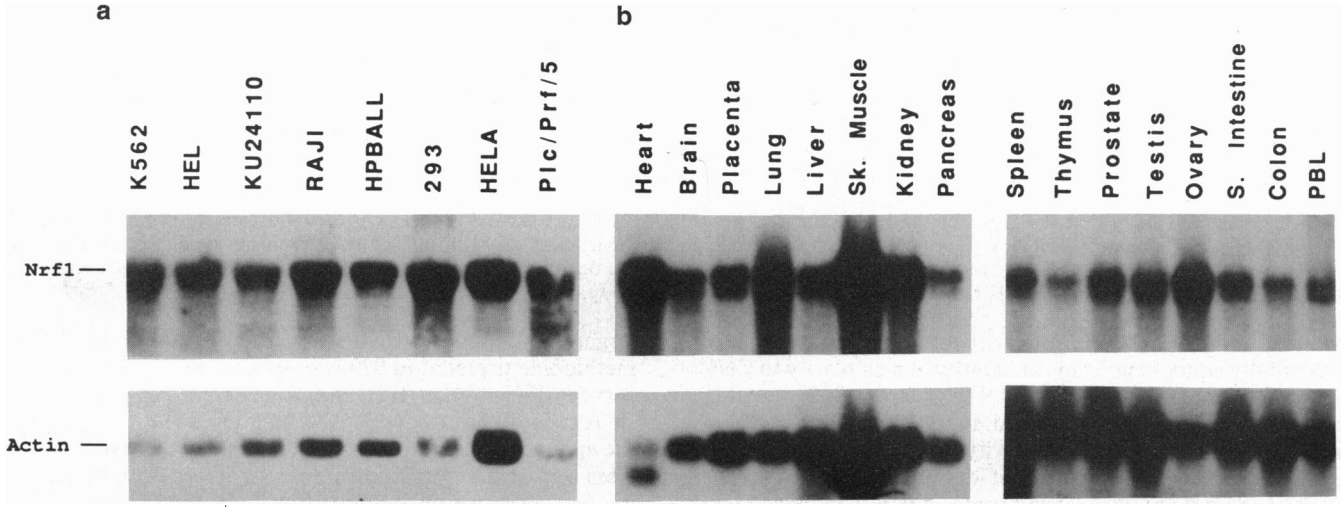
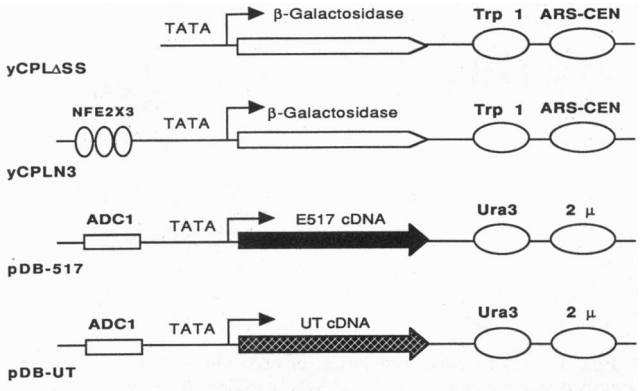


FIG. 4. Northern blot analysis of human cell lines and tissues. (*a*) RNA blot of various human cell lines. K562, HEL, and KU2241 are erythroleukemic cell lines. Raji and HPBALL are B- and T-cell lines, respectively. 293, HeLa, and Plc/Prf/5 cells are human embryonic kidney, fibroblast, and liver cell lines, respectively. The two Nrf1 transcripts (seen with lighter exposure of the autoradiogram) are ≈5 kb. Each sample contained 10 μg of total RNA. Hybridization to human β-actin probe is shown as the control. (*b*) RNA blot analysis of RNA from various human tissues (Clontech). Lanes contained 2 μg of poly(A)$^+$ RNA. Nrf1 and β-actin transcripts are indicated.

AATAAA, the 3′ untranslated region contains an unusual polyadenylylation motif (ATGAAA) at nt 4394–4400 identical to the m-*junD* motif (34) and thus provides alternative termination sites. Northern blots containing poly(A)$^+$-selected RNA from various human tissues were also examined. High levels of transcripts were seen in heart, skeletal muscle, kidney, lung, and ovary, and lower levels were detected in placenta, liver, brain, pancreas, spleen, thymus, prostate, testis, small intestine, colon, and peripheral blood leukocytes (Fig. 4*b*). It appears that Nrf1 is expressed ubiquitously albeit at different levels.

Transcriptional Activation by Nrf1. To further examine the regulatory role of Nrf1 for the NF-E2/AP1 element, yeast were transformed with various combinations of reporter and effector plasmids and assayed for β-galactosidase activity. The β-galactosidase reporter plasmids were placed under the control of *CYC1* minimal promoter plus or minus three tandem copies of NF-E2/AP1 binding site, designated yCPLN$_3$ and yCPLΔSS, respectively, in place of the *CYC* upstream activation sequence elements. Yeast expression plasmid pDB517 containing the original clone isolated from the cross-complementation experiment described above was cotransformed with the reporter constructs. A cDNA clone designated pDBUT and isolated during selection for double transformants but incapable of conferring resistance on subsequent platings of JCN1 on neomycin-supplemented plates was used as a negative control. As shown in Fig. 5, only pDB517 and yCPLN$_3$ double transformants showed a large enhancement of β-galactosidase activity. In contrast, yCPLN$_3$ alone and double transformants containing pDBUT and yCPLN$_3$ or pDB517 and yCPLΔSS plasmids showed no enhancement. Thus, the E517 protein product activates via the NF-E2/AP1 element.



Effector Plasmid	Reporter Plasmid	β-Gal activity
	yCPLN3	5.3 ± 4.0
pDB-517	yCPLN3	292.5 ± 46.5
pDB-517	yCPLDSS	5.0 ± 1.0
pDB-UT	yCPLN3	1.4 ± 0.5

FIG. 5. Nrf1 protein product activates transcription in yeast. (*Upper*) Schematic representations of plasmids. (*Lower*) β-Galactosidase activities of yeast (see text for explanation). β-Galactosidase activities are expressed as units (34).

DISCUSSION

We have isolated a human gene that encodes a protein, Nrf1, with homology to the bZIP family of proteins by cross-species complementation in yeast. The deduced amino acid sequence is notable for domains characteristic of bZIP proteins and for its remarkable homology to mouse NF-E2. Although the nucleotide sequence of Nrf1 predicts a protein of 742 aa with an expected size of 81 kDa, we have not observed such a protein in immunoblots of K562 whole cell or nuclear extracts using two rabbit polyclonal antibodies against synthetic peptides. Instead, a protein of 65 kDa was detected. In agreement with these results, the endogenous protein was detected by immunoblot experiments at 65 kDa, similar to products derived from rabbit reticulocyte extracts of *in vitro*-transcribed RNA from the partial cDNA clone. Although the nucleotide sequence revealed two internal AUGs to be in good consensus positions for translational initiation, it is not clear whether the smaller peptide is the result of preferential usage of these internal start codons. If an internal AUG is used, the predicted 5′ untranslated region would be >1.5 kb. The function, if any, of such a large untranslated region is unknown. A possible explanation for the size disparity between the protein observed and that expected from the gene sequence is the deficiency of post-translational modification in the reticulocyte lysate system. However, we excluded this possibility by showing that the

endogenous and *in vitro*-translated proteins display similar mobility in SDS/PAGE. Thus, we believe that this protein migrates anomolously due to intrinsic properties. Atypical behavior in SDS/PAGE gels has been observed for c-fos, c-myc, and protein 4.1 (35–37).

The ability of Nrf1 to activate transcription through the tandem NF-E2/AP1 repeat in HS2 was confirmed in K562 (J.Y.C., unpublished data) and yeast cells. Expression of Nrf1 from the transfected plasmid increased the expression of a cotransfected reporter plasmid in a NF-E2/AP1-motif-dependent manner. A striking feature of the region immediately N-terminal to the bZIP domain is the abundance of acidic residues flanked by multiple serine and threonine residues that is reminiscent of activation domains of Sp1 and Gal4 (31). A marked reduction of activity was seen when this region was truncated (J.Y.C., unpublished data).

In contrast to NF-E2, which is restricted to the erythroid, megakaryocytic, and mast cell lineages (23), we have found that Nrf1 is expressed ubiquitously. Thus, the cellular distribution of Nrf1 and NF-E2 shows an interesting parallel to the GATA-binding protein family (38). Although expression of GATA-1 and murine NF-E2 is restricted to similar lineages, other members of the family such as GATA-2, 3, and 4 are more widely expressed. Whether both Nrf1 and NF-E2 demonstrate similar binding specificities is yet to be determined. However, the remarkable conservation of the putative DNA binding domain (basic region) between NF-E2 and Nrf1 suggests that they probably bind to very similar, if not identical, sequences. What then is the mechanism for transactivating specificity between Nrf1 and NF-E2? Perhaps sequences outside the bZIP domains impart specificity by providing appropriate protein–protein interaction in the transcription complex. Such a mechanism has been postulated for Oct1 and Oct2 proteins (39). Alternatively, the specificity may be determined by association with their respective dimerization partner, as Nrf1 and NF-E2 are bZIP proteins. The ubiquitous distribution of Nrf1 raises the question as to what role, if any, it plays in globin gene expression. The range of promoters containing a NF-E2 binding sites includes several genes involved in the synthesis of heme and the iron-storage protein ferritin (23). As heme and iron play diverse roles including oxygen transport, prostaglandin synthesis, inactivation of oxygen radicals, and as prosthetic groups in cytochrome P450 enzymes, they are found in all tissues (40). Hence, we speculate that Nrf1 may play a role in the gene expression of heme biosynthetic enzymes and the iron storage protein ferritin.

We thank Dr. L. Guarente for the pDB20 plasmid, and Drs. K. Shannon and B. Yen for helpful discussion and reading of the manuscript. This work was supported in part by National Institutes of Health Grant DK16666. Y.W.K. is an Investigator of the Howard Hughes Medical Institute.

1. Tuan, D., Solomon, W., Li, Q. & London, I. M. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 6384–6388.
2. Forrester, W. C., Thomson, J., Elder, T. & Groudine, M. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 1359–1363.
3. Orkin, S. H. (1990) *Cell* **63**, 665–672.
4. Van der Ploeg, L. H. T., Konings, A., Oordt, M., Roos, D., Bernini, L. & Flavell, L. A. (1982) *Nature (London)* **283**, 637–642.
5. Taramelli, R., Kioussis, D., Vanin, E., Bartram, K., Groffen, J., Hurst, J. & Grosveld, F. G. (1986) *Nucleic Acids Res.* **14**, 7017–7029.
6. Curtin, P. T., Pirastu, M., Kan, Y. W., Gobert-Jones, J. A., Stephens, A. D. & Lehmann, H. H. (1985) *J. Clin. Invest.* **76**, 1554–1558.
7. Driscoll, M. C., Dobkin, C. S. & Alter, B. P. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 7470–7474.
8. Grosveld, F., Blomvan, Ossendelft, G., Greaves, D. & Kollias, G. (1987) *Cell* **51**, 975–985.
9. Behringer, R. R., Ryan, T. M., Reilly, M. D., Asakura, T., Palmiter, R. D., Brinster, R. L. & Townes, T. M. (1989) *Science* **245**, 971–973.
10. Ney, P. A., Sorrentino, B. P., McDonaugh, K. T. & Neinhuis, A. W. (1990) *Genes Dev.* **4**, 993–1006.
11. Moi, P. & Kan, Y. W. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 9000–9004.
12. Collis, P., Antoniow, M. & Grosveld, F. (1990) *EMBO J.* **9**, 233–240.
13. Philipsen, S., Talbot, D., Fraser, P. & Grosveld, F. (1990) *EMBO J.* **9**, 2159–2167.
14. Talbot, D., Philipsen, S., Fraser, P. & Grosveld, F. (1990) *EMBO J.* **9**, 2169–2178.
15. Curtin, P. T., Lui, D., Lui, W., Chang, J. C. & Kan, Y. W. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 7082–7086.
16. Ikuta, T. & Kan, Y. W. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 10188–10192.
17. Strauss, E. C. & Orkin, S. H. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 5809–5813.
18. Landschulz, W. H., Johnson, P. F. & McKnight, S. C. (1988) *Science* **240**, 1759–1764.
19. Mignotte, V., Eleouet, J. F., Raich, N. & Romeo, P. H. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 6548–6552.
20. Lee, M. G. & Nurse, P. (1987) *Nature (London)* **327**, 31–35.
21. Fikes, J. D., Becker, D. M., Winston, F. & Guarente, L. (1990) *Nature (London)* **346**, 291–294.
22. Becker, D. M., Fikes, J. D. & Guarente, L. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 1968–1972.
23. Andrews, N. C., Erdjument-Bromage, H., Davidson, M. B., Tempst, P. & Orkin, S. H. (1993) *Nature (London)* **362**, 722–727.
24. Mohler, J., Vani, K., Leung, S. & Epstein, A. (1991) *Mech. Dev.* **34**, 3–10.
25. Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab. Press, Plainview, NY).
26. Genetics Computer Group (1991) GCG Sequence Analysis Software Package (GCG, Madison, WI), Version 7.0.
27. Guthrie, C. & Fink, G. R. (1991) *Methods Enzymol.* **194**, 182–187.
28. Miller, J. H. (1972) *Experiments in Molecular Genetics* (Cold Spring Harbor Lab. Press, Plainview, NY).
29. Gaurente, L. & Hoar, E. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 7860–7864.
30. Geitz, R. D. & Sugino, A. (1988) *Gene* **74**, 527–534.
31. Mitchell, P. J. & Tjian, R. (1989) *Science* **245**, 371–381.
32. Pascal, E. & Tjian, R. (1991) *Genes Dev.* **5**, 1646–1656.
33. Kozak, M. (1987) *Nucleic Acids Res.* **15**, 8125–8148.
34. Hirai, S. I., Ryseck, R. P., Mechta, F., Bravo, R. & Yaniv, M. (1989) *EMBO J.* **8**, 1439–1443.
35. Van Beveren, A., Straaten, F. V., Curran, T., Muller, R. & Verma, I. M. (1983) *Cell* **32**, 1241–1255.
36. Watt, R. A., Shatzman, A. R. & Rosenberg, M. (1985) *Mol. Cell. Biol.* **5**, 448–456.
37. Conboy, J., Kan, Y. W., Shohet, S. B. & Mohandas, N. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 9512–9516.
38. Orkin, S. H. (1992) *Blood* **80**, 575–581.
39. Tanaka, M. & Herr, W. (1990) *Cell* **60**, 375–386.
40. Abraham, N. G. (1991) *Blood Rev.* **5**, 19–28.
41. Chan, J. Y., Han, X.-L. & Kan, Y. W. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 11366–11370.