Order Number 9315490

Subtractive cloning of an erythroid-specific transcription factor and other murine erythroleukemia cell genes

Miller, Ira J., Ph.D.

City University of New York, 1993

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

Subtractive Cloning of an Erythroid-Specific Transcription Factor and other Murine Erythroleukemia Cell Genes

by

Ira J. Miller

A dissertation submitted to the Graduate Faculty in Biomedical Sciences in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York.

1993

This manuscript has been read and accepted for the Graduate Faculty in Biomedical Sciences in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

12/7/92
Date

Chair of Examining Committee

12/7/92
Date

Executive officer

Dr. Francesco Ramirez
Dr. Jonathan Licht
Dr. Xin Yuan Fu
Dr. Frank Costantini
Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

SUBTRACTIVE CLONING OF AN ERYTHROID-SPECIFIC TRANSCRIPTION FACTOR AND OTHER MURINE ERYTHROLEUKEMIA CELL GENES

by

Ira J. Miller

Adviser: Dr. James Bieker

The process of erythroid cell specialization involves the coordinate control of structural, enzymatic and regulatory genes. The most popular murine model of erythroid differentiation is the Friend murine erythroleukemia (MEL) cell, which undergoes terminal maturation upon chemical induction. We used the Friend model to ask what genes are expressed in erythroid cells and not in a macrophage cell line by performing subtractive hybridization. We isolated ten novel sequences and eight previously known nonglobin genes that are differentially expressed. We concentrated on the most narrowly expressed transcripts based on Northern blot analysis of tissue RNAs. The most erythroid-specific clone in the group is a 1.4kb transcript with a heterogeneous startsite, expressed only in the spleen and bone marrow- the physiologic organs of adult murine hematopoiesis. It is not expressed in T-cell, B-cell monocyte or macrophage lines, but is present at low levels in mast cell lines. We propose that it encodes a transcription factor on the basis of the deduced aminoacid sequence. The carboxy terminal domain contains three TFIIIA-like $(cys)_2$-$(his)_2$ zinc finger motifs highly homologous to the those of the Krüppel family of DNA-binding proteins, prompting us to name it EKLF- for Erythroid

Krüppel-like Factor. The amino 2/3 of the EKLF is acidic and proline rich, both features of functional transcriptional activating domains. On the basis of its homology to Sp1, EKLF is a candidate gene for a previously described factor that binds to "CAC" sites within erythroid specific promoters.

Two other novel sequences displayed provocative homologies. One of them contains a repeated motif found in several guanosine-binding proteins and in the Purkinje cell specific protein L7, and it may play a role in second messenger signaling. On Northern blots,it detects transcripts of 4kb in brain, heart and ovary, and 2.5kb in the heart. The other is a 1,567 nucleotide transcript encoding a protein homologous to the acidic nuclear X-linked lymphocyte-regulated protein. The MEL gene is also present in T-cell line EL-4 and at low levels in most tissues.

ACKNOWLEDGEMENTS

I would like to acknowledge my preceptor Jim Bieker whose patience and faith in my ultimate success are rare teaching qualities in the current competitive scientific environment. He never discouraged me from following my own inclinations nor found fault in my failures, and was always available for advice. I am also obliged to Ute Hochgeshwender and Miles Brennan for their technical advice and moral support which catalyzed my productivity after a slow start. I am grateful to Bill Scher for instructions on culture and induction of various MEL cell strains which he provided, to Jay Unkeless for the J774 line, and particularly to Ed Siden for passing to me all of the other cell lines used, along with recommendations for their culture conditions. I thank Francesco Ramirez for ongoing advice and the members of his lab for sharing their reagents and samples. I also acknowledge my cohorts Tina, Stephanie, and Cherie for providing the lab with humor and a sense of perspective. I am also grateful to Jonathan Licht for his help with DNA-binding experiments currently in progress.

I thank my father who serves as a source of personal inspiration, and my mother for letting me get a Ph.D. I am beholden to my wife, Sari, for getting me to follow a normal circadian schedule, and both of our close families for their love and support.

Table of Contents

Table of Contents (continued)

List of Figures

List of Figures (continued)

I. INTRODUCTION

The pattern of genes physiologically expressed in each cell results from the interplay of several variables which exert their influence during the process of cellular differentiation. These variables include the state of the cell's chromatin, the interaction between its signal transduction apparatus and its hormonal, cellular, extracellular and metabolic milieu, its genome, and its history. All of these determinants can contribute to the establishment of a set of transcription factors, presumably unique to each cell type, that directs its characteristic pattern of gene expression and constrains the cell's potential for subsequent differentiation (62). For example, a pluripotent bone marrow stem cell gives rise to hormonally responsive progenitors of lymphoid and myeloid lineages (65), which become further diversified to form all of the specialized cellular components of the blood. At each stage of differentiation, establishment and maintenance of the new phenotype is thought to be brought about in part by repatterning the combinatorial array of regulatory factors (89). In particular, the control of globin gene expression within the erythroid lineage has provided numerous insights into the molecular components involved in such cellular specialization.

The Friend MEL cell. The success of studies of erythroid gene regulation has relied on the availability of large amounts of transcriptionally active erythroid cells. The most popular murine model is the Friend Murine Erythroleukemia cell (MEL), which is derived from an erythroid precursor cell immortalized by infection with

Friend Virus (reviewed in (53)). The development of Friend MEL cell lines began with an attempt made in 1956 by Charlotte Friend to prove that viruses could cause tumors in adult mice as well as in neonates. She discovered a cell-free filtrate that caused massive erythroblastosis within days of injection into adult mice, with eventual outgrowth of a few malignant erythroid clones (29). These cells, blocked at a stage prior to commitment to terminal differentiation, were subsequently reared for *in vitro* passage as MEL cells. The filtrate was eventually shown to contain a complex of two biologically and physically separable viruses- a replication competent Friend Murine Leukemia Virus (MuLV), and a replication defective Friend Spleen Focus Forming Virus (SFFV) (61, 88). Neither virus carries an oncogene, but the SFFV component contains a glycosylated fusion protein, GP-55, resulting from a recombination event that occurred between open reading frames of envelope protein coding regions of the Friend MuLV and the endogenous murine Mink Cell Focus Forming Virus (22)(reviewed in (4)). GP-55 has been shown to confer a continuous replicative signal upon cells that express the erythropoietin receptor, which partly explains the target specificity of Friend Virus complex (44). Other changes shown to frequently take place in the malignant transformation process are the inactivation of the p53 (44) gene and activation of either of two members of the ETS gene family of transcription factors (5, 63). Most of the MEL cell lines continually give rise to a small percentage of terminally differentiating erythroid cells (70), but the utility of the Friend model is in the susceptibility of these cells to chemical induction. In response to a variety of agents (e.g. DMSO), cells first become committed to differentiate and then proceed with a maturation process that includes production of hemoglobin (30) and coordinate

regulation of other erythroid-specific genes (28). The process of commitment requires protein synthesis (42), and it is characterized by the accumulation of pivotal mRNAs (43) in response to a signal transduced in part by a decrease in activity of protein kinase C (55, 56).

Erythroid-specific transcription factors. Using the Friend system and other models of erythroid gene regulation, the presence of several unique transcription factors has been deduced from the interaction between their distinct binding sites and erythroid-specific components of cells or cell extracts. One of these proteins, GATA-1, was cloned from chicken (24, 72) and mouse (95) after purification on the basis of affinity for its DNA target site, a nucleotide hexamer present abundantly in the globin gene regulatory regions. Since then GATA-1 has been shown to be autoregulated (35, 96) and to be crucial for erythropoiesis (73). GATA-1 sites in promoters and enhancers are essential for tissue-specific expression of globins (33, 78) and of other erythroid-specific genes (15, 25). More recently, it has been shown that GATA-1 sites play a crucial role in the formation of the locus control regions (LCRs) that activate the α-(37) and β- (47) globin gene clusters. GATA-1 is also present in megakaryocytes (54, 79) and in mast cells (54), and GATA sites are required for mast cell expression of carboxypeptidase A (100).

A second factor shown to have a role in erythroid-specific gene expression is NF-E2. This protein binds to AP-1 sites in the erythroid porphobilinogen deaminase promoter and activates its expression (60). NF-E2 target sequences are also present in the sites of DNAse hypersensitivity within the globin LCRs (37, 47), and they have been shown to cooperate with GATA-1 sites in the formation of active LCR

domains, the powerful long distance enhancers that potentiate transcription from the globin gene clusters in a developmentally appropriate manner (92, 93). Although an LCR domain segment containing a GATA-1 site seems to be sufficient for low level expression of stable transgenes in a position independent manner, the presence of NF-E2 sites dramatically augments this basal activity (11, 93). Like GATA-1, NF-E2 is also present in megakaryocytes (79).

The goal of the present study was to identify genes for cellular components that are specific to the erythroid lineage. Some of these molecules, like hemoglobin, are effectors of erythroid cell function. Others, like GATA-1, have a role in the internal regulation of erythroid identity. To find these erythroid-specific genes, we used subtractive hybridization to isolate cDNAs specific to the Friend MEL cell. Subtractive cloning techniques have enabled others to isolate many transcriptionally regulated genes without regard to function (e.g. Myo D (20)). Here we discuss the isolation of several previously undescribed MEL-specific gene products isolated by this approach as described in detail in the "Materials and Methods" section. The "Results A" section presents the expression patterns and sequences of these genes, except that data concerning one gene product are presented separately in "Results B." The latter section concerns a novel erythroid-specific zinc finger protein whose putative role is explored in the "Discussion" section. On the basis of the homology of its finger region to that of a Drosophila body pattern-determining gap gene we have named it EKLF, for erythroid Krüppel-like factor. The last section of the thesis suggests experiments to test its predicted physiological role.

II. MATERIALS AND METHODS

Cell lines and culture. Friend MEL strain DS-19, a gift of W. Scher, was grown in T-flasks in Glasgow G-MEM (Gibco) plus 10% Fetal bovine serum (FBS). Differentiation was induced in 15% FBS by the addition of 5mM hexamethylenebisacetamide (HMBA,Sigma) for 48h and monitored by staining for hemoglobin with acidic benzidine (66). The J774 macrophage line, a gift of J. Unkeless, was cultured in D-MEM (Whittaker) plus 10% FBS in Nunclon tissue culture dishes (Nunc) and detached during passage by resuspension in phosphate buffered saline with 10mM EDTA. Macrophage cell line M1, myelomonocytic cell lines WEHI-3, T cell line EL-4, Pre-B cell lines F-4 and 3-1, and mast cell lines (84) 10P2, 10P8, 10P12, and 11P62, all generously provided by E. Siden, were cultured in RPMI + 10% FBS + 50μM 2-ME in Falcon 1058 petri dishes. All media contained L-glutamine, and were supplemented with 250units/ml penicillin-G and 200μg/ml streptomycin sulfate (Sigma). Cells were grown in a humidified chamber at 37°C, 5% C02 and were harvested at a density of 1 X 10^6 to 1.5 X 10^6 cells/ml except for 10P8, 10P12, and 11P62, which were harvested at 2 X10^5 to 6 X 10^5 cells/ml.

RNA preparation and analysis. Except as noted, standard molecular biology techniques were performed (7, 51). Total cellular RNA from D6xB2 mouse tissues or from all but one cultured cell line was isolated by a modification of a urea/LiCl method (2) in that SDS was added to 0.2% immediately before homogenizing. RNA

from DS-19 cells was prepared using a guanidium isothiocyanate/CsCl protocol (49). Poly A^+ RNA was selected by two serial passages through oligo dT sepharose (New England Biolabs) (51).

For Northern blotting, samples of total cellular RNA (10 μg/lane) were denatured and electrophoresed through a formaldehyde/agarose gel and blotted by capillary action in 10X SSC onto Hybond-N (Amersham) membranes. Shadows of the 18S and 28S ribosomal bands were visualized under short wave UV light and used as size markers and to confirm complete RNA transfer. Crosslinking to the membrane was done at 900 μJoule in a Stratalinker^{TM}. Membranes were hybridized at 65°C in a solution containing 6XSSC, 0.5% SDS, 20mg/ml hydrolized and denatured salmon sperm DNA, 10% dextran sulfate, and >2X10^6 CPM/ml random hexamer primed probe (Boehringer Mannheim). Filters were washed at high stringency (0.1X SSC, 0.1% SDS, 65°C). For Northerns, the duration of hybridization was minimized to between 2 and 4 h in order to preserve the integrity of the RNA. Membranes were stripped in boiling water and reprobed several times.

Primer extension of EKLF cDNA was performed by annealing 5μg Poly A^+ MEL DS-19 RNA with 0.5ng (100,000 CPM) ^{32}P kinased synthetic oligonucleotide C10 R3 (5' GATGGAGGGTAAGACAGTAT 3', nt 115-135, Fig. 16) in a volume of 10μl for 2h at 53°C in a buffer containing 0.6M NaCl, 40 mM PIPES pH6.5, 1mM EDTA. After ammonium acetate/ethanol precipitation, the annealed primer was extended using MMLV Reverse Transcriptase at 42°C for 2h. Products were analyzed on a 6% acrylamide sequencing gel.

Primer extension results were verified by S1 nuclease analysis, which required

isolation of the genomic EKLF clone. 300,000 plaques from λ-Dash II mouse genomic library (Stratagene) were screened with an EKLF cDNA plasmid insert (7). A liquid culture lysate of one positive plaque was used to prepare CsCl step gradient purified phage for DNA extraction. An end labeled oligonucleotide probe homologous to a region close to the 5' end of the EKLF cDNA was used to identify the Bam HI fragment from the genomic recombinant that contained the 5' end of the gene. This segment was subcloned into pGem-1(Promega), and the genomic sequence upstream of the most 5' end of the cDNA determined. An end-labeled single-stranded antisense probe of 252 bases overlapping the putative 5' end was generated by 30 cycles of Taq polymerase (Promega) -mediated extension of a kinased antisense primer (C10-R3) using 1μg of TaqI-digested plasmid DNA template. The full length probe extension product was precipitated and purified on a 6% acrylamide urea denaturing gel. Samples of either 100μg whole RNA, 10μg poly A^+ RNA or 100μg tRNA were hybridized with 50,000 CPM of probe in a volume of 40μl overnight at 65°C in a buffer containing 1M NaCl, 0.2mM EDTA, and 50mM HEPES pH7.6. Each hybridization reaction was then divided into six aliquots, and S1 nuclease digestion was carried out on each aliquot in a 300μl volume for 1h at either 30°C or 37°C and with either 100u/ml, 300u/ml or 1000u/ml S1 nuclease. The buffer contained 0.28M NaCl, 50 mM sodium acetate, 4.5 mM $ZnSO_4$ and 20mg/ml denatured salmon sperm DNA. After addition of 0.4μg tRNA as carrier, digestion products were precipitated and resolved on a 6% acrylamide sequencing gel.

DNA analysis. Genomic DNA from White Leghorn Rooster (S.P.A.F.A.S.), Strongylocentrotus pupuratus (sea urchin, Pacific Biomarine), and human were a

gift of T. Fasy. Mouse DNA was a gift of K. Andrikopoulos. Drosophila melanogaster DNA from adult (ry506) flies and Xenopus laevis DNA from adult erythrocytes were isolated by standard methods. Digested DNA was transferred by capillary action in 10X SSC to a Hybond-N membrane. The filter was hybridized as for the RNA blots and washed at high stringency, 0.2X SSC, 0.1% SDS, 65^{o}C. Sequenase enzyme (USB) was used in the dideoxy chain termination method for sequencing as specified by the supplier. Oligonucleotide primers were synthesized according to the accumulated sequence information. In addition to the subtracted cDNA library described below, MEL clones were purified from two other λ-Zap II libraries, one prepared by Stratagene, the other a gift of A. Marks.

Chromosomal localization. DNA samples from mouse/hamster chromosomal hybrid cell lines were provided by Peter Lalley. In a blind experiment, we carried out 30 cycles of PCR on 1μg of each hybrid DNA sample using two different sets of oligonucleotides derived from the EKLF cDNA sequence under conditions pretested to produce a ds DNA product of appropriate size from mouse genomic DNA and not from hamster DNA. After resolution on a TBE agarose gel, the ethidium bromide stained gel was visualized under UV, and products of the correct size were scored positive. The results were tabulated and forwarded to Dr. Lalley for interpretation.

Isolation of erythroid-specific cDNA clones.

A) Subtracted cDNA library production. Gene products expressed in monocyte-macrophage J774 were depleted from MEL DS-19 cDNA, and the

remaining DS-19-specific cDNA was used to construct a cDNA library according to the following procedure:

1. Antisense DS-19 cDNA was produced by reverse transcription of 50μg/ml DS-19 poly A+ RNA at 37°C with 10,000 units/ml MMLV Reverse Transcriptase (BRL) in the buffer provided by the supplier with the enzyme plus 0.5 mM dNTP's (incl. 5-methyl-dCTP), 1000 units/ml RNAsin (Promega), 50μg/ml actinomycin C_1(D) (Boehringer Mannheim), 40 mg/ml phospho oligo dT(Pharmacia), and 250 μCi/ml (α-^{32}P) dCTP (NEN). After phenol/chloroform extraction of the cDNA synthesis reaction, template and product were precipitated in ethanol. DS-19 template RNA was removed by hydrolysis at 68°C in 0.1M NaOH for 20 min., and the solution was neutralized by addition of 1/10 volume sodium acetate.
2. Poly A+ RNA from J774 was prepared as described above under "RNA analysis." To also eliminate the globin genes from the subtracted cDNA population, sense strand RNA was transcribed in vitro from α- and β-globin-containing pBluescript II (Stratagene) plasmids using T3 polymerase as specified by the supplier (Stratagene). A mixture of 114μg J774 poly A+ RNA and 0.5 μg each globin RNA ("driver RNA") was treated twice under a sunlamp with photoactivatable biotin (Clontech) according to the manufacturer's recommendations. This facilitated purification of the remaining nonhybridized cDNA, by allowing free RNA and RNA:cDNA hybrids to be selectively removed by binding to streptavidin and phenol:chloroform extraction (86). Driver RNA

and cDNA were then combined and precipitated in one tube.

3. A single round of cDNA selection was performed as follows: a) driver RNA was hybridized to 9μg DS-19 antisense cDNA in a 30μl reaction containing hybridization buffer (50mM HEPES pH7.6, 2mM EDTA, 0.5M NaCl) plus 0.2% SDS and 1mM desferoxamine at 65°C for 20h; b) 300μl of hybridization buffer was added at room temperature, and RNA plus RNA:cDNA hybrids were removed by 5 cycles of binding with 5μg streptavidin (BRL) for 1 min. at room temperature, followed by phenol/chloroform extraction; c) each cycle, the streptavidin/nucleic acid complex-containing interface was back extracted with 25μl hybridization buffer and aqueous phases were pooled; d) the mass of the remaining (subtracted) cDNA was determined to be 1% of the total as measured by Cerenkov counting.
4. The second strand was synthesized by hairpin loop priming of the subtracted cDNA by Klenow fragment at 12°C overnight in a buffer containing 40mM HEPES pH6.9, 65 mM KCl, 16mM $MgCl_2$, 1mM DTT, 0.5 mg/ml BSA, and 250 units/ml Klenow enzyme. The ds cDNA was blunt ended with mung bean nuclease, ligated to Eco RI linkers, digested with Eco RI and size selected on a Sepharose CL-4B column (Pharmacia). Ligation to phage arms and packaging were performed using the Lambda Zap II/Gigapack II cloning kit (Stratagene).

The resulting library consisted of 8000 primary plaques of which only 20% contained inserts based on PCR analysis of randomly picked plaques. Insert sizes averaged 1kb and ranged from 0.5 to 2kb. We chose to proceed with screening this library despite its poor titer because the 1500 true recombinant plaques obtained

are representative of about 150,000 unselected gene products.

B) Screening the MEL-Specific Library. Our method for identifying the members of the library was a modification of Palazzolo *et al.* (69) and is outlined in Figure 1, as follows:

STEP 1 Insert-containing clones were picked without further purification by screening the amplified lambda phage library at the low density of 5000 plaques/150mm plate with a complex probe made by random hexamer-primed labeling of single-stranded subtracted cDNA prepared as described above.

STEP 2. These cDNA-containing phage were individually placed in the wells of microtiter plates, eluted in SM + 7% DMSO, and stored at -80°C. Clones were preliminarily named according to their well designation. The individual members of the library were then sorted to eliminate redundant members. To do this, we used probes made from subtracted library phage by "in vivo" subcloning of λ-Zap II inserts into pBluescript SK$^-$ using R408 helper phage as described in the Stratagene manual.

STEP 3. Eco RI cDNA inserts from alkaline lysis minipreps of these plasmids were purified from agarose gels using the Geneclean DNA isolation kit (Bio 101). "Dot blot" arrays of phage DNA bound to filters were created by carefully pipetting a small amount of phage stock from each microtiter well onto top agarose-embedded bacteria, and incubating uninverted at 37°C until 5mm diameter plaques formed.

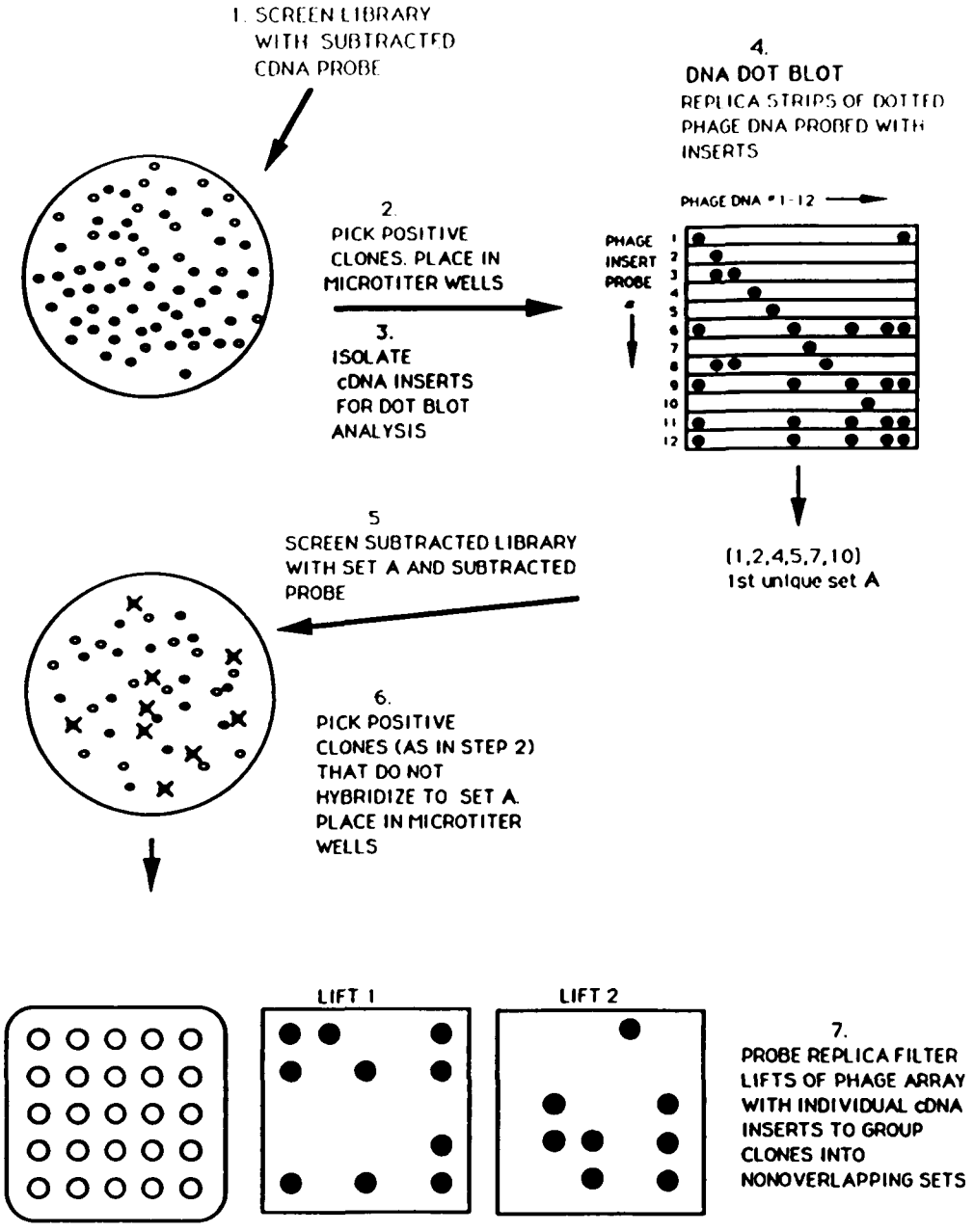

FIG. 1. Flowchart of the scheme used to identify the individual members of the subtracted cDNA library.

STEP 4. In the first round of screening these plaque arrays were absorbed onto multiple Hybond-N nylon filters (Amersham) each of which was probed with an Eco RI insert from one of the blotted phage. Membranes were hybridized at 65^{o}C in a solution containing 6XSSC, 0.5% SDS, 20mg/ml hydrolized and denatured salmon sperm DNA, 10% dextran sulfate, and >2X10^{6} CPM/ml random hexamer primed probe (Boehringer Mannheim). Filters were washed at high stringency (0.1X SSC, 0.1% SDS, 65^{o}C). UV crosslinking of the DNA to the membrane was done in a Stratalinker^{TM} at the "autocrosslink" setting while the filter was wet. This approach eliminated cross-hybridizing members from further consideration, leaving five unique, differentially expressed members (e.g. results of step 4). Sequences derived from these clones were then used to scan the GenBank (9) data bank.

STEP 5. We next used a pool of probes from this group to screen replica filters of the original library to avoid reselecting them in the second round of subtracted probe screening. Because the clones we chose in the first round were likely to be the most abundant, this early step eliminated 50% of the library from further consideration.

STEPS 6, 7. Successive rounds of sorting were then performed on this reduced population by probing replica phage arrays with individual Eco RI inserts.

As we proceeded, we were successful in obtaining increasingly rare differentially expressed gene products. Identification of the next 13 cDNAs eliminated 92% of the clones not identified in the first round. If the remaining 4% of the library consisted

of entirely differentially expressed gene products, then these would be derived from 0.04% of the MEL messenger RNAs. This portion of the library is likely to include some rarer gene products that are differentially expressed and potentially interesting, but none of the next six cDNA inserts we identified proved to detect messages that were expressed differentially when checked on Northern blots. At that point our productivity was rapidly decreasing and we elected to stop searching. In total we identified 18 unique cDNA sequences by this procedure.

We experimented with several modifications of the existing protocols for direct cloning of subtracted cDNAs. The first alteration was the inclusion of the iron chelator desferoxamine in the hybridization reaction in an effort to combat the thermal degradation of nucleic acid that might be due to ferrous-EDTA complexes. However, we did not observe that addition of this reagent had a protective effect. The second change we made was the use of 5-methyl dCTP in the synthesis of the first strand of the subtracted cDNA in order to protect the linkered cDNA from internal digestion with Eco RI. Unfortunately, DNA hemimethylated on cytosine residues does serve as a substrate for Eco RI endonucleolytic cleavage (inferred from discussion in (87)), and in fact some of the cDNA fragments we cloned proved to have been ligated to the vector at an internal Eco RI site.

Finally, a crucial step taken in our approach to screening the library was necessitated by the low percentage of the phage that contained cDNA inserts. We chose not to employ the method used by Palazzolo (69) to identify nonrecombinant phages by probing with an oligonucleotide that spans the Eco RI site of the polylinker, because the blue/white assay for β-galactosidase indicated that 90% of

the clones had some alteration in this region. Therefore, to overcome the problem of identifying cDNA insert-containing clones from this library, we simply screened it with a subtracted cDNA probe generated in the same way as the material used to make the library. We wished to concentrate primarily on the more abundant clones in any case, because from a kinetic standpoint they were the most likely to be differentially expressed. (Since rare sequences are present at low concentrations in the hybridization reaction they do not hybridize as efficiently as the more abundant gene products. Therefore, low abundance gene products that are not differentially expressed can contaminate the subtracted cDNA population. Many studies employ two successive rounds of subtractive hybridization in an attempt to further reduce the frequency of these unwanted cDNAs.)

Deciphering a spurious cloning artifact. Clone C10 was the first to be identified of the cDNA clones encoding EKLF (erythroid Krüppel-like factor, see section IIIB). Analysis of this clone presented several inconsistencies with the notion that it contains an insert derived from a single gene product, and the following observations revealed that bordering the 5' end of the EKLF sequence, clone C10 contains an unrelated cDNA fragment (data not shown):

1) Northern analysis of tissue RNA using the complete C10 insert produced a complicated pattern of bands of different sizes in different tissues, including brain (see section IIIA), none of which correspond to the length of the C10 insert.
2) The complete C10 insert is 600 bases longer than the full-length EKLF

sequence (see section IIIB).

3) Two large nonoverlapping open reading frames were deduced from the complete C10 plasmid sequence.

4) Using MEL cDNA we were unable to generate a PCR product that spans the two open reading frames.

5) A clone from a mouse brain cDNA library and five mouse genomic clones which all hybridize to the complete C10 insert do not hybridize to the EKLF portion of C10.

6) Based on sequence information from a genomic EKLF clone, the EKLF portion of clone C10 extends through 48 bp of 5' untranslated region, well upstream of an in-frame stop codon preceding the first methionine codon of the longest EKLF open reading frame (Fig. 16). Further upstream, the sequences of the EKLF genomic clone and the C10 clone diverge.

Therefore, we placed the 5' end of EKLF at the point of concurrence with the genomic sequence. The other part of the original C10 insert was renamed C10A (see section III A).

III. RESULTS

A. General MEL-Restricted cDNAs.

We identified erythroid-specific genes by applying the method of subtractive hybridization to create a library enriched for sequences restricted to the DS-19 MEL cell. As we were interested in finding genes that are as specific as possible to this bone marrow derived lineage, we used the J774 monocyte-macrophage line as the source of RNA for the subtractive hybridization. We reasoned that this myeloid descendant would have more gene products in common with the MEL cell than a more distantly related lymphoid or non-marrow-derived cell.

Estimation of differentially expressed gene prevalence in MEL. Although no direct quantitation of the cellular levels of any of these RNAs was performed, a general estimate of the efficiency of our screen can be made. In total, we isolated 18 cDNA fragments derived from gene products which are differentially expressed between MEL and J774 cells. If additional differentially expressed gene products went undetected, then the sum of their prevalences can be no more than 0.04% based on calculations made in the "Materials and Methods" section. But, as evidenced by a random sampling, this remaining population consists mostly of nondifferentially expressed rare sequences that simply escaped subtraction, as described in "Materials and Methods." (Based on the binomial distribution, the odds of choosing six nondifferentially expressed clones at random from a population with

more than 32% inserts is less than 10%, see Appendix A.) Therefore we believe that the remaining unstudied differentially expressed RNA population has a total prevalence of about 0.01%. We gain further confidence in this estimate by considering the number of times we identified each particular gene in our library. It stands to reason that in our normal routine for picking differentially expressed clones, the more times we identified any particular gene product the less likely the possibility that another product of equal abundance has gone undetected. Thirteen of the eighteen gene products were found in at least two independent clones. In particular, the prevalence in MEL cells of one of the rarer messages found, EKLF, was estimated to be about 1/20,000 based on its frequency in an unsubtracted MEL library. Since our sorting protocol revealed two independent EKLF clones in our subtracted cDNA library, we can be 75% confident that no other clone with the same frequency is present in our library. Based on all of these observations, a conservative estimate would be that we have detected the majority of MEL-specific RNAs with abundance of 1:10,000 or more.

Identification of known genes. The gene product fragments we obtained were characterized by partial sequencing and by Northern blot analysis of their tissue distribution. Most were not detectable in J774. In addition, a search of the GenBank database with partial sequences derived from these clones revealed that ten of them contained totally novel sequences. The other eight were identified to be genes previously cloned from the rat or the mouse. Some of these were expected findings: the two transcripts of the Friend Virus complex (16), the early erythroid carbonic

anhydrase I (27), GATA-1 (95) and c-myb (6). On the other hand, we were surprised to find the mRNA for a selenium binding liver protein (3), which is highly expressed in the MEL cell line and not detected in the J774 cell line (Fig. 2). This protein binds selenium coordinately, not covalently through the presence of selenocysteine, and its function is unknown. However, we can rationalize its presence with the notion that it may be necessary for development of the high levels of selenocysteine-containing glutathione peroxidase present in the mature erythrocyte. The latter enzyme is crucial for preventing hemoglobin inactivation by free radical-mediated oxidation of the iron atom in heme. Another known gene product that we found to display strict differential expression was a cDNA from the murine homologue of the rat serine protease inhibitor type 2 gene (41, 99). We can think of no obvious explanation for erythroid expression of this growth hormone-regulated gene. In addition to these seemingly normal transcription products, we detected an unusual cDNA clone, B12, whose 3' end contained a 70 bp stretch with an almost perfect match to the 5' end of the rat glutathione s-transferase transcript (90). This cDNA could not have been derived from a normal transcription product of the mouse homologue, because the MEL transcript detected on a Northern blot is 10-15kb (Fig. 3), much larger than the endogenous mRNA of the rat (700 bases). One possible explanation for this clone is that it is an artifact of transcription off an LTR of a Friend virus that inserted near the glutathione-s-transferase gene in DS-19 cells. This hypothesis could be tested by examining the DS-19 genomic DNA. None of the previously known transcripts were pursued further.

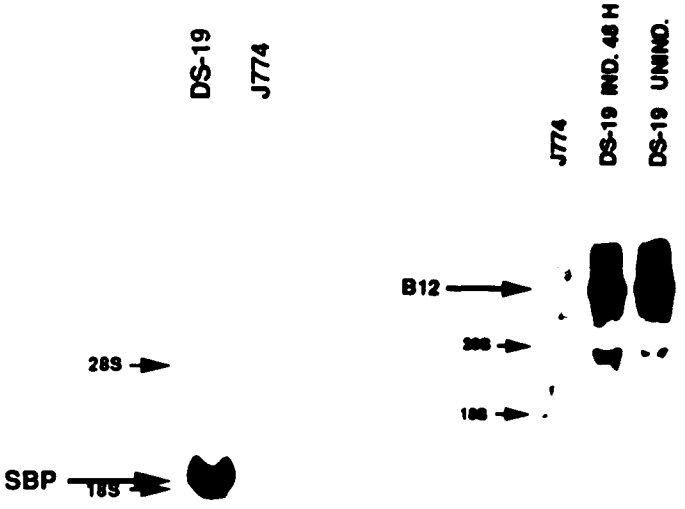

FIG. 2 (left). Northern blot analysis of Selenium binding protein (SBP) message levels in DS-19 and J774. Total RNA (10 ug) from each indicated cell line was probed with a random hexamer-labeled Eco RI cDNA insert from the subtracted library which was identified as SBP. Positions of the ribosomal RNA bands are indicated.

FIG. 3 (right). Northern blot analysis of message hybridizing to Eco RI insert of clone clone B12, described in text. Samples are the same as in Figure 2, with an extra lane containing RNA from DS-19 MEL cells induced to differentiate for 48 hours with HMBA. Positions of the ribosomal RNA bands are indicated

Tissue expression patterns of new MEL-specific cDNAs. We examined the tissue-specific expression patterns of the subtracted cDNAs in order to concentrate our efforts on the genes whose expression is most restricted to erythroid cells. The novel sequences we detected varied in the extent to which they were differentially expressed between the J774 and the DS-19 cell lines as well as in their expression among the tissues examined. No general trend of tissue-specific distribution was evident. Of the previously undescribed gene products, only one, EKLF, satisfied our criteria for erythroid-restricted expression at this level of analysis, being present only in bone marrow and spleen (see section B). Some of the other subtracted cDNA probes detected products with relatively restricted, though not necessarily erythroid patterns of expression. For example, C10A hybridizes to RNA species from only brain, heart and ovary. Interestingly, by Northern analysis, the size of this message in the ovary and brain is approximately 4kb, but the signal detected in the heart is only about 2.5kb (Fig. 4). We do not yet know whether the products detected in the heart and brain are derived from the same gene as the C10A probe.

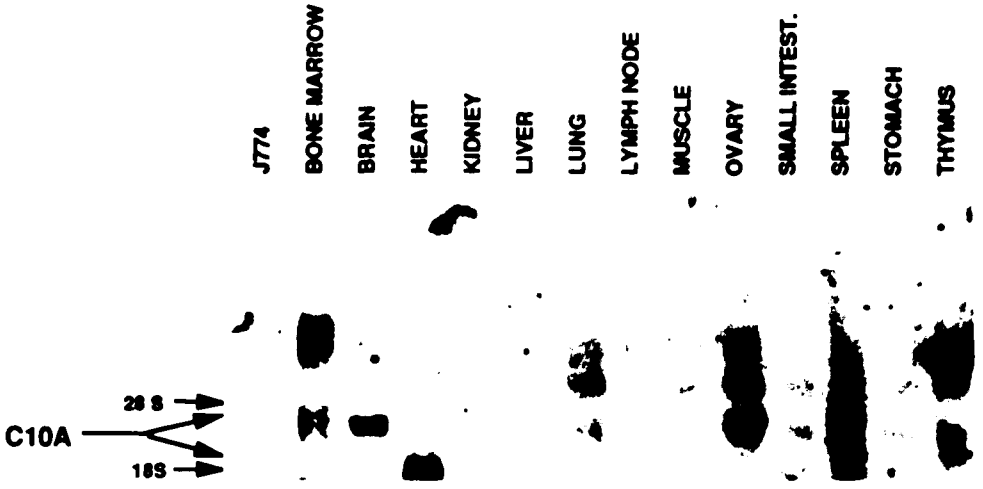

FIG. 4. Northern blot analysis of message hybridizing to C10A, the 5' end of clone C10. Total RNA (10 ug) from each indicated tissue was probed with random hexamer-labeled cDNA derived from the C10 Eco RI insert. The lower portion of the blot and the MEL RNA lane containing signals derived from the 3' end of this clone have been removed. Positions of the ribosomal RNA bands are indicated.

Two other products are relatively MEL-restricted. The first of these is A12, which detects two RNA species that are both expressed at low levels in all tissues examined (Fig. 5), but at much higher levels in MEL cell, myelomonocytic line WEHI-3, and T cell line EL-4 (Fig. 6). The smaller species was not detectable in J774, and was only detectable in one of the four mast cell lines. The other product

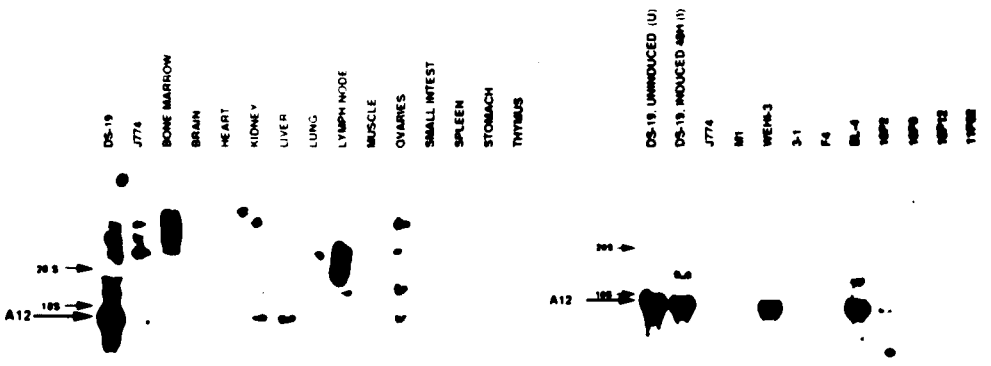

FIG. 5. (left) and FIG 6 (right). Northern blot analysis of message hybridizing to A12. Total RNA (10 ug) from each indicated tissue (left) or bone marrow-derived cell line (right) was probed with random hexamer-labeled cDNA derived from the A12 Eco RI insert. DS-19 is the MEL cell line from which EKLF was derived and was included as a positive control, and J774 is the non-EKLF-containing monocyte-macrophage cell line used for the original subtractive hybridization procedure and was included as a negative control. Positions of the ribosomal RNA bands are indicated.

whose expression is MEL-confined is A10, which is not detectable at all in J774 and is expressed at levels ten-fold higher in MEL cells than in all tissues where it is expressed (Fig. 7). In addition, A10 is down-regulated during differentiation.

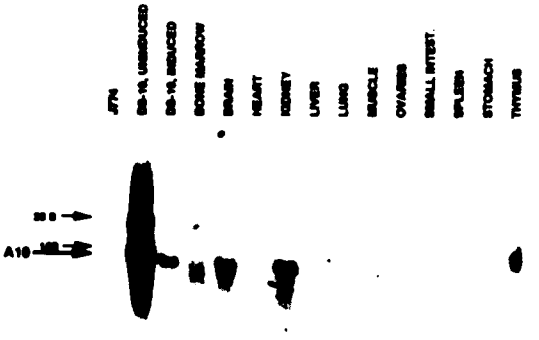

FIG. 7. Northern blot analysis of message hybridizing to A10. Total RNA (10 ug) from each indicated tissue was probed with random hexamer-labeled cDNA derived from the A10 Eco RI insert. In addition to the control RNAs of Figure 5, a sample of RNA from induced MEL (48h timepoint) was included. Positions of the ribosomal RNA bands are indicated.

The rest of the recovered subtracted cDNAs were found in a variety of tissues. G1 is expressed at about 5-fold higher levels in MEL than in J774 and is detectable in all tissues examined (Fig. 8). H5 is not detectable at all in J774 and has a scattered tissue distribution (Fig. 9). E4 is only about three times more abundant in MEL

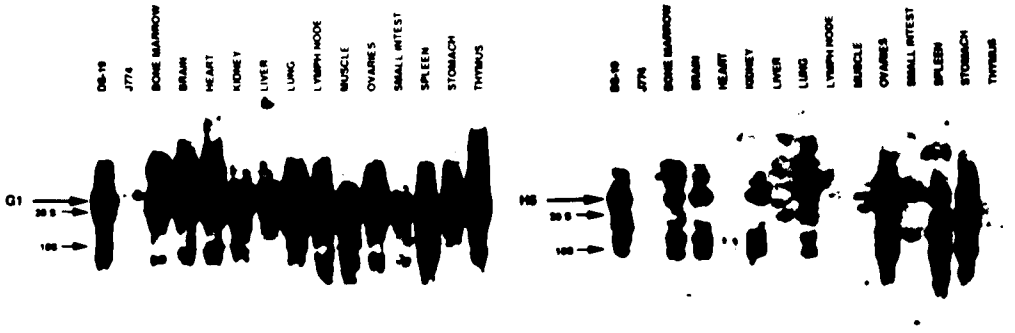

FIG. 8 (left). Northern blot analysis of message hybridizing to the Eco RI insert of clone G1. RNA samples are the same as in Figure 5. Positions of the ribosomal RNA bands are indicated.

FIG. 9 (right). Northern blot analysis of message hybridizing to the Eco RI insert of clone H5. RNA samples are the same as in Figure 5. Positions of the ribosomal RNA bands are indicated.

than J774 and is expressed in all tissues examined except the heart and liver (Fig. 10). D10 is found everywhere except J774 and bone marrow (Fig. 11). The tissue

FIG. 10 (left). Northern blot analysis of message hybridizing to the Eco RI insert of clone E4. RNA samples are the same as in Figure 5. Positions of the ribosomal RNA bands are indicated.

FIG. 11 (right). Northern blot analysis of message hybridizing to the Eco RI insert of clone D10. RNA samples are the same as in Figure 7. Positions of the ribosomal RNA bands are indicated.

expression patterns of G6 (Fig. 12) and H9 (not shown) were not assessed.

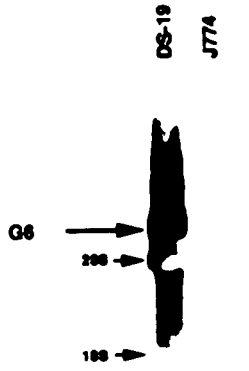

FIG. 12. Northern blot analysis of message hybridizing to the Eco RI insert of clone G6. RNA samples are as in Figure 2. Positions of the ribosomal RNA bands are indicated.

Indeed, we are puzzled by the tissue distribution of most of the differentially expressed gene products. In particular, we had expected that the novel MEL-derived sequences which are not expressed at all in J774 would tend to be more erythroid-restricted than they turned out to be. In particular, we expected highest levels of expression in bone marrow and spleen because those are the sites of physiological erythropoiesis in the adult mouse (58). But of all the detected cDNAs which survived the subtraction, only carbonic anhydrase type I (see section B) and EKLF (see section B) strictly satisfy this criterion, although the GATA-1 transcription factor (95) would have shown a similar expression pattern on Northern analysis of tissue samples. In fact, in the case of clone D10 the opposite is true: bone marrow levels are lower than those in all other tissues examined. Also, clone H5 which is not detected in J774, reaches the highest levels of expression in the ovaries. Moreover,

we cannot even find a tissue that consistently expresses the set of genes in question at low levels like J774 does.

These observations are not meant to belittle the number of physiologically erythroid-restricted messages. Of course, the cDNAs for α- and β-globin were known in advance to be strongly differentially expressed between the two lines and were specifically eliminated during the subtraction, as described in "Materials and Methods." In addition, many effector molecules of erythroid function including the cytoskeletal protein spectrin and the membrane protein Band III, do not become prevalent until after terminal differentiation of MEL cells (28), so we did not expect to find them. Nevertheless, we find the lack of clear erythroid restriction for this set of genes to be puzzling.

Several reasons can be given to account for the varied tissue expression patterns of the subtracted gene products. One possibility is that their regulation is governed by a complex interplay of variables that is not consistently patterned from tissue to tissue among the genes examined. For if they were regulated by a simple combinatorial array of transcription factors, then one would expect that some tissues would tend to consistently have low or high levels of many of the genes studied. Another possible reason for these genes lacking a predictably confined tissue expression pattern could be that some of the gene products may, in fact, be limited to a lineage defined subset of cells, but that particular lineage is present to varying degrees in different organs (e.g. smooth muscle). RNA samples from more uniform cell populations would be needed to discern their distribution on a cellular level. Alternatively, some of these gene products may simply be specifically down-

regulated in the macrophage lineage. However, if the recovery of an MEL-specific cDNA is due to a viral insertion event in MEL cells that activated transcription of a nearby gene, then we would expect higher levels of expression in MEL than in most mouse tissues, and not necessarily absence of expression in J774. In addition, the transcript size might be altered on a Northern. With the possible exception of B12, none of the products meet all of these criteria. Despite the lack of a consistent trend of expression among these genes, the analysis itself served to direct further energy to pursuing the one novel clone whose expression was highlighted by this analysis: EKLF (see section B).

In retrospect, it seems that as a source of RNA for the subtractive procedure, the J774 cell line had its advantages and its drawbacks. To its merit, a single round of hybridization removed 99% of the starting MEL cDNA population. We can therefore conclude that J774 expresses the vast majority of RNAs present in MEL. On the other hand, we could have used RNA from a more distantly related tissue such as liver or brain as the source of driver. Then we would expect that a higher percentage of MEL cDNAs would have survived the subtraction. Maybe such a population would tend to show a more generalized trend of tissue distribution by Northern blot analysis, being confined to myeloid-related organs in general.

Sequence analysis of the subtracted cDNAs. As discussed in "Materials and Methods," the clones that we obtained contained cDNA inserts fragmented by thermal and endonucleolytic cleavage. For our purposes we considered the gene products with the most restricted expression to be the most interesting. Therefore, partial nucleotide sequences were obtained for all of the differentially expressed clones, but

only the inserts of clones C10A, A12, and EKLF (section B) were sequenced to completion in both directions. In addition, as clone A12 was determined to be missing regions on both 5' and 3' ends, the entire open reading frame of the message was reconstructed by analyzing four additional overlapping clones from an unsubtracted MEL cDNA library. The partial sequences of the other novel clones detected are provided in the Appendix.

The nucleotide and predicted amino acid sequence of A12 is presented in Figure 13. The 5' end of the sequence shown is corroborated by two independent clones purified from an unsubtracted MEL cDNA library. However, we concluded that each of these other A12-containing clones also has within it an unrelated cDNA ligated head to head with the A12 gene product, based on the observation that poly(A) tracts were present on both ends of the insert. Therefore, the 5' end presented is taken from the point of divergence between two independent cDNA clones. The length of the reconstructed sequence is 1567 nucleotides, which is close to the size predicted from the Northern blot. The 3' untranslated end is 832 nucleotides long, and a polyadenylation signal is present 20 bases before the 3' end. The sequences of the two clones were in agreement except that one contained an insert of 15 bases relative to the other. We do not know whether this insertion resulted from alternative splicing or from different alleles. In either case, the protein product is likely to be unaffected by the insertion because it does not change the reading frame and is likely to be in the 5' untranslated region anyway. The longest open reading frame predicts a

```
  1 CTGCCTTGTGGAGACGAGGGACTAGGACTCCTATCTAACACAAGAAGATTCTAGCT ATG TCA AGC AGA AAA AGG AAG GCC ACT GAC ACT
  1                                                          M   S   S   R   K   R   K   A   T   D   T

 90  GCT GGC AGG CAC TCA AGG ATG GAT CCA AAT CTC TCC TCA GAT GAC AGC CAG AAC CCA GGT GCA GTC GCA GCA GCT
 12   A   G   R   H   S   R   M   D   P   N   L   S   S   D   D   S   Q   N   P   G   A   V   A   A   A

165  AAC AGA GAA GTC CTT GAT GCT GGT AGG GAG GAC ATT ATT TCC TCA GGA ACA GAG AGG CAA CAG GCC AGG AAG GAA
 37   N   R   E   V   L   D   A   G   R   E   D   I   I   S   S   G   T   E   R   Q   Q   A   R   K   E

240  AAA CAG GAC TTG GTC CAA GAA TTC GAA GAG CCA CGT AAC AAA GTT CTT CAG GAA AAC AGA AAA AAA TTC TCA AGG
 62   K   Q   D   L   V   Q   E   F   E   E   P   R   N   K   V   L   Q   E   N   R   K   K   F   S   R

315  ATC ATG ACC TCA TCT TTC AGT GCC ATG GAG GTC AAA ATT AAG GAT GTT CTG AAA ACC CAC TGT GAA GAG AGG CAG
 87   I   M   T   S   S   F   S   A   M   E   V   K   I   K   D   V   L   K   T   H   C   E   E   R   Q

390  AAA CTT TGT CAG GAC TAT TCT CTT CAG TTT ACA AAT TTG AAT AGG AAG TTA ACT TCG GAT GCA TAC AAA CTC AAG
112   K   L   C   Q   D   Y   S   L   Q   F   T   N   L   N   R   K   L   T   S   D   A   Y   K   L   K

465  AAA CAT GCC GAA ACA CTC TCT AAT ATG TTT ATG GAG CAA CAG AAG TTT ATT CAT GAA AGT CTC ACT CTT CAG AAG
137   K   H   A   E   T   L   S   N   M   F   M   E   Q   Q   K   F   I   H   E   S   L   T   L   Q   K

540  AAC AGA ATG AAG GAA TTT AAA TCA CTG TGT GAA AAA TAC TTG GAG AAA CTG GAG GTA CTG AGG GAT TCT CGG GGA
162   N   R   M   K   E   F   K   S   L   C   E   K   Y   L   E   K   L   E   V   L   R   D   S   R   G

615  AAT TCC ATC GCT GAA GAG CTG AGA CGT CTT ATA GCC ACC TTG GAA ATC AAA CTT CTG ATG CTG CAT AAC CAG CAA
187   N   S   I   A   E   E   L   R   R   L   I   A   T   L   E   I   K   L   L   M   L   H   N   Q   Q

690  AAG ACT GCT GCT CCT CCA CAG TCT CTC CTG GAT GTG TTA TTC TCA TAAAACTTTGGAGCACAAAAGCCGATATGAGGGGAGCAGT
212   K   T   A   A   P   P   Q   S   L   L   D   V   L   F   S

775  ATAATCATCTGGGTGAAGCTGCTAGTGCTGATGATTGCAGCTTTGGTGCCTGGATCTTCTCCTGTCTTCCTCCTGTTCCTGAAGATGCCCCCGTTCTGTTA

875  AATCCAAATAAACCCTGGGATTTGTGTGACAGCAGTGCAAATATGTCTGTAGGGAACTCCAGCCCTCAGGTAGAGATGGCAGGCTTTCCTTGGTAGCATC

975  GCAATTACTGGTTAGCACACATTGGTGTTGATACAGTAGATTATAATACCTGAACATCATCTCCCCAAATTGGAAGGCACACAAGTGGGCTATTGTTGTC

1075  GTGTCCAGATAAATAATTAGAGAAGATACAGGCTTTCTTTGACATTCTGCTCTGAAATGTGTTCAAGTGATGGAGCCTATGAAACTGACCTTCAGCCTCC

1175  AAAGGCAGGGTGCTCTTAATAGTTAAGGTAGTAGAATTCATCAAACCTCCCGNGTAAGAAAGATGTGAGTAACGTTGGCTTGAATCAGTGAAAATGCCAT

1275  TTGCAGACAACATTAAAAGGTCTTTCAATTAAATGAATCTGTCACACTAAGAGGTGATTCTGAACATGAATACAGCTTAATTCGCCATCATTTATTAATC

1375  CCTCACTTGTATGGATAGGAGTAACTTAGAATGTTTTGTTTGTGGAATTTATCGGCCATGTAGTATGGGATCATCAGGATATATATATGAGGTCTCGTGT

1475  ACAGTTAAAGCTCGCGTCCTGTGATTTCCTGTATGGATACCCTGGGTAATCCTTACTTGGGAACGTTTGATTTAATAAACATGAAATAGTTTC
```

FIG. 13. Nucleotide sequence of the A12 cDNA clone. Numbers on the left refer to the nucleotide sequence (upper) and the deduced amino acid sequence (lower). An in-frame stop codon upstream of the first methionine is underlined. The extra 15 nucleotides present in one clone are boxed. The codon with the best consensus for translation initiation (36) is boxed.

protein of 227 amino acids, and an upstream in-frame stop codon testifies that no protein-coding sequence is missing. However, the first three methionine codons are not within the context of a good match to the consensus sequence for translation initiation (40). The major translation product may result from initiation at nucleotide 339, which is within a good context for translational initiation, giving rise to product of 15.6 kDa MW, with an estimated pI of 9.27.

The A12 primary sequence alone may provide a clue to its structure and possible protein interactions. Chou-Fasman and Robin-Garnier algorithms predict that the protein is almost entirely α-helical, and alignment to NBRF shows similarity to the α-helical portions of the myosin heavy and light chains. However, the most significant NBRF database homology (FIG. 13A) was to an X-linked lymphocyte-regulated (XLR) protein (83), the first identified member of a family of 50 to 75 genomic sequences. The XLR protein is an acidic nuclear protein (32) with a postulated role in the regulation of lymphocyte differentiation (31). Like A12, XLR is extensively α-helical. The presence of A12 mRNA in T cell line EL-4 as well as the opposite charges of the two proteins raises the intriguing possibility that, like the tails of myosin chains, the two proteins form a coiled coil, stabilized by ionic interactions. The subunits of many heterodimeric proteins have this sort of structural homology, which probably reflects divergence from an ancestral sequence after gene duplication events like those which gave rise to the XLR family. As shown above, the A12 transcript is found at high levels in MEL and in a T-cell line, but at low levels in all tissues examined. Its sequence homology to XLR and its limited lineage distribution therefore suggest that the A12 gene product may be involved in

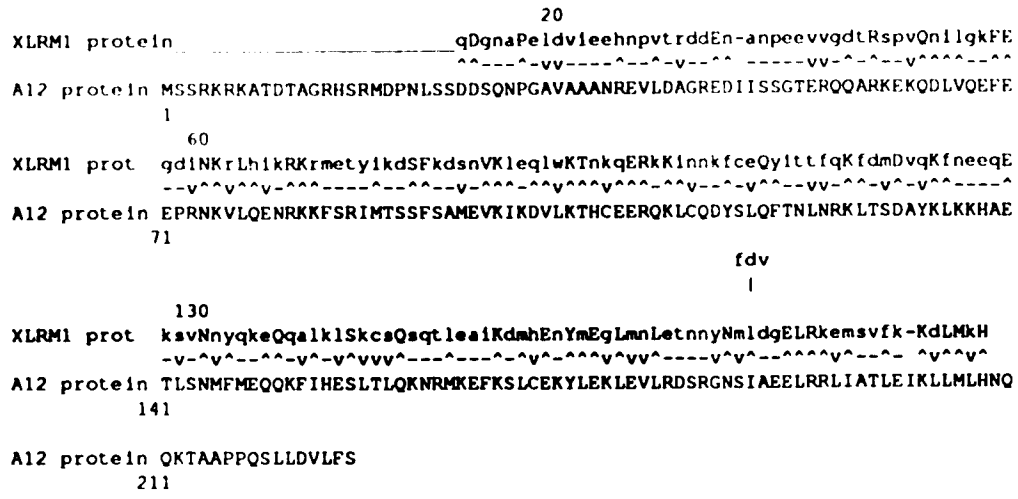

FIG. 13A. Alignement deduced A12 protein sequence with that of the pM1 member of the X-linked lymphocyte regulatory factor (XLRM1). Numbers correspond to amino acids of XLRM1 (upper) and A12 (lower). Changes between amino acids is signified as similar (∧), neutral (-) or dissimilar (v). Amino acid identities are shown capitalized in the XLR sequence.

regulation of a stage of erythroid or T-cell differentiation.

Another cDNA we have analyzed in detail is C10A, whose nucleotide and deduced amino acid sequence are presented in Figure 14. The cDNA fragment we obtained overlaps the carboxy terminal encoding portion of the gene, as evidenced by the stop codon at position 267. However, no poly(A) tail is present, indicating that part of the 3' untranslated stretch is missing.

```
  1 G CTC CCC CTG GAC CGG GAG CAG AAT GGA GAG ACC CAC CAC ACA GGG GAC TGG AGG GGT CCA GGC AGG GAC TCG
  1    L   P   L   D   R   E   Q   N   G   E   T   H   H   T   G   D   W   R   G   P   G   R   D   S

 74 CTT CCC CTC CCC ATG AGG AGC AGG AAG TAC CAG GAA GGT CCA GAT GCC ATT GAG AGG AGA CCC CGG GAG GGT AGT
 25  L   P   L   P   M   R   S   R   K   Y   Q   E   G   P   D   A   I   E   R   R   P   R   E   G   S

149 CAC TCT CCA CTG GAC AGT GCT GAT GTA AGG GTA CAG GTG CCT CGT ACG GGT ATT CCC AGG GCC CCG TCT TCT GAC
 75  H   S   P   L   D   S   A   D   V   R   V   Q   V   P   R   T   G   I   P   R   A   P   S   S   D

224 GAG GAG TGT TTC TTT GAC CTG CTG AGT AAG TTC CAG AGC AGT CGC ATG GAT GAC CAG CGC TGT CCC CTG GAG GAA
100  E   E   C   F   F   D   L   L   S   K   F   Q   S   S   R   M   D   D   Q   R   C   P   L   E   E

299 GGC CAG GCT GGG GCT GCT GAG GCC ACA GCT GCC CCA TCC GTG GAG GAT AGA GCA GCT CAG TCC TCC GTG ACA GCT
125  G   Q   A   G   A   A   E   A   T   A   A   P   S   V   E   D   R   A   A   Q   S   S   V   T   A

374 TCA CCA CAG ACA GAG GAG TTC TTT GAC CTC ATT GCC AGC TCC CAG AGC CGC CGG CTG GAC GAC CAG AGG GCT AGC
150  S   P   Q   T   E   E   F   F   D   L   I   A   S   S   Q   S   R   R   L   D   D   Q   R   A   S

449 GTA GGC AGC CTG CCT GGG CTA CGC ATC ACC CTC AAC AAT GTG GGG CAC CTC CGA GGC GAC GGG GAC GCC CAG GAG
175  V   G   S   L   P   G   L   R   I   T   L   N   N   V   G   H   L   R   G   D   G   D   A   Q   E

524 CCG GGG GAT GAG TTT TTC AAC ATG CTT ATC AAA TAC CAG TCC TCC AGG ATT GAT GAC CAG CGC TGC CCA CCC CCT
200  P   G   D   E   F   F   N   M   L   I   K   Y   Q   S   S   R   I   D   D   Q   R   C   P   P   P

599 GAT GTG CTG CCC CGT GGC CCT ACC ATG CCT GAT GAG GAT TTC TTC AGC CTT ATC CAG AGG GTG CAG GCT AAG CGG
225  D   V   L   P   R   G   P   T   M   P   D   E   D   F   F   S   L   I   Q   R   V   Q   A   K   R

674 ATG GAT GAG CAG CGT GTG GAC CTT GCT GGG AGT CCA GAG CAA GAG GCC AGT GGG CTG CCT GAT CCC CAG CAG CAG
250  M   D   E   Q   R   V   D   L   A   G   S   P   E   Q   E   A   S   G   L   P   D   P   Q   Q   Q

749 TAT CCA CCG GGT GCC AGC TAAGGCCTCGCCCCTACAGCCAGCCATACCCTACTCTGGACTCTGTAGGCTCACGGTTGTCCACAGTGGCCATGAT
275  Y   P   P   G   A   S

824 CCCCC
```

FIG. 14. Nucleotide sequence of the C10A cDNA clone. Numbers on the left refer to the nucleotide sequence (upper) and the deduced amino acid sequence (lower).

The most similar protein to C10A in the NBRF database is the cerebellar Purkinje-cell-specific protein L7, a cytoplasmic protein that is also found in the rod bipolar cells of the retina (7). The L7 transcript is absent in Purkinje cerebellar disease (53). The C10A protein contains a motif which is repeated four times within the cDNA fragment that we have cloned. The L7 protein contains two similar motifs (Fig. 15). When a second search of NBRF was carried out with this motif alone, a different list of protein homologies emerged. After L7, the proteins that contain the most similar motifs include the GTP-binding ral transforming protein (64), the rap1 GTPase activating protein (68), and cyclic GMP phosphodiesterase (63). In those proteins the motif is not repeated.

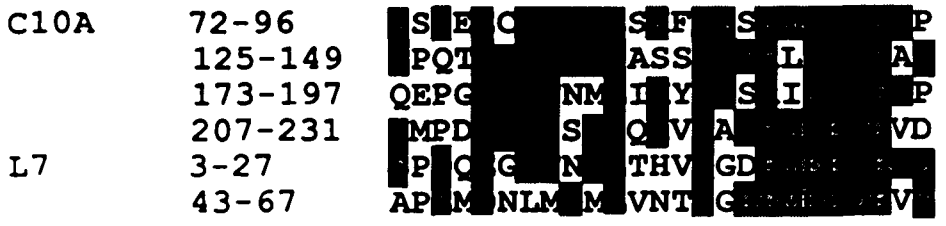

FIG. 15. Amino acid sequence alignment of the repeated motif of C10A and the Purkinje L7 protein (7, 53). Amino acids similar among at least three repeats are shaded. Amino acid numbering of C10A corresponds to Figure 14 and of L7 to the longer published sequence (7).

As mentioned above, transcripts hybridizing to C10A were detected in the brain, ovary and heart. Both the brain and the heart are electrochemically active tissues. Based on the pattern of expression and identities of proteins with this motif, one can imagine that the C10A protein functions in a specialized signaling pathway that may involve a guanine-containing entity. The detection of a message in the ovary that hybridizes to C10A and the presence of the transcript in MEL cells raises the possibility that C10A is involved in a pathway that is common to hematopoietic and germ cells like the c-Kit-mediated response (68). The fact that transcripts of different sizes are found in the heart and in the brain implies that there is either a family of genes with this repeated motif or that alternative splicing may tailor the specific properties of this protein in the heart.

B. EKLF

Notwithstanding the potential importance of the gene products described above, the one with the most exciting prospects is EKLF, based on its sequence and distribution.

Isolation of a putative transcription factor. As described in Materials and Methods, the first identified EKLF sequence was derived from the 3' portion of the cDNA clone provisionally named C10. This original EKLF clone did not contain a poly A sequence, but isolation of overlapping clones enabled us to determine the complete sequence of the EKLF transcript. The sequence of the missing 3' end (400 nucleotides) of EKLF was determined from two corroborating clones independently isolated from a custom made λ-Zap II cDNA library (Stratagene) as well as from an overlapping clone from the subtracted library. Sequence analysis of 16 additional cDNA clones revealed that none of their 5'-ends extended as far upstream as the original C10 clone. The complete reconstructed cDNA sequence based on these data is presented in Figure 16.

Primer extension and S1 nuclease protection analyses were undertaken in order to further define the EKLF transcription start site in MEL and spleen RNA. Both procedures revealed that transcription initiates heterogeneously (Fig. 17), with two major start sites mapping to positions 41 and 55. The transcript that gave rise to the 5'-end of the C10 cDNA clone must have resulted from a minor upstream start site, as no product corresponding to the beginning of this sequence was detected by primer extension or S1 nuclease protection. This sequence is also present in the genomic clone, and thus is not a cloning artifact (data not shown). We conclude from these data that the

FIG. 16. (Next page) Nucleotide sequence of the EKLF cDNA clone. Numbers on the left refer to the nucleotide sequence (upper) and the deduced amino acid sequence (lower). The major transcription start sites are indicated with arrows (nucleotides 41 and 55). An in-frame stop codon upstream of the first methionine is underlined (nucleotides 42-44). The added amino terminal peptide that would result from translation initiation at the first methionine is italicized (amino acids 1-18). Prolines in the amino terminal domain are shown boldface and underlined. The three TFIIIA-like $(Cys)_2$-$(His)_2$ zinc fingers in the carboxy domain are bracketed to include the region of homology to the Krüppel family of transcription factors.

```
1    GTGGGCAGACAGGAGCCCTCCAAGAAACTTTCCTAGCCTCATAGCCC ATG AGG CAG AAG AGA GAG AGG AGG CCT GAG GTC CAG GGT
1                                                    M   R   Q   K   R   E   R   R   P   E   V   Q   G

87   GGA CAC CAG CCA GCC ATG GCC TCA GCT GAG ACT GTC TTA CCC TCC ATC AGT ACA CTC ACC ACC CTG GGA CAG TTT
14   G   H   Q   P   A   M   A   S   A   E   T   V   L   P   S   I   S   T   L   T   T   L   G   Q   F

162  CTG GAC ACC CAG GAG GAC TTC CTC AAG TGG TGG CGG TCT GAG GAG ACG CAG GAT TTG GGG CCG GGG CCC CCG AAT
49   L   D   T   Q   E   D   F   L   K   W   W   R   S   E   E   T   Q   D   L   G   P   G   P   P   N

237  CCT ACG GGG CCG TCC CTT CAC GTG AGT CTG AAA TCG GAG GAC CCT TCC GGA GAG GAC GAT GAG AGG GAC GTG ACC
64   P   T   G   P   S   L   H   V   S   L   K   S   E   D   P   S   G   E   D   D   E   R   D   V   T

312  TGT GCG TGG GAC CCG GAT CTT TTC CTT ACA AAC TTT CCA GGT TCC GAG TCT CCC GGC ACT TCC CGG ACC TGT GCC
89   C   A   W   D   P   D   L   F   L   T   N   F   P   G   S   E   S   P   G   T   S   R   T   C   A

387  CTG GCG CCC AGC GTG GGG CCA GTG GCA CAC TTC GAG CCG CCT GAG TCT CTG GGC GCC TAC GCG GGT GGC CCA GGG
114  L   A   P   S   V   G   P   V   A   Q   F   E   P   P   E   S   L   G   A   Y   A   G   G   P   G

462  TTG GTG ACT GGG CCT TTG GGC TCC GAG GAG CAC ACA AGC TGG GCG CAC CCG ACT CCG AGA CCC CCA GCC CCT GAA
139  L   V   T   G   P   L   G   S   E   E   H   T   S   W   A   H   P   T   P   R   P   P   A   P   E

537  CCC TTC GTG GCC CCT GCC CTG GCC CCG GGA CTC GCT CCC AAG GCT CAG CCC TCG TAC TCC GAC TCG CGA GCG GGC
164  P   F   V   A   P   A   L   A   P   G   L   A   P   K   A   Q   P   S   Y   S   D   S   R   A   G

612  TCC GTA GGG GGC TTC TTC CCG CGG GCG GGG CTT GCG GTG CCC GCA GCT CCA GGC GCC CCC TAT GGG CTG CTG TCG
189  S   V   G   G   F   F   P   R   A   G   L   A   V   P   A   A   P   G   A   P   Y   G   L   L   S

687  GGA TAC CCC GCG CTG TAC CCC GCG CCA CAG TAC CAA GGC CAC TTC CAG CTC TTT CGC GGG CTC GCG GCG CCT TCT
214  G   Y   P   A   L   Y   P   A   P   Q   Y   Q   G   H   F   Q   L   F   R   G   L   A   A   P   S

762  GCT GGT GGG ACG GCG CCC CCT TCC TTC TTG AAT TGT CTG GGA CCT GGG ACT GTG GCC ACA GAA CTC GGG GCC ACT
239  A   G   G   T   A   P   P   S   F   L   N   C   L   G   P   G   T   V   A   T   E   L   G   A   T

837  GCG ATC GCC GGA GAC GCA GGC TTG TCC CCG GGA ACT GCG CCG CCC AAA CGC AGC CGG CGA ACT TTG GCA CCT AAG
264  A   I   A   G   D   A   G   L   S   P   G   T   A   P   P   K   R   S   R   R   T   L   A   P   K

917  AGG CAG GCG GCA CAT ACG TGC GGG CAC GAA GGC TGC GGG AAG AGC TAC TCC AAG AGC TCG CAC CTC AAG GCG CAC
289  R   Q   A   A   H   T   C   G   H   E   G   C   G   K   S   Y   S   K   S   S   H   L   K   A   H

987  CTG CGC ACG CAC ACG GGA GAG AAG CCT TAT GCC TGC TCC TGG GAC GGC TGT GAC TGG AGG TTC GCT CGC TCA GAC
314  L   R   T   H   T   G   E   K   P   Y   A   C   S   W   D   G   C   D   W   R   F   A   R   S   D

1062 GAA CTG ACG CGC CAC TAC CGG AAG CAC ACT GGA CAT CGT CCC TTC TGC TGT GGC CTC TGC CCA CGT GCT TTT TCA
339  E   L   T   R   H   Y   R   K   H   T   G   H   R   P   F   C   C   G   L   C   P   R   A   F   S

1137 CGC TCT GAC CAC TTA GCT CTG CAC ATG AAG CGT CAC CTC TGAGTGATCCTGCACAAGGACTGGGGATGAAATAAGAGTGGATCCAAG
364  R   S   D   H   L   A   L   H   M   K   R   H   L

1224 GACCGTATCCCAAAAGATGAGCCATTATATAGTCCTACCCAGATCAAAAACTGACCAGAAGACCATACAAAGGAGCCTTCAGGACAAACCTCACATGTCC

1324 TCAGGGAGCCCCACACATGGCCCCACAGACCCAGCAATATAGACCACCAGATAAATCAACTCAAATGGACCCCTAGACCAGAGGTGTGACCCTGTGTCCT

1424 GGACGCAGATGGACTGGGGTGAGATTTCCTAAGATCTAGAAGGGAGCTTCACACTGTGCCCATCCGCTAGGATTGTTGTCGTTACTATAAAAATTTCCCA

1524 TATAAAAAAAAAA
```

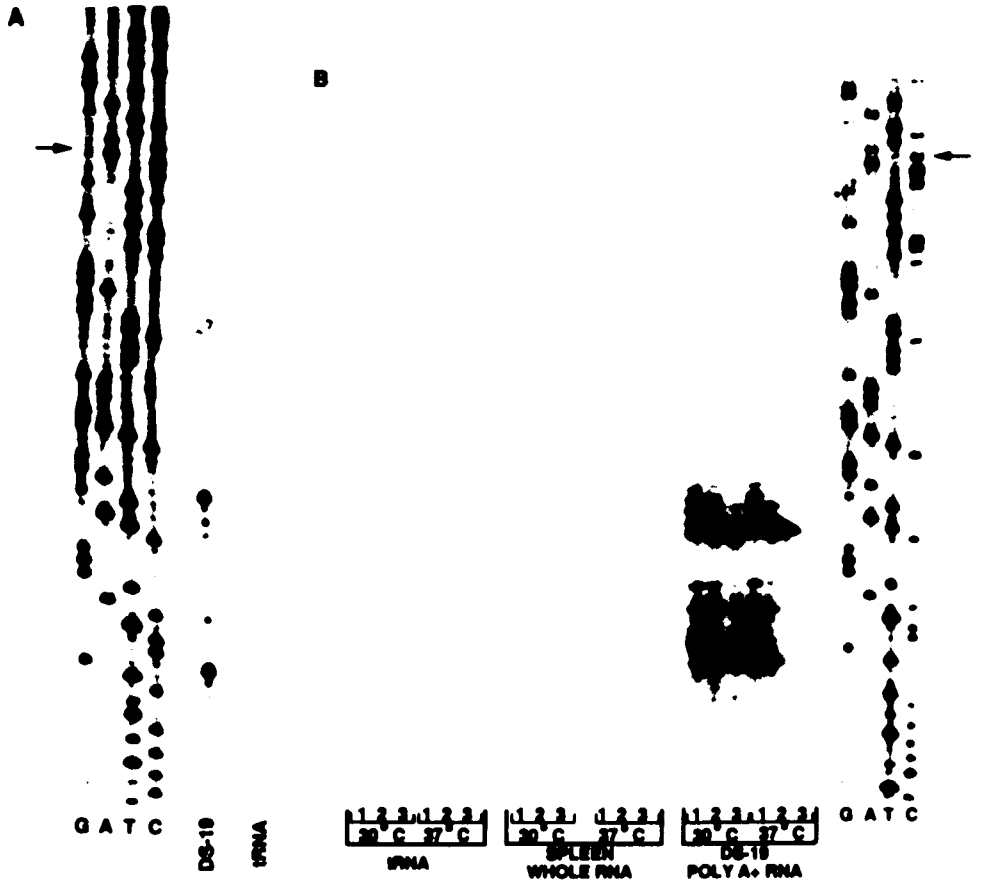

FIG. 17. Determination of the EKLF transcription start site. (A) Primer extension analyses were performed as described in the text with oligonucleotide C10-R3 and 5μg DS-19 poly A^+ RNA or tRNA as indicated. (B) DS-19 poly A^+ RNA (1.7μg/lane), spleen whole RNA (16.6 μg/lane), or tRNA (16.6 μg/lane) were hybridized to the 252-nucleotide probe that was prepared with oligonucleotide C10-R3 as described in the text. Protected products are shown that resulted from S1 nuclease digestion at (lane 1) 100 U/ml, (lane 2) 300 U/ml, or (lane 3) 1000 U/ml. Digestion was performed at either 30°C or 37°C as indicated. The corresponding EKLF (antisense) sequencing ladder using oligonucleotide C10-R3 as primer is shown adjacent to each figure. Arrows indicate the beginning of the sequence shown in Figure 16.

major transcripts are 1,486 and 1,472 nucleotides in length excluding the poly(A) tail.

Inspection of the sequence in Figure 16 reveals potential open reading frames beginning at each of the two predicted in-frame methionines (amino acids 1 and 19). The major transcript that begins at nucleotide 41 could potentially encode a protein of 376 amino acids if translation initiated at its first methionine codon. However, this codon is flanked by a poor match to the Kozak consensus sequence for translation initiation, in that neither a purine three bases before the ATG nor a G immediately following it are present (40). On the other hand, the codon at amino acid 19 is within a good match to the Kozak consensus sequence. Therefore, we believe that the major protein product begins there and is 358 amino acids in length (37,755 Da). The 3'-untranslated region is 350 nucleotides long and contains a polyadenylation signal 17 nucleotides upstream of the poly(A) tail.

The predicted major translation product has an estimated pI of 7.1. Most of the basic residues are in the carboxy terminal 25% of the protein. This includes three TFIIIA-like $(Cys)_2$-$(His)_2$ zinc finger motifs, implicating a region likely to function as a DNA-binding domain. A closer inspection of this sequence reveals that it is most similar to the finger motifs of the Drosophila gap gene Krüppel (80), Sp1 (38), the Wilm's tumor gene product (WT-1) (10), and three early response genes (ERGs) that are almost identical in their DNA-binding domains: NGFI-A (14), Krox-20 (13) and NGFI-C (18) (Fig. 18). Based on the strong homology of the zinc finger region of this new erythroid-cell-derived gene product to this family of DNA binding proteins, it was named EKLF for erythroid Krüppel-like factor.

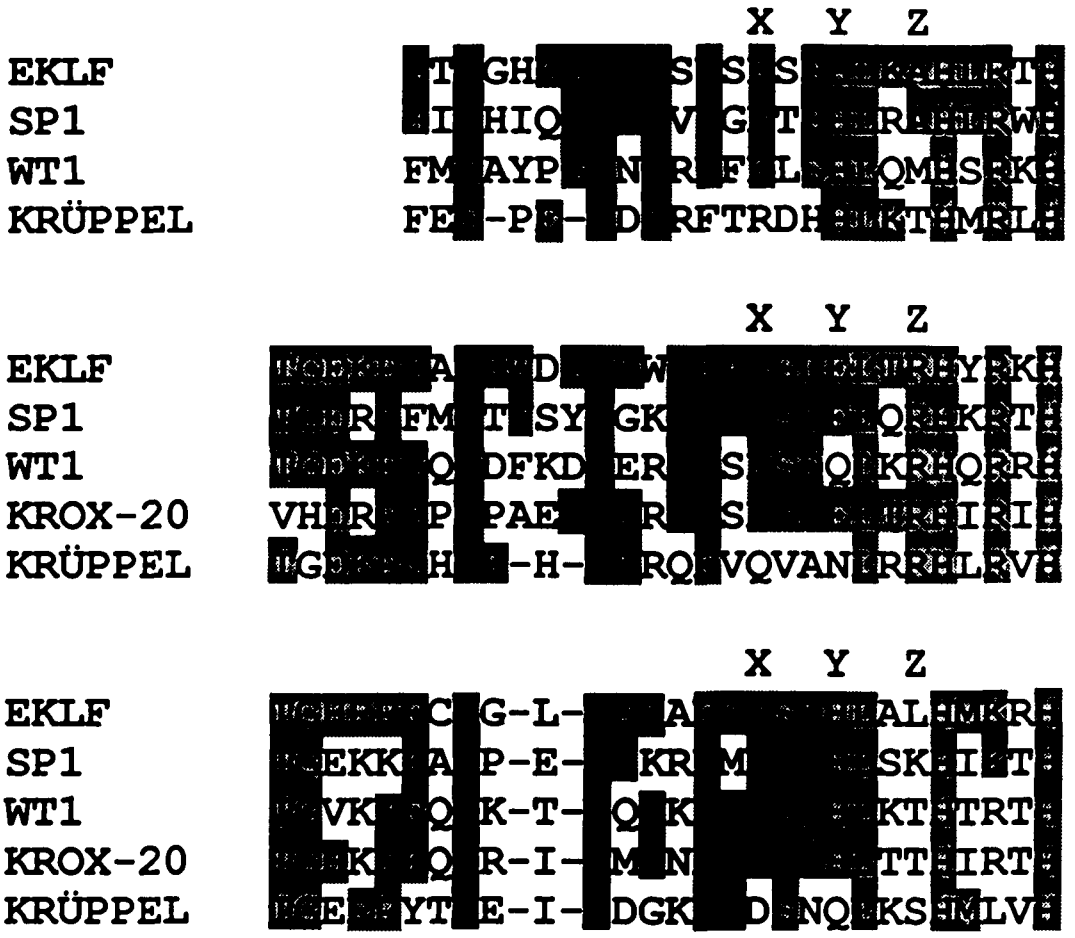

FIG. 18. Amino acid sequence alignment of the three EKLF zinc fingers with those of four closely related proteins. Amino acids identical to those of EKLF are shaded. Human Sp1 fingers 1, 2, and 3 (34), human Wilm's tumor gene product (WT-1) fingers 1, 2, and 3 (9), Drosophila Krüppel fingers 2, 3, and 4 (67), and mouse Krox-20 fingers 1 and 2 (13) are shown. X, Y, and Z denotes positions where basic amino acids have been shown to interact with G residues on the DNA in a variety of zinc finger proteins (35) and is described in the Discussion.

The amino terminal portion of EKLF (amino acids 1-292) bears no significant homology to any other proteins in the NBRF data base. This region, however, has two features characteristic of transcription factors. First, EKLF contains a high percentage (15%) of proline residues. By analogy to several other transcription factors (including WT-1, CTF (57), and NGFI-C), this region may function as a transcriptional activation domain (57). However, unlike the consecutive proline residues within WT-1, the prolines of EKLF seem to be scattered evenly, with never more than two in a row. In this respect, the amino terminal domain of EKLF is more like that of NGFI-C (18). EKLF and NGFI-C also share three similarly spaced clusters of four consecutive amino acids within the stretch of 71 residues adjacent to the DNA binding domain. Thus, it is possible that these two factors have similar or analogous protein interactions. A second feature is that the region bounded by amino acids 76-94 contains a net charge of -8 that may also behave as a transcriptional activation domain (77). These two features of the region adjacent to the zinc fingers are consistent with the notion that EKLF functions as a transcription factor.

Expression of EKLF mRNA is restricted predominantly to the erythroid lineage. In order to examine the tissue-specific pattern of EKLF mRNA expression, Northern blots of RNA from various mouse tissues were probed with the insert from the EKLF cDNA clone (Fig. 19). A transcript migrating slightly faster than the 18S ribosomal band and of the same size as that in the MEL cell was detected only in the samples from bone marrow and spleen (Fig. 19A). As seen in Figure 19C, this expression colocalizes with that of the erythroid-specific carbonic anhydrase type I. We

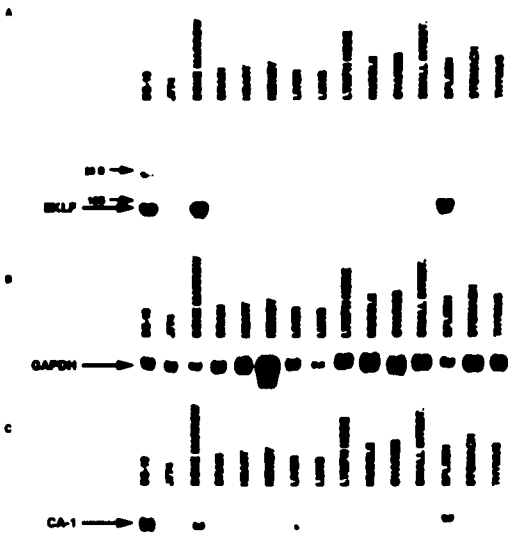

FIG. 19. Northern blot analysis of EKLF message distribution in various mouse tissues. Total RNA (10 ug) from each indicated tissue was probed with random hexamer-labeled cDNA derived from (A) EKLF, (B) rat glyceraldehyde phosphate dehydrogenase (GAPDH; used as a control for RNA loading), or (C) mouse carbonic anhydrase I (CA-1). The same filter was used for all analyses. DS-19 is the MEL cell line from which EKLF was derived and was included as a positive control, and J774 is the non-EKLF-containing monocyte-macrophage cell line used for the original subtractive hybridization procedure and was included as a negative control. Positions of the ribosomal RNA bands are indicated in (A).

conclude that EKLF is expressed within the normal organs of erythropoiesis in the adult mouse.

A more refined analysis of EKLF message distribution was achieved by probing a Northern blot containing RNA from a variety of bone marrow-derived cell lines (Fig. 20). The EKLF message does not appear in macrophage cell line M1, myelomonocytic WEHI 3, T cell line EL-4, or Pre-B cell lines F-4 and 3-1. Overexposure of the x-ray film reveals that the message is present in mast cell lines 10P2, 10P8, 10P12, and 11P62 (Fig. 20A), although we estimate that level of EKLF message is approximately twenty- to thirty-fold higher in the erythroid line than in the mast cell lines (data not shown). In addition, unlike c-myb, GATA-1, and carbonic anhydrase 1, the level of EKLF message in undifferentiated MEL cells is the same as the level in MEL cells that have been induced to differentiate with HMBA for 48 hours (Fig. 20). Approximately 62% of the induced cells were producing hemoglobin at that time as assayed by benzidine staining (data not shown). These results demonstrate the extraordinary specificity of erythroid EKLF expression, with only barely detectable levels present in mast cells. In addition, the levels of EKLF do not appear to be influenced by the onset of terminal differentiation.

The data in Figure 20 also show analyses of the levels of other transcripts in the bone marrow-derived cell lines. The message for transcription factor GATA-1, a factor crucial to erythropoiesis (73), is expressed at similar levels in the mast cell lines and in the erythroid DS-19 cell lines (Fig. 20B). This strongly contrasts with that of the EKLF mRNA, which is much more erythroid-restricted, and indicates that EKLF and GATA-1 are independently regulated. To eliminate the possibility that EKLF expression

FIG. 20. Northern blot analysis of EKLF message distribution in mouse bone marrow-derived cell lines. Total RNA (10 ug) from each of the indicated cell lines was probed with random hexamer-labeled cDNA derived from (A) EKLF, (B) mouse c-myb and mouse GATA-1, (C) rat GAPDH (as a control for RNA loading), or (D) mouse carbonic anhydrase I. The cell lines used were: mouse erythroleukemia cell line DS-19, monocyte-macrophage cell line J774, macrophage cell line M1, myelomonocytic cell line WEHI-3, Pre-B cell lines 3-1 and F4, T cell line EL-4, and mast cell lines 10P2, 10P8, 10P12, and 11P62. RNA samples from DS-19 were prepared both before and after a 48 hour induction with HMBA. The autoradiograph in (A) was overexposed to show the low signal present in the mast cell lines. Positions of the ribosomal RNA bands are indicated in (A). Abbreviations are as in Figure 19.

correlates with the degree of cellular maturation rather than with erythroid identity, the panel of cell lines was probed with c-myb as a marker for immature hematopoietic cells (48). Again in contrast to EKLF, all of the lines except for J774 express c-myb (Fig.20B). Finally, to assess whether EKLF expression correlates strictly with the erythroid phenotype, the cell line panel was probed for carbonic anhydrase I expression (Fig. 20D). Signal was detected only in MEL cells. This shows that although mast cell lines express c-myb and GATA-1 with levels equal to DS-19, the lack of carbonic anhydrase I expression argues against the possibility of a leaky erythroid phenotype that might have accounted for EKLF expression in these cells.

Evolutionary conservation of EKLF. A Southern blot of DNA from a variety of species was probed with EKLF and washed at high stringency to minimize cross-hybridization to other zinc finger proteins. A relatively simple pattern of hybridization was detected in human and mouse DNA samples as well as in those of more distantly related chicken, frog, and fruitfly (Fig. 21). Note, however, that due to the small size of the Drosophila genome, twenty times as many genome-equivalents are present in the Drosophila lane than in the mouse lane since we loaded the same mass of DNA. No signal was detected in the sea urchin sample. The same bands were visible only in the mouse and human DNA samples upon reprobing the blot with an EKLF probe that was missing the zinc finger region (data not shown). These data indicate that there is probably only one or two cross-hybridizing genes in each species examined resulting from conservation of the zinc finger region, rather than the proline-rich region of EKLF.

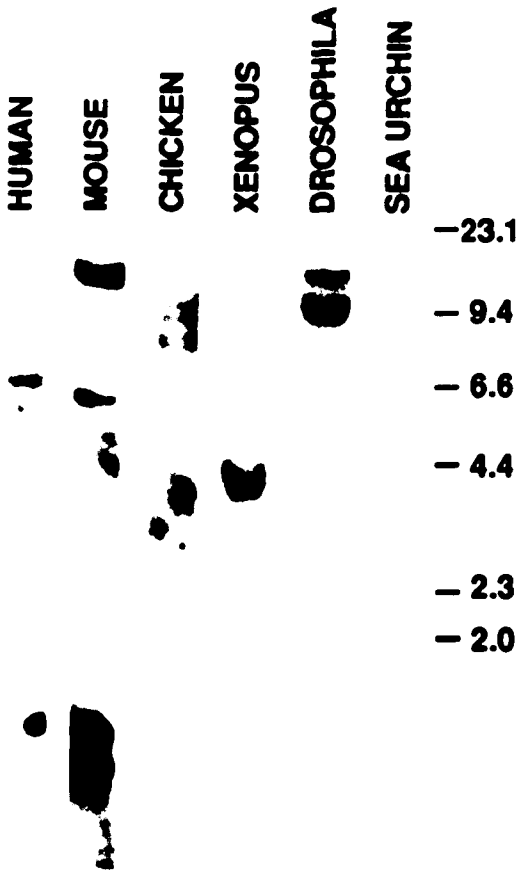

FIG. 21. Southern blot hybridization of EKLF tc DNA frcm various species. Eco R1-digested DNA (5 ug) from the indicated sources was probed with random hexamer-labeled EKLF cDNA and washed at high stringency as described in the text. Positions of molecular weight standards (in kB) are inJicated to the right of the figure.

Chromosomal Mapping. Chromosomal mapping by PCR analysis of mouse/hamster hybrid cell line DNA samples was carried out to determine whether EKLF is a candidate gene for any erythroid phenotypes that have been mapped in the mouse. It was tempting to speculate that EKLF and Fv-2 might be colocalized on chromosome 9 (46), since Fv-2 plays a role in the kinetics of erythroid differentiation (91) and is one of several loci that govern susceptibility to Friend Virus infection. However, results of the analysis excluded EKLF localization on chromosome 9 (data not shown).

IV. DISCUSSION

The molecular events that confer the ability to express lineage-specific genes upon an initially uncommitted, pluripotent stem cell continues to be a major question in cell differentiation. We have approached this problem by utilizing subtractive hybridization to enrich for transcripts specific to the erythroid lineage. The present study describes a gene isolated by this approach that, based upon its deduced protein sequence and tissue distribution, may be directly involved in the process of erythroid cell specialization.

EKLF is a putative erythroid-specific transcription factor. The EKLF cDNA described here encodes a ~1500 nt transcript present in moderate amounts in the MEL cell. The 5'-end of the transcript is heterogeneous, although two start sites account for the large majority of the transcripts detected by either S1 nuclease or primer extension analysis of MEL and spleen RNA. Minor initiation sites might also be used as evidenced by the fact that the original EKLF clone contains 5'-sequences that were not detected by either of the 5'-end analyses above. Multiple transcription start sites are not unusual for tissue-regulated genes, as exemplified by the heterogeneous 5'-ends of both c-myb (6) and GATA-1 (95).

Inspection of the deduced EKLF protein sequence reveals a number of interesting features. First, there are two potential translation initiation sites. The additional 18 amino acid piece resulting from use of the upstream initiation site could have a significant effect on the local charge of the amino terminus, making it more basic. However, we feel it is unlikely that the first methionine is used since the surrounding

sequence is far from optimal for translation initiation. Rather, it is more likely that this potential upstream protein start site may fall into the "encumbered leader sequence" phenomena reviewed recently by Kozak (40).

A second feature of the protein is the proline-rich amino domain, which in addition contains a short acidic stretch of amino acids. Such regions have been noted in a variety of transcription factors. The functional importance of acidic activating regions has been extensively studied (77). In addition, the neutral proline-rich region from CTF can confer transcriptional activation to an unrelated DNA-binding module and may mediate interactions with its own set of transcription adaptors (94).

The EKLF protein is also relatively high in serine/threonine content (15%). Many of these residues are potential sites for phosphorylation. Recent experiments have shown the importance of both kinase and phosphatase activities in regulation of CREB activity (34).

Lastly, the region of the EKLF protein most indicative of its function is the carboxy terminus, which contains three TFIIIA-like zinc fingers. In addition to fulfilling the conserved amino-acid requirements for a true zinc finger motif (26), the inter-finger region homology places the EKLF protein into the Krüppel-like family of transcription factors (82). The presence of the zinc finger domain is the most compelling piece of evidence that EKLF functions as a transcription factor in vivo. Whether EKLF functions as a repressor (as Krüppel (45) and WT-1 (50) do) or as an activator (like Sp1 (38) and NGFI-C (18)) is an important question that will be answered by future functional studies of EKLF.

The EKLF zinc-finger structure predicts potential DNA target sites. The similarity between the DNA binding domains of EKLF and other Krüppel family members, especially Sp1, is compelling evidence for proposing a specific role for EKLF in binding Sp1-like target sites. Both Sp1 and EKLF have three zinc fingers. The greatest homology between the regions is seen when the fingers are aligned in consecutive order (Fig. 18). Between Sp1 and EKLF, the lengths of each homologous finger is preserved as well as the spacing between them. Additionally, the Sp1 and EKLF fingers have the same basic amino acid residues at eight of nine critical X, Y, and Z positions (see Figure 18; (39). X-ray crystallography of a complex of the NGFI-A (Zif268) finger region bound to DNA (71) has shown that when histidine or arginine residues are present in these positions of the finger α-helix, they form hydrogen bonds with guanine nucleotides in the target sequence. Thus, it is highly probable that EKLF and Sp1 bind to an overlapping set of target sites. Based on Klevit's predictions for finger interactions with G residues, EKLF would be predicted to bind a site with the sequence [CCN CNC CCN] on the C-rich strand (39).

The literature is replete with observations of Sp1-like binding sites in erythroid-specific gene promoters and locus control regions. The best studied class of Sp1-like elements with the potential for being important in cell-specific transcription has the core sequence [CACCC]. Some of the erythroid CAC sites in the literature are perfect matches to the predicted EKLF target sequence. One of these, the β-globin CAC site in mice and humans, [CCA CAC CCT], is located at -90 with respect to the transcription start site. In early studies of the mouse ß-globin promoter, an intact CAC site was shown to be necessary for efficient transcription both in HeLa cells and in MEL cells, based on transient transfection assays (12, 17, 59). Expression in HeLa cells, however,

was dependent upon linkage of the promoter to the SV40 enhancer element. When the corresponding human promoter was linked to the locus control region and assayed in stable MEL transfectants, it was found that the proximal CAC box as well as the CCAAT box were required for activity (1). Extracts from erythroid MEL and K562 cells as well as nonerythroid HeLa cells (21) all confer footprint and band shift patterns to this region, but only the extract from induced K562 cells gives rise to a unique, rapidly migrating band that is not seen in the nonerythroid cell extract (52). Furthermore, this complex is not seen when the promoter contains a mutant CAC site with the C to G transversion that gives rise to a rare form of ß-thalassemia (67). The activity in HeLa cells is probably due to the presence of TEF-2 (19), a nonubiquitous factor present in HeLa and CV1 cell extracts that binds selectively to an identical CAC sequence in the GT-1C motif of the SV-40 enhancer as well as to the mouse ß-globin promoter CAC box. (98). Consistent with Mantovani's observations in K562 cells, a CAC site [CCC CAC CCT] at -135 in the chicken ß-globin promoter has been shown to be preferentially protected from DNAse digestion using extracts from transcriptionally active but not quiescent erythrocytes (23, 36). Sp1 also binds to this site, but this CAC binding factor has been shown to be distinct from Sp1 by virtue of the band pattern it forms in a mobility shift assay and by the fact that its binding for CAC is not readily competed by an oligonucleotide containing an Sp1 site (36). All of these results imply the existence of an erythroid-specific CAC binding factor. Another CAC site with a perfect match to the predicted EKLF binding site, [CCC CAC CCT], is found in the erythroid-specific promoter for human and mouse porphobilinogen deaminase. This segment contains adjacent CAC and GATA-1 sites. Mutation of either one reduces promoter activity ten-fold in

transient transfection assays (25).

Defining the consensus sequence for an erythroid-specific CAC binding activity is not as simple as it would seem from these studies. Similar target sites have been found in other erythroid promoters that have the [CACCC] core but do not match as well to our predicted target site. Two of these are found near GATA-1 sites in the promoters of human glycophorin and rat pyruvate kinase genes (60). The ζ-globin promoter also has a divergent CAC element, [CAC CAC CCC T], which is found at -95 with respect to the transcription start site. This segment has been shown to be both necessary and sufficient for activity in MEL but not HeLa cells when combined with the minimal CCAAT and TATA box-containing promoter (97). In contrast, a seemingly more ubiquitous protein other than Sp1 causes footprinting and mobility shifts of the CAC motifs in the promoters of human ε-globin [CTC CAC CCC T] (33) and chicken GATA-1 (35), as well as in the LCR's of human α-globin [CTC CAC CCC C] (see FP3b in (37)) and β-globin [CCC CAC CCC] (74). To confuse matters more, in the case of the α-globin LCR, this ubiquitous factor has been shown not to be TEF-2 (37). Some authors have preferred the view that the erythroid CAC binding protein is the same as TEF-2, yet there is no definitive biochemical data that any of the CAC box binding factors in extracts of HeLa, CV1, and MEL cells are the same proteins. Clearly, further studies are required to determine to which (if any) of the aforementioned motifs EKLF binds in vitro and whether these sites mediate trans-activation (or repression) by EKLF in vivo.

EKLF is primarily expressed in erythroid tissues. The distribution of EKLF message within normal mouse tissues is consistent with what one would expect for an

erythroid-specific product. In particular, limitation of EKLF expression to bone marrow and spleen coincides with the known sites of adult murine hematopoiesis (58). Use of cell lines to further establish the subpopulation of hematopoietic cells that express EKLF reveals that in addition to the erythroid cell, EKLF is present at low levels in mast cell lines. Based on the frequency of EKLF-positive clones in the unsubtracted cDNA library, the EKLF mRNA frequency is estimated to be about 1:20,000 in the MEL cell, which would correlate to an abundance of about 1:500,000 in the mast cell lines. EKLF is not in any of the other myeloid or lymphoid lines tested, even though all of these (except J774) are expressing c-myb. The close relationship among the erythroid, mast cell, and megakaryocyte lineages is underscored by the co-expression in these cells of the erythropoietin receptor (EpoR; cf. (101)) and of GATA-1 (54), a transcription factor known to be important in regulation of erythroid-specific genes. However, our observation that EKLF levels are significantly reduced in mast cells relative to MEL cells implies a role for EKLF in further discriminating among the potential patterns of gene expression in these close lineages. Investigation of EKLF expression in mouse bone marrow cells and spleen tissue sections by in situ methodology will determine whether EKLF is expressed in megakaryocytes similar to EpoR and GATA-1.

The mRNA population particular to any cell type is determined in part by its characteristic array of transcription factors. Commitment of a common precursor cell to the erythroid, mast, and megakaryocyte lineages may thus require a change in expression level of a member of this transcription factor subset. In this context, changes in the level of EKLF, in conjunction with that of GATA-1 and NF-E2, may help define and maintain the erythroid-specific pattern of gene expression. Thus, it will be important to determine when any changes in EKLF expression is first detectable among

members of this lineage and whether such changes correlate with the onset of lineage commitment. In addition, the ability to interfere with EKLF expression by antisense strategies in cultured cells as well as by targeted gene disruption in vivo will illuminate the role EKLF plays in this process.

V. Where To Now?

At this time, the gene products we have identified are of unexamined function. Although we can speculate as to the role of EKLF, several further studies will be required in order to pin down its exact role in erythropoiesis. Most of the other cDNAs have not been sequenced to completion, and the decision as to which of the clones to pursue will be dictated by their degree of erythroid restriction as well as by clues to their function obtained by the homologies detected through further sequencing. For example, further analysis of C10A has been begun by purification of cross-hybridizing clones from the heart and brain.

As for EKLF, several areas of study should be pursued. The first is a set of experiments to define the biochemical role of the EKLF protein. This function can be partly assessed in vitro by evaluating its binding affinity to various specific sites using DNA mobility shift and DNAse I footprinting assays and also through the use of binding site selection protocols. We have begun these experiments with purified GST-EKLF fusion protein expressed in bacteria. In other biochemical studies, antibodies generated to this protein can be used to quantify EKLF levels in erythroid cells and to determine whether EKLF is post-translationally modified, especially by phosphorylation. Another way to examine EKLF function is by altering its

expression in vivo. After determining its binding site, transfection into nonerythroid cells can address its role as a transcription factor. Can EKLF activate transcription from a globin promoter reporter-construct in transient transfection assays or from endogenous erythroid genes in stable transfectants? In a different set of transfection studies, suppressing EKLF function in erythroid cells can possibly be achieved by expression of antisense EKLF RNA or by introduction of antisense oligonucleotides. Alternatively, overexpression of a mutant EKLF with a nonfunctional activation domain might compete with normal EKLF activity. Thus, by down-regulation of EKLF or by interference with its function, maybe we will be in a position to ask whether EKLF is necessary for terminal differentiation of MEL cells.

A different line of experiments may add to our understanding of the molecular events that orchestrate the sequential steps of cellular differentiation. We propose that the common precursor to the erythroid, mast cell and megakaryocyte has the potential to develop various mutually exclusive arrays of lineage-specific transcription factors. The commitment to a specific lineage may be accompanied by a change in the expression of a member of this transcription factor subset. In a continuation of the Northern blot analyses we have started, the steady-state levels of EKLF message in the existing human megakaryocyte cell lines and in multipotential progenitor cells can easily be assessed. Also, in situ hybridization of normal mouse bone marrow and spleen may shed light on the homeostatic distribution of EKLF message. An alternative way to test this hypothesis is to focus on the temporal order of onset of EKLF expression in erythroid progenitors in vivo, especially relative to GATA-1, and c-myb. More elegant studies could be done on ES cell embryoid bodies

to compare onset of expression of EKLF to that of GATA-1 (85). An analogous system that may be useful in determining the temporal order of onset of these transcription factors using in situ hybridization studies is the spleen colony assay. Irradiated mice could be sacrificed at several intervals after bone marrow rescue, possibly with cell sorted populations, and checked for expression of various factors. Eventually, this knowledge may open up the possibility for reprogramming a myeloid cell's identity through an ordered manipulation of its nuclear factors.

Ordering their transcriptional onset may provide indirect evidence for interplay among the erythroid transcription factors, but direct clues to the events leading to EKLF expression may also be found more quickly through biochemical analysis of the EKLF promoter. What are the important protein-binding elements? Is EKLF autoregulated? EKLF may turn out to be induced after GATA-1, in keeping with its more erythroid-limited expression pattern. If so, does GATA-1 promote EKLF expression? It is hoped that eventually the information gained regarding the cell-specific nuclear factors that control transcription will mesh with our growing understanding of extracellular and cytoplasmic signaling events. The result will be a cohesive picture of how progenitor cells become committed to the erythroid lineage and then maintain a specific identity through the ordered action of multiple hormonal stimuli and nuclear factors.

APPENDIX A

From the binomial distribution, the probability of obtaining a sample of X positive results after N trials is dependent upon the probability p of obtaining a positive result each time, and is given as:

$$P(X) = \frac{N!}{X!\,(N-X)!}\, p^{X} (1-p)^{N-X}$$

The odds of getting 0 positive in 6 trials is:

$$P(0) = \frac{N!}{0!\,(N-0)!}\, p^{0} (1-p)^{N-0}$$

$$= \frac{6!}{6!}\, (1-p)^{6} \qquad = (1-p)^{6}$$

For p = 0.40 , P(0) = 0.05. If 40% of the remaining population contains differentially expressed clones, there is a 5% chance of getting 6 nondifferentially expressed clones out of 6 samples.

For p = 0.32, P(0) = 0.10

APPENDIX B

The following sequences were derived using the universal sequencing primers and used to scan the Genbank database for sequence homologies.

Clone A10, using M13 Sequencing primer

TTTACTAGGTATTAGTTCAGGCAGATGCCTGGTTCTGCGATCCCTCTCCCTATAGGTCACAGGATCGGA
TCGGGTTTCTTCAGAAGAGCCTGTGGGGCTGAGAGTCTGAACCCCAGGGTCAGAAGCAGAACTGGAAAG
CCTCCAATACATGTGTTTCCATGTTTCCAGGCTGGGCAGCAGGGAGGACAGGCCCATGACATGCCAAGG
GCTCTCCGTCAGACACCACGGAGGTTTGGTAGGC

Clone A10, using M13 Reverse Sequencing primer

TTGGCCTCGGCAGAATCGGAAGAGAGGTGGCCACCCGAATGCAATCCTTTGGAATGAAGACTGTAGGCT
ATGACCCCATCATCTCTCCTGAAGTCGCCGCCTCCTTTGAGTGTTCAGCAGCTGCACGCTGGAGGAGAT
CTGGGACCTCTCTGTGACTTTATAACTGTCCATGCCCCACTCCTGNNNNNNCTCTACCACAGGCTTCT
GAATGACAGCAGCACCTTTAGACTCAAGTCTGCAAGAAAGGTGTACAGTGGTGAACTGTGGCTCGAGGA
GGCATTGTAGATGAAGGTGCCTGCTG

Clone D5, using M13 Sequencing primer

GGGTTGCTGGAGGGAAGGCTGGGCTGTCAGGAGCAAAGGATGGATGTTCCTCTGGGGGAACCACAGGCC
ACGAGGGGGCCGGAGTAGGACTGGGAGACTGAAAAGGGCAGAAGGGGACAAGGGCTCAGGGAGTGGACA
AGAAGACCT

Clone D10,using M13 Sequencing primer

GGTGAAAAGGGCCTTGACTATGGAGGGGTTGCCGGAGAGTGGTTCTTCCTCATCTGAAGGAAATGTTCA
ACCCTTACTACGGCTGTTTGAATATGCTGCTACGGATAATTACACGCTACAGATAATCCTAACTCGGCT
T

Clone E4, using M13 Sequencing primer

CGGAAGGGAAACACTGTGCCAAGCCTAAGAGGATTACCCTCCAGGGTCTGTAACTTTATTTTCTGTAGC
TGCGGCTTCAGCAGCTGGCATCTTATGGACAAGGCGTTTAGCATGGGAGCCAGTTAGGAGTCAGTGGCA
AAGCGACTTACCAGAACAGACACTACTGAGTAGATCTCGTTGCACTGTCACTCACTAATACGATTTCAC
TGTAGTACTCACGCGCAGTCACAAGCTACGTCTGATCC

Clone G1, using M13 Sequencing primer

CTCAAGTTGGAAATTAATCCCAGCGAGCCACTCCATTCTTGTCGTACNCTCTGAAACCGAAGCCCGCA
GCATTGTTAATGGCGTCAGCTAAGGTCCATGCAAAATAGTACTTAGGTCTGGCAGCCAAGAGAGAGACG
TACAGATAGGTGGCTTTGGTCGGCCACGACGACGCAGTGGCTTGGAAATGCTCATCGATGTTGTACTCC
ACGGGTAGATGTTGGAGATGGTCAGGTGGATAATAAGGAGAGTCCACAGACCAGGAGCTTCTCCGTAAC
TGCTGCATTTGGAGATGGATCTGCTTCCCATGCT

Clone H5, using M13 Sequencing primer

CCTTCACAGACGATAGAAANCNGTTCAAACNGTGCNGGCTACCTGCTGCTACCCAGTCAGTCAGGAGA
TAGGNCCTCTGACGTTAGATGCCGTGTNGCCATCTAACATGTATTGCTAGCCTTAAGGAAGCAGCCAG
AGAAGAGCTGGAGTTTCTTNCTCGTGTGATTTTTAAATGGAGTTCAAAGGTTGNNGTTGNGTNCTGAC
TCCAGGNACGACCAGTGTTGA

Clone H9, using M13 Sequencing primer

GAGTAAAGACATGGATGTTGTGCCATATTTTCTTTTTACAGATAGTATAAGTGGGGAAGACTAGGTAAT
AGCCTCAAAGTAGAGAAGACATCAAGCAAAGGCAACTGAGAAAATGTCTGTGTGGCATTGTGGCTATGG
TTGACCATCTCATTCAGGGTGCTTCTAAAACAGCCCTCNCATCTGTCTACCAGGAGTTAGCCTATTCT
TGTTGATGAACATAGGGTGCTCATCAGTNTAGTAACTAGGAATGGGAACAGTTCTCCAAACAGTNTGT
AANGTGACGTGCTGT

REFERENCES

1. **Antoniou, M. and F. Grosveld.** 1990. β-Globin dominant control region interacts differently with distal and proximal promoter elements. Genes Dev. **4:** 1007-1013.

2. **Auffray, C. and F. Rougeon.** 1980. Purification of mouse immunoglobulin heavy chain RNAs from total myeloma tumor RNA. Eur. J. Biochem. **107:** 303-324.

3. **Bansal, M. P., T. Mukhopadhyay, J. Scott, R. G. Cook, R. Mukhopadhyay and D. Medina.** 1990. DNA sequencing of a mouse liver protein that binds selenium: implications for selenium's mechanism of action in cancer prevention. Carcinog. **11:** 2071-2073.

4. **Ben-David, Y. and A. Bernstein.** 1991. Friend virus-induced erythroleukemia and the multistage nature of cancer. Cell **66:** 831-834.

5. **Ben-David, Y., E. B. Giddens, K. Litwin and A. Bernstein.** 1991. Erythroleukemia induction by Friend Murine lekemia virus: insetional activation of a new member of the ets gene family, FLi-1, closely linked to c-ets-1. Genes Dev. **5:** 908-918.

6. **Bender, T. P. and M. Kuehl.** 1986. Murine myb protooncogene mRNA: cDNA sequence and evidence for 5' heterogeneity. Proc. Natl. Acad. Sci. USA **83:** 3204-3208.

7. **Berger, S. L. and A. R. Kimmel.** 1987. Guide to molecular cloning techniques. Academic Press, San Diego

8. **Berrebi, A. S., J. Oberdick, L. Sangameswaran, S. Christakos, J. I. Morgan and E. Mugnaini.** 1991. Cerebellar Purkinje cell markers are expressed in retinal bipolar neurons. J. Comp. Neurol. **308:** 630-649.

9. **Bilofsky, H. S. and C. Burks.** 1988. The GenBank (R) Genetic Sequence Data Bank. Nucleic Acids Res. **16:** 1861-1864.

10. **Call, K. M., T. Glaser, C. Y. Ito, A. J. Buckler, J. Pelletier, D. A. Haber, E. A. Rose, A. Kral, H. Yeger, W. H. Lewis, C. Jones and D. E. Housman.** 1990. Isolation and characterization of a zinc finger polypeptide gene at the human chromosome 11 Wilms' tumor locus. Cell **60:** 509-520.

11. **Caterina, J., T. Ryan, K. M. Pawlik, R. D. Palmiter, R. L. Brinster, R. R. Behringer and T. M. Townes.** 1991. Human β-globin locus control region: analysis of the 5' DNase I hypersensitivity site HS 2 in transgenic mice. Proc. Natl. Acad. Sci. USA **88**: 1626-1630.

12. **Charnay, P., P. Mellon and T. Maniatis.** 1985. Linker scanning mutagenesis of the 5'-flanking region of the mouse β-major-globin gene: sequence requirements for transcription in erythroid and nonerythroid cells. Mol. Cell. Biol. **5**: 1498-1151.

13. **Chavrier, P., P. Lemaire, O. Revelant, R. Bravo and P. Charnay.** 1990. The segment specific gene krox 20 encodes a transcription factor with binding sites in the promoter region of the Hox 1.4 gene. EMBO J. **9**: 1209-1218.

14. **Chavrier, P., M. Zerial, P. Lemaire, J. Almendral, R. Bravo and P. Charnay.** 1988. A gene encoding a protein with zinc fingers is activated during G0/G1 transition in cultured cells. EMBO J. **7**: 29-35.

15. **Chiba, T., Y. Ikawa and K. Todokoro.** 1991. GATA-1 transactivates erythropoietin receptor gene, and erythropoietin receptor-mediated signals enhance GATA-1 gene expression. Nucleic Acids Res. **19**: 3843-3848.

16. **Clark, S. P. and T. W. Mak.** 1983. Complete nucleotide sequence of an infectious clone of Friend spleen focus-forming provirus: gp55 is an envelope fusion glycoprotein. Proc. Natl. Acad. Sci. USA **80**: 5037-5041.

17. **Cowie, A. and R. M. Meyers.** 1988. DNA sequences involved in transcriptional regulation of the mouse β-globin promoter in murine erythroleukemia cells. Mol. Cell. Biol. **8**: 3122-3128.

18. **Crosby, S. D., J. J. Puetz, K. S. Simburger, T. J. Fahrner and J. Milbrandt.** 1991. The early response gene NGFI-C encodes a zinc finger transcriptional activator and is a member of the GCGGGGGCG (GSG) element-binding protein family. Mol. Cell. Biol. **11**: 3835-3841.

19. **Davidson, I., J. H. Xiao, R. Rosales, A. Staub and P. Chambon.** 1988. The HeLa cell protein TEF-1 binds specifically and cooperatively to two SV40 enhancer motifs of unrelated sequence. Cell **54**: 931-942.

20. **Davis, R. L., H. Weintraub and A. B. Lassar.** 1987. Expression of a single transfected cDNA converts fibroblasts to myoblasts. Cell **51**: 987-1000.

21. **deBoer, E., M. Antoniou, V. Mignotte, L. Wall and F. Grosveld.** 1988. The human β-globin promoter; nuclear protein factors and erythroid specific induction of transcription. EMBO J. **7**: 4203-4212.

22. **Dressler, S., M. Ruta, M. J. Murray and D. Kabat.** 1979. Glycoprotein encoded by the Friend spleen focus forming virus. J. Virol. **30**: 762-764.

23. **Emerson, B. M., J. M. Nickol and T. C. Fong.** 1989. Erythroid-specific activation and derepression of the chick β-globin promoter in vitro. Cell **57**: 1189-1200.

24. **Evans, T. and G. Felsenfeld.** 1989. The erythroid-specific transcription factorEryf1: A new finger protein. Cell **58**: 877-885.

25. **Frampton, J., M. Walker, M. Plumb and P. R. Harrison.** 1990. Synergy between the NF-E1 erythroid-specific transcription factor and the CACCC factor in the erythroid specific promoter of the human porphobilinogen deaminase gene. Mol. Cell. Biol. **10**: 3838-3842.

26. **Frankel, A. D. and C. O. Pabo.** 1988. Fingering too many proteins. Cell **53**: 675.

27. **Fraser, P. J. and P. J. Curtis.** 1986. Molecular evolution of the carbonic anhydrase genes: calculation of divergence time for mouse carbonic anhydrase I and II. J. Mol Evol **23**: 294-299.

28. **Fraser, P. J. and P. J. Curtis.** 1987. Specific pattern of gene expression during induction of mouse erythroleukemia cells. Genes Dev. **1**: 855-861.

29. **Friend, C.** 1957. Cell free transmission in adult Swiss mice of a disease having the character of a leukemia. J. Exp. Med. **105**: 307-318.

30. **Friend, C., W. Scher, J. G. Holland and J. Sato.** 1971. Hemoglobin synthesis in murine virus-induced leukemia cells in vitro: stimulation of erythroid differentiation by dimethyl sulfoxide. Proc. Natl. Acad. Sci. USA **68**: 378-382.

31. **Garchon, H. J.** 1991. The XLR (X-linked lymphocyte regulated) gene family (a candidate locus for an X-linked primary immune deficiency). Immunodefic. Rev. **2**: 283-302.

32. **Garchon, H. J. and M. M. Davis.** 1989. The XLR gene product defines a novel set of proteins stabilized in the nucleus by zinc ions. J. Cell Biol. **108**: 779-787.

33. **Gong, Q.-H., J. Stern and A. Dean.** 1991. Transcriptional role of a conserved GATA-1 site in the human ε-globin gene promoter. Mol. Cell. Biol. **11**: 2558-2566.

34. **Hagiwara, M., A. Alberts, P. Brindle, J. Meinkoth, J. Feramisco, T. Deng, M. Karin, S. Shenolikar and M. Montminy.** 1992. Transcriptional attenuation following cAMP induction requires PP-1-mediated dephosphorylation of CREB. Cell **70**: 105-113.

35. **Hannon, R., T. Evans, G. Felsenfeld and H. Gould.** 1991. Structure and promoter activity of the gene for the erythroid transcription factor GATA-1. Proc. Natl. Acad. Sci. USA **88**: 3004-3008.

36. **Jackson, P. D., T. Evans, J. M. Nickol and G. Felsenfeld.** 1989. Developmental modulation of protein binding to β-globin gene regulatory sites within chicken erythrocyte nuclei. Genes Dev. **3**: 1860-1873.

37. **Jarman, A. P., W. G. Good, J. A. Sharpe, G. Gourdon, H. Ayyub and D. Higgs.** 1991. Characterization of the major regulatory element upstream of the human α-globin gene cluster. Mol. Cell. Biol. **11**: 4679-4689.

38. **Kadonaga, J. T., K. R. Carner, F. Masiarz and R. Tijian.** 1987. Isolation of cDNA encoding transcription factor Sp1 and functional analysis of the DNA binding domain. Cell **51**: 1079-1090.

39. **Klevit, R. E.** 1991. Recognition of DNA by Cys_2, His_2 zinc fingers. Science **253**: 1367, 1395.

40. **Kozak, M.** 1991. An analysis of vertabrate mRNA sequences: intimations of translational control. J. Cell Biol.

41. **Le Cam, A., G. Pages, P. Auberger, G. Le Cam, P. Leopold, R. Benarous and N. Glaichenhaus.** 1987. Study of a growth hormone-regulated protein secreted by rat hepatocytes: cDNA cloning, anti-protease activity and regulation of its synthesis by various hormones. EMBO J. **6**: 1225-1237.

42. **Levenson, R. and D. Housman.** 1979. Developmental program of murine eerythroleukemia cells. J. Cell Biol. **82**: 715-725.

43. **Levenson, R., J. Kernen and D. Housman.** 1979. Synchronization of MEL cell commitment with cordycepin. Cell **18**: 1073-1078.

44. **Li, J.-P., A. D. D'Andrea, H. F. Lodish and D. Baltimore.** 1990. Activation of cell growth by binding of Friend spleen focus corming virus gp55 glycoprotein to the erythropoietin receptor. Nature (London) **343**: 762-764.

45. **Licht, J. D., M. J. Grossel, J. Figge and U. M. Hansen.** 1990. Drosophila Krüppel protein is a transcription factor. Nature (London) **346**: 76-79.

46. **Lilly, F.** 1970. Fv-2: Identification and location of a second gene governing the spleen focus response to Friend leukemia virus in mice. J. Natl. Cancer Inst. **45**: 163-169.

47. **Lowrey, C. H., D. M. Bodine and A. W. Nienhuis.** 1992. Mechanism of DNase I hypersensitivity site formation within the human locus control region. Proc. Natl. Acad. Sci. USA **89**: 1143-1147.

48. **Luscher, B. and R. Eisenman.** 1990. New light on myc and myb. Part II. myb. Genes Dev. **4**: 2235-2241.

49. **MacDonald, R. J., G. H. Swift, A. E. Przybyla and J. M. Chirgwin.** 1987. Isolation of RNA using guanidium salts. Methods Enzynol. **152**: 219-227.

50. **Madden, S. L., D. M. Cook, J. F. Morris, A. Gashler, V. P. Sukhatme and F. J. Rauscher.** 1991. Transcriptional repression mediated by the Wilm's tumor gene product. Science **253**: 1550-1553.

51. **Maniatis, T., E. F. Fritsch and J. Sambrook.** 1982. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

52. **Mantovani, R., N. Malgaretti, S. Nicolis, B. Gigloni, P. Comi, N. Cappellini, M. T. Bertero, F. Calgaris-Cappio and S. Ottolenghi.** 1988. An erythroid specific nuclear factor binding to the proximal CACCC box of the β-globin gene promoter. Nucleic Acids Res. **16**: 4299-4313.

53. **Marks, P. A. and R. A. Rifkind.** 1978. Erythroleukemic differentiation. Ann. Rev. Biochem. **47**: 419-448.

54. **Martin, D. I. K., L. I. Zon, G. Mutter and S. Orkin.** 1990. Expression of an erythroid transcription factor in megakaryocytic and mast cell lineages. Nature (London) **344**: 444-447.

55. **Melloni, E., S. Pontremoli, G. Damiani, P. Viotti, M. Patrone, R. A. Rifkin and P. A. Marks.** 1989. Differential expression of protein kinase C isozymes and erythroleukemia cell differentiation. J. Biol Chem. **264**: 18414-18418.

56. **Melloni, E., S. Pontremoli, R. Michetti, O. Sallo, A. G. Cakiroglu, J. F. Jackson, R. A. Rifkin and P. A. Marks.** 1987. Protein kinase C. activity and hexamethylenebissacetamide-induced erythroleukemia cell differentiation. Proc. Natl. Acad. Sci. USA **84**: 5282-5286.

57. **Mermod, N., E. A. O'Neill, T. J. Kelly, R. Kelly and R. Tijian.** 1989. The proline-rich transcriptional activator of CTF/NF-1 is distinct from the replication and DNA binding domain. Cell **58**: 741-753.

58. **Metcalf, D.** 1988. The molecular control of blood cells. Harvard University Press, Cambridge.

59. **Meyers, R. M., K. Tilly and T. Maniatis.** 1986. Fine structure genetic analysis of a β-globin promoter. Science **232**: 613-618.

60. **Mignotte, V., J. F. Eleouet, N. Raich and P.-H. Romeo.** 1989. Cis- and trans-acting elements involved in the regulation of the erythroid promoter of the human porphobilinogen deaminase gene. Proc. Natl. Acad. Sci. USA **86**: 6548-6552.

61. **Mirand, E. A., R. A. Steeves, L. Avila and J. T. Grace.** 1968. Spleen focus formation by polycythemic strains of Friend leukemia virus. Proc. Soc. Exp. Biol. Med. **127**: 900-904.

62. **Mitchell, P. J. and R. Tjian.** 1989. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. Science **245**: 371-378.

63. **Moreau-Gachelin, F., A. Tavitian and P. Tamborin.** 1988. Spi-1 is a putative oncogene is virally induced murine erythroleukemias. Nature (London) **1988**: 277-283.

64. **Nordquist, D. T., C. A. Kozak and H. T. Orr.** 1988. cDNA cloning and characterization of three genes uniquely expressed in cerebellum by purkinje neurons. J. Neurosci. **8**: 4780-4789.

65. **Olofsson, T. B.** 1991. Growth regulation of hematopoietic cells. Acta Oncol. **20**: 889-902.

66. **Orkin, S., F. I. Harosi and P. Leder.** 1975. Differentiation in erythroleukemic cells and their somatic hybrids. Proc. Natl. Acad. Sci. USA **72**: 98-102.

67. **Orkin, S. H., H. H. Kazazian Jr., S. E. Antonarakis, S. C. Goff, C. D. Boehm, J. P. Sexton, P. G. Waber and P. J. V. Giardina.** 1982. Linkage of β-thalassaemia mutations and β-globin gene polymorphisms with DNA polymorphisms in the human β-globin gene cluster. Nature (London) **296:** 627-631.

68. **Orr-Urtreger, A., A. Avivi, Y. Zimmer, D. Givol, Y. Yarden and P. Lonai.** 1990. Developmental expression of c-kit, a proto-oncogene encoded by the W locus. Devel. **109:** 911-923.

69. **Palazzolo, M. J., D. R. Hyde, K. VijayRaghavan, K. Meklenburg, S. Benzer and E. Meyerowitz.** 1989. Use of a new strategy to isolate and characterize 436 drosophila cDNA clones corresponding to RNAs detected in adult heads but not embryos. Neuron **3:** 527-539.

70. **Patuleia, M. C. and C. Friend.** 1967. Tissue culture studies on murine virus-induced leukemia cells: isolation of single cells in agar liquid medium. Cancer Res. **27:** 726-730.

71. **Pavletich, N. P. and C. O. Pabo.** 1991. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. Science **252:** 809-816.

72. **Perkins, N. D., R. H. Nicolas, M. A. Plumb and G. H. Goodwin.** 1989. The purification of an erythroid protein which binds to enhancer and promoter elements of haemoglobin genes. Nucleic Acids Res. **17:** 1299-1314.

73. **Pevny, L., M. C. Simon, E. Robertson, W. H. Klein, S.-F. Tsai, V. D'Agati, S. Orkin and F. Costantini.** 1991. Erythroid differentiation in chimeric mice blocked by a targeted mutation in the gene for transcription factor GATA-1. Nature (London) **349:** 257-260.

74. **Philipsen, S., D. Talbot, P. Fraser and F. Grosveld.** 1990. The β-globin dominant control region: hypersensitive site 2. EMBO J. **9:** 2159-2167.

75. **Pittler, S. J., W. Baehr, J. J. Wasmuth, D. G. McConnell, M. S. Champagne, P. VanTuinen, D. Ledbetter and R. Davis.** 1990. Molecular Charqacterization of Human and Bovine Rod Photoreceptor cGMP phosphodiesterase α-subunit and chromosomal localization of the human gene. Genomics **6:** 260-267.

76. **Polakas, P. G., R. F. Weber, B. Nevins, J. R. Didsbury, T. Evans and R. Snyderman.** 1989. Identification of the ral and rac1 gene products, low molecular mass GTP-binding proteins from human platelets. J. Biol Chem. **264:** 16383-16389.

77. **Ptashne, M.** 1988. How eukariotic transcriptional activators work. Nature (London) **335**: 683-689.

78. **Reitman, M. and G. Felsenfeld.** 1988. Mutational analysis of the chicken b-globin enhancer reveals two positive-acting domains. Proc. Natl. Acad. Sci. USA **84**: 6267-6271.

79. **Romeo, P.-H., M.-H. Prandini, V. Joulin, V. Mignotte, M. Prenant, W. Vainchenker, G. Marguerie and G. Uzan.** 1990. Megakaryocytic and erythrocytic lineages share specific transcription factors. Nature (London) **344**: 447-449.

80. **Rosenberg, U. B., C. Schröder, A. Preiss, A. Kienlin, S. Côté, I. Riede and H. Jäckle.** 1986. Structural homolgy of the product of the drosophila Kruppel gene with Xenopus transcription factor IIIA. Nature (London) **319**: 336-339.

81. **Rubinfeld, B., S. Munemitsu, R. Clark, L. Conroy, K. Watt, W. J. Crosier, F. McCormic and P. Polakis.** 1991. Molecular Cloning of a GTPase Activating Protein Specific for the Krev-1 protein p21rap1. Cell **65**: 1033-1042.

82. **Schuh, R., W. Aicher, U. Gaul, S. Côté, A. Preiss, D. Maier, E. Seifert, U. Nauber, C. Schröder, R. Kemler and H. Jäckle.** 1986. A conserved family of nuclear proteins containing structural elements of the finger protein encoded by Krüppel, a Drosophila segmentation gene. Cell **47**: 1025-1032.

83. **Siegel, J. N., C. A. Turner, D. M. Klinman, M. Wilkinson, A. D. Steinberg, C. L. MacLeod, W. E. Paul, M. M. Davis and D. I. Cohen.** 1987. Sequence analysis and expression of an X-linked, lymphocyte regulated gene family (XLR). J. Exp. Med. **166**: 1702-1715.

84. **Siegel, M. L., R. Horowitz, T. D. Morris, R. M. D. Venere and E. J. Siden.** 1985. Isolation and characterization of Abelson murine leukemia virus-transformed mast cell lines from migestation embryonic placenta. Eur. J. Immunol. **15**: 1136-1141.

85. **Simon, M. C., L. Pevny, M. V. Wiles, G. Keller, F. Costantini and S. Orkin.** 1992. Rescue of erythroid development in gene targeted GATA-1- mouse embryonic stem cells. Nature Genetics **1**: 92-98.

86. **Sive, H. L. and T. S. John.** 1988. A simple subtractive technique employing photoactivatable biotin and phenol extraction. Nucleic Acids Res. **22**: 10937.

87. **Sorge, J. A. and L. A. Blinderman.** 1989. ExoMeth sequencing of DNA: Eliminating the need for subcloning and oligonucleotide primers. Proc. Natl. Acad. Sci. USA **86**: 9208-9212.

88. **Steeves, R. A. and E. A. Mirand.** 1969. Separation of members of the Friend virus complex by sucrose gradient centrifugation. Proc. Amer. Assoc. Cancer Res. **10**: 86.

89. **Struhl, K.** 1991. Mechanisms for diversity in gene expression patterns. Neuron **7**: 177-181.

90. **Suguoka, Y., T. Kano, A. Okuda, M. Sakai, T. Kitagawa and M. Muramatsu.** 1985. Cloning and the nucleotide sequence of rat glutathione S-transferase P cDNA. Nucleic Acids Res. **13**: 6049-6057.

91. **Suzuki, S. and A. A. Axelrad.** 1980. Fv-2 locus controls the proportion of erythropoietic progenitor cells (BFU-E) sythesizing DNA in normal mice. Cell **19**: 225-236.

92. **Talbot, D. and F. Grosveld.** 1991. The 5' HS2 of the globin locus control region enhances transcription through the interaction of a multimeric complex binding at two functionally distinct NF-E2 bindig sites. EMBO J. **10**: 1391-1398.

93. **Talbot, D., S. Philipsen, P. Fraser and F. Grosveld.** 1990. Detailed analysis of the site 3 region of the human β-globin dominant control region. EMBO J. **9**: 2169-2178.

94. **Tanese, N., B. F. Pugh and R. Tijan.** 1991. Coactivators for a proline-rich activator purified from the multisubnit human TFIID complex. Genes Dev. **5**: 2212-2224.

95. **Tsai, S.-F., D. I. K. Martin, L. I. Zon, A. D. D'Andrea, G. G. Wong and S. H. Orkin.** 1989. Cloning of cDNA for the major DNA-binding protein of the erythroid lineage through expression in mammalian cells. Nature (London) **339**: 446-451.

96. **Tsai, S.-F., E. Strauss and S. H. Orkin.** 1991. Functional analysis and *in vivo* footprinting implicate the erythroid transcription factor GATA-1 as a positive regulator of its own promoter. Genes Dev. **5**: 919-931.

97. **Watt, P., P. Lamb, L. Squire and N. Proudfoot.** 1990. A factor binding GATAAG confers tissue specificity on the promoter of the human ζ-globin gene. Nucleic Acids Res. **18**: 1339-1350.

98. **Xiao, J.-H., I. Davidson, M. Macchi, R. Rosales, M. Vigneron, A. Staub and P. Chambon.** 1987. In vitro binding of several cell-specific and ubiquitous nuclear proteins to the GT-I motif of the SV40 enhancer. Genes Dev. **1**: 794-807.

99. **Yoon, J.-B., H. C. Towle and S. Seelig.** 1987. Growth hormone induces two mRNA species of the serine protease inhibitor gene family in rat liver. J. Biol Chem. **262**: 4284-4289.

100. **Zon, L., M. Gurish, R. Stevens, C. Mather, D. Reynolds, K. F. Austen and S. Orkin.** 1991. GATA-binding transcription factors in mast cells regulate the promoter of the mast cell carboxypeptidase A gene. J. Biol Chem. **266**: 22948-22953.

101. **Zon, L. I., H. Youssoufian, C. Mather, H. F. Lodish and S. H. Orkin.** 1991. Activation of the erythropoietin receptor promoter by transcription factor GATA-1. Proc. Natl. Acad. Sci. **88**: 10638-10641.