MOLECULAR AND CELLULAR BIOLOGY, July 1993, p. 4011–4022
0270-7306/93/074011-12$02.00/0

Vol. 13, No. 7

DNA-Binding Specificities of the GATA Transcription Factor Family

LINDA J. KO AND JAMES DOUGLAS ENGEL*

Department of Biochemistry, Molecular Biology and Cell Biology, Northwestern University, Evanston, Illinois 60208-3500

Received 9 March 1993/Returned for modification 8 April 1993/Accepted 21 April 1993

Members of the GATA family of transcription factors, which are related by a high degree of amino acid sequence identity within their zinc finger DNA-binding domains, each show distinct but overlapping patterns of tissue-restricted expression. Although GATA-1, -2, and -3 have been shown to recognize a consensus sequence derived from regulatory elements in erythroid cell-specific genes, WGATAR (in which W indicates A/T and R indicates A/G), the potential for more subtle differences in the binding preferences of each factor has not been previously addressed. By employing a binding selection and polymerase chain reaction amplification scheme with randomized oligonucleotides, we have determined the binding-site specificities of bacterially expressed chicken GATA-1, -2, and -3 transcription factors. Whereas all three GATA factors bind an AGATAA erythroid consensus motif with high affinity, a second, alternative consensus DNA sequence, AGATCTTA, is also recognized well by GATA-2 and GATA-3 but only poorly by GATA-1. These studies suggest that all three GATA factors are capable of mediating transcriptional effects via a common erythroid consensus DNA-binding motif. Furthermore, GATA-2 and GATA-3, because of their distinct expression patterns and broader DNA recognition properties, may be involved in additional regulatory processes beyond those of GATA-1. The definition of an alternative GATA-2–GATA-3 consensus sequence may facilitate the identification of new target genes in the further elucidation of the roles that these transcription factors play during development.

Many transcription factors have been found to be members of highly related multifactor families, and thus their specificity of action must be addressed in order to ascertain their respective functions. Consequently, it is critical to determine which factor, from an array of factors with closely related DNA-binding motifs, acts upon a particular recognition site from a variety of sites with similar sequences. Equally important is the elucidation of the means by which this discrimination is achieved; a variety of mechanisms by which related transcription factors are targeted to distinct regulatory elements have been discovered. Differences in DNA-binding properties direct the zinc finger estrogen and progesterone receptors to their appropriate targets (7, 44), as is the case with the *Antennapedia*-related homeodomain proteins *Ultrabithorax* and *Deformed* (8, 9). The related retinoic acid, thyroid hormone, and vitamin D receptors exemplify a unique solution to the problem of differential target recognition by discriminating between the spacing and orientation of closely related binding sites of these obligate dimers (27, 45). For the dimeric basic helix-loop-helix proteins MyoD and E2A, the choice of dimerization partner has been shown to dictate binding-site preference (2). Furthermore, some factor families have indistinguishable binding specificities, and specific protein-protein interactions mediated through regions outside of the DNA-binding domain are proposed to result in promoter targeting, as exemplified by the POU proteins Oct-1 and Oct-2 (40) and the basic helix-loop-helix proteins myogenin and MRF4 (5).

GATA-1 was originally identified as an erythroid cell-specific DNA-binding protein that bound to a WGATAR consensus sequence (in which W indicates A/T and R indicates A/G) found in the regulatory regions of many globin and nonglobin erythroid-specific genes (4, 13, 20, 24, 41, 46). After GATA-1 was cloned (11, 42), it was later found to be but one member of a transcription factor family related by their high degree of amino acid identity throughout the two-zinc-finger DNA-binding domain (48). The expression pattern of each of the GATA family members appears to be highly evolutionarily conserved among vertebrates: GATA-1 is expressed only in cells of the myeloid lineage (erythroid cells, mast cells, and megakaryocytes [25, 34, 42, 48]), with the notable exception of an abundant testis-specific form transcribed from an alternate promoter (18). GATA-2 is expressed in a wide variety of tissues, and GATA-3 is most abundantly expressed in T lymphocytes and in the developing central nervous system (21a, 48). Each of the chicken GATA factors, cGATA-1, -2, and -3, has been shown to bind to the GATA site found in the mouse α-globin promoter (TGATAA), and all three factors are expressed during avian, murine, and human erythropoiesis (22, 48). Therefore, each factor has the potential to bind to the same sites within downstream target gene regulatory elements. Determination of which specific GATA factor is bound to any given promoter or enhancer region is thus critical to understanding the role that each of these factors plays during development and differentiation. We therefore postulated that each factor might have a different, distinguishable binding specificity still encompassed within the WGATAR consensus that could allow each member of the GATA family to fulfill distinct functions.

When considering the further observation that the expression patterns of the various GATA family members are sometimes overlapping but decidedly distinct, the issue of binding specificity takes on further significance, since the cell types of most abundant expression differ for the various GATA proteins. Although all three GATA factors have been

* Corresponding author.

shown to bind a consensus sequence based on regulatory elements from within erythroid cell-specific genes (the WGATAR erythroid consensus), GATA-3, for example, may recognize a different site in T lymphocytes, the cell type of its highest expression (17, 19, 21, 30, 31). For this reason, we felt that the recognition element(s) of the GATA factors mediating transcriptional regulation in nonerythroid cells should be examined in detail.

Utilizing a binding-site selection procedure, we show here that the bacterially expressed chicken GATA proteins have virtually identical binding specificity for the originally defined erythroid GATA consensus site. However, cGATA-2 and -3 (but not cGATA-1) can also recognize a new alternative binding site with high affinity, consistent with the possibility that cGATA-2 and cGATA-3 regulate a broader constellation of target genes than cGATA-1.

MATERIALS AND METHODS

Bacterial expression of the cGATA-1, -2, and -3 proteins. cGATA-1, -2, and -3 proteins were prepared by introducing the relevant portions of the cGATA cDNAs (48) into the bacterial expression vector pGEX-2T (Pharmacia). The cGATA-1 cDNA clone was digested to completion with *Eco*RI (which cleaves at the 3′ end of the cDNA) and then partially digested with *Nco*I (which cleaves in two positions, 25 and 562, in the cGATA-1 cDNA sequence [48]). The CCATGG *Nco*I recognition sequence at the 5′ end of this fragment contains the initiation methionine for the cGATA-1 protein. The resulting 1.0-kb fragment corresponding to the entire coding region plus the 3′ untranslated region was isolated, and the protruding ends were filled with the Klenow fragment of DNA polymerase I (Promega), then ligated to the pGEX-2T vector digested with *Sma*I, and treated with calf intestinal alkaline phosphatase (Boehringer Mannheim). The resulting construct was confirmed by DNA sequencing. The cGATA-2 cDNA was digested with *Nco*I at positions 408 to 1800, thereby excising almost the complete coding region (410 to 1810). The CCATGG at the 5′ end of the fragment corresponds to the initiation methionine of the cGATA-2 protein. The ends were filled with the Klenow fragment of DNA polymerase I and ligated into *Sma*I-digested pGEX-2T vector (described above). The cGATA-3 bacterial expression construct was made in two steps: first, the 3′ end of the cGATA-3 cDNA from the *Bam*HI site at nucleotide 675 to the *Sma*I site in the downstream polylinker of the original vector (pGEM 7; Promega) was cloned into the corresponding sites of pGEX-2T. Second, a 522-bp region at the 5′ end of the cDNA was amplified by the polymerase chain reaction (PCR) using a downstream primer corresponding to nucleotides 692 to 672 (48) and an upstream primer corresponding to nucleotides 171 to 123 that introduced a *Bam*HI site (underlined) and an *Nco*I site (boldface) not present in the original cDNA:

GCAGCGAGCGGGATCCATGGAG primer
GCAGCGAGCGCGGAAGATGGAG cDNA

This PCR product was digested with *Bam*HI and inserted into the pGEX vector containing the 3′ coding sequences of cGATA-3. The resulting clone was confirmed by DNA sequencing.

The cGATA-1, -2, and -3 expression constructs were each transformed by electroporation into BL21 cells (39). Expression of each of the fusion proteins was carried out according to published procedures (38). Partially purified cGATA-1 was prepared from mature adult chicken erythrocytes (10, 15) by DNA cellulose chromatography followed by wheat germ agglutinin-agarose affinity chromatography (50).

Binding site selection procedure. (i) Generation of probes for gel mobility shift assay. The following synthetic oligonucleotides were used for the binding site selections:

```
A     TCCGAATTCCTACAG
NGAT  TCCGAATTCCTACAGGACNNNGATNNNNACTTGTTCACATGTAGACTGCAATGGTACCGTCT
B                                                     ACGTTACCATGGCAGA
```

Direct sequencing of the NGAT oligonucleotide with primers A and B confirmed that the starting material contained an equal representation of all four bases at each of the randomized positions; equal-intensity bands were detected in sequencing lanes for each of the four bases (data not shown). Results from the first binding-site selection experiment indicated that the original pool of NGAT was a mixture of 63-mers due to a deletion in the N region and full-length 64-mers. The latter were gel isolated and used for the second experiment.

(ii) Individual selection reactions. For each round of selection, the starting oligonucleotides were amplified for 30 cycles and then radiolabeled in a subsequent step. Because of the sensitivity of PCR when many cycles are performed, special attention was given to ensure the absence of contamination in these reactions. For each set of amplifications, a parallel control with no added template was included and duplicate reactions were performed. To one reaction 10 μCi of ^{32}P-labeled dCTP was added, and 10 μl of labeled product was electrophoresed on a 10% denaturing gel to verify that the major species recovered was the predicted 64-bp oligonucleotide and that it was detected only in a template-dependent fashion. After confirmation of a lack of contamination, the unlabeled amplification products were fractionated on a Sephadex G-25 column. Of 100 μl of selection reaction mixture, 10 μl was then used as a template for extending 500 ng (each) of primers A and B in the presence of 10 μCi of α-^{32}P-labeled dCTP. The products of these reactions were again passed over Sephadex G-25 columns and then used directly as probes in gel mobility shift assays.

As a mobility standard for the first round of selection, the 30-bp double-stranded MαP GATA site from the mouse α-globin promoter MαP$_{30}$ (29) was cloned into the *Bam*HI site of pGEM4 and a 63-bp *Eco*RI-*Sal*I fragment was excised, end labeled with [γ-^{32}P]ATP by using T4 polynucleotide kinase, and used as a probe, MαP$_{63}$, to determine the mobility of the 64-bp NGAT complexed with each of the GATA factors.

(iii) Gel mobility shift assays. Probes prepared as described above were used in 20-μl binding reaction mixtures contain-

ing 25 mM HEPES (*N*-2-hydroxyethylpiperazine-*N'*-2-ethanesulfonic acid) (pH 7.9), 80 mM NaCl, 5 mM $MgCl_2$, 0.1 mg of bovine serum albumin per ml 10 mM dithiothreitol, and 2.5% Ficoll. Each reaction mixture was incubated for 30 min at 0°C before electrophoresis as described previously (48). Bands corresponding to the GATA protein-DNA complex (the mobility of which was determined by comparison with the binding reactions with the $M\alpha P_{63}$ probe; see above) were excised from the dried gel and placed in 200 μl of 10 mM Tris, pH 7.5. The DNA was allowed to elute from the gel slice during a 3-h incubation at 37°C. Ten microliters of the eluted DNA was then used directly in a PCR to amplify the selected sites.

The same process was repeated for a total of four rounds of selection. While the initial rounds of selection were performed at lower salt concentrations, in the subsequent rounds salt concentrations were raised. The first selection experiment with cGATA-1 was performed by one round of selection at 80 mM NaCl followed by three rounds at 200 mM NaCl. All other experiments were performed using two rounds of selection at 80 mM NaCl followed by two rounds at 200 mM NaCl.

(iv) Cloning and sequencing of selected sites. After four rounds, the pools of oligonucleotides selected by each GATA factor were PCR amplified a final time before being cleaved by *Kpn*I and *Eco*RI (recognition sites in oligonucleotides A and B, above) and cloned into a *Kpn*I-*Eco*RI-cleaved, phosphatase-treated pBSIIKS vector (Stratagene). Denatured double-stranded templates containing individual sites were sequenced from the T7 promoter by dideoxy chain termination (6, 35).

Gel mobility shift assays with individual binding sites. Oligonucleotides generated from the cloned, selected sites were PCR amplified for 12 cycles starting with 200 ng of plasmid containing the single sites to generate probes for gel mobility shift assays. The amplification products were separated from the larger plasmid templates by electrophoresis of the samples on a 4% low-melting-point agarose gel (NuSieve GTG agarose; FMC) and isolation of the 64-bp PCR products. The fragments were resuspended in 10 mM Tris (pH 7.5)–1 mM EDTA, and 4% of the reaction mixture was labeled with Klenow fragment plus [α-^{32}P]dCTP as described above.

As a standard for gel shifts with selected cloned sites, the 30-bp MαP site (42) was kinase labeled and used to determine the amount of each protein required to give a shifted band of comparable intensity. This same quantity of protein was then used in binding reactions with individual sites. Gel mobility shifts used to determine equilibrium binding characteristics were performed in buffer containing 200 mM NaCl.

Dissociation rate analysis. Thirty-nucleotide double-stranded DNAs corresponding to the selected sequences B20 and CC21 (see Fig. 2A and 3B) were synthesized, and the dissociation rates for the three GATA factors from these two representative binding sites were determined as follows. For each off rate, a master binding reaction equivalent to four reactions as described above was used, with the following exceptions: the NaCl concentration was lowered to 40 mM in order to retard the off-rate kinetics, and the probe used was 1 ng of kinase-labeled, double-stranded B20 or CC21. Reactions were allowed to come to equilibrium for 30 min at 0°C before the addition of a 100-fold excess of synthetic, unlabeled double-stranded MαP competitor oligonucleotide at time zero. Aliquots (10 μl) were removed at the times indicated in Fig. 6 and loaded onto a gel running at 50 V. A separate, single reaction including a 100-fold excess of competitor was allowed to come to equilibrium to determine the end point of dissociation. After loading of all of the time points, the gel was run at 175 V for the remainder of the run. Complex formation at each time point was determined by densitometric scanning on a Molecular Dynamics PhosphorImager. For quantitative evaluation of the kinetic and equilibrium binding experiments, the amount of complexed DNA and protein from the coelectrophoresed final equilibrium value sample was subtracted from intermediate time point values. The K_d was determined by plotting the equation ln(fraction bound) = $-K_d t$ by utilizing the Cricket Graph software program to determine the best fit.

Relative equilibrium affinities. Synthetic oligonucleotides corresponding to the GATA sites within selected sites B20 and CC21 (see Results) were quantitated with a spectrophotometer, diluted to appropriate concentrations, and used for titration as competitors in binding reactions with the $M\alpha P_{30}$ site probe. Quantitation of relative affinities was determined by densitometry using a Molecular Dynamics PhosphorImager.

RESULTS

Bacterial expression of the cGATA factors. In order to obtain pure and abundant quantities of the GATA proteins to study their relative binding characteristics, each of the chicken GATA factors was expressed in bacteria as a glutathione *S*-transferase (GST) fusion protein. Purified GST-cGATA fusion proteins were shown to specifically bind to a high-affinity site from the mouse α-globin promoter ($M\alpha P_{30}$; see Materials and Methods and Fig. 3), while the full-length GST protein alone did not (data not shown). The fusion proteins were used in all of the following studies unless noted otherwise.

Binding-site selection. We initially postulated that the GATA factors might exert differential effects dictated by similar yet unique binding specificities. A number of GATA factor-regulated genes have been identified by the presence of a consensus recognition sequence (4, 13, 20, 24, 41, 46), but which GATA factor regulates any particular sequence in vivo is not yet resolved. Therefore, we initiated these experiments by fixing the most critical sequence determinant for GATA factor recognition and then determined the binding specificities of the three factors at adjacent nucleotide residues in an attempt to correlate the site recognition properties of each of the factors with the known *cis* regulatory elements.

The initial erythroid consensus binding site contained a GATA core by one analysis (13) and a GAT by another (46), each flanked by other nucleotides with some degree of degeneracy. Confirmation of the necessity of the first three nucleotides was demonstrated by methylation interference experiments with human GATA-1, suggesting that the most critical contact point identified in the recognition sequence was the G (46); subsequent studies with cGATA-1 confirmed this conclusion (36). Other experiments showed that mutation of the first A or T resulted in an 80- or 60-fold-lower affinity of factor binding, respectively, while mutation of the second A led to only a 20-fold decrease in affinity (29), consistent with the GAT consensus derivation. On the basis of these observations, we chose to undertake the binding-site selection experiment utilizing an oligonucleotide with a central GAT core DNA sequence bordered by random nucleotides (NGAT; see Materials and Methods). Randomized positions are designated −3 to −1 (5′ of the GAT core),

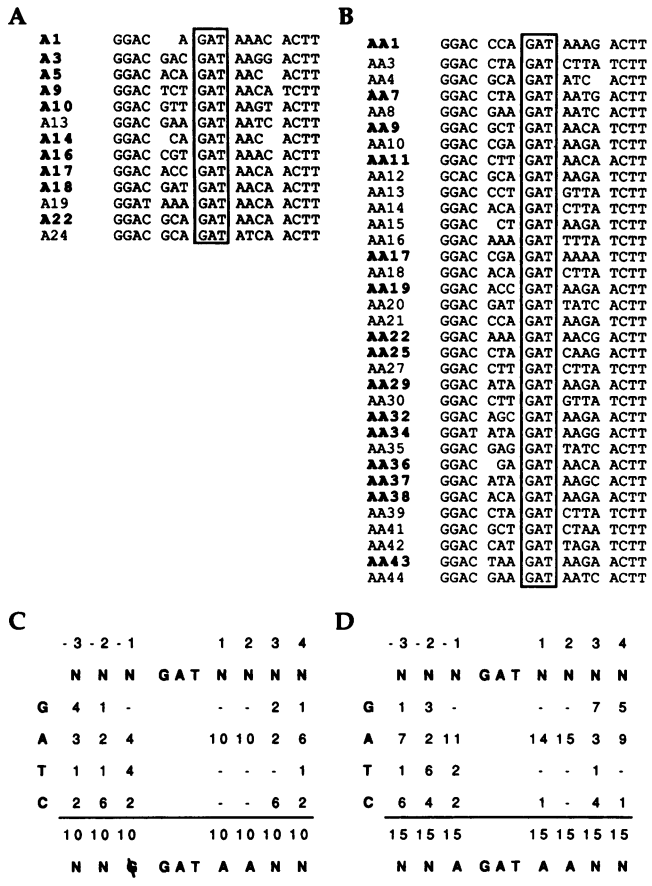

FIG. 1. Binding-site selection of cGATA-1. (A and B) Sequences of cloned sites after four rounds of selection for cGATA-1 binding in the first (A) and second (B) experiments. The sites identified in boldface are the single sites used for the analysis shown in panels C and D. (See Results for a description of the basis for exclusion from single-site analysis). The central GAT core encoded in NGAT (see Materials and Methods) is indicated by the boxes. Deletions are represented by gaps in the sequence. (C and D) Frequency of recovery of each nucleotide at each selected position and the derived consensus recognition sequence in the first (C) and second (D) experiments. To be considered part of the consensus, the threshold level of representation of a nucleotide at a given position was arbitrarily set at >60%. In the case of deletions, the adjacent fixed nucleotides next in sequence were considered the selected nucleotide. −, no single sites were recovered with that nucleotide at that position.

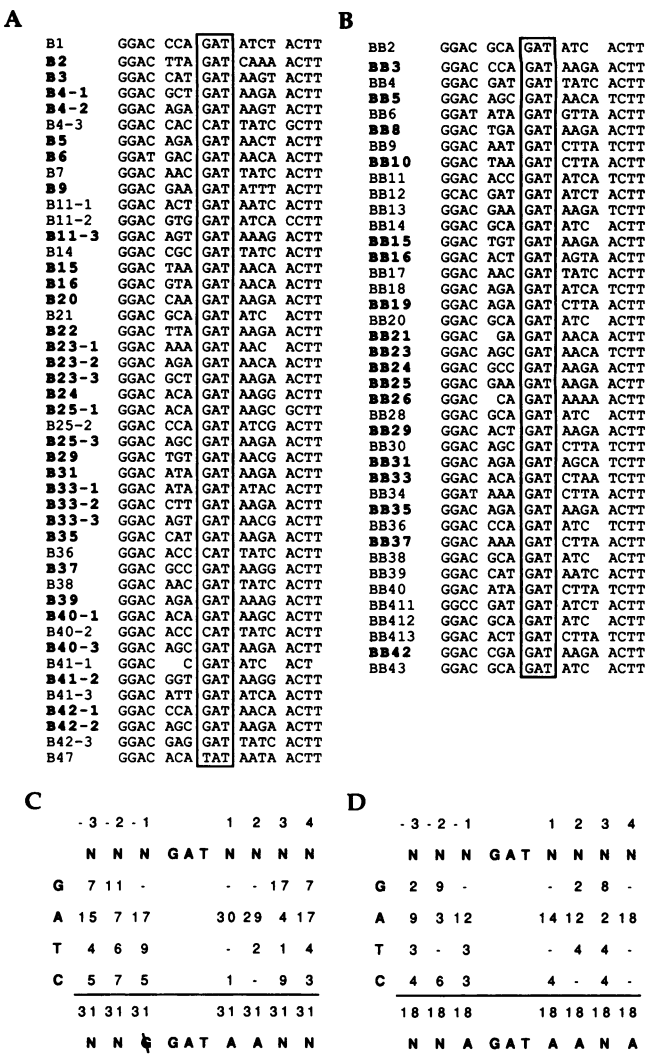

FIG. 2. Binding-site selection of cGATA-2. (A and B) Sequences of cloned sites after selection for cGATA-2 binding in the first (A) and second (B) experiments. The sites identified in boldface are the single sites used for the analysis shown in panels C and D. Sites identified as, e.g., B4-3 indicate that in clone B4, there was an insertion of three independent 64-bp oligonucleotides as a result of the cloning procedure (see Materials and Methods). (C and D) Frequency of recovery of each nucleotide at each selected position and the derived consensus recognition sequence in the first (C) and second (D) experiments. See the legend to Fig. 1 for details of the analysis.

and +1 to +4 (3′ of the GAT core). This initial assumption for the necessity of the GAT motif for the binding of the GATA factors was further borne out by the results (detailed below) in which selected oligonucleotides displaying mutations in the GAT core were found to contain a GAT motif elsewhere in the randomized portions of the oligonucleotide.

To determine the binding-site specificity of the various GATA factors, a selection and amplification procedure was employed (1, 2). Radiolabeled, double-stranded NGAT was synthesized and used as a probe in gel mobility shift assays with GST fusion proteins of each cGATA-1, -2, and -3. GATA factor-binding sites were selected by isolating the lower-mobility band of the protein-DNA complex separated by gel electrophoresis, followed by amplification by PCR using primers A and B corresponding to the ends of NGAT (Materials and Methods). The process was repeated for four consecutive rounds of selection, the pools of selected oligonucleotides were then cloned, and individual sites were sequenced.

The sites selected by GST–cGATA-1, -2, and -3, in two independent experiments, are shown in Fig. 1A and B, 2A and B, and 3A and B, respectively. Because of the cloning strategy employed (incorporating different restriction sites at either end of the selected sites in the A and B primers; see Materials and Methods), trimer or pentamer ligations in the cloning step were recovered in some cases; each sequence thus identified was analyzed as an independent site (Fig. 2B and 3B).

Several of the sites selected by either of the GATA proteins contained the sequence AGATCTTATCTT (e.g.,

A

C3-1	GGAC	CGC	GAT	AAGA	ACTT
C3-2	GGAC	AGC	GAT	TATC	ACT
C3-3	GGAC	CGT	GAT	ATCA	ACTT
C5-1	GGAC	AAA	GAT	AATA	ACTT
C5-2	GGAC	ATG	GAT	AATC	ACTT
C5-3	GGAC	AAA	GAT	AAGA	ACTT
C7-1	GGAC	CGA	GAT	AATA	ACTT
C7-2	GGAC	AAC	GAT	AAAA	ACTT
C7-3	GGAC	CGG	GAT	AACT	ACTT
C8-1	GGAC	AAT	GAT	TATC	ACTT
C8-2	GGAC	GTG	GAT	AAT	ACTT
C8-3	GGAC	ACA	GAT	AACA	ACTT
C9-1	GGAC	CGC	GAT	TATC	ACTT
C9-2	GGAC	AGA	GAT	AAGG	ACTT
C9-3	GGAC	GTC	AAT	AATC	ACTT
C10-1	GGAC	GGC	GAT	TAAT	ACTT
C10-3	GGAC	ACT	GAT	AAGT	ACTT
C11-1	GGAC	AGC	GAT	AATC	ACTT
C11-2	GGAC	AGA	GAT	ATCC	ACTT
C11-3	GGAC	GCT	GAT	AAGA	ACTT
C11-4	GGAC	ATA	AAT	ATC	ACT
C11-5	GGAC	TGA	GAT	AAGA	ACTT
C12-1	GGAC	AAC	GAT	TCT	ACTT
C12-2	GGAC	GT	GAT	AAGA	ACTT
C12-3	GGAC	AGT	GAT	GATC	ACTT
C12-4	GGAC	AGA	GAT	CAAG	ACTT
C14-1	GGAC	GAA	GAT	AACA	ACTT
C14-2	GGAC	CCC	CAT	AATC	ACTT
C14-3	GGAC	ACG	GAT	ATCA	ACTT
C15-1	GGAC	GCA	GAT	ATCG	ACTT
C15-2	GGAC	TTA	GAT	CTTA	ACTT
C15-3	GGAC	ATG	GAT	ATC	ACTT
C17	GATC	GC	GAT	ATC	ACTT
C18-1	GGAC	CCT	TAT	TATC	ACTT
C18-2	GGAC	TAA	GAT	ACCA	ACTT
C18-3	GGAC	GCC	GAT	AATC	ACTT
C19-1	GGAC	ACA	GAT	TATC	ACTT
C19-2	GGAC	AGC	GAT	AAGA	ACTT
C19-3	GGAC	ATT	GGT	GGTC	ACTT
C21	GCAC	AGA	GAT	AAC	ACTT
C22	GGAC	GTA	GAT	AAGA	ACTT
C23-1	GGAC	CAA	GAT	TAGG	ACTT
C23-2	GGAC	TGA	GAT	AATC	ACTT
C23-3	GGAC	ACA	GAT	AAAT	ACTT
C24	GGAT	GA	GAT	TAA	AGTT
C26-1	GGAT	AC	GAT	AAGA	ACTT
C26-2	GGAC	ACA	CAT	ATC	ACT
C26-3	GGAC	GAT	GAT	AAGG	ACTT
C27-1	GGAT	AGA	GAT	ATT	ACTT
C27-2	GGAC	GGG	GAT	ATCA	ACTT
C27-3	GGAC	AGA	GAT	AAGA	ACTT
C28-1	GGAC	AAA	GAT	ATCC	ACTT
C28-2	GGAC	AAC	CAT	ATCC	ACTT
C28-3	GGAC	AGA	GAT	AACA	ACTT
C29-1	GGAG	ATA	GAT	ATCT	ACTT
C29-2	GGAC	TGA	GAT	CTTA	ACTT
C29-3	GGAC	CAG	GAT	TATC	ATCC
C30	GGAC	GAA	GAT	CTTT	ACTT
C31-1	GGAC	AGC	AAT	TATC	ACTT
C31-2	GGAC	AGA	GAT	AACG	ACTT
C32	GGAC	TGT	GAT	AACC	ACTT
C34-1	GGAC	AGG	GAT	AATC	ACTT
C34-2	GGAC	CCC	GAT	ATCA	TCTT
C34-3	GGAC	TAA	GAT	AATC	ACTT
C35-1	GGAC	AGA	GAT	AAT	ACTT
C35-2	GGAC	CGA	GAT	AAAG	ACTT
C35-3	GGAC	AGA	GAT	AAGA	ACTT
C37-1	GGAT	ACT	GAT	ATCG	ACTT
C37-2	GGAC	AGA	GAT	AAGA	ACTT
C37-3	GGAC	GAT	GAT	AATC	ACTT
C39-1	GGAC	GCA	AAT	AAC	ACTT
C39-2	GGAC	CGA	GAT	AAG	ACTT
C39-3	GGAC	GCA	GAT	AATA	ACTT
C41	GGAG	ATA	GAT	TATC	ACTT
C43	GGAC	GTT	GTT	CATC	ACTT
C44	GGAC	CAT	GAT	AATC	ACTT
C45	GGAC	CGA	GAT	ATC	ACTT
C46	GGAC	AGT	GAT	ATTT	ACTT
C47-1	GGAC	ACT	AAT	ATC	ACTT
C47-2	GGAC	CGC	GAT	ATC	ACT
C47-3	GGAC	AAC	GAT	CTTA	ACTT
C48-1	GGAT	AAA	GAT	CTCT	ACTT
C48-2	GGAC	GGG	GAT	TATC	ACT
C48-3	GGAC	AGC	GAT	AATC	ACTT

B

CC1	GGAC	ACT	GAT	AAAC	ACCT
CC2	GGAC	ACA	GAT	TAAA	ACTT
CC3-1	GGAC	CGA	GAT	CTTA	TCTT
CC3-2	GGAC	GTT	GAT	AAGA	TCTT
CC3-3	GGAC	ATA	GAT	CTTA	TCTT
CC4	GGAC	ACA	GAT	CTAA	ACTT
CC6	GGAC	TAA	GAT	AAAA	TCTT
CC7	GGAC	ATA	GAT	AATA	ACTT
CC13	GGAG	ATA	GAT	CTTA	ACTT
CC14	GGAC	GAA	GAT	GTTA	TCTT
CC15	GGAC	GAA	GAT	AAGA	TCTT
CC16	GGAC	CAT	GAT	CGAT	ACTT
CC18	GGAC	AAC	GAT	CTTA	ACTT
CC20	GGAC	ATT	GAT	CTTA	ACTT
CC21	GCAC	CAA	GAT	CTTA	ACTT
CC23	GGAC	TAA	GAT	CTTA	ACTT
CC24	GGAC	GCA	GAT	CTTA	ACTT
CC26	GGAC	AGA	GAT	TATA	ACTT
CC27	GGAC	AGC	GAT	CTTA	ACTT
CC28	GGAC	AAA	GAT	CTAG	ACTT
CC30	GGAC	CT	GAT	AAGA	ACTT
CC31	GGAC	AAA	GAT	TAAA	ACTT
CC33	GGAC	TGA	GAT	CTTA	ACTT
CC35	GGAC	AAA	GAT	CTTT	ACTT
CC36	GGAC	AGA	GAT	CTTA	ACTT
CC40	GGAC	AAA	GAT	CTAC	ACTT
CC41	GGAG	ATT	GAT	CTGT	ACTT
CC42	GGAT	GAT	GAT	AATC	ACTT
CC44	GGAC	AGA	GAT	TTTA	ACTT
CC45	GGAC	GAT	GAT	ATTA	TCTT
CC46	GGAC	AAA	GAT	CAA	ACTT
CC47	GGAC	ATA	GAT	AAGA	ACTT
CC48	GGAC	CAG	GAT	TATC	ACTT

C

	-3	-2	-1		1	2	3	4
	N	N	N	GAT	N	N	N	N
G	8	23	2		1	-	15	6
A	20	10	26		31	34	6	27
T	6	3	7		4	5	12	6
C	7	5	6		5	2	8	2
	41	41	41		41	41	41	41
	N	N	A	GAT	A	A	N	A

D

	-3	-2	-1		1	2	3	4
	N	N	N	GAT	N	N	N	N
G	4	5	-		-	-	5	1
A	17	11	18		8	11	8	21
T	3	5	6		4	15	13	2
C	2	5	2		14	-	-	2
	26	26	26		26	26	26	26
	A	N	A	GAT	C/A	T/A	N	A

FIG. 3. Binding-site selection of cGATA-3. (A and B) Sequences of cloned sites after selection for cGATA-3 binding in the first (A) and second (B) experiments. The sites identified in boldface are those utilized in panels C and D. Sites identified as, e.g., C3-3 indicate that in clone C3, there was an insertion of three independent oligomers as a result of the cloning procedure. (C and D) Frequency of recovery of each nucleotide at each selected position and the derived consensus recognition sequence in the first (C) and second (D) experiments. See the legend to Fig. 1 for details of the analysis.

AA14, BB33, and CC3-3), in which a second GATA site was present on the bottom strand (TATC). Several other selected sites were found to have similar patterns of a GATA sequence in the inverse orientation on the bottom strand:

TABLE 1. Recovery frequency of erythroid consensus and other binding sites from GATA factor-selected oligonucleotides

Factor	Sequence	% of sites with sequence (no. with sequence/no. of sites) in: Expt 1	Expt 2
cGATA-1	**GAT AA**	100 (10/10)	93 (14/15)
cGATA-2	**GAT AA**	94 (29/31)	67 (12/18)
	GAT CT	0	22 (4/18)
cGATA-3	**GAT AA**	71 (29/41)	27 (7/26)
	GAT CT	10 (4/41)	50 (13/26)
	GAT TA	7 (2/41)	12 (3/26)

GATATC (e.g., AA4, BB2, and C26-2) or **GATTATC** (e.g., AA20, B14, and C8-1). Because of the possibility that each GAT sequence (one on the top and one on the bottom strand) serves as a separate recognition site, these particular products might be considered double sites with inverted orientation and are considered separately below. Other oligonucleotides (recovered less frequently) contained two GATA sequences in a direct repeat orientation (e.g., AA34). The recovery rate of double sites was very frequent and is not surprising given that double sites would be theoretically predicted to be of twofold-higher affinity than single sites. However, to address the specificity of each factor for a single binding site, data from recovered (potential) multiple sites were excluded from the analyses described immediately below.

Sites with mutations in the GAT core were also excluded from the analysis. During several rounds of amplification, the intrinsic incorporation error frequency of *Taq* polymerase occasionally mutated the GAT core, yet these oligonucleotides were still selected by GATA factor binding (e.g., B4-3 and C9-3). Of a total of 15 oligonucleotides in this category, 11 were found to contain a GAT sequence elsewhere in the oligonucleotide (generally on the bottom strand, within the +1 to +4 randomized positions). This observation further supports the initial assumption that the GAT core sequence would be critical for GATA factor binding.

To identify the relative contributions of specific nucleotides at defined positions in determining the ability of a site to be recognized, the single binding-site sequences recovered were compiled and analyzed for the frequency of encountering any nucleotide at each randomized nucleotide position, and a consensus was derived for each of the factors based on the most highly favored nucleotide at each position and the frequency at which it was recovered (Fig. 1C and D, 2C and D, and 3C and D). The nucleotides immediately 3′ to the GAT core were the most highly selected, with each of the factors selecting **GATAA** at a very high frequency (Fig. 1C and D, 2C and D, and 3C and D), consistent with the canonical WGATAR erythroid consensus. In two independent experiments, GATA-1 very specifically selected **GATAA**, while GATA-2, in comparison, had a lower specificity for those same nucleotide positions, and GATA-3 selected this same site with the lowest frequency. In the first experiment, 100, 94, and 71% of the single sites selected by GATA-1, -2, and -3, respectively, contained a **GATAA** motif. A similar trend was apparent with the single sites in the second experiment, in which 93, 67, and 27% of the sites selected by GATA-1, -2, and -3, respectively, displayed the identical sequence (Table 1). Thus, all three cGATA factors selected an erythroid consensus **GATAA** site at a relatively high frequency.

Inspection of the recovered sites also revealed that other,

non-WGATAR consensus sites were also selected at a significant frequency by the GATA-2 and -3 factors. Sites containing a **GATCT** motif accounted for 22% of the single sites selected by GATA-2 in the second experiment and 10 and 50% of GATA-3-selected sites in the first and second experiments, respectively (Table 1). Another site, **GATTA**, was recovered at a much lower but reproducible frequency by GATA-3. These three types of sites, **GATAA**, **GATCT**, and **GATTA**, accounted for the majority of single sites recovered in the selection assay (Table 1) and further show that the nucleotide identity at a given position is not independent of one selected at a neighboring position. By chi-square analysis, the dependence of the identity of the +2 position upon the +1 identity exceeds the 0.005 level of significance for the **GATAA** sites selected by all three factors in the first experiment and for the **GATCT** sites selected by GATA-3 in the second experiment. Although the total number of sites selected with the **GATTA** motif is too low to effectly employ statistical analysis, inspection of the sequences recovered suggests a dependence between the +1 and +2 identities for this motif as well.

The fact that there exist numerical differences in the representation of **GATAA** and **GATCT** sites in the first and second experiments may indicate that early selection events limited the population of oligonucleotides carried through subsequent rounds of amplification and selection. However, the general conclusions derived from the two independent experiments are internally consistent: all three GATA factors select canonical **GATAA** sites, but in addition, GATA-2 and GATA-3 have a broader specificity, also selecting novel, alternative consensus sites.

The previously derived GATA consensus, WGATAR (13, 46), defined the −1 position as A or T, and indeed, for each GATA factor, an adenine was recovered at the highest frequency at this position and thymine was the base selected next most often. Surprisingly, binding sites with a C at this position were also recovered, but at an even lower frequency (Fig. 1C and D, 2C and D, and 3C and D). An unanticipated result found was that not only are certain nucleotides selected, but clearly some nucleotides also appear to be prohibited. For all three GATA factors, there is a virtual absence of guanine at positions −1 and +1; of 143 single sites sequenced, only 2 contained a G at position −1 (C8-2 and C7-3; Fig. 3A) and only 1 contained a G at +1 (C12-3; Fig. 3A). Not surprisingly, they were selected by GATA-3, the factor with the greatest apparent latitude in binding-site preference. Finally, although the +2 position is defined by the erythroid consensus as an obligate purine (**WGATA**R), the binding-site data indicate an almost exclusive selection for an A or T at this position, not G or C (Fig. 1C and D, 2C and D, 3C and D).

Overall, it does not appear that any particular nucleotide is favored at position −3 or −2, 5′ to the GAT core. In contrast, the nucleotides 3′ to the core appear to contribute significantly to the ability of a binding site to be recognized by a particular GATA factor. As discussed above, the +1 and +2 positions are highly selected, and while other nucleotides are tolerated, a preference for an A at position +4 is also apparent for all three factors (Fig. 1C and D, 2C and D, and 3D). Although from the compiled analysis of the selected sites there does not appear to be any overall preference either for or against any particular nucleotide at position +3, some sequence motifs are more frequently observed among the selected sites (e.g., **GATAAGA** or **GATCTTA**), suggesting that the selection for 3′ nucleotides in creating a GATA factor-binding site is not independent of the identities at other nearby positions.

Because of the concern that the binding specificities of the bacterially overexpressed fusion protein might not necessarily reflect the specificities of the native, endogenous proteins, selections were also performed in parallel with partially purified cGATA-1 isolated from mature adult chicken erythrocytes (see Materials and Methods). The identical consensus was derived (NNA**GATAA**NN), and selected sites from this procedure are denoted with the prefix W in Fig. 4 and Table 2.

In summary, GATA-1 selected the most highly defined binding site, NNA**GATAA**NN, while GATA-2 and -3 selected a broader range of sites. That GATA-1 recognizes only a subset of potential sites for GATA-2 and -3 was also confirmed by using each of the third-round-selected oligonucleotide pools from the second experiment in gel mobility shift assays with all three factors. The pool of sites selected by GATA-1 were bound equally well by all three transcription factors, while pools of sites selected by GATA-2 and GATA-3 gave a strong shifted band with the reciprocal factor but only a weak one with GATA-1 (data not shown).

Analysis of individual selected GATA factor-binding sites. Although one may derive significant information from compiled results, the function of these transcription factors ultimately involves their interaction with single, specific sites. Therefore, several of the selected single sites were analyzed for the ability to be recognized by the three GATA factor family members. Individual sites were amplified from subclones by PCR, radiolabeled, and used in gel mobility shift assays with the bacterially expressed GATA fusion proteins.

The amount of each protein to be used in each gel mobility shift assay was arbitrarily defined as the quantity required to yield comparable intensities of lower-mobility bands indicating protein-DNA complex formation by using the $M\alpha P_{30}$ probe (see Materials and Methods). The relative concentrations of all three proteins used in these binding reactions, when visualized by a Coomassie-stained gel, yielded roughly comparable amounts of proteins (data not shown). While affinities of the three proteins for the $M\alpha P_{30}$ site have not been rigorously proven by determination of the amount of specific DNA-binding activity in each of the three different bacterially expressed extracts, the similar dissociation rates of GATA-1, -2, and -3 from a WGATAR consensus site (see below) suggest that the affinities of all three are comparable.

As would be predicted from the selection analysis, different DNA sequences are recognized with different affinities by GATA-1, -2, and -3. All three factors bind equally well to an A**GATA**A**G**A site, but alterations of positions +1 to +3 (to A**GATCTT**A) result in sites that can be recognized by GATA-2 and GATA-3 but not by GATA-1 (Fig. 4 and Table 2; compare B20 and CC21 or W5 and CC24). In both examples, the binding of GATA-2 and -3 to the A**GATAA**GA site was of greater affinity than that to the A**GATCTT**A site. Similarly, the sequence A**GATTA** is bound weakly by GATA-2 and -3 but only minimally by GATA-1 (CC2 and CC31).

Changes at position −1 can affect recognition by the factors even when the site conforms to the **GATAAGA** consensus derived from the selections; thus, CA**TGAT**AAGA (B35) is recognized by each of the GATA factors with higher affinity than either CA**C**GATAAGA (W19) or CA**A**GATAAGA (B20), even though the latter also conforms to the originally defined erythroid consensus sequence (Fig. 4 and Table 2). The sequence GC**A**GATAAGA (W5) was recognized by all three GATA factors strongly, and compar-

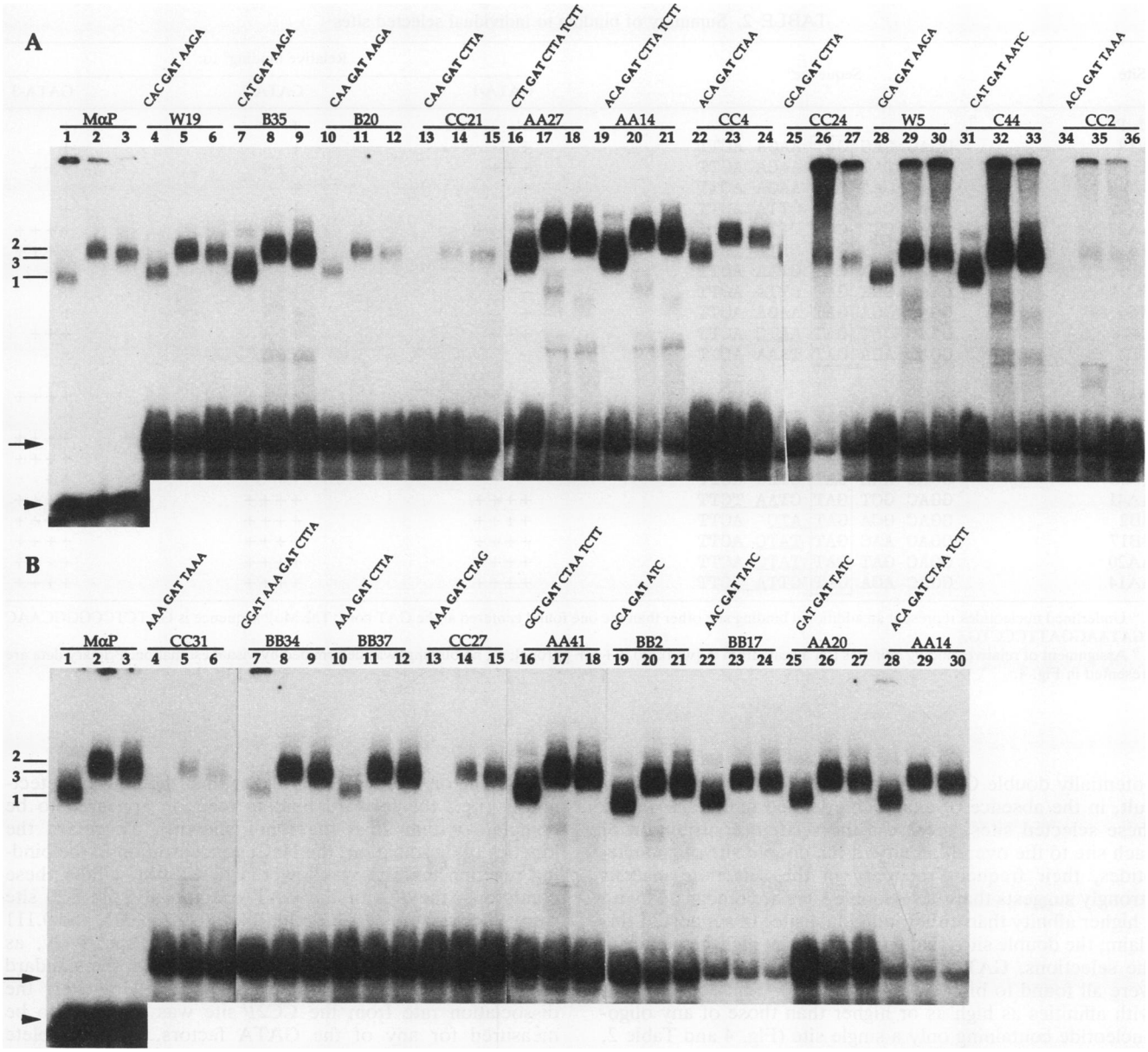

FIG. 4. Binding of the GATA factors to individual selected sites. Cloned GATA-binding sites identified by the selection procedure (Fig. 1 to 3) were amplified by PCR, radiolabeled, and used as probes in gel mobility shift assays with GATA-1, -2, and -3. The amount of protein was normalized to give an equivalent intensity of a shifted band by the $M\alpha P_{30}$ probe and used in binding reactions with equal numbers of counts of each probe. Individual sites are identified by name (e.g., B35; Fig. 1, 2, and 3), and the sequence of each is also shown above each set of lanes. The GAT core is shown in boldface, bordered by selected nucleotides. Flanking fixed regions are shown if mutations in the region created new binding sites. The migration of the free $M\alpha P_{30}$ probe is indicated by the arrowhead, and that of the individual site probes is indicated by the arrow. Markers 1, 2, and 3 show the GATA factor-DNA complexes of GATA-1, -2, and -3, respectively, with each probe. Gels represent two independent experiments; comparisons can be made about relative affinities of binding within an experiment but not between experiments. (A) cGATA-1 (lanes 1, 4, 7, 10, 13, 16, 19, 22, 25, 28, 31, and 34), cGATA-2 (lanes 2, 5, 8, 11, 14, 17, 20, 23, 26, 29, 32, and 35), and cGATA-3 (lanes 3, 6, 9, 12, 15, 18, 21, 24, 27, 30, 33, and 36); (B) cGATA-1 (lanes 1, 4, 7, 10, 13, 16, 19, 22, 25, and 28), cGATA-2 (lanes 2, 5, 8, 11, 14, 17, 20, 23, 26, and 29), and cGATA-3 (lanes 3, 6, 9, 12, 15, 18, 21, 24, 27, and 30).

ison of the binding results between B20 and W5 indicates that although there was no overall preference for nucleotide identity at −3 and −2 in the analysis of compiled data, at a specific site changes at these positions can affect the affinity with which it will be recognized by GATA-binding proteins. Similarly, changes in positions +3 and +4 can also alter binding affinity, as demonstrated in the comparison between BB37 (AAAGATCTTA) and CC27 (AAAGATCTAG) (Fig. 4B and Table 2). Consistent with the selection results, although positions +1 and +2 are critically important in determining whether a GATA factor will bind to a particular sequence, flanking positions also significantly contribute to the equilibrium affinity of the factor for the site.

Several of the sites recovered from the selection were

TABLE 2. Summary of binding to individual selected sites

Site	Sequence[a]	Relative binding[b] to:		
		GATA-1	GATA-2	GATA-3
MαP		+	+	+
W19	GGAC CAC GAT AAGA ACTT	+	+	+
B35	GGAC CAT GAT AAGA ACTT	+++	+++	+++
B20	GGAC CAA GAT AAGA ACTT	+	+	+
CC21	GCAC CAA GAT CTTA ACTT	−	+	+
AA27	GGAC CTT GAT CTTA TCTT	++++	++++	++++
AA14	GGAC ACA GAT CTTA TCTT	++++	++++	++++
CC4	GGAC ACA GAT CTAA ACTT	++	++	++
CC24	GGAC GCA GAT CTTA ACTT	−	+	+
W5	GGAC GCA GAT AAGA ACTT	+	+	+
C44	GGAC CAT GAT AATC ACTT	+++	+++	+++
CC2	GGAC ACA GAT TAAA ACTT	−	±	±
MαP		++++	++++	++++
CC31	GGAC AAA GAT TAAA ACTT	±	+	+
BB34	GGAT AAA GAT CTTA ACTT	+	++++	++++
BB37	GGAC AAA GAT CTTA ACTT	+	++++	++++
CC28	GGAC AAA GAT CTAG ACTT	−	++	++
AA41	GGAC GCT GAT CTAA TCTT	++++	++++	++++
BB2	GGAC GCA GAT ATC ACTT	++++	++++	++++
BB17	GGAC AAC GAT TATC ACTT	++++	++++	++++
AA20	GGAC GAT GAT TATC ACTT	++++	++++	++++
AA14	GGAC ACA GAT CTTA TCTT	++++	++++	++++

[a] Underlined nucleotides represent an additional binding site other than the one found centered at the GAT core. The MαP sequence is GATCTCCGGGCAAC TGATAAGGATTCCCTG.

[b] Assignment of relative binding affinities (on a scale from ± [weak] to ++++ [strong]; −, no binding) was determined by visual estimation. Primary data are presented in Fig. 4.

potentially double GATA-binding sites. Although it is difficult, in the absence of examining defined mutations within these selected sites, to assess the relative contribution of each site to the overall affinity of the double-site oligonucleotides, their frequent recovery in the selection analysis strongly suggests that these double sites are recognized with a higher affinity than either individual site. In support of this claim, the double sites that were found at a high frequency in the selections, **GATATC**, **GATTATC**, and **GATCTTATC**, were all found to bind GATA-1, -2, and -3 all equally well, with affinities as high as or higher than those of any oligonucleotide containing only a single site (Fig. 4 and Table 2, AA41, BB2, BB17, AA20, and AA14).

In summary, gel mobility shift assays examining the relative affinities of each GATA factor for individual selected sites generally confirmed the original selection results: as a rule, the affinity of each type of site was roughly proportional to its representation in the pool of selected sites. GATA-1 bound well to **GATAAGA** sites but very poorly to sites with the motif **GATCTTA** or **GATTA**, consistent with the lack of representation of the last two types of sites in those selected by GATA-1 (Fig. 1 and 4). However, GATA-2 and GATA-3 have considerably broader recognition abilities in that they can bind with high affinities to both the canonical erythroid consensus sites and the newly defined, alternative consensus sites (Fig. 2 through 4).

Determination of GATA factor-DNA dissociation rates. To investigate the biochemical basis for the differences in equilibrium binding affinities described above, association and dissociation rates were analyzed for each of the GATA factors for two binding sites representative of the erythroid and alternative consensus sites recovered from selection, B20 (CAA**GATAAGA**) and CC21 (CAA**GATCTTA**), respectively. The on rates were all higher than the limit of detectability (i.e., the forward binding reaction appeared to be complete within 20 s [data not shown]). To retard the kinetics of dissociation, the NaCl concentration in the binding reaction mixture was lowered to 40 mM. Under these conditions, the K_ds of the GATA factors for the B20 site were all comparable (0.106 ± 0.014, 0.099 ± 0.001, and 0.111 ± 0.003 min^{-1} for GATA-1, -2, and -3, respectively, as averages from two independent experiments ± the standard error of the mean; Fig. 5D, E, and F). However, the dissociation rate from the CC21 site was too high to be measured for any of the GATA factors, with complete dissociation occurring within 20 s of the addition of competitor (Fig. 5A, B, and C and data not shown).

Equilibrium affinity of the GATA factors for an erythroid consensus versus an alternative consensus GATA-binding site. Because we sought to quantitate the relative affinity of cGATA-1, -2, and -3 for a pair of matched sites representing the two types of sequences identified during the selection process, B20 and CC21 sites were used as specific competitors in gel mobility shift assays as representative of the erythroid and alternative consensus sites, respectively. As indicated in Fig. 6A, cGATA-1 had a 16-fold-higher affinity for the B20 site than the CC21 site, whereas for both cGATA-2 and cGATA-3, the B20 site displayed only a 2-fold-higher affinity for binding than did the CC21 site (Fig. 6B and C). These results are consistent with the binding-site selection data, in which the dramatic difference in affinity of GATA-1 resulted in recovery of only erythroid consensus sites. With cGATA-2 and -3, however, a more comparable affinity between the two types of sites allowed the selection of the alternative recognition sequence for these GATA factors. Given the similar relative affinities defined here by

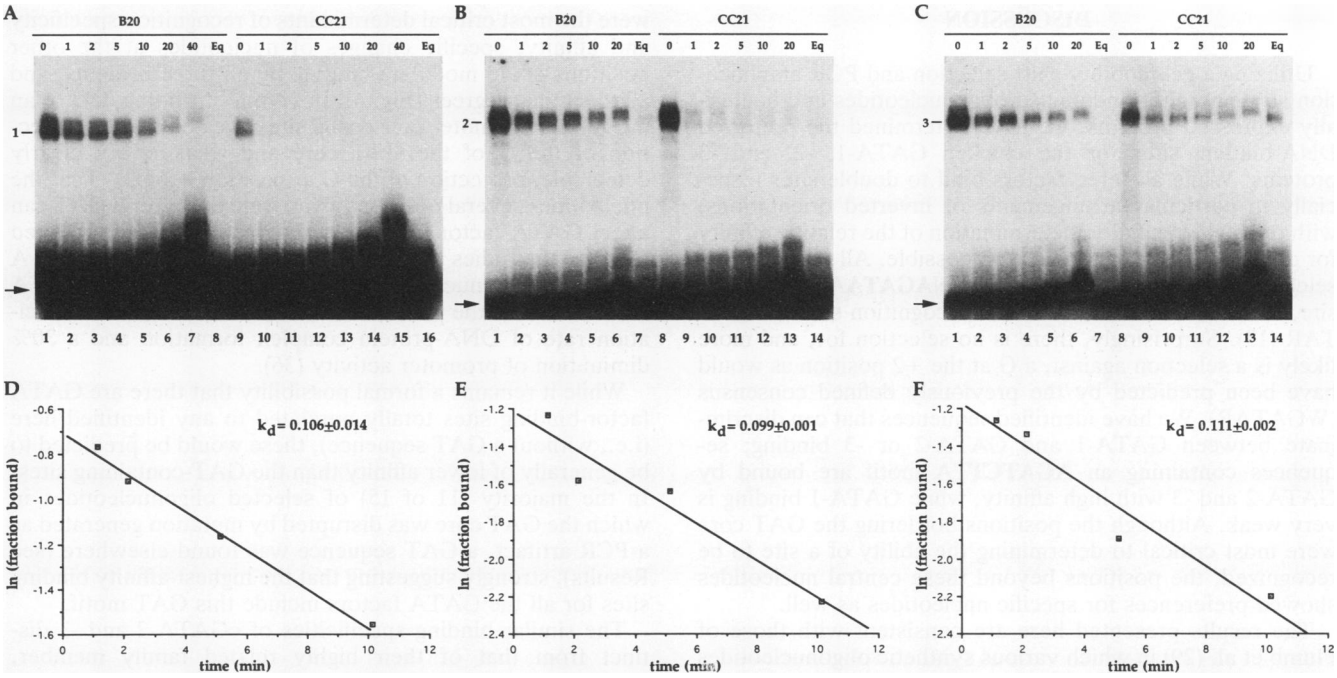

FIG. 5. Dissociation rate constant determination for the B20 and CC21 sites. Four nanograms of probe was used in binding reactions with cGATA-1 (A), cGATA-2 (B), and cGATA-3 (C). First, the radiolabeled binding sites plus factor were allowed to come to binding equilibrium. A 100-fold excess of unlabeled MαP$_{30}$ was then added to the reaction, and at the times indicated (in minutes) above each lane, aliquots of the reaction mixture were loaded onto a native polyacrylamide gel during electrophoresis. The migration of the free probe is indicated by the arrow, and markers 1, 2, and 3 show the GATA factor-DNA complexes of GATA-1, -2, and -3, respectively, with each probe. Lanes Eq, the endpoint of the competitions in binding reactions with competitor after each was allowed to reach equilibrium. (D, E, and F) Quantitative evaluation of the results of GATA-1 (D), GATA-2 (E), and GATA-3 (F) dissociation from B20. The K_d shown is the average of two independent determinations ± the standard error of the mean. See Materials and Methods for details of the kinetic analysis.

GATA-2 and -3, why would cGATA-2 not select the alternative (GATCT) sites as frequently as cGATA-3 (Table 1)? One possibility is that the analysis of only the B20 and CC21 sites is too limited and that these results are not be easily extrapolated to sites with differing identities at the −1 to −3 positions. Thus, GATA-2 could recognize the erythroid consensus site at a much higher relative affinity compared with that of the alternative site if both binding sites contained residues different from those within the two selected sites examined here.

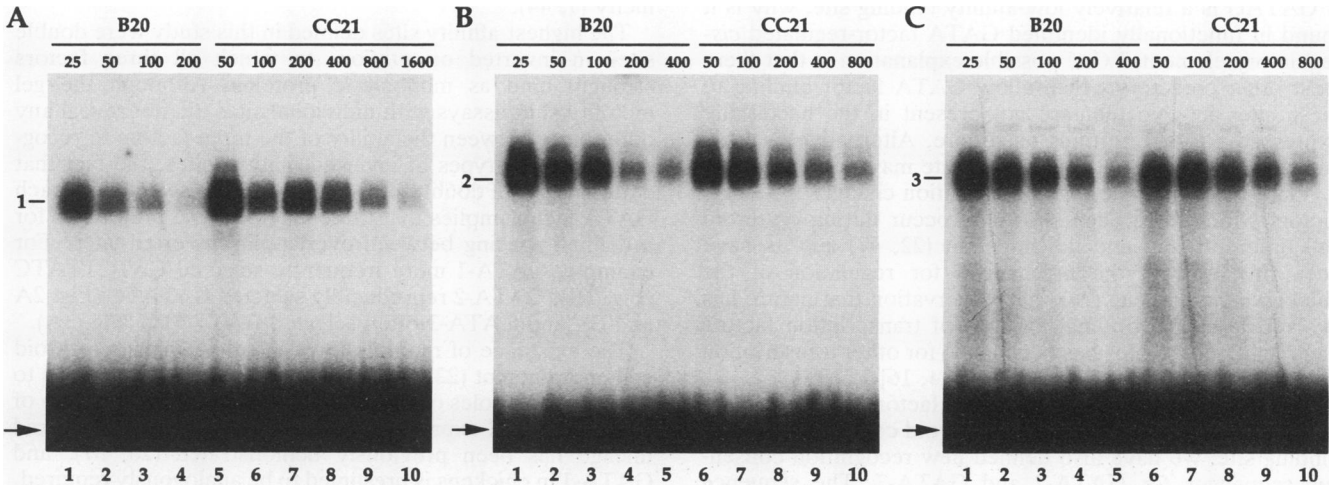

FIG. 6. Relative affinities of cGATA-1, -2, and -3 for sites B20 and CC21. Binding reactions were performed with GATA-1 (A), -2 (B), and -3 (C); the MαP$_{30}$ probe; and increasing amounts (indicated in nanograms above each lane) of either B20 or CC21 binding-site competitor. The migration of the free MαP$_{30}$ probe is indicated by the arrow, and markers 1, 2, and 3 show the GATA factor-DNA complexes of GATA-1, -2, and -3, respectively.

DISCUSSION

Utilizing a gel mobility shift selection and PCR amplification strategy with randomized oligonucleotides and bacterially expressed proteins, we have determined the preferred DNA-binding sites for the chicken GATA-1, -2, and -3 proteins. While all three factors bind to double sites (especially in particular arrangements of inverted orientations) with the highest affinity, determination of the relative affinity for preferred single sites was also possible. All three factors selected an erythroid consensus **NNAGATAANN** binding site, in accord with the canonical recognition site of WGATAR (13). Surprisingly, there is no selection for, and more likely is a selection against, a G at the +2 position as would have been predicted by the previously defined consensus (WGATAR). We have identified sequences that can discriminate between GATA-1 and GATA-2 or -3 binding; sequences containing an **AGATCTTA** motif are bound by GATA-2 and -3 with high affinity, while GATA-1 binding is very weak. Although the positions bordering the GAT core were most critical to determining the ability of a site to be recognized, the positions beyond these central nucleotides showed preferences for specific nucleotides as well.

The results presented here are consistent with those of Plumb et al. (29) in which various synthetic oligonucleotides with mutations in the mouse α-globin promoter GATA site, MαP, were assessed for their ability to compete for binding with a wild-type MαP site. Mutation of the two adenines at positions +1 and +2 were shown to dramatically reduce competition of binding (20- and 4-fold, respectively), but the +3 site could be altered with no difference in ability to compete with the wild type, consistent with the presence of an N at this position in the consensus derived here. Furthermore, competition with an MαP site mutated to a **GATAG** and the **GATAG** site from the chicken β^H-globin promoter was shown to compete four- to sixfold less well than the wild-type **GATAA** MαP site.

Previously identified GATAG sites have been found in the chicken α-globin enhancer, the chicken β^H-globin promoter, the human β-globin enhancer and $^A\gamma$-globin promoter, and the chicken β-globin enhancer (12, 13, 20, 24, 29, 33, 46); the last two sites have also been shown to be required for full erythroid transcriptional activity of these elements (23, 32). If GATAG is a relatively low-affinity binding site, why is it found in functionally identified GATA factor-regulated *cis*-regulatory elements? One possible explanation is that there exist other cofactors which allow GATA factor binding to these sites in vivo that are not present in the bacterially expressed purified proteins used here. Alternatively, these sites may have evolved so that the site may be more easily regulated, for example, by concentration changes in GATA factors which have been shown to occur during erythroid cell differentiation and development (22, 47) and as have been proposed in the mechanism for regulation of the chicken ρ-globin gene (26). The observation that nature has evolved less than optimal pairings of transcription factors and their cognate sites has been made for other transcription factors (e.g., for CACC and c-*myc* [14, 16]).

In addition to showing that all three factor family members have similar preferences for an erythroid consensus GATA-binding site, we have also defined new recognition consensus sequences for GATA-2 and GATA-3. The sequence **AGATCTTA** is a high-affinity site while **AGATTA** is a relatively low-affinity site for both GATA-2 and GATA-3; GATA-1 does not bind to either type of site well. Although the precise nucleotide identities of positions −1, +1, and +2 were the most critical determinants of recognition specificity and affinity, specific changes of nucleotides at the other positions could modulate binding by all three proteins, and all to similar degrees (Fig. 4). In vivo footprinting data of an mGATA-1 promoter consensus site revealed strong protection of the G of the GAT core and weaker, but clearly detectable, protection of the G at position +3 (43). That the nucleotides several positions away from the central GAT can affect GATA factor function through that site is evidenced by recent studies of the cGATA-1 promoter region. A deletion of the nucleotide at the −3 position of the GATA site at −139 of the promoter results in an increased dissociation rate of DNA-protein complex formation and a 30% diminution of promoter activity (36).

While it remains a formal possibility that there are GATA factor-binding sites totally unrelated to any identified here (i.e., without a GAT sequence), these would be predicted to be generally of lower affinity than the GAT-containing sites. In the majority (11 of 15) of selected oligonucleotides in which the GAT core was disrupted by mutation generated as a PCR artifact, a GAT sequence was found elsewhere (see Results), strongly suggesting that the highest-affinity binding sites for all the GATA factors include this GAT motif.

The similar binding specificities of cGATA-2 and -3 distinct from that of their highly related family member, cGATA-1, was not unanticipated; in the 107 amino acids of homology in the fingers and adjacent basic region (which is required for cGATA-1 DNA binding [49]) of all three factors, there are only three residues which differ between cGATA-2 and -3, but there are 18 differences in the cGATA-1 sequence compared with the cGATA-2 sequence (11, 48). In an analysis of the function of the two zinc fingers in mouse and chicken GATA-1, the C-terminal finger has been identified as containing the critical DNA-binding function, while the N-terminal finger plays a role in stabilizing protein-DNA complex formation (23, 49). Five of the amino acid differences between cGATA-1 and cGATA-2 are found in the C-terminal finger, and notably, one of the differences is found between the first pair of cysteines in the C-terminal finger (at position 165 in cGATA-1 [11]). Significantly, a single amino acid change in this "knuckle" region between the conserved cysteines in the steroid hormone receptors has been shown to be sufficient to alter DNA-binding specificity (7, 44).

The highest-affinity sites defined in this study were double sites in inverted orientation, to which all three factors strongly bind as monomeric proteins. Although the gel mobility shift assays with individual sites did not reveal any differences between the ability of the three factors to recognize the three types of inverted binding sites, the fact that certain types of double sites appeared to be selected by each GATA factor implies that there may be some preference for a defined spacing between overlapping inverted sites. For example, GATA-1 more frequently selected **GATCTTATC** (Fig. 1B), GATA-2 reproducibly selected **GATATC** (Fig. 2A and B), and GATA-3 often selected **GATTATC** (Fig. 3A).

The existence of multiple GATA factors in the erythroid cell environment (22, 48) poses an interesting question as to the functional roles of cGATA-1, -2, and -3. The necessity of mouse GATA-1 expression for development of the erythroid lineage has been previously demonstrated (28, 37), and GATA-1 in chickens is presumed to be analogously required. It is interesting to note that cGATA-3 expression seems to correlate with later stages of differentiation in erythrocytes and nervous system tissues (15a, 22, 48). Given the broader recognition ability of GATA-3, it is conceivable that upon its

induction during late erythroid differentiation, GATA-3 could replace GATA-1 at some erythroid consensus binding sites, maintaining GATA factor function at those target genes but additionally recognizing another battery of alternative sites which regulate genes expressed later in differentiation. The identification in this study of a high-affinity, noncanonical binding site for GATA-2 and GATA-3, AGAT CTTA, may facilitate the identification of downstream target genes for these factors in erythroid as well as nonerythroid cells (3).

The determination of which GATA factor acts at a given consensus site appears to be dictated by some aspect of factor function beyond simple differences in site recognition properties. The octamer binding proteins Oct-1 and Oct-2, which also have identical binding specificities, have been proposed to display promoter selectivity by interactions of their activation domains with other proteins at the transcriptional initiation complexes (40). A similar model, in which specificity is localized to regions outside the DNA-binding domain, has also been explored for the differential activation abilities of MRF4 and myogenin on the muscle creatinine kinase enhancer (5). A mechanism imparting differential specificity may also be operative in this case as well; experiments are currently under way to test this hypothesis.

ACKNOWLEDGMENTS

We thank Brett Andres for indispensable technical assistance; Mark Leonard, Zhuoying Yang, and Jon Widom for technical advice and valuable discussions; and Katie George, Kevin Foley, and Matt Roth for critical reading of the manuscript. We also thank Christina Ko for the contribution of reagents critical to the successful completion of this project.

This work was supported by an NIH NRSA training grant award to Northwestern University (L.J.K.; GM 08061) and an NIH research grant (GM 28896).

REFERENCES

1. **Blackwell, T. K., L. Kretzner, E. M. Blackwood, R. N. Eisenman, and H. Weintraub.** 1990. Sequence-specific DNA binding by the c-Myc protein. Science **250:**1149–1151.
2. **Blackwell, T. K., and H. Weintraub.** 1990. Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. Science **250:**1104–1110.
3. **Briegel, K., K.-C. Lim, C. Plank, H. Beug, J. D. Engel, and M. Zenke.** Ectopic expression of a conditional GATA-2/estrogen receptor chimera arrests erythroblast differentiation in a hormone-dependent manner. Genes Dev., in press.
4. **Catala, F., E. deBoer, G. Habets, and F. Grosveld.** 1989. Nuclear protein factors and erythroid transcription of the human $^{A}\gamma$-globin gene. Nucleic Acids Res. **17:**3811–3826.
5. **Chakraborty, T., and E. N. Olson.** 1991. Domains outside of the DNA-binding domain impart target gene specificity to myogenin and MRF4. Mol. Cell. Biol. **11:**6103–6108.
6. **Choi, O.-R., and J. D. Engel.** 1986. A 3′ enhancer is required for temporal and tissue-specific transcriptional activation of the chicken adult β-globin gene. Nature (London) **323:**731–734.
7. **Danielsen, M., L. Hinck, and G. M. Ringold.** 1989. Two amino acids within the knuckle of the first zinc finger specify DNA response element activation by the glucocorticoid receptor. Cell **57:**1131–1138.
8. **Dessain, S., C. T. Gross, M. A. Kuziora, and W. McGinnis.** 1992. Antp-type homeodomains have distinct DNA binding specificities that correlate with their different regulatory functions in embryos. EMBO J. **11:**991–1002.
9. **Ekker, S. C., D. P. von Kessler, and P. A. Beachy.** 1992. Differential DNA sequence recognition is a determinant of specificity in homeotic gene action. EMBO J. **11:**4059–4072.
10. **Emerson, B. M., J. M. Nickol, P. D. Jackson, and G. Felsenfeld.** 1987. Analysis of the tissue-specific enhancer at the 3′ end of the chicken adult β-globin gene. Proc. Natl. Acad. Sci. USA **84:**4786–4790.
11. **Evans, T., and G. Felsenfeld.** 1989. The erythroid-specific transcription factor Eryf1: a new finger protein. Cell **58:**877–885.
12. **Evans, T., and G. Felsenfeld.** 1991. *trans*-activation of a globin promoter in nonerythroid cells. Mol. Cell. Biol. **11:**843–853.
13. **Evans, T., M. Reitman, and G. Felsenfeld.** 1988. An erythrocyte-specific DNA-binding factor recognizes a regulatory sequence common to all chicken globin genes. Proc. Natl. Acad. Sci. USA **85:**5976–5980.
14. **Fisher, D. E., L. A. Parent, and P. A. Sharp.** 1993. High affinity DNA-binding Myc analogs: recognition by an α helix. Cell **72:**467–476.
15. **Gallarda, J. L., K. P. Foley, Z. Yang, and J. D. Engel.** 1989. The β-globin stage selector element factor is erythroid-specific promoter/enhancer binding protein NF-E4. Genes Dev. **3:**1845–1859.

15a. **George, K., J. Kornhauser, and J. D. Engel.** Unpublished observations.

16. **Hartzog, G. A., and R. M. Myers.** 1993. Discrimination among potential activators of the β-globin CACCC element by correlation of binding and transcriptional properties. Mol. Cell. Biol. **13:**44–56.
17. **Ho, I.-C., L.-H. Yang, G. Morle, and J. M. Leiden.** 1989. A T-cell-specific transcriptional enhancer element 3′ of Cα in the human T-cell receptor α locus. Proc. Natl. Acad. Sci. USA **86:**6714–6718.
18. **Ito, E., T. Toki, H. Ishihara, H. Ohtani, L. Gu, M. Yokoyama, J. D. Engel, and M. Yamamoto.** 1993. Erythroid transcription factor GATA-1 is abundantly transcribed in mouse testis. Nature (London) **362:**466–469.
19. **Joulin, V., D. Bories, J.-F. Eleouet, M.-C. Labastie, S. Chretien, M.-G. Mattei, and P.-H. Romeo.** 1991. A T-cell specific TCR δ DNA binding protein is a member of the human GATA family. EMBO J. **10:**1809–1816.
20. **Knezetic, J. A., and G. Felsenfeld.** 1989. Identification and characterization of a chicken α-globin enhancer. Mol. Cell. Biol. **9:**893–901.
21. **Ko, L. J., M. Yamamoto, M. W. Leonard, K. M. George, P. Ting, and J. D. Engel.** 1991. Murine and human T-lymphocyte GATA-3 factors mediate transcription through a *cis*-regulatory element within the human T-cell receptor δ gene enhancer. Mol. Cell. Biol. **11:**2778–2784.

21a. **Kornhauser, J., K. George, and J. D. Engel.** Unpublished observations.

22. **Leonard, M. W., K.-C. Lim, and J. D. Engel.** Expression of the GATA transcription factor family during early erythroid development and differentiation. Submitted for publication.
23. **Martin, D. I. K., and S. H. Orkin.** 1990. Transcriptional activation and DNA-binding by the erythroid factor GF-1/NF-E1/Eryf1. Genes Dev. **4:**1886–1898.
24. **Martin, D. I. K., S.-F. Tsai, and S. H. Orkin.** 1989. Increased γ-globin expression in a nondeletion HPFH mediated by an erythroid-specific DNA-binding factor. Nature (London) **338:**435–438.
25. **Martin, D. I. K., L. I. Zon, G. Mutter, and S. H. Orkin.** 1990. Expression of an erythroid transcription factor in megakaryocytic and mast cell lineages. Nature (London) **344:**444–447.
26. **Minie, M., T. Kimura, and G. Felsenfeld.** 1992. The developmental switch in embryonic ρ-globin expression is correlated with erythroid lineage-specific differences in transcription factor levels. Development **115:**1149–1164.
27. **Naar, A. M., J.-M. Boutin, S. M. Lipkin, V. C. Yu, J. M. Holloway, C. K. Glass, and M. G. Rosenfeld.** 1991. The orientation and spacing of core DNA-binding motifs dictate selective transcriptional responses to three nuclear receptors. Cell **85:**1267–1279.
28. **Pevny, L., M. C. Simon, E. Robertson, W. H. Klein, S.-F. Tsai, V. D'Agati, S. H. Orkin, and F. Costantini.** 1991. Erythroid differentiation in chimaeric mice blocked by a targeted mutation in the gene for transcription factor GATA-1. Nature (London) **349:**257–260.

29. **Plumb, M., J. Frampton, H. Wainwright, M. Walker, K. Macleod, G. Goodwin, and P. Harrison.** 1989. GATAAG; a cis-control region binding an erythroid-specific nuclear factor with a role in globin and non-globin gene expression. Nucleic Acids Res. **17**:73–92.
30. **Redondo, J. M., S. Hata, C. Brocklehurst, and M. S. Krangel.** 1990. A T cell-specific transcriptional enhancer within the human T cell receptor δ locus. Science **247**:1225–1229.
31. **Redondo, J. M., J. L. Pfohl, and M. S. Krangel.** 1991. Identification of an essential site for transcriptional activation within the human T-cell receptor δ enhancer. Mol. Cell. Biol. **11**:5671–5680.
32. **Reitman, M., and G. Felsenfeld.** 1988. Mutational analysis of the chicken β-globin enhancer reveals two positive-acting domains. Proc. Natl. Acad. Sci. USA **85**:6267–6271.
33. **Reitman, M., E. Lee, H. Westphal, and G. Felsenfeld.** 1990. Site-independent expression of the chicken β^A-globin gene in transgenic mice. Nature (London) **348**:749–752.
34. **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.
35. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.
36. **Schwartzbauer, G., K. Schlesinger, and T. Evans.** 1992. Interaction of the erythroid transcription factor cGATA-1 with a critical autoregulatory element. Nucleic Acids Res. **20**:4429–4436.
37. **Simon, M. C., L. Pevny, M. V. Wiles, G. Keller, F. Costantini, and S. H. Orkin.** 1992. Rescue of erythroid development in gene targeted GATA-1-mouse embryonic stem cells. Nat. Genet. **1**:92–98.
38. **Smith, D. B., and K. S. Johnson.** 1988. Single-step purification of polypeptides expressed in *Escherichia coli* as fusions with glutathionine S-transferase. Gene **67**:31–40.
39. **Studier, F. W., A. H. Rosenberg, J. J. Dunn, and J. W. Dubendorf.** 1990. Use of T7 RNA polymerase to direct expression of cloned genes. Methods Enzymol. **185**:60–89.
40. **Tanaka, M., J.-S. Lai, and W. Herr.** 1992. Promoter-selective activation domains in Oct-1 and Oct-2 direct differential activation of an snRNA and mRNA promoter. Cell **68**:755–767.
41. **Trainor, C. D., S. J. Stamler, and J. D. Engel.** 1987. Erythroid-specific transcription of the chicken histone H5 gene is directed by a 3′ enhancer. Nature (London) **328**:827–830.
42. **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.
43. **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.
44. **Umesono, K., and R. M. Evans.** 1989. Determinants of target gene specificity for steroid/thyroid hormone receptors. Cell **57**:1139–1146.
45. **Umesono, K., K. K. Murakami, C. C. Thompson, and R. M. Evans.** 1991. Direct repeats as selective response elements for the thyroid hormone, retinoic acid, and vitamin D3 receptors. Cell **65**:1255–1266.
46. **Wall, L., E. deBoer, and F. Grosveld.** 1988. The human β-globin gene 3′ enhancer contains multiple binding sites for an erythroid-specific protein. Genes Dev. **2**:1089–1100.
47. **Whitelaw, E., S.-F. Tsai, P. Hogben, and S. H. Orkin.** 1990. Regulated expression of globin chains and the erythroid transcription factor GATA-1 during erythropoiesis in the developing mouse. Mol. Cell. Biol. **10**:6596–6606.
48. **Yamamoto, M., L. J. Ko, M. W. Leonard, H. Beug, S. H. Orkin, and J. D. Engel.** 1990. Activity and tissue-specific expression of the transcription factor NF-E1 multigene family. Genes Dev. **4**:1650–1662.
49. **Yang, H.-Y., and T. Evans.** 1992. Distinct roles for the two cGATA-1 finger domains. Mol. Cell. Biol. **12**:4562–4570.
50. **Yang, Z., M. W. Leonard, L. J. Ko, K. M. George, M. Yamamoto, and J. D. Engel.** 1991. Transcription factors implicated in β-globin gene switching, p. 249–265. *In* G. Stamatoyannopoulos and A. W. Nienhuis (ed.), The regulation of hemoglobin switching. Johns Hopkins University Press, Baltimore.