The EMBO Journal vol.12 no.13 pp.4993-5005, 1993

The two zinc finger-like domains of GATA-1 have different DNA binding specificities

David J.Whyatt[1], Ernie deBoer[1] and Frank Grosveld[1,2]

Laboratory of Gene Structure and Expression, National Institute for Medical Research, Mill Hill, London NW7 1AA, UK
[1]Present address: Erasmus University, Department of Cell Biology, Postbus 1738, 3000 Rotterdam, The Netherlands
[2]Corresponding author

Communicated by F.Grosveld

The GATA-1 transcription factor has been shown to be important in the regulation of globin and non-globin genes in erythroid, megakaryocytic and mast cell lineages. It is a member of a family of GATA proteins which both overlap in their expression patterns and bind the motif (A/T)GATA(A/G). The GATA family of proteins are also members of the superfamily of zinc finger-like domain proteins and have two similar domains of the type Cys-X_2-Cys-X_{17}-Cys-X_2-Cys which direct the DNA binding of the protein. A random oligonucleotide selection procedure has been employed to further elucidate the mechanism of GATA-1-DNA recognition. The resulting oligonucleotides were tested for binding activity to both wild-type and mutant GATA-1 proteins. Two classes of GATA-1-DNA interaction have been defined, the first requiring only the carboxy finger of GATA-1 to bind and having the motif GAT(A/T), the second requiring both finger domains to bind and having the core motif (T/C)AAG. By using sequence comparison and depurination analysis it is concluded that the two finger-like domains of GATA-1 have different DNA binding recognition motifs. Binding of GATA-1 to GAT(A/T) motifs is associated with transcriptional activation of linked genes. The only known (T/C)AAG motif is in the distal CAAT-box promoter region of the human $^{A}\gamma$-globin gene where the binding of GATA-1 appears to regulate the correct developmental suppression of γ-globin expression.

Key words: DNA recognition/GATA-1/globin/transcription factors/zinc finger

Introduction

The murine GATA-1 transcription factor (Wall *et al.*, 1988; Martin *et al.*, 1989; Plumb *et al.*, 1989) is a DNA binding protein found in the erythroid, mast and megakaryocytic cell lineages (deBoer *et al.*, 1988; Martin *et al.*, 1990; Romeo *et al.*, 1990). It has also recently been described in the testis, where it is expressed in a subset of cells in the seminiferous tubules (Ito *et al.*, 1993). It contains two zinc finger-like domains of the configuration Cys-X_2-Cys-X_{17}-Cys-X_2-Cys (Tsai *et al.*, 1989), which have been shown to direct DNA binding of the protein (Martin and Orkin, 1990). Homologous GATA-1 proteins exist in the chicken (Evans and Felsenfeld, 1989), human (Trainor *et al.*, 1990), frog (Zon *et al.*, 1991) and the nematode *Caenorhabditis elegans* (Spieth *et al.*, 1991). The carboxy zinc finger-like domain alone is conserved in regulatory factors for nitrogen metabolism in *Aspergillus* (Kudla *et al.*, 1990), *Neurospora* (Fu and Marzluf, 1990) and *Saccharomyces cerevisiae* (Cunningham and Cooper, 1991).

GATA-1 has been shown to be a member of a multigene family that share homology in their zinc finger-like domains. GATA-2 and -3 have been cloned from the chicken, frog and human, although as yet only GATA-3 has been reported in the mouse (Yamamoto *et al.*, 1990; Ko *et al.*, 1991; Lee *et al.*, 1991; Zon *et al.*, 1991; Dorfman *et al.*, 1992). In the chicken, GATA-2 is widely expressed in various tissues including erythroid cells, and GATA-3 is found in the adult brain, embryonic kidney, T cells and a distinct subset of erythroid cells, i.e. adult reticulocytes (Yamamoto *et al.*, 1990). GATA-4 has recently been described in the mouse, where it is expressed in the heart, intestinal epithelium, primitive endoderm and gonads (Arceci *et al.*, 1993).

GATA-1 binds a core DNA consensus, GATA, in the promoter and enhancer sequences of various erythroid genes such as globin, porphobilinogen deaminase, carbonic anhydrase 1 and erythropoietin receptor genes (deBoer *et al.*, 1988; Brady *et al.*, 1989; Mignotte *et al.*, 1989a,b; Frampton *et al.*, 1990; Chiba *et al.*, 1991). GATA-1 is required for the correct differentiation of the erythroid lineage (Pevny *et al.*, 1991) and is a potent transcriptional regulator of its target genes. It is capable of transactivating promoters containing GATA-1 binding sites in transient assays (Martin and Orkin, 1990; Evans and Felsenfeld, 1991; Chiba *et al.*, 1991), and in similar experiments it has been implicated as a positive regulator of its own promoter (Nicolis *et al.*, 1991; Tsai *et al.*, 1991; Schwartzbauer *et al.*, 1992). Interestingly, the erythroid specific activity of the glycophorin B promoter has been shown to be dependent on GATA-1 mediated displacement of a repressor (Rahuel *et al.*, 1992). GATA-1 also binds in the human β-globin locus control region (LCR) (Philipsen *et al.*, 1990; Talbot *et al.*, 1990; Pruzina *et al.*, 1991) and is of vital importance in the activity of hypersensitive site 3 of the LCR in directing high-level expression of linked genes in transgenic mice (Philipsen *et al.*, 1993).

Although GATA-1 binds to the sequence WGATAR (where W = A or T and R = A or G) (Yamamoto *et al.*, 1990), this does not appear to be the complete definition of the GATA-1 consensus DNA recognition sequence. Firstly, not all sequences containing a WGATAR site will efficiently bind GATA-1 in a bandshift assay (Evans *et al.*, 1988). Evidence has accumulated that sequences either side of this core consensus also mediate GATA-1 DNA recognition (Yang and Evans, 1992; Schwartzbauer *et al.*, 1992). It has been shown that the carboxy zinc finger domain is absolutely necessary for DNA binding to the GATA motif and that the amino zinc finger may mediate recognition of subtle

variations around this core site (Martin and Orkin, 1990; Yang and Evans, 1992; Schwartzbauer *et al.*, 1992). Secondly, GATA-1 appears to bind regions where no WGATAR motif is present. Possibly the most important example of this phenomenon is in the human γ-globin distal CAAT-box, where GATA-1 DNA binding activity has been shown to be functionally important in the correct developmental regulation of this gene in transgenic mice (Berry *et al.*, 1992). The single point mutation in this region which reduces GATA-1 binding was first identified in the human condition known as Greek hereditary persistence of fetal haemoglobin (HPFH) (Collins *et al.*, 1985; Gelinas *et al.*, 1985). This condition has clinical importance, because elevated levels of γ-globin in adult life can alleviate β-thalassaemia and sickle cell anaemia. Thirdly, although the proteins of the GATA family overlap in their expression patterns, it has been suggested that they regulate different target genes in a tissue-specific manner (Ko *et al.*, 1991). Thus, because all GATA proteins seem to recognize a core GATA motif, differential DNA recognition of target sites would be expected to be found in sequences outside this core motif.

The system first developed to define SRF and c-fos binding sites (Pollock and Treisman, 1990) was exploited to fully define which DNA sequences GATA-1 recognizes. Crude nuclear extracts were first bound to a random pool of oligonucleotides and the resulting complexes were immunoprecipitated. Specific binding sites were then amplified by PCR, subcloned and sequenced as described by Pollock and Treisman (1990). Using these selected oligonucleotides, the role the different finger-like domains have in GATA-1−DNA interaction was tested utilizing mutant GATA-1 proteins. Thus, it is shown that GATA-1 can bind two different DNA sequence motifs. The first sequence recognized contains the core motif GAT(A/T) and binding by GATA-1 does not require the amino zinc finger-like domain. The second sequence to which GATA-1 will bind contains the core motif (T/C)AAG. Both the amino and carboxy zinc finger-like domains of the GATA-1 protein are required to direct stable interaction with this sequence.

a SINGLE "GATA" CONTAINING SITES:

		BANDSHIFT:
CLONE 1	GATCCTGTCGAGGTTGTCCTTATAACGTCGACGATAGAGGCG	nd
CLONE 21	GATCCTGTCGGGTAAGGCCTGATGAACAGATAAGGGGAGGCG	nd
CLONE 20	GATCCTGTCGTGGATAGGGGACTAATGGACTGTTACGAGGCG	+++
CLONE 10	GATCCTGTCGATACGGCCGTGGTAGGGTGGGGTTAGAGGCG	+
CLONE 22	AATTCGCCTCCCCAGATAGCGTCATAGTCCCCACCCCGACAG	nd
CLONE 14	AATTCGCCTCCTAGAGATAAGGATGCCCACCACGCCGACAG	nd
CLONE 12	AATTCGCCTCTAGGATCACTCCCTAGAGATAGTGGACGACAG	nd
CLONE 11	GATCCTGTCGGGAGGAAAGTAATGTACCGATATAATGAGGCG	++
CLONE 26	GATCCTGTCGACCCGAGGTGGTCAGTGTGGATACGTGAGGCG	+
CLONE 31	GATCCTGTCGGCCACTGCACGGCGAAGTGATAAGGGGAGGCG	+++
CLONE 38	GATCCTGTCGATGATAGTGGCTTATTCCCCCGGTCGAGGCG	+++
CLONE 45	GATCCTGTCGATAGTGGTAGTCAAGGGGCTAGGCGGGAGGCG	nd
CLONE 62	GATCCTGTCGGCAAAGATAAGGGTTTGAGGGGGAGGGAGGCG	+++
CLONE 68	GATCCTGTCGGCATGATAGTGGGCTAACGGAGCAGGGAGGCG	+++
CLONE 84	GATCCTGTCGTGCGCGGGGTGCGGAGCTTGATAACGGAGGCG	nd
CLONE 83	GATCCTGTCGCTCCCAACCCCACTTCATCGTAGATAGAGGCG	nd
CLONE 78	GATCCTGTCGATAATGTGGTTGCCGGTGGCCTTCCGGAGGCG	++
CLONE 96	GATCCTGTCGGCGCCACGGGTAGCGAACCTTGATAGGAGGCG	+
CLONE 92	GATCCTGTCGGTTATTGATAGGATCACATACGGCTGGAGGCG	nd
CLONE 79	GATCCTGTCGATAGGAGTAGTGCCTGTTACGGTGACGAGGCG	nd
CLONE 53	GATCCTGTCGATACTAAAATTGTCTGCCGGTTGAGAGAGGCG	nd
CLONE 76	GATCCTGTCGGCCTCGTGGTAAACTACGTGATAGCAGAGGCG	nd
CLONE 88	GATCCTGTCGGTTTTCGTACCATGAGATACACGGCTGAGGCG	nd
CLONE 61	AATTCGCCTCGCATTCAGATACGCTATTGATGTCCCCGACAG	+++
CLONE 9	GATCCTGTCGCGGGAAGATGTTGACGATATAGCAGGGAGGCG	+++
CLONE 67	GATCCTGTCGTAGTACCTTGGGACCAGGCAAGATAGGAGGCG	nd
CLONE 71	GATCCTGTCGATAAGACAAGAAGATCGCAGAGCGTGAGGCG	++
CLONE 63	GATCCTGTCGGGATAGGTGGATGTCAGGGGATGGTAGAGGCG	nd
CLONE 2	GATCCTGTCGATAAGTGTGCAGGGTAGGACTTGTTCGAGGCG	++
CLONE 3	AATTCGCCTCCTGTAAAAGATAAGCCTGCCCGCGACCGACAG	+++
CLONE 27	AATTCGCCTCGGACCCCGCCCATGGTCCCGATAACTCGACAG	+

TOTAL NUMBER OF SEQUENCES = 31

b SINGLE "GATT" CONSENSUS SITES :

		BANDSHIFT:
CLONE 16	GATCCTGTCGGGAGATTCATGCAGAGGCCGATGCGAGAGGCG	nd
CLONE 19	AATTCGCCTCTAACGTGATTATGGCTCCTCCCCCATCGACAG	nd
CLONE 47	GATCCTGTCGGGATGATTGGCAGCATTACGGTGCCTGAGGCG	nd
CLONE 91	GATCCTGTCGGATTAGCTAGGGTTGGTAGCTGGGCTGAGGCG	nd
CLONE 64	AATTCGCCTCGGGAGGATTATAACACCACACATCGACGACAG	nd
CLONE 90	GATCCTGTCGAGTGCGGTAGGCTGATTAGAGACGTCGAGGCG	nd
CLONE 85	AATTCGCCTCCCACCACAGATTACACTGGAGCGCGCCGACAG	nd
CLONE 59	GATCCTGTCGGAGGCAGGGAGATTAGCATTTGTCGGGAGGCG	+++
CLONE 94	GATCCTGTCGTGTGATGATCTGGATTGTGGCCTGGGAGGCG	+++
CLONE 13	GATCCTGTCGATTAGCGGTCCTCATTGAGTGGACTGGAGGCG	+
CLONE 32	AATTCGCCTCGTGAAGATTACTGGCACCGCATGCCAGCGACAG	+++
CLONE 54	GATCCTGTCGCGATTGCCGGGAAGGTCGAGAAGCAAGAGGCG	nd
CLONE 89	GATCCTGTCGTGATCGACGGATTACGGGGGAAGGCTGAGGCG	nd

TOTAL NUMBER OF SEQUENCES = 13

c DOUBLE "GATA" CONSENSUS SITES:

		BANDSHIFT:
CLONE 18	GATCCTGTCGATACCGCCATATCTGGGATGTCTGATGAGGCG	nd
CLONE 17	GATCCTGTCGCAGATAGTGATAGCACCTGCACTGTGGAGGCG	nd
CLONE 4	AATTCGCCTCCCGATACTGGCGTATAACGGCTATCACGACAG	++
CLONE 6	AATTCGCCTCGATATCCCTAGTACGATCTGGGCTATCGACAG	nd
CLONE 30	GATCCTGTCGATACAGGAGTTGGCGGCGGGGAGATAGAGGCG	+++
CLONE 43	GATCCTGTCGATAAGGCGCTATAGAAGCATATCGAGGAGGCG	nd
CLONE 58	GATCCTGTCGGGGGTCGAGGATAGGATAGATGCCTTGAGGCG	nd
CLONE 98	GATCCTGTCGTCATCGTAGATACTAGCCTTCTTATCGAGGCG	nd
CLONE 52	GATCCTGTCGGGGAGATAGGCATTGACTATATCGTGAGGCG	nd

TOTAL NUMBER OF SEQUENCES = 9
TOTAL NUMBER OF SITES = 18

d DOUBLE "GATT" CONSENSUS SITES:

		BANDSHIFT:
CLONE 81	GATCCTGTCGCCGCGCTCACTACGATTGGGATTATGGAGGCG	+++
CLONE 80	GATCCTGTCGGCTAAGTGACGGATTCGATTAGCGAGAGGCG	nd
CLONE 34	GATCCTGTCGTGTACTGATTAGCGCAACAGGATTGGGAGGCG	nd
CLONE 56	AATTCGCCTCGATTGACGTCGGATTTGGGTACTAGCGACAG	++

TOTAL NUMBER OF SEQUENCES = 4
TOTAL NUMBER OF SITES = 8

e BOTH "GATT" AND "GATA" CONSENSUS SITES:

CLONE 74	GATCCTGTCGATAAGGTTAGATTGGAGTGGGCACCGGAGGCG	nd
CLONE 73	GATCCTGTCGATGGCGAAATCACGATAACGGCTAATGAGGCG	nd
CLONE 72	GATCCTGTCGATAGGGATGTCCTGATGGATTGACTCGAGGCG	+
CLONE 69	GATCCTGTCGAGGATCTGATAGGGCTGGGATTGTCTGAGGCG	+++
CLONE 97	GATCCTGTCGGGATAAAGTGGCCGCATGGAATCATGAGGCG	nd
CLONE 57	GATCCTGTCGATTGGGGGGTGAGGATACTTTAAGTTGAGGCG	nd
CLONE 8	GATCCTGTCGATACGGGGGTACGAGCGGCCCCGATTGAGGCG	nd
CLONE 5	GATCCTGTCGGATATAGGAGGTGCAGGCCACAATCGGAGGCG	nd
CLONE 25	GATCCTGTCGGTCTCCCTGGAATGGATAAGATTATAGAGGCG	nd
CLONE 40	GATCCTGTCGGTGGATAAAGATTGTTAGGTCGTGGGGAGGCG	nd
CLONE 35	GATCCTGTCGAGGCATGCAGATTGTGTTGATAAATGGAGGCG	nd
CLONE 44	AATTCGCCTCCTCCATAAAACCTTAACGATAAATCTCGACAG	nd

TOTAL NUMBER OF SEQUENCES = 12
TOTAL NUMBER OF "GATA" SITES = 12
TOTAL NUMBER OF "GATT" SITES = 12

f BOTH "GATA" AND DOUBLE "GATT" CONSENSUS SITES:

CLONE 82	GATCCTGTCGATTATGATTGTCCGATATTTGACAGCGAGGCG	nd
CLONE 87	GATCCTGTCGCAAGATAACGGCAGAATCGAGGATTGGAGGCG	nd

TOTAL NUMBER OF SEQUENCES = 2
TOTAL NUMBER OF "GATA" SITES = 2
TOTAL NUMBER OF "GATT" SITES = 4

g SEQUENCES WITH NEITHER "GATA" NOR "GATT" SITES:

		BANDSHIFT:
CLONE 75	GATCCTGTCGGTGCGGAGTGGCAGACTCAGGCATTCGAGGCG	-
CLONE 70	AATTCGCCTCGACCCTGGCATCAGTTACACTCTCCCCGACAG	++
CLONE 15	AATTCGCCTCCATTTGGCGACAAGATCTTAGCCAGTCGACAG	++
CLONE 33	AATTCGCCTCCAGCCAGGCAGACCGGCCCGTGATGCGACAG	-
CLONE 37	AATTCGCCTCTGACAGTTAGGACCGTGCGACCTTTTCGACAG	nd
CLONE 28	AATTCGCCTCTTATTTCCTATGTAGATCGCGACGGGCGACAG	++
CLONE 23	GATCCTGTCGAGCGATCAGCGACCCCACGGCCAGAGGAGGCG	-
CLONE 66	AATTCGCCTCCAAGAACCCCGGGACAACGACTCCCACGACAG	-
CLONE 7	GATCCTGTCGGACAGTCCGACCATGCGGCGAGCGACGAGGCG	-
CLONE 95	AATTCGCCTCCGACCCACGTGCCAGAGTCCCCTGCTCGACAG	-
CLONE 86	AATTCGCCTCTAGTTGGGAAGGTTTGCCGCACCGCCCGACAG	-
CLONE 24	GATCCTGTCGGGCCGAGCTAAAAGTTTTTATGGTGTGAGGCG	-

TOTAL NUMBER OF SEQUENCES = 12

Fig. 1. Listing of clones subcloned and sequenced from the enriched pool. The random segment of the input oligonucleotide pool is shown in bold type. GATA and GATT motifs are underlined and GATG motifs are outlined. The clone numbers are indicated on the left and binding activity on the right (see Materials and methods) +++ = strong binding activity (>50% of probe bound), ++ = medium binding activity (10–50% of probe bound), + = weak binding activity (<10% of probe bound), − = no binding activity, nd = not done. (**a**) Single GATA motifs, (**b**) single GATT motifs, (**c**) double GATA motifs, (**d**) double GATT motifs, (**e**) both GATT and GATA motifs, (**f**) both GATA and double GATT motifs and (**g**) sequences with neither GATA nor GATT motifs.

Results

Crude nuclear extract prepared from murine erythroleukaemia cells was incubated with a random pool of oligonucleotides, in which a region of 26 random bp was flanked by two defined regions, 25 bp in length, containing *Bam*HI and *Eco*RI restriction sites (see Materials and methods). Specific complexes were immunobound to protein A–Sepharose beads using a rat mAb against murine GATA-1. The bound oligonucleotides were then isolated and amplified by PCR. An aliquot of this material was put back into a binding reaction and further enriched by immunoprecipitation and then re-amplified. After three rounds of binding, immunoprecipitation and amplification, a binding activity was detectable in the resulting oligonucleotide pool, as assayed by gel retardation analysis. This activity could be abolished using a 100-fold excess of an unlabelled GATA-1 binding oligonucleotide (derived from the human β-globin −200 promoter region). The protein–DNA complex was also recognized and its gel mobility reduced by the GATA-1 specific antibody, confirming that the binding activity was caused by GATA-1 (data not shown).

The enriched oligonucleotide pool was then digested with *Bam*HI and *Eco*RI restriction endonucleases, subcloned into the Bluescript plasmid and sequenced; 83 clones were sequenced. In 12 clones it was found that there were only 25 bp in the randomly defined region of the clone (rather than the expected 26), and these were sequenced in both directions to avoid erroneous sequence readings caused by sequencing artefacts. Those which were confirmed to contain only 25 bp in this region were probably products of infidelity in the elongation reactions, either in the oligonucleotide synthesis or in the PCR amplification.

The resulting sequences were then subdivided and sequence aligned in various groups. The following clones were found: 31 containing a single GATA motif (Figure 1a); 13 containing a GATT motif (Figure 1b); nine containing a double GATA motif (Figure 1c); four containing a double GATT motif (Figure 1d); 12 containing both a GATA and a GATT motif (Figure 1e); two containing a single GATA and a double GATT motif (Figure 1f); and 12 containing neither GATA nor GATT sites (Figure 1g).

When the occurrence of these sequences was analysed for statistical bias using a χ-squared test, only GATA and GATT appeared to have been selected for in the enrichment procedure. The sequence GATG was also found at a higher than expected frequency, although this appeared to be statistically insignificant (Table I).

Biases around the core GATA and GATT motifs were investigated by aligning those sequences which contained only one of either of these motifs (Tables II and III), because a second motif on any one oligonucleotide may have arisen by chance rather than by selection. A second constraint on this analysis was not to include sequences which extended into the 'defined' regions (up to 5 bp either side of the core motif), because such sequences would also have given a false bias.

As shown in Table II, the biases around a GATA motif mainly occurred 3′ of the core motif. Taking the first G of GATA as position 0, the only statistically significant bias 5′ of GATA was a preference for an A residue at position −1 and a selection against a G residue (defined here as $P < 0.05$ in a χ-squared test). Downstream of the core there was a preference for G residues at positions 4, 5, 6 and 7, and a selection against a T at position 4 and a C at position 5. There was no bias either 5′ of position −1 or 3′ of position 7. When this bias is compared with that found amongst 48 GATA identified *in vivo* (see Table IV), the agreement appears to be close. The only two remarkable differences are (i) a bias for a C at position −2 *in vivo* where none was

Table I. Preliminary selection data: nine different tetramers with the sequence GANN are analysed for significant abundance using a χ^2 test

Sequence	Expected	Observed	Significant to $P < 0.05$ (χ^2 test)
GATA	17	53	Yes
GATT	17	37	Yes
GATG	17	24	No
GATC	8	8	No
GACA	17	15	No
GACT	17	11	No
GACC	17	16	No
GACG	17	10	No
GAGG	17	18	No

Total number of clones = 83.
Total random sequence = 2148 bp.

Table II. Single GATA consensus sites, excluding overlaps into defined ends [total number = 19(5′) + 20(3′)]

	Position													
	−5	−4	−3	−2	−1	0	1	2	3	4	5	6	7	8
Site														
G	7	2	4	5	1	20	–	–	–	9	11	11	10	4
A	3	5	8	5	9	–	20	–	20	6	2	4	2	6
T	6	8	2	5	5	–	–	20	–	1	6	2	4	6
C	3	4	5	4	4	–	–	–	–	4	1	3	4	4
Statistical bias (χ^2 test), $P < 0.05$														
for	–	–	–	–	A	G	A	T	A	G	G	G	G	–
against	–	–	–	–	G			(all others)		T	C	–	–	–

An alignment of single GATA motif sequences from Figure 1A, excluding those sequences which extended into the defined regions of the original oligonucleotide pool, i.e. leaving 19 sequences with no overlap five bases 5′ to the core GATA and 20 sequences with no overlap five bases 3′ to the core GATA. Position numbers are assigned with the G residue of GATA as position 0. Beneath the position numbers is the number of times a particular base (shown left) occurred at this position. Below is shown a χ^2 statistical analysis indicating a significant bias for or against a particular nucleotide at that position.

found in the selected pool, and (ii) a preference for an A *in vivo* over a G in the selected pool. It should be noted that not all those GATA sites included for analysis as *in vivo* sites have been shown either to be functionally important or actually to bind GATA-1.

The biases around a core GATT motif are shown in Table III. A selection for an A residue was found at positions −5 and 4, and a G residue at positions −3 and 7. A T residue was selected against at position 4. Whether any importance should be placed on the biases found 5′ to the core motif is arguable, because although significance was shown using a χ-squared test, very few sequences were analysed at these positions and only four in seven bases were an A at position −5 and only four in seven bases a G at position −3. This compares with the significant positions 3′ of the core, where an A occurred in eight out of 12 sequences at position 4 and a G occurred in seven out of 12 sequences at position 7. When compared with three GATT motifs found in the human β-globin locus, which have been shown to bind GATA-1 (see Table III), two out of three of these sites contained an A at position 4, but no other agreement was found.

Binding analysis

When the subcloned oligonucleotides were tested in a gel retardation assay (see Materials and methods), all sequences tested containing either a GATA or a GATT motif were found to have DNA binding activity (Figure 1a–f), although with a quantitative variability. This was presumably caused by either the affects of sequences around the core motif or the presence of a second site increasing the potential target sites on any particular oligonucleotide, although these possibilities were not investigated. Of the 11 sequences tested which did not contain a GATA or a GATT motif, three showed binding activity. Those clones which showed no binding activity may have appeared in the enriched pool as a background signal or may have lost sequences important for their binding activity during the subcloning procedure, either by mutation or loss of sequence flanking the restriction sites.

At this point it was unclear what GATA-1 was recognizing in those three sequences which had binding activity yet did not have either GATA or GATT motifs. GATA-1 did not appear to be binding to GATG or GATC motifs *per se*, because these motifs were found in both binding and non-

Table III. Single GATT consensus sites, excluding overlaps into defined ends [total number = 7(5′)+ 12(3′)]

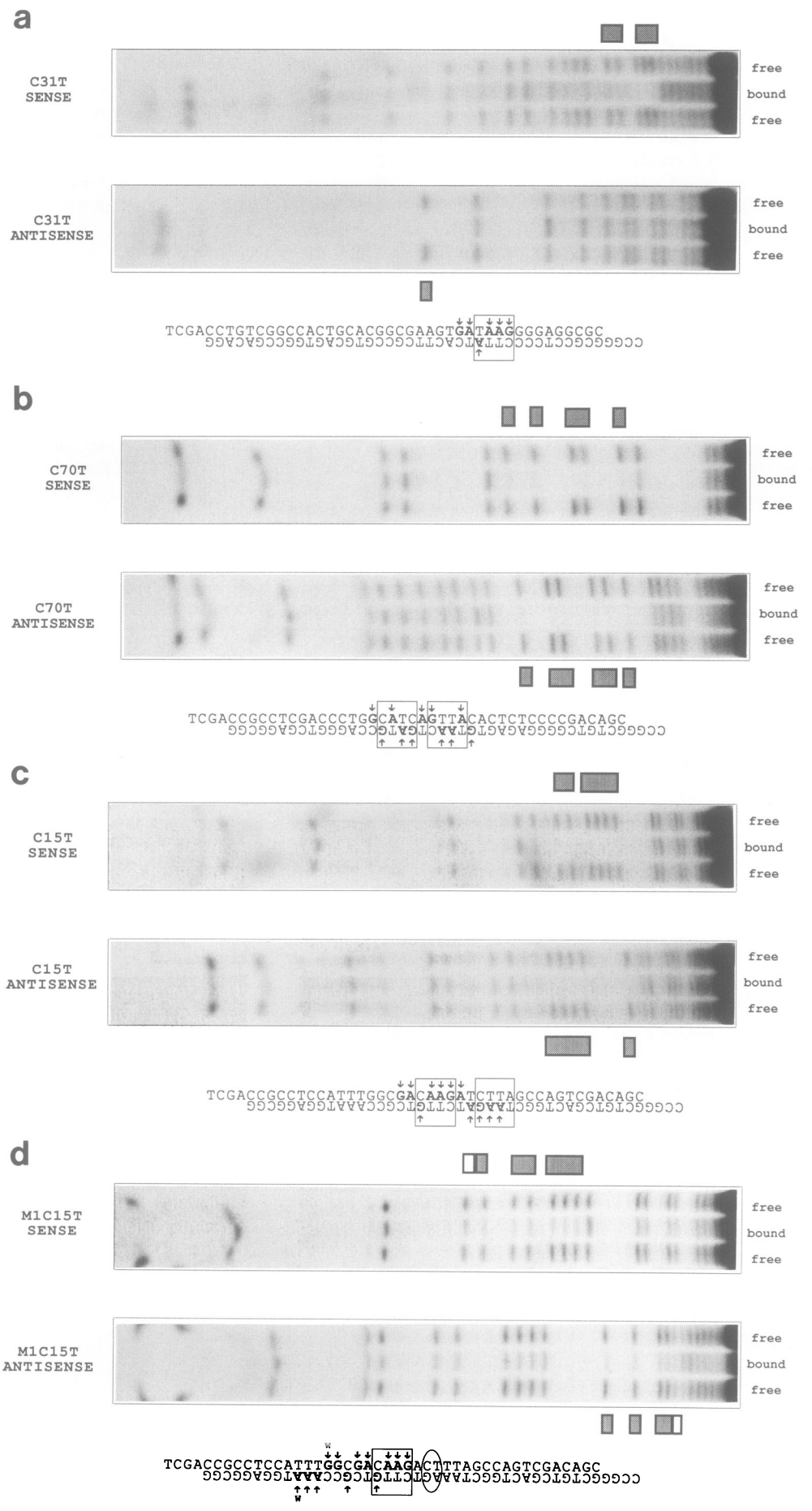
	Position													
	−5	−4	−3	−2	−1	0	1	2	3	4	5	6	7	8
Site														
G	2	3	4	2	2	12	–	–	–	3	4	3	7	4
A	4	1	1	2	3	–	12	–	12	8	1	3	3	2
T	–	2	–	1	2	–	–	12	–	–	3	2	1	2
C	1	1	2	2	–	–	–	–	–	1	4	4	1	4
Statistical bias (χ^2 test), $P < 0.05$														
for	A	–	G	–	–	G	A	T	T	A	–	–	G	–
against	(numbers too low to define)						(all others)			T	–	–	–	–

A similar analysis to that in Table II, aligning single GATT sites from Figure 1B, excluding overlaps into defined ends. Below are shown three motifs of the GATT type found in the human β-globin locus.
−71 β-globin promoter, AGTAGATTGGCCAA (deBoer *et al.*, 1988); 3′ β-globin enhancer, TTAAGATTAGCATT; β-globin intragenic enhancer, AACATGATTAGCAA (Wall *et al.*, 1988)

Table IV. GATA consensus sites found *in vivo* (total number = 48)

	Position												
	−4	−3	−2	−1	0	1	2	3	4	5	6	7	8[a]
Site													
G	11	12	9	2	48	–	–	–	16	24	18	16	14
A	15	10	11	25	–	48	–	48	27	11	12	15	8
T	10	15	3	19	–	–	48	–	4	5	10	4	12
C	1	11	25	2	–	–	–	–	1	8	8	13	13
Statistical bias (χ^2 test), $P < 0.05$													
for	–	–	C	A/T	G	A	T	A	A	G	G	–	–
against	–	–	T	G/C	–	(all others)			T/C	T	–	T	–

[a]Only 47 bases analysed at this position.
A similar analysis to that shown in Table II, aligning GATA motifs which have been identified from *in vivo* sequences. Source references are deBoer *et al.* (1988), Evans *et al.* (1988), Wall *et al.* (1988), Martin *et al.* (1989) Plumb *et al.* (1989), Frampton *et al.* (1990), Philipsen *et al.* (1990), Talbot *et al.* (1990), Chiba *et al.* (1991), Gong *et al.* (1991), Pruzina *et al.* (1991), Tsai *et al.* (1991), Fong and Emerson (1992), Rahuel *et al.* (1992) and Schwartzbauer *et al.* (1992).


a
C31T
SENSE
free
bound
free
C31T
ANTISENSE
free
bound
free
TCGACCTGTCGGCCACTGCACGGCGAAGTGATAAGGGGAGGCGC
CCGGGCCGCTCCCCTTATTCACTTCGCCGTGCAGTGGCCGACAGG
b
C70T
SENSE
free
bound
free
C70T
ANTISENSE
free
bound
free
TCGACCGCCTCGACCCTGGCATCAGTTACACTCTCCCCGACAGC
CCGGGCTGTCGGGGACGTGTAATGCCAGTGCCAGGGTCGAGGCGG
c
C15T
SENSE
free
bound
free
C15T
ANTISENSE
free
bound
free
TCGACCGCCTCCATTTGGCGACAAGATCTTAGCCAGTCGACAGC
CCGGGCTGTCGACTGGCTAAGATCTTGTCGCCAAATGGAGGCGG
d
M1C15T
SENSE
free
bound
free
M1C15T
ANTISENSE
free
bound
free
TCGACCGCCTCCATTTGGCGACAAGACTTTAGCCAGTCGACAGC
CCGGGCTGTCGACTGGCTAAACTCTTGCCCAAATGGAGGCGG

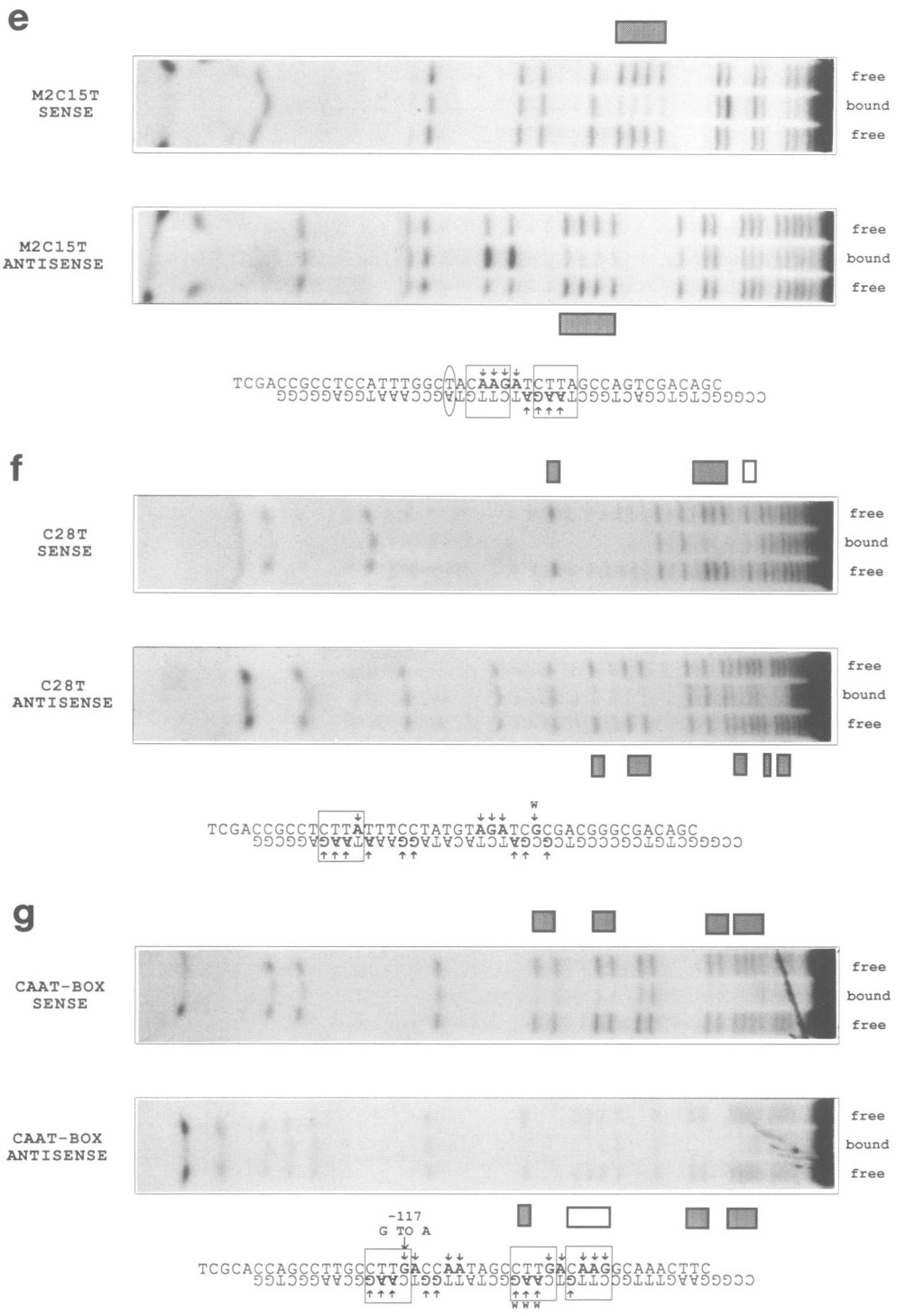


Fig. 2. Radiolabelled probes were depurinated using formic acid, bound to crude nuclear extract from MEL cells, and then electrophoresed through a 4% polyacrylamide gel. Free and bound probe was eluted, cleaved with piperidine and electrophoresed through a denaturing 15% polyacrylamide gel. The free and bound fractions are displayed. A strong footprint is indicated by a filled block and a weak footprint by an empty block. Also shown is the sequence of each probe and the footprint, indicated by both bold type and arrows (w = weak footprint); top = sense strand, bottom = antisense strand. (T/C)AAG motifs are boxed. (**a**) C31T; (**b**) C70T, the GATG and TAAC sequences are boxed; (**c**) C15T; (**d**) M1C15T, the mutated bases are circled; (**e**) M2C15T, the mutated bases are circled; (**f**) C28T; (**g**) CAAT-box sequence, derived from the human $^{A}\gamma$-globin gene promoter, the −117 mutation is indicated.

binding oligonucleotides (compare clone 70 with clones 33 and 7, compare clones 15 and 28 with clone 23 (Figure 1g).

Depurination analysis

To address the problem of GATA-1 binding to oligonucleotides containing neither GATA nor GATT motifs, a depurination interference assay was used to determine the sequences which were being recognized by GATA-1 (Wall *et al.*, 1988).

As shown in Figure 2a, the single GATA motif probe, C31T, has six purine residues important for binding activity (five on the sense and one on the antisense strand) which

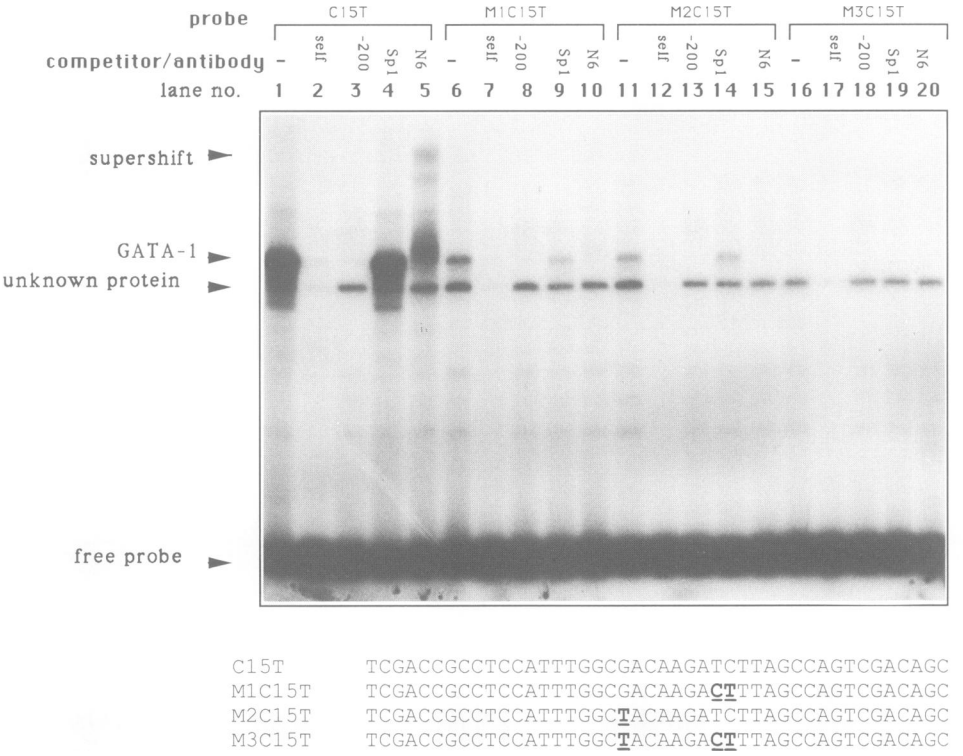
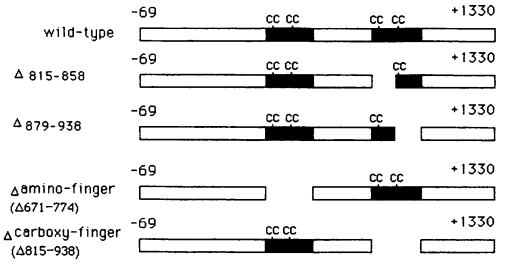


Fig. 3. Gel retardation assays (see Materials and methods). The free probe, the mobility of the bound GATA-1 complex and the mobility of the antibody–GATA-1–probe complex are indicated. Probes C15T and the point mutations of C15T, M1C15T, M2C15T and M3C15T, were radiolabelled and bound to 2 μg of crude MEL extract (lanes 1–20). A 100 times excess of competitor was included in the binding reactions in lanes 2, 7, 12 and 17 (self competition), lanes 3, 8, 13 and 18 (−200 human β-globin promoter GATA-1 binding site), and lanes 4, 9, 14 and 19 [non-GATA-1-specific competitor Sp1, derived from a known Sp1 binding site in the β-globin LCR (Philipsen *et al.*, 1993)]. Anti-GATA-1 antibody N6 was included in lanes 5, 10, 15 and 20.

correspond to the sequence GATAAG. This result is in close agreement with what was predicted to be important for binding around a GATA core (see above and Tables II and IV).

Probe C70T (Figure 2b) appears to be bound by GATA-1 over a large region containing the GATG motif, although the footprint extends 6 bp 5′ and 1 bp 3′ to this motif, encompassing a GTTA (or TAAC on the other strand) motif. This suggests that GATA-1 recognition of a core GATG motif may be more dependent on the surrounding sequences than on the recognition of a GATA motif, which would explain why not all GATG motifs will bind GATA-1. The sequence GATG is crucial for binding activity, because mutation of this sequence to GAAG abolishes activity (data not shown).

Probe C15T (Figure 2C) also shows a relatively broad footprint corresponding to the sequence GACAAGATCTT. As shown in Figure 3, C15T also binds one other higher mobility protein, of unknown identity, which footprints a region at one end of the probe away from the GATA-1 footprint (data not shown). Mutation of the GATC sequence to GACT (M1C15T) reduces, but does not abolish, binding (Figure 3), and also shifts the footprint slightly 5′ to the sequence TTTGGCGACAAG (Figure 2D). Mutation of the GACAAG sequence to TACAAG (M2C15T) also reduces, but does not abolish, binding (Figure 3), and results in the footprinted region centring over the sequence AAGATCTT (Figure 2E). These results suggest that although GATC contributes to the recognition site, it is not the crucial sequence for recognition. When the site contains both the



sequence no.	footprint	binding activity		
		w-type	Δamino	Δcarboxy
C31T	GATAAG	YES	YES	NO
C13T	GTCGATT	YES	YES	NO
C70T	GTAACTGATGC	YES	NO	NO
C15T	GACAAGATCTT	YES	NO	NO
M1C15T	TTTGGCGACAAG	YES	NO	NO
M2C15T	AAGATCTT	YES	NO	NO
C28T	G-GATCT-----GGAAATAAG	YES	NO	NO
CAAT-box	CTTGACCAA-----CTTGACAAG	YES	NO	NO

Fig. 4. The cDNA fragments and deletions used in the coupled *in vitro* transcription/translation experiment are shown (see Materials and methods). The numbers corresponding to the base pairs as designated by Tsai *et al.* (1989). CC = cysteine pair. A summary of the interaction of the wild-type, amino finger deleted and carboxy finger deleted GATA-1 with various target sites is shown in the table.

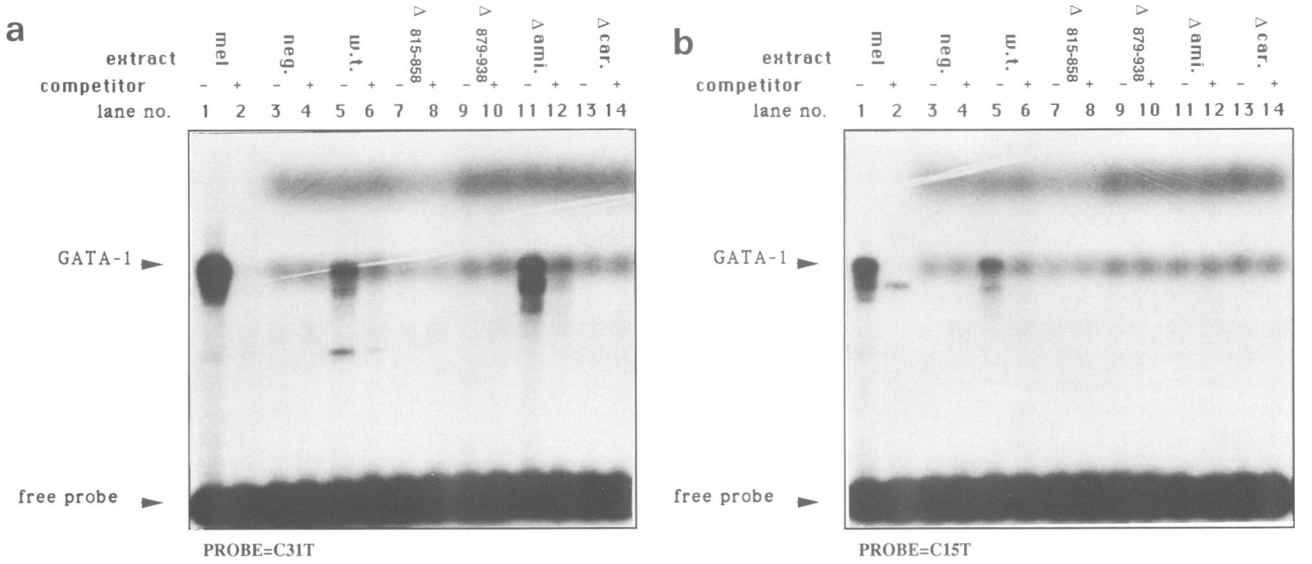


Fig. 5. Gel retardation assays using: mel = crude MEL extract; rabbit reticulocyte lysate transcribed/translated with: neg. = vector alone, w.t. = wild-type GATA-1 cDNA, Δ815−858 = GATA-1 cDNA with region containing third cysteine pair deleted (base pairs 815−858), Δ879−938 = GATA-1 cDNA with region containing fourth cysteine pair deleted (base pairs 879−938), Δamino = GATA-1 cDNA with region containing both cysteine pairs of the amino zinc finger-like domain deleted (base pairs 671−774), Δcarboxy = GATA-1 cDNA with region containing both cysteine pairs of the carboxy zinc finger-like domain deleted (base pairs 815−938). Odd numbered lanes contained no competitor in the binding reaction, even numbered lanes contained a 100× excess of competitor (−200 human β-globin GATA-1 binding site). The mobility of the free probe and bound GATA-1 is indicated. Note that the background shadow at the same position as GATA-1 in the lanes containing reticulocyte lysate is presumably a result of [^{35}S]methionine-labelling of an unrelated protein that migrates at this position in native polyacrylamide gel electrophoresis. Both experiments were performed at the same time using identical extracts, and electrophoresed on the same polyacrylamide gel. The following were radiolabelled and bound to the reagents described above: (**a**) C31T; and (**b**) C15T.

aforementioned mutations (M3C15T), binding activity is abolished (Figure 3).

Probe C28T (Figure 2F) contains a bipartite footprint, spaced in a manner similar to the CAAT-box footprint [Figure 2G; this sequence being derived from the human γ-globin promoter (with additional *SalI*/*Xma*I linkers) which also binds two other proteins CP1 and NFE6 (Berry *et al.*, 1992)], being 5 bp apart. A single point mutation of the CAAT-box probe in the footprinted region abolished binding activity (Berry *et al.*, 1992), this mutation being equivalent to the position −117 (in the human γ-globin promoter) G→A mutation which is associated with HPFH (Collins *et al.*, 1985; Gelinas *et al.*, 1985).

When sequences C15T, C28T and the CAAT-box are analysed for a consensus recognition sequence, the depurination assays raise two points. First, all three sequences contain a footprinted region which is larger than that seen on a canonical GATA probe (C31T, Figure 2a); in two cases this footprint is bipartite (CAAT-box and C28T). Second, all three sequences contain repeats of (T/C)AAG which appears to be footprinted. In almost all cases this sequence defines one end of the footprint, except in C15T where the 3′ end contains an inverted TAAG (the footprint only encompassing the AAG). Mutation of part of the footprinted region of C15T (M2C15T, Figure 2E) reveals a binding site which contains two opposed (T/C)AAGA repeats. Mutation of this part of C15T (M1C15T, Figure 2d) results in a footprint which is now defined at one end by (T/C)AAG. Interestingly, the footprint of probe C70T contains a motif similar to this, i.e. TAAC (see Figure 2b).

It has been shown that the zinc finger-like domain of GATA-1 towards the carboxy end of the protein (referred to hereafter as the carboxy finger) will interact with a GATA motif on its own (Martin and Orkin, 1990; Yang and Evans, 1992; Omichinski *et al.*, 1993). Furthermore, it has also been suggested that the zinc finger-like domain towards the amino end of the protein (amino finger) may interact with bases around the core GATA motif (Martin and Orkin, 1990; Yang and Evans, 1992). Thus, a GATAAG site could be interpreted as two overlapping recognition sequences where GATA is recognized by the carboxy finger and TAAG is recognized by the amino finger. This would imply that C28T, C15T and the CAAT-box oligonucleotides are recognized via their (T/C)AAG motifs using the amino finger domain, with the carboxy finger stabilizing the interaction via interactions with G residues 5′ to the (T/C)AAG motif. The only exception to this arrangement would be site M2C15T, which is a weak binding site. In this case, the carboxy finger may be stabilizing via an interaction with a C or a T, which of course cannot be detected by the depurination assay. In the case of probe C70T, the large footprint on this oligonuceotide may suggest that its binding site is related to that of type (T/C)AAG. The motif GATG, unlike GATA or GATT, did not appear to always bind GATA-1, suggesting a role for flanking sequences. Thus, a sub-optimal carboxy finger recognition site, such as GATG, may be compensated for by a flanking sub-optimal amino finger recognition site, such as TAAC.

Interaction with GATA-1 DNA binding domains

To test if the zinc finger-like DNA binding domains of GATA-1 indeed have differential recognition specificities, various deletions were made in the murine GATA-1 cDNA and the resulting templates *in vitro* translated (see Materials and methods) using rabbit reticulocyte lysate (Figure 4). The resulting lysates were then tested for DNA binding activity.

As shown in Figure 5a, deletion of the amino finger had

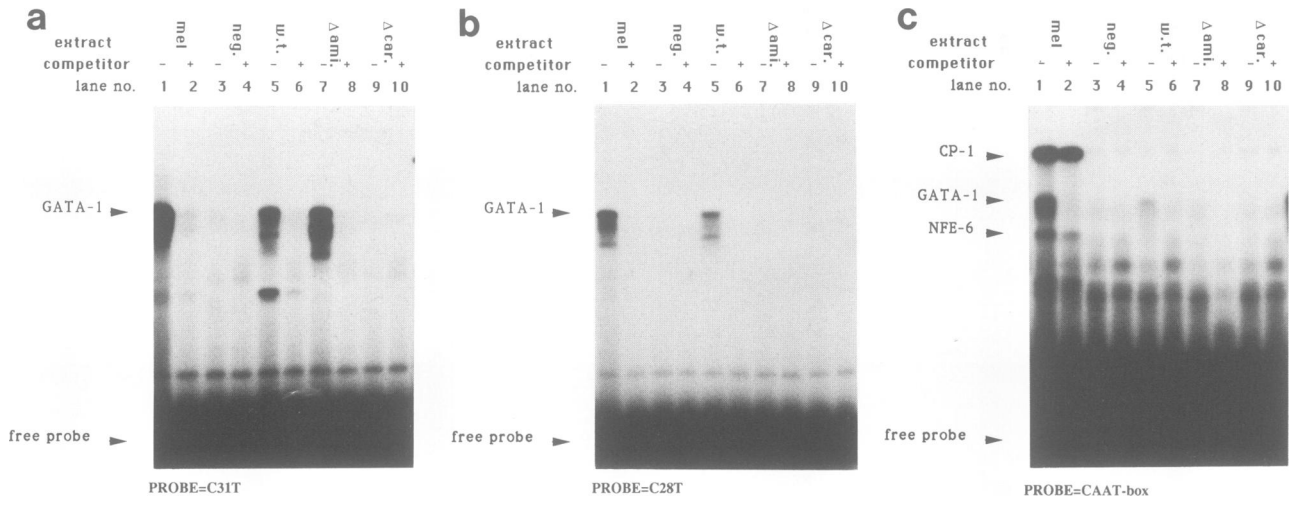


Fig. 6. Gel retardation assays identical to those described in Figure 5, although not including samples Δ815–858 or Δ879–938. All three experiments were performed at the same time using identical extracts, and electrophoresed on the same polyacrylamide gel. The mobilities of the free probe and bound GATA-1 are indicated. The following were radiolabelled and bound to reagents as described above: (**a**) C31T; (**b**) C28T; and (**c**) CAAT-box. Note in (c) the presence of two other binding proteins in crude MEL extract CP1 and NFE6 (Berry *et al.*, 1992). Longer exposures of this panel to X-ray autoradiography did not meaningfully alter this image.

no effect on the DNA binding activity using probe C31T. Deletion of the carboxy finger or either of the regions containing the cysteine pairs in this region abolished binding activity to C31T. An identical result was achieved using probe C13T [a GATT-type motif, confirmed by depurination analysis (data not shown)]. When C15T was used as the probe, deletions in (or of) either finger-like domain abolished binding (Figure 5b).

Using C31T as a control probe again, it can be seen in Figure 6 that C28T and the CAAT-box probes are not bound by GATA-1 if either of the zinc finger-like domains are absent. Probes containing the GATG recognition site (C70T) will not bind the amino finger deletion (data not shown), giving a result similar to that obtained by using probe C15T. The results are summarized in Figure 4. It should be noted that the *in vitro* translated wild-type protein bound relatively weakly to the CAAT-box probe, raising the possibility that a secondary modification, which is important for GATA-1 binding, is occurring inefficiently in the reticulocyte lysate. Preliminary data show that phosphorylation of GATA-1 is important for its binding activity, suggesting that such a modification may play a specific role in binding to the CAAT-box probe.

Discussion

A binding site enrichment protocol (Pollock and Treisman, 1990) has been used to define a consensus for GATA-1 DNA binding site recognition. The optimal site appears as AGATAGGGG centred around a core GATA motif. The sequence bias found in GATA core motifs and described in vertebrate genomes is C(A/T)GATAAGG. These two sequences are in close agreement in that the sequence bias appears to be for purines 3′ of the GAT core (on the same strand), of the form AAGG or AGGG. This extends on the previous canonical consensus site WGATAR (Yamamoto *et al.*, 1990) and is similar to the consensus proposed by Plumb *et al.* (1989), i.e. GATAAG.

Two other GAT motifs have been defined which bind GATA-1, i.e. GATT and a GATG sequence. Both these sites have been found in the human β-globin locus and appear to bind GATA-1 (deBoer *et al.*, 1988; Wall *et al.*, 1988; Philipsen *et al.*, 1990; Talbot *et al.*, 1990). The GAT(A/T) core motifs examined are capable of binding GATA-1 independent of the amino finger of GATA-1. It has been shown previously that the carboxy finger of GATA-1 alone will direct sequence-specific DNA binding to a GATA motif (Martin and Orkin, 1990; Schwartzbauer *et al.*, 1992; Yang and Evans, 1992; Omichinski *et al.*, 1993), and that the role of the amino finger in this context may be in site discrimination rather than it having a binding activity of its own. The amino finger will affect the stability of binding by GATA-1 to GATA motifs, but to different extents depending on the flanking sequences (Martin and Orkin, 1990; Yang and Evans, 1992). The stability of interaction supplied by the amino finger appears crucial in C70T, which contains a sub-optimal site, i.e. GATG. In this case the amino finger may be recognizing a flanking TAAC sequence (see below).

Another group of non-canonical GATA-1 binding motifs has also been defined, with the core consensus (T/C)AAG. In this case the amino finger of GATA-1 is critical for binding activity. The carboxy finger is still required, possibly to stabilize interactions via upstream residues. These results suggest two different binding specificities for the two GATA-1 zinc finger-like domains, with the carboxy finger recognizing GAT(A/T) and the amino finger recognizing (T/C)AAG. Differential specificities have previously been well characterized in zinc finger-like domains of a number of proteins, such as Krox-20, Sp1 and Zif268 (Nardelli *et al.*, 1991, 1992; Pavletich and Pabo, 1991; Desjarlais and Berg, 1993). The (T/C)AAG motif is also found as the 3′ half of strong binding sites such as GATAAG, suggesting that both finger domains may have a role in such sites, although such a close juxtaposition may not allow both finger domains to interact with the DNA simultaneously. Footprinting analysis of GATA-1 lacking an amino finger bound to such a site supports this view, in that the footprint only changes from GATAAG (wild-type) to GATAA (no amino finger) (data not shown). This compares with the protein Sp1, where three zinc finger-like domains of the type

Cys_2His_2 define a 9 bp site with each finger interacting with three bases (Desjarlais and Berg, 1992). GATA-1 binding motifs may also be formed by the juxtaposition of two sub-optimal motifs, for example the site in C70T where the carboxy domain may be recognizing a GATG site and the amino finger may be recognizing a TAAC site, to give a single binding motif which is dependent on both finger domains.

When this work was completed, two papers were published which support the interpretation of the data presented here, both of which used random oligonucleotide selection procedures (Ko and Engel, 1993; Merika and Orkin, 1993). Both groups confirmed that GATA-1 selected for GATA and GATT motifs, as would be expected from previous analysis of vertebrate gene loci. Ko and Engel (1993) suggested two extended motifs, the first being GATAAG which is recognized by GATA-1, −2 and −3. The second motif, GATCTTA, was only recognized by GATA-2 and -3, and not GATA-1; however, they found one motif which was selected for by GATA-1 of the form GATCAAG. In the latter case, Ko and Engel (1993) suggested that GATC is a core recognition motif for two of the GATA proteins (GATA-2 and -3), although they have not confirmed this by footprinting analysis. GATA-1 did not appear to be strongly selecting for the (T/C)AAG motif, yet because they based their analysis on varying 7 bp around an invariant GAT core, they were not able to form these sites as the repeat motif which appears in the recognition sites presented in this paper. Both GATA-2 and -3 had a strong preference for a (T/C)AAG motif, particularly in the absence of a GAT(A/T/G) motif, suggesting that both proteins may share with GATA-1 a bipartite DNA binding recognition motif. Thus, the conclusion by Ko and Engel (1993) that GATC is recognized like GATA or GATT may have to be reinterpreted. The alternative explanation of their data would suggest that the GATC motif is not recognized by GATA-2 and -3, but a selection for (T/C)AAG could account for its presence. Because of the structure of their random oligonucleotides, this motif could be formed in such a way as to make the last C of the GATC motif the opposite base to the last G of the (T/C)AAG motif, thus yielding the apparent motif of GATCTTA. Merika and Orkin (1993) found that GATA-1 would also recognize a GATG or a GTTA (TAAC on the other strand) motif in some contexts. Unfortunately, only one flanking base either side of each motif was shown when tested for binding activity, and no footprinting analysis was attempted on these sites. Thus, neither the contribution of flanking sequences nor the juxtaposition of GATG and TAAC motifs could be shown from the data presented. However, it is worth noting that some GATG and GTTA motifs were found to bind to GATA-1 only in the presence of both finger domains, consistent with the data discussed here.

Two different classes of interaction of GATA-1 with a target motif have been defined. The first occurs with target sequences containing GAT(A/T) motifs and is not dependent on the amino zinc finger-like domain. The second occurs with sequences containing a core (T/C)AAG motif and is dependent on the amino zinc finger-like domain. An intermediate type of site, i.e. C70T which contains two sub-optimal sites (GATG and TAAC), appears to require both finger-like domains to bind. This suggests a model, as shown in Figure 7, where the two fingers have overlapping DNA

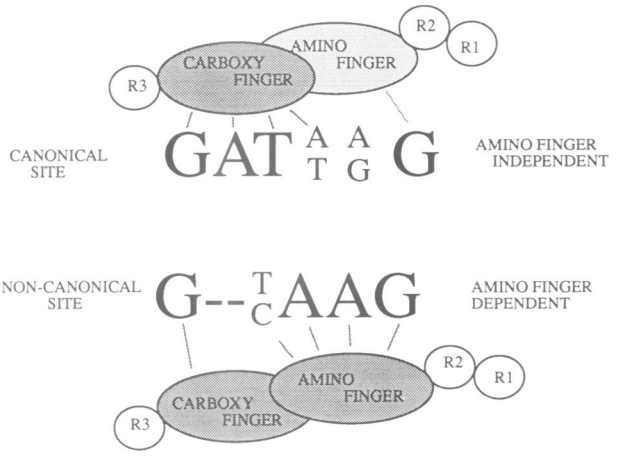


Fig. 7. Model suggested by gel retardation of mutant GATA-1 proteins. The possible interaction of GATA-1 with a canonical GATA or GATT motif which does not require the amino finger domain is displayed. Also shown is the possible interaction of GATA-1 with a non-canonical motif which lacks a carboxy finger recognition sequence, but instead uses the motif (T/C)AAG and the amino finger to direct binding. The carboxy finger is still needed to stabilize binding in this case. The 'R' circles represent the repeat domains of the protein as defined by Trainor *et al.* (1990).

recognition sequences. The sequences containing GATA or GATT define a motif which is recognized by GATA-1 lacking an amino finger, suggesting that these nucleotides define the sequence specificity of the carboxy finger. For any other sequence to bind a GATA-1 protein, both the motif (T/C)AAG and an intact amino finger are necessary, yet not sufficient. An exception to this would be site C70T which appears to contain both a poor carboxy finger recognition site (GATG) and a poor amino finger recognition site (TAAC). These two motifs together may synergize to give a working recognition site. The carboxy finger alone will not bind to (T/C)AAG sequences, but appears to facilitate stable interaction, possibly via interaction with residues which appear to be footprinted 5′ to the core (T/C)AAG motif (see Figure 2).

These two forms of GATA-1−DNA interaction suggest a biochemical explanation of the different functions for GATA-1 *in vivo*. Binding of GATA-1 to sites of the type GAT(A/T) has been correlated to transcriptional activation in the β-globin locus (deBoer *et al.*, 1988; Philipsen *et al.*, 1993), the GATA-1 gene itself (Nicolis *et al.*, 1991; Tsai *et al.*, 1991; Simon *et al.*, 1992), the porphobilinogen deaminase gene (Mignotte *et al.*, 1989a,b), the erythropoietin receptor gene (Chiba *et al.*, 1991), and indirect activational activity via the displacement of a repressor in the glycophorin B promoter (Rahuel *et al.*, 1992). It has been shown that GATA-1 recognition of this type of motif does not require the amino finger domain (Martin and Orkin, 1990; Yang and Evans, 1992). It has also been shown that a GATA-1 protein lacking the amino finger domain will act as a transactivator (Martin and Orkin, 1990), although there is some evidence for the amino finger domain containing a transactivation domain (Yang and Evans, 1992).

In the case of (T/C)AAG binding motifs, only one has been found *in vivo* (Berry *et al.*, 1992), and this appears to define a developmentally specific repressor activity. In this case, GATA-1 binds in the human γ-globin CAAT-box to a non-canonical site containing (T/C)AAG repeats. A

single G→T point mutation at position −117 in the γ-globin promoter results in the γ-globin gene being incorrectly expressed in adult transgenic mice carrying β-globin locus transgenes (Berry *et al.*, 1992). This mutation is found in the human population where it is associated with a condition known as HPFH (Collins *et al.*, 1985; Gelinas *et al.*, 1985). Of the proteins known to bind in this region, only GATA-1 was significantly altered in its binding activity. Berry *et al* (1992) showed that GATA-1 binding was abolished by this mutation. It can therefore be suggested that interaction of the amino finger of GATA-1 with a non-canonical recognition sequence, i.e. (T/C)AAG, may result in a DNA−protein structure which precludes the activational potential of domains such as the amino finger of the GATA-1 protein, inducing repression of subject genes. Alternatively, employing a different domain to direct DNA binding may present a different domain of the protein to potential transcriptional co-factors. Either possibility may be the case in the γ-globin CAAT-box.

In two of the three non-canonical GATA-1 binding sites defined here, the (T/C)AAG appeared as a repetitive motif, suggesting that a single motif will not support detectable binding and that increasing target sites may enhance binding activity many fold. The multimerization of this motif in the γ-globin CAAT-box may be functionally important, because it has been shown that multiple GATA-1 binding sites are required for GATA-1 to transactivate reporter genes in erythroid cells, though not in heterologous cells (Evans and Felsenfeld, 1991). Whether GATA-1 molecules will bind DNA cooperatively is not known.

In conclusion, the canonical recognition site of GATA-1 has been extended to AGATAGGG, it has been shown that GATA-1 will also bind to GATT and GATG sites, and a new subclass of GATA-1 binding sites have been defined with a core consensus of (T/C)AAG which requires both zinc finger-like domains of the GATA-1 protein to stably interact. Furthermore, it is suggested that the interaction of GATA-1 with (T/C)AAG sites may be a functionally distinct event, possibly involved in transcriptional repression.

Materials and methods

Preparation of nuclear extracts

Crude nuclear extracts were prepared from mouse erythroleukaemia (MEL) cells as described previously (deBoer *et al.*, 1988) which is a modification of the method of Gorski *et al.* (1986). The final $(NH_4)_2SO_4$ pellets were redissolved in 1 ml of buffer D (20 mM HEPES, pH 7.9, 20% glycerol, 0.1 M KCl, 0.2 mM dithiothreitol, 0.5 mM EDTA, 0.5 mM phenylmethylsulfonyl-fluoride) per 3 g of starting material. Extracts contained 10−30 mg proteins per ml and were stored frozen under liquid nitrogen.

Binding site enrichment protocol

Enrichment for binding sites from a random oligonucleotide pool was essentially done as previously described (Pollock and Treisman, 1990). Briefly, three oligonucleotides were used: R76, 5′ CAGGTCAGTTCAGCGGATCCTGTCG(G/A/T/C)$_{26}$GAGGCGAATTCAGTGCAACTGCAGC 3′; primer F, 5′ GCTGCAGTTGCACTGAATTCGCCTC 3′; primer R, 5′ CAGGTCAGTTCAGCGGATCCTGTCG 3′.

The random sequence oligonucleotide R76 was rendered double stranded by primed synthesis using primer F and Klenow fragment, radiolabelled during the elongation reaction by the inclusion of ^{32}P-radiolabelled dCTP. Binding reactions (10 μl) contained 2 μl of crude MEL nuclear extract (diluted to 1 μg/μl in buffer D), 1 μl of double stranded oligonucleotide probe (1 ng of R76 or subsequent amplification products), 2 μg of poly (dI):poly (dC), 1 μl of 10 × binding buffer [50 mM Tris−Cl, pH 8.0, 5 mM dithiothreitol, 5 mM EDTA, 250 mM NaCl and 10% Ficoll (Pharmacia)], 1 μl of rat anti-mouse GATA-1 IgG mAb N6 (a 1:4 dilution of the growth medium of a confluent culture of the myeloma cell line producing this antibody, source: J.D.Engel, Evanston) and 1 μl of mouse anti-rat IgG2a mAb (1:10 dilution of ascites fluid, Sigma product no. R-0761).The binding reaction was incubated at room temperature for 15 min, then added to 20 μl of protein A−Sepharose beads (Pharmacia) and incubated overnight at 4°C. The beads were then washed three times with 1 × binding buffer, and the bound oligonucleotides eluted, purified and then amplified by PCR using primers R and F as previously described (Pollock and Treisman, 1990). Amplified product (1 ng) was then put back into a binding reaction and the procedure repeated. A 0.1 ng aliquot of the product of each round of enrichment was tested in a gel retardation assay for binding activity (see below). After three rounds of enrichment a binding activity was detectable, and the enriched pool was digested with *Bam*HI and *Eco*RI and subcloned into Bluescript pKS+. Clones were then sequenced using standard KS and SK primers (Stratagene).

Gel retardation analysis

Probes were prepared from selected clones by digesting Bluescript KS+ subclone plasmids with *Bam*HI and *Eco*RI and filling in using Klenow fragment and ^{32}P-radiolabelled dGTP and dATP. Probe fragments were then purified from 8% polyacrylamide gels.

Probes were also prepared with *SalI*−*XmaI* linkers, designated CT (clone number as referred to in text) except the CAAT-box probe, and were synthesized as single stranded oligonucleotides. They were end labelled with polynucleotide kinase and ^{32}P-radiolabelled γATP, and then annealed as previously described (deBoer *et al.*, 1988).

The sequences of probes not referred to in the text (sense strand only shown) are: −200, CGAGGCCAAGAGATATATCTTAGAGGGAGT (deBoer *et al.*, 1988); and Sp1, AAATAGTCCCGCCCCTAACTCCGCCCAT (Philipsen *et al.*, 1993).

Gel retardation assays were performed as previously described (deBoer *et al.*, 1988). Briefly, a 10 μl reaction was set up as follows: 0.1 ng of probe was added to 1 μl of 10 × binding buffer, 2 μg of poly (dI):poly (dC) and 2 μl of extract [either 2 μg crude MEL nuclear extract or rabbit reticulocyte lysate (see below)]. Also added, as indicated in Results, was a 100 × excess of cold competitor DNA or 1 μl of antibody N6 (prepared as above). Reactions were then incubated at room temperature for 20−30 min. After addition of 1/10th volume of 20% Ficoll containing 0.05% Xylene blue and 0.05% bromophenol blue, the samples were run on a 4% polyacrylamide gel. The gel was dried, then exposed to X-ray film.

Depurination analysis

Depurination analysis was essentially performed as described by Wall *et al.* (1988). Briefly, radiolabelled oligonucleotide probes were depurinated with formic acid as described by Maxam and Gilbert (1980), then bound to crude MEL extract in a scaled-up reaction (10×) as described above. Free and bound fractions of the probe were separated on a 4% polyacrylamide gel, eluted and purified. After cleavage with piperidine (Maxam and Gilbert, 1980), the samples were separated on 15% polyacrylamide sequencing gels.

In vitro translation of GATA-1 cDNAs

The GATA-1 cDNA constructs were obtained from L.Wall (Montreal). Each mutant was constructed by PCR and subcloned into the *Xba*I and *Eco*RI sites in the Bluescript pKS+ polylinker (Stratagene). Each plasmid (3 μg) was linearized with Asp718 and then transcribed and translated using the TNT™ T7 coupled reticulocyte lysate system (Promega) in the presence of radiolabelled [^{35}S]methionine. Resulting lysates were then run on SDS−polyacrylamide gels and the protein yields quantitated by phosphoimaging. Equivalent amounts of each protein were then used in the gel retardation analysis (see above).

Acknowledgements

We are grateful to Lee Wall (Montreal) for sharing the GATA-1 mutant constructs, Doug Engel (Evanston) for making the GATA-1 antibody available, Meera Berry and Sjaak Philipsen for help and advice and Colin Young for growing large batches of cells. D.W. was partly supported by ICI (UK). The work was supported by ICI (UK) and the MRC (UK).

References

Arceci,R.J., King,A.A.J., Simon,M.C., Orkin,S.H. and Wilson,D.B. (1993) *Mol. Cell. Biol.*, **13**, 2235−2246.

Berry,M., Dillon,N. and Grosveld,F. (1992) *Nature*, **358**, 499−502.

Brady,H.J.M., Sowden,J.C., Edwards,M., Lowe,N. and Butterworth,P.H.W. (1989) *FEBS Lett.*, **257**, 451−456.

Chiba,T., Ikawa,Y. and Todokoro,K. (1991) *Nucleic Acids Res.*, **19**, 3843–3848.
Collins,F.S., Metherall,J.E., Yamakawa,M., Pan,J., Weissman,S.M. and Forget,B.G. (1985) *Nature*, **313**, 325–326.
Cunningham,T.S. and Cooper,T.G. (1991) *Mol. Cell. Biol.*, **11**, 6205–6215.
deBoer,E., Antoniou,M., Mignotte,V., Wall,L. and Grosveld,F. (1988) *EMBO J.*, **7**, 4203–4212.
Desjarlais,J.R. and Berg,J.M. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 7345–7369.
Desjarlais,J.R. and Berg,J.M. (1993) *Proc. Natl Acad. Sci. USA*, **90**, 2256–2260.
Dorfman,D.M., Wilson,D.B., Bruns,G.A.P. and Orkin,S.H. (1992) *J. Biol. Chem.*, **267**, 1279–1285.
Evans,T. and Felsenfeld,G. (1989) *Cell*, **58**, 877–885.
Evans,T. and Felsenfeld,G. (1991) *Mol. Cell. Biol.*, **11**, 843–853.
Evans,T., Reitman,M. and Felsenfeld,G. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 5976–5980.
Fong,T.C. and Emerson,B.M. (1992) *Genes Dev.*, **6**, 521–532.
Fu,Y.-H. and Marzluf,G.A. (1990) *Mol. Cell. Biol.*, **10**, 1056–1065.
Frampton,J., Walker,M., Plumb,M. and Harrison,P.R. (1990) *Mol. Cell. Biol.*, **10**, 3838–3842.
Gelinas,R., Endlich,B., Pfeiffer,C., Yagi,M. and Stamatoyannopoulos,G. (1985) *Nature*, **313**, 323–325.
Gong,Q., Stern,J. and Dean,A. (1991) *Mol. Cell. Biol.*, **11**, 2558–2566.
Gorski,K., Carneiro,M. and Schibler,U. (1986) *Cell*, **47**, 767–776.
Ito,E., Toki,T., Ishihara,H., Ohtani,H., Gu,L., Yokoyama,M., Engel,J.D. and Yamamoto,M. (1993) *Nature*, **362**, 466–468.
Ko,L.J. and Engel,J.D. (1993) *Mol. Cell. Biol.*, **13**, 4010–4022.
Ko,L.J., Yamamoto,M., Leonard,M.W., George,K.M., Ting,P. and Engel,J.D. (1991) *Mol. Cell. Biol.*, **11**, 2778–2784.
Kudla,B., Caddick,M.X., Langdon,T., Martinez-Rossi,N.M., Bennett,C.F., Sibley,S., Davies,R.W. and Arnst,H. (1990) *EMBO J.*, **9**, 1355–1364.
Lee,M., Temizer,D.H., Clifford,J.A. and Quertermous,T. (1991) *J. Biol. Chem.*, **266**, 16188–16192.
Martin,D.I.K. and Orkin,S.H. (1990) *Genes Dev.*, **4**, 1886–1898.
Martin,D.I.K., Tsai,S.F. and Orkin,S.H. (1989) *Nature*, **338**, 435–438.
Martin,D.I.K., Zon,L.I., Mutter,G. and Orkin,S.H. (1990) *Nature*, **344**, 444–447.
Maxam,A. and Gilbert,W. (1980) *Methods Enzymol.*, **65**, 499–560.
Merika,M. and Orkin,S.H. (1993) *Mol. Cell. Biol.*, **13**, 3999–4010.
Mignotte,V., Eleouet,J.F., Raich,N. and Romeo,P.H. (1989a) *Proc. Natl Acad. Sci. USA*, **86**, 6548–6552.
Mignotte,V., Wall,L., deBoer,E., Grosveld,F. and Romeo,P.H. (1989b) *Nucleic Acids Res.*, **17**, 37–54.
Nardelli,J., Gibson,T.J., Vesque,C. and Charnay,P. (1991) *Nature*, **349**, 175–178.
Nardelli,J., Gibson,T. and Charnay,P. (1992) *Nucleic Acids Res.*, **20**, 4137–4144.
Nicolis,S., Bertini,C., Ronchi,A., Crotta,S., Lanfranco,L., Moroni,E., Giglioni,B. and Ottolenghi,S. (1991) *Nucleic Acids Res.*, **19**, 5285–5291.
Omichinski,J.G., Trainor,C., Evans,T., Gronenborn,A.M., Clore,G.M. and Felsenfeld,G. (1993) *Proc. Natl Acad. Sci. USA*, **90**, 1676–1680.
Pavletich,N.P. and Pabo,C.O. (1991) *Science*, **252**, 809–817.
Pevny,L., Simon,M.C., Robertson,E., Klein,W.H., Tsai,S.F., D'Agati,V., Orkin,S.H. and Costantini,F. (1991) *Nature*, **349**, 257–260.
Philipsen,S., Talbot,D. Fraser,P. and Grosveld,F. (1990) *EMBO J.*, **9**, 2159–2167.
Philipsen,S., Pruzina,S. and Grosveld,F. (1993) *EMBO J.*, **12**, 1077–1085.
Plumb,M., Frampton,J., Wainwright,H., Walker,M., Macleod,K., Goodwin,G. and Harrison,P. (1989) *Nucleic Acids Res.*, **17**, 73–92.
Pollock,R. and Treisman,R. (1990) *Nucleic Acids Res.*, **18**, 6197–6204.
Pruzina,S., Hanscombe,O., Whyatt,D., Grosveld,F. and Philipsen,S. (1991) *Nucleic Acids Res.*, **19**, 1413–1419.
Rahuel,C., Vinit,M., Lemarchandel,V., Cartron,J. and Romeo,P. (1992) *EMBO J.*, **11**, 4095–4102.
Romeo,P.H., Prandini,M.H., Joulin,V., Mignotte,V., Prenant,M., Vainchenker,W., Marguerie,G. and Uzan,G. (1990) *Nature*, **344**, 447–449.
Schwartzbauer,G., Schlesinger,K. and Evans,T. (1992) *Nucleic Acids Res.*, **20**, 4429–4436.
Simon,M.S., Pevny,L., Wiles,M.V., Keller,G., Constantini,F. and Orkin,S.H. (1992) *Nature Genetics*, **1**, 92–98.
Spieth,J., Shim,Y.H., Lea,K., Conrad,R. and Blumenthal,T. (1991) *Mol. Cell. Biol.*, **11**, 4651–4659.
Talbot,D., Philipsen,S., Fraser,P. and Grosveld,F. (1990) *EMBO J.*, **9**, 2169–2177.
Trainor,C.D., Evans,T., Felsenfeld,G. and Boguski,M.S. (1990) *Nature*, **343**, 92–96.
Tsai,S., Martin,D.I.K., Zon,L.I., D'Andrea,A.D., Wong,G. and Orkin,S.H. (1989) *Nature*, **339**, 446–451.
Tsai,S., Strauss,E. and Orkin,S.H. (1991) *Genes Dev.*, **5**, 919–931.
Wall,L., deBoer,E. and Grosveld,F. (1988) *Genes Dev.*, **2**, 1089–1100.
Yamamoto,M., Ko,L.J., Leonard,M.W., Beug,H., Orkin,S.H. and Engel,J.D. (1990) *Genes Dev.*, **4**, 1650–1662.
Yang,H. and Evans,T. (1992) *Mol. Cell. Biol.*, **12**, 4562–4570.
Zon,L.I., Mather,C., Burgess,S., Bolce,M.E., Harland,R.M. and Orkin,S.H. (1991) *Proc. Natl Acad. Sci. USA*, **88**, 10642–10646.

Received on July 15, 1993; revised on July 29, 1993