An efficient method for isolating putative promoters and 5'-transcribed sequences from large genomic clones.

D P Mortlock, M R Nelson and J W Innis

Genome Res. 1996 6: 327-335
Access the most recent version at doi:10.1101/gr.6.4.327

GENOME METHODS

An Efficient Method for Isolating Putative Promoters and 5′-transcribed Sequences from Large Genomic Clones

Douglas P. Mortlock,[1] Matthew R. Nelson,[1] and Jeffrey W. Innis[1–3]

[1]Departments of Human Genetics and [2]Pediatrics, University of Michigan, Medical Science II, M4708, Ann Arbor, Michigan 48109-0618

Efficient strategies to isolate promoters and flanking exons from large genomic clones would facilitate the assembly of transcription units, complement existing techniques to isolate expressed sequences, and provide 5′ regulatory elements. We have developed a rapid and simple method to isolate promoters from large mammalian genomic DNA clones by exploiting the abundance of binding sites for the ubiquitous transcription factor Sp1 in gene promoters. Using this method, putative promoter sequences with Sp1-binding sites are enriched ~100-fold from fragmented P1 clone DNA. Based on the abundance of Sp1-binding motifs in promoters, we predict that a significant subset of vertebrate promoters could be isolated by this method.

Contigs of genomic clones are being assembled for large segments of human, mouse, and other vertebrate genomes, and such clones often become essential reagents for the study of a wide variety of biological processes, as well as the identification of mutations that cause disease in humans or in animal models. Both regulatory and coding sequence information are needed to investigate gene function fully.

The need to identify coding sequences within genomic clones led to the development of methods such as cDNA selection (Lovett et al. 1991; Korn et al. 1992; Tagle et al. 1993) and exon trapping (Buckler et al. 1991). To generate the most complete transcription maps from genomic regions, the use of several complementary approaches is beneficial, as no single technique is able to isolate all gene-coding sequences. Few techniques are capable of isolating 5′ ends of transcription units, or gene regulatory elements. Strategies for isolating gene promoters will greatly increase the ability to identify cDNAs that contain complete 5′ coding regions and would accelerate studies of function and transcriptional regulation of individual genes.

Transcription factors have been used to extract fragments selectively from complex mixtures of DNA, to identify consensus binding sequences (Gronostajski et al. 1984; Kinzler and Vogelstein 1989, 1990; Lelong et al. 1989; Buckler et al. 1991). The abundance of transcription factor binding sites in human DNA has been shown to be greater in promoter than in nonpromoter sequences (Bucher 1990; Prestridge and Burks 1993). Therefore, we sought to develop a method (which we term promoter capture) that exploits this difference, by using transcription factors to extract promoters from large genomic clones.

The mammalian transcription factor Sp1 (Briggs et al. 1986) binds to motifs containing the GC box (the DNA motif GGGCGG), although high-affinity binding of Sp1 to DNA requires specific bases outside the GC box, extending its site preference to ~10 bases in vitro (Kadonaga et al. 1986; Thiesen and Bach 1990). Because this motif is GC-rich, close matches to the consensus should be rare except in promoters and CpG islands. It has been shown that the GC box is common in vertebrate gene promoters and in CpG islands in general (Kadonaga et al. 1986; Gardiner-Garden and Frommer 1987; Thiesen and Bach 1990), but close matches to an optimized Sp1-binding site consensus cluster ~50–200 bp 5′ of transcription start sites (Bucher 1990). GC boxes are present in CpG islands associated with both tissue-specific and housekeeping gene promoters (Gardiner-Garden and Frommer 1987). Sp1 is highly conserved between human and mouse, is expressed ubiquitously (Saffer et al. 1991), and regulates a wide variety of genes.

[3]Corresponding author.
E-MAIL innis@umich.edu; FAX (313) 763-3784.

6:327–335 ©1996 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/96 $5.00

For these reasons, we developed a promoter isolation strategy using Sp1, and applied it to four murine and one human phage P1 clones.

RESULTS

Development of GST–Sp1 Selection Conditions

We initially explored an Sp1-selection methodology based on an antibody "sandwich" approach using commercially available purified human Sp1, anti-Sp1 polyclonal antibodies, and protein A–Sepharose as an attachment matrix for the antibody. This method demonstrated the ability of Sp1 to selectively bind DNA restriction fragments with known Sp1-binding sites but not vector fragments. To increase efficiency, we subsequently developed a selection scheme using a glutathione *S*-transferase (GST)–Sp1 fusion protein, which allowed the protein to be conjugated directly to glutathione–agarose beads (Fig. 1). This method proved more reliable and yielded greater recovery of bound DNA. All selection experiments described below used this GST–Sp1 selection protocol.

The GST–Sp1 fusion protein contains the zinc finger DNA-binding and carboxy-terminal domains of human Sp1 fused to GST (Perkins et al. 1994). To optimize selection with GST–Sp1, we performed DNA-binding experiments using end-labeled restriction fragments from a plasmid, p0.7, which contains a fragment of the murine ADA gene promoter inserted into the polylinker of pTZ19R. This sequence includes a GC-rich region upstream of the transcription start site that contains four GC box motifs and complexes strongly with human Sp1 in footprinting experiments (Innis et al. 1991). A 300-mM KCl wash was optimal for reducing nonspecific binding while maximizing the relative recovery of the 818-bp ADA promoter fragment (Fig. 2A). These results are consistent with the KCl elution profile of native Sp1 observed in DNA affinity chromatography, where Sp1 elutes from concatamerized high-affinity binding sites at just over 300 mM KCl (Briggs et al. 1986). To test the DNA-binding capacity of a typical GST–Sp1 protein preparation, we performed selections as in Figure 2A, but with increasing amounts of input DNA (Fig. 2B). As the total input DNA increased, the recovered labeled DNA decreased only slightly, indicating that the DNA-binding capacity of the GST–Sp1 was not exhausted even with 10 μg of input DNA. From quantitating the amount of recovered labeled DNA, we concluded that ~150–200 ng of the 818-bp fragment was recovered from a 10-μg selection experiment. Therefore, selection was effective over a wide range of input DNA quantities.

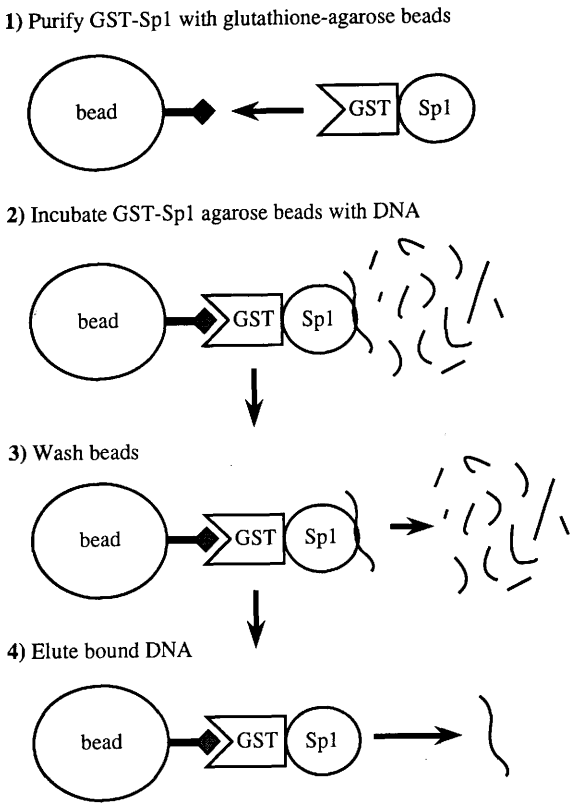

Figure 1 Schematic diagram of GST–Sp1 selection of DNA fragments or plasmids. All steps are performed in microcentrifuge tubes using a "batch" protocol. The GST–Sp1 fusion protein is first purified from a crude bacterial lysate using glutathione–agarose beads (step 1). The beads are pre-equilibrated with binding buffer, DNA is added, and binding is allowed to occur (step 2). The beads are then washed several times with a 300 mM KCl buffer that removes unbound DNA (step 3). Finally, the remaining bound DNA is removed by elution at 65°C (step 4).

Enrichment was also observed when plasmids were used as input DNA. This allowed additional rounds of selection to be performed on circular subcloned DNA (data not shown).

Application of Selection to Genomic Clones

To determine whether GST–Sp1 selection could enrich fragments with Sp1-binding sites from a more complex mixture of DNA, we applied GST–Sp1 selection to four phage P1 clones from an incomplete contig containing the murine *Hoxa* gene cluster (Fig. 3), as well as a human P1 con-

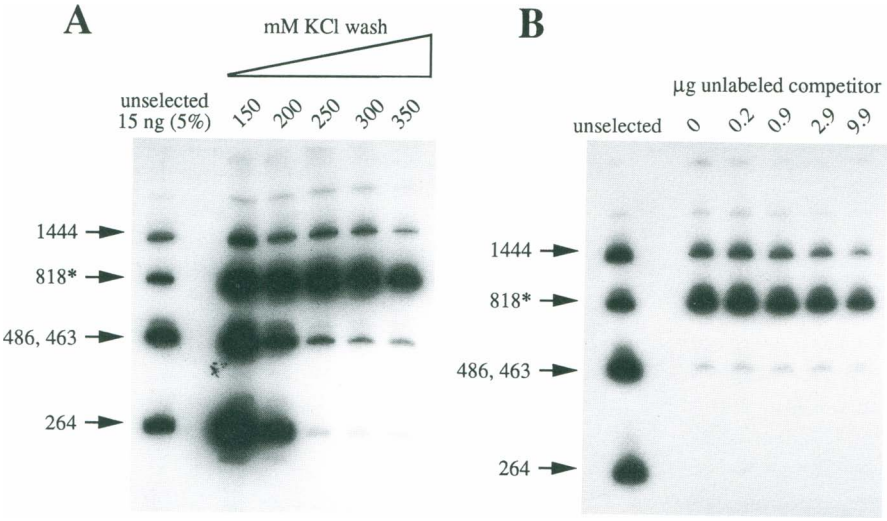

Figure 2 Determination of wash conditions and DNA-binding capacity for GST–Sp1 selection. Shown are end-labeled DNA fragments subjected to 4% polyacrylamide nondenaturing gel electrophoresis. (*A*) Fifteen nanograms of labeled, unselected p0.7 restriction fragments are shown in the *left* lane. Fragments (300 ng) were used as input DNA for each of five selection experiments; selected DNAs were eluted, precipitated, and loaded in the lanes at *right*. Selections were identical except for wash buffer KCl concentration as indicated. Size in base pairs of restriction fragments is shown to the *left* of the gel. The asterisk denotes the 818-bp fragment that has several high-affinity Sp1-binding sites. Not shown is a 32-bp fragment that was not bound by GST–Sp1. (*B*) Demonstration of GST–Sp1 DNA-binding capacity. End-labeled p0.7 fragments were mixed with varying amounts of unlabeled p0.7 fragments and subjected to GST–Sp1 selection. (*Left* lane) Five nanograms of unselected end-labeled fragments. An aliquot of 100 ng of end-labeled fragments was used as input DNA for five selections, using 300 mM KCl wash conditions. In four selections, unlabeled fragments were added to bring the total amount of DNA to between 300 ng and 10 µg. Selected DNAs were loaded in the lanes at *right*.

taining part of the BMP6 gene. Because P1 clone inserts average 80–100 kb (Sternberg 1990), these five clones together likely contain >0.4 Mb of insert genomic DNA.

We estimated that each clone would yield ≥250 *Sau*3AI restriction fragments. P1 *Sau*3AI fragments were selected with GST–Sp1, in separate binding reactions for each P1 clone, and eluted fragments were subcloned into the *Bam*HI site of pTZ18R. The resulting first-round library was amplified in *Escherichia coli*. Selected library plasmids were harvested from transformed colonies and used as input DNA for two successive rounds of GST–Sp1 selection. The rationale for the number of selection rounds used is described below (see Discussion).

From each P1 clone, a final selected library was generated, and subclones from each library were analyzed by restriction digestion. Enriched clones were readily identifiable as those that had identical insert sizes. Inserts from these clones were used as probes to determine more precisely the subclone abundance in both shotgun *Sau*3AI libraries and selected libraries.The abundance of individual *Sau*3AI inserts in shotgun libraries was typically between 0.3–0.5%, agreeing with the predicted number of fragments from P1 clone digests. Enriched subclones were sequenced fully or partially and subjected to BLASTN and BLASTX homology searches. No selected subclone inserts

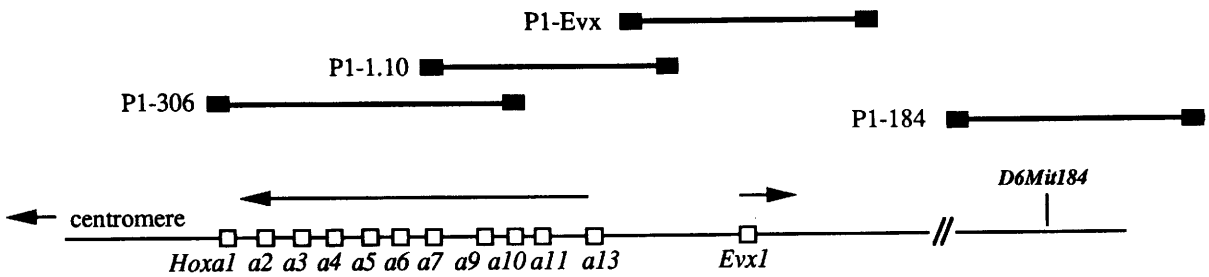

Figure 3 P1 contig of the murine *Hoxa* gene cluster on chromosome 6. Three overlapping clones comprise the *Hoxa* locus; a fourth P1 clone, containing *D6Mit184*, lies ~0.3 cM distal to the locus. A minimum of 12 known genes are contained in the contig, including all *Hoxa* genes and the *Evx*1 gene.

had sequences derived from the P1 cloning vector (pAD10SacBII), as determined by either Southern blot analysis or by an observed lack of P1 vector sequence homology.

Characteristics of GST–Spl-selected Subclones

Table 1 shows a summary of enriched clones, sorted by P1 clone of origin. From the 5 P1 clones, 11 enriched subclones were identified that represented ≥6% of a selected library. Subclone sA-40 showed sequence identity to the promoter of the *Hoxa4* gene and contains a *Sau*3AI fragment of 754 bp that includes 246 bp of sequence 5′ to the transcription start site, as well as 508 bp of the first exon. We anticipated that this sequence would be highly enriched by GST–Sp1, because this promoter is known to contain several Sp1-binding sites between −237 and −47 bp relative to the transcription start site (Chavrier et al. 1990).

Southern blot analysis identified sA-18 to be a *Bam*HI fragment between the *Hoxa2* and *Hoxa1* genes. This fragment contains a CpG island and two GC boxes that lie ~1.9 and 2.2 kb 5′ to the *Hoxa1* transcription start site. Subclone s2-6 originates from between exons 1b and 2 of the *Hoxa9* gene (Benson et al. 1995), as confirmed by PCR (data not shown). This fragment maps to the region of overlap between P1-306 and P1-1.10 and was enriched to a similar degree from each clone (representing 6.0% and 5.1% of the final libraries, respectively). Interestingly, the sequence of s2-6 shows partial homology (75% over a 63-bp region) to the 5′ end of the guinea pig *Hoxa9* cDNA clone GPK5 (Rubin et al. 1987). However, the homologous murine sequence contains multiple small deletions, insertions, and stop codons that preclude the existence of a conserved reading frame.

Subclone s3-10 (Fig. 4A) contains a highly GC-rich insert with six GC boxes. Southern blot experiments showed that this insert hybridizes to the same 3.0-kb *Xba*I fragment from P1-1.10 that contains the *Hoxa13* homeo box. Subsequently, we have subcloned the entire *Hoxa13* gene region from P1-1.10 and determined that s3-10 lies ~2 kb 5′ to the homeo box. Most of the mammalian Hox cluster genes have a similar gene structure, comprised of two exons and a single intron, with the homeo box usually present in the second exon (Galliot et al. 1989; Tan et al. 1992; Benson et al. 1995; Hsieh-Li et al. 1995). Sequence of the genomic region immediately 3′ to s3-10 (relative to transcription of *Hoxa13*) showed an uninterrupted open reading frame (ORF) with significant

Table 1. Subclones isolated by GST–Sp1 selection

Clone of origin	Subclone	Identity	No. of GC boxes	CpG island	Size (bp)	Percent in final selection
P1-306	sA-40	*Hoxa4* 5′ end	5	yes	790	32.0
	sA-18	*Hoxa1* 5′ region	2	yes	1227	17.7
	s2-6	*Hoxa9* 5′ region	2	yes	419	6.4
P1-1.10	s3-10	*Hoxa13* 5′ end	6	yes	390	45.5
	s3-29	GAG repeat	0	no	276	18.0[a]
P1-Evx	s3-3	N.D.	N.D.[b]	N.D.	~1300	20.2
	s3-11	transposon-like element	0	no	391	27.8
P1-184	s3-2	N.D.	4	yes	503	70.0[a]
	s3-15	N.D.	0	no	289	6.0[a]
	s3-41	N.D.	≥2[b]	N.D.	~1200	12.0[a]
P1-BMP6 (human)	s3-1	protein disulfide isomerase EST $P(N) = 2.2 \times 10^{-9}$	≥2[b]	yes	~2100	32.3

P1 clones 306, 1.10, Evx, and 184 are from murine chromosome 6, and P1-BMP6 is a human P1 clone containing the BMP6 gene. (N.D.) Not determined. CpG island status of insert sequences was determined by criteria described previously (Larsen et al. 1992). The right-hand column shows the final percentage of indicated clones in final selected libraries; all data were determined by hybridization to lifts of >100 colonies, except as indicated.

[a]Percent of 50 clones analyzed by restriction digests.

[b]Subclones that were partially sequenced.

A. P1-1.10 #3-10

```
GATCGGGGTCCCTGTGTTCCCAGCCCCACCCGCCCTAGTTTCTCCCCCCCACCCACCCCC 60
                                  1
GCCTCCTCCGTGGGTGACTGCTGCGCGGCCGCGCGGCGGGAGCTAGAGCTGGCGCAGAGG 120
ACCGCGCGGGCGGCGGCGGGGACGGCGGGCCCAGCGCGGAGCCGAGCCGGAGCCTCCCGC 180
      2
TTTGCATACGCCGGGGGCGGGCCGGGGGCGGGCCAATGGGCGGCCGCCTGGCGCGACCCC 240
                3         4         5
GCAGGGCGCGACCCGCCCCCCTCGCTCGGGCCCCGGCCCCGCGCGCCGCGCTCTTCACTT 300
          6
CTTGGGGGCTTTTTAAAACAGCGCCACTGGGGTCTTCTCCATGCGGCTCGGGCTATGACA 360
GCCTCCGTGCTCCTCCACCCCCGCTGGATC                                390
```

B. P1-184 #3-2

```
GATCGCAGACGCCGCCGCAAGCACCCGGGTCCGCTCGAGGGGCGGGATTTGGGTGGGGCG 60
                                  1                    2
GAGAACCGGGCCCGCCCCCCTCGGAGTGCGGGCCCCGCCCCCTCTCTCCCTCCTTTTCCCCA 120
          3                   4
ACGGACCGCAATGATTTAGAAGTTCAGGAATCCCACGTGACGTCACCGGGGGGGTGATGT 180
TATGTACTCTAAATAGAATGTATTGTAATCTTTGCTCAGTCCAACGTGGTCCCTTCACCC 240
GGGCTCGGCTCTTGCCTTCTCCACACTTGTTTGGTAAGTTCCTAAAAATACGGCGCGGCC 300
ACTTCTGGGCCTTTTGCATTTAGCCACTTTCGCTGCACTTGCCGCGCGAGCCGACTCCGG 360
AGCGTCCTTGTTTGCAGGGGTTGGTTCATCGCTGCAGCCGCGAGGTTACTTAGTCCAATG 420
GGCAGGTCACAGGGCAAGGAGTAGAAAGGATGCGATGCATTTTTTCCCCCTTGATTTGTG 480
TTTGTAAACACTTTGGCGGATC                                        503
```

C. P1-BMP6 #3-1

```
GATCCAGAGCGCCGCGCACTTCGTCATGTTCTTCGCGCCCTGGTAACGCCGCGCAGGCCG 60
GCGCCCTGCCCCGCACTTGGCGCTTCGCGGGGCGTGGGGCGCTTGGGCGCCGCGGGCGGG 120
                                                  1
GGGTCCCGGGGCCACGTGGACGCCGACGGCTCTGGTCTGCGGCGTGCGTGCTACGGGCGC 180
GGGGCTCGGGGCTCAGGGGTGCGGGGTGTAATACTGGGGGCGCGGGGCCGGGAGCGCGGG 240
GACCCTGCGGGTTCGGGGCTCCGGCTGGCGGCGAGCGGCGGGCGCGGCCGCCGCGGCTCG 300
GGGCCGGGAGCCGGGGACTGCAGGGGGACCGAGGGCTTGGGGCTCCGGGTCGGGAGGCGT 360
TAACTCA                                                       367
```

Figure 4 Sequence of selected inserts s3-10 (*A*), s3-2 (*B*), and s3-1 (*C*). GC boxes (GGGCGG or CCGCCC) are underlined. The sense strand of s3-10 relative to the *Hoxa13* homeo box is shown. The complete *Sau*3AI fragment sequence is shown for s3-10 and s3-2; partial sequence of one end of the s3-1 insert is shown. A second GC box was present at the opposite end of s3-1 (not shown). The identical match within the first 44 bases of s3-1 to the sense strand of a human protein disulfide isomerase EST is underlined.

homology to the axolotl *Hoxa13* (Gardiner et al. 1995) and chicken *Hoxd13* (Rogina and Upholt 1993) genes. Reverse transcription–PCR (RT–PCR) experiments indicated that this ORF is transcribed, and primer extension experiments on embryonic mRNA showed several *Hoxa13* transcription initiation sites within s3-10, between bases 249 and 303 (data not shown). These data indicate that the likely translational start codon for *Hoxa13* is encoded at base 355 of s3-10. We will present the structure and sequence of the murine *Hoxa13* gene in a separate publication.

Subclone s3-2 (Fig. 4B) showed no significant data base homology matches or putative exon-coding regions; however, its sequence contains part of a CpG island and four GC box motifs. Partial sequence of clone s3-1 (Fig. 4C) showed a 44-bp direct match to a human expressed sequence tag (EST) (EMBL/GenBank no. Z21669). A BLASTX search showed that this EST contains an ORF that shares homology with a family of protein disulfide isomerases, with highest matches to the mouse, rat, and human ER60 proteins [$P(N) < 1 \times 10^{-25}$ for each].

Two other clones, s3-29 from P1-1.10 and s3-11 from P1-Evx, did not contain GC boxes or CpG-island-like sequences, although they were enriched significantly by GST–Sp1. However, clone s3-29 contains a GAG trinucleotide repeat. The repeat consists wholly of purines on one strand and pyrimidines on the other (GAGGAG/CTCCTC), which is similar to the GC box motif (GGGCGG/CCGCCC). Clone s3-11 showed BLASTN homology [highest match was to GenBank no. U17095, $P(N) = 2.8 \times 10^{-110}$] to a group of rodent transposon-like elements, and contains four imperfect repeats containing the motif TGGGTGTG, which is also similar to the GC box. Although these do not appear to be strong Sp1-binding motifs, we hypothesize that their reiteration allows more efficient Sp1 binding, as there were no GC boxes in these subclones. It has been shown that Sp1 can sometimes bind to imperfect GC box-like motifs (Kadonaga et al. 1986); therefore, concatamerized variants of this sequence might be able to compete effectively with consensus binding motifs in GST–Sp1 selection experiments.

In summary, of the 11 clones identified by GST–Sp1 selection, 1 is a previously characterized promoter (sA-40), 1 is a newly identified promoter (s3-10), 2 are CpG islands that map to intergenic or intronic regions of the *Hoxa* cluster, 1 is an anonymous murine CpG island, 1 is a human CpG island that contains part of a protein disulfide isomerase gene coding region, and 2 clones contain repetitive motifs of putative non-consensus Sp1-binding sites.

DISCUSSION

We found that promoter capture, using GST–Sp1 selection, is a quick way to isolate putative promoters from P1 clones. One GST–Sp1 selection experiment takes ~2 hr, and identification of enriched clones can be accomplished with minimal equipment. Starting from the source P1 DNA, selection and identification of enriched subclones can be accomplished within 1 week, and multiple P1 clones can be processed simultaneously to increase throughput. Between one and four en-

riched subclones were generated per P1; therefore, only a small number of subclones need to be pursued as putative promoters. The average size of the selected subclones is 800 bp, which is long enough to be an efficient probe for cDNA screening and physical mapping purposes, yet short enough to be fully sequenced using vector primers. We are exploring the use of bacterial artificial chromosome (BAC) and yeast artificial chromosome (YAC) clone DNAs, or pooling strategies with cosmid DNA, to extend the capabilities and flexibility of this technique to varied cloning projects. In a test of GST–Sp1 selection on YAC DNA, five rounds of selection were applied to total yeast DNA *Sau*3AI fragments from a murine YAC strain that contains subclone s3-10; hybridization data indicate that this sequence was enriched ~1600-fold, indicating that GST–Sp1 selection could be used to extract promoter sequences from YACs.

Our goal in using GST–Sp1 selection was to quickly isolate the few fragments from each P1 clone that had the strongest Sp1-binding capability. To show that enrichment occurs with each round of selection, we determined the abundance of the sA-40 and s3-10 subclones at each step in the selection process by hybridization (Table 2). The most dramatic enrichment occurred during the first selection on restriction fragments, whereas the second selection showed about a threefold enrichment for either sequence. The third selection did not further enrich s3-10, but sA-40 was enriched from 22% to 32%. Although the enrichment in the third round may vary, it adds minimal time and effort to the overall selection process and probably aids in the removal of unwanted subclones.

The *Hoxa* contig contains all 11 *Hoxa* genes, as well as the *Evx1* gene; therefore, it contains at least 14 promoters, including two alternate *Hoxa10* promoters (Benson et al. 1995) and at least one *Hoxa11* antisense promoter (Hsieh-Li et al. 1995), as well as a complex assortment of regulatory elements, exons, introns, and intergenic regions. Promoter sequences >100 bp upstream of the transcription start site are currently known for the *Hoxa1, Hoxa2, Hoxa4, Hoxa10* (both promoters), *Hoxa11*, and *Hoxa13* (this paper) genes, as well as *Evx1*. The 5′ region of each of these genes contains a CpG island; however, only the *Hoxa4* and *Hoxa13* promoters contain a cluster of GC boxes in sequences immediately upstream of the transcription start site. All other known promoters within this contig lack GC box clusters, although the *Hoxa1* promoter contains a single GC box, 25 bp 5′ to the start site (LaRosa and Gudas 1988) and a single GC box lies 126 bp upstream of the *Hoxa10* exon 1a 5′ end (Gail Benson, pers. comm.). The fact that several GST–Sp1-selected clones containing GC box clusters were enriched, whereas several other CpG islands that lack such clusters were not, indicates the ability of GST–Sp1 selection to discriminate between islands that do or do not contain Sp1-binding sites.

GST–Sp1 selection should be a useful addition to the growing repertoire of methods capable of isolating promoter sequences. Two other approaches for isolating potential promoters from genomic clones have been described. The first appro-ach is directed at the isolation of CpG islands, which are associated with a large fraction of mammalian promoters. Islands can be isolated using rare-cutting restriction enzyme sites (John et al. 1994; Valdes et al. 1994), denaturing gradient gel electrophoresis (DGGE) (Shiraishi et al. 1995), or from total genomic DNA, with a methyl-CpG-binding protein (Cross et al. 1994). Because CpG islands are sometimes located in 3′ regions of transcription units (Gardiner-Garden and Frommer 1987), some islands may not overlap

Table 2. Enrichment Increases with Multiple Rounds of Selection

			Rounds of selection		
Subclone	P1 of origin	Shotgun	1	2	3
sA-40	P1-306	0.4% (5/1216)	7.2% (83/1150)	22% (63/285)	32% (127/389)
s3-10	P1-1.10	0.3% (4/1445}	14.9% (183/1227)	46% (345/750)	45% (80/176)

Shown are the abundances of two subclones in selected libraries after various rounds of selection, as assayed by hybridization. The number of positive colonies relative to the total number screened is shown in parentheses with each percentage value.

gene promoters. A second approach to promoter isolation exploits the ability of some subcloned promoters to drive transcription of reporter genes in transfected cells (Toyoda et al. 1995), although the percentage of promoters that could be identified in this way is not clear.

The utility of GST–Sp1 selection depends on the fraction of mammalian gene promoters that have clusters of Sp1-binding sites. Approximately 57% of human genes are associated with CpG islands, and of these genes, ~70% have islands at the 5′ end (Larsen et al. 1992). A survey of 66 such promoters showed an average of about 3 GC boxes 5′ to the transcription start site (Gardiner-Garden and Frommer 1987). Given this data, we predict that up to 30%–40% of all promoters, on average, will have suitable clusters of Sp1 sites that could be isolated using GST–Sp1 selection. We applied selection to five P1 clones representing ~250 kb of nonoverlapping murine DNA and a human P1 clone of ~80kb. Including subclones s3-2 and s3-1, we identified four actual or putative promoter sequences from the total of 330 kb analyzed, or one per 83 kb. This brings the presumed number of promoters present in the clones analyzed to at least 16, and four actual or putative promoters were isolated (25%). The application of promoter capture to other regions of the human or mouse genomes will fully address the overall efficiency of this methodology.

Not all promoters will be identified using GST–Sp1 selection; however, we have developed GST–Sp1 selection as a complementary technique to existing gene and promoter isolation methods. To extend promoter capture to selection of non-Sp1-regulated promoters, it might be implemented using other transcription factors as DNA-selecting agents. The most useful factors for this would be those that bind to sequence motifs of sufficient length and complexity to make them rare in nonpromoter genomic DNA and, ideally, would bind to a significant fraction of promoters, making them generally applicable for promoter isolation projects. Alternatively, selection using proteins with more gene-specific regulatory effects might be useful in locating distant enhancer elements associated with target genes.

METHODS

Plasmids and Genomic Clones

p0.7 is a pTZ19R-derived plasmid containing bases from −147 to +515 relative to the cap site of the murine adenosine deaminase gene (Innis et al. 1991). To obtain labeled restriction fragments for experiments to determine binding conditions, p0.7 was digested with *Hin*dIII and *Taq*I, and fragments were end-labeled with [α-^{32}P]-dGTP by Klenow polymerase overhang fill-in. The pGEX–KT–Sp1(FD) plasmid (Perkins et al. 1994) is an *E. coli* expression vector that encodes the GST–Sp1 fusion protein. The vector pGEX–KT has been described previously (Hakes and Dixon 1992). P1 clones from the murine *Hoxa* locus contig are described elsewhere (Innis et. al. 1996). P1–BMP6 was isolated from a human P1 library by PCR screening using an STS from the 3′-untranslated region of the BMP6 gene (GenBank no. M60315) and confirmed by Southern hybridization. P1 DNA was prepared from bacterial cultures by alkaline lysis followed by purification with Qiagen maxiprep kits.

Preparation of GST–Sp1 Lysate

The *E. coli* XL1-BLUE strain containing pGEX–KT–Sp1(FD) was inoculated into 25 ml of Luria broth (LB) with 100 μg/ml of ampicillin, and shaken overnight at 37°C. Cells were then pelleted, resuspended in 500 ml of fresh LB with ampicillin, and allowed to grow to an OD_{600} of 0.6. IPTG was then added to 0.66 mM to induce GST–Sp1 expression, and the culture was shaken for 3 hr. The cells were then pelleted, washed once in PBS, and resuspended in 100 ml of buffer Q at 4°C [20 mM HEPES (pH 7.9), 10 mM $MgCl_2$, 20% glycerol, 0.01% NP-40, 1 mM DTT, 0.5 mM PMSF, 2.5 U/ml DNase I, 1 μg/ml of leupeptin]. The cell resuspension was placed on ice and sonicated three times with a standard microtip (Fisher 550 Sonic Dismembrator) for 15 sec each, with a 1-min pause between bursts. Disodium EDTA (pH 8.0), was added to 1 mM to inhibit DNase I, and the lysate was centrifuged at 20,000*g*, at 4°C, for 15 min. Finally, the supernatant was divided into 1-ml aliquots in 1.5-ml microcentrifuge tubes, snap-frozen on dry ice/ethanol, and stored at −80°C.

Purification of GST–Sp1 Protein

For each selection, 1 ml of frozen GST–Sp1 lysate was thawed on ice. A 40-μl aliquot of a 1:1 slurry of glutathione–agarose beads and 0.1 M NaCl (stored at 4°C; Sigma cat. no. G4510) was added to the thawed lysate, which was then gently shaken for 1 hr on ice. The beads were pelleted briefly, the supernatant was discarded, and the beads were washed (gently mixing by inversion, and recentrifuging briefly to pellet beads each time) three times with 500 μl of ice-cold IW buffer [20 mM HEPES (pH 7.9), 75 mM KCl, 2.5 mM $MgCl_2$, 0.01% NP-40, 1 mM DTT, 0.5 mM PMSF] and twice with 500 μl of ice-cold Sp1-binding buffer [24 mM Tris-HCl (pH 8.0), 10% glycerol, 50 mM KCl, 5 μM $ZnSO_4$, 0.5 μM DTT, 6.25 mM $MgCl_2$, and 0.25 mM disodium EDTA (pH 8.0)], keeping the tubes on ice between centrifugations.

DNA Selection with GST–Sp1 Agarose Beads

Between 100 ng and 5 μg of DNA restriction fragments, or 10 μg of plasmid DNA, was resuspended in 50–100 μl of Sp1-binding buffer, and chilled on ice. This binding mixture was added to pre-equilibrated GST–Sp1 agarose beads and incubated for 10 min on ice, with gentle mixing of the tube every few minutes to resuspend the agarose beads.

The tubes were respun as above and the supernatant was discarded. The beads were then washed five times with 500 μl of ice-cold Sp1 wash buffer (identical to the Sp1-binding buffer above, except with 300 mM KCl). Finally, the agarose beads were resuspended in 200 μl of elution buffer [300 mM sodium acetate (pH 7.0), 1 mM EDTA], vortexed briefly, and incubated at 65°C for 15 min to heat inactivate GST–Sp1. Eluted DNA was precipitated with ethanol and 5 μg of glycogen.

Amplification of Subcloned, Selected Fragments

Portions of subclone ligations or GST–Sp1-selected plasmids were transformed by electroporation into *E. coli* XL1-BLUE cells, which were plated to a density of 2000–3000 colonies on 15-cm (LB + 100 μg/ml of ampicillin) agar plates. The colonies were gently scraped with a clean spatula into 5 ml of PBS and pelleted, and plasmid DNA was purified from the cell pellet using a standard commercial miniprep kit (Wizard Minipreps, Promega).

Analysis of GST–Sp1-enriched Libraries

For restriction analysis, 50 colonies from each selected library were picked randomly from transformant plates and inoculated into 300 μl of TYGPN medium (20 grams/liter of tryptone, 10 grams/liter of yeast extract, 5 grams/liters of sodium monophosphate, 0.8% glycerol, 0.1% potassium nitrate) in a 96-well format, and incubated for 2 days at 37°C. Plasmid DNAs were prepared in plate format by alkaline lysis, resuspended in 10–20 μl of TE (10 mM Tris-HCl/1 mM EDTA) at pH 8.0, and 6 μl of each was used for 8-μl restriction digests with *Xba*I and *Sac*I, which flank the *Bam*HI cloning site in the pTZ18R polylinker. Digested DNA was separated by electrophoresis in 1.4% agarose gels. For hybridization analysis, colonies were transferred directly to nitrocellulose or nylon membranes and the membranes were screened with [α-^{32}P]-dGTP-labeled probes prepared from purified subclone inserts by random-primed labeling (Feinberg and Voglestein 1983). Hybridizations were performed in a 7% SDS/0.25 M sodium phosphate buffer (pH 7.2), with 50 μg/ml of sheared, denatured herring sperm DNA, at 65°C. Membranes were washed with 0.1× SSC/0.1% SDS at 65°C for final stringency.

Sequencing

DNA sequencing was performed by the University of Michigan Biomedical Research Core Facilities DNA Sequencing Core with either T7 or M13 forward primers or custom primers. Sequences of GST–Sp1-selected inserts were deposited in GenBank (accession numbers U50600-U50602).

ACKNOWLEDGMENTS

We thank Kristen Kazen-Gillespie and Laura Post for excellent technical assistance and helpful advice. We especially thank Neil Perkins and Gary Nabel for providing the pGEX KT Sp1(FD) vector and GST–Sp1 purification protocols. We also thank Jeffrey Chamberlain, Ronald Koenig, Jerome Gorski, and David Burke for reading the manuscript and providing helpful comments. This work was supported by grants from the Phoenix Memorial Fund, Rackham Graduate School, and the Office of the Vice President for Research of the University of Michigan. J.W.I. also wishes to acknowledge support from the Departments of Human Genetics and Pediatrics. D.P.M. is supported by a University of Michigan Research Partnership Program Fellowship. The sequence data described in this paper have been submitted to the GenBank data library under accession nos. U50600–U50602.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Benson, G.V., T.E. Nguyen, and R.L. Maas. 1995. The expression pattern of the murine *Hoxa-10* gene and the sequence recognition properties of its homeodomain reveal specific properties of *Abdominal B*-like genes. *Mol. Cell. Biol.* **15:** 1591–1601.

Briggs, M.R., J.T. Kadonaga, and S.P. Bell. 1986. Purification and biochemical characterization of the promoter-specific transcription factor, Sp1. *Science* **234:** 47–52.

Bucher, P. 1990. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.* **212:** 563–578.

Buckler, A.J., D.D. Chang, S.L. Graw, J.D. Brook, D.A. Haber, P.A. Sharp, and D.E. Housman. 1991. Exon amplification: A strategy to isolate mammalian genes based on RNA splicing. *Proc. Natl. Acad. Sci.* **88:** 4005–4009.

Chavrier, P., C. Vesque, B. Galliot, M. Vigneron, P. Dollé, D. Duboule, and P. Charnay. 1990. The segment-specific gene *Krox-20* encodes a transcription factor with binding sites in the promoter region of the *Hox-1.4* gene. *EMBO J.* **9:** 1209–1218.

Cross, S.H., J.A. Charlton, X. Nan, and A.P. Bird. 1994. Purification of CpG islands using a methylated DNA binding column. *Nature Genet.* **6:** 236–244.

Feinberg, A.P. and B. Voglestein. 1983. A technique for radiolabeling DNA restriction fragments to high specific activity. *Anal. Biochem.* **132:** 6–13.

Galliot, B., P. Dollé, M. Vigneron, M.S. Featherstone, A. Baron, and D. Duboule. 1989. The mouse Hox-1.4 gene: Primary structure, evidence for promoter activity and expression during development. *Development* **107:** 343–359.

Gardiner, D.M., B. Blumberg, Y. Komine, and S.V. Bryant. 1995. Regulation of *Hoxa* expression in developing and regenerating axolotl limbs. *Development* **121:** 1731–1741.

Gardiner-Garden, M. and M. Frommer. 1987. CpG islands in vertebrate genomes. *J. Mol. Biol.* **196:** 261–282.

Gronostajski, R., K. Nagata, and J. Hurwitz. 1984. Isolation of human DNA sequences that bind to nuclear factor I, a host protein involved in adenovirus DNA replication. *Proc. Natl. Acad. Sci.* **81:** 4013–4017.

Hakes, D.J. and J.E. Dixon. 1992. New vectors for high level expression of recombinant proteins in bacteria. *Anal. Biochem.* **202:** 293–298.

Hsieh-Li, H.M., D.P. Witte, M. Weinstein, W. Branford, H. Li, K. Small, and S.S. Potter. 1995. *Hoxa11* structure, extensive antisense transcription, and function in male and female fertility. *Development* **121:** 1373–1385.

Innis, J.W., D. Moore, S. Kash, V. Ramamurthy, M. Sawadogo, and R. Kellems. 1991. The murine adenosine deaminase promoter requires an atypical TATA box which binds transcription factor IID and transcriptional activity is stimulated by multiple upstream Sp1 binding sites. *J. Biol. Chem.* **266:** 21765–21772.

Innis, J.W., S.M. Darling, K. Kazen-Gillespie, L.C. Post, D.P. Mortlock, and T. Yang. 1996. Orientation of the Hoxa complex and placement of the *Hd* locus distal to *Hoxa2* on mouse chromosome 6. *Mamm. Genome* (in press).

John, R.M., C.A. Robbins, and R.M. Myers. 1994. Identification of genes within CpG-enriched DNA from human chromosome 4p16.3. *Hum. Mol. Genet.* **3:** 1611–1616.

Kadonaga, J.T., K. Jones, and R. Tjian. 1986. Promoter-specific activation of RNA polymerase II transcription by Sp1. *Trends Biochem. Sci.* **11:** 20–23.

Kinzler, K. and B. Vogelstein. 1989. Whole genome PCR: Application to the identification of sequences bound by gene regulatory proteins. *Nucleic Acids Res.* **17:** 3645–3653.

——— 1990. The GLI gene encodes a nuclear protein which binds specific sequences in the human genome. *Mol. Cell. Biol.* **10:** 634–642.

Korn, B., Z. Sedlacek, A. Manca, P. Kioschis, D. Konecki, H. Lehrach, and A. Poustka. 1992. A strategy for the selection of transcribed sequences in the Xq28 region. *Hum. Mol. Genet.* **1:** 235–242.

LaRosa, G.J. and L.J. Gudas. 1988. Early retanoic acid-induced F9 teratocarcinoma stem cell gene ERA-1: Alternate splicing creates transcripts for a homeo box-containing protein and one lacking the homeobox. *Mol. Cell. Biol.* **8:** 3906–3917.

Larsen, F., G. Gundersen, R. Lopez, and H. Prydz. 1992. CpG islands as gene markers in the human genome. *Genomics* **13:** 1095–1107.

Lelong, J., G. Prevost, K.I. Lee, and M. Crepin. 1989. South Western blot mapping: A procedure for simultaneous characterization of DNA binding proteins and their specific genomic DNA target sites. *Anal. Biochem.* **179:** 299–303.

Lovett, M., J. Kere, and L. Hinton. 1991. Direct selection: A method for the isolation of cDNAs encoded by large genomic regions. *Proc. Natl. Acad. Sci.* **88:** 9628–9632.

Perkins, N.D., A.B. Agranoff, E. Pascal, and G.J. Nabel. 1994. An interaction between the DNA-binding domains of RelA(p65) and Sp1 mediates Human Immunodeficiency Virus gene activation. *Mol. Cell. Biol.* **14:** 6570–6583.

Prestridge, D. and C. Burks. 1993. The density of transcriptional elements in promoter and non-promoter sequences. *Hum. Mol. Genet.* **2:** 1449–1453.

Rogina, B. and W.B. Upholt. 1993. Cloning of full coding chicken cDNAs for the homeobox-containing gene Hoxd-13. *Nucleic Acids Res.* **21:** 1316.

Rubin, M.R., W. King, L.E. Toth, I.S. Sawczuk, M.S. Levine, P. D'Eustachio, and M.C. Nguyen-Huu. 1987. Murine *Hox-1.7* homeo-box gene: Cloning, chromosomal location, and expression. *Mol. Cell. Biol.* **7:** 3836–3841.

Saffer, J.D., S.P. Jackson, and M.B. Annarella. 1991. Developmental expression of Sp1 in the mouse. *Mol. Cell. Biol.* **11:** 2189–2199.

Shiraishi, M., L.S. Lerman, and T. Sekiya. 1995. Preferential isolation of DNA fragments associated with CpG islands. *Proc. Natl. Acad. Sci.* **92:** 4229–4233.

Sternberg, N. 1990. Bacteriophage P1 cloning system for the isolation, amplification, and recovery of DNA fragments as large as 100 kilobase pairs. *Proc. Natl. Acad. Sci.* **87:** 103–107.

Tagle, D., M. Swaroop, M. Lovett, and F. Collins. 1993. Magnetic bead capture of expressed sequences encoded within large genomic segments. *Nature* **361:** 751–753.

Tan, D., J. Ferrante, A. Nazarali, X. Shao, C.A. Kozak, V. Guo, and M. Nirenberg. 1992. Murine *Hox-1.11* homeobox gene structure and function. *Proc. Natl. Acad. Sci.* **89:** 6280–6284.

Thiesen, H. and C. Bach. 1990. Target detection assay (TDA): A versatile procedure to determine DNA binding sites as demonstrated on Sp1. *Nucleic Acids Res.* **18:** 3203–3209.

Toyoda, A., J. Kusuda, H. Maeda, and K. Hashimoto. 1995. An enhanced promoter trap protocol. *Mamm. Genome* **6:** 426–428.

Valdes, J.M., D.A. Tagle, and F.S. Collins. 1994. Island rescue PCR: A rapid and efficient method for isolating transcribed sequences from yeast artificial chromosomes and cosmids. *Proc. Natl. Acad. Sci.* **91:** 5377–5381.

Received November 7, 1995; accepted in revised form March 1, 1996.