Whole genome PCR: application to the identification of sequences bound by gene regulatory proteins

Kenneth W.Kinzler and Bert Vogelstein

The Oncology Center, The Johns Hopkins University School of Medicine, 424 N. Bond Street, Baltimore, MD 21231, USA

ABSTRACT
A strategy is described that allows the isolation of DNA sequences that can bind to gene regulatory proteins. Total genomic DNA is first converted to a form that is suitable for amplification by the polymerase chain reaction (Whole Genome PCR), and the DNA sequences of interest are selected by binding to the regulatory protein and immune precipitation. Because sequences recovered from the selection step can be amplified by PCR, the selection process can be designed for maximum enrichment with little concern about recovery. Furthermore, the selection process can be repeated as often as necessary. Sequences recovered after amplification can be cloned and/or used as hybridization probes. As a test of this strategy, we selected human sequences that bound to Xenopus transcription factor IIIA (TFIIIA). Seven clones were isolated that were on the average 94% identical to the previously described 61 bp binding site of TFIIIA. This strategy could be adapted to isolate sequences that can be selected by any physical or biological method.

INTRODUCTION
Perhaps the oldest problem encountered in biochemistry is the inevitable loss of desired molecules during their purification. This problem is exacerbated when the initial concentration of the desired molecule is low and several sequential steps are required to achieve purity. DNA sequences in particular pose a significant problem in this regard, since individual sequences of interest (for example, those that bind to specific proteins) often represent less than 0.01% of the total sequences present (1). In this report, we show how the purity-yield problem can be circumvented during the isolation of selectable DNA fragments. This strategy involves converting total genomic DNA to a form which can be amplified *ad infinitium* by the polymerase chain reaction (whole genome PCR). DNA sequences of interest are selected by binding to a solid support, and the selected DNA reamplified. Thus even miniscule amounts of selected sequences can be recovered efficiently. The selected DNA can then be cloned into prokaryotic vectors and appropriate recombinants chosen by hybridization to a probe that identifies sequences enriched by the procedure. As an illustration of the technique, we show that human DNA sequences which bind to the *Xenopus laevis* transcription factor IIIA (TFIIIA, reviewed in 2) can be easily identified and cloned using this strategy.

MATERIALS AND METHODS
Preparation of DNA for whole genome PCR
High molecular weight DNA (4 μg) from human lymphocytes was cleaved to completion with MboI (BRL) or sheared to an average size of 200–300 bp (mass average) by sonication on ice. The ends of the DNA were made blunt by incubation with the Klenow fragment
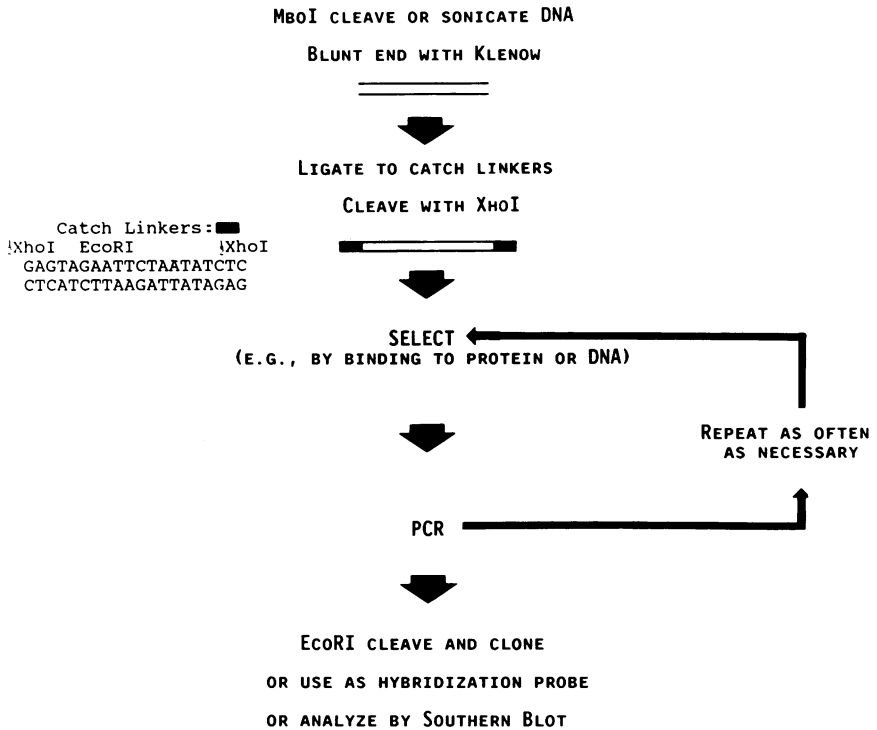
MBoI CLEAVE OR SONICATE DNA

BLUNT END WITH KLENOW

LIGATE TO CATCH LINKERS

CLEAVE WITH XHoI

```
     Catch Linkers:■
 !XhoI  EcoRI         !XhoI
   GAGTAGAATTCTAATATCTC
   CTCATCTTAAGATTATAGAG
```

SELECT
(E.G., BY BINDING TO PROTEIN OR DNA)

REPEAT AS OFTEN
AS NECESSARY

PCR

EcoRI CLEAVE AND CLONE

OR USE AS HYBRIDIZATION PROBE

OR ANALYZE BY SOUTHERN BLOT

Figure 1: *Schematic of the Whole Genome PCR applied to selection of DNA sequences.*

of DNA polymerase I (4 units, Pharmacia) in 100 μl of buffer (20 mM Tris-HCI, pH7.4, 10mM $MgCl_2$, 120 μM each of dATP, dCTP, dGTP and dTTP) for 20 minutes at 37°C. DNA was then extracted one time with buffered phenol/chloroform (PC9, 3 parts phenol: 2 parts chloroform: 2 parts 500 mM Tris-HCl, pH 9.0, 20 mM EDTA, 10mNaC1) and precipitated with 2.5 volumes of ethanol in the presence of 2.5M NH_4Ac and 40 μg glycogen (Boehringer Mannheim). The DNA was then ligated to 30 μg of 'catch linkers' using ten units of T4 ligase (BRL) in 100 μl of ligase buffer supplied with the enzyme. The catch linkers consisted of 15 μg of phosphorylated catch A oligomer (5'-GAGTAGAATTCTAATATCTC-3') and 15 μg of phosphorylated catch B oligomer (5'-GAGATATTAGAATTCTACTC-3').

To cleave 'catch linkers' ligated to themselves, the ligation was diluted to a final volume of 1200 μl in buffer (50 mM Tris-HCl, pH 8.0, 50 mM NaC1, 10 mM $MgCl_2$, 100 μg/ml acetylated BSA [BRL]) and incubated for two hours at 37°C with 1,130 units XhoI (BRL). After restriction, the 'catch-linked' DNA was extracted with PC9 and ethanol precipitated as described above.

Immune precipitation of TFIIIA/DNA complexes

TFIIIA was prepared from 7S particles (a gift from D. Brown) by treating with RNase A (0.5 μg RNase A/5 μg 7S particles) for five minutes at room temperature in 50 μl of DNA binding buffer (50 mM Hepes [N-[2[Hydroxyethyl] piperazine-N-[2-ethanesulfonic acid]], pH 7.5, 50 mM KCl, 5 mM $MgCl_2$ 10 μM $ZnSO_4$, 1 mM dithiothreitol, 20%
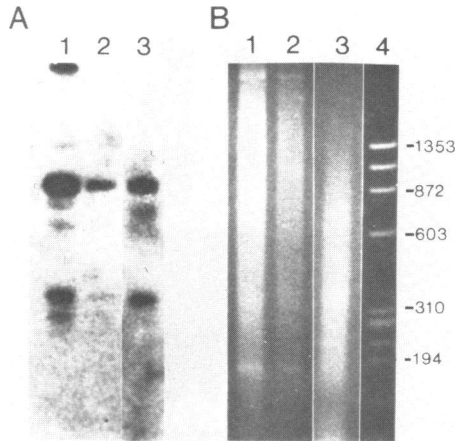
Figure 2: *Southern blot analysis demonstrating the feasibility of whole genome PCR.* Panel A, Southern blot analysis using a RB cDNA as a hybridization probe; panel B, Ethidium bromide staining pattern of gel analyzed in panel A. Lane 1, 4.5 µg of MboI cleaved human lymphocyte DNA; lane 2, 0.5 µg of MboI cleaved human lymphocyte DNA; lane 3, DNA from 0.5 ng of 'catch-linked' human lymphocyte DNA after an approximately 5,000 fold amplification by PCR and MboI cleavage (see Materials and Methods); lane 4, 250 ng of HaeIII cleaved PhiX DNA markers (BRL). The sequences detected in the Southern blot are present at one copy per haploid genome.

glycerol). DNA binding was performed in 20 µl of DNA binding buffer with 400 ng TFIIIA, 500 ng 'catch-linked' sonicated human DNA and 2 µg of poly (dI-dC) (Pharmacia) as competitor. The binding and all subsequent steps were carried out at room temperature. After incubating 30 minutes, 200 µl of immune precipitation buffer (50 mM Hepes, pH 7.5, 150 mM KCl, 5mM MgCl$_2$, 10 µM ZnSO$_4$, 1% Triton X-100, 0.05% SDS) and 5 µl of rabbit anti-TFIIIA antisera (a gift from D. Brown) were added. After incubating an additional hour, the mixture was transferred to a new tube containing 7.5 mg of Protein-A-Sepharose (Sigma). The protein-A-Sepharose (100 mg) was hydrated in 15 ml TN buffer (10 mM Tris-HCl pH 7.5, 150 mM NaCl) for 15 min. at 4°C and washed two times with TN buffer before using. Protein-A-sepharose incubations were carried out for one hour at room temperature with end-over-end mixing. The bound complexes were recovered by centrifigation for one minute in a microfuge and washed two times with 1.5 ml of immune precipitation buffer. DNA was freed from the complex by incubating 50 min. at 50°C in dissociation buffer (500 mM Tris-HCl, pH 9.0, 20 mM EDTA, 10mM NaCl, 0.2% SDS). Protein-A-sepharose was removed by centrifigation and the supernatant containing DNA was extracted with PC9 and ethanol precipitated as described above. Various amounts of the recovered DNA (1/2, 1/20, 1/200) were amplified by PCR using the catch linkers as primers. Reactions yielding 100−400 ng of DNA were then used as target DNA for another round of selection or to construct a library.

Polymerase Chain Reaction
PCR reactions (*3*) were carried out in 50 µl of PCR buffer (10 mM Tris-HCl, 50 mM KCl, 2.5 mM MgCl$_2$, 10 mM B-mercaptoethanol, 200 µM each of dATP, dCTP, dGTP and dTTP, 0.2 mg/ml gelatin) using 350 ng each of catch A and catch B oligomers and 2.5 units of Taq Polymerase (Cetus). A one min. incubation at 90°C, a two min. incubation
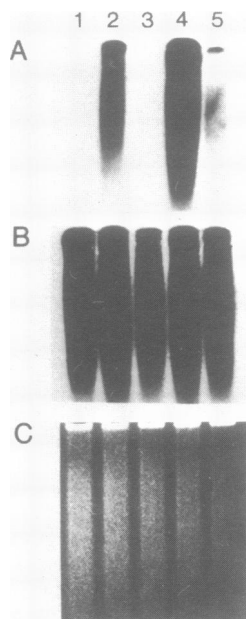
Figure 3: *Southern blot analysis demonstrating selection of 5S RNA related sequences by binding to TFIIIA.* Panel A, Southern blot using a probe for 5S RNA (pXbs201); panel B, same blot as in panel A hybridized to total human DNA as probe; panel C, ethiduim bromide staining pattern of gel analyzed in panel A and B. The following samples were amplified from catch-linked sonicated human DNA as described in Materials and Methods. Lane 1, amplified DNA before selection; lane 2, DNA selected (1×) with TFIIIA and anti-TFIIIA sera and then amplified; lane 3, DNA selected with TFIIIA and non-immune antisera and then amplified; lane 4, same preparation as in lane 2, further selected with TFIIIA and anti-TFIIIA sera and then re-amplified; lane 5, same preparation as in lane 2 further selected with TFIIIA and non-immune antisera and then amplified.

at 45°C and a two min. incubation at 70°C were repeated 25 times using a robot to move the reactions between water baths.

Southern Blot Analysis

For Southern blot analysis of MboI cleaved DNA, 1 ng of catch-linked MboI cleaved DNA was amplified by PCR as described above. One half of the PCR product (25 μl) was diluted to 100 μl in PCR buffer and amplified for two more cycles with 5 units Taq polymerase and 1.4 μg each of Catch A and Catch B oligomers. These additional two cycles were performed to ensure that the PCR product was double stranded. This product was then extracted with PC9 and ethanol precipitated as described above. The resulting DNA (1−3 μg) was cleaved with MboI, separated by electrophoresis through an agarose gel, transferred to nylon, and hybridized as previously described (4). The hybridization probe was the 3.8kb EcoRI restriction fragment of the RB cDNA clone pGEM-1-Rb generously provided by Wen-Hwa Lee (5).

For Southern blot analysis of sonicated DNA, 10 to 50 ng of previously amplified DNA was re-amplified for 25 cycles. Under these conditions, the capacity of the reaction to produce DNA is exceeded and generally uniform quantities of DNA were produced. One fifth of this DNA was separated by electrophoresis through an agarose gel, transferred to nylon and hybridized as described above. The hybridization probe was pXbs201,
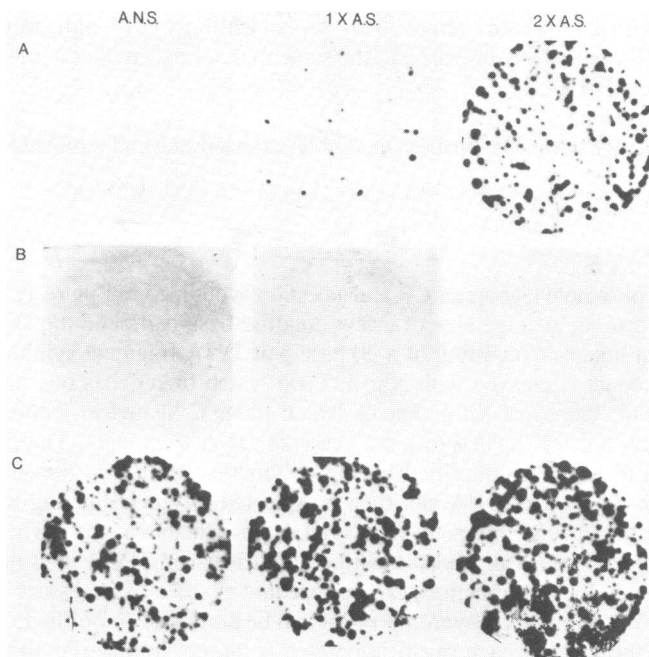
Figure 4: *Plaque hybridization demonstrating enrichment of selected sequences.* Row A, plaques lifts that were hybridized to twice amplified and selected sequences (2×A.S.); row B, the same lifts as in A re-hybridized to amplified non-selected sequences (A.N.S.); row C, the same lifts as in A re-hybridized to total human lymphocyte DNA. Column A.N.S., library constructed from amplified but non-selected sequences; column 1×A.S., library constructed from amplified selected sequences; column 2×A.S., library constructed from twice amplified and selected sequences; all plates contained approximately 5,000 recombinant phage clones.

containing a 249 bp HindIII/BamHI fragment of the somatic 5S RNA gene fom *X. borealis* generously provided by D. Brown (*6*).

Library construction and screening

PCR products to be used for cloning were extracted with PC9 and ethanol precipitated as described above. The amplified DNA was then cleaved with EcoRI, extracted with PC9 and precipitated with one-half volume of isopropanol in the presence of 0.5M NaClO$_4$. The isopropanol precipitation removes the linkers because DNA smaller than 200 bp is not precipitated under the conditions (*7*). For phage libraries, 1−5 ng of EcoRI cleaved amplified DNA was ligated to 500 ng of EcoRI-cleaved, dephosphorylated lambda gt10 arms (Promega Corporation) and packaged using commercially prepared packaging extracts (Gigapack Plus, Stratagene). Plaques were lifted onto nylon filters (Colony/Plaque Screen, NEN) and hybridized as previously described (*4*). For plasmid libraries, 1−5ng of EcoRI cleaved amplified DNA was ligated to 100 ng EcoRI cleaved, dephosphorylated pBluescript KS M13 + (Stratagene) and transfected into XL-1 Blue by the CaCl$_2$ method. Colonies were lifted onto nylon filters (*8*) and hybridized as previously described (*4*).

Probes for screening the library were prepared as follows: For amplified DNA, 10−50 ng of the PCR product was labeled by the random primer method in 25 μl (*9*), boiled for 10 min. in 200 μl of a solution containing 5×SSC and 1 mg/ml sheared human placental

DNA, and repeated sequences removed by pre-annealing for 15 min. at 68°C (*10*). A total human DNA probe was prepared in the same manner except that it was pre-annealed to salmon sperm DNA .

Nucleotide sequencing
Nucleotide sequencing was performed on double-stranded plasmid templates as previously described (*11*).

RESULTS

Whole genome PCR
The technique for Whole Genome PCR and selection is outlined in Figure 1. Total genomic DNA is cleaved to an average size of a few hundred base pairs and the DNA fragments ligated to 'catch linkers' consisting of a 20 base pair DNA fragment synthesized *in vitro*. The ligation product is cleaved with Xho I. Each 'catch linker' has one half of an XhoI site at its termini; therefore XhoI cleaves 'catch linkers' ligated to themselves but will not cleave 'catch linkers' ligated to most genomic DNA fragments. The linked DNA is now in a form that can be amplified by the polymerase chain reaction using the catch oligomers as primers. The DNA can then be selected (e.g. via binding to a protein or nucleic acid) and the bound fragments recovered by any suitable method. The small amount of DNA fragments specifically bound can be amplified using PCR, and these two steps (selection followed by amplification) can be repeated as often as necessary to achieve the desired purity. Because the recovered product can be amplified (*a*mplified, *s*elected DNA [A.S. DNA]), the yield at each purification step is largely immaterial and the steps can be chosen on the basis of the highest stringency or specificity available. Finally, the A.S. DNA fragments can be cleaved with EcoRI and cloned. Clones of interest are identified by hybridization of recombinants to radioactively labelled A.S. DNA.

To demonstrate the feasibility of Whole Genome PCR, total human genomic DNA was cleaved with MboI and ligated to catch linkers. The equivalent of 0.5 ng of genomic DNA was then amplified by PCR, cleaved with MboI and analyzed by Southern blot analysis (Figure 2). Figure 2A (lane 3) demonstrates that single copy sequences could be detected with as little as a 0.5 ng of starting DNA after an approximately 5,000 fold amplification by PCR. There appeared to be a bias toward amplifying smaller fragments (Figure 2A, lane 3; Figure 2B, lane 3) which was expected based on the properties of Taq polymerase and PCR.

Selection of sequences that bind TFIIIA
As test of this strategy, we attempted to identify the human DNA fragments which bind to TFIIIA purified from *Xenopus laevis* oocytes. TFIIIA is a prototype DNA binding protein; it contains nine copies of the zinc finger amino acid motif which mediate binding to 5S ribosomal RNA (5S rRNA) genes. Five-hundred ng of catch-linked DNA was bound to TFIIIA, and the bound DNA precipitated by incubation with anti-TFIIIA antiserum and protein-A sepharose. The amount of recovered DNA was too small to visualize upon agarose gel electrophoresis. When amplified with PCR however, the A.S. DNA could be visualized (Figure 3C, lane 2) and was found to be enriched for 5S rRNA gene related sequences by Southern blotting (Figure 3A, lane 2). Human 5S rRNA is similar in sequence to *Xenopus* 5S rRNA, and we therefore expected that enrichment of 5S rRNA genes would reflect a successful purification. After another cycle of binding and PCR, the enrichment for

```
OOCYTE   TGATCTCAGAAGCGATACAGGGTCGGGCCTGGTTAGTACCTGGATGGGAGACCGCCTGGGA

         ............................................................
CAT-1    TGATCTCAGAAGCGATACAGGGTCGGGCCTGGTTAGTACCTGGATGGGAGACCGCCTGGGA  100%

         ............................................................
CAT-4    TGATCTCAGAAGTGATACAGGGTCGGGCCTGGTTAGTACCTGGATGGGAGACCGCCTGGGA  98%

         ............................................................
CAT-13   TCTACTCAGAAGCGATACAGGGTCGGGCCTGGTTAGTACCTGGATGGGAGACCGCCTGGGA  95%

         ... .. .....................................................
CAT-16   TAATATCTCAAGCGATACAGGGTCGGGCCTGGTTAGTACCTGGATGGGAGACCGCCTGGGA  93%

         ...... ..........................   .......................
CAT6-4   TGATCTGTGAAGTGATACAGGGGCAGGCCTGGTTAGTACCTGGATGGGAGACCGCCTGAGA  90%

         ...... ..........................   .......................
CAT-12   TGATCTGTGAAGTGATACAGGGGCAGGCCTGGTTAGTACCTGGATGGGAGACCGCCTGAGA  90%

         .. ... ..........  .............   .........................  ..
CAT-9    TGAACTGTGAAGTGATACAAGGTCAGGCCTGGTTAGTACCTGGATGGGAGACCGCCTGAGA  89%
```

Figure 5:*Nucleotide sequence of TFIIIA binding sites.* The nucleotide sequence of the TFIIIA binding site in *Xenopus laevis* oocyte type 5S rRNA (*18*) is shown at the top of the figure (oocyte). Cat-1, Cat-4, Cat-9, Cat-12, Cat-13, and Cat-16 were clones isolated from a plasmid library by screening with 2×A.S. sequence as probe. Cat6-4 was plaque purified from a phage library and the insert subcloned into pBluescript for sequencing. Two dots above a base indicates identity to the TFIIIA binding site. The number to the right of the Cat clones indicate the percent identity of the clones to the *Xenopus* TFIIIA binding site defined previously (*12*). The inserts from Cat-1, Cat-4, Cat-9, Cat-12, Cat-13, Cat-16, and Cat6-4 were respectively 143, 121, 128, 100, 97, 88 and 198 base pairs in length. Only the TFIIIA binding site motifs within these clones are shown in the figure.

sequences hybridizing to 5S RNA genes increased significantly (Figure 3A, lane 4). Control experiments using normal rabbit serum instead of TFIIIA antiserum were used to show the specificity of the purification procedure (Figure 3, lane 3 & 5). From the intensity of hybridization, we estimated that this procedure resulted in an enrichment of approximately a thousand fold after two cycles of TFIIIA binding and amplification.

The enriched sequences could be cloned into lambda phage or plasmid vectors through EcoRI sites present in the catch linkers attached to each fragment (Figure 1). Any desired number of clones could be obtained, since a large amount of A.S. DNA could be produced. Because no selection step is perfect, however, we expected that some of the resultant clones should contain sequences binding to TFIIIA while others should contain random sequences that coincidentally copurified with the TFIIIA binding sequences. To identify clones containing TFIIIA binding sites, they were screened with a hybridization probe generated from A.S. DNA. Because A.S. DNA should be significantly enriched for TFIIIA binding sequences, this probe should preferentially hybridize to clones containing such sequences. A control probe consisting of PCR amplified but *not* TFIIIA *s*elected DNA (A.N.S. DNA) was used as a control for this hybridization.

The results in Figure 4A show that DNA selected and amplified through two cycles (2×A.S.) did not hybridize to any of 5,000 clones prepared from the A.N.S. DNA, but the 2×A.S. probe hybridized to approximately 1% of the clones prepared from 1×A.S. DNA, and to approximately 10% of the clones prepared from 2×A.S. DNA. When A.N.S. DNA was used as a probe, no hybridization was seen in any of the phage libraries (Figure 4B). Hybridization with a total human DNA probe indicated the presence of human DNA in clones prepared from all three DNA samples (Figure 4C). Thus the hybridization to

A.S. DNA effectively identified clones containing sequences which were enriched by the procedure.

Seven clones that hybridized to $2 \times$ A.S. DNA were isolated and sequenced. All seven clones contained sequences similar to the 61 base pair TFIIIA binding site (21) previously identified in *Xenopus laevis* oocyte type 5S rRNA genes (Figure 5). The average identity was 94% over the 61 bp region and ranged from 89% to 100%.

DISCUSSION

The procedure described above successfully identified human sequences containing the binding site for TFIIIA. Physical methods have been used before to identify DNA sequences that bind to specific proteins (e.g., *13,14*), and the new strategy should greatly facilitate such efforts in the future. The major advantages are (i) the purification process can be repeated several times without fear of exhausting product yield, and (ii) following cloning, the selected sequences can be identified by hybridization to a probe made from A.S. DNA. The success of the technique of course will depend on the quality of the reagents (i.e. DNA binding protein and antibody), the relative affinities of the protein for specific versus random DNA sequences, and the number of times the binding site is repeated within the genome. It should be noted that most DNA binding proteins yet identified, like TFIIIA, bind to consensus sequences which are repeated in the genome. The total number of TFIIIA binding sequences present in the human genome is not known for certain, although previous estimates suggest that there are approximately 2×10^3 human 5S RNA genes (*15*). Moreover, the binding of TFIIIA to its recognition site is relatively weak [Kd = 10^{-9} (*16*)]. Therefore, the demonstrated success of the procedure for isolating TFIIIA binding sequences suggests that it will also be applicable to other DNA binding proteins.

Finally, the general strategy outlined in Figure I could also be used, with minor modifications, to identify and clone DNA sequences selectable by other methods. One particularly important example would be the identification of sequences that are deleted in inherited diseases or in neoplastic cells. Methods for selection of human DNA sequences that do not hybridize to the DNA from normal cells have been reported (*17*). The identification of such sequences might be facilitated by employing the procedures described here, substituting hybridization selection for protein binding in the selection step.

ACKNOWLEDGEMENTS

REFERENCES

1. Wingender, E. (1988) Nucleic Acids Research 16, 1879−1902.
2. Wolffe, A.P. and Brown D.D. (1988) Science 241, 1626−1632.
3. Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B. and Erlich, H.A. (1988) Science 239, 487−491.
4. Ruppert, J.M., Kinzler, K.W., Wong, A.J., Bigner, S.H., Kao, F., Law, M.L., Seuanez, H.N., O'Brien, S.J. and Vogelstein, B. (1988) Molecular and Cellular Biology 8, 3104−3113.
5. Lee, W.H., Bookstein, R., Hong, F., Young, L.J., Shew, J.Y., Eva, Y.-H and Lee, P. (1987) Science 235, 1394−1399.
6. Bogenhagen, D.F., Sakonju, S. and Brown, D.D. (1980) Cell 19, 27−35.

7. Haymerle, H., Herz, J., Bressan, G.M., Frank, R. and Stanley, K. (1986) Nucleic Acids Research 14, 8615−8624.
8. Buluwela, L., Forster, A., Bochm, T. and Rabbits, T.H. (1989) Nucleic Acids Research 17, 452.
9. Feinberg, A.P. and Vogelstein, B. (1984) Analytical Biochemistry 137, 266−267.
10. Sealey, P.G., Whittaker, P.A. and Southern, E.M. (1985) Nucleic Acids Research 25, 1905−1922.
11. Kraft, R., Tardiff, J., Krauter, K.S. and Leinwand, L.A. (1988) Biotechniques 6, 544−549.
12. Sakonju, S. and Brown, D.D. (1982) Cell 31, 395−405.
13. Pollwein, P., Wagner, S. and Knippers, R. (1987) Nucleic Acids Research 16 9741−9759.
14. Biedenkapp, H., Borgmeyer, U., Sippel, A.E. and Klempnauer, K.-H. (1988) Nature 335, 835−837.
15. Hatlen, L. and Attardi, G. (1971) Journal of Molecular Biology 56, 535−553.
16. Hanas, J.S., Bogenhagen, D.F. and Wu, C.-W. (1983) Proc. Natl. Acad. Sci. 80, 2142−2145.
17. Kunkel, L.M., Monaco, A.P., Middlesworth, W., Ochs, H.D. and Latt, S.A. (1985) Proc. Natl. Acad. Sci. 82, 4778−4782.
18. Wolters, J. and Erdmann, V.A. (1988) Nucleic Acids Research, 16, Supplement r1−r70.

This article, submitted on disc, has been automatically converted into this typeset format by the publisher.