© 1997 Oxford University Press   Nucleic Acids Research, 1997, Vol. 25, No. 4   781–786

Libraries for genomic SELEX

Britta S. Singer[1], Timur Shtatland[1], David Brown[1,+] and Larry Gold[1,2,*]

[1]Department of Molecular, Cellular, and Developmental Biology, University of Colorado, Boulder, CO 80309-0347, USA and [2]NeXstar Pharmaceuticals, Inc., 2860 Wilderness Place, Boulder, CO 80301, USA

Received October 23, 1996; Revised and Accepted December 17, 1996

ABSTRACT

An increasing number of proteins are being identified that regulate gene expression by binding specific nucleic acids *in vivo*. A method termed genomic SELEX facilitates the rapid identification of networks of protein–nucleic acid interactions by identifying within the genomic sequences of an organism the highest affinity sites for any protein of the organism. As with its progenitor, SELEX of random-sequence nucleic acids, genomic SELEX involves iterative binding, partitioning, and amplification of nucleic acids. The two methods differ in that the variable region of the nucleic acid library for genomic SELEX is derived from the genome of an organism. We have used a quick and simple method to construct *Escherichia coli, Saccharomyces cerevisiae*, and human genomic DNA PCR libraries that can be transcribed with T7 RNA polymerase. We present evidence that the libraries contain overlapping inserts starting at most of the positions within the genome, making these libraries suitable for genomic SELEX.

INTRODUCTION

Interactions of proteins with DNA and RNA are at the heart of gene expression regulation. It has become clear that this regulatory network is intricate, and that we are only starting to understand its full scope. In addition to proteins for which binding nucleic acids seems to be the primary function *in vivo*, other proteins have dual functions of which one is the capacity to bind to nucleic acid [for instance, (1–3) and see the lists in (4,5)]. Some of these proteins are involved in gene regulation, while the function of nucleic acid binding in others remains unknown. Several proteins that bind two or more different nucleic acids are involved in gene regulation [e.g., (3,6,7)]. Undoubtedly, there are other protein–nucleic acid interactions that have yet to be identified. For most proteins, RNA ligands can be selected that bind with nanomolar affinities (affinities that are certainly high enough to elicit a response *in vivo*). Among these proteins are many not thought of as RNA- or DNA-binders (8). Based on this evidence, we have suggested (9) that a wide range of proteins affect gene expression by interacting with nucleic acids *in vivo*; indeed, we hypothesized that a complete description of the workings of the cell must include a 'linkage map' that describes the interactions between proteins and nucleic acids in the life of the cell. The discovery of interactions like these requires a global search method, a method such as genomic SELEX.

Genomic SELEX is an extension of SELEX (8,10,11). In SELEX, nucleic acids that bind tightly to a protein of interest are identified through successive rounds of binding, partitioning and amplification. In SELEX as originally developed, the library contains $10^{14–15}$ random sequences. PCR amplification requires that the nucleic acid sequences of interest be flanked by fixed sequence primer annealing sites. A T7 promoter is included in one of the primer annealing sites so that the library can be expressed as RNA.

In genomic SELEX, the libraries contain sequences derived from the genome of the organism of interest flanked by fixed regions that allow PCR amplification and transcription. The success of genomic SELEX is critically dependent on the quality of the starting library. Ideally, the library should be fully representative of the genome of interest and the various genomic inserts should be equally represented. In this article we present a method of library construction, and two independent methods to test library quality. We also present the results of these tests on genomic libraries that we constructed from human, *Saccharomyces cerevisiae*, and *Escherichia coli* DNA.

MATERIALS AND METHODS

Library construction

DNAs from human placenta (Type XIII) and *E.coli* B were purchased from Sigma. *Saccharomyces cerevisiae* DNA was purified from strain S288C using equilibrium density gradients in CsCl (12).

Figure 1 provides an overview of library construction. We used random priming on denatured genomic DNA that had been sheared by sonication only enough to reduce viscosity. Primer B_{ran} (12 μM final concentration) and genomic DNA (0.17 mg/ml final concentration, 25 mg total) were mixed and incubated at 93°C for 3 min, then quickly chilled on ice. Klenow (0.9 U per ml final concentration) and 300 μM dNTPs (final concentration) were added and the reaction was incubated on ice for 5 min, at 25°C for 25 min and at 50°C for 5 min. The low temperature step facilitates annealing of the primer's random nine nucleotides, while the 50°C step allows Klenow extension through hairpins in single-stranded DNA. Four successive spins through Microcon-10 filters (Amicon, MA) removed ~60% of the primers from the first step. [Kirk Jensen, personal communication, reported that Microspin S-400 HR Columns (Pharmacia) removed 95% of the primers during construction of a library using the method described here.] Second strand synthesis with

*To whom correspondence should be addressed. Tel: +1 303 546 7605; Fax: +1 303 546 7603; Email: gold@nexstar.com

+Present address: Ambion, Inc., 2130 Woodward St., Suite 200, Austin, TX 78744, USA

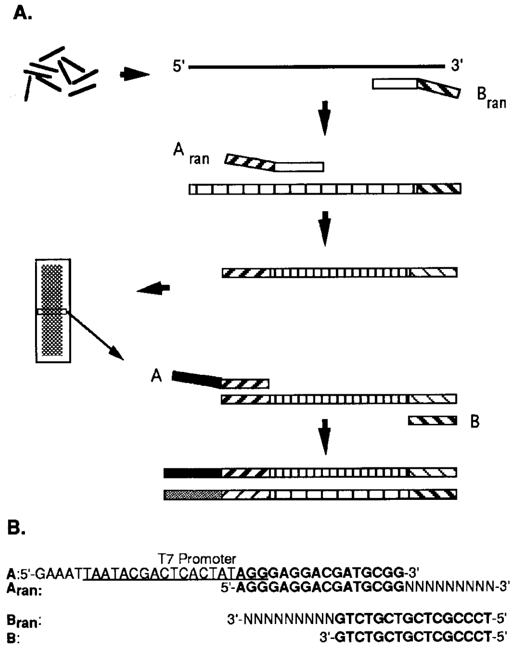
A.
5'
3'
B ran
A ran
A
B
B.
T7 Promoter
A:5'-GAAATTAATACGACTCACTATAGGGAGGACGATGCGG-3'
Aran: 5'-AGGGAGGACGATGCGGNNNNNNNNN-3'
Bran: 3'-NNNNNNNNNGTCTGCTGCTCGCCCT-5'
B: 3'-GTCTGCTGCTCGCCCT-5'

Figure 1. (**A**) Diagram of library construction. The upper left corner depicts sheared, denatured genomic DNA (bold lines). The arrows indicate the subsequent steps in library construction. The next panel shows annealing with primer B_{ran}. The fixed sequence is crosshatched, and the random nine nucleotides are represented by a white box. The newly synthesized strand (vertical stripes) in the next picture serves as template for second strand synthesis primed by A_{ran}. Here again, the white box represents the random nucleotides and the crosshatched section is fixed sequence. The products of that reaction are run on a denaturing gel and various fractions are electroeluted. The desired template molecule is shown with the fixed sequence of primer A_{ran} (crosshatched) and the complement of the fixed sequence of primer B_{ran} (lighter crosshatching) flanking the new template strand (closely spaced vertical lines). Amplification with primer A adds a T7 promoter (black). The resulting double-stranded library molecules have the form shown in the final picture: inserts (vertical lines) flanked by priming sequences (the complements of primers are depicted as lighter than the primers). (**B**) Sequences of primers used in library construction. The sequences common to primers A and A_{ran} and to B and B_{ran} are in bold. The T7 promoter is underscored. We have made one yeast library using different fixed sequences; see Materials and Methods.

primer A_{ran} was the same as first strand synthesis. The reaction products were separated on a denaturing gel and fractions of various sizes (with genomic inserts ranging from about 40–700 nucleotides) were electro-eluted according to the protocol provided by Isco, Inc. (Lincoln, NE). The yield was ≥1 pmol of each fraction. Since the human genome has about 6×10^9 nucleotides, each position potentially served as the starting nucleotide of an insert of every given size in ≥100 molecules of the human library (and correspondingly more for the less complex genomes of *E.coli* and *S.cerevisiae*). Next we amplified the library by PCR using primers of completely fixed sequences, one of which adds a T7 promoter (primer A); thus the library can be expressed as either RNA or DNA.

Removing primer B_{ran} with Microcon filters reduces the fraction of the molecules that have primer B_{ran} at both ends, and the extension protocol should not give rise to primer A_{ran} at both ends; nonetheless, some of the library molecules did in fact contain the same priming site at both ends. These unwanted molecules were eliminated from the population by a single cycle of transcription (which requires the T7 promoter from primer A), reverse-transcription (which requires primer B) and PCR.

All of the libraries described here have been generated using the primers shown in Figure 1 except for one yeast library. The promoter-containing primer (A) in this library is 5′-GAAATTAATACGACTCACTATAGGGAGACAAGAATAAACGCTCAA-3′ (promoter underscored) and primer B is 5′-GCCTGTTGTGAGCCTCCTGTCGAA-3′.

PCR

Primers. Single copy gene primer sequences were selected from GenBank version 90.0 (8/95). We chose primer sequences with predicted annealing temperatures close to 72°C, and whenever possible, within 1°C of each other. We used 72°C as the PCR extension temperature. Thus we were often able to eliminate the annealing step, and we may have additionally gained a measure of specificity of priming. We calculated annealing temperatures in degrees centigrade (T_A) by the formula of Wu *et al.* (13):

$$T_A = 1.46(A + T + 2C + 2G) + 22,$$

where A, T, C and G are the numbers of the corresponding nucleotides in the primer.

Primers were obtained from Operon (CA). Biotinylated primers were synthesized incorporating three ‘biotin-ON’ phosphoramidites at the 5′ end.

Reactions. A typical 100 µl reaction contained 1–10 µM of each of the two primers, 50 mM KCl, 10 mM Tris–HCl, pH 9.0 (25°C), 0.1% Triton X-100, 3 mM $MgCl_2$, 50 µM each of the four dNTPs, 0.05 U/ml Taq DNA polymerase. PCR mix without dNTPs and polymerase was assembled at room temperature and overlaid with 50 µl of mineral oil. It was heated to 95°C for 5 min, then dNTPs and polymerase were added. Thereafter cycling proceeded for the requisite number of cycles of 93°C for 30 s and 72°C for 45 s (plus a 10 s annealing step in between, if necessary, at a temperature determined by the above formula).

Distribution of the end-points of genomic inserts

We used a biotinylated genome-specific primer and a library primer (A or B) to generate by PCR a set of overlapping ‘sub-fragments’ that contain only one fixed library primer annealing site and variable extents of genomic insert (Fig. 2B).

Because the library primer anneals to one strand of *every* double-stranded library molecule, the reaction that yields exponential amplification of the desired set of overlapping sub-fragments simultaneously yields linear amplification of one strand of every molecule in the library. The number of cycles (n) required for the desired exponentially amplified material to approach the quantity of the undesired linearly amplified molecules can be calculated by solving for n in the following equation:

$$\begin{array}{ccc} \text{Linear} & & \text{Exponential} \\ 2N + 2N(A–1)n & = & sA^n \end{array}$$

Given N base pairs in the genome, there is a maximum of 2N distinct fragments in the library (one fragment starting at each position). The number of sub-fragments (s) that can be exponentially amplified is a function of both the size of the genomic library insert and the size of the genome-specific primer. A represents the amplification occurring during each PCR cycle; for complete doubling, A = 2; in practice 1.5 < A < 2. Thus, for the human genome (N = 3 x 10^9), a library with 60 amplifiable

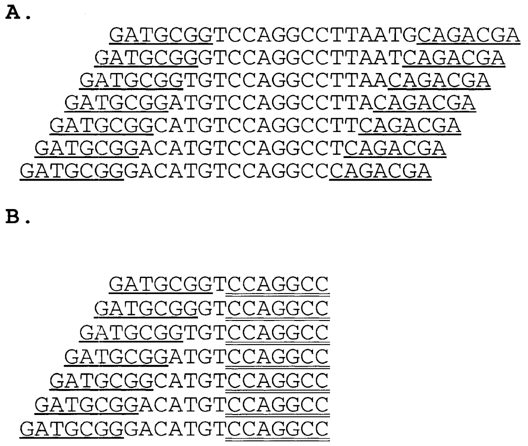
A.
GATGCGGTCCAGGCCTTAATGCAGACGA
GATGCGGGTCCAGGCCTTAATCAGACGA
GATGCGGTGTCCAGGCCTTAACAGACGA
GATGCGGATGTCCAGGCCTTACAGACGA
GATGCGGCATGTCCAGGCCTTCAGACGA
GATGCGGACATGTCCAGGCCTCAGACGA
GATGCGGGACATGTCCAGGCCCAGACGA
B.
GATGCGGTCCAGGCC
GATGCGGGTCCAGGCC
GATGCGGTGTCCAGGCC
GATGCGGATGTCCAGGCC
GATGCGGCATGTCCAGGCC
GATGCGGACATGTCCAGGCC
GATGCGGGACATGTCCAGGCC

Figure 2. (**A**) Idealized representation of library molecules derived from one genomic site, flanked by seven nucleotide fixed sequences, underscored. In the realistic library, the insert sizes vary somewhat, but ideally every genomic nucleotide is present as the 'first' and 'last' nucleotide of human sequence adjacent to the fixed sequences (the primer annealing sites; see Fig. 1) in inserts of a given length. (**B**) A set of sub-fragments generated during analysis of the distribution of end points in a library. Varying portion of the insert are flanked by one genome-specific primer (double underscored) and one library primer (underscored). Since the genomic insert in the library shown is 14 nucleotides (see A) and the primer is 7 nucleotides, the maximum number of different sub-fragments that can be generated here is 7. In practice, we used two nested genome-specific primers, and thus the number of sub-fragments generated is determined by the positions of both genome-specific primers, since the insert must contain priming sites for both of them in order for the sub-fragment to be amplified in this experiment (see Materials and Methods).

sub-fragments (s = 60), and 1.5-fold amplification per cycle, 54 cycles (n) are required for the desired exponentially amplified material to approximate the number of the undesired linearly amplified molecules. If A = 2, n is decreased to 32. In practice, the number of cycles used was somewhere between these two values.

PCR products that contained the genome-specific biotinylated primer were batch-purified on streptavidin beads. Two hundred microlitres of drained ImmunoPure immobilized streptavidin (Pierce, IL) were washed in an Eppendorf tube 5 times with 500 µl binding buffer (50 mM NaCl, 10 mM Tris–HCl, pH 7.5, 1 mM EDTA). One hundred microlitres of the chloroform-extracted PCR products were mixed with 500 µl binding buffer and then added to streptavidin beads. Binding was carried out at room temperature with rocking for 30 min. Unbound DNA was removed by washing 4 times with 500 µl binding buffer. DNA complementary to the biotinylated strand was eluted by denaturation at 37°C for 15 min with 400 µl of 0.15 M NaOH and then for 5 more min with another 500 µl of 0.15 M NaOH. NaOH was neutralized with an equimolar amount of acetic acid. DNA was ethanol precipitated using 40 µg glycogen as a carrier.

The PCR/streptavidin purification steps were done a total of 2–3 times in order to isolate the desired set of sub-fragments. Generally, after the first PCR, the native polyacrylamide gel showed that the predominant product was the size of the intact library plus some of even lower mobility. After the second PCR/streptavidin purification, the products generally formed a smear of approximately the predicted size range (e.g. predicted 70–90 base pairs, observed 50–100 base pairs). The second or third PCR was carried out using a nested genome-specific primer. For example, genome-specific primers for *E.coli metB* sub-fragments were, for PCR 1, BBBAATGTCAGGCACCAGAGTAAA, for PCR 2, BBBACCAGAGTAAACATTGTGTTAAT. Here, B stands for biotin, and the shared sequence isunderscored.

The final PCR amplification of this material used uracil-primers for cloning into the CloneAmp pUC18 system (Gibco BRL, MD). In some cases, the uracil-primer was the nested primer. We sequenced the plasmids by the Sanger method.

Single copy gene PCR analysis of the library

We chose primers predicted to anneal only to single copy genes and carried out PCR as described in the section 'PCR' above. We typically used 3 pmol library DNA and a corresponding amount of genomic DNA. For instance, in a library with genomic insert DNA of length 60, 3 pmol library molecules contains 120 ng insert. Hence for the control we used 120 nge genomic DNA from which the library was made.

RESULTS

Library construction

Genomic libraries have been generated previously using a variety of methods, including restriction digestion and ligation (14), mechanical fragmentation and blunt-end ligation (15), mechanical fragmentation and enzymatic 'tailing' (He *et al*., in preparation), and PCR amplification using a single primer with a fixed 5′ end and random bases at the 3′ end (16), the method most similar to that reported here.

Figure 1 shows the approach we used to construct our libraries. Human, yeast, or *E.coli* genomic DNA was denatured and annealed to an oligo with nine random nucleotides at the 3′ end and a fixed sequence at the 5′ end. After annealing, which ideally is distributed randomly, the oligo was extended with Klenow. Another randomized oligo with a different 5′ fixed sequence was added to the products of the first reaction, annealed and extended in the same way. We ran the extended reaction products on a denaturing gel to fractionate by size. Each fraction became the basis of a library with a different length of genomic insert. The library was completed by PCR amplification that added a T7 promoter to one of the primer annealing sites.

Distribution of the end-points of genomic inserts

The library should contain an overlapping set of inserts for every segment of the genome (Fig. 2A). In order to test this notion, we developed a novel technique that allows us to examine in detail the sequences of such overlapping inserts. This method shows us the distribution of end-points in a specific region of genomic sequence, and allows us to determine the sequence fidelity of the library, both within and outside of the nine base pair region derived from random priming during library construction. Figure 2B shows a hypothetical set of overlapping sub-fragments that would be generated during this analysis. They are called 'sub-fragments' because they include only one end of the corresponding library fragments (Fig. 2A). Each sub-fragment has one library and one genome-specific primer annealing site. During the PCR that generates this set of sub-fragments, the library primer amplifies (linearly) every molecule in the library. Thus, it is necessary to biotinylate the genome-specific primer so that the desired sub-fragments can be isolated by binding to streptavidin-coated beads. Additional cycles of PCR using a 'nested' genome-specific primer eliminate any remaining background.

Table 1. Analysis of the nine positions adjacent to the 5′ fixed primer annealing site in various libraries

Organism	exo	Gene	No.	%	Percent incorrect at each position								
			tested	correct	9	8	7	6	5	4	3	2	1
E.coli	–	met B	60	8	0	12	13	13	13	21	28	46	43
Yeast	–	NDC1	44	18	2	5	5	7	12	21	24	30	39
Yeast	–	NDC1*	49	32	2	4	8	6	16	24	20	36	32
Human	+	ada	43	49	0	0	9	12	10	15	0	14	33
Human	+	U1A	43	44	0	0	14	3	6	0	10	23	37

We aligned sub-fragments by the fixed primer annealing sites and looked at the nine nucleotides of the insert immediately adjacent. These nine nucleotides are involved in random priming during library construction. The numbering is from the fixed sequence toward the middle of the molecule, thus position 1 is closest to the fixed sequence. Both yeast libraries and the *E.coli* library were constructed with Klenow fragment lacking the 3′ exonuclease, as indicated. The column '% correct' displays the percentage of clones that are matched to genomic sequence in *all* positions. The isolated sub-fragments vary in size, as predicted (see Fig. 2). All of the sub-fragments reported have at least one nucleotide of genomic sequence in addition to the genomic primer sequence. Sub-fragments with fewer than nine nucleotides between the library primer annealing site and the genome specific priming site do not yield information about all positions; thus fewer than the total 'No. tested' may contribute to the percentages reported for these positions. Several of the sub-fragments examined give a better match to genomic sequence if gaps or bulges are allowed in one or both sequences. Since our analysis did not allow gaps and bulges, it reports a higher percentage of incorrect bases in the positions closer to the fixed sequence than would otherwise be the case. When gaps and bulges are allowed, usually no sub-fragment had more than three mismatches between library and genomic sequence. *This library uses different fixed sequences from the others (see Materials and Methods).

The end-points of genomic insert sequences in sub-fragments isolated from single copy genes in four libraries are shown in Figure 3. While we did not find an end-point at every possible position, well over half of them are represented. The largest region in which we found no end-points is only nine nucleotides long. This expanse is small relative to the size of the insert size to be chosen for most purposes.

We examined the nine nucleotides adjacent to primer A in libraries from all three organisms. These nine nucleotides are involved in annealing to the randomized sequence of the primer during library construction. As primer annealing at low temperature is imprecise, we expected misannealing to generate mutations in this region. The results are shown in Table 1. The human library was generated using the Klenow fragment of *E.coli* DNA polymerase I with intact 3′ exonuclease, the so-called proofreading exonuclease (exo$^+$). The yeast and *E.coli* libraries were made with Klenow lacking that exonuclease (exo$^-$). As was expected, the library generated with exo$^+$ Klenow is more accurate in these positions. The downside of the exo$^+$ polymerase is that it might yield a greater over-representation of molecules with genomic inserts adjacent to regions that are similar in sequence to the fixed regions of the library primers, since the entire random region could be removed by this enzyme, leaving only annealing of the fixed region to genomic DNA.

Incorrect bases in the priming region reduce the effective size of the genomic insert. (By 'insert' we mean the part of the library molecule between the two fixed sequences.) The effective size of the insert should be reduced by 4 ± 4 nucleotides in the human library that we made, and by 6 ± 6 nucleotides in the *E.coli* library. In practice, it is wise to work with a library that has inserts long enough to make mistakes in the priming regions irrelevant.

This method of testing the library also shows how well the sequences of library inserts match the published genomic sequence. In the region excluding the nine nucleotide stretches adjacent to the library primers, two of the 60 *E.coli* sub-fragments sequenced had one point mutation each, whereas in the human library four out of 86 sub-fragments sequenced had one point mutation each. Thus the libraries that we have generated are sufficiently accurate for use in genomic SELEX.

Single copy gene PCR analysis of the library

We also tested the library quality with a more traditional method, using PCR to amplify various genomic segments. Each amplified segment spans almost the entire genomic insert length of the library. If some genomic segment is missing from the library because of the way the library was constructed, it will be amplified from the original genomic DNA, but not from the library DNA. With all four libraries, the observed size of the PCR products was as predicted from the GenBank sequences (Fig. 4). Because the library contains overlapping inserts, and because the size of genomic segments amplified in these experiments approximates the size of the insert, most of the molecules with one genome-specific priming site do not contain the other genome-specific priming site. Thus, for a given genomic region, most library molecules are not PCR-amplifiable. As expected, the yield was lower with the library DNA as a template than with the genomic DNA (from which library was made), under otherwise identical conditions.

Escherichia coli library. A total of 13 tested segments, 60 base pairs each, were all amplified both from the *E.coli* genomic DNA and from the library. Five segments were amplified from the dam (DNA adenine methylase) gene, four segments from the *bgl* operon (involved in utilization of sugars, beta-glucosides), and one segment each from *metB* (involved in methionine biosynthesis), the *ilvGMEDA* operon (involved in isoleucine/valine biosynthesis), *corA* (Mg^{2+} transport protein), and the ribosomal RNA gene (this segment is the only one tested that is not from a single copy gene).

We were concerned that sequences predicted to form stable hairpins in ssDNA might result in the under-representation in the completed library of molecules that include those sequences or that are adjacent to them. Such a hairpin could obstruct primer annealing or extension during library construction. To address this question, we amplified an rRNA gene segment inside a region that is predicted to form a long hairpin, thus possibly preventing random primer annealing during library construction. Based on the number of cycles it takes to amplify the segment from the

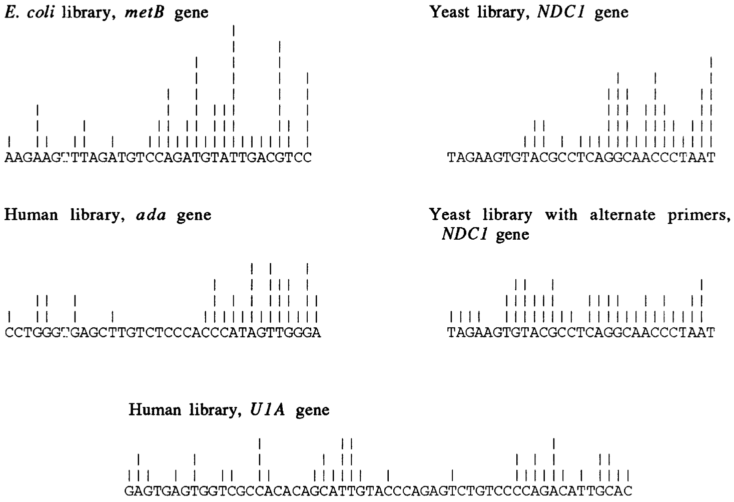
E. coli library, *metB* gene
AAGAAGTTTAGATGTCCAGATGTATTGACGTCC
Yeast library, *NDC1* gene
TAGAAGTGTACGCCTCAGGCAACCCTAAT
Human library, *ada* gene
CCTGGGTGAGCTTGTCTCCCACCCATAGTTGGA
Yeast library with alternate primers, *NDC1* gene
TAGAAGTGTACGCCTCAGGCAACCCTAAT
Human library, *U1A* gene
GAGTGAGTGGTCGCCACACAGCATTGTACCCAGAGTCTGTCCCCAGACATTGCAC

Figure 3. Distribution of end-points of genomic inserts. The sub-fragments whose end-points are shown here are the same as those used for the analysis shown in Table 1. An end-point is the last genomic nucleotide in each library sub-fragment. The sub-fragments with end-points at a genomic position are indicated by| above this position; the number of |s is equal to the number of sub-fragments with the same end-point. The sequences shown are the sense strand, displayed 5′ to 3′. The 3′ nucleotide shown is adjacent to the internal (nested) genome-specific primer used for the analysis. Analyses of end-points in the human*ada* and *U1A* genes used two different primer sets each; in both cases, one primer is partially included in the sequence shown.

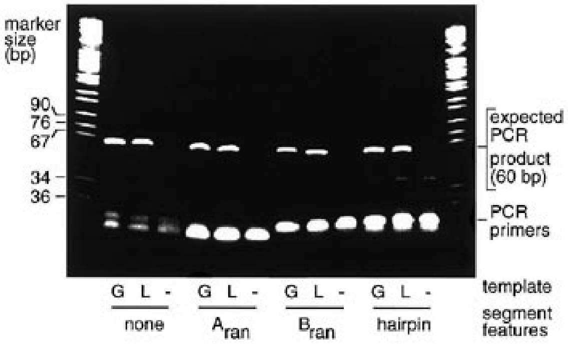

Figure 4. PCR amplification with genome-specific primers of *E.coli* genomic DNA (G), library DNA (L) and no DNA (–). In each set (G, L, –) the PCR was carried out for the same number of cycles. Segment features: none – *dam* gene segment; A_{ran} – segment downstream from the site in the genome similar to library primer A_{ran} fixed sequence, therefore potentially over-represented in the library; B_{ran} – same as above, for library primer B_{ran}; hairpin – rRNA gene segment located within a long hairpin predicted to form when the library is made; therefore potentially under-represented in the library.

library DNA to approximately the same level as from the genomic DNA, the sequence does appear to be somewhat under-represented in the library (data not shown).

Another potential problem is over-representation of molecules with insert sequences adjacent to regions that are able to pair with the fixed regions of the library primers. Because not only the randomized part of the primer, but also the fixed part can anneal to these sites, inserts downstream from these sites could be over-represented in the library. The predicted annealing sites for the fixed sequences were found by FASTA (17,18) search of the available *E.coli* genome. The *ilvGMEDA* and *corA* segments are both downstream from the predicted annealing sites for the fixed sequences of the primers used to construct the library (primers A_{ran} and B_{ran}, respectively, see Fig. 1). Our PCR experiments indicate that they might indeed be somewhat over-represented (data not shown). Both this over-representation and the under-representation discussed in the preceding paragraph are too small to be seen in Figure 4.

Yeast library. PCR primers designed to amplify 140–150 nucleotide segments of seven different *S.cerevisiae* single copy genes (*MPS1*, *MPS2*, *NDC1*, *MOB1*, *MOB2*, *GSP1*, and *YTX1*) were used to test the quality of the yeast genomic library. All seven segments were amplified both from the genomic DNA and from the library. This confirms, at least at the level of this analysis, that the libraries contain the same DNA segments as the genome from which they were derived.

Human library. Seven out of 10 segments tested were amplified from both library and genomic DNAs. Four *ada* (adenosine deaminase) gene segments were amplified. One *ada* segment was amplified from neither genomic DNA nor from the library; this result probably indicates that the optimal PCR conditions for their amplification have yet to be determined. Three segments from mitochondrial ATPase subunit 6 gene were amplified from both library and genomic DNAs and two adjacent segments were amplified from neither. No special efforts were made to remove mitochondrial DNA from the starting genomic DNA preparation; however, it is possible that mitochondrial DNA was absent from our library and that the three ATPase segments were amplified *only* because of their homology to some nuclear DNA.

DISCUSSION

Libraries such as the ones we describe in this paper may be useful for a variety of genomic SELEX applications, e.g., to find RNA or DNA that binds a particular protein or another biologically important molecule of interest, such as an antibiotic, a cofactor, a mono- or polysaccharide, or to find RNA or DNA that is cleaved or otherwise covalently modified by a protein, a metal ion or any other molecule. The described method of library construction by randomized primer extension is easy and robust (it has worked for human, *S.cerevisiae*, and *E.coli* DNA). Although we have constructed only genomic DNA libraries, the method is adaptable to cDNA as well.

By PCR with genomic primers, we amplified from the library DNA template *all* segments that we amplified from genomic DNA. No specific loss of segments can be attributed to the library construction process.

A 'perfect' library has molecules with inserts of a given size that begin at every nucleotide of the genome. Because of this, distribution of end-point analysis provides a more sensitive and rigorous test of the library quality. We have shown that the libraries reported here are virtually complete. If we had sequenced additional sub-fragments, it is likely that we would have discovered additional end-points. However, even if inserts with the end-points that we failed to find are indeed missing in their respective libraries, the libraries are sufficiently comprehensive to contain every genomic binding site represented by many distinct inserts in an appropriately long library. Not all positions have equal fractions of molecules that start there, but the level of the imperfection is insufficient to affect the outcome of an *in vitro* selection (19; Vant-Hull *et al*., in preparation).

Sequencing showed that there are relatively few mutations in the library except in the nine nucleotide region immediately adjacent to the library primer annealing sites. Errors in this region make the library somewhat shorter by reducing the portion of the library molecules that is identical to the genomic sequences. These errors, as well as the adjacent fixed sequences, may affect binding of genomic inserts during the subsequent SELEX, and this remains a shortcoming in this method of library construction. However, all other published libraries have fixed sequences too. We are currently developing methods to overcome this limitation (Shtatland *et al*., in preparation).

We are not aware of any other published methods for rigorously testing the quality of a genomic library. Our methods may thus be a useful experimental tool in assessing the quality of a library. So far, only one other library was tested in our lab, and was found to be comparable in quality (He *et al*., in preparation).

If our libraries are used for RNA SELEX, one should keep in mind that not all genomic DNA is transcribed into RNA. Any RNAs not expressed *in vivo* will have little or no biological relevance; however, very tight binding to an RNA sequence not thought to be transcribed may indicate that the sequence is transcribed after all. Binding sites that are present only in spliced or edited RNA do not exist in our libraries. On the other hand, libraries made from cDNA do not include introns and intron–exon boundaries, sites that may be important in regulation of splicing [reviewed in (20)]. Moreover, cDNA libraries reflect transcription in some particular stage of development, and may thus yield incomplete answers for certain biological questions.

We have discussed both our expectations from genomic SELEX and the early results from some of the genomic SELEX experiments underway in our lab (9). We have used genomic SELEX to discover binding sites for the bacteriophage MS2 coat protein in the *E.coli* genome (Shtatland *et al*., in preparation) and binding sites for human U1A protein within human RNA (Singer *et al*., in preparation). We have also performed genomic SELEX using human basic fibroblast growth factor (a protein not known to bind RNA *in vivo*) and human genomic RNA; this SELEX yielded a single RNA winner that has a nM K_d (He *et al*., in preparation).

Genomic SELEX is conceived to be analogous to the yeast two-hybrid system (21) as a rapid screen for any protein–nucleic acid or metabolite–nucleic acid interaction that occurs *in vivo*; in short, we expect genomic SELEX to provide a nucleic acid 'linkage map' for such interactions and note that a nucleic acid 'linkage' made plausible through genomic SELEX can be tested directly in organisms using the familiar research tools of molecular biology.

ACKNOWLEDGEMENTS

We thank D. Burke, R. Gutell, K. Jensen, K. Krauter, A. Payano-Baez, and B.Vant-Hull for thoughtful comments and useful discussions. We thank A. Wright, C. Nislow, H. Chial, F. Luca, A. Monterrosa, S. Jacobsen, and A. Schutz for oligonucleotides, D. Smith for suggesting the primer sequences, S. Creighton for suggesting conditions for Klenow, and D. Lorenz for help with the figures. This work was supported by NIH grant GM19963 to L.G., and by funds from NeXstar Pharmaceuticals, Inc., for which we are grateful.

REFERENCES

1 Chu, E., Voeller, D. M., Jones, D. M., Takechi, T., Maley, G. F., Maley, F., Segal, S. and Allegra, C. J. (1994) *Mol. Cell. Biol.*, **14**, 207–213.
2 Singh, N. K., C. D. Atreya, and Nakhasi, H. L. (1994) *Proc. Natl. Acad. Sci. USA.*, **91**, 12770–12774.
3 Fester, T., and Schuster, W. (1995) *Biochem. Mol. Biol. Int.*, **36**, 67–75.
4 Hentze, M.W. (1994) *Trends Biochem. Sci.*, **19**, 101–103.
5 Kyrpides, N.C., and Ouzounis, C.A. (1995) *J. Mol. Evol.*, **40**, 564–569.
6 van Gelder, C. W., Gunderson, S. I., Jansen, E. J., Boelens, W. C., Polycarpou-Schwarz, M., Mattaj, I. W., and van Venrooij, W. J. (1993) *EMBO J.*, **12**, 5191–5200.
7 Tasheva, E. S., and Roufa, D. J. (1995) *Genes Dev.*, **9**, 304–316.
8 Gold, L., Polisky, B., Uhlenbeck, O., and Yarus, M. (1995) *Annu. Rev. Biochem.*, **64**, 763–797.
9 Gold, L., Brown, D., He, Y.-y., Shtatland, T., Singer, B.S., and Wu, Y. (1997) *Proc. Natl. Acad. Sci. USA*, **94**, 59–64.
10 Tuerk, C., and Gold, L. (1990) *Science*, **249**, 505–510.
11 Ellington, A. D., and Szostak, J. W. (1990) *Nature*, **346**, 818–822.
12 Phillipsen, P., Stotz, A., and Scherf, C. (1991) *Methods Enzymol.*, **194**, 169–182.
13 Wu, D. Y., Ugozzoli, L., Pal, B. K., Qian, J., Wallace, R. B. (1991) *DNA Cell Biol.*, **10**, 233–238.
14 Sompayrac, L., and Danna, K. J. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 3274–3278.
15 Kinzler, K.W., and Vogelstein, B. (1989) *Nucleic Acids Res.*, **17**, 3465–3653.
16 Grothues, D., Cantor, C. R., and Smith, C. L. (1993) *Nucleic Acids Res.*, **21**, 1321–1322.
17 Pearson, W.R., and Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 2444–2448.
18 Genetics Computer Group, *Program Manual for the Wisconsin Package*, 8th ed. 1994, Madison, WI.
19 Irvine, D., Tuerk, C., and Gold, L. (1991) *J. Mol. Biol.*, **222**, 739–61.
20 Manley, J.L., and Tacke, R. (1996) *Genes Dev.*, **10**, 1569–1579.
21 Evangelista, C., Lockshon, D., and Fields, S. (1996) *Trends Cell Biol.*, **6**, 196–199.