Functional Cloning, Sorting, and Expression Profiling of Nucleic Acid-Binding Proteins

Y. Ramanathan, Haibo Zhang, Virginie Aris, et al.

Genome Res. 2002 12: 1175-1184
Access the most recent version at doi:10.1101/gr.156002

Letter

Functional Cloning, Sorting, and Expression Profiling of Nucleic Acid-Binding Proteins

Y. Ramanathan,[1] Haibo Zhang,[1,2] Virginie Aris,[1,2] Patricia Soteropoulos,[1,3] Stuart A. Aaronson,[4] and Peter P. Tolias[1,3,5]

[1]Center for Applied Genomics, Public Health Research Institute, International Center for Public Health W420M, Newark, New Jersey 07103, USA; [2]Center for Computational Biology and Bioengineering, New Jersey Institute of Technology, New Jersey 07102, USA; [3]Department of Microbiology and Molecular Genetics, University of Medicine and Dentistry of New Jersey-New Jersey Medical School, Newark, New Jersey 07103, USA; [4]The Derald H. Ruttenberg Cancer Center, Mount Sinai School of Medicine of New York University, New York, New York 10029, USA

A major challenge in the post-sequencing era is to elucidate the activity and biological function of genes that reside in the human genome. An important subset includes genes that encode proteins that regulate gene expression or maintain the structural integrity of the genome. Using a novel oligonucleotide-binding substrate as bait, we show the feasibility of a modified functional expression-cloning strategy to identify human cDNAs that encode a spectrum of nucleic acid-binding proteins (NBPs). Approximately 170 cDNAs were identified from screening phage libraries derived from a human colorectal adenocarcinoma cell line and from noncancerous fetal lung tissue. Sequence analysis confirmed that virtually every clone contained a known DNA- or RNA-binding motif. We also report on a complementary sorting strategy that, in the absence of subcloning and protein purification, can distinguish different classes of NBPs according to their particular binding properties. To extend our functional annotation of NBPs, we have used GeneChip expression profiling of 14 different breast-derived cell lines to examine the relative transcriptional activity of genes identified in our screen and cluster analysis to discover other genes that have similar expression patterns. Finally, we present strategies to analyze the upstream regulatory region of each gene within a cluster group and select unique combinations of transcription factor binding sites that may be responsible for dictating the observed synexpression.

[The following individual kindly provided reagents, samples, or unpublished information as indicated in the paper: M. Stempher.]

DNA sequencing data derived from the human genome suggests that the total number of genes is ~30,000–40,000, and that only a fraction of these have been annotated functionally. Surprisingly, a comparison between human genes predicted by the publicly funded genome consortium (Lander et al. 2001) and those proposed by Celera Genomics (Venter et al. 2001) reveals that the two groups are in agreement for only ~12,000 genes (Hogenesch et al. 2001). In silico approaches for gene prediction on the basis of ontology classification have been applied to model organisms such as *Drosophila melanogaster*, but have yielded an underestimation of the total number of genes encoded by the fruit fly genome (Adams et al. 2000). In other studies, searching for paralogs of known cancer-causing human genes has also met with limited success (Futreal et al. 2001). A more promising approach has been the elucidation of transcript expression profiles using DNA chips/microarrays. This technique has revolutionized the manner by which we study differences in gene expression between two cell populations, as it enables the simultaneous monitoring of thousands of genes (Liotta and Petricoin 2000). Whereas expression profiling produces an overwhelming quantity of data, modeling biological problems and assigning genes to pathways requires prior knowledge of protein function or activity. Thus, integrative approaches involving computational prediction, microarray technology, and functional characterization are required for continued gene discovery and pathway elucidation. These observations indicate that a major challenge of the post-sequencing era is to assign function to tens of thousands of genes that are either unrelated to known genes or for whom functional data do not currently exist.

An important subset of genes for which we seek information on function and/or activity includes those that encode nucleic acid binding proteins (NBPs), which act to regulate gene expression or maintain the structural integrity of the genome. As a group, NBPs are encoded by 13.5% of the predicted human genes (Venter et al. 2001). Perturbation of NBPs' expression or activity can significantly alter the entire genomic expression profile of a cell. This is underscored in diseases such as cancer that result as genetic and epigenetic alterations accumulate in individual cells. In addition to point mutations, cancer-associated genetic alterations include chromosomal deletions, amplifications, and translocations (Lengauer et al. 1998; Rowley 1998). For example, changes in the activity or expression of oncogenes or tumor suppressor genes often lead to the development of cancer (Hanahan and Weinberg 2000). Biochemical and genetic efforts have been instrumental in identifying genes that are primary targets of cancer-causing alterations, resulting in the discovery of nearly 100 oncogenes and 30 tumor suppressor genes, many of

[5]Corresponding author.
E-MAIL tolias@phri.org; FAX (973) 854-3453.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.156002.

which encode NBPs (Futreal et al. 2001). In addition to oncoproteins and tumor suppressors, alterations in the activity or expression of proteins that function in DNA repair can also lead to cancer (Hoeijmakers 2001).

With respect to their activity, NBPs can be distinguished on the basis of their binding properties for particular nucleic-acid types (DNA or RNA) and conformations [i.e., single stranded (ss), double stranded (ds), hairpins, etc,]. The best characterized are transcription factors (such as p53 and E2F) that bind to dsDNA and regulate gene expression (Dyson 1998; Ryan et al. 2001). ssDNA-binding proteins constitute another important group and are usually involved in genome maintenance; defects in DNA repair can contribute to a predisposition for many cancers (for review, see Hoeijmakers 2001). Another major class of NBPs includes RNA-binding proteins that function in post-transcriptional processing, RNA stability, and protein biosynthesis. Many clinically relevant transcripts are protected from degradation by RNA-binding proteins (for review, see Chen and Shyu 1995). For example, c-*myc* transcripts are protected from degradation by a ssRNA-binding protein whose levels are elevated in colorectal cancer and the corresponding gene amplified in breast cancer (Doyle et al. 2000; Ross et al. 2001). Modulation of another RNA-binding protein CstF by the tumor suppressor BRCA1/BARD1 complex has been shown recently to affect polyadenylation in certain tumor cells (Kleiman and Manley 2001).

Technologies that can identify genes encoding NBPs are valuable tools for functional annotation of genomic data. The electrophoretic mobility shift assay is a strategy that has been used for many years to identify proteins that bind a DNA/RNA element of interest. Supershifting the mobility of protein-nucleic acid complexes with antibodies can reveal the identity of the bound proteins. The antibody that caused the supershifting phenomenon can be used to clone the corresponding gene from an expression library. If antibodies are not available, protein purification followed by amino acid sequencing enables the design of degenerate primers that can be used for cloning the gene. These methods are time consuming and low throughput, designed to clone one gene at a time.

Alternative strategies to identify or study targets of NBPs include PCR-based sequential binding and amplification of oligonucleotides to determine nucleic acid-binding specificity for individual or pools of purified proteins (for review, see Quellette and Wright 1995). If the original proteins are unknowns, the identified binding sites can then be used as bait to isolate them from cDNA expression libraries. A powerful approach to study protein-nucleic acid interaction in vivo combines chromatin immunoprecipitation (ChIP) analysis with DNA microarrays. This method begins with chemical cross-linking of proteins bound to DNA in living tissue followed by immunoprecipitation (Orlando 2000). The precipitated DNA can be labeled and used as probes on DNA microarrays to monitor the effect of binding of transcriptional activators on gene expression (Ren et al. 2000; Iyer et al. 2001). This approach is powerful in understanding global regulatory networks in prokaryotes and yeast whose genomes are small and intergenic regions are amenable to microarray studies. Similar studies reporting the use of microarray expression analysis combined with ChIP have been applied recently to higher eukaryotes (Ren et al. 2002; Weinmann et al. 2002). However, complete representation of intergenic regions from higher eukaryotes on microarrays is still a major challenge.

A potential strategy for cloning NBPs is the yeast one-hybrid screen (Fields and Song 1989). Although this method has the advantage of screening under relatively native in vivo conditions, false positives are common (due to binding of endogenous yeast NBPs), especially when the binding site is designed to bind a spectrum of NBPs. Functional cloning of a desired DNA-binding protein by screening a phage expression library using a specific DNA sequence as a bait was first reported by Singh et al. (1988). A similar strategy for cloning an RNA-binding protein was used by Qian and Wilusz (1993).

In the present study, we show the feasibility of functional screening of NBPs from cDNA libraries using a novel DNA substrate as bait within a modified expression cloning strategy. We have also developed a complementary nitrocellulose filter-binding method that sorts the activity encoded by each clone isolated in the screen directly without purification and reveals information on their binding properties for particular types and conformations of nucleic acids. To further our functional annotation of each of the unique genes obtained from our screen, we searched for other genes that had related expression profiles across a series of 14 breast-derived cell lines (12 cancer and 2 normal epithelial cell lines) as determined with Affymetrix GeneChips. We present strategies to analyze coordinate expression of gene clusters to determine the candidate unique combination of transcription factors that may be controlling the observed expression of the group.

RESULTS

Rational and Screening for NBP Activities

Traditionally, biochemists have exploited the ability of both DNA- and RNA-binding proteins to nonspecifically bind to ss- and dsDNA columns to help purify them. The major limitation of this approach has been its low throughput and the involvement of many tedious steps from protein purification and sequencing to degenerate probe design and library screening. To eliminate these problems, we replaced this biochemical approach with a molecular strategy that simultaneously enables both the identification and cloning of nucleic acid-binding activities. The strategy is a modification of the cDNA expression cloning procedure, which involves screening phage libraries that are plated and transferred onto nitrocellulose filters for activities that bind a novel universal substrate (shown in Fig. 1). This substrate is a 55-mer DNA oligonucleotide designed to form several predicted secondary structural features that could be bound by a variety of NBPs. The structural features include the following: (1) a linear ss region of random sequence with a free 5′ end, (2) a structure resembling a replication gap consisting of the above-mentioned linear random ss region adjacent to a duplex composed of unique base pairs, (3) a region within the duplex that contains an extra nucleotide and a mismatch, and (4) a hairpin structure.

Using radioactively labeled universal substrate as bait, our intent was to identify a variety of NBPs through λ gt11 cDNA library expression screening of their activity. The libraries were derived from cell line SW480 (colorectal adenocarcinoma) and normal fetal lung tissue. These two libraries were used to examine the applicability of the screen to identify NBPs from diverse tissue sources. Phage clones from each library were used to infect *Escherichia coli* strain Y1090. After IPTG induction, proteins expressed by the phage were transferred onto nitrocellulose filters, subjected to sequential denaturation and renaturation in serial dilutions of guanidine hydrochloride, and used for in vitro-binding reactions with ^{32}P-labeled universal substrate (Fig. 1).

Figure 1 Binding substrates. The sequence and the predicted secondary structure of different binding substrates used in the screening and sorting experiments. The region containing extra nucleotides is denoted by outlined text, mismatches are denoted in bold. (Pu) Purine; (N) degenerate base. Not shown in the figure is poly (I)·poly (C) (see Methods).

Because the orientation of each cDNA insert in the λ gt11 expression library is random, and because only one of the three open reading frames can produce functional protein, we expected only one of every six clones screened to actually express a protein. We decided to screen ~400,000 clones from each library. Ninety-one positive clones were obtained from the SW480 library and seventy-nine from fetal lung. The 3′ and 5′ ends of cDNA inserts were sequenced using vector-specific primers. Clone identities were determined using BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) and conserved protein domains were searched in the PFAM database (http://pfam.wustl.edu/hmmsearch.shtml). Table 1a lists the clones obtained from the primary screen and the corresponding number of hits from each library. Of 170 cDNA clones sequenced, 30 were unique. As only a single clone contained no known nucleic-acid binding motifs, the screen did not appear to identify false positives, implying that the universal substrate was an effective ligand for screening DNA- and RNA-binding proteins. The positive clones isolated in each screen are associated with diverse cellular functions, many are involved in RNA metabolism, and nearly one-third are unique. In addition to identifying known NBPs, the screening strategy also identified six novel transcripts (one from the adenocar-

Table 1a. List of Clones Obtained from the Primary Screen

Clone no.	Gene name	UniGene no.	No. of hits		Motif
			Adeno	Fetal	
1	PTB	Hs. 172550	3	0	RRM
2	HNRPDL	Hs. 170311	1	7	RRM
3	YB-1/dpbB	Hs. 74497	19	8	CSD
4	Nucleolin	Hs. 79110	3	1	RRM
5	Topoisomerase 3	Hs. 91175	3	0	Several
6	CSDA/dbpA	Hs. 198726	13	1	CSD
7	ASF	Hs. 73737	1	0	RRM
8	SFRS5	Hs. 166975	1	1	RRM
9	HNRPAB	Hs. 81361	3	0	RRM
10	CPSF4 (30K)	Hs. 6351	5	0	zf (CCCH)
11	PEG10	Hs. 137476	0	1	bZip and zf (CCHC)
12	CIRBP	Hs. 119475	1	7	RRM
13	**Novel 405_111**	NA	0	1	RRM
14	ZNF9	Hs. 2110	0	1	zf (CCHC & NF-X1)
15	HNRPM	Hs. 79024	2	4	RRM
16	SPRS4	Hs. 76122	1	0	RRM
17	**Novel 346_711**	NA	1	0	Unknown
18	RBMS2	Hs. 20938	0	1	RRM
19	EIF4H	Hs. 180900	0	1	RRM
20	PABPC1	Hs. 172182	13	5	RRM & PABP
21	PABPC4	hs. 169900	10	1	bZip, RRM & PABP
22	HNRPA 1	Hs. 249495	4	14	RRM
23	HNRPD	Hs. 303627	0	10	RRM
24	**Novel 405_311**	NA	0	1	R3H
25	**Novel 412_111**	NA	0	1	Coiled coil DBD
26	**Novel 418_321**	NA	0	1	RRM
27	MSSP-2	Hs. 241567	0	1	RRM
28	RBMS1	Hs. 241567	0	4	RRM
29	U2AF65	Hs. 7655	1	0	RRM
30	**Novel 421_111**	NA	0	1	RRM

The first column refers to the clone number, second column to the gene identity and third column to the corresponding UniGene number. Novel genes are represented in bold and NA refers to not applicable. The fourth column provides information on the corresponding number of hits of each gene from the two library screens (Adeno library made from adeno-carcinoma cell line and Fetal library made from fetal-lung tissue). The fifth column provides information on the predicted nucleic acid-binding motif present in the protein encoded by the corresponding gene (as determined by searching the PFAM database).

Table 1b. Description of Novel Clones

Novel clones	Description
405_111	Splice variant of cold inducible RNA-binding protein (Hs. 119475)
346_711	Splice variant of a hypothetical protein FLJ10261 (Hs. 26176)
405_311	Homology to human KIAA1002 protein (NM_014925)
412_111	Splice variant of Positive Cofactor 4-PC4 (Hs. 356473)
418_321	Splice variant of RNA-binding motif single stranded interacting protein-RBMS3 (Hs. 158446)
421_111	Homology to Musashi homolog 2 mRNA (NM_138962)

The first column refers to the novel clones and the second column to a short description of homology of the corresponding novel clone to known sequence. Accession/UniGene numbers are indicated in parenthesis.

cinoma and five from the fetal lung library) consisting of both novel genes and unreported splice variants (see Table 1b). Overall, this screening method is effective in cloning a variety of DNA and RNA-binding proteins.

Sorting the Activity of NBPs

Nucleic acid-binding proteins can be further distinguished on the basis of their ability to bind DNA or RNA and particular conformations (ss, ds, hairpins, etc.). It is thus desirable to further characterize clones obtained from the primary screen and to further distinguish their individual nucleic acid-binding activities. To do this, sorting of the 30 unique nucleic acid-binding proteins obtained from our primary screens was attempted by use of a similar filter-binding strategy. For this purpose, we tested the binding characteristics of these unique clones to seven additional sorting oligonucleotides, each having a specific predicted conformation (Fig. 1). Six of these display features contained in distinct regions of the universal substrate such as (1) dsDNA, (2) mismatched dsDNA, (3) mismatched dsDNA with a hairpin, (4) DNA hairpin, (5) random ssDNA, and (6) ssRNA. An additional sorting substrate was used in the form of poly (I)·poly(C), a synthetic dsRNA analog. Our goal was to use this battery of oligonucleotides as baits to sort the activity of the clones obtained from the primary screen into useful categories on the basis of their individual nucleic acid-binding properties.

A multiplicity of infection of 10^3 of each purified positive clone was used to infect *E. coli* Y1090 and plated on a 150-mm plate. Filter-binding assays were performed as described above, except that each filter was cut into eight equal slices and each slice was allowed to bind one of the eight end-labeled binding substrates. Random clones from the adenocarcinoma library were used as a negative control and human Staufen, a dsRNA-binding protein was used as a positive control for poly(I)·poly(C) binding. Results of the sorting experiment are shown in Figure 2.

Each block represents an independent experiment. As expected, all 30 proteins bind the universal substrate. A total of 21 bound to ssDNA, 15 to mismatched DNA hairpin, 11 to ssRNA, 9 to a DNA hairpin, 8 to dsDNA, 7 to mismatched dsDNA, and 3 to poly (I)·poly(C). No protein was capable of binding to all eight substrates, indicating the sufficient diversity of the sorting substrates. In summary, these results indicate that the activity of nucleic acid-binding proteins can be rapidly sorted by this simple filter-binding assay. Knowledge of each protein's binding property is informative in designing new experiments for further functional characterization. For example, clone 35, which bound to mismatched dsDNA and not to a perfect duplex, may be involved in the recognition of damaged DNA.

Synexpression Studies using GeneChip Profiling

The results presented above show the feasibility of our novel approach to clone and further characterize the activity of large numbers of NBPs. To extend the functional annotation of individual clones that were identified, we took advantage of the availability of the human genome sequence and the ability to monitor the mRNA expression of thousands of genes with microarrays. Our rationale was to use microarray data to identify genes that are regulated in a similar manner as our NBP of interest and to analyze the *cis*-regulatory elements of these genes. This may enable the identification of distinct binding sites for unique combinations of transcription factors that regulate this group of genes as a whole, providing further insight into their biological function and regulation.

Using GeneChip microarrays, we examined the normalized transcription profiles of 12,500 genes in 14 breast-derived cell lines (Table 2) and uncovered distinct clusters of genes with similar expression patterns to the 24 known genes that were obtained from our primary screen for NBPs (see Table 1a). This analysis was performed using `GeneSpring` 4.11 with a Pearson correlation and threshold set above 0.95. Six clusters (of the 24 examined) that contained from 5 to 50 members were selected for further analysis. The results are depicted in Figure 3, in which each graph illustrates the expression profile of one of the six selected positive clones in the 14 breast cell lines along with that of other probe sets that represent additional coexpressed genes.

Computational whole-genome analysis suggests that known DNA-binding sites are significantly enriched in upstream promoter regions compared with other regions and random sequences (Aach et al. 2001). We thus retrieved 4 kb of DNA sequence upstream from the transcriptional start site of all the clustered genes (from the NCBI human genome database) and searched for transcription factor binding sites within each 4-kb upstream region using TRANSFAC database at a default threshold of 85 (http://molsun1.cbrc.aist.go.jp/research/db/TFSEARCH.html). Whereas many transcription factor binding sites were found within each of the upstream 4-kb regions, only those that had at least one putative binding site in the 4-kb upstream sequence of all members within a cluster (or synexpression group) were considered for further analysis. Only 31 of ~1500 relevant binding sites in the database met this criteria (Table 3). Examination of the results presented in Table 3 led to the following hypotheses: First, transcription factors such as AML-1a, deltaEF1, GATA-1, GATA-2, GATA-3, Ik-2, Lyf-1, MZF1, Nkx-2.5, CdxA, and SRY are probably not conferring expression specificity because their binding sites are present in the 4-kb upstream sequence of all the genes from all six clusters; secondly, it is possible that different combinations of the remaining 20 putative binding sites may determine the specificity of each synexpression group. Although some transcription factors (such as USF) are ubiquitously expressed, they have been shown to exert cell-specific regulation of different genes in association with

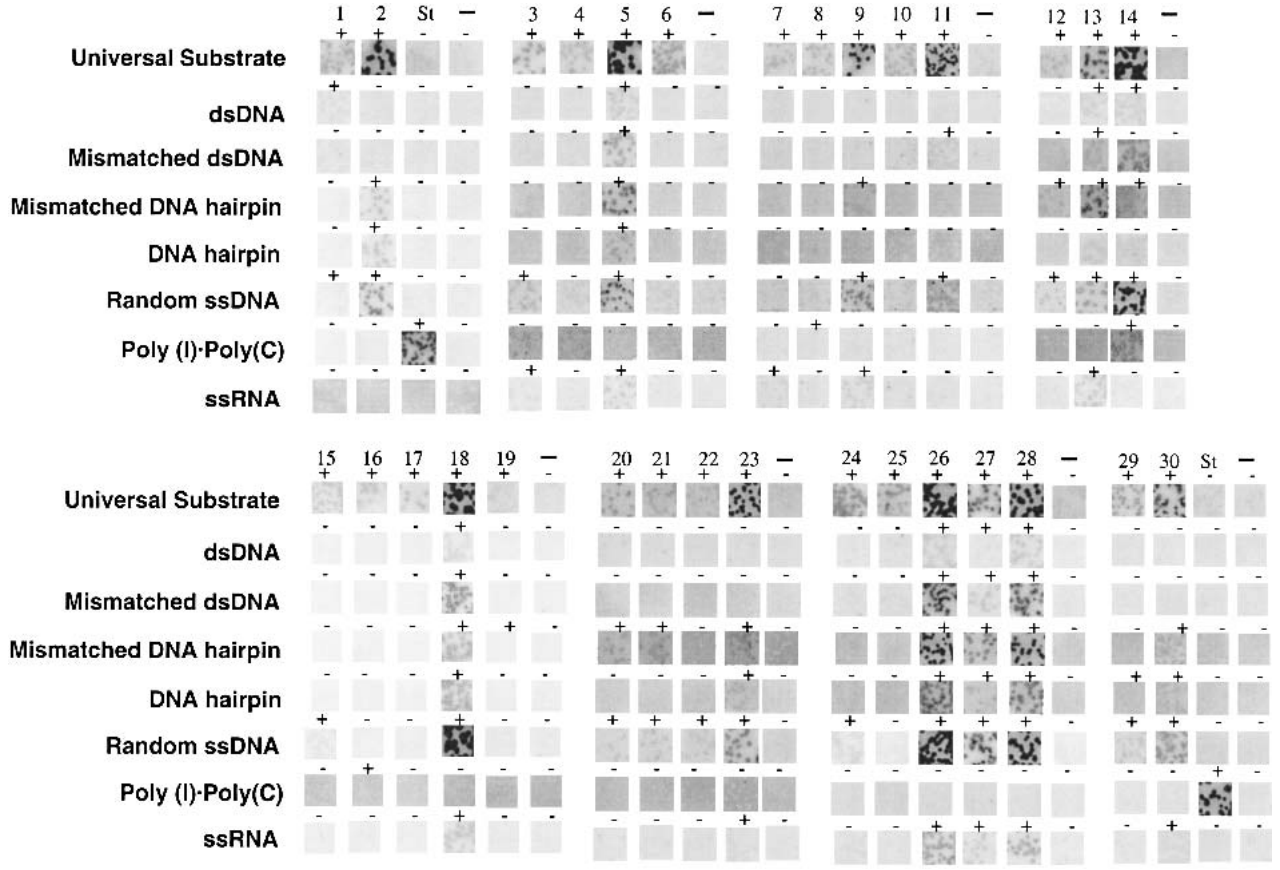

Figure 2 Sorting nucleic acid binding proteins. Each block represents an independent experiment. Numbers at *top* refer to gene identity (see Table 1a); (St) Staufen; negative control. (+/−) The binding activity of each clone to the corresponding binding substrate.

other factors depending upon the cell type that they are expressed in (Andrews et al. 2001). Finally, some transcription factors such as CREB, Sox-5, p300, Pbx-1, CDP, NF-κB, and GATA-X contain binding sites unique to only one of the six synexpression groups.

Table 2. Breast Cancer Lines

Cell lines	ATCC number	Tumorigenicity
MCF-7	HTB-22	Yes
AB5589	Lab stock	No
SK-BR-3	HTB-30	Yes
MDA-MB-468	HTB-132	Yes
T-47D	HTB-133	Yes
MCF-10A	CRL-10317	No
BT-20	HTB-19	Yes
BT-474	HTB-20	Yes
ZR-75-1	CRL-1500	Yes
MDA-MB-134 V1	HTB-23	Yes
MDA-MB-361	HTB-27	Yes
MDA-MB-231	HTB-26	Yes
MDA-MB-157	HTB-24	Yes
Breast Cancer-3	Lab stock	Yes

The first column lists the fourteen cell lines used for expression profiling and the second column indicates the corresponding ATCC number where available. The third column refers to the tumorigenicity of the individual cell line.

We then addressed the possibility that the presence of these 31 transcription factor binding sites in the 6 promoter clusters was occurring by chance. To do this, we first counted the total number of binding sites for each transcription factor as it occurred within an entire cluster and divided by the number of genes in that cluster. This gave a measure of the average number of binding sites for each transcription factor within the 4-kb upstream region of each gene within a synexpression group. We then compared this number with the average number of binding sites that were expected to occur for each transcription factor by chance in random sequences. For this calculation, 6 sets of 100 different 4-kb random sequences were generated (using JAVA random number generator), but the base composition of the promoter cluster was maintained. A search for transcription factor binding sites in each of these 100 random sequences was done, and the average occurrence of each binding site in the random sequence was calculated. The results are presented in Figure 4, in which we have plotted the ratio of the average occurrence of each transcription factor binding site in each promoter cluster to the number of times it is expected to occur by chance, expressed as a percentage difference. It appears from this analysis that only 8 (Table 4) of 31 (Table 3) predicted transcription factor binding sites are occurring at over twice the frequency that is expected by chance (Fig. 4). This analysis suggests that only a subset of the predicted transcription factor binding sites that occur in each

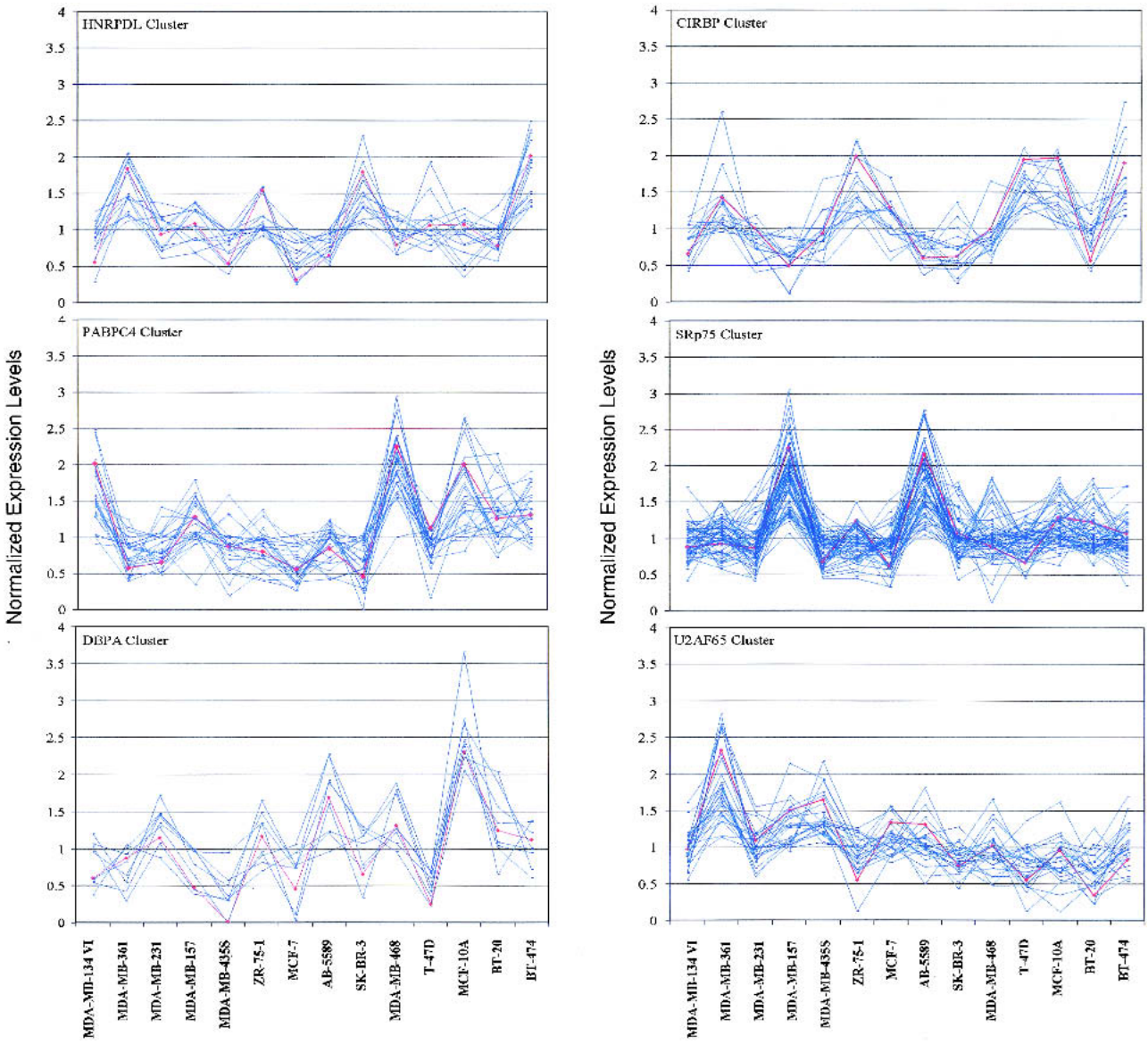

Figure 3 Clustering genes with similar expression patterns in 14 different breast-derived cell lines. Each panel displays groups of genes (in blue) with similar expression patterns in the indicated (normal and cancer) cell lines to the query gene (in red) denoted at *top*. GeneChips were scaled to 2500 and each gene was median centered by computing its median expression level across all of the cell lines (provided that this value was at least 10 and that the ratio of the absolute expression and median was at least 0.001). Selection was done by use of a Pearson correlation with a threshold >0.95 with `GeneSpring` software 4.11.

cluster (as shown in Table 4) may be responsible for synexpression.

DISCUSSION

Major goals of genomic research in the post-sequencing era include efforts to discover and functionally annotate novel genes and to further characterize known genes. This study highlights the use of a broad combination of innovative molecular and computational techniques such as activity expression cloning and sorting, microarray clustering, and computational promoter analysis, to functionally characterize large numbers of NBPs.

Current in silico approaches to identify genes are useful but suffer from several predictive failures including the identification of false positives and false negatives (Venter et al. 2001) or the inability to detect unknown but real genes (Futreal et al. 2001; Murray and Marks 2001). Because the function of the majority of the genes encoded by the human genome is not known (Lander et al. 2001) and 40% of the predicted proteins cannot be ascribed with distinct molecular functions (Venter et al. 2001), there is a need to develop new methods to solve these problems. Even if a gene is accurately predicted, the value of genomic information can be realized only when it is coupled with further biochemical characterization (Tupler et al. 2001).

Over the last decade, the cloning of DNA-binding proteins by screening expression libraries with specific binding

Table 3. Transcription Factor Binding Sites in the 4-kb Upstream Sequence of Clustered Genes

CIRBP (11)	DBPA (8)	HNRPDL (13)	PABPC4 (18)	SRP75 (41)	U2AF65 (20)
AML-1a	AML-1a	AML-1a	AML-1a	AML-1a	AML-1a
δEF1	δEF1	δEF1	δEF1	δEF1	δEF1
GATA-1	GATA-1	GATA-1	GATA-1	GATA-1	GATA-1
GATA-2	GATA-2	GATA-2	GATA-2	GATA-2	GATA-2
GATA-3	GATA-3	GATA-3	GATA-3	GATA-3	GATA-3
Ik-2	Ik-2	Ik-2	Ik-2	Ik-2	Ik-2
Lyf-1	Lyf-1	Lyf-1	Lyf-1	Lyf-1	Lyf-1
MZF1	MZF1	MZF1	MZF1	MZF1	MZF1
Nkx-2.5	Nkx-2.5	Nkx-2.5	Nkx-2.5	Nkx-2.5	Nkx-2.5
SRY	SRY	SRY	SRY	SRY	SRY
CdxA	CdxA	CdxA	CdxA	CdxA	CdxA
	AP-1	AP-1	AP-1	AP-1	AP-1
c-Ets-1		c-Ets-1	c-Ets-1	c-Ets-1	c-Ets-1
Oct-1	Oct-1	Oct-1	Oct-1		Oct-1
USF	USF			USF	USF
HNF-3b		HNF-3b	HNF-3b		
S8	S8	S8			
	C/EBPb1	C/EBPb1	C/EBPb1		
C/EBP			C/EBP		
	CRE-BP		CRE-BP		
HFH-2		HFH-2			
HSF2	HSF2				
	STATx		STATx		
		Sp1			Sp1
CREB					
	GATA-X				
NF-κB					
			p300		
		Pbx-1			
			Sox-5		
	CDP				

The top row displays the positive clones selected for synexpression studies. Numbers in parenthesis refer to the number of genes that have similar expression patterns to the corresponding gene in 14 breast cell lines as determined using Gene Spring software (see Results). Binding sites for transcription factors present in the 4-kb upstream region of all the members of a cluster group are represented in each column.

sites has become an established technique for researchers who study transcriptional regulation (Singh et al. 1988). This method was designed to clone individual DNA-binding proteins using an oligonucleotide that has multiple copies of a binding site of interest. Using a modification of this approach, we show the feasibility of cloning not one but a spectrum of NBPs. This is based on a novel oligonucleotide binding substrate (Fig. 1), designed to form several structural features, which enabled the activity cloning of 30 unique NBPs having diverse cellular functions (Table 1a). A majority of the NBPs obtained in this screen were RNA-binding proteins. One explanation of this result is that RNA-binding proteins are more abundant in the cell compared with DNA-binding proteins. Alternatively, as our approach cannot detect NBPs that bind only as heteromers, this would also diminish the identification of DNA-binding transcription factors. In addition to known NBPs, the screen also identified six novel clones, suggesting that this is a reasonable approach for gene discovery.

Comparison of positive clones that were identified by our screen derived from normal fetal lung tissue versus the adenocarcinoma cell line revealed an asymmetrical distribution for some NBPs. An example is the very high representation of the transcription factor YB-1 (19 of 91) in the adenocarcinoma cell line (Table 1a). YB-1 is known to be expressed at high levels in several cancers, including colorectal carcinoma (Shibao et al. 1999). In contrast to YB-1, HNRPA1 is highly represented in the fetal lung library screen (14 of 80). This protein is primarily involved in the packaging and transport of mRNA (Nakielny et al. 1997) and is also thought to have a role in the formation of specific myometrial protein species in parturition (Pollard et al. 2000). From these two examples, it appears that our screening method is not only effective in cloning a spectrum of NBPs but also provides useful information regarding the tissue-specific distribution of expressed transcripts.

Several fusion clones were obtained in both library screens (data not shown). None of the six fusion clones obtained from the adenocarcinoma library were represented more than once, suggesting that they may have arisen as in vitro artifacts during library construction. However, we cannot rule out the possibility that some of these may actually represent translocations or other genomic rearrangements. One of the six independent fusion clones obtained from the fetal lung library was represented twice. This clone represents a fusion between the gene encoding cold inducible RNA-binding protein (Chromosome19) and the gene encoding collagen1 (Chromosome17). The functional significance of this fusion remains to be elucidated.

Using a battery of binding substrates shown in Figure 1 and a further modification of our screening procedure, the activity of NBPs obtained from our screen could be further characterized on the basis of their specific nucleic acid-binding properties. This one step filter-binding-based sorting assay is a rapid alternative for preliminary characterization of binding activities compared with other approaches that require tedious and time-consuming subcloning and protein purification strategies for every clone.

The approach of clustering genes that show similar expression patterns across many microarray experiments and to identify common *cis*-regulatory elements has been applied to lower eukaryotes such as yeast (Roth et al. 1998; Spellman et al. 1998; Pilpel et al. 2001). Using normalized microarray data from 14 breast-derived cell lines (see Results section and Fig. 3), we clustered genes displaying similar expression patterns to each of the genes identified in our screen. Known upstream *cis*-regulatory elements unique to each cluster were identified from the transcription factor binding site database. The presence of unique combinations of putative binding sites for certain transcription factors in the upstream regulatory region of each gene cluster is in agreement with combinatorial gene regulation strategies thought to be used by eukaryotes (Smale 2001), suggesting that each gene cluster may be regulated by a unique mechanism. Our results are consistent with this hypothesis, as only 31 of ~1500 binding sites examined were present in these 6 clusters. When the frequency of occurrence of these transcription factor binding sites in real sequences for each cluster of coexpressed genes was compared with randomly generated sequences (see Results), without exception, all six clusters had transcription factor binding sites that were enriched, which lends credibility to this approach. However, only 8 of the 31 transcription factor binding sites are predicted to occur greater than twofold in real clusters compared with random sequences. Although this observation may suggest that only a smaller subset (compare Tables 3 and 4) of transcription factors may be responsible for synexpression, caution needs to be exercised when interpreting these results,

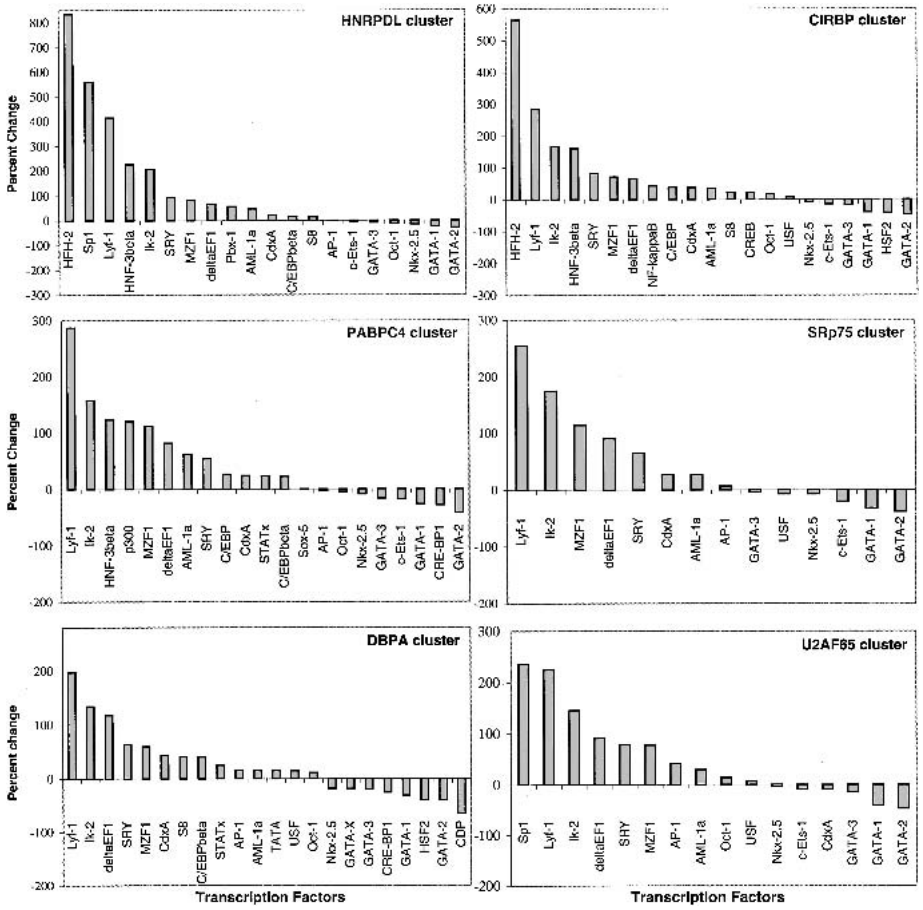

Figure 4 Relative abundance of transcription factor binding sites in experimentally derived versus random DNA sequence clusters. Each panel refers to a cluster of genes coregulated with a queried gene. The ratio of the relative abundance of each transcription factor binding site in a real (experimentally derived) compared with a random DNA sequence cluster is expressed as a percent change on the Y-axis. Transcription factors are plotted on the X-axis.

as some of the biologically relevant transcription factors may be excluded. This analytical strategy may provide leads to decipher the function of poorly characterized and novel genes by virtue of their being coexpressed with genes that are known to function in particular processes. In silco approaches such as these will be instrumental in acting as filters to guide the direction of wet laboratory research strategies in efforts to understand regulatory networks that operate in complex eukaryotes such as humans.

Table 4. Computational Filtering of Putative Transcription Factor Binding Sites in Clustered Promoters

CIRBP (11)	DBPA (8)	HNRPDL (13)	PABPC4 (18)	SRP75 (41)	U2AF65 (20)
Ik-2	Ik-2	Ik-2	Ik-2	Ik-2	Ik-2
Lyf-1	Lyf-1	Lyf-1	Lyf-1	Lyf-1	Lyf-1
HNF-3b		HNF-3b	HNF-3b		
			MZF1	MZF1	
HFH-2		HFH-2			
		Sp1			Sp1
	ΔEF1				
			p300		

This table represents filtered data that appears in Table 3, displaying only binding sites for transcription factors that are enriched by at least 100% in the real (experimentally derived) versus random DNA sequence clusters.

METHODS

Binding Substrates and Primers

Primers for sequencing and binding substrates for screening and sorting experiments were custom synthesized either at Integrated DNA Technologies, Inc. or at the University of Medicine and Dentistry of New Jersey Molecular Resource Facility. The sequence of the binding substrates is presented in Figure 1. Poly (I)·Poly(C) was purchased from Amersham Pharmacia Biotech, Inc.

Preparation of ^{32}P-Labeled Binding Substrates

The binding substrates were end labeled with [γ^{32}P]ATP (NEN Life Science Products) using T4 polynucleotide kinase (Sambrook et al. 1989). Oligonucleotides were denatured at 20°C above their Tm and allowed to renature by slow cooling except for random ssDNA and ssRNA (see Fig. 1), which, after denaturing for 5 min, were chilled on ice to prevent renaturing. Typically, a 20-µL reaction contained 50 picomoles of oligo, 13 µL of [γ^{32}P]ATP (130 µCi), 2 µL of 10× kinase buffer, and 1 µL of T4 polynucleotide kinase. The reaction mixture was incubated at 37°C for 45 min. Reaction was stopped with 16 µL of stop solution (40% glycerol; 100 mM EDTA; 0.02% SDS), and the unincorporated nucleotides were removed from the reaction by passing through 1 mL of Bio-gel P30 resin (Bio-Rad Laboratories).

λgtII Expression Libraries and cDNA Screening

The human fetal lung and the human colorectal adenocarcinoma cDNA libraries were constructed in the expression vector λgt11 by Clontech Laboratories, Inc. Each library contains ~10^6 independent clones with an average insert size of 1.7 kb. *E. coli* strain Y1090 was used as a host for screening the library and for phage DNA extraction.

cDNA library screening was performed as a modification of a protocol originally described in Vinson et al. (1988). Twenty milliliters of LB medium (supplemented with 10 mM $MgSO_4$ and 0.2% Maltose) was inoculated with *E. coli* strain Y1090. The culture was grown for 6 h at 37°C with agitation. The cells were collected by centrifugation and resuspended in 1/2 vol of 10 mM $MgSO_4$. Infection was performed in sterile snap-cap polystyrene tubes (USA Scientific). For primary screening, 300 µL of *E. coli* cells were mixed with phage (at 1–2 × 10^4 pfu for each 150 × 15-mm plate) and incubated

for 20 min at 37°C. Eight milliliters of molten top agarose (LB-Mg with 0.7% agarose) prewarmed to 45°C was added and the mixture was vortexed and poured onto LB agar plates (containing 50 µg/mL of ampicillin). Petri dishes were incubated inverted for 4 to 5 h at 37°C until pinpoint plaques appeared. Nitrocellulose membrane circles (Schleicher & Schuell, BA85; pore size 0.45 µm) were labeled with a pencil and soaked in 10 mM IPTG for 30 min. Membranes were air dried on 3-mm Whatman paper for 20–30 min and placed gently on the plates containing pinpoint plaques. Plates were incubated for an additional 10–12 h at 37°C, and chilled at 4°C for 1 h. Orientation of the filter with respect to the plate was marked by making three asymmetric holes through the filter with a needle dipped in India ink. Membranes were removed from the top agarose and air dried for 20–30 min. on 3-mm Whatman paper. Membrane-bound proteins were subjected to denaturation and renaturation by sequential 10-min incubation cycles with agitation at 4°C with 6 M, 3 M, 1.5 M, 0.375 M, and 0.185 M guanidine hydrochloride in Binding buffer [25 mM HEPES (pH7.9); 25 mM NaCl; 5 mM $MgCl_2$; 0.5 mM DTT; 50 µM $ZnCl_2$]. The membranes were then washed twice for 45 min with binding buffer, blocked with 5% BSA in Binding buffer, and rinsed twice with Binding buffer containing 0.25% BSA. For binding assays with radiolabeled oligonucleotides, membranes were incubated with ^{32}P-labeled binding substrate (1–2 $\times$ 10^6 Cerenkov counts/min/mL) in binding buffer with gentle agitation, initially for 1 h at 4°C, followed by 2 h at room temperature. After three consecutive washes with binding buffer containing 0.1% Triton X-100 in a shaker, the membranes were air dried for 20–30 min. on 3-mm Whatman paper, wrapped with Saran wrap, and exposed to film for autoradiography or to a PhosphorImager screen for imaging. Positive plaques were picked and suspended in 1 mL of SM buffer [50 mM Tris-Cl (pH 7.5), 100 mM NaCl, 8 mM $MgSO_4$, 0.01% gelatin], and stored in 2 drops of chloroform. The eluted phages were used for secondary screening by a method exactly as detailed above ,except that a smaller petri plate (100 $\times$ 15 mm) was used. For sorting experiment, each NBP clone was plated on 150 $\times$ 15 mm LB amp plate at a density of 800–1000 plaques. As a negative control, the library itself was plated to get a similar number of plaques. The filter-binding assay in the cDNA screen was performed as detailed above, except that each filter is cut into eight equal slices. Eight individual binding reactions were performed, each containing one of the seven labeled sorting substrates depicted in Figure 1 and with poly (I)·poly(C).

λ DNA Preparation and Sequencing

λ plate lysate preparation and DNA extraction were performed as described in Sambrook et al. (1989) with the exception that NZCYM and λ diluent were replaced with LB medium and SM buffer, respectively.

Phage lysates were prepared by plating 2–3 $\times$ 10^4 pfu on 150 $\times$ 15 mm LB/Mg agarose plates. After overnight incubation at 37°C, the top agarose was harvested and resuspended in 2 mL of SM buffer and a few drops of chloroform, followed by vigorous vortexing. Lysate was collected after centrifugation.

For DNA extraction, ~4 mL of lysate (~10^{10} pfu) was dechloroformed and treated with DNaseI (100 units/mL final) and RNaseA (20 µg/mL final) at 37°C for 1h. The lysate was incubated at 4°C for 1 h with NaCl to a final concentration of 1 M. Debri was removed by centrifugation and the supernatant was incubated at 4°C for 1 h with polyethylene glycol (8000) to a final concentration of 10%. Phage particles were centrifuged at 10,000*g* for 15 min, and each pellet was resuspended in 1 mL of SM buffer. EDTA and SDS were added to a final concentration of 20 mM and 0.5%, respectively and incubated for 10 min at 68°C. Aliquots (600 µL) were transferred to microfuge tubes, and phage DNA was purified by extracting twice with phenol and once with chloroform. DNA was precipitated with ethanol and the pellets were resuspended in TE (10 mM Tris-Cl and 1 mM EDTA). The DNA pellets were treated with RNaseA (20 µg/mL final) and further processed by phenol extraction and ethanol precipitation.

λ forward (GGTGGCGACGACTCCTGGAGCCCG or GGTCTGCGCTGCGGGACG) and λ reverse (TTGACAC CAGACCAACTGGTAATG) sequencing primers were used to determine the end sequences of positive clones. Sequencing was performed using Big Dye Terminator version 2.0 with BAC modifications, and the reaction products were run on an ABI 373 automated DNA sequencer (PE-ABI). Sequences were analyzed using `BLAST` (http://www.ncbi.nlm.nih.gov/BLAST/). Conserved domain searches were performed using the PFAM database (http://pfam.wustl.edu/hmmsearch.shtml).

GeneChip Expression Profiling and Data Analysis

Expression profiling using Affymetrix GeneChips was performed as described previously (Carmel et al. 2001). Approximately 5 µg of total cellular RNA from 14 breast cell lines (Table 2) was prepared from Trizol homogenates, followed by ethanol precipitation, and used for each target. RNA was reverse transcribed into cDNA with a T7 promoter-containing primer using Superscript II (Invitrogen Life Technologies). After phenol-chloroform extraction and ethanol precipitation, the cDNA was used as a template in a biotin-labeled in vitro transcription reaction (Enzo BioArray, Affymetrix). The resulting target cRNA was purified on RNeasy columns (QIAGEN) and fragmented for hybridization to Affymetrix U95A GeneChips that contain probes representing ~12,500 human genes. Hybridization was done overnight at 45°C for 16 h on a GeneChip Hybridization Oven 640 (Affymetrix). The GeneChips were then processed on the Affymetrix GeneChip Fluidics Workstation 400, following the EukGE-WS2v4 protocol, except that only 7 µg of fragmented cRNA was added to the hybridization cocktail. The GeneChips were scanned on a Hewlett Packard GeneArray Scanner, and the results analyzed using Affymetrix `Microarray Analysis Suit` 4.0. Each chip was scaled to an arbitrarily selected average difference value of 2500. Clustering for synexpression was done using a standard Pearson correlation coefficient (`GeneSpring` 4.11; Silicon Genetics). The 4-kb upstream sequences relative to the transcription start site were retrieved from NCBI human genome database. Searches of transcription factor binding sites were done using the TRANSFAC database at a default threshold setting of 85 (Heinemeyer et al. 1998) (http://molsun1.cbrc.aist.go.jp/research/db/TFSEARCH.html).

ACKNOWLEDGMENTS

We thank Steven Ghanny for technical assistance, Martha Stempher for the AB5589 cell line, and Robert J. Donnelly of the New Jersey Medical School Molecular Resource Facility for DNA sequencing and primer synthesis. This work is supported by NIH grant CA83213 from the National Cancer Institute awarded to P.P.T. The Center for Applied Genomics is supported in part with R&D Excellence grant 00-2042-007-21 from the New Jersey Commission on Science and Technology awarded to P.P.T.

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

Aach, J., Bulyk, M.L., Church, G.M., Comander, J., Derti, A., and Shendure, J. 2001. Computational comparison of two draft sequences of the human genome. *Nature* **409:** 856–859.

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskens, R.A., Galle,

R.F., et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185–2195.
Andrews, G.K., Lee, D.K., Ravindra, R., Lichtlen, P., Sirito, M., Sawadogo, M., and Schaffner, W. 2001. The transcription factors MTF-1 and USF1 cooperate to regulate mouse metallothionein-I expression in response to the essential metal zinc in visceral endoderm cells during early development. *EMBO J.* **20:** 1114–1122.
Carmel, J.B., Galante, A., Soteropoulos, P., Tolias, P., Recce, M., Young, W., and Hart, R.P. 2001. Gene expression profiling of acute spinal cord injury reveals spreading inflammatory signals and neuron loss. *Physiol. Genomics* **7:** 201–213.
Chen, C.Y. and Shyu, A.B. 1995. AU-rich elements: Characterization and importance in mRNA degradation. *Trends Biochem. Sci.* **20:** 465–470.
Doyle, G.A., Bourdeau-Heller, J.M., Coulthard, S., Meisner, L.F., and Ross, J. 2000. Amplification in human breast cancer of a gene encoding a c-myc mRNA- binding protein. *Cancer Res.* **60:** 2756–2759.
Dyson, N. 1998. The regulation of E2F by pRB-family proteins. *Genes & Dev.* **12:** 2245–2262.
Fields, S. and Song, O. 1989. A novel genetic system to detect protein-protein interactions. *Nature* **340:** 245–246.
Futreal, P.A., Kasprzyk, A., Birney, E., Mullikin, J.C., Wooster, R., and Stratton, M.R. 2001. Cancer and genomics. *Nature* **409:** 850–852.
Hanahan, D. and Weinberg, R.A. 2000. The hallmarks of cancer. *Cell* **100:** 57–70.
Heinemeyer, T., Wingender, E., Reuter, I., Hermjakob, H., Kel, A.E., Kel, O.V., Ignatieva, E.V., Ananko, E.A., Podkolodnaya, O.A., Kolpakov, F., et al. 1998. Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. *Nucleic Acids Res.* ***26:*** 362–367.
Hoeijmakers, J.H. 2001. Genome maintenance mechanisms for preventing cancer. *Nature* **411:** 366–374.
Hogenesch, J.B., Ching, K.A., Batalov, S., Su, A.I., Walker, J.R., Zhou, Y., Kay, S.A., Schultz, P.G., and Cooke, M.P. 2001. A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes.*Cell* **106:** 413–415.
Iyer, V.R., Horak, C.E., Scafe, C.S., Botstein, D., Snyder, M., and Brown, P.O. 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409:** 533–538.
Kleiman, F.E. and Manley, J.L. 2001. The BARD1-CstF-50 interaction links mRNA 3′ end formation to DNA damage and tumor suppression. *Cell* **104:** 743–753.
Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
Lengauer, C., Kinzler, K.W., and Vogelstein, B. 1998. Genetic instabilities in human cancers. *Nature* **396:** 643–649.
Liotta, L. and Petricoin, E. 2000. Molecular profiling of human cancer. *Nat. Rev. Genet.* **1:** 48–56.
Murray, A.W. and Marks, D. 2001. Can sequencing shed light on cell cycling? *Nature* **409:** 844–846.
Nakielny, S., Fischer, U., Michael, W.M., and Dreyfuss, G. 1997. RNA transport. *Annu. Rev. Neurosci.* **20:** 269–301.
Orlando, V. 2000. Mapping chromosomal proteins in vivo by formaldehyde-crosslinked- chromatin immunoprecipitation. *Trends Biochem. Sci.* **25:** 99–104.
Pilpel, Y., Sudarsanam, P., and Church, G.M. 2001. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.* **29:** 153–159.
Pollard, A.J., Sparey, C., Robson, S.C., Krainer, A.R., and Europe-Finner, G.N. 2000. Spatio-temporal expression of the trans-acting splicing factors SF2/ASF and heterogeneous ribonuclear proteins A1/A1B in the myometrium of the pregnant human uterus: A molecular mechanism for regulating regional protein isoform expression in vivo. *J. Clin. Endocrinol. Metab.* **85:** 1928–1936.
Qian, Z. and Wilusz, J. 1993. Cloning of a cDNA encoding an RNA binding protein by screening expression libraries using a northwestern strategy. *Anal. Biochem.* **212:** 547–554.
Quellette, M.W. and Wright, W.E. 1995. Use of reiterative selection for defining protein-nucleic acid interactions. *Curr. Opin. Biotechnol.* **6:** 65–72.
Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., et al. 2000. Genome-wide location and function of DNA binding proteins. *Science* **290:** 2306–2309.
Ren, B., Cam, H., Takahashi, Y., Volkert, T., Terragni, J., Young, R.A., and Dynlacht, B.D. 2002. E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. *Genes & Dev.* **16:** 245–256.
Ross, J., Lemm, I., and Berberet, B. 2001. Overexpression of an mRNA-binding protein in human colorectal cancer. *Oncogene* **20:** 6544–6550.
Roth, F.P., Hughes, J.D., Estep, P.W., and Church, G.M. 1998. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* **16:** 939–945.
Rowley, J.D. 1998. The critical role of chromosome translocations in human leukemias. *Annu. Rev. Genet.* **32:** 495–519.
Ryan, K.M., Phillips, A.C., and Vousden, K.H. 2001. Regulation and function of the p53 tumor suppressor protein. *Curr. Opin. Cell. Biol.* **13:** 332–337.
Sambrook, J., Fritsch, E.F., and Maniatis, T. 1989. *Molecular Cloning: A laboratory manual*. 2nd edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
Shibao, K., Takano, H., Nakayama, Y., Okazaki, K., Nagata, N., Izumi, H., Uchiumi, T., Kuwano, M., Kohno, K., and Itoh, H. 1999. Enhanced coexpression of YB-1 and DNA topoisomerase II α genes in human colorectal carcinomas. *Int. J. Cancer* **83:** 732–737.
Singh, H., LeBowitz, J.H., Baldwin, Jr., A.S., and Sharp, P.A. 1988. Molecular cloning of an enhancer binding protein: Isolation by screening of an expression library with a recognition site DNA. *Cell* **52:** 415–423.
Smale, S.T. 2001. Core promoters: Active contributors to combinatorial gene regulation. *Genes & Dev.* **15:** 2503–2508.
Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9:** 3273–3297.
Tupler, R., Perini, G., and Green, M.R. 2001. Expressing the human genome. *Nature* **409:** 832–833.
Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.
Vinson, C.R., LaMarco, K.L., Johnson, P.F., Landschulz, W.H., and McKnight, S.L. 1988. In situ detection of sequence-specific DNA binding activity specified by a recombinant bacteriophage. *Genes & Dev.* **2:** 801–806.
Weinmann, A.S., Yan, P.S., Oberley, M.J., Huang, T.H., and Farnham, P.J. 2002. Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. *Genes & Dev.* **16:** 235–244.

WEB SITE REFERENCES

http://molsun1.cbrc.aist.go.jp/research/db/TFSEARCH.html; site for putative transcription factor-binding sites in a query sequence.
http://pfam.wustl.edu/hmmsearch.shtml; Pfam; protein domain site.
http://www.ncbi.nlm.nih.gov/BLAST/; sequence search site.

Received February 5, 2002; accepted in revised form May 29, 2002.