Discovering distinct genes represented in 29,570 clones from infant brain cDNA libraries by applying sequencing by hybridization methodology.

A Milosavljevic, M Zeremski, Z Strezoska, et al.

Genome Res. 1996 6: 132-141
Access the most recent version at doi:10.1101/gr.6.2.132

RESEARCH

Discovering Distinct Genes Represented in 29,570 Clones from Infant Brain cDNA Libraries by Applying Sequencing by Hybridization Methodology

Aleksandar Milosavljevic,[1,4] Marija Zeremski,[1,5] Zaklina Strezoska,[1,5] Danica Grujic,[1,6] Hristem Dyanov,[1,7] Shawna Batus,[1,3] David Salbego,[1] Tatjana Paunesku,[1] M. Bento Soares,[2] and Radomir Crkvenjakov[1,3,8]

[1]Genome Structure Group, Center for Mechanistic Biology and Biotechnology, Argonne National Laboratory, Argonne, Illinois 60439; [2]Department of Psychiatry, College of Physicians and Surgeons of Columbia University and New York State Psychiatric Institute, New York, New York 10032

To discover all distinct human genes and to determine their patterns of expression across different cell types, developmental stages, and physiological conditions, a procedure is needed for fast, mutual comparison of hundreds of thousands (and perhaps millions) of clones from cDNA libraries, as well as their comparison against data bases of sequenced DNA. In a pilot study, 29,570 clones in duplicate from both original and normalized, directional, infant brain cDNA libraries were hybridized with 107–215 heptamer oligonucleotide probes to obtain oligonucleotide sequence signatures (OSSs). The OSSs were compared and clustered based on mutual similarity into 16,741 clusters, each corresponding to a distinct cDNA. A number of distinct cDNAs were successfully recognized by matching their 107-probe OSSs against GenBank entries, indicating the possibility of sequence recognition with only a few hundred randomly chosen oligomers.

An intermediate and currently feasible step in the Human Genome Project is the sequencing of cDNA fragments, which are referred to as expressed sequence tags, (ESTs) (Adams et al. 1991).

EST strategy relies on a one-at-a-time random, direct sampling of cDNA libraries, with the result that every second to third sequence is needlessly resequenced if directly prepared libraries are used. As the number of sequenced cDNAs grows, the resequencing problem will inevitably worsen. To reduce resequencing, libraries are typically normalized by biochemical procedures (Soares et al. 1994). In the normalized libraries, the relative abundances of the most frequent and of the rarest cDNAs are equalized to a large degree, increasing ≤20-fold the chance of finding the rarest cDNAs (Soares et al. 1994). Although the normalization improves the chance of finding a new gene by two- to threefold (Drmanac at al. 1994; Soares et al. 1994), the total number of clones that need to be sequenced to compile a nearly complete catalog of human genes would still exceed ≤10-fold the total number of distinct genes. The EST resequencing problem is exacerbated by the fact that many genes are expressed in multiple libraries. Two cDNA clones may not even be recognized as originating from the same mRNA by the EST strategy because the sequenced fragments may not overlap.

To quantitatively study the expression of genes across different cell types, developmental stages, and physiological conditions, hundreds of thousands (and perhaps millions) of clones from a number of potentially highly redundant non-normalized cDNA libraries must be comparatively studied. Redundancy may also be profitably employed for complete cDNA sequencing: An average 3- to 6-kb mRNA can be efficiently sequenced by current methods based on ≤10

Present addresses: [3]Hyseq, Inc., Sunnyvale, California 94086; [4]CuraGen Corp., Branford, Connecticut 06405; [5]Department of Genetics, University of Illinois at Chicago, Chicago, Illinois 60612; [6]Department of Medicine RN 320, Beth Israel Hospital, Boston, Massachusetts 02215; [7]2562 Chelsea Drive, no. 101, Woodridge, Illinois 60517.
[8]Corresponding author.
E-MAIL crk@sbh.com; FAX (408) 524-8141.

6:132–141 ©1996 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/96 $5.00

overlapping cDNAs from various libraries. This all puts a premium on a method that would enable rapid and economical mutual clone comparisons and comparisons of clones against already sequenced DNA.

It has been estimated that oligomer sequence signatures (OSSs) consisting of 100–1000 probes would make possible precise mutual comparisons of clones, as well as comparisons of clones against DNA sequence data bases (Drmanac et al. 1991; Lenon and Lehrach 1991). The hybridization data production lines that are being developed for the purpose of sequencing by hybridization (SBH) can be employed easily for rapid generation of OSSs (Drmanac et al. 1992, 1993; Meier-Ewert et al. 1993; Drmanac and Drmanac 1994; Grujic et al. 1994). Densely arrayed clones can be examined at a 10- to 100-fold higher rate and much more economically than by standard sequencing, thus achieving the data throughputs required for exhaustive gene discovery and for detailed studies of expression patterns.

Here we present a pilot study in which the newly developed methods for mutual clone comparisons (Milosavljevic et al. 1995) and for comparisons of clones against known DNA sequences (Milosavljevic 1995) have been applied in an analysis of 29,570 cDNAs from the recently developed human infant brain cDNA libraries (Soares et al. 1994). For this study, $>10^7$ individual clone/probe hybridization measurements were collected. A SBH-based study utilizing different data analysis methods confirms our findings about the structure of these cDNA libraries (S. Drmanac, N.H. Stavropoulos, I. Labat, J. Vonan, B. Hauser, M.B. Soares, and R. Drmanac, in prep.). Our joint results demonstrate unique opportunities in genome-scale cDNA analysis that are afforded by large-scale hybridization experiments.

RESULTS

Reliability of OSS Clustering

Hybridization experiments, OSS scaling, and clustering were performed as described in Methods. The clustering algorithm grouped the clones into disjointed clusters based on mutual similarity of their OSSs.

Two independent approaches were applied to estimate clustering error: (1) Individual cDNAs were spotted in duplicate and (2) groups of highly overlapping control clones of known sequence were spotted along with the cDNAs. The degree of false separation into disjointed clusters of signatures that come from identical or highly overlapping clones, as well as the degree of false joining of nonoverlapping control clones into identical clusters, was used as an estimate of clustering error.

The clustering error, estimated as the percentage of dots that do not occur together with their duplicate in the same cluster, ranges from 1% on small filters to 5.5% on medium filters (Table 1). These results indicate the high reproducibility of hybridization experiments and the low clustering error rate for all filter formats and array densities. (The actual failure rate of clustering is even smaller because the signatures whose duplicates were eliminated owing to missing values were also counted as clustering failures. The slight loss of accuracy on medium and large filters may be justified by an ~10-fold increase in hybridization throughput per filter.)

In addition to the cDNAs, each of the filters contained several independent amplifications of a set of 46 clones of known sequence as controls (Pizzuti et al. 1992). The 1- to 2-kb control clones cover a 12-kb portion of the human dystrophin gene intron segment. The dots that correspond to each of these control clones were obtained from

Table 1. Summary of Filter Types Used in Hybridization Experiments

Filter format	Total types	Dots/filter	Dots scored	cDNAs scored	False separation of repeated dots (%)
Small	8	3,456	24,353	11,078	1.02
Medium	2	7,776	13,461	6,810	5.49
Large	1	31,104	23,448	11,682	4.45

A fraction of dots spotted on each type of filter contained control clones: 11.1% on small filters, 3.7% on medium filters, and 1.25% on large filters.

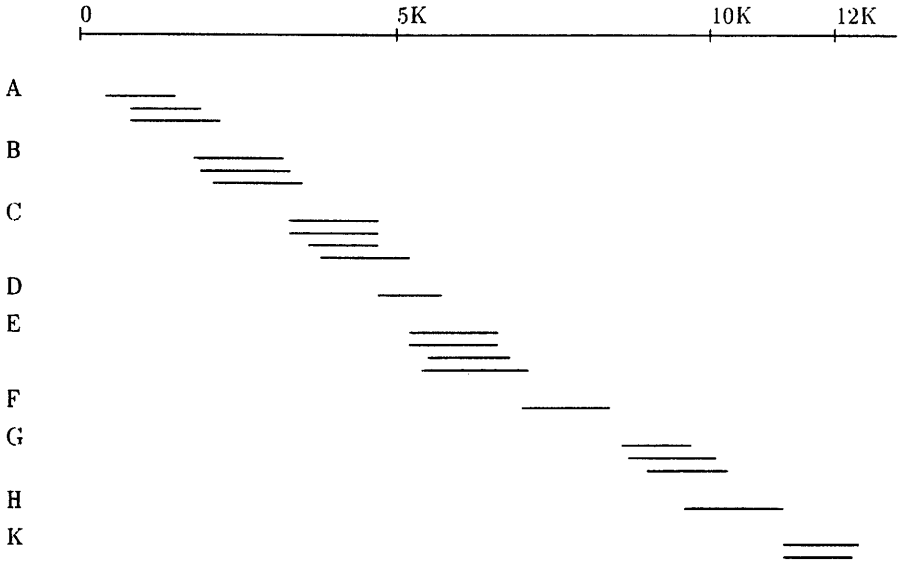

Figure 1 The set of control clones from the dystrophin gene intron used as a benchmark for clustering. The clones are selected so that there are nine groups of clones (denoted A–K) with a 50% overlap within groups and <50% overlap across groups.

a number of independent amplifications. A subset of these clones that are actually used as a benchmark is depicted in Figure 1. The control clones enabled us to test whether identical or highly similar clones are grouped together by the clustering algorithm despite the fact that they are spotted on different filters and hybridized separately. The clustering of control clones spotted on all eight small filter types is summarized in Table 2. Only 0.13% of nonoverlapping clones were falsely joined together into the same cluster, and only 4.5% of overlapping clones were falsely split across different clusters.

The low rate of false joining is further supported by the fact that none of the clusters in the final clustering of all the signatures contained both a cDNA clone and a control clone. This was expected because the control clones come from an intronic region. Because tens of thousands of cDNA clones were clustered together with a thousand control clones, a probability of false joining is <1 in 10 million pairwise comparisons.

The low error rate indicates the ability of the clustering algorithm to recognize significantly overlapping or homologous clones even if they are spotted on different filters. The observed grouping of overlapped control clones demonstrates that contigs for complete sequencing of long cDNAs can be assembled.

The pattern of false splitting (Table 2) indicates that for most large clusters there are a few "satellite" clusters containing apparently dissimilar signatures. The rate of 4.5% indicates that the number of distinct cDNA clones in the final clustering experiment may be slightly overestimated. A number of small satellite cDNA clusters may contain clones that are highly similar to clones from a larger cluster but are not detected as such.

Abundance Structure of cDNA Libraries

The next step was the clustering of clones from human infant brain cDNA libraries. The measured abundance structures of 11,078 independent clones from the original library and 10,340 from the normalized library are shown in Table 3. The average abundance of an individual cDNA was 0.02% for the original library and 0.01% for the normalized; the original library contained threefold more clusters that achieve abundance of 0.1% or more. Apparently, normalization did not affect the total number of moderate- and low-abundance clusters while at the same time it significantly reduced the number of high-abundance clusters. Clone identification (described below) allowed us to show that the normalization was extremely successful for the most frequent mRNAs: The abundances for α-tubulin, elongation factor α1, and cytoskeletal γ-actin mRNAs fell 35-, 32-, and 10-fold, respectively (Table 4). Our results indicate that the biochemical normalization did not increase the number of distinct cDNAs in a randomly drawn sample of ~10,000 clones by more than a factor of 1.9. An approximately twofold increase is also expected on statistical grounds (Drmanac et al. 1994).

Discovering Distinct Genes

The goal of the final cDNA clustering experiment was to count the total number of distinct genes represented in our sample by clustering together signatures of all 29,570 clones from both original

Table 2. Clustering of Control Clones

	A	B	C	D	E	F	G	H	K	Total
1.	131									131
2.	2									2
3.		126								126
4.		7								7
5.		2								2
6.			150							150
7.			2							2
8.			2							2
9.			2							2
10.			5							5
11.			2							2
12.			2					1		3
13.				52						52
14.					164					164
15.					1					1
16.					7					7
17.						44				44
18.							148			148
19.							5			5
20.								40		40
21.									89	89
22.									4	4
Discarded	23	20	28	4	37	10	9	7	15	153
Total	156	155	193	56	209	54	162	48	108	1141

Clustering of control clones depicted in Fig. 1. Columns correspond to groups of overlapping clones. Rows correspond to clusters obtained by the algorithm based on signatures obtained from all eight small filters. The error of false separation, estimated by dividing the number of clones in splinter clusters by the total number of clones outside "garbage" clusters is 4.5%. Theoretical analysis presented in Milosavljevic et al. (1995) can be used to show that the probability of false joining of two signatures is at most 10^{-5}, which is consistent with our experimental result: one false joining (0.13%) in cluster 12 resulted from >100,000 pairwise comparisons. However, even the single false joining may be a result of sample contamination rather than clustering error.

and normalized libraries. The clustering algorithm produced 16,741 clusters, each corresponding to a distinct cDNA. A total of 12,363, or 74%, of all clusters contained single clones. Some of these singletons might represent nuclear leakage or cloning artifacts.

To show further that the cDNA clustering procedure is reliable, cDNAs from several clusters were also sequenced from both ends. The sequencing confirmed homogeneity within clusters. For example, end sequencing of four clones from a γ-actin cluster and five clones from elongation factor α1 cluster revealed that the clones from the same cluster share essentially the same sequence, except that 5′ ends of individual clones started at variable positions within the gene (≥200 bp apart).

cDNA Sequence Recognition

A further step in the characterization of cDNAs was a systematic recognition of the genes that give rise to the identified clusters. For that purpose, the lists of oligomers that were putatively identified as occurring in particular clones were compared against known DNA sequences applying a newly developed sequence recognition method (Milosavljevic 1995) described in Methods. Owing to the small number of oligomer probes that were hybridized, the search was limited to the most abundant genes; a data base consisting of oligomer lists for the 100 most frequent cDNAs (as determined by cluster size) from the original (non-normalized) infant brain library were queried by sequences of 195 genes expected

Table 3. cDNA Abundances

Abundance class (%)	cDNA species: calculated for rat brain library	cDNA species: human infant brain library original (11,078 clones)	cDNA species: human infant brain library normalized (10,340 clones)
>1	13	8	0
0.5	11	13	4
0.2	23	84	21
0.1	146	150	51
0.05	456	443	429
0.02	290	668	1032
<0.01	—	3529	6703
Total species	—	4895	8246
Average abundance	—	0.02	0.01

cDNA abundances in original and normalized infant brain libraries, as compared with calculated abundances for rat brain (Table 7 in Milner and Sutcliffe 1983). The abundances are estimated within a factor of 2; e.g., some of the 84 cDNAs species assigned to abundance class 0.2 may belong to abundance classes 0.1 or 0.5. For the purpose of average abundance calculation, abundances >1% and <0.01% are rounded off to 1% and 0.01%, respectively. Low abundances could not be calculated precisely in the rat brain study because of small sample size.

to be highly to moderately expressed in the brain, as described in Methods. All 21 matches that gave a relative score of 10 bits or more were considered further. It was hypothesized that these matches are attributable either to sequence identity or to high sequence similarity.

To test the putative identifications, the 5′ ends of 18 clones (average length of 300 bp) were obtained by single-pass sequencing on an ABI sequencer; the three remaining clones (of a total of 21) could not be sequenced because of technical problems. The sequences were then used in a BLAST search against GenBank. The sequenced fragments, BLAST matches, and the hypothesized sequences were then aligned pairwise. In 11 cases (of a total of 18), the BLAST search confirmed our putative identification (Table 5).

The main reason why the remaining seven identifications were not confirmed by BLAST searches was the incompleteness of our set of 195 selected genes. In four of the seven cases, identical sequences were present in GenBank but absent from our selected set; consequently, BLAST discovered the four exact matches whereas our search resulted in four significant similarities, which could be confirmed by alignments of sequenced fragments and hypothesized sequence. There was only one case where a sequence returned by BLAST search was also present in our selected set but was not recognized.)

In addition to the GenBank search, the gel-sequenced fragments were also used as queries in a search of dbEST. All of the searches identified either highly similar or identical sequences, indirectly confirming that we have chosen frequently expressed genes for our comparisons.

Cross-correlation with Independently Obtained OSSs and cDNA Abundance Data

As a preliminary test of the possibility of cross-correlation of oligonucleotide sequence signatures (OSS) data across different laboratories, the hybridization experiments reported by S. Drmanac, N.H. Stavropoulos, I. Labat, J. Vonau, B. Hauser, M.B. Soares, and R. Drmanac, (in prep) were coordinated with our experiments so that a portion of clones from the other study were hybridized with the same set of 107 probes. A total of ~22,000 signatures from the two laboratories were clustered together using the method described by Milosavljevic et al. (1995). A significant degree of cross-correlation was confirmed by the correct grouping of control clones of known sequence from the two data sets (data not shown). Correspondence between a number of clusters obtained in the two studies could be established, and the abundance information could be integrated, as shown in Table 4.

Table 4. Abundances of Identified cDNAs (%)

Clone	Original		Normalized	Original/ normalized
	OSS (%)	gel (%)	OSS (%)	
HUMTABAK, α-tubulin	2.58/1.7*	2.7	0.074/0.082*	35
HUMEF1AR, elongation α1-factor	2.83/2.5*	3.1	0.088/0.033*	32r
HSACTCGR, cytoskeletal γ-actin	0.47/0.24*	0.43	0.022*	10*
HUMCOR2M, cytochrome bc1	0.044*		0.072*	0.61*
HUMVDAC2X, voltage dependent channel	0.018/0.022*		0.066*	0.33*
RNU03417, olfactomedin	0.045/0.044*		0.299*	0.15*
HUMMTCG, mitochondrial, 1700–2100 bp	0.72/0.48*		0.24/0.30*	2.9/1.6*
HUMHSP90R, 90-kD heat shock protein	0.180	0.18	0.027	6.7
HUMUB, ubiqutin	0.14	0.06	0.022	6.5
HUMTHRA2A, thyroid hormone receptor	0.14	0.12	0.011	13
HUMGNPAS, G(s) α	0.14	0.06	0.016	8.4
HUMGBR, G(s) β	0.14			
HUMARF1BA, ADP-ribosylation factor	0.091	0.12	0.011	8.3
HUMCAM, calmodulin	0.091	0.43	0.038	2.4
HUMHEXKIN, hexokinase 1	0.081	0.06	0.005	16.2
HSMMAR, macmarcks mRNA	0.072			
HSEF1GMR, elongation factor 1-γ subunit	0.26	0.24		
HUMXT00951, anonymous EST	0.081		0.011	7.4
HUMMTTRNA, mitochondrial tRNA genes, 2660–3100 bp	0.099			

Abundances of identified clones in original and normalized cDNA libraries, as determined by OSS clustering and gel-based sequencing (Adams et al. 1993). Asterisks denote abundances estimated based on clustering of OSSs obtained by S. Drmanac, N.A. Stavropoulos, I. Labat, J. Vonau, B. Hauser, M.B. Soares, and R. Drmanac (in prep.); asterisks in the last column were obtained as ratios of values from columns 2 and 4. The missing values could not be estimated reliably because of missing data, uncertain correspondences, and other problems.

A number of clusters were identified by gel sequencing of individual members and by subsequent data base searches; by cross-correlating the data from the two laboratories, a number of clusters could be identified without sequencing. Moreover, the published abundance information that is obtained through EST gel sequencing of the same library could be compared against abundances obtained by our clustering experiments. The results (shown in Table 4) indicate an excellent agreement between EST gel sequencing and our results on the abundances of elongation factor α1, α-tubulin, and γ-actin cDNAs, the first two being the most prevalent species in the library. The sevenfold smaller sample size in the gel-based study does not allow a comparison of the abundances of remaining cDNAs. The extent of agreement of the abundances obtained from the independent hybridization experiments of the two laboratories demonstrates the possibility of cross-correlation of independently obtained OSSs.

Comparison of cDNA Abundance Data with Independently Obtained Estimates for Rat Brain

To obtain a correct estimate of a gene expression pattern, it is important to use cDNA libraries that do not exhibit cloning and amplification bias and contain a low percentage of contaminating sequences. Several EST studies (Kahn et al. 1992; Adams et al. 1993; Matsubara and Okubo 1993) have found that the libraries used in our study show lower bias and contamination than that of other libraries examined. To check for a possible cloning bias cDNA, abundance measurements were compared with the earlier estimates for rat brain obtained on mRNA directly (Milner and Sutcliffe 1983). Table 3 summarizes the results of the comparison. Assuming similar abundance structure in the rat and the original human libraries, the agreement between the abundances of highly and moderately frequent cDNAs indicates that the original library does not exhibit any major bias.

Table 5. Recognition of Sequence Identities and Similarities

Hypothesized	Top GenBank score
Exact identifications	
1. HUMEFIAR, elongation factor α1	HUMEFIAR, elongation factor α1
2. HSMMAR, MacMarcks mRNA	HSMMAR, MacMarcks mRNA
3. HUMCAM, calmodulin	HUMCAM, clamodulin
4. HUMHEXKIN, hexokinase 1	HUMHEXKIN, hexokinase 1
5. HUMTHRA2A, thyroid hormone α-receptor	HUMTHRA2A, thyroid hormone α-receptor
6. HUARF2BA, ADP-ribosylation α-factor	HUARF2BA, ADP-ribosylation α-factor
Significant similarities	
7. HSEFIGMR, elongation γ1-factor	HUMPANCAN, pancreatic tumor-related protein
8. HUMGFAP, glial fibrilary protein	RNDP150, rat dynein-associated protein
9. HUMBADPT, adaptin b	HUMHBP, HDL binding protein
10. HUMG19P1A, 80K-H protein	anonymous EST
11. HUMCAMPPK, cAMP-d kinase Ia	HUM4AI, initiation factor 4 Ia

Recognition of sequence identities and similarities by comparison of 100 oligomer lists against 195 GenBank entries. Of 18 putative recognitions, 6 identities and 5 significant similarities were confirmed by single pass gel sequencing and subsequent BLAST search of GenBank. An identity was considered confirmed if it occurred as the top-scoring entry of a BLAST search. The numbers of mutations indicated under Sequence similarities include single-pass sequencing errors.

DISCUSSION

The partial cDNA sequence information obtained by SBH methodology and the standard gel-based sequencing of ESTs do not give the same kind of sequence information: Hybridization experiments sample the entire cDNA sequence, whereas an average gel-based sequencing experiment gives information about only one-fifth of a full-length cDNA.

We should emphasize that the clustering of clone sequence signatures obtained by SBH methodology does entail a certain degree of probabilistic error. However, it has been shown (Milosavljevic et al. 1995) that the error diminishes rapidly with new hybridization experiments (larger OSSs). In the experiments described here, ~5%–10% of the total number of cDNAs may be assigned erroneously to small clusters that are separate from a large true cluster. On the other hand, false joining of different cDNAs into the same cluster was virtually absent.

The clustering accuracy demonstrated here suffices for efficient selection of cDNA clones for further characterization: For example, by picking a single representative from each cluster, a catalog of cDNAs for complete sequencing can be created; all distinct cDNAs would be represented in such a catalog and only 5%–10% would appear more than once. This procedure may be viewed as in silico normalization of cDNA libraries.

The results shown in Table 5 demonstrate first successful recognition of cDNA sequences based on hybridization experiments. It was predicted that signatures obtained by hybridization with a few hundred probes would suffice for accurate data base matching (Drmanac et al. 1991; Lenon and Lehrach 1991). The initial successes in recognizing cDNA sequences described in this paper, as well as experiments with control clones (Milosavljevic 1995), strongly support this early prediction.

An interesting application of the sequence recognition method is the identification of already sequenced cDNAs. As the number of sequenced cDNAs grows, an ever-increasing number of cDNAs will become recognizable by this method, thus obviating the need for the much more costly and time-consuming gel-based sequencing. This procedure may be viewed as in silico subtraction of already sequenced clones from cDNA libraries. The partial sequence data obtained by the SBH and EST approaches are presently not commensurate, because SBH probes sample the entire cDNA sequence while the single-pass sequencing on average gives information only about one-fifth of a full-length cDNA. However, the contigs of overlapped ESTs (Adams et al. 1995), if of a sufficient length, could be used to cross-reference the results of the two methods on the same libraries.

METHODS

Hybridization Experiments

Clones from the original and normalized human infant brain cDNA libraries (Soares et al. 1994) were arrayed and immobilized on nylon filters in the form of PCR products of plasmid inserts by applying previously described techniques (Drmanac and Drmanac 1994; Drmanac et al. 1992, 1993, in prep.; Grujic et al. 1994).

In the following, we say that two physical filters are of the same type if they contain the same cDNAs that are spotted using the same pattern; that is, two replicas of filters belong to the same filter type. Each type of filter contained a set of clones spotted in duplicate. A total of 11 filters was created (see Table 1): Eight small filter types contained 3456 dots each, arrayed on a 8 × 12-cm surface; two medium filter types contained 7776 dots each, arrayed on a 8 × 12-cm surface; and one large filter type contained 31,104 dots arrayed on a 16 × 24-cm surface.

A common set of 107 heptamers was hybridized with each of the 11 distinct filter types. Figure 2 presents an example of a hybridization experiment. Several physical copies of each type of filter were prepared to run parallel hybridization experiments. Each physical copy of a particular type of filter was hybridized with a different subset of heptamer probes (probe list available on request). The results obtained with all filter replicas were pooled to give OSSs consisting of hybridization intensities for the entire probe set.

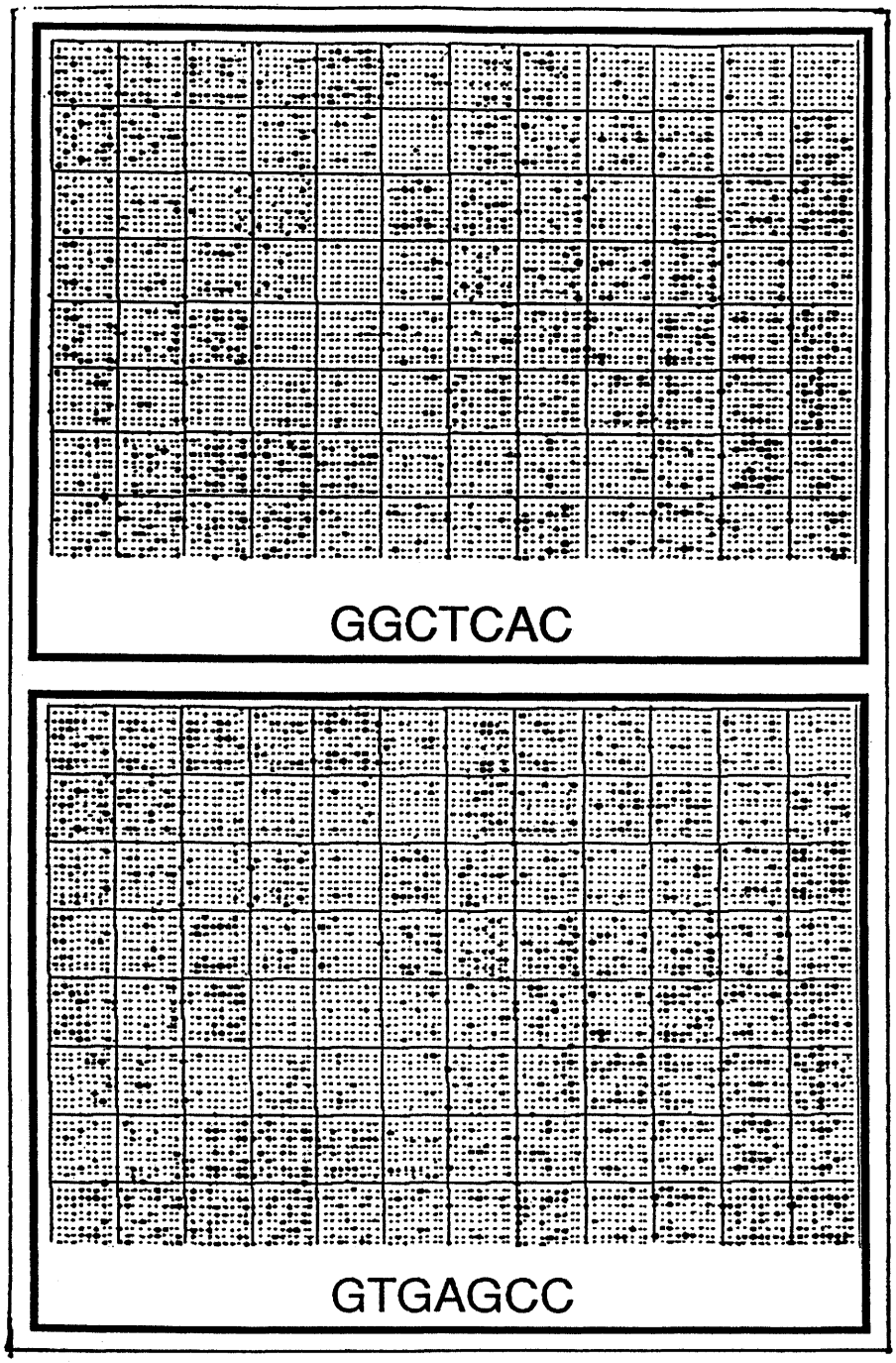

Figure 2 The 8 × 12-cm filter replicas containing 7776 dots hybridized with two complementary heptamer probes GGCTCAC and GTGAGCC. The computer-drawn grid on the scanned image facilitates visualization of duplicate samples displaced by five rows within the same square and column (except for every fifth row within each square, which contains duplicates in horizontal displacement). The high coincidence of hybridization intensities of duplicates, as well as the high degree of congruency across the two images, indicate a high degree of reproducibility of individual hybridization experiments.

Scaling and Clustering of OSSs

To achieve reproducibility of individual clone/probe hybridization intensities across different filters and despite variations in experimental conditions, two scaling steps were performed: Mass-scaling provides reproducibility across different dots on the same filter where as rank-scaling provides reproducibility across different filters.

To estimate relative molarity in individual dots, each

physical copy of a filter was hybridized with a mass probe, which consisted of an oligomer that was complementary to the primer region of cDNA PCR products. In the mass-scaling step, to factor out differences in the molarity of the cDNA in individual dots, the hybridization intensities of each probe were divided by the hybridization intensities of the mass probe. The dots that did not give hybridization intensities with the mass probe above a prespecified threshold were considered empty and all hybridizations with them were discarded from further analysis.

In the rank-scaling step, the mass-scaled intensities of each dot were replaced by their rank value among all others on the same filter. Assuming that each filter contains large numbers of clones that are randomly picked from the same library, the rank-scaling step provides reproducibility across different filters, despite the usual variations in experimental conditions (for more details, see Milosavljevic et al. 1995). Sequence signatures for particular clones were compiled by pooling together rank-scaled hybridization intensities across different physical copies of a particular filter type. The signatures that were missing >25% of the hybridization values (owing to empty dots) were discarded from further analysis. A clustering algorithm (Milosavljevic et al. 1995) was then applied to group the signatures into disjointed clusters according to their mutual similarities.

DNA Sequence Recognition

The DNA sequences were recognized by comparing a list of oligomers that were identified to occur in them against known DNA sequences.

For each hybridization signature consisting of 107 hybridization intensities, a list consisting of the 28 oligomer probes (roughly one-quarter of 107) that exhibit the highest intensities was compiled. The list of 28 oligomers was augmented by the additional 28 reverse complementary oligomers (because both strands of the PCR products were hybridized and the orientation of oligomers could not be resolved).

To identify the most frequent cDNAs, the clones from the original library were clustered at high stringency (to maximize homogeneity of clones within clusters) and a single representative from the 100 largest clusters was selected. The corresponding 100-oligomer lists were used as a data base for searches using known DNA sequences as queries. A total of 195 searches involving gene sequences of average length of 2.5 kb were performed against the data base consisting of the 100-oligomer lists.

The significance of similarity between a sequence and a list was determined by the algorithmic significance method: For each sequence and a candidate oligomer list, it was estimated how many bits of information about the sequence were revealed by the list. Every bit of information implies a twofold improvement in significance value of the particular match (for details, see Milosavljevic 1995). Two parameters were considered for each query sequence: the top score with a particular oligomer list, and the difference between the top score and the second highest score, which were termed absolute and relative scores, respectively. A few inconsistencies (different sequences matching the same clone) were resolved by taking the one with the highest absolute score.

ACKNOWLEDGMENTS

We thank R. Drmanac for sharing SBH methods before publication and for use of the hybridization data of his group. Donna Muzny and Richard Gibbs of the Baylor College of Medicine Human Genome Center kindly provided prerelease sequence data and M13 clones used in the experiments. We also thank David E. Nadziejka for editorial assistance. cDNA library construction was supported by grant No. DE-FG02-91ER61233 from the U.S. Department of Energy (DOE) to M.B.S. This work was supported by the U.S. DOE, Office of Health and Environmental Research, under contract W-31-109-ENG-38. The sequence recognition software is available from Argonne Industrial Technology Development Center under accession number ANL-SF-94080. For help in its use and further information contact A.M.

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

Adams, M., J. Kelley, J. Gocayne, M. Dubnick, M. Polymeropoulos, H. Xiao, C. Merril, A. Wu, B. Olde, R. Moreno, A. Kerlavage, W. McCombie, and J.C. Venter. 1991. Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* **252:** 1651–1656.

Adams, M., M. Soares, A. Kerlavage, C. Fields, and J.C. Venter. 1993. Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nature Genet.* **4:** 373–380.

Adams, M., A. Kerlavage; [76 authors], and J.C. Venter. 1995. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* (Suppl.) **377:** 3–17.

Drmanac, S. and R. Drmanac. 1994. Processing of cDNA and genomic kilobase-sized clones for massive screening, mapping and sequencing by hybridization. *BioTechniques* **17:** 328–336.

Drmanac, R., G. Lennon, S. Drmanac, I. Labat, R. Crkvenjakov, and H. Lehrach. 1991. Partial sequencing by hybridization: Concept and applications in genome analysis. In *The First International Conference on Electrophoresis, Supercomputing and the Human Genome*, pp. 60–74. World Scientific, Singapore, Malaysia.

Drmanac, R., S. Drmanac, I. Labat, R. Crkvenjakov, A. Vicentic, and A. Gammell. 1992. Sequencing by hybridization: Towards an automated sequencing of one million M13 clones arrayed on membranes. *Electrophoresis* **13:** 566–573.

Drmanac, R., S. Drmanac, I. Labat, A. Vicentic, A. Gemmell, N. Stavropoulous, and J. Jarvis. 1993. SBH and the integration of complementary approaches in the mapping, sequencing, and understanding of complex genomes. In *The Second International Conference on*

Bioinformatics, Supercomputing, and Complex Genome Analysis. pp. 120–134. World Scientific, Singapore, Malaysia.

Drmanac, R., S. Drmanac, I. Labat, and N. Stavropoulous. 1994. Requirements in screening cDNA libraries for new genes and solutions offered by SBH technology. In *Identification of transcribed sequences* (ed. U. Hochgeschwender and K. Gardiner), pp. 239–251. Plenum Press, New York, New York.

Grujic, D., Z. Strezoska, and R. Crkvenjakov. 1994. High throughput PCR procedure for up to 6-kb lengths of DNA, *BioTechniques* **17:** 291–294.

Khan, A., A. Wilcox, M. Polymeropoulos, J. Hopkins, T. Stevens, M. Robinson, A. Orpana, and J. Sikela. 1992. Single pass sequencing and physical and genetic mapping of human brain cDNAs. *Nature Genet.* **2:** 180–185.

Lennon, G., and H. Lehrach. 1991. Hybridization analyses of arrayed cDNA libraries. *Trends Genet.* **7:** 314–317.

Matsubara, K. and K. Okubo. 1993. cDNA analyses in the human genome project. *Gene* **135:** 265–274.

Meier-Ewert, S., E. Maier, A. Ahmadi, J. Curtis, and H. Lehrach. 1993. An automated approach to generating expressed sequence catalogues. *Nature* **361:** 375–376.

Milner, R. and J. Sutcliffe. 1983. Gene expression in rat brain. *Nucleic Acids Res.* **11:** 5497–5520.

Milosavljevic, A. 1995. DNA sequence recognition by hybridization to short oligomers. *J. Comput. Biol.* **2:** 355–370.

Milosavljevic, A., Z. Strezoska, M. Zeremski, D. Grujic, T. Paunesku, and R. Crkvenjakov. 1995. Clone clustering by hybridization. *Genomics* **27:** 83–89.

Pizzuti, A., M. Pieretti, R. Fenwick, R. Gibbs, and C.T. Caskey. 1992. A transposon-like element in the deletion-prone region of the dystrophin gene. *Genomics* **13:** 594–600.

Soares, M., M. Bonaldo, P. Jelene, L. Su, L. Lawton, and A. Efstratiadis. 1994. Construction and characterization of a normalized cDNA library. *Proc. Natl. Acad. Sci.* **91:** 9228–9232.

Received October 31, 1995; accepted in revised form February 6, 1996.