Original Contributions

Mammalian Genome

Mammalian Genome 3: 609–619, 1992

# Hybridization fingerprinting of high-density cDNA-library arrays with cDNA pools derived from whole tissues

**Thomas M. Gress,* Jörg D. Hoheisel, Gregory G. Lennon, Günther Zehetner, and Hans Lehrach**

Imperial Cancer Research Fund, P.O. Box 123, 44 Lincoln's Inn Fields, London WC2a 3PX, UK

Received May 25, 1992; accepted July 2, 1992

**Abstract.** As part of an integrated mapping and sequencing analysis of genomes, we have developed an approach allowing the characterization of large numbers of cDNA library clones with a minimal number of experiments. Three basic elements used in the analysis of cDNA libraries are responsible for the high efficiency of this new approach: (1) high-density library arrays allowing thousands of clones to be screened simultaneously; (2) hybridization fingerprinting techniques to identify clones abundantly expressed in specific tissues (by hybridizations with labeled tissue cDNA pools) and to avoid the repeated selection of identical clones and of clones containing noncoding inserts; and (3) a computerized system for the evaluation of hybridization data. To demonstrate the feasibility of this approach, we hybridized high-density cDNA library arrays of human fetal brain and embryonal *Drosophila* with radiolabeled cDNA pools derived from whole mouse tissues. Fingerprints of the library arrays were generated, localizing clones containing cDNA sequences from mRNAs expressed at middle to high abundance (>0.1–0.15%) in the respective tissue. Partial sequencing data from a number of clones abundantly expressed in several tissues were generated to demonstrate the value of the approach, especially for the selection of cDNA clones for the analyses of genomes based on expressed sequence tagged sites. Data obtained by the technique described will ultimately be correlated with additional transcriptional and sequence information for the same library clones and with genomic mapping information in a relational database.

## Introduction

A goal of the genome project is the understanding of the genetic information of organisms. One route to achieve this is the analysis of genomic DNA by physical mapping and sequencing. However, it is not yet practical to sequence the entire genome and not possible to identify the genomic sequence elements regulating the pattern of transcription of sequences in different cell types. Therefore, it is necessary to experimentally obtain information on the pattern of expression. An analysis of transcribed sequences offers a significant enrichment for informative sequences and can more easily be correlated with the level of function, the protein. As an essential part of the analysis of the genome, we have started to analyze libraries of cDNA clones with the goal of identifying their pattern of transcription, to determine partial sequence information, and ultimately to correlate the genomic mapping, transcriptional and sequence information into a global data set (Lehrach et al. 1990).

To obtain information on the pattern of transcription experimentally, one can follow two approaches: (1) a determination of the number of clones of a specific type in cDNA libraries by oligomer fingerprinting (digital) and (2) the analysis of the normalized intensity of hybridization of labeled cDNA to cDNA clones (analog). In the present study, we report the results of applying the second approach. This method has allowed the characterization of vast numbers of cDNA clones with techniques that proved successful for genomic mapping purposes (Lehrach et al. 1990; Nizetic et al. 1991; Hoheisel et al. 1991a,b) based on the use of library arrays spotted at high clone density with the help of a robotic device. Hybridization fingerprinting analysis with total cDNA pools derived from different tissues was used to identify clones on the cDNA library arrays containing mRNA sequences expressed at middle to high abundance. Difficulties usually encountered when using tissue cDNA pool probes to screen cDNA libraries can be eliminated by control

The nucleotide sequence data reported in this paper have been submitted to the EMBL database and have been assigned the accession numbers X65374–X65393 and X65268–X65275.

**Present address:* Medizinische Klinik I, Universität Ulm, Robert Koch Str. 8, 7900 Ulm, FRG

*Offprint requests to:* T.M. Gress

experiments; a number of relevant problems must be considered. 97% of the mass of mammalian brain PolyA-mRNA is comprised of a maximum of 20% of the PolyA-mRNA complexity (Bantle and Hahn 1976), so that identical, abundantly expressed clones are repeatedly selected. Each cDNA pool hybridization produces a bulk of data, which can prove difficult to analyze, thus limiting the total number of clones that can be assessed in one single experiment (Derman et al. 1981; Crampton et al. 1980). Finally, complex probes usually generate a high amount of background and unspecific hybridization. Repetitive sequences from nuclear RNA (Hochgeschwender et al. 1989) and polyA-tails (Dworkin and Dawid 1980; Sargent and Dawid 1983) in cDNA pool probes make the interpretation of hybridization results difficult.

Our approach can be used to select library clones containing cDNA sequences from mRNAs of specific abundance classes for a variety of purposes, e.g., in the context of the Human Genome Project for the generation of expressed sequence-tagged sites (ESTs; Adams et al. 1991, 1992). To test the system on a limited number of clones, we decided to start with the analysis of clones abundantly expressed in several tissues that most likely code for proteins involved in structural and regulatory functions in every cell. Partial sequencing data for a number of clones was generated by a single run of double-stranded sequencing from the 5′ end of the clones. Results obtained in the fingerprinting and partial sequencing experiments can be correlated with additional data available, or produced for the same cDNA library clones as part of our integrated approach for the analysis of genomes (e.g., fingerprinting with simple and complex probes, as summarized in Lennon and Lehrach 1991) and to our genomic mapping data (Lehrach et al. 1990).

## Materials and methods

Unless stated otherwise, all standard experimental procedures were performed according to Sambrook and colleagues (1989).

### *cDNA libraries*

Two cDNA libraries were available for these studies, one prepared from 0- to 4-h *Drosophila* embryos (Brown and Kafatos 1988; Hoheisel et al. 1991a) and one prepared from human fetal brain (HFB; Lennon and Lehrach 1991). Both libraries were originally primed at the 3′ end with oligo d(T). Clones were individually transferred into 96-well microtiter dishes. Clones of 96 microtiter dishes were spotted by a robotic device onto 22 × 22-cm Hybond N plus membranes (Amersham). Filters were incubated and processed as described previously (Nizetic et al. 1991; Hoheisel et al. 1991a,b). Two sets of HFB cDNA-filters (18432 clones total) and one set of *Drosophila* cDNA-filters (9216 clones total) were available.

### *RNA-preparation*

RNA was extracted from a set of mouse tissues (C57 black: liver, ovary, testis, heart, and kidney) and from human fetal brain (obtained from the tissue bank at the Royal Marsden Hospital, London). Total RNA was prepared according to Chomczynski and Sacchi (1987), and polyA-RNA was selected by oligo d(T) cellulose chromatography according to standard procedures.

### *Generation of labeled cDNA pools*

First-strand cDNA was synthesized from 5 μg polyA-RNA with oligo d(T) primers (Pharmacia) and M-MLV $H^-$ reverse transcriptase (superscript, BRL) according to the manufacturer's instructions. No radioactive nucleotide was used for the main first-strand synthesis. One μl of the main reaction was supplemented with 1 μCi of radioactive dCPT for a separate pilot reaction. The labeling procedure was continued only if the bulk of first-strand cDNA was larger than 600–700 bp, as assessed on an alkaline gel. The RNA/cDNA hybrid pool was transferred to siliconized tubes and precipitated with sodium acetate and ethanol. To remove the RNA strand, the pellet was resuspended in 100 μl 0.1 M NaOH and heated to 68°C for 20 min. After neutralization with 2 μl 6 M HCl and 10 μl 2 M Tris-HCl pH 7.6, the single-stranded cDNA was precipitated with ethanol. Of the first-strand cDNA 1 μg was labeled with random hexamer primers and Klenow polymerase. High specific activities of up to $5 \times 10^8$ cpm/μg were achieved by use of 100 μCi of radioactive dCTP per reaction. Klenow polymerase was inactivated by two extractions with phenol/chloroform/isoamyl alcohol (50/49/1). The labeled cDNA pools were precipitated with sodium acetate and ethanol and competed with 200 μg of sonicated human placental DNA (Sigma) and 20 μg polyU homopolymer (Pharmacia) for approximately 2.5 h as described elsewhere (Sealy et al. 1985; Litt and White 1985). cDNA pool probes were used immediately.

### *cDNA pool hybridization*

Hybridizations and prehybridizations were carried out in bags. cDNA library filters were prehybridized in 50 ml 6 × SSC (1 × SSC: 150 mM NaCl, 15 mM sodium citrate), 5× Denhardt's solution, 0.5% SDS, 100 μg yeast t-RNA/ml, 50 μg sonicated total human DNA/ml, 10 μg polyA homopolymer/ml (Pharmacia), and 50% formamide at 42°C for at least 16 h. Two to three times $10^8$ cpm radiolabeled cDNA pool probe were added to 80 ml of prewarmed fresh hybridization solution prepared as above. Hybridization was carried out at 42°C for at least 48 h, with no more than three filters hybridized simultaneously in one bag. Filters were washed two times for 5 min at room temperature in 2 × SSC/0.1%SDS, once for 30 min at 68°C in 2 × SSC/0.1%SDS, and once for 30 min at 68°C in 1 × SSC/0.1%SDS. Filters were briefly blotted dry and covered with Saran Wrap (Genetic Research Instrumentation, GRI). X-ray films (Kodak XAR) were exposed for 1–2 days at −70°C by use of intensifying screens.

### *Image analysis*

Filters were scanned on a PhosphorImager (Molecular Dynamics) for quantitative analysis of signal intensities. The scanned 16-bit images were stored in files (approximately 15 Mb per image), transferred to a MikroVAX II computer, and analyzed with a Kontron IPS image analysis system. The analysis software was programmed in "C" using functions from the Kontron Image analysis library (Günther Zehetner, unpublished). After background corrections the positive spots were first identified as objects by use of four convolution filters, and then the integrated optical density of each object was calculated. A 96 × 96 grid was fitted over the binary image, and for each object its position within the grid was determined. This allowed assignment of certain spotting position to each object. Results were stored in a file for further analysis. Data obtained by image analysis were normalized for the different amounts of DNA contained in each individual in situ clone on the library arrays. Hybridizations of the library arrays with vector DNA were analyzed as described above. The resulting optical density (OD) values for each individual clone were divided by the OD value of the clone with the highest signal intensity on the filter. These factors were used to normalize each OD value from other hybridizations to the same filter. Hybridization results were stored in a relational database

(Reference Library Database; Lehrach et al. 1990; G. Zehetner, unpublished) together with the name of the corresponding cDNA clone and its microtiter plate location.

### Control hybridizations

Probes were added at 2–3 × $10^6$ cpm/ml. Filters were usually hybridized overnight and washed as described for cDNA pool hybridizations. Vector DNA, total human placental DNA, total mouse DNA, and polyA-homopolymer were used as controls. PolyA-homopolymer was labeled with T4-polynucleotide kinase. The other control probes were labeled by random hexamer priming. Results were analyzed as described for cDNA pool hybridizations.

### Northern blotting

As determined by $OD_{260}$ measurement and ethidium bromide staining, 2.5 μg, of PolyA mRNA from five mouse tissues (brain, liver, ovary, heart, testis) were size fractionated on 1.2% agarose, 8% formaldehyde denaturing gels and transferred to Hybond N membranes (Amersham). A set of cDNA clones identified with cDNA pool hybridizations were labeled by random hexamer priming, competed with 10 μg/ml polyA-homopolymer as described above, and added to the hybridization solution at 1 × $10^7$ cpm/ml. Prehybridizations and hybridizations were carried out in 0.5 mM sodium phosphate (pH 7.2), 7% SDS, 1 mM EDTA, 1% bovine serum albumin, 100 μg yeast tRNA, and 10 μg polyU-homopolymer/ml (Pharmacia). Filters were washed down to 0.1 × SSC/0.1%SDS at 68°C and exposed for 3 days at −70°C by use of intensifying screens.

### Backhybridization of cDNA clones to the library filters

Inserts of cDNA clones identified by cDNA pool hybridization were individually amplified by a polymerase chain reaction (PCR) with the primers T3 (ATTAACCCTCACTAAAGGGA) and T7 (GTAATACGACTCACTATAGG) for HFB cDNA-clones and NB40L (GAATAAACGCTCAACTTTCCCACC) and NB40R (AGACCGGAATTCGGCGGCCGC) for Drosophila cDNA clones. PCR was carried out in 100 μl 1.5 mM $MgCl_2$, 50 mM KCl, 10 mM Tris-HCl pH 8.3, 0.01% [w/v] gelatin as reaction buffer, and 2.5 units *Taq* polymerase (AmpliTaq, Cetus). A few cells from a single colony of the particular cDNA clone were used as template. 30 cycles of 1 min at 94°C, 1 minute at 50°C, and 2 min at 72°C were performed in a Perkin Elmer thermal cycler. PCR products were fractionated on a 1% agarose gel and classified into three length groups (0.4–0.8 kb, 0.9–1.5 kb, >1.6 kb). Up to 10 PCR products belonging to the same size group were pooled and digested with the enzymes used for cloning the cDNAs (*Not*I/*Eco*RI or *Xho*I/*Eco*RI) in order to separate any vector sequence from the PCR product. Several dilutions of these digested cDNA clone pools were tested in labeling reactions to find the optimal conditions. Usually a 1:5 dilution allowed identification of the clones of interest and kept background hybridization low. cDNA-clone pools were labeled by random hexamer priming. The competition and hybridization conditions were as described for Northern blots. Filters were washed at highest stringency (0.1 × SSC, 0.1%SDS, 68°C) and analyzed as described for whole tissue cDNA pool hybridizations.

### Sequencing

Plasmid DNA was prepared according to standard alkaline lysis protocols. In addition, the plasmid DNA was precipitated once with PEG and extracted twice with phenol and twice with chloroform to improve the quality of the DNA template. Double-stranded DNA was sequenced with a T7 Sequencing Kit (Pharmacia). One single run of sequencing from the 5′ end of each selected individual cDNA clone was carried out with T3 or T7 (depending on the cloning orientation) for HFB cDNA clones NB40L for *Drosophila* cDNA clones as primers. Reactions were run on a 6% PAGE, 8 M urea gel. The running conditions allowed an analysis of the first 100–400 bp. Sequencing data were analyzed manually. Nucleic acid (Genembl) and protein databases (NBRF, Swiss Protein Bank) were initially screened with the UWGCG software package (Devereux et al. 1984), with the program FASTA. Protein bank searches were carried out only for DNA sequences showing no significant match. The program TRANSLATE was used for conversion of DNA sequences in all six possible reading frames. Protein sequences showing no significant matches with the program FASTA were used to screen the protein database OWL110.0 with the program "PROSEARCH" (J.F. Collins and A.F.W. Coulson, University of Edinburgh, UK) on a Sun computer.

## Results

### cDNA pool hybridizations and image analysis

cDNA pools from four mouse tissues (heart, ovary, liver, kidney) and from human fetal brain were hybridized to both sets of filters (HFB and *Drosophila*). Hybridization results were analyzed by conventional autoradiography and with a PhosphorImager (Molecular Dynamics). Autoradiographs were used for gross evaluation of hybridization results and for a preliminary visual comparison of different hybridization data. The number of clones hybridizing with each tissue-cDNA pool was larger than could be handled by these conventional methods. Figure 1a shows an entire cDNA library filter (*Drosophila*) after hybridization with a tissue cDNA pool to provide an overview of the amount of data produced in one single experiment. For more efficient analysis, a PhosphorImager was used

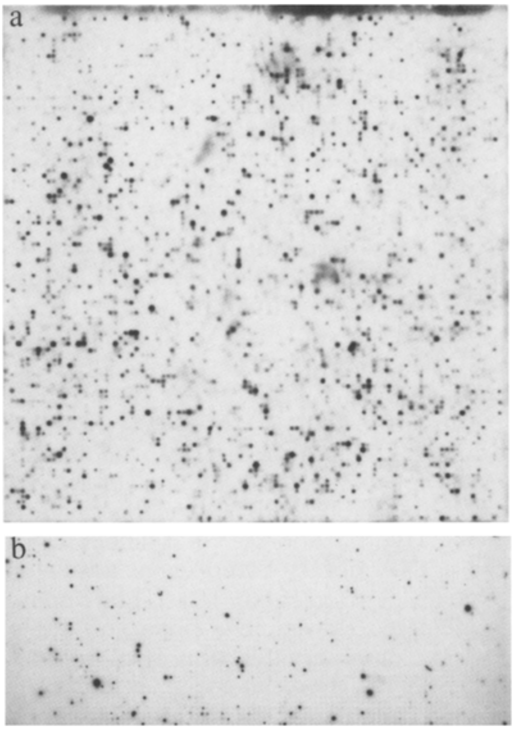

**Fig. 1.** (a) Hybridization of a human fetal brain cDNA pool probe to a *Drosophila melanogaster* cDNA library filter. (b) Poly-A control hybridization. Only the lower third of the complete filter is shown.

for evaluation of hybridization results and to transfer data into a computer database. This system was not only more efficient in analyzing and comparing hybridization data, but also allowed the determination of grey values for each single clone hybridizing with a particular cDNA pool. The storage of grey values for each hybridizing clone is extremely important, considering the complexity of a cDNA pool probe, which consists of approximately 30,000 different sequences with a wide range of different abundances (Sutcliffe 1988). To standardize the analysis system, we adjusted grey value thresholds to allow the identification of all clones classified as positive by the visual analysis of a subset of 1000 clones per library filter and hybridization. For elimination of the influence of colony growth (e.g., different DNA content of individual in situ clones or wells with no growth at all) on the final grey values, the data were vector normalized. To avoid mistakes due to hybridization background (e.g., glove dust may generate background resembling true hybridization signals) and to increase the reliability of the computerized measurement, we applied highly stringent conditions in the analysis and normalization process. Hybridization signals that could not be identified with an original library clone beyond any doubt were omitted from this large-scale analysis. These procedures reduced the total number of clones that could be assessed per filter and experiment (Table 1), but the degree of reliability is of higher priority than the absolute number of clones. Figure 2 shows typical grey value patterns for a small subset of individual HFB cDNA clones as obtained by the image analysis. Each cDNA pool hybridization result was compared with data obtained from the control experiments on the same set of filters. A PolyA-homopolymer probe hybridized to a considerable number of clones identified with cDNA pools (30–40% for HFB and 7–12% for *Drosophila;* Table 1, Figs. 1b and 3e). Competition of the filter with PolyA and preannealing of the probe with PolyU reduced this background but a comparison with the PolyA homopolymer hybridization was still essential. A few clones positive with the PolyA homopolymer were sequenced from the 3′ end and revealed PolyA stretches of more than 200 nucleotides (e.g., clone 5, Fig. 3). Clones hybridizing with cDNA pools and with total human (Fig. 3d) or total mouse DNA (Fig. 3c) were likely to contain repetitive sequences and were not taken into consideration. A considerable fraction (6–10%) of clones on the HFB cDNA library array hybridized with total human DNA. This was not unexpected, as the library was constructed from total PolyA-mRNA (nuclear and cytoplasmic RNA). Hybridization of total human DNA to *Drosophila* cDNA library filters gave a smaller number of signals (3–4%; Table 1B). Total mouse DNA probes detected only a few clones on either Drosophila (10–15 clones per filter) or HFB-filters (40–50 clones per filter). cDNA pools generated from mouse rather than from human tissues were preferentially used to reduce background hybridization signals because of human-specific repetitive sequences. On average, 28% of all HFB library clones and 23% of the cDNA *Drosophila* clones hybridized to at least one cDNA pool (Table 1). A large number of these positives could not be used for further analysis as they hybridized with one of the control probes as well (up to 45% for HFB and up to 16% for *Drosophila*). Clones not hybridizing with the controls contain cDNA sequences from mRNAs expressed at middle to high abundance (>0.1%; see estimates in paragraph *d* below) in the tissue used to prepare the cDNA pool. The results of hybridizations with cDNA pools derived from different tissues were analyzed as described and compared. Table 1 shows the overlaps between different hybridization results (cDNA pools and controls) under high stringency conditions for the image analysis. Approximately 50% of the clones (HFB and *Drosophila*) listed in Table 1 showed hybridization signals with at least two cDNA pools. These clones were selected for the characterization experiments described as follows.

### *Backhybridizations to Northern blots*

cDNA clones hybridizing with more than two different tissue cDNA pools were hybridized to PolyA Northern blots containing a panel of mRNAs from different mouse tissues. Approximately 80% of these clones detected discrete bands between approximately 5.0 and 0.5 kb on the PolyA mRNA blots. Two clones with insert sizes of 2.5 and 2.8 kb hybridized to transcripts of only 160 bp of length, similar to the brain-specific ID sequences described by Sutcliffe and co-workers (1982). In contrast to the aforementioned ID sequence, the 160-bp band was present in all tissues, but showed the strongest expression in brain and liver. Fifteen percent of the selected clones, mostly showing weak signals in the cDNA pool hybridizations, did not hybridize to any of the tissues on the Northern blot panel and were probably not abundant enough for detection. Clone hybridizing with more than two tissue-cDNA pools detected mRNAs in all tissues on the Northern blot, even in those tissues that gave no hybridization signal with cDNA pools. The conclusions to be drawn from the Northern blot experiments are: (1) cDNA clones detected by cDNA pool hybridization hybridize back to Northern blots, thus proving the specificity of the method and (2) Most cDNA clones hybridizing with at least two cDNA pools are abundantly expressed in all tissues on the Northern blot panel. Some clones may be expressed around the limit of detection (approximately 0.1–0.15%) in a few tissues. Only tissue cDNA pools containing the respective sequence above this limit will generate a hybridization signal on the colony filters. Figure 4 shows a few examples of Northern blot hybridizations. A detailed explanation is given in the figure legend.

### *Sequence analysis*

Our conventional double-strand DNA sequencing from the 5′ end usually produced readable sequences between 50 and 200 nucleotides in a single run (for

**Table 1.** This table summarizes the results of the cDNA pool hybridizations with both the HFB (A) and the *Drosophila* (B) cDNA library as obtained by the computerized image analysis. The tables outline the total number of clones hybridizing with one cDNA pool and the overlaps between different cDNA pools. Only vector-normalized data is shown. On average two cDNA pool hybridizations show 50% overlapping clones. (A): Human fetal brain cDNA library, total number of clones analyzed per filter = 5225. cDNA pools used: (1) human fetal brain, (2) mouse ovary, (3) mouse kidney, (4) mouse liver, (5) mouse heart, (6) total human genomic DNA control, and (7) polyA homopolymer control. (B:) *Drosophila* cDNA library; total number of clones analyzed per filter = 5276. cDNA pools used: (1) human fetal brain, (2) mouse ovary, (3) mouse kidney, (4) mouse heart, (5) total human genomic DNA control, and (6) polyA homopolymer control.

A

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|
| 1121 | 544 | 615 | 378 | 672 | 89 | 439 | 1 |
| | 787 | 493 | 263 | 441 | 41 | 311 | 2 |
| | | 1017 | 350 | 500 | 56 | 341 | 3 |
| | | | 741 | 316 | 37 | 214 | 4 |
| | | | | 1017 | 87 | 441 | 5 |
| | | | | | 319 | 95 | 6 |
| | | | | | | 754 | 7 |

B

| 1 | 2 | 3 | 4 | 5 | 6 | |
|---|---|---|---|---|---|---|
| 715 | 463 | 359 | 398 | 28 | 52 | 1 |
| | 722 | 396 | 378 | 29 | 60 | 2 |
| | | 631 | 291 | 25 | 50 | 3 |
| | | | 675 | 29 | 71 | 4 |
| | | | | 55 | 4 | 5 |
| | | | | | 132 | 6 |

average, see legend of Table 2). Sequences longer than 50 nucleotides were found to be most useful for database searches. A clone was considered to be homologous to the identified gene only if a match of more than 80% was detected over more than 50 nucleotides, or if more than 80% identity for at least 10 amino acids of the hypothetical protein was found. Most database matches were to genes with structural and regulatory ("housekeeping") functions in cells. A summary of the sequencing results is shown in Table 2. Only a small number of *Drosophila* cDNA clones was sequenced to demonstrate the applicability of the approach to the *Drosophila* genome. Sequences with no significant database match and more than 50 nucleotides were submitted to the EMBL databank.

### *Backhybridization of characterized clones*

Repeated characterization of abundantly expressed clones is likely to happen with primary (rather than normalized) cDNA libraries. To reduce this time-consuming work, a strategy was devised allowing the identification of all positions of already characterized clones. cDNA inserts were individually PCR amplified, pooled, and hybridized back to the library filters. Only inserts with approximately the same length were pooled, to avoid differences in the specific activity of each sequence within the pool. cDNA clone pool hybridizations to cDNA library filters generated a large amount of background and unspecific hybridization. It was essential to digest the pooled, PCR-amplified cDNA inserts with the cloning enzymes to separate the rest of the vector sequences from the insert before starting the labeling reaction. Without the restriction digestion, even a pool hybridization with three cDNA clones was almost identical with a vector hybridization, which made identification of specific clones impossible. With this method, up to 10 cDNA clones could by hybridized as a pool to the cDNA libraries, allowing a clear identification of all specific clones (Fig. 5a and b). On average, one clone in a pool hybridized to 10–14 clones on each HFB library filter (Fig. 5a and b), thus allowing a gross calculation of the abundance levels of hybridizing clones. The abundance of clones present 10–14 times in 9216 clones (one library filter) was estimated to be 0.1–0.15%.

## Discussion

In projects with the final goal of characterizing large numbers of cDNA clones, it is desirable to develop strategies that ensure a maximal output of information from every experiment and that reduce the redundancies of a random selection of cDNA clones. As we have shown in our work with genomic libraries (Nizetic et al. 1991; Hoheisel et al. 1991a,b; Craig et al. 1990), such large-scale projects can most easily be performed with library arrays spotted at high clone density with a robotic device and a variety of hybridization fingerprinting techniques. In the present study hybridization fingerprinting of high-density gridded cDNA libraries with non-normalized tissue cDNA pool probes permitted us to obtain data on roughly one-third of the cDNA library clones with only a few hybridizations. Clones hybridizing with cDNA pools contain cDNA sequences from mRNAs expressed at middle to high abundance (>0.1–0.15%) in the tissue used to generate the probe. On average, cDNA clones

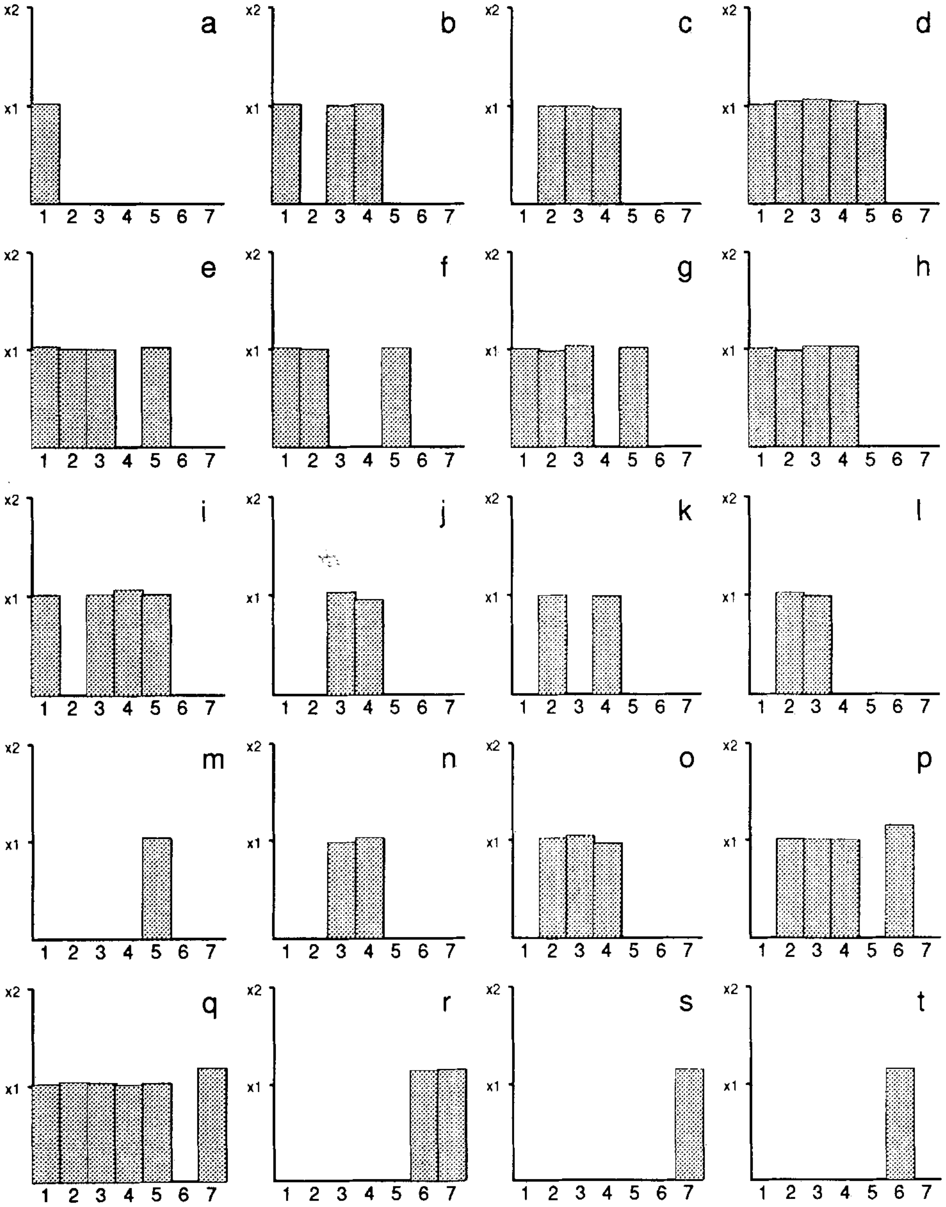

**Fig. 2.** Typical grey value patterns as produced by different cDNA pool and control hybridizations for a subset of HFB library clones are shown. To make results obtained in different hybridizations for individual clones comparable, grey values are expressed as a multiple of the mean grey value (up to ×2 as indicated on the y-axis) of all hybridization signals obtained in the particular hybridization. Each individual graphic shows the results for one single clone; the grey values obtained by different hybridizations are shown as **bars** (cDNA pools: 1 = human fetal brain, 2 = mouse ovary, 3 = mouse kidney, 4 = mouse liver, 5 = mouse heart; controls: 6 = total human DNA, 7 = polyA homopolymer). a, b, c, d, g, h, i, j, k, n, and o show clones containing sequences with no significant database match. Clones p, q, r, s, and t hybridize with one of the controls or with both cDNA pools and controls. Homologies to known genes could be detected for the remaining clones: e = cytochrome oxidase III; f = ATPase6; l = mitochondrial mRNA 13; m = sorcin.

selected in this approach hybridized back to 10–14 clones on the same library filter. Assuming that up to 3000 clones per filter may hybridize with one single-tissue cDNA pool, a maximum of 300 cDNAs from different mRNA species expressed at middle to high abundance could be identified. Our estimates are in agreement with data from mRNA-complexity studies performed by Hastie and Bishop (1976). These authors estimated that a mouse liver cell with a total of 500,000 mRNAs contains 12,000 sequences present at 15 copies (low abundance), 350 sequences present at 300 copies (middle abundance), and 10 sequences present at 12,000 copies (highly abundant clones). Clones not identified with this fingerprinting technique potentially represent clones containing rare mRNA sequences, which have been described to contain a large fraction of tissue-specific genes (Sutcliffe 1988), or clones containing no insert or noncoding inserts. To distinguish these two groups and to identify clones erroneously identified as positive in a tissue cDNA pool experiment owing to unspecific hybridization, we included a series of control experiments in the strategy. In particular, a control hybridization with a PolyA-homopolymer proved essential, as up to 45% of identified clones may appear positive by hybridization of labeled PolyA tails to large stretches of PolyT in the

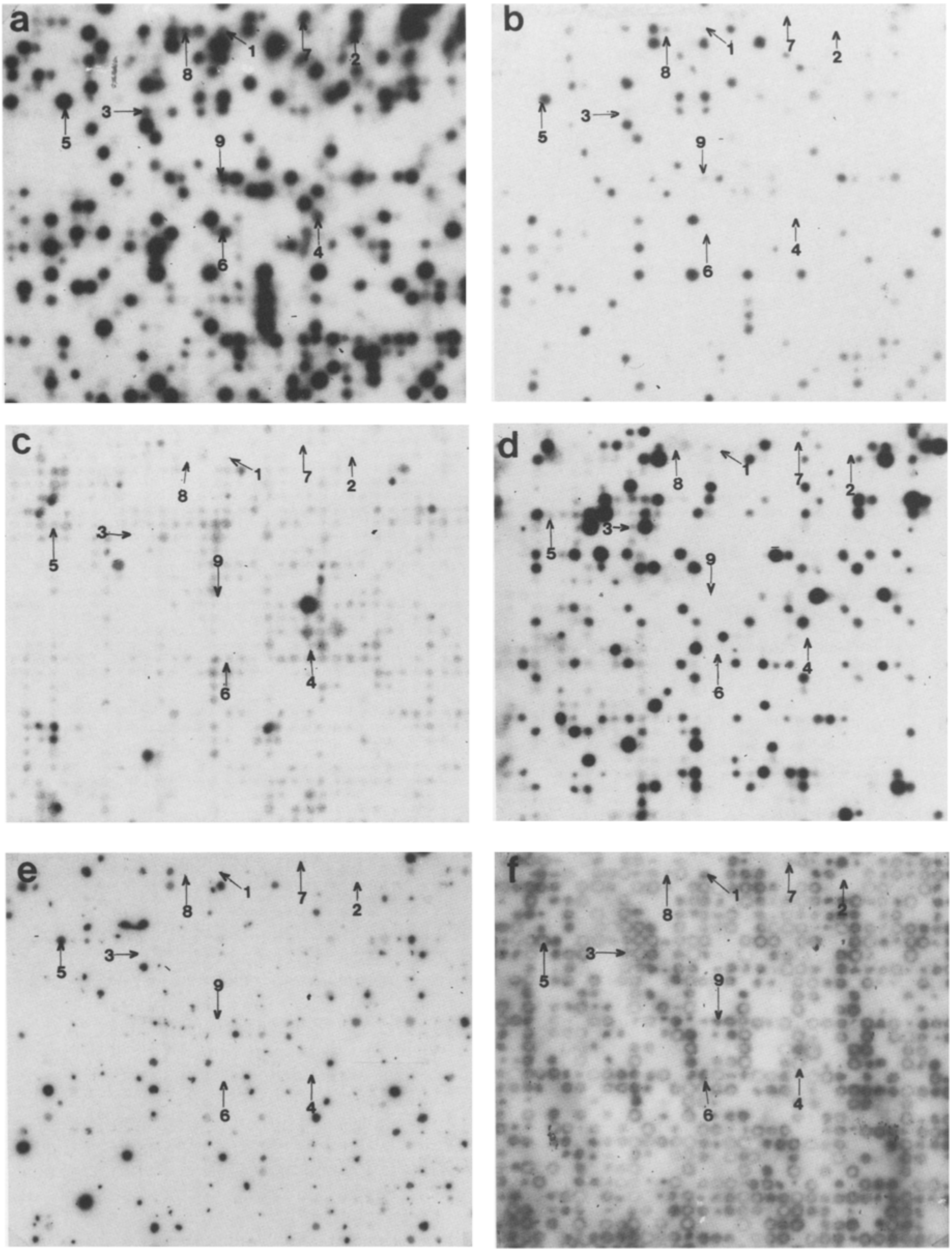

**Fig. 3.** A section derived from the human fetal brain cDNA library (coordinates: y55–y85/1–x32) was hybridized with a set of different probes. **a,** cDNA pool human fetal brain; **b,** cDNA pool adult mouse liver; **c,** total mouse DNA control hybridization; **d,** total human DNA control hybridization; **e,** polyA homopolymer control hybridization; **f,** vector DNA control hybridization. The **arrows** point on a set of clones that have been used for further analysis on Northern blots, and by sequencing. 1, clone b51/ubiquitin; 2, clone b37/β-tubulin; 3, clone b43/calcium-binding protein sorcin; 4, clone b41/16s ribosomal RNA; 5, clone b47/false positive owing to a poly-A tail of more than 250 nucleotides; 6, clone b45/5′ new sequence; 7, clone b53/5′ new sequence; 8, clone b59/5′ new sequence; 9, clone b44/5′ new sequence. Northern blot hybridizations for clones 1, 2, and 6 are shown in Fig. 4. Analyzed clones show no or only weak background hybridization signals in controls c, d, and e. Clones 1, 3, 5, 6, 7, 8, and 9 show signals of different intensity in a and b, due to the different levels of expression in the respective tissue. 2 and 4 are not present in b. The vector hybridization of the same filter (f) shows that clone 2 did not grow on this filter; clone 4 may be expressed below the level of detection of the method. Clone 5 appears as one of the strongest positives in a and b, and was identified as false positive in control e; sequencing of the 3′ end revealed a polyA stretch of more than 250 nucleotides.

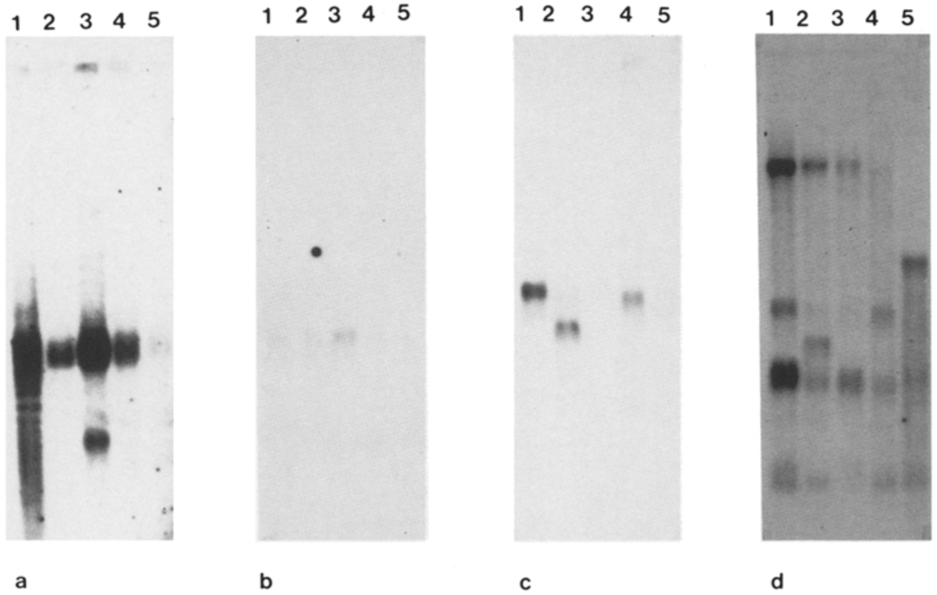

**Fig. 4.** Hybridization of a set of cDNA clones identified with cDNA pool hybridizations to Northern blots containing 2.5 μg polyA RNA per lane of the following adult mouse tissues: 1, brain; 2, heart; 3, testis; 4, ovary; 5, liver. The **lines on the right side** of the autoradiographs represent the size markers (from the top: 9.5 kb, 7.5 kb; 28 s rRNA = 4.7 kb, 4.4 kb, 2.4 kb; 16s rRNA = 1.8 kb, 1.4 kb). **a:** clone b37 identified as β-tubulin by sequencing showed the typical expression patterns of the MF3 β-tubulin isoform (Wang et al. 1986). Background resembling weak bands on this autoradiography is due to the conditions of exposure; the RNA in lane 1 shows slight degradation. **b:** clone b54 with no significant sequence match showed weak bands in all displayed tissues. **c:** clone d17 (*Drosophila* cDNA clone) showed an expression pattern typical for actin (Cleveland et al. 1980). The clone was later identified as cytoplasmic actin by 5′ sequencing. **d:** clone b51 hybridizes to three bands in all five mouse tissues. The smaller bands of 0.6 kb and 1.1 kb are identical for all five mouse tissues. The third band was 2.5 kb in mouse liver and 4.3 kb in the remaining tissues. The weak bands visible between the 16s and 1.4-kb marker bands represent remnants from a previous hybridization with actin (filters were not stripped to avoid loss of RNA). Sequence analysis showed a highly significant match with the ubiquitin multigene family (Wiborg et al. 1985). The 0.6 kb (UBI A), 1.1 kb (UBI B), and 2.5 kb (UBI C) bands equal the expression pattern described for ubiquitin by Wiborg and co-workers (1985). The most likely explanations for the larger 4.3-kb band in four mouse tissues are that it might represent either unspliced, immature mRNA, a new gene, or a new isoform of ubiquitin. We displayed Northern blot hybridizations from a set of clones with well-known sequences to demonstrate the conformity of our sequence data with the obtained expression patterns.

cloned cDNA. cDNA pool probes were preferentially prepared from a different species (e.g., mouse) than the cDNA library (human) to reduce unspecific hybridization due to human-specific repetitive sequences (e.g., Alu). Control hybridization with total genomic DNA allowed the detection of additional clones positive owing to hybridization with other labeled repetitive sequences in the pool probe. An efficient elimination of all "false positives" is possible only with the strategy presented here. Other groups have encountered the same problems and have tried to reduce them by different competition techniques (Dworkin and Dawid 1980; Crampton et al. 1980). Our experiments demonstrate that even a maximal competition with PolyA/PolyU-homopolymers and genomic DNA is not sufficient to eliminate these background problems completely. Control hybridization with a PolyA homopolymer has been used before (Sargent and Dawid 1983; Dworkin and Dawid 1980), but as library screens were mainly performed with the technique originally devised by Grunstein and Hogness (1975) and no computerized image analysis was available, the efficiency of these approaches was limited. As an alternative approach to reducing this background, we are in the process of testing oligo-dT primers for the synthesis of the first-strand DNA anchored at the boundary between PolyA tail and 3′ untranslated region, by adding an A, G, or C to the 3′ end of the primer.

In addition to allowing the screening of thousands of clones at a time, the use of cDNA library arrays spotted at a high clone density offers the possibility for an automated analysis of hybridization fingerprinting data. The automated data analysis we employed was based on a computerized image analysis system and a relational database. A computerized image analysis system is essential for the efficient handling of this large amount of hybridization fingerprinting data. A relational database is necessary to establish correlations between results obtained in different fingerprinting and control experiments. The structure of the database used in our laboratory has recently been summarized (Lehrach et al. 1990). Much time and effort was invested in this part of the approach; in particular, the standardization and optimization of the image analysis system proved demanding. A large scale of grey values is generated in one single-tissue cDNA pool hybridization, and the determination of adequate grey value thresholds allowing one to distinguish between "positives and negatives" in each individual experiment is not a simple matter. At present, the most reliable method to standardize and validate the system is to use highly stringent image analysis conditions adjusted to the results of visual analysis of a subset of clones in each hybridization. Further optimization and standardization of this system will be essential for a large-scale application. Information produced by hy-

Table 2. Partial sequencing data for HFB and embryonal *Drosophila* cDNA clones. Matches of at least 80% over more than 50 nucleotides were considered significant. The average sequence length was 115 ± 50 for *Drosophila* clones and 125 ± 60 for HFB clones. For peptide database screens the sequence was considered homologous to the known peptide only if an identity of more than 80% for at least ten amino acids was found. Clones hybridizing with more than two tissue cDNA pools were selected for sequencing. Sequences longer than 50 nucleotides, showing no significant database match, were submitted to the EMBL database and were assigned the following accession numbers X65374–X65393 (HFB) and X65268–X65275 (*Drosophila*).

| Human fetal brain | | *Drosophila* | |
|---|---|---|---|
| unknown | 29 | unknown | 9 |
| sorcin | 1 | cytoplasmatic actin | 1 |
| ubiquitin | 1 | E75 protein | 1 |
| cytochrome oxid. III | 1 | vector | 2 |
| cytochrome oxid. II | 1 | Total | 13 |
| Asparagin t-RNA | 1 | | |
| hnRNP core protein | 1 | | |
| ATPase 6 | 1 | | |
| alpha 1 globin | 1 | | |
| β tubulin | 1 | | |
| ADP ribosylation factor 3 | 1 | | |
| ribosomal protein L37a (rat) | 1 | | |
| ribosomal protein S19 (rat) | 1 | | |
| ribosomal protein L30 (rat) | 1 | | |
| ribosomal protein L19 (rat) | 1 | | |
| 16s rRNA | 1 | | |
| polyA tail | 4 | | |
| vector | 3 | | |
| Total | 51 | | |

bridization fingerprinting will prove valuable for several purposes (Lennon and Lehrach 1991). cDNA pool hybridizations have been used to screen cDNA libraries for sequences differentially expressed in tumorous versus nontumorous tissue (Shiosaka et al. 1982, 1987; Augenlicht et al. 1987; Jacobs and Birnie 1980). With the strategy presented in this paper, the information extracted from such experiments could be maximized. We have constructed a cDNA library from a pancreatic tumor cell line for this purpose, high-density in situ library filters have been produced, and preliminary fingerprinting experiments are in progress.

The approach presented here will be of special value in selecting clones for the generation of expressed sequence tagged sites (ESTs) for mapping and sequencing the human genome. A random selection of cDNA clones from libraries may lead to a high proportion of redundancies by the repeated selection of the same highly abundant clones, or of clones containing no inserts or noncoding inserts (Adams et al. 1991). The use of cDNA pool hybridizations in combination with conventional cDNA library filter lifts, as described by Höög (1991), represents an improved strategy for selecting cDNA clones. Nevertheless, serious disadvantages exist, including the difficulty in comparing results with previous experiments (controls and hybridization fingerprints) and the inability to simultaneously analyze more than a small number of clones plated at low density. Although we selected only mRNAs expressed at middle to high abundance, the time and effort necessary to generate ESTs from all 300 sequences of this abundance class would be significantly lower and could be completed in the foreseeable future. To prove the feasibility of this approach, we generated a number of ESTs for a well-defined and compact group of mRNA sequences. More than 50% of the total number of clones detected in tissue cDNA pool hybridizations reacted with probes generated from at least two different tissues. Sequences abundantly expressed in several tissues are likely to code for regulatory and structural proteins essential for the functioning of every cell. A number of clones from this group were selected for a single run of

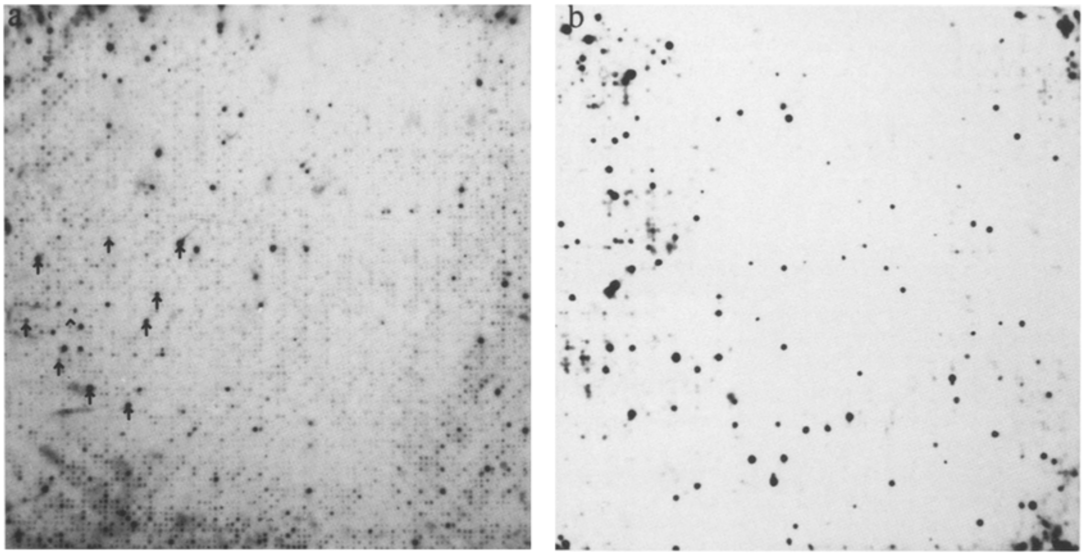

Fig. 5. Backhybridization of a pool of ten PCR-amplified cDNA inserts to both human fetal brain cDNA filters (**a**: set 1, **b**: set 2). *Arrows* point on the clones used for the generation of the PCR-pool probe. The pool probe identified nine of the original ten clones used. Comparison with the vector hybridization showed that the only original clone (**arrowhead**) not identified by the backhybridization did not grow on this particular filter. A total of 287 clones was detected on both filters.

sequencing of double-stranded DNA from the 5′ end of the inserts. We sequenced the 5′ rather than the 3′ ends, because the coding sequence is more likely to be at this end of the clone (e.g., Sutcliffe 1988). As expected, a proportion of these sequences showed significant homologies to genes with constitutive functions. No significant homology could be found for more than 50% of the sequences (Table 2). These sequences may represent as yet unknown genes.

Data produced by the cDNA pool hybridizations will form part of our integrated approach for the mapping of genomes and will supplement data obtained by hybridization with complex and simple probes, and by oligonucleotide hybridization partial sequencing (Lennon and Lehrach 1991). Each technique will generate a set of information for each individual clone, thus characterizing it, not only with a short sequence tag, but also by information concerning abundance levels and additional hybridization fingerprinting data.

We conclude that the approach presented in this paper will be helpful for the characterization of a large number of cDNA clones, especially for the mapping projects based on the generation of ESTs. Three new basic elements used in the screening of cDNA libraries are responsible for the high efficiency of this approach: (1) the use of high-density library arrays which allow thousands of clones to be screened simultaneously; (2) the use of hybridization fingerprinting techniques to identify abundantly expressed clones and to eliminate spurious clones (no insert or noncoding inserts); and (3) the use of a computerized system for the evaluation of hybridization data.

*Acknowledgments.* We wish to express our gratitude to Linda MacCarthy for performing the image analysis of the hybridization results; Hugh Griffiths for the computer analysis of the image analysis data; Stefan Beck and Paul Freemont for valuable advice concerning the database searches; Roger Cox and Mathias Hoglund for helpful discussions; Catherine Knights and Tobias Hatter for excellent technical assistance; Janet Roden, Lisa Gellen, Christal Douglas, and Anita Beitverda for producing library filters; and Daniel Nelson for a critical review of this manuscript. T.M.Gress was supported by a fellowship of the Deutsche Forschungsgemeinschaft.

## References

Adams, M., Kelly, J., Gocayne, J., Dubnick, M., Polymeroppoulos, M., Xiao, H., Merril, C., Wu, A., Olde, B., Moreno, R., Kerlavage, A., McCombie, R., and Venter, J.C.: Complementary DNA sequencing: expressed sequence tags and Human Genome Project. *Science 252:* 1651–1656, 1991.

Adams, M., Dubnick, M., Kerlavage, A., Moreno, R., Kelley, J., Utterback, T., Nagle, J., Fields, C., and Venter, C.: Sequence identification of 2375 human brain genes. *Nature 355:* 632–634, 1992.

Augenlicht, L., Wahrman, M., Halsey, H., Anderson, L., Taylor, J., and Lipkin, M.: Expression of cloned sequences in biopsies of human colonic tissue and in colonic carcinoma cells induced to differentiate in vitro. *Cancer Res 47:* 6017–6021, 1987.

Bantle, J., and Hahn, W.: Complexity and characterization of polyadenylated RNA in the mouse brain. *Cell 8:* 139–150, 1976.

Brown, N. and Kafatos, F.: Functional cDNA libraries from Drosophila embryos. *J Mol Biol 203:* 425–437, 1988.

Chomczynski, P. and Sacchi, N.: Single step method of RNA isolation by acid guanidinium-thiocynate-phenol-chloroform extraction. *Anal Biochem 162:* 156–159, 1987.

Cleveland, D., Lopata, M., MacDonald, R., Cowan, N., Rutter, W., and Kirschner, M.: Number and evolutionary conservation of α- and β-tubulin and cytoplasmic β- and γ-actin. Genes using specific cloned cDNA probes. *Cell 20:* 95–105, 1980.

Craig, A., Nizetic, D., Hoheisel, J., Zehetner, G., and Lehrach, H.: Ordering of cosmid clones covering the herpes simplex virus type I (HSV-I) genome: a test case for fingerprinting by hybridization. *Nucleic Acids Res 18:* 2653–2660, 1990.

Crampton, J., Humphries, S., Woods, D., and Williamson, R.: The isolation of clones cDNA sequences which are differentially expressed in human lymphocytes and fibroblasts. *Nucleic Acids Res 8:* 6007–6017, 1980.

Derman, E., Krauter, K., Walling, L., Weinberger, C., Ray, M., and Darnell, J., Jr.: Transcriptional control in the production of liver-specific mRNAs. *Cell 23:* 731–739, 1981.

Devereux, J., Haeberli, P., and Smithies, O.: A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res 12:* 387–395, 1984.

Dworkin, M and Dawid, I.: Construction of a cloned library of expressed embryonic gene sequences from *Xenopus laevis*. *Dev Biol 76:* 435–448, 1980.

Grunstein, M. and Hogness, D.: Colony hybridization: a method for the isolation of clones DNAs that contain a specific gene. *Proc Natl Acad Sci USA 72:* 3961–3965, 1975.

Hastie, N. and Bishop, J.: The expression of three abundance classes of messenger RNA in mouse tissues. *Cell 9:* 761–774, 1976.

Hochgeschwender, J., Sutcliffe, G., and Brennan, M.: Construction and screening of a genomic library specific for mouse chromosome 16. *Proc Natl Acad Sci USA 86:* 8482–8486, 1989.

Hoheisel, J.D., Lennon, G.G., Zehetner, G., and Lehrach, H.: Use of high coverage reference libraries of *Drosophila melanogaster* for relational data analysis. A step towards mapping and sequencing of the genome. *J Mol Biol 220:* 903–914, 1991a.

Hoheisel, J.D., Drmanac, R., Larin, Z., Lennon, G.G., Monaco, A., Zehetner, G., and Lehrach, H.: Use of high coverage libraries for an integrated analysis of genomic DNA. Adv Mol Genet 4: 125–132, 1991b.

Höög, C.: Isolation of a large number of novel mammalian genes by a differential cDNA library screening strategy. *Nucleic Acids Res 19:* 6123–6127, 1991.

Jacobs, H. and Birnie, G.D.: Post-transcriptional regulation of messenger abundance in rat liver and hepatoma. *Nucleic Acids Res 8:* 3087–3120, 1980.

Lehrach, H., Drmanac, R., Hoheisel, J., Larin, Z., Lennon, G., Monaco, A., Nizetic, D., Zehetner, G., and Poustka, A.: Hybridization fingerprinting in genome mapping and sequencing. *In* K. Davies and S.M. Tilghman (eds.); *Genome Analysis, Vol. I: Genetic and Physical Mapping,* pp. 39–81, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 1990.

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

Litt, M. and White, R.: A highly polymorphic locus in human DNA revealed by cosmid-derived probes. *Proc Natl Acad Sci USA 82:* 6206–6210, 1985.

Nizetic, D., Zehetner, G., Monaco, A., Gellen, L., Young, B., and Lehrach, H.: Construction, arraying and high-density screening of large insert libraries of human chromosomes X and 21: their potential use as reference libraries. *Proc Natl Acad Sci USA 88:* 3233–3237, 1991.

Sambrook, J., Fritsch, E.F., and Maniatis, T.: *Molecular Cloning, A Laboratory Manual,* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1989.

Sargent, T. and Dawid, I.: Differential gene expression in the gastrula of *Xenopus laevis*. *Science 222:* 135–139, 1983.

Sealy, P., Whittaker, P., and Southern, E.: Removal of repeated sequences from hybridisation probes. *Nucleic Acids Res 13:* 1905–1922, 1985.

Shiosaka, T. and Saunders, G.: Differential expression of selected genes in human leukemia leukocytes. *Proc Natl Acad Sci USA 79:* 4668–4671, 1982.

Shiosaka, T., Tanaka, Y., and Kobayashi, Y.: Preferentially expressed genes in stomach adenocarcinoma cells. *Br J Cancer 56:* 539–544, 1987.

Sutcliffe, J.G.: mRNA in the mammalian central nervous system. *Annu Rev Neurosci 11:* 157–198, 1988.

Sutcliffe, J.G., Milner, R.J., Bloom, F.J., and Lerner, R.: Common 82-nucleotide sequence unique to brain RNA. *Proc Natl Acad Sci USA 79:* 4942–4946, 1982.

Wang, D., Villasante, L.S., and Cowan, N.: The mammalian β-tubulin repertoire: hematopoietic expression of a novel heterologous β-tubulin isotype. *J Cell Biol 103:* 1903–1910, 1986.

Wiborg, O., Pedersen, M., Wind, A., Berglund, L., Marcker, K., and Vuust, J.: The human ubiquitin multigene family: some genes contain multiple directly repeated ubiquitin coding sequences. *EMBO J 4:* 755–759, 1985.