Requirements in screening cDNA libraries for new genes and solutions offered by SBH technology

R. Drmanac, S. Drmanac, I. Labat, and Nick Stavropoulos

Integral Genetics Group

Center for Mechanistic Biology and Biotechnology

Argonne National Laboratory

9700 South Cass Avenue

Argonne, Il 60439-4833

Requirements in screening cDNA libraries for new genes and solutions offered by SBH technology, R. Drmanac, S. Drmanac, I. Labat, and Nick Stavropoulos, Integral Genetics Group, Center for Mechanistic Biology and Biotechnology, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, Il 60439-4833

ABSTRACT

Under different assumptions about the total number of genes, the number of housekeeping and tissue-specific genes, and the difference in the number of mRNAs per cell for functional and nonfunctional genes, significantly different results can be expected from screening random cDNA clones. We have developed gene expression models as a guide for interpretation of experimental results. For statistical, biological, and technical reasons, the search for 100,000 plus genes and discrimination between nonfunctional, housekeeping, and tissue-specific genes requires the analysis of up to 10 million clones from 20 to 50 tissues. Oligonucleotide hybridization of dense clone blots is an inexpensive and fast way to screen such large clone sets. Our preliminary results on control clones and thousands of cDNA clones from an infant brain library demonstrate the feasibility of the method.

INTRODUCTION

The number of mammalian genes is usually estimated to be 50,000 to 100,000 (or sometimes as high as 150,000) (1). Studies of gene expression have found that the number of mRNAs transcribed from a gene can vary a few thousand times and that there is an exponential increase of the number of genes having a small number of mRNA copies (2-4). Also, the existence of 200–300 cell types each expressing several hundred specific (usually highly expressed) genes is estimated. There is no clear-cut definition of what has to be considered as an functional gene or whether there is a transcription leakage of nonfunctional genes. These factors can seriously influence the identification of new or tissue-specific genes via the screening of cDNA libraries.

Partial sequencing by hybridization (SBH) of random cDNA clones as a way to catalog and define tissue specificity of genes was proposed four years ago (5). Instead of tens of thousands of probes necessary for complete sequencing if a single genome is analyzed (6) or 3000 to 4000 if the data from similar genomes are integrated (7,8), 100-1000 probes are sufficient for partial SBH. In the last three years, successes of single pass gel sequencing of cDNAs have strongly demonstrated the usefulness of incomplete sequences (9-12). Partial SBH is, in principle, a less expensive approach with the ability to analyze millions of clones. SBH has been proven in a blind experiment (13) and recently we have developed a hybridization data production line to score up to 32 probes on 30,000 dots per day (14,15). Similar facilities are under development by Hans Lehrach's group (16).

In this paper, we present several models of gene expression and analyze the main factors which can influence the hunt for new genes via the screening of random cDNA libraries. The basic steps in the preparation and use of dense DNA dot arrays are described, and some results that demonstrate the feasibility and efficiency of gene inventorying by oligonucleotide hybridization are presented. Furthermore, partial SBH and single-pass gel sequencing are compared and a gene analysis scheme that combines the two approaches is discussed.

RESULTS AND DISCUSSION

1. Quantitative Models of Gene Expression: Implications for Screening cDNA Libraries

There are several possibilities for the distribution of genes in terms of the number of their mRNAs per cell in a homogenous tissue. We defined four models based on six gene expression levels (Table 1). Six expression levels representing averages for related parts of the distribution are a crude approximation of reality. The level with five messengers per cell represents genes with 1 to 10 messengers per cell, the 30 messengers level comprises genes having 11 to 90 mRNAs per cell and so on. Levels below 0.1 mRNA per cell have no significant influence on the screening. The models (except model 4) do not differ in the two highest expression levels. Furthermore, since there is a general agreement that a gene is not functional if there is less than one messenger per cell, the genes belonging to the levels of 0.1 and 0.5 mRNA per cell will be considered as transcription leakage. This is not true for complex tissues consisting of several cell types. If 10% of cells represent one type some functional genes may have only 0.1 mRNA per cell of whole tissue. The basic differences among the models are the level of expression (leakage) for genes that are not functional in the given cell type and partition of housekeeping and tissue-specific genes between low (five mRNAs on average) and moderate (30 mRNAs on average) levels of expression.

Model 1 represents the case with a gap between the expression level of the functional and the nonfunctional (leaky) genes. In models 2 through 4 the gap is progressively eliminated. Consequently, less than 1% (in model 1) or up to 15.2% (in model 4) of cDNA clones represents genes without function in the given cells, assuming a total of 112,000 genes, of which 12,000 are housekeeping and cell type specific genes. If the number of genes is 150,000 and if model 4 turns out to be correct, then up to 20% clones may represent genes non-specific for the given cells.

Leakage prevents the discrimination of active and inactive genes, but it offers an opportunity to define the catalog of all genes without analysis of all tissues. If 100,000 clones are screened per tissue, then 79-86% of functional genes will be represented with at least one clone. (For 200,000 clones, the range is 91-97%. The percentages are calculated by summing number of genes in 10^5 or 2×10^5 clones for the four highest expression levels and dividing by 12,000). By analyzing 10 cell types (a total of one million clones), most of the housekeeping genes will be recognized by occurrence in two or more cell types. A fraction of cell type specific genes having a few mRNAs per cell (or, due to statistical reasons), will be represented with one clone and can not be discriminated from non-functional genes. Furthermore, 14% in model 1 ($1 - [120 \times 0.6 + 180 \times 0.99 + 50]/350$) and 16% in model 3 ($1 - [240 \times 0.76 + 60 + 50]/350$) of tissue-specific genes will not be found in 100,000 clones. On the other hand, 30, 54 and 77% of the genes which do not function in any of the 10 analyzed tissues can be found in models 2, 3, and 4, respectively (the calculation is done by the formula described in the footnote to Table 1 where g = 100,000 and c is 10 times the total number of clones in the library of 100,000 clones, which are expected to represent genes from the 0.1 and 0.5 mRNA-per-cell levels).

If 10,000 clones are screened per cell type, only 27% in model 4 and 38% in model 1 of housekeeping and cell type specific genes will be found in one library. Most of the genes from the 5- and 30-mRNA levels will be represented by a single clone and can not be discriminated from inactive genes. Furthermore, the number of clones representing one gene will be a very

inaccurate measure of its expression level due to statistical factors. Cell type specific and housekeeping genes can be potentially distinguished if 30-40 cell types are analyzed. Thus, by screening a small number of clones per cell type, the majority of cell type specific genes will not be found or recognized, and the expression level can not be determined accurately for most of the genes. To be able to obtain these two types of data, 200,000 or more clones have to be analyzed per cell type if any one of the models is correct and if ordinary libraries are used.

The only differences between cDNA libraries of a cell type and a complex tissue is the reduction of clone frequency for genes specific for one cell type in the tissue. In this case, tissue specific genes having a few mRNAs per cell of one cell type may not be possible to discriminate from the leaky genes. The models allow calculation of the expected clone redundancy and the expected number of genes, which will be represented by a certain number of clones if a given number of clones is analyzed per tissue of known complexity. We are planning to develop a program to test the influence of particular variables and to assess the agreement between experimental and expected results.

2. Biological, Technical and Statistical Reasons Impose Screening Several Million cDNA Clones

The data collected by screening random cDNA libraries allow, in principle, the identification of tissue-specific genes and an estimate of their expression levels. The accuracy of the findings depends on various biological, statistical, and technical factors that influence the preparation and screening of cDNA libraries. The impact of statistical factors and transcription leakage can be anticipated by the described models. A few other biological facts can be taken in consideration. Cryptic promoters or transcription termination sites probably exist. Furthermore, incomplete splicing (small introns remain in some mRNAs) or trans-splicing (17–19) can occur. More than 2 Gb of non-coding sequences and hundreds of thousands of primary transcripts per nucleus can produce enormous numbers of "new genes" by very rare transcription or splicing errors. It is not impossible that 20% of the mRNAs representing thousands of genes or "gene like" sequences can be present in a cell without function in that cell or in any other cell. The cells will waste much less energy for this level of error than for the transcription of intron sequences and the number of proteins for any of these unnecessary mRNAs will be below the level which can influence cell functions.

Technical problems in library preparation (contamination with genomic DNA or external mRNAs or DNAs, chimeric clones, false primed clones) and in library screening (deletions or recombinations during clone amplification or PCR, cross-talk of the wells, external contamination, sequencing or hybridization errors) will add further uncertainty in the meaning of the data. By summing the expected levels of all these types of error, we estimate that up to 30% of the cDNA clones in a library can be artifacts or can represent genes nonfunctional in the given tissue. The screening of large sets of clones from various tissues is one possibility that will discriminate between artifacts and real genes or between functional and leaky genes. It may be necessary to consider only clones found at least twice. This requirement will increase the number of clones to be screened several-fold.

Can normalized libraries help? The two highest expression levels represent more than 50% of the clones, and in normalized libraries these genes will be represented by a significantly smaller fraction of clones. The average clone redundancy can be reduced up to twofold. Our preliminary screening of 10,000 clones showed a reduction in the number of redundant clones from 40% in the ordinary library to 20% in the normalized infant brain library (12) constructed

by Bento Soares (Columbia University). On the other hand, very rare transcripts (comprising 2–10% of the mRNAs, which may represent transcription leakage) can be found in the normalized libraries by screening 5- to 10-times-smaller number of clones.

The basic disadvantage of normalized libraries is the loss of information about the expression level of genes. This information can not be determined by genome sequencing. Because of this loss, and a relatively small saving in the number of clones, fully representative cDNA libraries can be more important than normalized libraries. Standard methods of preparation of cDNA libraries can introduce a bias for very short and very long mRNAs because of a narrow cDNA size selection or a reduced transformation efficiency, respectively.

Even in a perfectly normalized library, 10 gene equivalents have to be screened to find the last percentage of genes. To find genes having low levels of expression, and especially to find them twice, as many as 100 gene equivalents (10 million instead of one million clones) may be necessary.

3. Massive Clone Screening by Oligonucleotide Hybridization

To implement DNA screening, mapping, and sequencing by oligonucleotide hybridization, we have developed facilities with a present capacity to score 8–32 probes (6–12 bases in length) per day on 30,000–120,000 DNA fragments spotted on nylon membranes (14,15,20). Procedures for the high throughput clone arraying, amplification, and spotting have been developed (S. Drmanac and R. Drmanac, in preparation). The procedures involve arraying genomic or cDNA clones (prepared in M13 or plasmid vectors) in multi-well plates (96, 384, or 864 wells) by picking plaques or colonies, or by dispensing in the wells an optimally diluted transformation mixture. Replica plates of the master plates are prepared by transferring 2 μl of the cultures in the wells filled with 100 μl of water. An array of metal pins is used to transfer liquid from all wells simultaneously (instead of row by row using a multichannel pipet).

In the next step, cloned inserts are amplified by PCR directly from the bacterial cultures without DNA isolation. PCR mixture is dispensed in the wells and 1 μl of each of diluted cultures is transferred by pin array into the corresponding wells. BioOvens (BioTherm, Fairfax, Virginia) are used for cycling six plates in parallel (14). Many parameters are optimized to be able to routinely produce 20 ng of 2-kb clones per 1 μl of PCR reaction in 90% of the cases (Fig.1).

Amplified DNA is used to prepare high-density dot blots. We have defined conditions to prepare well-defined dots using an array of metal pins 0.3 mm in diameter. Interestingly, DNA can be spotted without removing the oil usually used to prevent the evaporation of PCR reaction mixtures. Membranes (15 × 23 cm, four 96-well plates) with 31,104 dots (4 × [(9×9×96)]) are routinely prepared using a Biomek1000 XYZ table (Beckman, Fullerton, CA). Several hardware modifications of the station have been made and specific software has been developed (I. Labat *et al.*, unpublished results). Up to 100 replica membranes can be prepared from 15-μl PCR reactions.

Membranes are hybridized with $(N)_{0-2}(B)_{6-10}(N)_{0-2}$ probes (N, degenerated positions; B, specific base positions) at 12 °C using 4 pM to 5 nM probe concentrations, and are washed for 20 min to 1h at 2-20 °C, depending on the length and GC content of the probes (15). Each filter is hybridized in a separate box with outlets for pumping out washing buffer. A setup which has four boxes fixed to a cooling plate and mounted on a shaker is presently in use. Filters can be reused over 50 times. The intensities of the hybridization signals are determined by our image

analysis program (DOTS) (20) from the files of pixel values generated by a PhosphorImager (Molecular Dynamics, Sunnyvale, CA). One dot span is covered by 30 pixels. Examples of very different patterns obtained with two probes scored on one array of cDNA clones are shown on Figure 2. The zinc finger consensus probe shown in this example demonstrated the possibility of defining a subset of genes that may encode specific protein motifs.

4. Recognition of Highly Similar cDNA Clones by Comparing Hybridization Signatures

Recently we developed user friendly software (SCORES) based on X windows (N. Stavropoulos and R. Drmanac, unpublished results) for data evaluation and normalization and for comparison of the generated oligonucleotide sequence signatures (OSS) (21). Instead of a less precise 0/1 scoring scheme, properly normalized hybridization intensities (scores) are used (R. Drmanac *et al.*, in preparation). OSS of the pairs of dots prepared in two independent PCR reactions from the same master clones are presented on Figure 3. The scores of a particular probe are very similar for the pairs of dots except in a few percent. For example, dot (7,75) has score of 1 and corresponding dot (7,78) a score of 14 with probe F9. Possible reasons are dust or the shadow of the strong surrounding dots, especially if a dot has a small amount of DNA, which is the case for this dot (7,78); its relative mass is only 9.

Similarities of clone signatures are defined by calculating the distance parameter (Fig.4). The distribution of the distance values for identical and random clones is shown. There is a significant separation of the two classes of pairs, which allows accurate identification of identical and similar pairs in a large set of clones. A. Milosavljevic (personal communication) motivated by the success of our score-based distance calculation procedure, has developed rank scaling of hybridization intensities and a different procedure for OSS comparisons. Also, we have defined an additional way to measure the similarity of the clone signatures (R. Drmanac *et al.*, in preparation). In this approach, the number of probes which have significantly similar scores for the given pair of clones is defined. The evaluation of the advantages and limitations of these approaches is in the progress.

Clones with significantly small distances are grouped (clustered) together. The result obtained by the clustering procedure applied on the data from a test experiment is presented in Table 2. For this test every clone was spotted twice. Clones having low mass and an insignificant number of positive scores (in this case less than 4) were eliminated from the comparisons. In this small set of 876 clones, 744 genes are represented (15% of the clones are redundant). We confirmed by restriction mapping (Figure 5) that clusters consist of similar (not identical) clones in the majority of cases. The differences among the similar clones can not be explained by variation in clone size only. These clones represent either highly similar members from the gene families or alternatively spliced (or maybe trans-spliced) messengers. In the clustering procedure used, clones which differ up to 30% in size (or have 30% — i.e., a few exons of non-corresponding sequences) are recognized as significantly similar.

5. A Scheme for Gene Analysis by Combining Partial SBH and Single-Pass Gel Sequencing

Approximately 100,000 human cDNA clones have been sequenced from one or both ends by single-pass gel sequencing (at a cost of about $3 million). Screening one million clones probably can not be done in less than five years. If the reads are 300 bp on average, 600 Mb will be produced (twofold coverage of the expected 300 Mb of expressed sequences). By a very

rough estimate based on the described models, we expect that less than 80% of the genes will be found. For most of them, only one piece will be determined, and it will be difficult to define the relationships of these clones and to identify artifacts. It is unlikely that more than a half a million cDNA clones from the all other organisms will be end-sequenced in the next five years using presently available resources.

The described partial SBH approach is very cost- and time-effective. With our existing facilities, half a million clones can be analyzed in one year with 200 probes for less than $1 per clone of total cost. Further automation of the clone-managing and probe-labeling procedures, and particularly hybridization steps can increase the screening throughput to two million clones per year (Intelligent Automation Systems, Boston, MA, is constructing a machine to automatically operate 24 boxes which would have a daily throughput of 3 million clone/probe scores). The types of data expected from large-scale screening and the types of probes planned to be used are listed in the Tables 3 and 4, respectively. We recently have screened 20,000 cDNAs from the mentioned human brain libraries with 260 probes, and data collection for an additional 40,000 clones is in process.

Partial SBH information is distributed over the whole insert and allows an estimate of the overall similarity of clone sequences. A smaller number of false positive or false negative clone matches is expected than if the comparisons are based on end sequences only. Also, 200 probes provide enough information to match the signatures with corresponding known gene sequences (unpublished results, 22). Because it is possible to screen several millions of clones in a few years, comprehensive gene catalogues with estimated expression levels for most of the genes in the analyzed tissues can be established. The enormous throughput can allow the elimination of artifacts by counting only cases represented by at least two clones. The main disadvantage of partial SBH based on a small number of probes is impossibility of translating OSSs into protein sequence. Significantly long stretches of amino acids can be defined if 1000 or more probes are scored.

Data collected by SBH screening can be useful in several ways. Representative clones can be used as probes for gene mapping by FISH or by screening YAC, BAC, or cosmid libraries. In addition, a minimal set of representative clones (one gene equivalent) can be spotted on membranes for screening by genomic probes (e.g., cosmid or YAC clones) or by mRNA populations expressed under various physiological conditions. This can simplify the identification of genes for genetic diseases.

Complete gene sequencing can be rationalized and significantly accelerated. First of all, representative clones from the new families can be selected for complete sequencing. The molecular genetics will benefit enormously from the studies of the new gene families. The family of homeobox genes is a suggestive example. The complete sequencing of long mRNAs will be facilitated by selecting displaced clones from the defined contigs.

In addition to cDNA screening, partial SBH with less than 1000 probes (5) and single-pass gel sequencing can be combined to provide inexpensive overviews of genomic sequences (15,20). Resolution of fine genome structures and the identification of genes and their probable functions are anticipated (5,8,15,20). Partial cDNA and genomic sequences from a few species can be generated to the end of this century for a half of the cost required to sequence the human genome completely. A "partial sequences first" approach gives an opportunity to practice "sequenetics" on the present level of development of sequencing techniques. An inexpensive full sequencing (15,20) and routine individual resequencing by next-generation methods (multiplex sequencing (23), directed sequencing by modular primers (24), capillary electrophoresis, mass

spectrometry (25), and fast SBH by compact arrays called "sequencing chips", reviewed in refs. 8 and 15) will further improve the accuracy of the predictions and extend the field of the genetic discoveries achievable by comparative sequence analysis.

ACKNOWLEDGMENTS

In a discussion, Dr. George Church (Harvard Medical School, Boston, MA) brought to our attention the transcription leakage problem. We are grateful for the thoughtful comments of Drs. Radomir Crkvenjakov and Frank Collart. We thank David E. Nadziejka for technical editing and Kay Bexson for technical assistance. Work supported by the U.S. Department of Energy, Office of Health and Environmental Research, under Contract No. W-31-109-ENG-38.

REFERENCES

1. I. Labat, N.A. Stavropoulos, and R. Drmanac, Search for re-estimated 150,000–200,000 human genes in unsequenced genomic fragments, in: "Genome Mapping and Sequencing," Cold Spring Harbor Laboratory, Cold Spring Harbor, NY (1993).

2. D.M. Chikaraishi, Complexity of cytoplasmic polyadenylated and nonpolyadenylated rat brain ribonucleic acids, Biochem.:18:3249–3256 (1979).

3. N. Chaudhari, and W.E. Hahn, Genetic expression in the developing brain, Science:220:924–928 (1983).

4. R.J. Milner, and J.G. Sutcliffe, Gene expression in rat brain, Nucl. Acid Res.:11:5497–5519 (1983).

5. R. Drmanac., G. Lennon, S. Drmanac, I. Labat, R. Crkvenjakov, and H. Lehrach, Partial sequencing by oligohybridization: concept and applications in genome analysis, in "Electrophoreses, Supercomputing and the Human Genome," C.R. Cantor and H.A. Lim, eds., World Scientific, Singapore (1991).

6. R. Drmanac, I. Labat, I. Brukner, and R. Crkvenjakov, Sequencing of megabase plus DNA by hybridization: theory of the method, Genomics:4:114–128 (1989).

7. R. Drmanac, How 1000 base pairs with 10% error become 10,000 base pairs of correct sequence: a felicitous marriage of the gel and hybridization sequencing methods. Abstracts of Genome Mapping and Sequencing Meeting, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY (1992).

8. R. Drmanac, and R. Crkvenjakov, Sequencing by hybridization (SBH) with oligonucleotide probes as an integral approach for the analysis of complex genomes, Int. J. Genome Res.:1:59–79 (1992).

9. M.D. Adams, J.M. Kelley, J.D. Gocayne, M. Dubnick, M.H. Polymeropoulos, H. Xiao, C.R. Merril, A. Wu, O. Olde, R.F. Moreno, A.R. Kerlavage, W.R. McCombie, and J.C. Venter, Complementary DNA sequence: expressed sequence tags and human genome project, Science:252:1651–1656 (1991).

10. A.S. Wilcox, A.S. Khan, J.A. Hopkins, and J.M. Sikela, Use of 3′ untranslated sequences of human cDNAs for rapid chromosome assignment and conversion to STSs: implications of an expression map of the genome, Nucl. Acids Res.:13:1837–1843 (1991).

11. K. Okubo, N. Hori, R. Matoba, T. Niiyama, A. Fukushima, Y. Kojima, and K. Matsubara, Large scale cDNA sequencing for analysis of quantitatiave and qualitatiave aspects of gene expression, Nature Genetics:2:173–179 (1992).

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

13. R. Drmanac, S. Drmanac, Z. Strezoska, T. Paunesku, I. Labat, M. Zeremski, J. Snoddy, W.K. Funkhouser, B. Koop, L. Hood, and R. Crkvenjakov, DNA sequence determination by hybridization: a strategy for efficient large-scale sequencing, Science:260:1649–1652 (1993).

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

15. R. Drmanac, S. Drmanac, J. Jarvis, and I. Labat, Sequencing by hybridization, in "Automated DNA Sequencing and Analysis Techniques," J.C. Venter, ed., Academic Press, London, in press.

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

17. T. Blumenthal, Mammalian cells can trans-splice? But do they?, BioEssays:15:347–348 (1993).

18. J.P. Bruzik, and T. Maniatis, Spliced leader RNAs from lower eukaryotes are *trans*-spliced in mammalian cells, Nature:360:692–695 (1992).

19. P.G. Zaphiropoulos, Differential expression of cytochrome P450 2C450 2C24 transcripts in rat kidney and prostate: evidence indicative of alternative and possibly trans splicing events, Biochemical and Biophysical Research Communications:192:778–786 (1993).

20. R. Drmanac, S. Drmanac, I. Labat, A. Vicentic, A. Gemmell, N. Stavropoulos, and J. Jarvis, SBH and the integration of complementary approaches in the mapping, sequencing, and understanding of complex genomes, in "Proceedings of Second International Conference

on Bioinformatics, Supercomputing and Complex Genome Analysis," H.A. Lim, J.W. Fickett, C.R. Cantor, and R.J. Robbins, eds., World Scientific, Singapore (1992).

21. G.S. Lennon, and H. Lehrach, Hybridization analyses of arrayed cDNA libraries. TIG:7:314–317 (1991).

22. A. Milosavljevic, Discovering sequence similarity by the algorithmic significance method, in "Proceedings of the First International Conference on Intelligent Systems for Molecular Biology," H. Hunter, D. Searls, and J. Shavlik, eds., AAAI Press, Menlo Park, CA (1993).

23. G.M. Church, and S. Kieffer-Higgins, Multiplex DNA sequencing, Science:240:185–188 (1988).

24. J. Kieleczawa, J.J. Dunn, and F.W. Studier, DNA sequencing by primer walking with strings of contiguous hexamers, Science:258:1787–1791 (1992).

25. L.M. Smith, The future of DNA sequencing, Science:262:530 (1993).

26. R. Drmanac, Z. Strezoska, I. Labat, S. Drmanac, and R. Crkvenjakov, Reliable hybridization of oligonucleotides as short as six nucleotides, DNA and Cell Biology:9:527–534 (1990).

FIGURE LEGENDS

Figure 1. cDNA inserts amplified by PCR. Row A from four 96-well plates is tested on an agarose gel by loading 3 µl from each well. One to two inserts per row (15% on average) give very weak or invisible bands. Tfl polymerase (Epicentre Technologies, Madison, WI) which gives under our conditions better yield than AmplyTaq (Perkin Elmer, Norwalk, CN) was used. The superiority of Tfl has been demonstrated by D. Grujic and R. Crkvenjakov (personal communication).

Figure 2. Hybridization patterns of two 7-mer probes with 7776 cDNA dots. Images represent one quarter of a filter (Gene Screen, NEN, Boston, MA) containing 31,104 dots. Grids are superimposed by our image analysis (DOTS) program. The first image is obtained by a coding-specific probe NNTGATGGTN, and the second by a zinc finger consensus probe (G,A)AAGCCNTTC. Hybridization conditions (15, 26) are not full-match-specific in the case of the zinc finger probe.

Figure 3. Oligonucleotide sequence signatures. The column labeled "Relative Mass" represents relative hybridization intensities obtained by a probe complementary to the coamplified vector sequence (mass probe) with dots containing DNA in comparison to the average signal of intentionally created empty dots. Columns 1–20 represent a subset of probes hybridized to one filter. Below the column number, the hybridization date and probe name are specified. Each row represent hybridization score values obtained with one dot specified by the row and column number on the filter. Pairs of dots separated by horizontal lines are generated from the same master clone by repeated PCR. Score values are adjusted for the difference in the amount of DNA using relative mass values (5). A score value of 1 represents dots with no detectable match.

Figure 4. Distribution of OSS distances for identical and random clone pairs. The histogram is generated by Histo program (J. Jarvis, unpublished information). The formula for the distance calculation is written under the histogram. P is number of probes; L and S are the larger and smaller score values obtained by a particular probe with the given pair of clones. Bars represent how many (specified in some cases) clone pairs have distances which fall in the given range.

Figure 5. Restriction mapping of two clones with highly similar hybridization signatures. Inserts are amplified by PCR, digested by *Alu* I and *Hae* III restriction enzymes, and separated on 2% agarose gel.

Table 1 Model distributions of genes in six levels of expression in a cell type

	Number of mRNAs per gene per cell					
	0.1	0.5	5	30	200	2000
Model 1						
Total genes	1,000	2,000	5,000	6,000	1,000	10
Cell type specific genes	0	0	120	180	45	5
Housekeeping genes	0	0	4,880	5,820	500	5
% Clones	0.0	0.2	6.0	42	47	5.0
% Genes in 10^5 clones[a]	2	10	60	99	100	100
No. Genes in 10^5 clones	20	200	3,000	5,940	1,000	10
Model 2						
Total genes	50,000	15,000	7,000	4,000	1,000	10
% Clones	1.3	1.9	9.0	31	52	5
% Genes in 10^5 clones	2.5	12	72	100	100	100
No. Genes in 10^5 clones	1,300	1,900	4,500	4,000	1,000	10
Model 3						
Total genes	80,000	40,000	9,000	2,000	1,000	10
Cell type genes	0	0	240	60	45	5
Housekeeping genes	0	0	8,760	1,940	500	5
% Clones	2.3	5.9	13	18	58	5.8
% Genes in 10^5 clones	2.9	14	76	100	100	100
No. Genes in 10^5 clones	2,300	5,500	7,000	2,000	1,000	10
Model 4						
Total genes	20,000	80,000	10,000	1,500	500	20
% Clones	0.7	14.5	18	16	36	14.5
% Genes in 10^5 clones	3	18	83	100	100	100
No. genes in 10^5 clones	700	14,000	8,300	1,500	500	20

[a]The percentages are calculated by equation $1-(1-1/g)^c$, where g is the total number of genes in the given level and c is the number of clones representing these genes.

Table 2 Profile of a cDNA-screening experiment

Probes	51
Dots:	3456
PCR samples	1728
Control clones	96
Low mass	590 (36%)
Analyzable	1042 (64%)
Less than four hits	166 (16%)
For pairwise comparison	876 (54%)
Single clone clusters	687 (78%)
Multiple clone clusters	57 (3.3 clones/cluster)
Distinct clusters	744 (85% of 876 clones)

Table 3 Benefits from a large-scale cDNA secreening by oligonucleotide hybridization

1.	Catalogs of genes and gene families
2.	Tissue and time expression pattern for most genes
3.	Compositional features (G+C, Alu, coding capacity, motifs)
4.	Identification of the longest clone for each mRNA
5.	Clone contigs for long mRNAs (random priming)

Table 4 Composition of an oligonucleotide set suitable for cDNA analysis

Type	No. specified bases	No. probes
Coding specific	7–8	40
Alu repeat	8–7	24
Gene domains	7–8	20
Extreme G/C or A/T	7–10	16
Protein motifs	7–9	40
Overlapped	7–8	34
Exceptional	5–6 and 9–15	26
		200

Figure 1

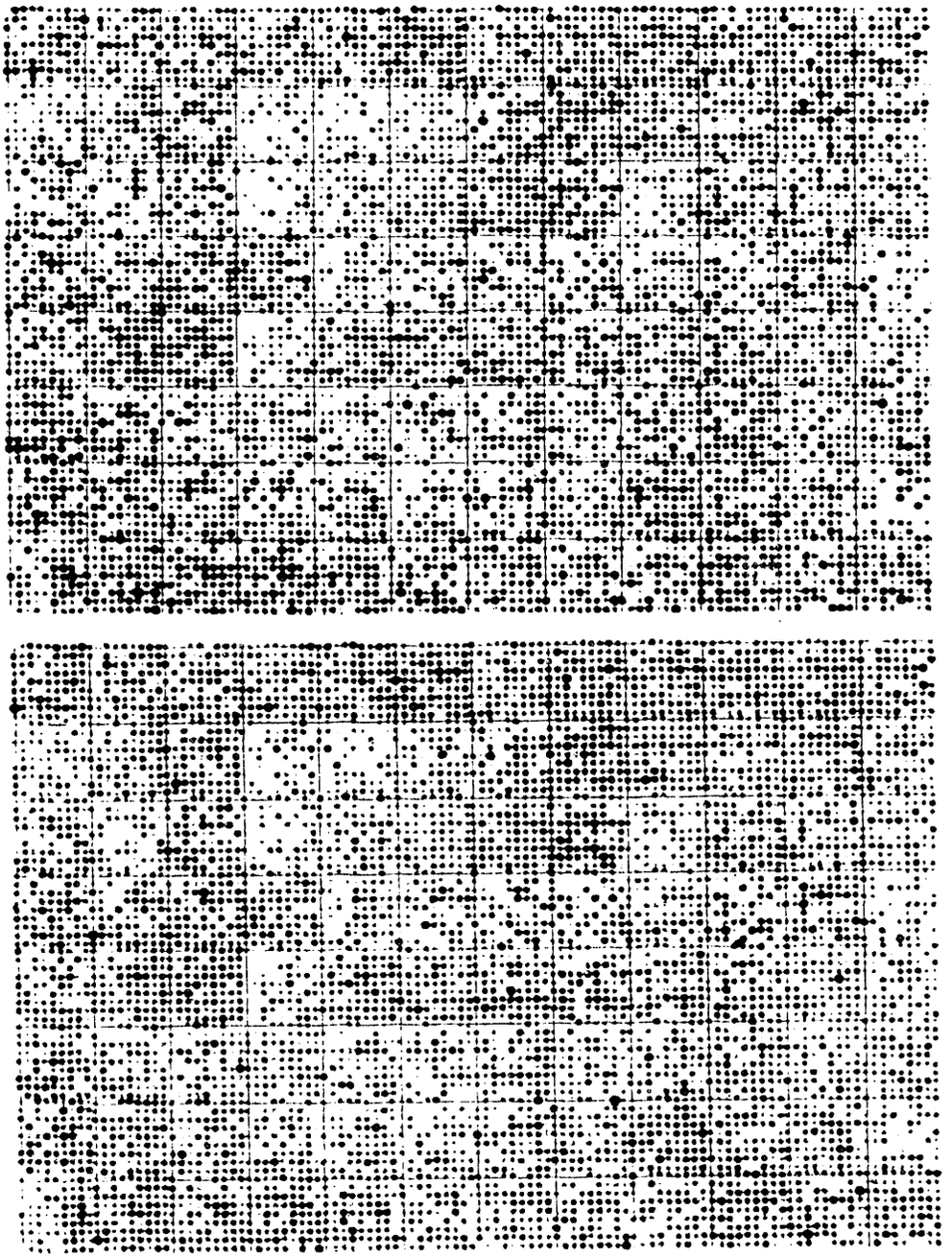

Figure 2

	Relative	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Date:	Mass	0415	0423	0425	0427	0428	0429	0430	0503	0504	0505	0507	0508	0511	0512	0514	0517	0519	0521	0524	0525
Probes:		7-25	12N	1I	F24	F3	E10	E30	7-10	7A	F21	8C	C8	F9	F11	5N	E37	7-31	7N	3N	E21
Clone..																					
(7, 3)	19	3	1	1	2	5	5	17	9	1	1	3	1	1	2	1	2	1	3	1	9
(7, 6)	12	3	1	1	3	5	4	18	10	1	3	3	3	2	3	1	2	2	4	1	8
(7, 12)	42	7	9	1	3	3	1	4	9	1	4	2	2	1	1	7	4	1	7	2	1
(7, 15)	32	7	7	1	3	4	2	5	4	1	5	3	1	2	1	5	3	1	4	2	1
(7, 21)	16	5	6	2	1	5	1	11	3	3	1	2	3	1	1	8	3	1	1	7	10
(7, 24)	12	11	10	3	3	10	2	14	5	4	2	6	4	2	2	13	4	3	2	10	15
(7, 30)	20	1	1	4	3	1	2	1	5	1	2	3	1	3	2	1	1	1	1	5	2
(7, 33)	14	2	2	3	4	1	2	4	6	2	2	4	1	4	2	1	1	1	1	8	3
(7, 39)	33	3	6	1	2	2	1	3	6	1	1	5	1	1	1	2	1	1	3	3	5
(7, 42)	23	3	16	1	2	4	1	3	6	1	2	5	1	1	2	1	1	1	6	3	7
(7, 48)	75	3	1	1	1	1	2	1	1	1	2	3	1	1	8	1	1	1	2	4	2
(7, 51)	41	3	1	1	1	1	5	1	1	1	3	5	3	1	10	1	1	3	4	6	2
(7, 57)	38	1	4	1	1	1	1	3	3	6	1	3	3	1	2	2	2	2	3	2	3
(7, 60)	13	1	4	3	2	1	2	4	3	5	2	3	5	4	7	3	2	4	3	2	3
(7, 66)	65	1	4	1	2	1	1	1	8	1	2	7	8	1	1	1	2	2	24	2	1
(7, 69)	60	1	5	1	3	1	1	1	7	1	2	7	8	1	1	1	2	2	24	3	1
(7, 75)	24	2	2	1	3	13	2	1	5	2	1	1	2	1	4	3	1	2	1	1	10
(7, 78)	9	2	2	6	4	7	1	4	5	2	1	1	2	14	3	4	1	4	1	2	6
(7, 84)	63	1	1	1	2	2	1	1	1	1	2	5	3	7	2	1	3	1	1	4	1
(7, 87)	66	1	3	1	2	2	2	1	1	1	2	4	3	8	3	1	3	1	1	5	3

Figure 3

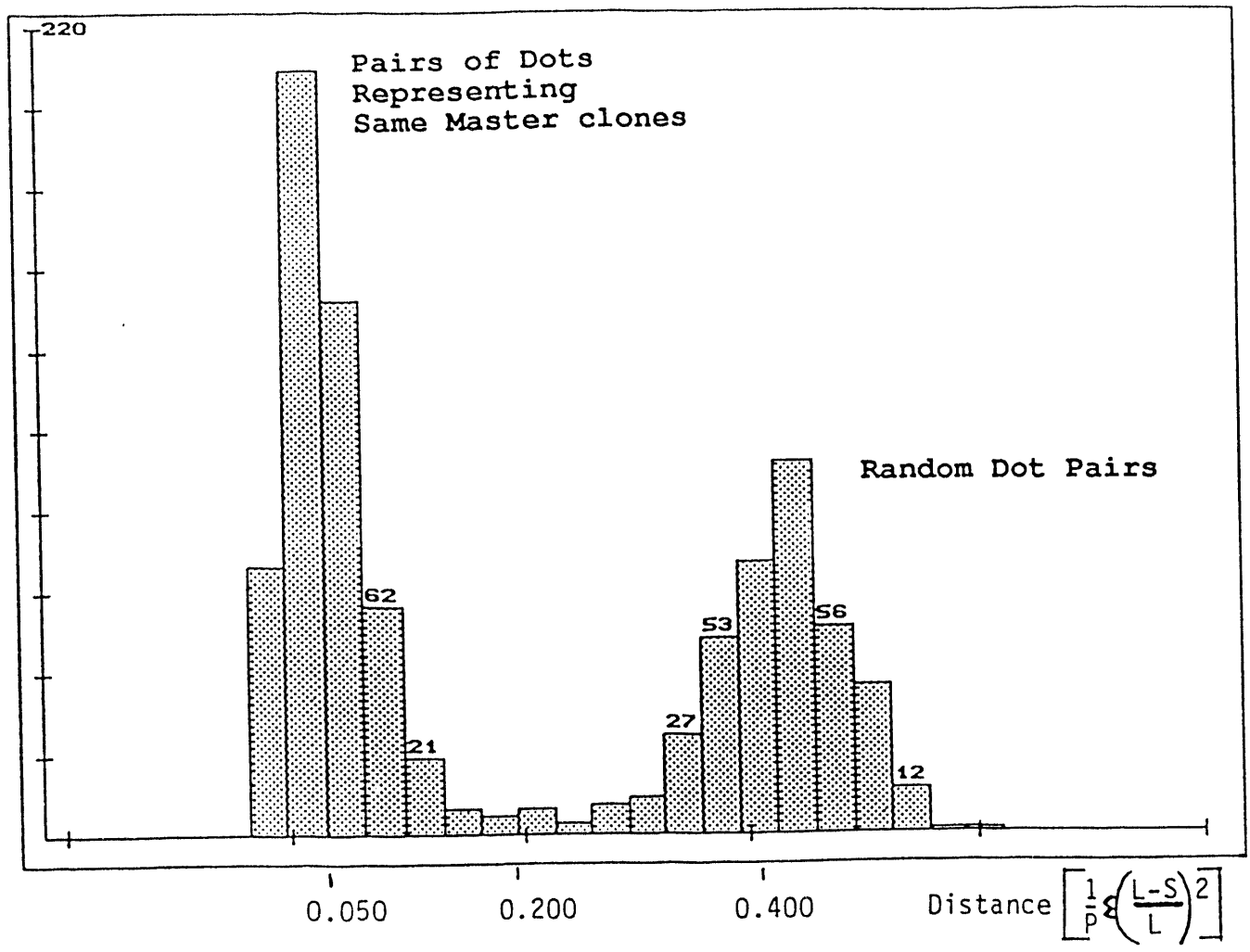
220
Pairs of Dots
Representing
Same Master clones
Random Dot Pairs
62
21
27
53
56
12
0.050
0.200
0.400
Distance $\left[\frac{1}{P}\Sigma\left(\frac{L-S}{L}\right)^2\right]$

Figure 4

Figure 5

DATE
FILMED
4/29/94
END