ELSEVIER

Genetic Analysis: Biomolecular Engineering
15 (1999) 209–215

GENETIC ANALYSIS
Biomolecular Engineering

www.elsevier.com/locate/gat

Negative selection: a method for obtaining low-abundance cDNAs using high-density cDNA clone arrays

Peter S. Nelson [a,b,*], Victoria Hawkins [a], Michel Schummer [a], Roger Bumgarner [a], Wai-Lap Ng [a], Trey Ideker [a], Camari Ferguson [a], Leroy Hood [a]

[a] *Department of Molecular Biotechnology, Box 357730 HSB, Room K360, University of Washington, Seattle, WA 98195, USA*
[b] *Department of Medicine, University of Washington, Seattle, WA 98195, USA*

Received 24 November 1998; received in revised form 29 March 1999; accepted 3 April 1999

Abstract

The identification of the entire complement of genes expressed in a cell, tissue, or organism provides a framework for understanding biological properties and establishes a tool set for subsequent functional studies. The large-scale sequencing of randomly selected clones from cDNA libraries has been successfully employed as a method for identifying a large fraction of these expressed genes. However, this approach is limited by the inherent redundancy of cellular transcripts reflecting widely variant levels of gene transcription. As a result, a high percentage of transcript duplications are encountered as the number of sequenced clones accrues. To address this problem, we have developed a negative hybridization selection method that employs the hybridization of complex cDNA probes to high-density arrays of cDNA clones and the subsequent selection of clones with a null or low hybridization signal. This approach was applied to a cDNA library constructed from normal human prostate tissue and resulted in the reduction of highly expressed prostate cDNAs from 6.8 to 0.57% with an overall decline in clone redundancy from 33 to 11%. The selected clones also reflected a more diverse cDNA population, with 89% of the clones representing distinctly different cDNAs compared with 67% of the randomly selected clones. This method compares favorably with cDNA library re-association normalization approaches and offers several distinct advantages, including the flexibility to use previously prepared libraries, and the ability to employ an iterative screening approach for continued accrual of cDNAs representing rare transcripts.

Keywords: cDNA; Library; EST; Normalization; Array

1. Introduction

One approach for identifying the complete repertoire of genes expressed in a given cell type, tissue, or organism involves the large-scale single-pass sequencing of clones randomly selected from cDNA libraries [1]. The sequences produced by such a strategy are termed expressed sequence tags (ESTs) and are capable of uniquely identifying a particular gene. Datasets of ESTs have proven useful for a variety of applications including the discovery of novel and tissue-specific genes [2].

cDNA libraries are usually constructed to faithfully represent the population of mRNAs that are present in the starting tissue. Thus, a cDNA library will reflect both the transcript diversity and redundancy inherent in the complex transcriptome of cellular mRNAs. A typical eukaryotic cell contains between 10^5 and 10^6 mRNA molecules transcribed from perhaps 10,000–30,000 different genes [3,4]. These mRNAs are distributed in three broad abundance classes: rare transcripts, consisting of $>10{,}000$ different species present at 1–15 copies per cell (40–45% of the mRNA mass), moderately abundant transcripts, consisting of approximately 1000 different species present at 100–500 copies per cell (40–45% of the mRNA mass), and highly abundant transcripts, consisting of 10–15 species present at more than 1000 copies per cell (10–20% of the mRNA mass). The distribution of this population means that efforts to profile all expressed genes from a given tissue using a random sequencing approach will quickly be faced with a high degree of redundancy and a loss of efficiency.

* Corresponding author.
E-mail address: psnels@u.washington.edu (P.S. Nelson)

1050-3862/99/$ - see front matter
PII: S1050-3862(99)00006-6

Methods designed to address this redundancy issue involve the construction of equalized or normalized cDNA libraries that simultaneously reduce the number of cDNAs derived from abundant transcripts and increase the representation of rare cDNAs in the population [5]. These methods employ reassociation kinetics that take advantage of the high sequence specificity of homologous nucleic acid hybridization. While this strategy has been successfully employed to construct normalized cDNA libraries for large-scale sequencing projects [6], the procedures are complex and may not always achieve the desired reduction in clone redundancy because some gene sequences are not normalized efficiently [5–7]. In addition, truncated clones are on occasion favored over their longest counterparts during the normalization process [7]. Finally, cDNAs containing repetitive sequences are not removed in the normalization processes involving reassociation kinetics [7]. While this may be useful in that rare species harboring DNA repeats will not be removed from the population, we have found that repetitive sequences comprise 80–90% of the single-pass sequence read of 5–10% of randomly selected cDNA clones [8]. As these repetitive sequences are generally not useful for database homology searches, and otherwise add little information, it would be desirable to remove these clones from the sequencing effort.

In this communication, we describe an alternative method for the selection of cDNA clones representing moderately abundant and rare species in a cDNA library. The approach entails replicate arraying randomly selected cDNA clones at high density on nylon membranes followed by hybridization with a complex total cDNA probe synthesized from the same starting mRNA as that used in constructing the cDNA library. cDNAs representing abundant species exhibit intense hybridization signals, while cDNAs representing moderately to lowly abundant transcripts exhibit low, or no hybridization signals. Selecting the 'negative' cDNAs for sequencing greatly reduces the redundancy of high-throughput sequencing efforts, and increases the diversity of represented clones.

2. Methods

2.1. Library construction and probe preparation

A directional cDNA library was constructed from polyA + RNA isolated from total normal human prostate tissue as described previously [8], and designated PN001. The library was not normalized or amplified. cDNA clones were plated onto 22×22-cm ampicillin-containing agar plates at a density of ≈ 3000 clones per plate. Individual clones were randomly picked into 384-well microtiter plates containing LB media with 8% glycerol and 100 μg/ml ampicillin. The plates were incubated overnight at 37°C in humidified chambers until growth was detected. Clones were replica-spotted onto pre-wetted nylon membranes (Schleicher and Schuell) using an automated robotic spotting tool (Q-bot, Genetix, Inc), or a hand-held spotting device [9]. After spotting, the membranes were placed on top of 22×22 cm LB agar plates and grown for 8–12 h at 37°C until colonies were visible, but not touching. The colonies were lysed and DNA bound to the membranes using a previously described colony lysis procedure [10]. The filters were air dried and UV cross-linked prior to hybridization.

2 μg of polyA + RNA extracted from normal human prostate were reverse transcribed in a 30 μl reaction containing 50 mM Tris–HCl (pH 8.3), 75 mM Kcl, 3 mM $MgCl_2$, 10 mM DTT, 200 μM each dATP, dGTP, and dTTP, 20 μM dCTP, 5 mCi/ml ^{32}P-dCTP (3000 Ci/mmol) (DuPont/NEN), 100 μg/ml oligo dT-30, and 20 U/μl Superscript II reverse transcriptase (Life Technologies). The reaction was incubated at 42°C for 30 min. Unincorporated nucleotides were removed using Sephadex G-50 spin column chromatography (Pharmacia).

2.2. Filter hybridization and clone selection

High density clone filters were prehybridized for 6 h at 42°C in glass tubes containing 20 ml of solution comprised of 4x SSC (1x SSC = 150 mM NaCl, 15 mM sodium citrate), 50% formamide, 10x Denhart's, 0.1% SDS, 8% dextran sulfate, 50 mM phosphate buffer (pH 7.2), 1 mM EDTA, and 50 μg/ml sheared salmon sperm DNA. Hybridization was performed with 1×10^6 cpm probe/ml of the same solution for 14 h at 42°C. The membranes were washed at 65°C twice for 10 min in 2x SSC/0.1% SDS, once for 15 min in 1x SSC/0.1% SDS, and once for 15 min in 0.1x SSC/0.1% SDS. Filters were exposed to phosphor capture screens (Molecular Dynamics) for 8–48 h, and the signals were quantitated on a phosphorimager (Molecular Dynamics).

Clones were selected for sequencing based either upon a visual inspection of a null hybridization signal, or were selected based upon a hybridization intensity threshold determined by image analysis. The phosphorimage gel file was analyzed using a spot-finding algorithm written in Labview programming language (National Instruments, Austin, TX) to find the spot center of mass and subtract local background (Roger Bumgarner, personal communication). The output of spot intensities was exported to an Excel spreadsheet, and the clones were ordered by intensity level. Clones with signal intensities within the bottom quartile of the averaged intensities were selected for sequence analysis. Several clones with intensities in the top 10% were also selected for sequencing.

Clone inserts were amplified by the PCR using primers homologous to vector sequences flanking the cDNA inserts: VN26: tttcccagtcacgacgttgta, and VN27: gtgagcggataacaatttcac as previously described [8]. The PCR products were separated on 1.2% TAE agarose gels, and products >500 bp ($\approx 92\%$ of the clones examined) were purified in 96-well format as described previously [11] and sequenced using fluorescent dye-labeled primers as suggested in the Prism cycle sequencing kit (Applied Biosystems, Foster City, CA). This group of sequences was designated PN001-NS, for negative selection.

2.3. Sequence assemblies and redundancy/diversity determination

Prostate EST sequences from the NCI Cancer Genome Anatomy Project (CGAP) libraries CGAP-PR21, representing a non-normalized normal prostate cDNA library, and CGAP-PR22 representing a normalized normal prostate cDNA library were obtained from the CGAP web site; http://www.ncbi.nlm.nih.gov/ncicgap. Statistics describing the library quality, composition, and EST clustering results using the program estcluster was obtained from the Washington University Sequencing web site; http://genome.wustl.edu/est/esthmpg.html. The sequencing failure rates for each library are reported on this website. The failure rates described in Section 3 reflect these reported rates minus the number of mitochondrial sequences as we do not consider these ESTs as sequence failures.

In order to determine the redundancy and diversity in each group of sequences; PR21, PR22, PN001, PN001-NS, the sequences were clipped of flanking vector sequence and repetitive sequences were masked using the Repeatmasker algorithm (http://ftp.genome.washington.edu/RM/RepeatMasker.html). Sequences shorter than 100 bp were not further analyzed. Each group of sequences was assembled into clusters using the contig assembly program CAP [12]. Each contig was searched against the National Center for Biotechnology Information (NCBI) Genbank and dbEST nucleotide sequence databases using the BLASTN algorithm [13]. Selected individual ESTs representing clones with high hybridization intensities were also compared with the Genbank and dbEST databases for sequence homology and putative gene identification.

3. Results

3.1. Hybridization results and sequence analysis

A representative high density cDNA nylon array hybridization with a radio-labeled normal prostate cDNA probe is shown in Fig. 1A. A wide range of clone hybridization intensities was observed. Approximately 10% of the clones exhibited a signal equivalent to the background hybridization signal. A total of 842 cDNA clones were selected representing clones in the lowest quartile of the averaged hybridization intensities (Fig. 1B). PCR amplification of 81 clones yielded insert sizes of less than 300 bp, and an additional 61 sequences were of poor quality for an overall success rate of 80%. These results are similar to our previously published results for randomly selected clones from PN001 of 83.5% [8], and to the overall sequencing success rates of 70% for PR21, and 73% for PR22.

The remaining 700 clones were searched for similarities with sequences in the Genbank and dbEST public databases using the BLASTN algorithm. Four mRNA species have previously been shown to be highly expressed in normal human prostate; elongation factor 1-alpha (EF1-alpha), prostate specific antigen (PSA), prostate secretory protein (PSP), prostate acid phosphatase (PAP) [8]. In addition, the transcript for semenogelin [14] is highly represented in EST projects from several prostate cDNA libraries (http://www.ncbi.nlm.nih.gov/ncicgap). Of the clones chosen by negative selection, only 0.6% (4 ESTs) corresponded to these five highly expressed transcripts (Table 1). In contrast, 7.9% of CGAP-PR21 ESTs, 6.8% of PN001 ESTs, and 2.5% of the normalized CGAP-PR22 ESTs were homologous to these highly abundant species (Table 1).

Additional abundant and redundant sequences in EST sequencing projects are comprised of interspersed repetitive sequences (SINEs, LINEs, LTRs), and sequences encoded by the mitochondrial genome. The representation of mitochondrial ESTs was reduced from 5.3% in the non-normalized CGAP-PR21 library to 2.1% in the normalized CGAP-PR22 library. However, the percentage ESTs containing repetitive sequences was increased from 3.4% in CGAP-PR21 to 6.3% in CGAP-PR22. In contrast, the negative-selection procedure reduced the mitochondrial ESTs from 5.8 to 2.1% and ESTs with DNA repeats from 9.6 to 5.7% in PN001 and PN001-NS respectively (Table 1).

A potential bias in the accumulation of redundant ESTs may be introduced by the depth of EST sampling in different libraries. In general, the number of redundant sequences will increase as a function of the number of ESTs sequenced. To address this bias, a total of 500 ESTs from each library were randomly selected and examined in a separate analysis (Table 2). The percentage of ESTs corresponding to highly expressed prostate transcripts was reduced in both the normalized and negative selection cohorts when compared to their normal unmanipulated library counterparts (Table 2). ESTs corresponding to mitochondrial and repetitive sequences were not included in this analysis as these data were not available separately for the CGAP library ESTs.

Ten clones with hybridization signals measured in the upper 10% of hybridization intensities were selected and sequenced (Fig. 1). Six of these clones were homologous to regions of the mitochondrial genome located between nucleotides 8074 and 9416 of the mitochondrial consensus sequence (gbIX62996). The remaining four clones were homologous to PSA, ferritin, and genes coding for the ribosomal phosphoprotein P0 and ribosomal protein L29; all previously shown to be present at moderate to high abundance in the human prostate [8].

3.2. Determination of library redundancy and diversity

Each library of ESTs was separately assembled into clusters consisting of overlapping and contiguous DNA sequences using the assembly program CAP [12]. The number of ESTs contributing to each cluster was

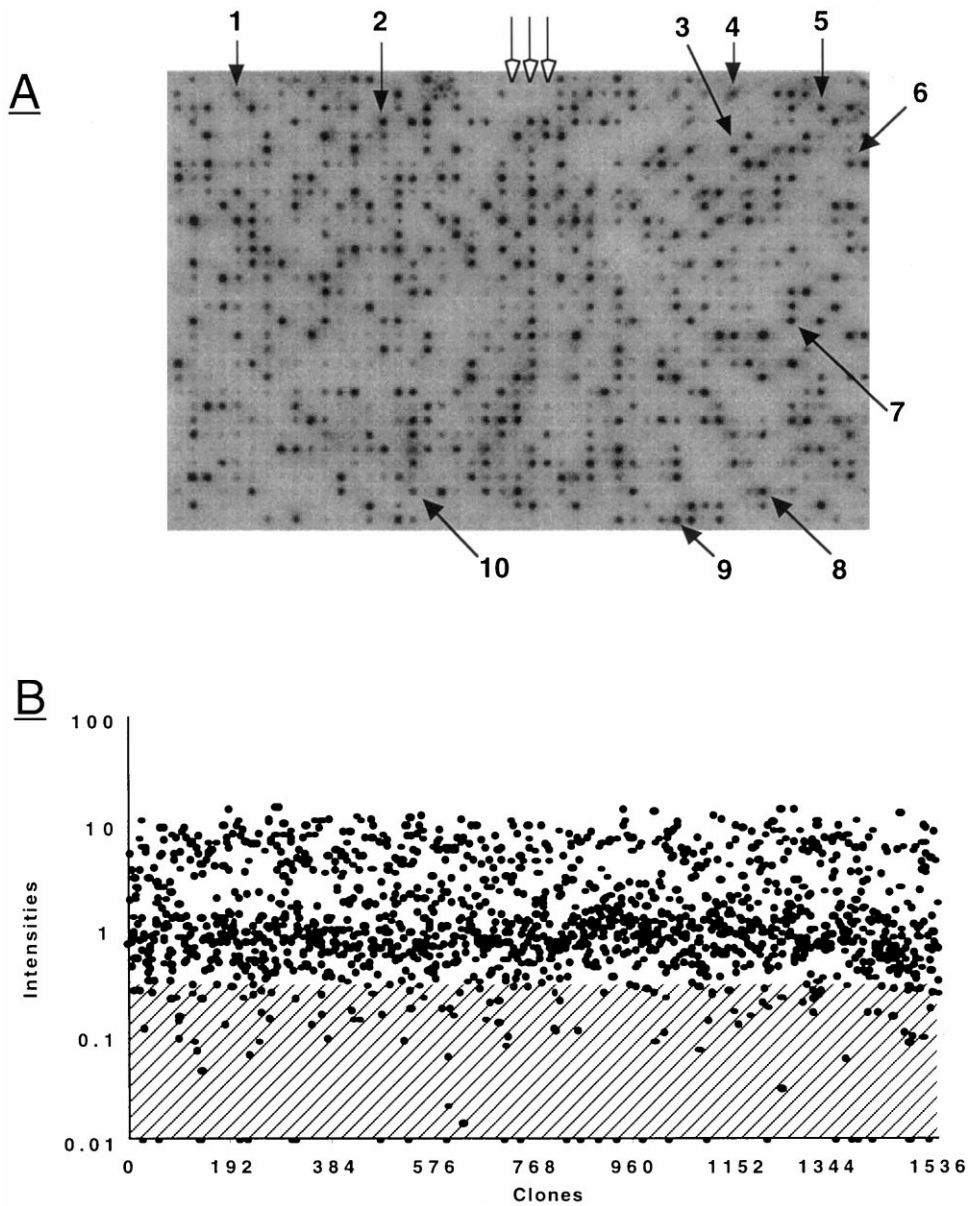

Fig. 1. Image of high density cDNA-clone array hybridization. (A) Hybridization of a 1536-clone array with a complex total cDNA probe reverse transcribed from normal human prostate mRNA. Open arrowheads identify three examples of negative-selection clones with a null or low hybridization signal; (a), phenol sulfotransferase (gbU52852), (b), ezrin (embX51521), (c), (EST D14696). Closed arrowheads identify ten clones exhibiting a high signal intensity that were selected at random. Sequencing and nucleotide database homology comparison identified these cDNAs as; 1, 4, 5, 6, 7, 8, mitochondrial genome (gbIX62996). 2. ribosomal phosphoprotein P0 (gbIM17885). 3. ribosomal protein L29 (embIZ49148). 9. ferritin (gbIM94859). 10. prostate specific antigen (gbIU17040). (B) Representation of the quantitative hybridization values on a logarithmic scale. Undetected spots have been assigned a value of 0.01 for graphing purposes. The shaded area represents clones with signal intensities in the bottom quartile of the entire group.

Table 1
Normalization and negative selection: a comparsion of two methods for reducing EST redundancy

Library	CGAP-PR21 No manipulation	CGAP-PR22 normalized	PN001 No manipulation	PN001-NS negative selection
# Sequences	1274	5445	1168	700
# Different species	731	3478	780	622
Abundant prostate ESTs				
EF-1 alpha	32	32	30	2
PSA	13	19	12	1
PSP	12	25	11	0
PAP	8	2	11	1
Semenogelin	36	59	15	0
Total abundant ESTs	101 (7.9%)	137 (2.5%)	79 (6.8%)	4 (0.6%)
Mitochondrial	68 (5.3%)	114 (2.1%)	62 (5.8%)	15 (2.1%)
Repetitive sequences	43 (3.4%)	346 (6.3%)	112 (9.6%)	40 (5.7%)
Total redundant ESTs (%)	484 (38%)	1846 (34%)	388 (33%)	78 (11%)

counted, and the consensus sequence of each cluster was searched against the public databases and annotated. Singleton ESTs that did not cluster were also searched for similarity with public database sequences, and singletons that exhibited homology to the same identified database sequence as a given cluster were counted as part of that cluster, and thus counted as redundant. Each cluster was counted as a different individual species, and each singleton that did not exhibit homology to a database sequence with similarity to a cluster was also counted as an individual species.

The number of different species for each library cohort is shown in Table 1 and reflects the sampled diversity of the library. The percentage of redundant ESTs for each library is also shown in Table 1 and Table 2. Negative selection decreased the percentage of redundant cDNAs from 33% in the randomly selected PN001 library to 11% in the PN001-NS cohort with a corresponding increase in the percentage of clones representing different species from 67 to 89%. A more accurate comparison of the results of normalization and negative selection is shown by comparing the first 500 ESTs sequenced from each library cohort (Table 2). Normalization increased the number of different EST species from 355 to 442, a change of 17%, and decreased the percentage of redundant ESTs from 29% to 11.6%. Negative selection increased the number of different EST species by 9% and decreased the percentage of redundant ESTs from 14 to 4.8%.

4. Discussion

Advances in high-density array hybridization technology have now made it possible to determine the expression patterns of thousands of genes simultaneously. Various array platforms have been developed including colony or DNA arrays on nylon membranes [15,16], DNA arrays on glass substrates [17], and arrays of synthesized oligonucleotides [18]. The capabilities of each of these methods would be greatly enhanced through the identification of non-redundant clone or sequence sets representing all genes potentially expressed in the biological system(s) under investigation. Such a non-redundant clone set would allow for the efficient design of arrays through the inclusion of cDNAs corresponding to rare transcripts and provide for the maximum use of array space.

The partial sequencing of clones from cDNA libraries to generate ESTs suitable for identifying and categorizing individual cDNAs has proven to be a valuable resource for array substrates and for defining cell and tissue transcript profiles. However, the clone redundancy inherent in cDNA libraries constructed using standard methods limits their usefulness for defining the medium and low abundance cDNAs in library population. A differential cDNA library screening strategy has been described as a means to selectively acquire rare cDNAs [19]. While useful, the method as described would not be suitable for screening tens of thousands of cDNAs due to the random spatial distri--

bution of plated clones leading to overlapping plaques. Accurate replicates are also difficult to produce that would be suitable for iterative screening.

We describe here an approach for increasing the diversity and reducing the clone redundancy in a population of cDNAs using quantitative hybridization and negative selection of clones arrayed at high density. This procedure has several advantages over other methods such as normalization and subtraction for reducing the variation in abundance among the clones in a cDNA library. Standard library construction methods are employed without the necessity for PCR, reassociation reactions, or column purification of single-stranded DNA as used in several of the normalization and subtraction methods [7,20,21]. Libraries previously made or purchased can be used without requiring new library construction. In addition, as redundant clones are identified, they can be pooled to comprise a secondary or tertiary generation probe for hybridizing to array replicates and thus eliminate cDNAs representing moderately abundant transcripts from further selection [22]. Finally, the negative selection array approach could be combined with a library normalization procedure to further enhance the efficiency of clone selection.

One limitation for the negative selection approach is the potential for selecting a higher percentage of clones without inserts as these clones would not be expected to generate a hybridization signal and thus would be chosen as negatives. This problem is highly dependent upon the quality of library construction, and a library with a high percentage of no insert clones would not be a suitable substrate. We did not observe a significant increase in the number of clones without inserts in the negative-selection cohort of PN001 clones.

Although the negative-selection process reduced the number of cDNAs homologous to highly expressed prostate transcripts, mitochondrial sequences, and interspersed DNA repeats, the method did not successfully remove all of these sequences. The selected PSA cDNA represents an alternatively spliced form of this transcript, and this may have accounted for the low hybridization signal leading to its selection. Although the mitochondrial genome encodes 37 genes [23], our EST classification annotated all mitochondrial sequences similarly as 'mitochondrial' based upon database homology to the mitochondrial consensus sequence. However, it is likely that transcripts derived from the 37 mitochondrial genes are present at different abundance levels, perhaps ranging from rare to abundant. Indeed, we have determined that 41% of the prostate mitochondrial ESTs are homologous to a region of the mitochondrial genome encoding the ATP synthase 6, cytochrome C oxidase III, and tRNA-gly genes. Mitochondrial cDNAs chosen as negatives in the selection process did not encompass this region, and thus may represent mitochondrial transcripts of low abundance. Including a mitochondrial genomic probe in a second round of negative selection may further reduce the selection of these sequences.

As picking the individual cDNA clones must be performed for sequencing projects involving either normalized libraries or the array-based approach, the major additional labor involved in the negative-array-selection process involves the actual gridding and processing of the arrayed clones. Robotic tools enhance the speed and accuracy of this procedure, but we have also described the feasibility and utility of using simple hand-held replicator devices that are available commercially for minimal cost [9]. Such devices can also be constructed from common materials (http://chroma.mbt.washington.edu/ARD/). Thus, this technique is broadly applicable for both large-scale production line approaches and small projects of limited scope.

Table 2
Normalization and negative selection: a comparsion of the first 500 library ESTs

Library	CGAP-PR21 No manipulation	CGAP-PR22 normalized	PN001 No manipulation	PN001 Negative selection
# Sequences	500	500	500	500
# Different species	355	442	430	476
Abundant prostate ESTs				
EF-1 alpha	11	5	12	2
PSA	5	2	3	1
PSP	6	2	3	0
PAP	0	0	5	1
Semenogelin	12	2	0	0
Total abundant ESTs	34 (6.8%)	11 (2.2%)	23 (4.6%)	4 (0.8%)
Total redundant ESTs (%)	145 (29%)	58 (11.6%)	70 (14%)	24 (4.8%)

Acknowledgements

We thank Steven Lasky, Carol Loretz, and Vilaska Nguyen for sequencing support. We thank Laurie Hassel for assistance with array construction.We thank Barbara Trask for critical review of the manuscript. This work received support from CaPCURE and a grant (K08 CA75173-01A1) from the National Cancer Institute (P.S.N.).

References

[1] Adams MD, Kelley JM, Gocayne JD, Dunnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, Kerlavage AR, McCombie WR, Venter JC. Science 1991;252:1651–6.
[2] Adams MD, Kerlavage AR, Fleischman RD, Fuldner RA, Bult CJ, Lee NH, Kirkness EF, Weinstock KG, Gocayne JD, White O. Nature 1995;377(Suppl 28):3–174.
[3] Bishop JO, Morton JG, Rosbash M, Richardson M. Nature 1974;250:199–204.
[4] Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, Vogelstein B, Kinzler KW. Science 1997;276:1268–72.
[5] Ko MSH. Nucleic Acids Res 1990;18:5705–11.
[6] Soares MB, Bonaldo MF, Jelene P, Su L, Lawton L, Efstratiadis A. Proc Natl Acad Sci USA 1994;91:9228–32.
[7] Bonaldo MF, Lennon G, Soares MB. Genome Res 1996;6:791–806.
[8] Nelson PS, Ng WL, Schummer M, True LD, Liu AY, Bumgarner RE, Ferguson C, Dimak A, Hood L. Genomics 1998;47(1):12–25.
[9] Schummer M, Ng W, Nelson PS, Bumgarner RE, Hood L. Biotechniques 1997;23:1087–92.
[10] Nizetic D, Drmanac R, Lehrach H. Nucleic Acids Res 1991;19:182.
[11] Wang K, Gan L, Boysen C, Hood L. Anal Biochem 1995;226:85–90.
[12] Huang X. Genomics 1992;14(1):18–25.
[13] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. J Mol Biol 1990;215:403–10.
[14] Lilja H, Abrahamsson PA, Lundwall A. J Biol Chem 1989;264(3):1894–900.
[15] Gress TM, Hoheisel JD, Lennon GG, Zehetner G, Lehrach H. Mammalian Genome 1992;3:359–71.
[16] Nguyen C, Rocha D, Granjeaud S, Baldit M, Bernard K, Naquet P, Jordan B. Genomics 1995;29:207–16.
[17] Schena M, Shalon D, Davis RW, Brown PO. Science 1995;270:467–70.
[18] Wodicka L, Dong H, Mittmann M, Ho M-H, Lockhart DJ. Nat Biotechnol 1997;15:1359–67.
[19] Hoog C. Nucleic Acids Res 1991;19(22):6123–7.
[20] Takahashi N, Ko MS. Genomics 1994;23(1):202–10.
[21] Gurskaya NG, Diatchenko L, Chenchik A, Siebert PD, Khaspekov GL, Lukyanov KA, Vagner LL, Ermolaeva OD, Lukyanov SA, Sverdlov ED. Anal Biochem 1996;240(1):90–7.
[22] Orr SL, Hughes TP, Sawyers CL, Kato RM, Quan SG, Williams SP, Witte ON, Hood L. Proc Natl Acad Sci USA 1994;91(25):11869–73.
[23] Wallace DCLM, Brown MD, Huoponen K, Torroni A. In: Aj C, editor. Human Gene Mapping 1995: a Compendium. Baltimore: Johns Hopkins University Press, 1995:910–54.