

## Subtraction of cap-trapped full-length cDNA libraries to select rare transcripts

Tomoko Hirozane-Kishikawa<sup>1,2</sup>, Toshiyuki Shiraki<sup>1,2</sup>, Kazunori Waki<sup>1,2</sup>, Mari Nakamura<sup>1,2</sup>, Takahiro Arakawa<sup>1</sup>, Jun Kawai<sup>1,2</sup>, Michela Fagiolini<sup>3</sup>, Takao K. Hensch<sup>3</sup>, Yoshihide Hayashizaki<sup>1,2</sup>, and Piero Carninci<sup>1,2</sup>

<sup>1</sup>RIKEN Yokohama Institute, Yokohama, <sup>2</sup>RIKEN, Wako Main Campus, and <sup>3</sup>RIKEN Brain Science Institute, Saitama, Japan

*BioTechniques* 35:510-518 (September 2003)

*The normalization and subtraction of highly expressed cDNAs from relatively large tissues before cloning dramatically enhanced the gene discovery by sequencing for the mouse full-length cDNA encyclopedia, but these methods have not been suitable for limited RNA materials. To normalize and subtract full-length cDNA libraries derived from limited quantities of total RNA, here we report a method to subtract plasmid libraries excised from size-unbiased amplified  $\lambda$  phage cDNA libraries that avoids heavily biasing steps such as PCR and plasmid library amplification. The proportion of full-length cDNAs and the gene discovery rate are high, and library diversity can be validated by in silico randomization.*

### INTRODUCTION

We have been preparing the mouse full-length cDNA encyclopedia (1,2) by selecting new cDNA clones by one-pass sequencing of full-length enriched cDNA libraries prepared with the cap-trapper strategy (3,4). To increase gene discovery, we have normalized and subtracted cDNAs before cloning by hybridizing biotinylated RNA/mRNAs to single-stranded cap-selected cDNAs (4). This method allowed for the removal of highly expressed mRNAs and cDNAs deriving from genes that had already been found. However, this method can only be used for relatively large tissues because it requires a relatively high quantity of mRNA (1). There are no methods to normalize and subtract full-length cDNAs deriving from tissues that are difficult to collect, such as preimplantation embryos, stem cells, or microdissected tissues, from which only a reduced quantity of RNA can be extracted. The methods developed by Bonaldo et al. (5) require plasmid libraries amplification, by which short inserts are preferentially amplified and therefore unsuitable for long-insert full-length cDNA libraries. To overcome such problems, we used the properties of our  $\lambda$  FLC vector family (6), which allows for unbiased amplification and excision into plasmid cDNA libraries

that retain the original mRNA size (2.5–3.0 kb). We have previously demonstrated that after solid-phase phage library amplification, the discovery of rare or new genes and library diversity improved (6). After unbiased bulk excision, plasmid libraries are used to produce single-stranded DNA (ssDNA), which is subjected to subtraction, followed by second-strand synthesis and electroporation into bacterial cells. This novel strategy allowed us to prepare, for the first time, full-length, large subtracted cDNA libraries from small microdissected tissues that were used for the mouse cDNA encyclopedia.

### MATERIALS AND METHODS

#### Construction of the Primary cDNA Library

RNA was extracted from microdissected neocortex in 18-, 24-, 28-, and 50-day mice, respectively, and four kinds of cDNA libraries (P18, P24, P28, and P50) were constructed as previously described (3,4,6) from 50  $\mu$ g each total RNA. Cap-trapper-selected, full-length first-strand cDNA was prepared with 1st BG primer (3,4) and ligated to a second linker by the single-strand linker ligation method (7). Second-strand cDNA was then synthesized

and digested with *Bam*HI/*Xho*I, and cDNA was ligated to the  $\lambda$  FLC-I vector (6). After packaging (MaxPlax™; EPICENTRE, Madison, WI, USA), cDNA libraries were amplified as previously described (8) in C600 cells on a semisolid phase (LB-agar plates supplemented with 10 mM MgSO<sub>4</sub>) at 50,000 plaques per 13-cm diameter dish. These controlled amplification conditions avoid competition among phage particles and reduce bias (6).

#### Tester Preparation

Phage particles (from 5.0 to 6.7  $\times 10^{10}$ ) from P18, P24, P28, and P50 libraries were infected in 10 mL BNN132 grown overnight ( $A_{600} = 0.5$ ) in 20 mM MgSO<sub>4</sub> and then cultured in LB medium at 30°C for 1 h. The plasmids were extracted with a Wizard® Plus Midipreps DNA Purification System (Promega, Madison, WI, USA). To remove *Escherichia coli* genome contamination, the plasmid was treated with 3 U Plasmid-Safe™ ATP-Dependent DNase (EPICENTRE) per microgram of DNA at 37°C for 1 h in the presence of 1 mM ATP. The sample was purified with 20  $\mu$ g proteinase K in 100  $\mu$ L reaction mixture that contained 0.2% sodium dodecyl sulfate (SDS) and 10 mM EDTA at 45°C for 15 min, extracted using phenol/chloroform and chloroform, and precipitated with isopropanol. The double-stranded plasmid DNA was converted to a single-stranded plasmid by the combined action of the *GeneII-ExoIII* enzymes from the GeneTrapper® cDNA Positive Selection System (Invitrogen, Carlsbad, CA, USA), according to the manufacturer's instructions (Figure 1). A site-specific endonuclease, *GeneII* binds to the fl origin in phagemid vectors and nicks specifically one of the strands of the supercoiled DNA. The nicked strand is then digested with *ExoIII* to generate single-stranded circles. Five micrograms of double-stranded plasmid DNA were nicked with 1  $\mu$ L *GeneII* at 30°C for 1 h in 20  $\mu$ L 1 $\times$  *GeneII* buffer. After the heat-inactivation of *GeneII* at 65°C for 5 min, 1  $\mu$ L reaction mixture was transferred for the agarose gel analysis. The remainder was chilled on ice, and after the addition of 2  $\mu$ L *ExoIII*, incubated at 37°C for 1 h. After

digestion with *GeneII* or *GeneII-ExoIII*, undigested double-stranded plasmid DNA was electrophoresed with 1  $\mu$ L each of the digested samples in a 0.8% agarose gel to confirm that the double-stranded DNA (dsDNA) was converted to ssDNA. After digestion with 20  $\mu$ g proteinase K in 100  $\mu$ L reaction mixture and extraction with phenol/chloroform then chloroform as described earlier, single-stranded plasmid was ethanol-precipitated and used as the tester.

**Driver Preparation**

The driver DNA was generated by PCR with the  $\lambda$  DNA from the library (Figure 1). The  $\lambda$  DNA was extracted as previously described (9). PCR amplification was performed with 25 ng  $\lambda$  DNA as template and 2.5 pmol each primer. The primers used for PCR were 1stBG-T3 (5'-CTCTCTCAAT-TAACCTCACTAAAGGGAGACCTTCTGGAGAGT<sub>16</sub>-3') and 19 linker-T7 (5'-TTTTTTTAAATACG-ACTCACTATAGGGAGAAGGT-GACTATAGAACCA-3'). PCR was carried out using 2.5 U *LA Taq*<sup>TM</sup> polymerase (Takara, Tokyo, Japan) in 1 $\times$  GC Buffer I (Takara) and 0.4 mM dNTPs in a volume of 100  $\mu$ L. The cycling conditions were 94°C for 3 min, 70°C for 5 min (addition of hot start polymerase), 94°C for 45 s, 55°C for 1 min, and 25 cycles of 68°C for 5 min and 65°C for 2 min. PCR-amplified fragments were verified by agarose gel electrophoresis. PCR products were then purified by proteinase K treatment and extracted with phenol/chloroform and chloroform, and the primers were removed with the diatomaceous earth DNA extraction of dsDNA (10). To prepare an antisense single-stranded driver, linear amplification was performed with 2.5 pmol 1stBG-T3 primers and up to 1.5  $\mu$ g amplified products as a template. PCR was carried out using 2.5 U *LA Taq* polymerase in 1 $\times$  GC Buffer I and 0.4 mM dNTPs. The cycling conditions were 94°C for 3 min, 70°C for 5 min (addition of hot start polymerase), 94°C for 45 s, 55°C for 1 min, and 5 cycles of 68°C for 5 min and 65°C for 2 min. After purification with proteinase K and phenol/chloroform extraction, the PCR products were

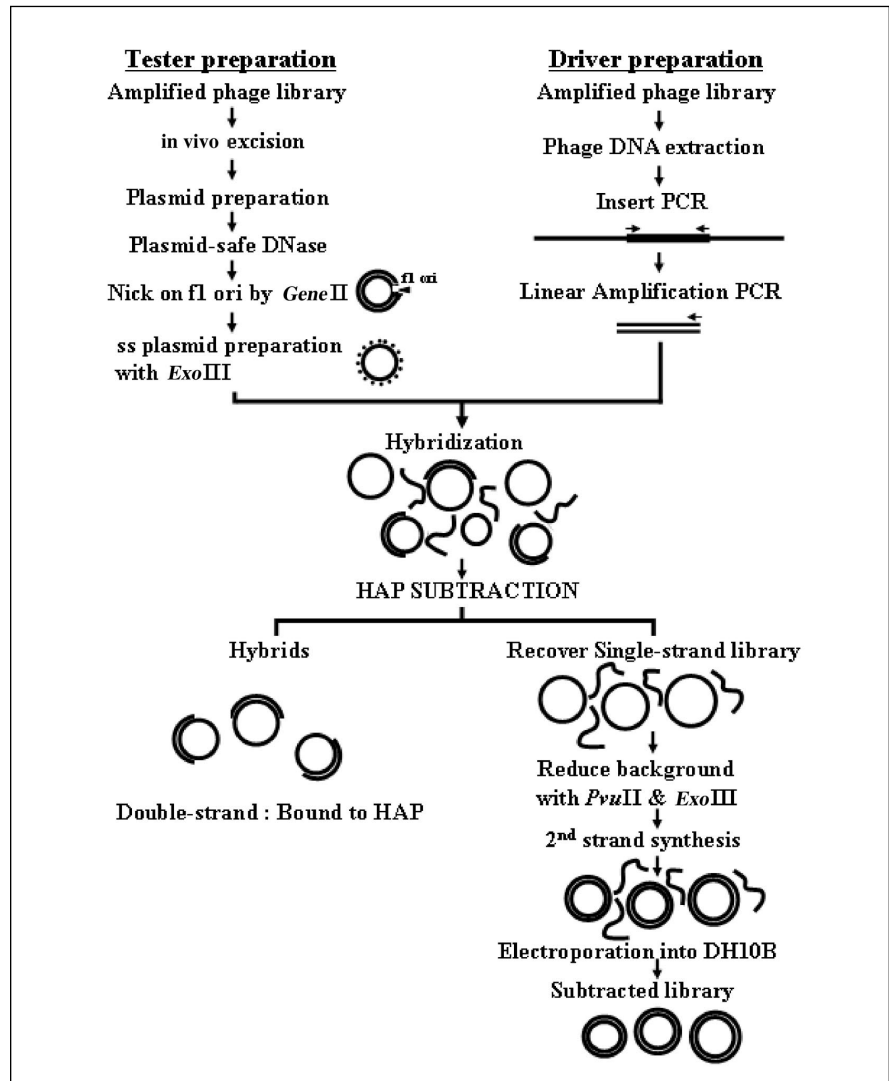


Figure 1. Representation of the library subtraction strategy. HAP, hydroxyapatite; ss, single-strand.

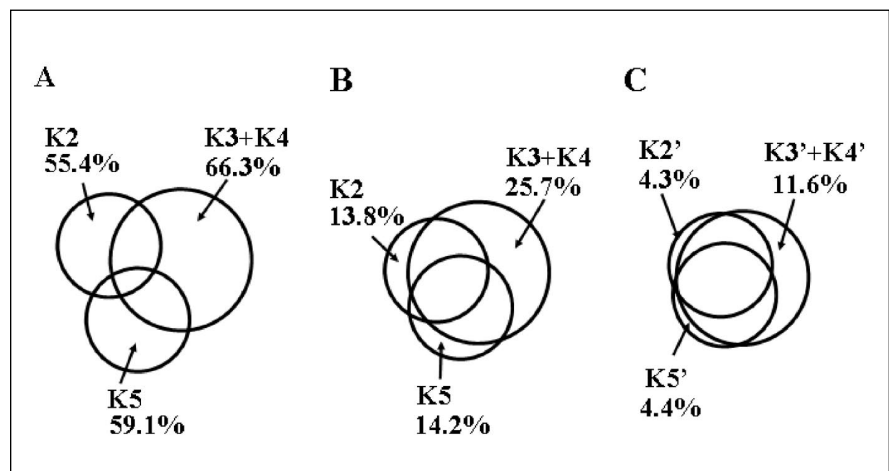


Figure 2. Validation of subtraction efficiency. (A) Ratio of K2, K3+K4, and K5 library-specific clusters plus singletons. (B) Ratio of K2, K3+K4, and K5 library-specific clusters (without singletons). (C) Ratio of computer-randomized K2', K3'+K4', and K5' library-specific clusters (without singletons).

ethanol-precipitated with 1/4 volume of 10 M ammonium acetate, dissolved in water, and used as a driver.

### Hybridization

A total amount of 10 (20-fold excess: low ratio) or 50  $\mu\text{g}$  (100-fold excess: high ratio) of linear amplification products (driver) from each library (Table 1) were mixed with 1.5 of 5' blocking oligonucleotide 19linker-T7 (5'-TTTTTTTAAATACGACTCAC-TATAGGGAG AAGGTGACAC-TATAGAACCA-3') and 1.5  $\mu\text{g}$  of 3' blocking oligonucleotide BlockBGT3GA (5'-AAAAAAAAAAAAAAAAACTCTC-CAGAAGGTCTCCCTTTAGT-GAGGGTTAATTGAGAGAG-3'). These oligonucleotides were used to block the portion of 5' and 3' PCR primer used for the driver preparation. The mixture was ethanol-precipitated, dissolved in 4.5  $\mu\text{L}$  hybridization buffer (50% deionized formamide, 25 mM HEPES, pH 7.5, 5 mM EDTA), and heated at 80°C for 3 min. Single-stranded plasmid (tester; 0.5  $\mu\text{g}$ ) was dissolved with 4.5  $\mu\text{L}$  heat-denatured driver mixture. An aliquot (0.45  $\mu\text{L}$ ) was transferred to another tube, kept in a freezer at -20°C during hybridization, and used as a control. One microliter of 5 M NaCl was then added to the driver/tester mixture, and the hybridization was performed at 4°C for 8 h (Cot 75, low ratio; Cot 200, high ratio). The Cot value was calculated as described by Anderson et al. (11).

### Capture of Single-Stranded DNA by Hydroxyapatite Treatment

Calibrated batches of Hydroxyapatite DNA Grade Bio-Gel<sup>®</sup> HTP Gel (HAP; Bio-Rad Laboratories, Hercules, CA, USA) were kept rigorously desiccated in a dry space, hydrated just before use with 0.12 M NaPi buffer (0.12 M sodium phosphate buffer, pH 6.8, 1 mM EDTA, 1% SDS), and boiled for 30 min. One milliliter of hydrated HAP (500  $\mu\text{L}$  bed vol; 0.2 g HAP) was transferred to a 2-mL tube and centrifuged for 10 s at 15,000 $\times g$ , followed by the removal of the supernatant. The hybridized sample library was diluted with 500  $\mu\text{L}$  HAP buffer, incubated at 65°C, and added to the HAP, mixed vigorously, and rotated at 60°C for 30 min. After

centrifugation for 10 s at 15,000 $\times g$ , the supernatant containing ssDNA was recovered in a new tube. The HAP was washed twice with 500  $\mu\text{L}$  HAP buffer, rotating at 60°C for 30 min, and the supernatants were mixed with the recovered single-stranded samples. In recent experiments, we performed two rounds of HAP treatment, and dsDNA hybrids were completely subtracted. The HAP-treated sample was extracted with phenol/chloroform and then chloroform and purified by cetyltrimethylammonium bromide (CTAB) precipitation. A final 0.1% concentration of CTAB was added to the sample. After 1 h of incubation at room temperature, the sample was centrifuged at 15,000 $\times g$  for 15 min, and the supernatant was removed and washed with 0.1% CTAB. The pellet was resuspended in 150  $\mu\text{L}$  7 M GuHCl, and 375  $\mu\text{L}$  ethanol were then added and precipitated. To remove any possible contaminating double-stranded plasmid DNA, the recovered sample was digested with 15 U *PvuII* in 100  $\mu\text{L}$  1 $\times$  M Buffer (Nippon Gene, Tokyo, Japan), pH 7.5, for 2 h at 37°C and treated with 65 U *ExoIII* for 30 min at 37°C. The subtracted sample was treated with 1  $\mu\text{L}$  20 mg/mL proteinase K in the presence of 0.2% SDS and 10 mM EDTA at 45°C for 15 min, extracted with phenol/chloroform and then chloroform, and ethanol-precipitated. The pellet was dissolved in 15.75  $\mu\text{L}$  water.

### Second-Strand Synthesis

To prepare the second-strand reaction, we mixed the subtracted single-stranded plasmid with 6  $\mu\text{L}$  100 ng/ $\mu\text{L}$  reverse primer (5'-TAACAATTTCA-CACAGGAAACA-3'), 30  $\mu\text{L}$  MasterAmp<sup>™</sup> Extra-Long PCR 2 $\times$  PreMix 2 (EPICENTRE), and 6  $\mu\text{L}$  5 mM [nicotinamide-adenine dinucleotide (NAD<sup>+</sup>); Sigma, St. Louis, MO, USA]. The reaction mixture was incubated at 50°C for 5 min and supplemented with 1.25 U FailSafe<sup>™</sup> PCR Enzyme Mix (EPICENTRE), 1.25 U ELONGASE<sup>®</sup> Enzyme Mix (Invitrogen), and 2.5 U Ampligase<sup>®</sup> DNA Ligase (EPICENTRE). The second-strand synthesis was performed as the following annealing-extension program: 50°C for 1 min, 68°C for 5 min, and 3 cycles

Table 1. Summary and Analysis of Data on Primary Libraries and Subtracted Libraries

Primary Library				Subtracted Library										
Primary Library ID	Stages (ages)	Titer	Insert Size (kb)	Library ID	Tester Library	Subtracting Driver Library	Insert Size (kb)	5' end Sequencing Passes	Clusters plus Singletons	Redundancy	Full-Coding Rate (%)	No. EST (%)	No. NT (%)	Unique (%)
P18	18 days	9.70E+05	2.7	K2	P18	P24, P28, P50	2.7	19,296	10,945	1.76	90.63	1269 (11.6)	1357 (12.4)	3140 (28.7)
P24	24 days	1.40E+06	2.5	K3	P24	P18, P50	2.6	16,094	9258	1.74	90.56	1242 (13.4)	1362 (14.7)	2827 (30.5)
P28	28 days	1.60E+06	2.6	K4	P28	P18, P50	2.9	21,223	11,912	1.78	90.73	1687 (14.2)	1807 (15.2)	3542 (29.7)
P50	50 days	1.70E+06	2.5	K5	P50	P18, P24, P28	2.5	20,347	11,875	1.71	89.64	1757 (14.8)	1908 (16.1)	4019 (33.8)

No. EST and No. NT: number and percentage of sequences without relevant similarity to expressed sequence tag (EST) or GenBank nucleotides (nonredundant GenBank, EMBL, DNA Data Bank of Japan, or Protein Data Bank sequences), respectively.  
 Unique indicates the number and percentage of clusters plus singletons that do not have relevant homology with RIKEN clones.  
 Full-coding rate denotes the percentage of clones that by alignment to mouse complete mRNA showed the presence of the starting ATG.

of 72°C for 1 min, 80°C for 1 min, and 65°C for 10 s. Double-stranded plasmid DNA was treated with 20 µg proteinase K in the presence of 0.2 % SDS and 10 mM EDTA, extracted with phenol/chloroform and then chloroform, and ethanol-precipitated. The plasmids were dissolved in 3 µL water and electroporated into ElectroMAX DH10B™ cells (Invitrogen) at 2.5 KV/cm.

**Sequence Analysis**

About 20,000 clones from each subtracted library (K2, K3, K4, and K5) were sequenced. The cDNA sequencing, sequence analysis, and clustering were performed as previously described (1,12,13). The novelty rate was calculated by comparing our sequences to the nonredundant nucleotide sequences [except for expressed sequence tags (ESTs)] that were updated on Jan. 14, 2003. The full-length rate was calculated as previously described (14), which essentially consists of aligning the ESTs produced from K2–K5 libraries to known mRNA from public databases that are annotated as mouse and complete coding sequences. To confirm the subtraction efficiency, we randomized in silico the library of origin of sequence data (15) and compared it with the experimental data. If the libraries are the same, then the number of library-specific clusters will be similar to the random assignment made by the computer. We assigned an uniform random number *P* (0.0 < *P* < 1.0) to all sequences from K2–K5 libraries and assigned each sequence to a randomized library as follows: range1 = *n* of sequences of K2/*n* of all sequences and range2 = *n* of sequences of K2, K3, and K4/*n* of all se-

quences. If (0.0 < *P* ≤ range1), then the sequence falls in the randomized library K2'; if (range1 < *P* ≤ range2), then the sequence falls in the randomized library K3'+K4'; and if (range2 < *P* ≤ 1.0), then the sequence falls in the randomized library K5' (Table 2).

All sequences in libraries K2–K5 that passed quality-control tests were submitted to the DNA Data Bank of Japan and therefore propagated to other public databases (GenBank® accession nos. BY237236–BY298532 and BY599747–BY660055).

**RESULTS AND DISCUSSION**

**Analysis of the Subtracted Libraries**

We performed library subtraction, starting from four primary libraries prepared from developing mouse brain neocortex (postnatal 18-, 24-, 28-, and 50-day stages) and assessed them by determining the titer, average size of the inserts, rate of full-length cDNA, redundancy (by sequencing the 5' end of the inserts), and isolation of new genes from ESTs, GenBank nucleotides, or RIKEN clones (Table 1). The titer of the subtracted libraries was one order lower than that of control non-subtracted libraries (data not shown), suggesting that the subtractions efficiently removed approximately 90% of the tester. The 5' end sequence redundancy of the subtracted libraries was 1.71–1.78 (Table 1). Furthermore, these subtracted libraries revealed the effective identification of unknown genes (28.7%–33.8%) compared to the RIKEN sequences (Table 1).

We further resequenced 3' ends from these libraries, which showed a redundancy of 1.995 after sequencing an average 9297 clones and a novelty rate of 20.3% per cluster (10.2% novel sequences/reaction). This was compared to the most similar normalized/subtracted cDNA library prepared with a proven method (4), the RIKEN library A8–300 (cortex, 10-day neonate mice), which showed slightly higher redundancy (2.14) but only after sequencing more clones (15,147). However, the A8–300 library showed lower novelty rate, 15.5% per cluster (approximately 7.23% novel sequences/reaction). Al-

though it is impractical to compare the novelty rate of cDNA libraries that are sequenced at different depths, our data suggest that the new methodology is at least comparable to the proven one (4) for gene discovery.

Library manipulations, such as amplification, conversion to single strand, hybridization, HAP chromatography, reconstitution of the double strand, and reinsertion into bacteria, have been invariantly associated with a dramatic decrease in the size of cDNA inserts and representation bias. Previously described normalized and subtracted libraries have relatively short inserts and a low rate of full-length cDNA (5,16). In this report, we have identified the experimental conditions by which large, full-length cDNA subtracted libraries can be obtained. Through rigorous testing (data not shown), our protocol has been shown to optimize the following: (i) the large size of bulk plasmid excision; (ii) optimal conversion to single-stranded plasmid library; (iii) full-length driver preparation; (iv) hybridization conditions and reagent treatments that do not degrade long full-length single-stranded plasmid DNA; (v) avoidance of the binding of long single-stranded plasmid DNA to HAP; (vi) novel purification from phosphate buffer; and (vii) efficient reconstitution of second strand, which is particularly difficult for long ssDNA. The average insert size of the K2–K5 subtracted libraries was 2.7, 2.6, 2.9, and 2.5 kb, which was indistinguishable from the original primary libraries, P18, P24, P28, and P50 (2.7, 2.5, 2.6, and 2.5 kb, respectively) (Table 1). Therefore, relevant size by library subtraction bias can be ruled out. Large size was shown to promote gene discovery when using full-length cDNA libraries (6). Moreover, we confirmed the full-coding rate by alignment with GenBank's mouse complete mRNA sequences (14). The full-coding rate of the K2–K5 subtracted libraries was 90.63%, 90.56%, 90.73%, and 89.64%, respectively (Table 1), which suggests that the clones of the subtracted libraries retain reasonably high full-length rate, which clearly differs from notably lower full-length rate values (20%–30%) previously obtained with other normalized and subtracted cDNA libraries (16).

### Subtraction Efficiency and Specificity

Clustering analysis of approximately 80,000 sequences from subtracted libraries showed that the number of clusters was low compared to singletons (clusters of clones that appeared only once), indicating the diversity of the subtracted libraries (Table 2). To confirm the subtraction efficiency, we first compared experimental data with computer-randomized data (Table 2). K3 and K4 subtracted libraries were integrated because these libraries were prepared using the same libraries (P18 and P50) as a driver (Table 1). The ratio of K2, K3+K4, and K5 library-specific clusters plus singletons was 55.4%, 66.3%, and 59.1%, respectively. In contrast, the ratio of computer-randomized K2', K3'+K4', and K5' library-specific clusters plus singletons was 46.3%, 58.3%, and 47.5%, respectively.

Since the subtracted libraries contained many singletons, we then compared the subtraction efficiency without calculating the singletons. The ratio of library-specific clusters (without singletons) was 13.8%, 25.7% and 14.2%, respectively. In contrast, the ratio of each randomized library-specific cluster (without singletons) was 4.3%, 11.6%, and 4.4%, respectively, which represents the distribution that we would have obtained if the libraries had the same complexity but no difference (Figure 2).

The library specificity of the clusters was significantly higher (up to 3.2-fold) in the subtracted libraries, even for such similar tissues, compared to computer-randomized data. This suggests that among them, 9.5% to 14.1% of the clusters are likely to represent developmental stage-specific clusters, apart from the singletons that appeared in only one condition that accounted for 55% to 66% of the groups. Further analysis is in progress to verify the distribution differences among clusters that appear in more than one library and then to verify the expression sequence variability among clusters of such cDNAs on microarrays. As an alternative to deep sequencing and statistical analysis, one could identify control genes that are known to be overexpressed or underexpressed in the analyzed tissues and per-

Table 2. Subtraction Efficiency

Library ID	Total Sequence	Total Clusters (Singletons)	Library-Specific Sequences	Library-Specific Clusters (%)	Total Clusters without Singletons	Library-Specific Clusters without Singletons (%)
K2	21,670	9681(4668)	6675	5360 (55.4)	5013	692 (13.8)
K3+K4	39,485	15,256 (8351)	13,065	10,124 (66.3)	6905	1773 (25.7)
K5	23,504	11,041(5777)	7538	6523 (59.1)	5264	746 (14.2)
K2'	21,665	10,912 (4787)	5354	5050 (46.3)	6125	263 (4.3)
K3'+K4'	39,498	16,473 (8696)	10,791	9601 (58.3)	7777	905 (11.6)
K5'	23,496	11,401(5136)	5721	5412 (47.5)	6265	276 (4.4)

K2, K3+K4, and K5 indicate the experimental data. K2', K3'+K4', and K5' indicate the computer-randomized data.

form semiquantitative RT-PCR on the libraries to verify subtraction. However, this method requires prior knowledge of gene expression and may not be straightforward in genome centers.

A major concern in cDNA library

subtraction is the nonspecific removal-related sequences. Although the specificity of HAP at 60°C to retain only specifically hybridized cDNAs has already been demonstrated (5), we have tested whether members of gene

families are spared by nonspecific hybridization. Among the 4690 clusters and 2076 singletons for which the 5' end sequence allowed for the annotation of a protein name, we checked the frequency of gene family members. For instance,

there were 115 different clusters plus 77 different singletons that were annotated as "kinases" (0.28% of the annotated clusters plus singletons). Clusters annotated as receptors, including known markers, appeared 68 times among clusters and 81 among singletons (0.21%). There were 32 zinc finger clusters and 19 singletons. Some zinc fingers appeared specifically in some libraries, while others did not. A cluster similar to zinc finger 35 appeared only in the K2 library (8 times), the zinc finger proteins 346 appeared only in K2 (5 times) and K5 (twice), while other zinc finger proteins appeared quite more randomly distributed, such as the most represented (zinc finger protein 9), which appeared 8, 6, 10, and 5 times, respectively, in K2–K5 libraries. Although these libraries were not prepared to calculate the absolute gene expression but the differential gene expression, and expression will be further validated by microarrays, this analysis suggests that cDNA family members are not dramatically (if at all) affected by subtraction and are present in our libraries.

Library subtraction is a very promising technology because it can be used with small samples for which a satisfactory quantity of mRNA is unavailable for subtraction before cloning. This technology is being used to further collect the remaining very rare full-length cDNAs derived from 246 cDNA libraries that we prepared previously using nonredundant rearranged drivers (1). These libraries include many microdissected libraries such as preimplantation embryos (1); for instance, see our UniGene Library 5391 (<http://www.ncbi.nlm.nih.gov/UniGene/lbrowse.cgi?ORG=Mm>), a two-cell stage library, library 12245, a blastocyst library, or other microdissected libraries available on the same page. This technique is also being adapted to normalize small-scale cDNA libraries by self-hybridizing a cDNA library at a 1:1 ratio and can be used in any other project to discover rarely expressed genes from amplified libraries.

#### ACKNOWLEDGMENTS

We thank Y. Ishii, D. Sasaki, H. Konno, and the other members of the

*RIKEN Genome Exploration Research Group Science Center for production of the sequencing data. This study has been supported by a Research Grant for the RIKEN Genome Exploration Research Project from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government to Y.H. This study has also been supported by a grant from Rijicho Renkei Kenkyu Fund to T.H., P.C., and M.F. for Interdisciplinary Collaboration.*

#### REFERENCES

- Carninci, P., K. Waki, T. Shiraki, H. Konno, K. Shibata, M. Ito, K. Aizawa, T. Arakawa, et al. 2003. Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. *Genome Res.* 13:1273-1289.
- Okazaki, Y., M. Furuno, T. Kasukawa, J. Adachi, H. Bono, S. Kondo, I. Nikaido, N. Osato, et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 420:563-573.
- Carninci, P. and Y. Hayashizaki. 1999. High-efficiency full-length cDNA cloning. *Methods Enzymol.* 303:19-44.
- Carninci, P., Y. Shibata, N. Hayatsu, Y. Sugahara, K. Shibata, M. Itoh, H. Konno, Y. Okazaki, et al. 2000. Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. *Genome Res.* 10:1617-1630.
- Bonaldo, M.F., G. Lennon, and M.B. Soares. 1996. Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res.* 6:791-806.
- Carninci, P., Y. Shibata, N. Hayatsu, M. Itoh, T. Shiraki, T. Hirozane, A. Watahiki, K. Shibata, et al. 2001. Balanced-size and long-size cloning of full-length, cap-trapped cDNAs into vectors of the novel  $\lambda$ -FLC family allows enhanced gene discovery rate and functional analysis. *Genomics* 77:79-90.
- Shibata, Y., P. Carninci, A. Watahiki, T. Shiraki, H. Konno, M. Muramatsu, and Y. Hayashizaki. 2001. Cloning full-length, cap-trapper-selected cDNAs by using the single-strand linker ligation method. *BioTechniques* 30:1250-1254.
- Sambrook, J., E.F. Fritsch, and T. Maniatis. 1989. *Molecular Cloning: A Laboratory Manual*, 2nd ed. CSH Laboratory Press, Cold Spring Harbor, NY.
- Manfioletti, G. and C. Schneider. 1988. A new and fast method for preparing high quality  $\lambda$  DNA suitable for sequencing. *Nucleic Acid Res.* 16:2873-2884.
- Boom, R., C.J. Sol, M.M. Salimans, C.L. Jansen, P.M. Wertheim-van Dillen, and J. van der Noordaa. 1990. Rapid and simple method for purification of nucleic acids. *J. Clin. Microbiol.* 28:495-503.
- Anderson, M. and B.D. Young. 1985. Quantitative analysis of solution hybridization, p. 73-111. In *Nucleic Acids Hybridization, A Practical Approach*, IRL Press, Oxford.
- Konno, H., Y. Fukunishi, K. Shibata, M. Itoh, P. Carninci, Y. Sugahara, and Y. Hayashizaki. 2001. Computer-based methods for the mouse full-length cDNA encyclopedia: real-time sequence clustering for construction of a nonredundant cDNA library. *Genome Res.* 11:281-289.
- Shibata, K., M. Itoh, S. Aizawa, S. Nagaoka, N. Sasaki, P. Carninci, H. Konno, J. Akiyama, et al. 2000. RIKEN integrated sequence analysis (RISA) system-384-format sequencing pipeline with 384 multicapillary sequencer. *Genome Res.* 10:1757-1771.
- Sugahara, Y., P. Carninci, M. Itoh, K. Shibata, H. Konno, T. Endo, M. Muramatsu, and Y. Hayashizaki. 2001. Comparative evaluation of 5'-end-sequence quality of clones in CAP trapper and other full-length-cDNA libraries. *Gene* 263:93-102.
- Press, W.H., B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling. 1988. *Numerical Recipes in C*. Cambridge University Press, Cambridge, UK.
- Marra, M., L. Hillier, T. Kucaba, M. Allen, R. Barstead, C. Beck, A. Blistain, M. Bonaldo, et al. 1999. An encyclopedia of mouse genes. *Nat. Genet.* 21:191-194.

Received 21 February 2003; accepted 27 May 2003.

Address correspondence to Piero Carninci, Genome Science Laboratory, RIKEN, Wako Main Campus, 2-1 Hirosawa Wako-shi, Saitama, 351-0198, Japan. e-mail: rgscerg@gsc.riken.go.jp