Normalization and Subtraction of Cap-Trapper-Selected cDNAs to Prepare Full-Length cDNA Libraries for Rapid Discovery of New Genes

Piero Carninci,[1] Yuko Shibata, Norihito Hayatsu, Yuichi Sugahara, Kazuhiro Shibata, Masayoshi Itoh, Hideaki Konno, Yasushi Okazaki, Masami Muramatsu, and Yoshihide Hayashizaki

Laboratory for Genome Exploration Research Group, RIKEN Genomic Sciences Center (GSC) and Genome Science Laboratory, RIKEN Tsukuba Institute, Core Research of Evolutional Science and Technology (CREST), Japan Science and Technology Corporation (JST), Tsukuba 305-0074, Japan

In the effort to prepare the mouse full-length cDNA encyclopedia, we previously developed several techniques to prepare and select full-length cDNAs. To increase the number of different cDNAs, we introduce here a strategy to prepare normalized and subtracted cDNA libraries in a single step. The method is based on hybridization of the first-strand, full-length cDNA with several RNA drivers, including starting mRNA as the normalizing driver and run-off transcripts from minilibraries containing highly expressed genes, rearrayed clones, and previously sequenced cDNAs as subtracting drivers. Our method keeps the proportion of full-length cDNAs in the subtracted/normalized library high. Moreover, our method dramatically enhances the discovery of new genes as compared to results obtained by using standard, full-length cDNA libraries. This procedure can be extended to the preparation of full-length cDNA encyclopedias from other organisms.

It has been tempting to prepare and use full-length cDNA libraries (Kato et al. 1994; Maruyama and Sugano 1994; Edery et al. 1995; Carninci at al. 1996, 1998; Carninci and Hayashizaki 1999) in large-scale gene discovery efforts incorporating one-pass sequencing that resemble the existing EST projects (Adams et al. 1991, 1995; Hillier et al. 1996; Marra et al. 1999). One advantage of such an approach is that most clones contain the complete coding sequence as well as the 5′ and 3′ untranslated regions (UTRs), thus dramatically accelerating the subsequent sequencing, biocomputation, and protein expression and other functional assays. However, generating full-length cDNA libraries has some inherent problems. The preparation of full-length cDNA is more efficient for short mRNAs than for long transcripts. In addition, cloning and propagation is more difficult for long cDNAs than short cDNAs, thus introducing further size bias. Using truncated cDNAs to retrieve the full-length cognate is impractical on the genomic scale; however, cDNAs in a standard library can be cloned in either their full-length or truncated forms, thus favoring discovery of at least one EST for any gene, regardless of its length.

Another problem associated with gene discovery reflects the nature of the cellular mRNA. Depending on their expression, mRNAs can be defined as superprevalent (or abundant), intermediate, or rare. In a typical cell, 5–10 species of superprevalent cDNA comprise at least 20% of the mass of mRNA, 500–2000 species of intermediately expressed mRNA comprise 40%–60% of the mRNA mass, and 10,000–20,000 rare messages may account for <20%–40% of the mRNA mass. This average distribution may vary markedly between tissue sources, and the presence of numerous highly expressed genes may further unbalance this distribution. Sequencing cDNAs from standard cDNA libraries is ineffective for discovering rarely expressed genes, when intermediately and highly expressed cDNAs would be sequenced redundantly.

We are working on the mouse full-length cDNA encyclopedia project, the ultimate goal of which is to collect at least one full-length cDNA for every expressed gene, regardless of the tissue (http://genome.rtc.riken.go.jp/). To this end, we wanted to remove not only redundant cDNAs but also sequences that were represented already in a previous library, thus accelerating the discovery of new, full-length cDNAs. Therefore, we wanted to develop a technology capable not only of normalizing the frequencies of full-length cDNAs from mRNAs belonging to the three different classes of expression but also of subtracting cDNAs that have already appeared in other libraries.

[1]Corresponding author.
E-MAIL rgscerg@rtc.riken.go.jp or carninci@rtc.riken.go.jp; FAX 81-298-369098.
Article and publication are at www.genome.org/cgi/doi/10.1101/gr.145100.

We considered several possible strategies that were based on the reassociation kinetics of nucleic acids, but none were amenable to full-length cDNA approaches. Existing technologies (Soares et al. 1994; Bonaldo et al. 1996) widely used for normalization and subtraction for large-scale gene discovery through EST approaches were unappealing to us, mainly because they were not easily applicable to long cDNA inserts. These protocols rely in fact on the reassociation of the nucleic acids in amplified plasmid libraries. However, plasmid libraries are associated with a cDNA-size cloning bias that manifests as an increased cloning efficiency of short cDNAs. In addition, during library amplification before normalization and/or subtraction, the growth of cDNA clones varies with plasmid length; therefore, long clones are underrepresented after bulk amplification of the library. This discrepancy would lead to underrepresentation of long cDNAs and difficulty in cloning long, rare cDNAs.

To avoid the problems related to amplification of libraries, we wanted to develop a technique to normalize and subtract cDNA before cloning. Published protocols did not lead to equal representation among clones of different sizes, maintain the length of long cDNAs after hybridization, or incorporate simultaneous normalization and subtraction of cDNAs. Therefore, methods based on PCR (Takahashi and Ko 1994; Diatchenko et al. 1996) in which long and otherwise difficult-to-amplify cDNAs are likely to be underrepresented were unsuitable for a full-length cDNA approach. Methods in which an immobilized nucleic cDNA driver on a solid matrix to subtract mRNA tester (Sasaki et al. 1994; Tanaka et al. 1996) were unsuitable for our purposes because of the risk of mRNA degradation before cDNA synthesis. In addition, the hybridization kinetics of nucleic acids immobilized on a solid phase (Tanaka et al. 1996) is slower than those for solution hybridization (Anderson and Young 1985). Libraries created with PCR- and solid matrix–based technologies were only partially characterized and showed sequence redundancy similar to that of nonnormalized cDNA libraries used in ESTs projects.

In addressing the normalization of full-length cDNAs, we felt that an aliquot of the mRNA initially used for the cDNA library preparation would be the ideal driver because it reflects the complexity of the first-strand cDNA tester. In addition, such a strategy could be extended easily to subtract sense mRNA sources from other tissues. Further, because cDNA cloning vectors commonly used in cDNA libraries construction carry the promoter sequences of T7 and T3 RNA polymerases, it would be easy to subtract cDNAs obtained from other libraries or pools of clones that had been already categorized by one-pass sequencing. Although frequently used to separate the hybridized driver and tester, hydroxyapatite chromatography requires strict temperature control, thus rendering the procedure technically demanding. Biotinylation of the mRNA driver is an easy alternative that is amenable to upscaling. Further, biotinylation can be coupled easily to streptavidin-phenol extraction (Barr and Emanuel 1990) or techniques using magnetic beads, provided that the reported cDNA degradation caused by photobiotinylated drivers (Fargnoli et al. 1990) is prevented. Here we present the first method for preparing normalized/subtracted libraries that also facilitates high-efficiency cloning of full-length cDNAs.

RESULTS

Strategy

In preliminary experiments, we first aimed at developing or adapting technologies that addressed the following points: high-efficiency removal of mRNA drivers; lack of cDNA size reduction after hybridization that would affect the frequency of full-length cDNAs; suitability for both normalization and subtraction; low cross-reactivity between similar but unidentical sequences; and being reproducible and amenable to upscaling both in terms of size of the driver and the number of libraries to be prepared.

Our general strategy (Fig. 1) involving hybridization of first-strand cDNA to mRNA has several advantages. This methodology is a modification of the previously proposed cDNA library preparation by Cap-Trapper (Carninci and Hayashizaki 1999) and accommodates the cloning of full-length, normalized-subtracted cDNA. Our method has the benefit of being amenable to using starting mRNA for the normalization process as well as to subtraction with an in vitro–transcribed RNA driver from any other directionally cloned cDNA library, preferably one made by using Cap-Trapper technology. After the subtraction/normalization step, cDNA is cloned.

Development of the Technology

We incorporated hybridization in formamide at 42°C because the mild temperature apparently did not lead to degradation of the cDNA after prolonged incubation (not shown). We checked whether relevant nonspecific hybridization occurred at these conditions to avoid removal of related but different sequences. We used two clones that share 76.8% identity in 1554 nucleotides, which also had long stretches of ~85% homology. These clones were the mouse full-length tubulin-M β5 and an unknown mouse cDNA that is 93% similar to the Chinese hamster mRNA for beta tubulin (clone B3T). Hybridization in 0.25 M NaCl gave excellent removal of specifically hybridized clones without cross-hybridization between the two clones.

One of our primary requirements was that our

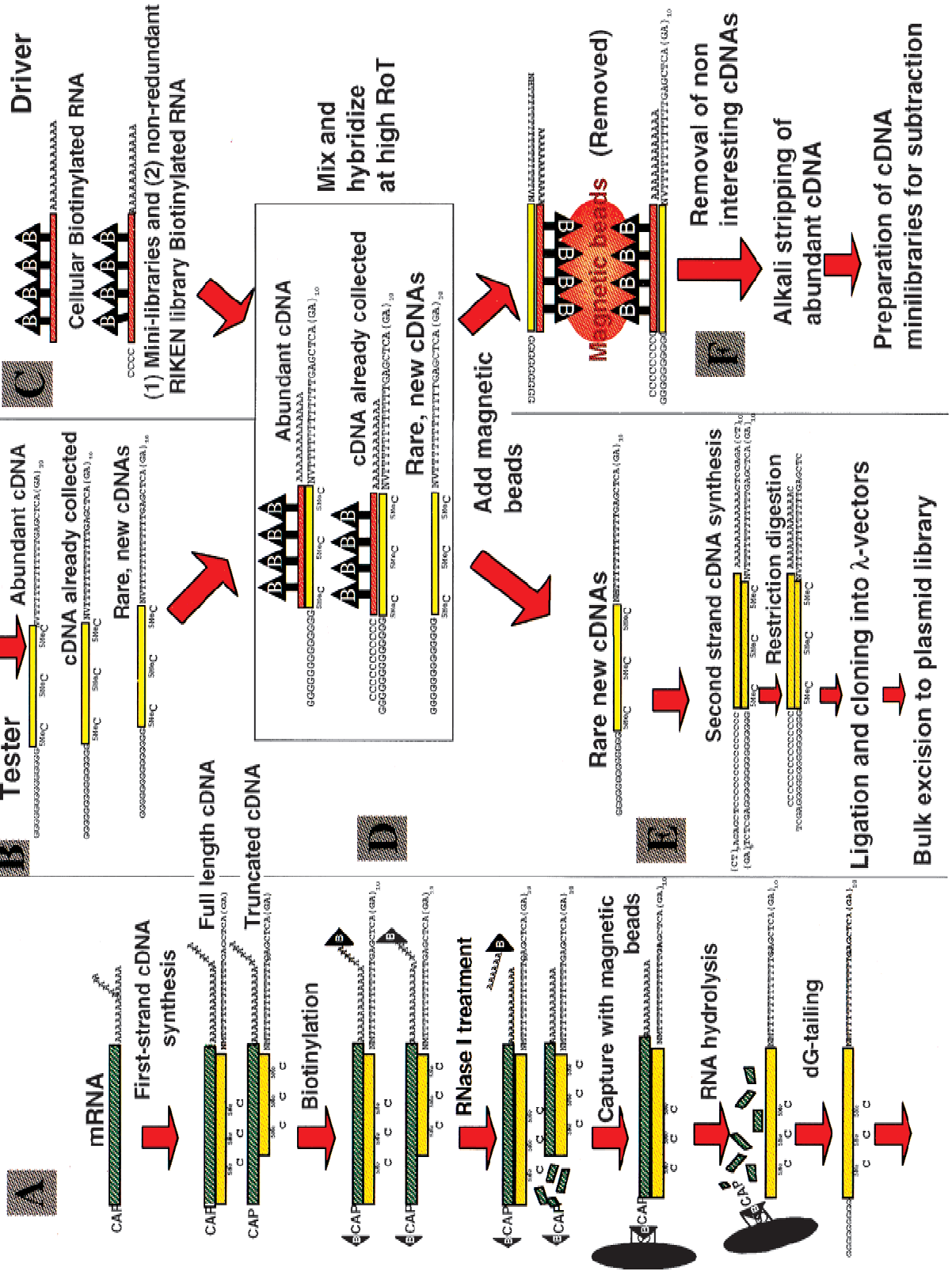

Figure 1 Schematic diagram of the normalized-subtracted cDNA preparation protocol. (*A*) general scheme for preparing full-length single-strand cDNA; (*B*) representation of various populations of tester cDNAs; (*C*) normalizing driver (cellular mRNA) and subtracting drivers (run-off transcripts); (*D*) hybridization; (*E*) rare/new cDNAs are used for second-strand cDNA preparation (normalized/subtracted cDNA library); (*F*) abundant cDNAs/unwanted cDNAs are removed and may be used for the preparation of minilibraries to implement subtraction.

method had to remove most of the RNA driver, a condition essential for removal of the tester-driver hybrid as well. Therefore, we tested whether the RNA biotinylation system afforded a high efficiency of labeling RNAs with biotin. The method that initially worked most efficiently to remove the driver was psoralen-biotinylation, which enabled removal of most of the biotinylated driver by using streptavidin beads. We verified the utility of our selected methods by hybridizing a 5-Kb tester cDNA, at the 3′-end of reeler cDNA (Hirotsune et al. 1995), to its RNA driver at RoT = 0.2 (for a detailed description of RoT see Anderson and Young 1985). This methodology led to removal of ~98%–99% of the starting cDNA, as measured by counting the radioactivity of the cDNA and by visualizing the intensity of electrophoresis smear. Because it performed as well as the psoralen-biotin system but was easier to use, we switched to the Mirus biotinylation kit (Panvera) for subsequent library preparation. To subtract the hybridized probe, magnetic porous glass (MPG) streptavidin beads (CPG) worked better in our hands than did other systems, such as the streptavidin-phenol technique; the streptavidin beads removed close to or >99% of tester-driver hybrid (not shown).

Reduction of the Frequency of Highly Expressed cDNAs

After preliminary experiments, we prepared several normalized and normalized/subtracted libraries (Table 1). cDNA libraries that were generated more recently were normalized and subtracted with both the minilibraries and the RNA drivers derived from the rearrayed nonredundant RIKEN cDNA encyclopedia to reduce wasteful resequencing of clones already represented.

We compared the second-strand cDNA from a standard pancreas cDNA library to its normalized/subtracted counterpart (Fig. 2). Normalization was performed at RoT = 10, and subtraction was accomplished by using a set of minilibraries, each of which contained 1000–2000 redundant, mainly abundant clones from liver, lung, brain, or placenta. We generated the minilibraries by cloning the highly expressed fraction of previously prepared, normalized cDNA libraries. Amplified cDNA minilibraries were then used to prepare the subtracting drivers (see Methods). The RoT for the subtracting drivers equaled 1 U for every 200 clones (e.g., RoT = 5 when 1000 clones were used). The average size of normalized, subtracted cDNA was longer than that of the nonnormalized, nonsubtracted cDNA, suggesting that degradation does not occur during the subtraction step and that long cDNAs are expressed more rarely than the short ones. We frequently have observed similar results with other cDNAs. In addition, the cDNAs corresponding to highly expressed cDNAs are absent from the normalized-subtracted library; the electrophoretic pattern demonstrates the cDNA normalization (Fig. 2). Working with full-length cDNA further helped to better visualize the removal of specific cDNAs. We did not further sequence the standard pancreas cDNA library because of the superprevalence of a very few cDNA species. Another way of demonstrating the benefit of normalization/subtraction is shown in Figure 3. We used first-strand cDNA from lung as a template and normalized or normalized/subtracted one aliquot and cloned another aliquot that was not normalized. Plaque hybridization of the normalized cDNA library and the standard counterpart suggested that the frequency of highly expressed genes was reduced in the normalized library. When we screened 10,000 plaques of the normalized lung library, the representation of elongation factor 1-α was reduced from 90 plaques in the control library to 10 in the normalized library, carbonyl reductase decreased from approximately 70 to 3, and uteroglobin was reduced from approximately 510 to 2 plaques. These results strongly suggest that the frequency of highly expressed cDNAs in the normalized library was much less than that in the control.

Increasing the Frequency of Discovering Rare Genes

To verify the enrichment of rare cDNAs, large-scale sequencing of the library is the most indicative test because we anticipated reduced sequence redundancy and increased discovery of new genes. We prepared several libraries (Table 1) and assessed them by checking the titer, average size of the cDNA inserts, presence of full-length cDNA, redundancy (by sequencing the 3′-ends of inserts), and recovery of new genes/ESTs. Assessing the degree of sequence redundancy was our final evaluation of the efficiency of the normalization/subtraction process. Standard libraries (series 22–, 23–, 26–, and 31–) prepared from an aliquot of the starting cDNA are shown for comparison (Table 1).

From Lib.32, we developed a new cloning vector that could incorporate long cDNA inserts, thereby increasing the efficiency of cloning for long cDNAs and facilitating bulk excision into a plasmid cDNA library by using the *cre-lox* system (P. Carninci, in prep.). Normalized/subtracted libraries incorporating this cloning system are deeper than previously prepared libraries, in which mainly short cDNAs were cloned. In a successful normalized-subtracted cDNA library (e.g., Lib.49 from testicular tissue) using our new cloning system, the redundancy of the sequences from 3′-ends was as low as 1.63 (calculated by dividing the number of different clusters by the total number of sequencing passes) after sequencing 8900 clones. Redundancies of <2.0 in >10,000–15,000 3′-end sequences can be expected in successful cDNA libraries from complex tissues (e.g., testis, brain, and thymus). Further, the normalized/subtracted cDNA libraries facilitated efficient and in-

creased recovery of unknown genes. For example, Libraries 22–100, 23–100, 26–100, and 31–100 produced about or more than twice the amount of new data per sequencing reaction than did the standard library counterparts 22–000, 23–000, 26–000, and 31–000 (Table 1, no-EST and no-NT columns).

Sequencing several cDNAs from various libraries reveals a relevant decrease in sequence redundancy in the normalized-subtracted library as compared to that in standard cDNA libraries (Fig. 4). Normalization increases the frequency of new gene discovery to almost twice that for standard libraries during a given sequencing effort. In comparison, subtraction with nonredundant, rearrayed drivers removes cDNAs redundant among various tissues and therefore improves the rate of new gene discovery during the course of the project.

To date, by sequencing 929,814 clones, we have been able to cluster 128,671 3′-end sequences into different groups. Because of the constant monitoring of the gene discovery rate per given cDNA library (P. Carninci, submitted), normalized/subtracted cDNA libraries were largely preferred over standard counterparts. In addition, 60,941 singletons (clusters of clones that appeared only once) were collected from 829,017 sequencing runs from normalized/subtracted cDNA libraries. Currently, we have rearrayed approximately 30,000 cDNA clones to be used for preparing RNA drivers for new cDNA libraries for the mouse cDNA encyclopedia project (http://genome.rtc.riken.go.jp/). Thus far we have prepared normalized, subtracted cDNA libraries at RoTs >200 that were producing 20%–30% of new sequences against our internal database by 3′-end reading when comparing against 70,000 different cDNA clusters.

Full-Length cDNA Rate

Of primary importance is that the full-length cDNA content is maximal after the normalization/subtraction steps. In fact, in Table 1, we can appreciate the proportion of full-length cDNAs in the various cDNA libraries. The evaluation was performed as summarized (Y. Sugahara et al., submitted). We then sequenced several hundreds of clones from the normalized-subtracted libraries. Sequences that hit "complete mRNA" sequences of mouse were aligned and checked for the presence of the initiator ATG. The presence of the initiator ATG was the factor used to assess the quality of 5′-ends instead of the exact overlap of our clones with published 5′ sequences. In fact, published "complete" sequences may differ from Cap-Trapper sequences because of differences in promoter/transcriptional start-site usage and cloning techniques. In contrast, the presence of the initiator ATG reliably shows that a given clone is practically full-length. In most of our cDNA libraries (Table 1), 80%–100% of clones included the first ATG, with an average of 88.1% in the libraries here presented. This average value goes close to standard Cap-Trapper cDNA libraries, where about 95% of full-coding cDNA was reported (Carninci et al. 1996), although the data set used for this previous analysis was different because we included the comparison of mouse to homologue genes of other vertebrates. A successful blastocyst cDNA library obtained with a cap-switch method (Sasaki et al. 1998) scored similarly (94% of clones contained the first ATG), but in this library we could cluster only 937 genes in 3995 sequencing passes. In data of another project (Marra et al. 1999), in three nonnormalized, full-length oligo-capping cDNA libraries (Maruyama and Sugano 1994), about 77% of clones contained the first ATG (Sugahara et al., submitted). By considering ESTs candidates as full-length at 5′ end when they match within 50 bp from a sequence annotated as full-length, these oligo-capping libraries were scored 65%–70% full-length at 5′ end (Marra et al. 1999), while ESTs from remaining standard and mainly normalized libraries (Bonaldo et al. 1996) scored about 27% full-length rate at 5′ end (Marra et al. 1999). Unfortunately, gene diversity from oligo-capping libraries was reduced: 2159/8231, 4463/21,594, and 2648/18,792 of "clusters versus classified ESTs" were obtained, respectively for the "mewa," "mkia," and "mlia" ibraries as clustered in the Unigene database (http://www.ncbi.nlm.nih.gov/UniGene/Mm.Home.html) on July 18, 2000.

Specificity of the Normalization-Subtraction Steps

To preliminarily evaluate the specificity of our normalization-subtraction protocol, we checked the 3′ sequences of our libraries for the presence of B1 repeats. B1 repeats are present in about 5% of the 3′-ends of cDNAs. We assumed that if the hybridization were nonspecific, the frequency of B1 repeats, which are highly homologous, would be greatly reduced in normalized-subtracted libraries because of the excess of driver carrying the B1-repeat sequence. The frequency of B1 repeats apparently does not vary between normalized-subtracted and control libraries (Table 2), suggesting that the specificity of subtraction was satisfactory. The incidence of B1 regions differs from the previously described 5% because the sequencing read-length does not span the entire 3′ UTR. Detailed analysis of full-length cDNA sequences will confirm the specificity of subtraction among gene-family members.

DISCUSSION

When we began this work, genome-scale characterization of full-length cDNAs was an important problem. Although the technology for generating full-length cDNA libraries had been described already (Kato et al.

Table 1. Summary of Data on cDNA Libraries

Library ID	Developmental stage/tissue	Pormalizing driver (Rot)	Subtracting driver (RoT)	Method	Titer	Insert size (kbp)	Sequencing	Speies	Redun-dancy	No. ESTs (%)		No. NT (%)		Unique (%)		Coding (%)
18–100	Adult/pancrea	mRNA (5)	ms1 (20)	1	8.20e + 04	1.2	13556	3402	3.98	307	(9.0)	873	(25.7)	442	(13.0)	(100.0)
22–000		(standard)	(standard)		5.90e + 04	0.88	1458	488	2.99	26	(5.3)	52	(10.7)	42	(8.6)	(82.1)
22–100	Adult/	mRNA (5)	ms1 (20)	1	3.50e + 05	1.21	4400	1932	2.28	120	(6.2)	324	(16.8)	196	(10.1)	(82.1)
22–104	stomach	mRNA (5)	ms1 (20), Nm1 (5)	1	2.00e + 05	1.13	3936	1862	2.11	144	(7.7)	302	(16.2)	207	(11.1)	(82.1)
23–000	Adult/	(standard)	(standard)		4.10e + 04		1179	556	2.12	30	(5.4)	50	(9.0)	36	(6.5)	76.8
23–100	tongue	mRNA (5)	ms1 (20), Nm1 (5)	1	4.10e + 04	1.44	10267	4017	2.56	410	(10.2)	992	(24.7)	586	(14.6)	76.8
24–100	ES cell	mRNA (5)	ms1 (20), Nm1 (5)	1	1.30e + 05	1.77	15226	4495	3.39	236	(5.3)	677	(15.1)	485	(10.8)	(88.6)
25–100	Embryo13/ liver	mRNA (5)	ms1 (20), Nm1 (5)	1	8.50e + 04	1.19	5448	1525	3.57	52	(3.4)	179	(11.7)	168	(11.0)	(92.2)
26–000	Embry10/	(standard)	(standard)		6.10e + 05	1.38	2108	1061	1.99	31	(2.9)	97	(9.1)	71	(6.7)	92.3
26–100	whole boy	mRNA (7.5)	ms1 (30), Nm1 (7.5)	1	5.00e + 05	1.32	11267	4722	2.39	330	(7.0)	870	(18.4)	582	(12.3)	92.3
28–100	Embryo 10 + 11/	mRNA (7.5)	ms1 (30), Nm1 (7.5)	1	8.80e + 05	1.29	6248	3411	1.83	190	(5.6)	450	(13.2)	271	(7.9)	(93.9)
28–104	whole body	mRNA (7.5)	ms1 (30), Nm1 (7.5)	1	8.80e + 05	1.38	9321	4335	2.15	293	(6.8)	672	(15.5)	453	(10.4)	(93.9)
31–000	Embryo/	(standard)	(standard)		4.90e + 04	1.22	488	369	1.32	12	(3.3)	30	(8.1)	23	(6.2)	(86.2)
31–100	head	mRNA (10)	ms1 (40), Nm1 (10)	1	4.20e + 05	1.55	7838	4229	1.85	344	(8.1)	682	(16.1)	494	(11.7)	(86.2)
32–304	Embryo14 + 17/ head	mRNA (10)	ms1 (40), Nm1 (10) Nm2 (10)	1	3.30e + 05	2.5	424	389	1.09	22	(5.7)	41	(10.5)	20	(5.1)	(88.2)
38–304	Embryo11/ placenta & extraembryonl c tissue	mRNA (10)	ms1 (40), Nm2 (10)	2	2.60e + 06	1.45	3657	2165	1.69	98	(4.5)	255	(11.8)	156	(7.2)	(100.0)
39–304	Embryo13/ whole body	mRNA (10)	ms1 (40), Nm1 (10) Nm2 (10)	2	2.10E + 05	2.47	348	319	1.09	14	(4.4)	33	(10.3)	22	(6.9)	(90.0)
49–304	Adult/testis	mRNA (10)	ms2 (90), Nm2 (10)	2	2.60E + 06	2.11	8900	5444	1.63	1102	(20.2)	1443	(26.5)	1214	(22.3)	(95.7)
52–304	Adult/ Xiphoid	total RNA (3)	ms2 (90), Nm2 (10)	2	7.30e + 05	2.69	272	256	1.06	12	(4.7)	21	(8.2)	15	(5.9)	(100.0)
53–304	Adult/ pituitary grand	total RNA (3)	ms2 (90), Nm2 (10)	2	2.10e + 06	2.38	8059	4658	1.73	411	(8.8)	640	(13.7)	833	(17.9)	(100.0)
54–304	Neonate6/ head	mRNA (10)	ms2 (90), Nm2 (10)	2	1.30e + 06	2.3	2663	2101	1.27	115	(5.5)	217	(10.3)	196	(9.3)	(90.0)
55–304	Neonate10/ head	mRNA (10)	ms2 (90), Nm2 (10)	2	1.70e + 06	2.18	603	525	1.15	39	(7.4)	83	(15.8)	44	(8.4)	77.3

Table 1. (*Continued*)

Library ID	Developmental stage/tissue	Formalizing driver (Rot)	Subtracting driver (RoT)	Method	Titer	Insert size (kbp)	Sequencing	Speies	Redundancy	No. ESTs (%)		No. NT (%)		Unique (%)		Coding (%)
56–304	Embryo6/ whole body	(subtracted only)	ms2 (90), Nm2 (10)	2	6.00e + 05	2.3	416	371	1.12	10	(2.7)	25	(6.7)	16	(4.3)	(100.0)
57–304	Embryo8/ whole body	(subtracted only)	ms2 (90), Nm2 (10)	2	1.20e + 06	1.91	19632	7758	2.53	778	(10.0)	1598	(20.6)	1155	(14.9)	(100.0)
58–304	Adult/ thymus	mRNA (10)	ms2 (90), Nm2 (10)	2	1.70e + 06	3.27	10259	6442	1.59	604	(9.4)	1074	(16.7)	1100	(17.1)	(80.0)
60–304	Embryo13/ testis	total RNA (5)	ms2 (90), Nm2 (10)	2	5.70e + 05		11079	6498	1.7	672	(10.3)	1122	(17.3)	1243	(19.1)	(75.0)
61–304	Embryo14/ thymus	(subtracted only)	ms2 (90), Nm2 (10)	2	4.80e + 05	4.13	206	196	1.05	9	(4.6)	24	(12.2)	16	(8.2)	(60.0)
62–304	Enbryo11/ head	mRNA (10)	ms2 (90), Nm2 (10)	2	3.30e + 05	2.19	2967	2374	1.25	149	(6.3)	265	(11.2)	256	(10.8)	(70.0)

The library ID is the identification number in the RIKEN database. The sublibrary ID is 000 for the standard library and 100, 104, or 304 for the subtracted/normalized cDNA libraries. In the subtracting or normalizing driver column, the value in brackets indicates the RoT used for each driver. The normalizing driver was always an aliquot of the starting RNA. The subtracting driver was mn1, minilibrary of liver, lung, brain, and placenta; mn2, minilibrary of liver, lung, brain, placenta, testis, pancreas, small intestine, stomach, and tongue; Nm1, RIKEN nonredundant minilibrary (4000 clones); Nm2, RIKEN nonredundant minilibrary (1600 clones). The Method column indicates the method used to label the mRNA driver. Method 1, chemical end-biotinylation ($Na10_4$-biotin hydrazide long arm for cap and 3′-end of mRNA), enzymatic biotinylation (during RNA polymerase treatment using biotin-UTP), and biotin-psoralen (all RNAs); Method 2: Label IT biotin labeling kit only. Sequencing indicates the number of 3′-end sequencing passes. Species indicates the number of different species (clusters) obtained. Redundancy is the obtained redundancy (sequence/species). No. ESTs and No. NT indicate the number and percentage of sequences from the cDNA library that do not have relevant homology with EST or GenBank nucleotides (nonredundant GenBank +EMBL +DDBJ +PDB sequences), respectively. The percentage value refers to the number of clusters (species) and not to the number of total successful sequence. Coding indicates the percentage of clones among the mouse, complete mRNA that had the initiating ATG, as evaluated by 5′-end sequencing. Values in parentheses are when clones are <50 (but at least 10).

1994; Maruyama and Sugano 1994; Edery et al. 1995; Carninci at al. 1996, 1998; Carninci and Hayashizaki 1999), several of these methods involve PCR amplification. The associated preferential amplification of specific subpopulations of cDNAs adversely affects the discovery of rare and/or difficult-to-amplify cDNAs (Maruyama and Sugano 1994). The Cap-Trapper technique does not require PCR, thus leading to production of relatively deep and unbiased libraries. However, this method was not optimized for efficient discovery of rare, full-length cDNAs by one-pass sequencing without the use of tactics such as normalization and subtraction. It was generally considered problematic to prepare normalized/subtracted cDNA libraries that are at the same time full length (Rubin et al. 2000). Here we describe the strategy and methodology we developed to prepare normalized-subtracted cDNAs for genomic-scale, full-length cDNA discovery. Our technique greatly improves on the previous situation, in which normalized-subtracted cDNA libraries typically carried primarily incompletely synthesized cDNAs. This report shows for the first time the possibility of undertaking full-length, genomic-scale gene discovery by using a sequencing approach, as we show that the size of the libraries and the proportion of full-length cDNA inserts is very satisfactory.

Regarding the Size of Rare mRNAs

We repeatedly have observed that in alkali gels and checks of plasmid size, subtracted-normalized cDNAs seem to be longer than inserts from standard libraries. This finding is probably not an artifact of our normalization-subtraction method because we have efficiently subtracted long cDNAs by using magnetic beads in test experiments; the rate of new gene discovery confirmed this trend. In addition, the inserts of normalized-subtracted cDNA libraries are not shorter than those of standard libraries prepared with the same starting RNA. This result suggests that our method preserves the integrity of the cDNA after subtraction—as was confirmed by subsequent sequence analysis. This observation further suggests that the average length of the rarely expressed mRNAs is longer than the average length of the bulk cellular mRNA. Protocols that favor production and cloning of long, full-length cDNA inserts seem to increase the rate at which new genes are discovered in full-length cDNA libraries. Of particular concern in the generation of full-length cDNA libraries is the difficulty of constructing vectors that clone short and long cDNAs with the same efficiency and stability of a long plasmid vector during the propagation of the cDNA library. Notwithstanding these difficulties, we expect that the sequencing redundancy of a satisfactory full-length, normalized/subtracted cDNA library will be <1.5 when at least 7000 clones are sequenced and <2.0 for 15,000 clones.

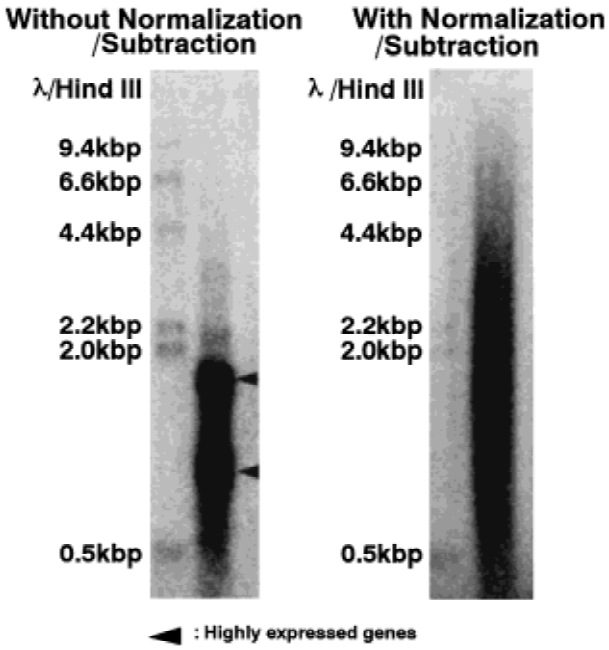

Figure 2 Visualization of removal of highly abundant full-length cDNAs. *Left,* second-strand cDNA prepared with control pancreas cDNA; *right,* cDNA prepared with an aliquot of the same pancreas cDNA after normalization/subtraction. Highly abundant cDNAs are indicated with an arrow and are removed in the normalized/subtracted cDNAs.

Relevance of Subtraction for Large-Scale Gene-Discovery Projects

Besides simple normalization, the key factor we are experiencing in the full-length cDNA gene discovery program is the importance of subtraction. Subtraction allows removal of already sequenced cDNAs as well as those that are predominantly expressed in other libraries. Subtraction helps keep the process of gene discovery efficient because resequencing of already-represented genes is reduced. Clearly this approach is facilitated by having the library production and sequencing centers in the same physical location so that feedback regarding clones to be used as drivers for subtraction occurs in a timely manner. Following this approach, we have prepared libraries subtracted with drivers corresponding to 30,000 different, previously sequenced cDNAs. In this situation, we are able to prepare libraries in which the rate at which new genes are discovered approached 25%–30% per successful sequencing reaction after clustering against a database of >80,000 3′-end sequences (not shown).

This rate of new gene discovery has to be considered extremely high after sequencing such a relevant part of 3′ ESTs. Subtraction removes 90%–95% of the mass of cDNA that might otherwise be represented in a library; this fact suggests that the rate of new gene discovery in an unsubtracted cDNA library would be 1.25%–3.0%. If we sequence 10,000 clones from a

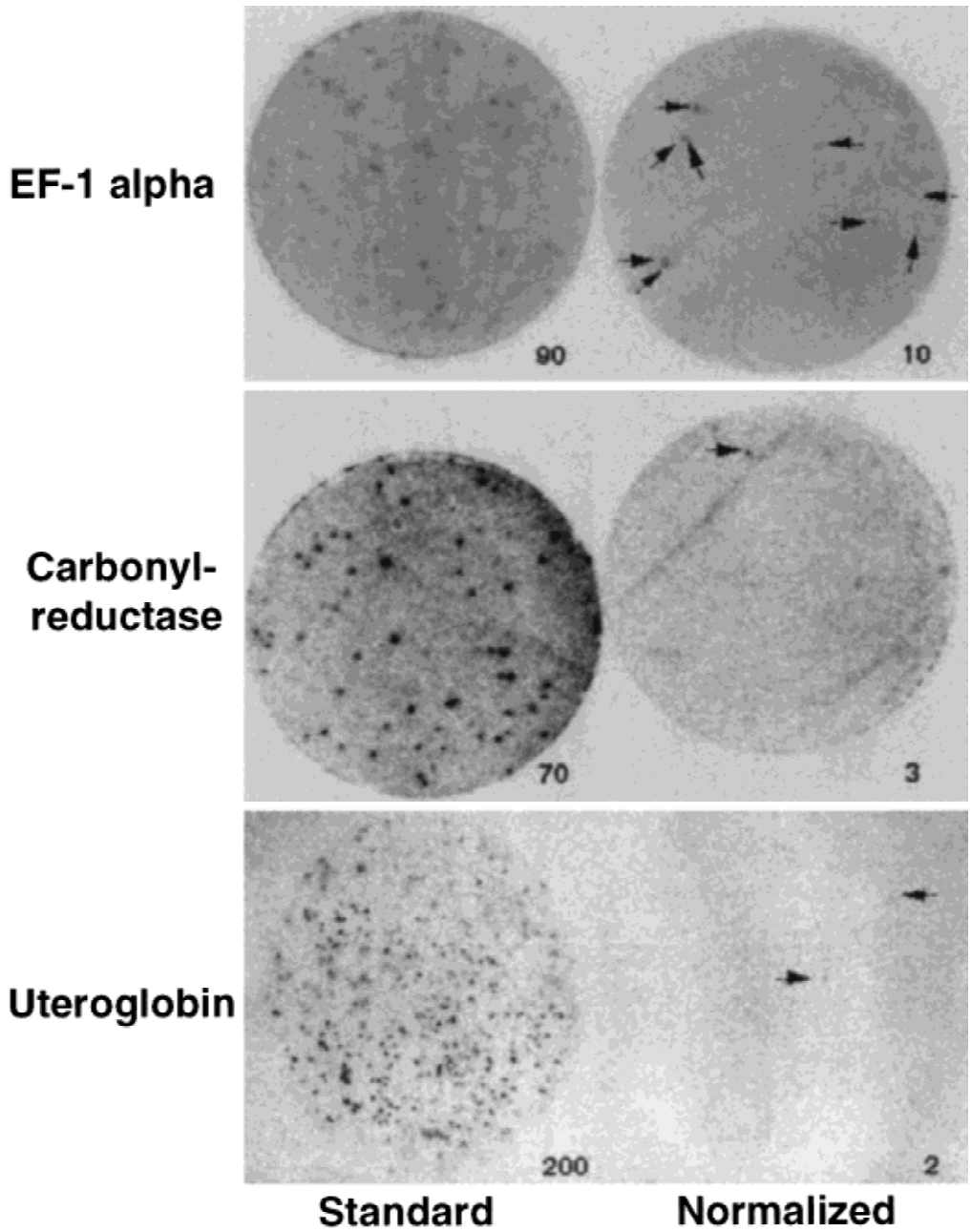

Figure 3 Plaque hybridization of replicas containing control lung cDNA library (left) or a normalized lung cDNA library (right). In the right panel (normalized), an arrow indicates the plaque we have counted.

given efficiently subtracted cDNA library, we would be able to clone cDNAs that are expressed once in 100,000–200,000 events with a 50% probability. For cDNAs that are expressed even more infrequently, other strategies such as the preparation of cDNA libraries from defined subregions of tissues are required (P. Carninci, in prep.). We expect that our proposed methodology will be useful in the collection of the remaining human full-length cDNAs as well as the generation of cDNA encyclopedias for other organisms. Our method might further be used for applications that would benefit from normalized, full-length cDNA libraries, such as expression cloning.

METHODS

Harvest of mRNA and all other preparatory steps were completed as described previously (Carninci and Hayashizaki 1999).

cDNA Synthesis

In a total volume of 24 µL, we combined 5–10 µg mRNA, 5 µg of the first-strand primer containing the *Bam*HI and *Sst*I re-

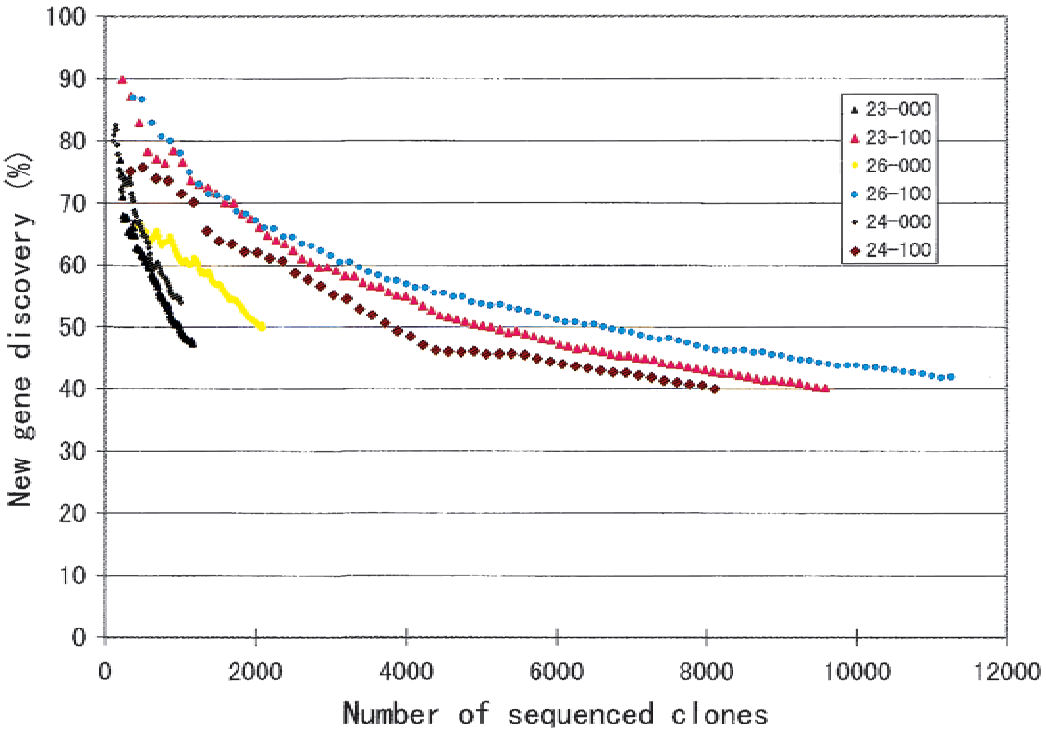

Figure 4 Sequencing redundancy (or the decrease in new gene discovery) increases sharply in standard cDNA libraries (-000 libraries), but in normalized/subtracted full-length cDNA libraries (-100 libraries), redundancy increases much more slowly. New genes (%) are referred as singleton (%) within a given cDNA library.

striction sites (5′-$(GA)_5$AAGGATCCAAGAGCTC$(T)_{16}$VN-3′), and 11.2 µL 80% glycerol. For liver and lung libraries and minilibraries, we instead used a primer containing an *Xho*I site (5′$(GA)_8$ACTCGAG$(T)_{16}$VN-3′), which generates inversely oriented cDNA libraries. The RNA-primer mixture was denatured at 65°C for 10 min. In parallel, we combined in a final volume of 76 µL; 18.2 µl 5× first-strand synthesis buffer; 9.1 µL 0.1 M DTT; 6.0 µL 10 mM (each) dTTP, dGTP, dATP, and 5-methyl-dCTP (instead of dCTP); 29.6 µL saturated trehalose (~80%, low metal content; Fluka Biochemika); and 10.0 µL Superscript II reverse transcriptase (200 U/µl). We placed 1.0 µl [α-^{32}P]dGTP in a third tube. The mRNA, glycerol, and primers were mixed on ice with the solution containing the Superscript, and an aliquot (20%) was quickly added to the tube containing the [α-^{32}P]dGTP. First-strand cDNA syntheses were performed in a thermocycler with a heated lid (e.g., MJ Research) according to the following program: step 1, 45°C for 2 min; step 2, gradient annealing: cool to 35°C over 1 min; step 3, complete annealing: 35°C for 2 min; step 4, 50°C for 5 min; step 5, 56°C for 60 min. Incorporation of radioactivity allowed us to estimate the yield of cDNA (Carninci and Hayashizaki 1999). The cDNA was treated with proteinase K, phenol/chloroform- and chloroform-extracted, and ethanol-precipitated by using ammonium acetate as the salt (Carninci and Hayashizaki 1999).

Table 2. Similar Presence of B1 Repeats in 200- and 300-Base Stretches in the 3-UTR of cDNAs from Subtracted/Normalized and Control Cap-Trapper cDNA Libraries

Library	Normalized-subtracted sublibrary	Control sublibrary
Embryo 18	1.3% (143/10970)	1.6% (4/244)
Stomach	1.3% (114/8840)	0.6% (10/1606)
Tongue	1.2% (131/10974)	0.7% (10/1408)
ES cells	1.4% (209/15220)	0.9% (10/1089)
Embryo 13-liver	0.7% (37/5521)	1.0% (7/718)
Embryo 10	1.8% (225/12724)	1.3% (30/2239)
Embryo 12-head	1.8% (160/8873)	1.4% (7/490)

The number of clones positive for B1 sequences versus the total number of sequences analyzed is indicated in parentheses.

cDNA Biotinylation

Before biotinylation, the diol group of the cap and 3′-end of RNA were oxidized in a final volume of 50 µL, containing the resuspended first-strand cDNA, 66 mM sodium acetate (pH 4.5), and 5 mM $NaIO_4$. Samples were incubated on ice in the dark for 45 min. cDNA was then precipitated by adding 0.5 µL of 10% SDS, 11 µL NaCl, and 61 µl of isopropanol. After incubation in the dark on ice for 45 min or at −20° or −80°C for 30 min, the sample was centrifuged for 10 min at 15,000 rpm. Finally we rinsed the cDNA twice with 70% ethanol and resuspended it in 50 µL of water. Subsequently, the cap was biotinylated in a final volume of 210 µL by adding 5 µL M sodium acetate (pH 6.1), 5 µL 10% SDS, and 150 µL of 10 mM biotin hydrazide long-arm (Vector Biosystem).

After overnight (10–16 hr) incubation at room temperature (22°–26°C), the cDNA was precipitated by adding 75 µL 1

M sodium acetate (pH 6.1), 5 µL 5 M NaCl, and 750 µL absolute ethanol and incubated on ice for 1 hr or at −20° to −80°C for 30 min. The cDNA was pelleted by centrifugation at 15,000 rpm for 10 min; we then washed the pellet once with 70% ethanol and once with 80% ethanol. We resuspended the cDNA in 70 µL 0.1× TE (1 mM Tris [pH 7.5], 0.1 mM EDTA).

Capture and Release of Full-Length cDNA

We combined 500 µL of MPG-streptavidin beads and 100 µg DNA-free tRNA and incubated the mixture on ice for 30 min with occasional mixing. The beads were separated by using a magnetic stand for 3 min, and the supernatant was removed. The beads then were washed three times with 500 µL washing/binding solution (2 M NaCl, 50 mM EDTA [pH 8.0]).

At the same time, we added 1 U of RNase I (Promega) per microgram of starting mRNA to the cDNA sample in the buffer supplied by the manufacturer (final volume, 200 µL); the sample was incubated at 37°C for 15 min. To stop the reaction, we put the sample on ice and added 100 µg tRNA and 100 µL of 5 M NaCl. To capture the full-length cDNA, we combined the biotinylated, RNase I-treated cDNA and the washed beads, which were resuspended in 400 µL of the washing/binding solution. After mixing, the tube was gently rotated for 30 min at room temperature. Full-length cDNA remained on the beads, and the shortened cDNAs did not. The beads were separated from the supernatant on a magnetic stirrer. We gently washed the beads to remove the nonspecifically adsorbed cDNAs: Two washes with washing/binding solution; one with 0.4% SDS, 50 µg/mL tRNA; one with 10 mM Tris-HCl (pH 7.5), 0.2 mM EDTA, 40 µg/mL tRNA, 10 mM NaCl, and 20% glycerol; and one with 50 µg/mL tRNA in water.

cDNA was released from the beads by adding 50 µL 50 mM NaOH, 5 mM EDTA and incubating for 10 min at room temperature with occasional mixing. The beads then were removed magnetically, and the eluted cDNA was transferred on ice to a tube containing 50 µL 1 M Tris-HCl, pH 7.5. The elution cycle was repeated once or twice with 50-µL aliquots of 50 mM NaOH, 5 mM EDTA until we recovered most of the cDNA (80%–90%, as measured by monitoring the radioactivity with a handheld monitor) from the beads.

To remove traces of RNA that could later interfere with the biotinylated RNA driver, we quickly added 100 µL of 1 M Tris-HCl, pH 7.0, and 1 µL RNase I (10U/µL) to the recovered cDNA on ice; the sample then was incubated at 37°C for 10 min. The cDNA was treated with proteinase K, phenol/chloroform-extracted, and back-extracted. We then added 2–3 µg glycogen and ethanol-precipitated the sample in a siliconized tube. Alternatively, the sample can be concentrated by using one round of ultrafiltration with a Microcon 100 (Millipore) for 40–60 min at 2000 rpm. If ethanol precipitated, the cDNA could be redissolved in 20 µl of 0.1 × TE.

CL-4B Spin-Column Fractionation of cDNA

We treated the cDNA samples with CL-4B chromatography (Carninci and Hayashizaki 1999) or an S-400 spin column (Amersham-Pharmacia) essentially as described by the manufacturer.

Oligo-dG Tailing of the First-Strand cDNA

We combined the cDNA sample, 5 µL of the 10× TdT buffer (2 M potassium cacodylate [pH 7.2], 10 mM $MgCl_2$, 10 mM 2-mercaptoethanol), 5 µL of 50 µ M dGTP, 5 µL of 10 mM $CoCl_2$, and 40 U terminal deoxynucleotidyl transferase in a final volume of 50 µL. Samples were incubated at 37°C for 30 min. At the end, reaction was stopped with EDTA 20 mM, cDNA digested with proteinase K, extracted with phenol chloroform, and ethanol precipitated. Sample was finally redissolved in TE. We checked the tail length as described (Carninci et al. 1999), after which the cDNA was used in a second-strand synthesis for use in check libraries (see later) or underwent normalization/subtraction.

Normalization Drivers

mRNA drivers comprising an aliquot of the starting mRNA were called "normalizing drivers." To calculate the concentration of the normalizing driver, we approximated the ribosomal/structural RNA contamination in the starting mRNA by assuming that the incorporation rate of the first-strand synthesis reflected the actual mRNA concentration, thus assuming 100% efficiency of priming and elongation. Assuming that the proportion of mRNA converted to first-strand cDNA corresponded to the effective mRNA concentration, we omitted accounting for less than full-length cDNAs—usually not all of the mRNA is primed. A slight excess of normalization driver was unlikely to interfere with the normalization process as dramatically as would a paucity of driver. Therefore, we assumed that the amount of mRNA in the sample was the same as the quantity of first-strand cDNA produced.

Subtraction Drivers

Subtracting drivers comprised bulk run-off transcripts prepared from cloned minilibraries and rearrayed libraries from the nonredundant RIKEN cDNA encyclopedia by using T7 and T3 RNA polymerases.

Minilibraries contain ~1000–2000 clones of cDNA deriving from a previous normalization experiment. By adapting the standard protocol, we prepared minilibraries from the captured aliquot (abundant cDNA fraction) that was the byproduct of normalization experiments. After normalization, the abundant cDNA fraction was removed from the beads with 50 mM NaOH/5 mM EDTA; after neutralization, second-strand cDNA was prepared. Cloning was accomplished in a way analogous to that previously described (Carninci and Hayashizaki 1999). Plasmid was then bulk-excised, and 1000–2000 clones per minilibrary were amplified on agarose/ampicillin. For driver preparation, we plated 20,000–50,000 colonies on SOB-agarose/ampicillin and incubated the plates overnight at 37°C. We scraped bacterial cells from the plate in the presence of resuspension solution (Wizard DNA extraction kit; Promega,) and later followed the manufacturer's protocol.

Preparation of the Nonredundant cDNA Library Driver

Single clones from the full-length cDNA encyclopedia (http://genome.rtc.riken.go.jp/) were rearrayed for the subtraction. From 384-well plates, rearrayed cDNAs were then plated on SOB-agarose/ampicillin plates. Plasmid extraction, DNA cleavage, and RNA preparation was performed as for minilibraries.

We digested the extracted plasmid at the 3′-end of the multiple cloning site by using *Pvu*I when the minilibrary was cloned with *Xho*I at the 3′-end site or *Sst*I when the library was cloned in the *Sst*I site. RNA was synthesized by using either T3

or T7 RNA polymerase (Life Technologies), depending on the map of the construct used to prepare the driver, to prepare sense run-off RNAs. We used T3 polymerase for *Pvu*I-cleaved minilibraries (up to 14), and T7 polymerase for *Sst*I-cleaved minilibraries (15 and following). RNA was prepared by using RNA polymerases (Life Technologies) according to manufacturer's instructions. Extensive digestion with 1–2 µL DNaseI (RQ1, RNase-free, Promega) was performed for 30 min. Proteinase K digestion was then performed, followed by extraction with phenol/chloroform and chloroform, and the cDNA was precipitated.

Biotin Labeling of Normalizing/Subtracting RNA Drivers

To further clean up RNA drivers before labeling, we used the RNeasy kit (QIAGEN) according to the instructions of the manufacturer. Subsequently, we used the Mirus nucleic acid biotinylation kit (Panvera) essentially as described by the manufacturer. For instance, 10 µg of the RNA mix was labeled by combining it with 10 µL of Label IT reagent and 10 µL of labeling buffer A, in a final volume of 100 µl. We incubated the reaction at 37°C for 1 hr, after which we precipitated the biotinylated RNA by adding 1/20-volume of 5M NaCl and two volumes of 99% ethanol. After standard ethanol precipitation, the pellet was washed once with 80% ethanol, resuspended in 20 µL of 1× Mirus labeling buffer A, and stored at −80°C until used. Alternatively, mRNA was labeled by using the psoralen-biotinylation kit (Ambion) according to the instructions of the manufacturer.

Normalization/Subtraction

The RNA drivers and cDNA were deproteinated by using proteinase K followed by phenol/chloroform extraction, chloroform extraction, and ethanol precipitation. Oligo-dG-tailed cDNA was used as a substrate, which was mixed with the RNA drivers and blocking oligonucleotides (biotin-dG_{16}) to hybridize to the C-stretch present in the subtracting driver and with oligo-dT primer to block the polyA sequences. Hybridization was carried out at RoT values of 5–500, depending on the experiment, in a buffer containing 80% formamide (from a deionized stock), 250 mM NaCl, 25 mM HEPES (pH 7.5), and 5 mM EDTA. Hybridization was carried out at 42°C in a dry oven; even volumes as small as 5 µL did not require mineral-oil overlays. After hybridization, we precipitated the sample by adding 2.5 volumes of absolute ethanol and incubated it for 30 min on ice. The sample was centrifuged for 10 min at 15,000 rpm and washed once with 70% ethanol; we carefully resuspended the cDNA in 10 µL of water on ice.

Removal of the Hybrid

In parallel, we prepared 50 µl CPG magnetic beads for each 1 µg of biotinylated driver RNA; 5 µL beads could bind >400 ng of biotinylated driver. To each 50 µL of beads, we added 10 µg tRNA as a blocking agent, then incubated the beads at room temperature for 10–20 min or on ice for 30–60 min with occasional shaking. We used a magnetic stand to remove the beads, which we washed three times with a large excess of 1 M NaCl, 10 mM EDTA and resuspended them in a volume of 1 M NaCl, 10 mM EDTA equivalent to the original volume of the bead suspension.

We mixed the blocked beads with the redissolved tester/driver mixture and incubated the entire sample at room temperature for 15 min with occasional gentle mixing. After removing the beads by using a magnetic stand for 3 min, we recovered the supernatant, which contained the single-strand normalized/subtracted cDNA. The beads were washed once with excess volume of binding buffer (1 M NaCl, 10 mM EDTA) to recover any remaining ssDNA. We measured the radioactivity of the labeled samples before and after the procedure in order to estimate the yield of normalization/subtraction.

To concentrate the cDNA solution to ~50 µL, we used Microcon 100 ultrafiltration as described by the manufacturer (Millipore). Subsequently, the cDNA was pelleted by using the standard isopropanol procedure; the pellet was resuspended in 44 µL of 0.1 × TE, to which 5 µL of RNase I buffer and 1 U RNase I were added, in a volume of 50 µL. Samples were then incubated for 20 min at 37°C, after which we added 400 µL of 0.2% SDS to inactivate the RNase I. Traces of degraded RNAs, blocking oligonucleotide, SDS, and buffer were removed by ultrafiltration with a Microcon 100 filter at 2000 rpm and 25°C until the volumes were reduced to <20 µL. The samples were desalted by adding 400 µl of 0.1 × TE then centrifuging as above for a total of three washes. We recovered the cDNA by inverting the filter in a new tube and centrifuging at 9000 rpm for 1 min.

Second-Strand cDNA Synthesis

For normalized/subtracted cDNA, the standard control libraries, and the minilibraries, the second-strand synthesis and cloning steps were the same. The *Xho*I-containing primer 5′-$(GA)_7$TTCTCGAGTTAATTAAATTAATC$_{13}$-3′ was prepared and purified by using standard techniques, as was the first-strand cDNA primer. For the lung and liver libraries and minilibraries, the *Sst*I-containing primer 5′-$(GA)_9$GAGCTCACT-AGTTTAATTAAATTAATC$_{11}$-3′ was used as the second-strand primer. To prepare the second-strand reaction, we mixed the oligo-dG-tailed cDNA with 6 µl of 100 ng/µL second-strand primer adapter, 6 µL of EX-Taq second-strand buffer (Takara), and 6 µL 2.5 mM (each) dNTPs. Hot-start priming then was performed by adding 3 µl of 5 U/µL ExTaq polymerase (Takara) at 65°C in a thermocycler. After mixing, the annealing temperature was reached by a negative ramp to 45°C for the *Xho*I primer and to 35°C for the *Sst*I primer. After 10 min at the annealing temperature, the second-strand cDNA was extended during incubation at 68°C for 20 min. The annealing-extension cycle was repeated once more, followed by a final elongation step at 72°C for 10 min. At the beginning of the hot-start, we mixed a 5 µL aliquot with 0.5 µL of [α^{32}P]dGTP or [α^{32}P]dCTP to follow the incorporation. We used the labeled aliquot at the end of reaction to visualize the cDNA and to calculate the second-strand yield (Carninci and Hayashizaki 1999).

cDNA Cloning

Second-strand cDNA was treated with proteinase K, extracted with phenol-chloroform and chloroform, and ethanol-precipitated according to standard procedures. We then cleaved the cDNA by using 25 U/µg each of *Sst*I and *Xho*I (lung and liver libraries and Lib.18–31) or *Bam*HI and *Xho*I (Lib.32–64). After the digestion, cDNA was treated with proteinase K, extracted with phenol-chloroform, and purified over a CL-4B spin column (Pharmacia). After ethanol precipitation, we cloned the cDNA essentially as described (Carninci and Hayashizaki 1999). The vector for cloning the cDNA Lib.32–64 will be described elsewhere (P. Carninci, in prep.).

Other Methods

Plaque hybridization was performed by using a random primer according to standard protocols (Sambrook et al. 1989). Alkali electrophoresis was performed as described (Sambrook et al. 1989). All autoradiography signals were visualized by using the Bas 2000 imaging system (Fuji).

Bacteria were picked with commercially available picking machines (Q-bot and Q-pix; Genetics, UK) and transferred to 384-microwell plates. Duplicate plates were used to prepare plasmid DNA. For plasmid DNA, 384-well plate were divided and grown in 4×96 deep well plates. After overnight growth, plasmids were extracted either manually (Itoh et al. 1997) or automatically (Itoh et al. 1999). Sequences typically were run on the RISA sequencing instrument (K. Shibata, in prep.); a few sequences were generated by using the Perkin Elmer-Applied Biosystems ABI 377. Sequencing primers were the M13 forward and reverse primers, and the main sequencing operation will be described in detail elsewhere (P. Carninci, in prep.).

Sequences for clustering were analyzed as follows. The poly-T (for 3′-end) and C-stretch (for 5′-end) regions were trimmed from the one-pass sequences. From the trimmed sequences, we selected 100-bp sequences to use as tag sequences. We used BLAST 2.0.9 to search for homology between the new tag and the database of nonredundant 100-bp tag sequences; sequences having BLAST parameters of $E = 1.0e^{-25}$ or lower were clustered together. When the tag sequence was not in the database, the tag was added to the database. When the database contained the tag, it was added to the member of identical group of the TAG. If the tag was found in the database and at the same time the shift was <10 bases, the overlap was >80 bases, with >90% identity in the overlap, the sequences were grouped together. In addition, the algorithm categorized sequences within the library as "new," "nonredundant," or "redundant," according to the previously defined criteria.

ACKNOWLEDGMENTS

We dedicate this work to Yuichi Sugahar, coauthor of this article, who died prematurely in an accident. We thank Claudio Schneider for multiple discussions and encouragement; Tomoko Hirozane, Toshiyuki Shiraki, and Kenjiro Sato for excellent technical contributions; and all members of the Genome Science Laboratory for collecting the data. This study has been supported by Special Coordination Funds and a Research Grant for the RIKEN Genome Exploration Research Project, CREST (Core Research for Evolutional Science and Technology), and ACT-JST (Research and Development for Applying Advanced Computational Science and Technology) of Japan Science and Technology Corporation (JST). Y.H. was funded by the Science Technology Agency in Japanese Government. This work was also supported by a Grant-in-Aid for Scientific Research on Priority Areas and Human Genome Program, from the Ministry of Education, Science and Culture, and by a Grant-in-Aid for a second Term Comprehensive 10-Year Strategy for Cancer Control from the Ministry of Health and Welfare to Y.H.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

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

Adams, M., Kerlavage, A., Fleischmann, R., Fuldner, R., Bult, C., Lee, N., Kirkness, E., Weinstock, K., Gocayne, J., White, O., et al. 1995. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides cDNA sequence. *Nature* **377:**(Suppl.) 3–174.

Anderson, M., and Young, B.D. 1985. Quantitative analysis of solution hybridisation. In *Nucleic Acids Hybridisation, a practical approach*, pp. 73–111. IRL Press, Oxford.

Barr, F. and Emanuel, B. 1990. Application of a subtraction hybridization technique involving photoactivatable biotin and organic extraction to solution hybridizationanalysis of genomic DNA. *Anal. Biochem.* **186:** 369–373.

Bonaldo, M., Lennon, G., and Soares, M. 1996. Normalization and subtraction: Two approaches to facilitate gene discovery. *Genome Res.* **6:** 791–806.

Carninci, P. and Hayashizaki, Y. 1999. High-efficiency full-length cDNA cloning. *Methods Enzymol.* **303:** 19–44.

Carninci, P., Nishiyama, Y., Westover, A., Itoh, M., Nagaoka, S., Sasaki, N., Okazaki, Y., Muramatsu, M., and Hayashizaki, Y. 1998. Thermostabilization and thermoactivation of thermolabile enzymes by trehalose and its application for the synthesis of full length cDNA. *Proc. Natl. Acad. Sci.* **95:** 520–524.

Carninci, P., Kvam, C., Kitamura, A., Ohsumi, T., Okazaki, Y., Itoh, M., Kamiya, M., Shibata, K., Sasaki, N., Izawa, M., et al. 1996. High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics* **37:** 327–336.

Diatchenko, L., Lau, Y., Campbell, A., Chenchik, A., Moqadam, F., Huang, B., Lukyanov, S., Lukyanov, K., Gurskaya, N., Sverdlov, E., et al. 1996. Suppression subtractive hybridization: A method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proc. Natl. Acad. Sci.* **93:** 6025–6030.

Edery, I., Chu, L., Sonenberg, N., and Pelletier, J. 1995. An efficient strategy to isolate full-length cDNAs based on an mRNA cap retention procedure (CAPture). *Mol. Cell Biol.* **15:** 3363–3371.

Fargnoli, J., Holbrook, N., Fornace Jr., A. 1990. Low-ratio hybridization subtraction. *Anal. Biochem.* **187:** 364–73.

Hillier, L., Lennon, G., Becker, M., Bonaldo, M., Chiapelli, B., Chissoe, S., Dietrich, N., DuBuque, T., Favello, A., Gish, W., et al. 1996. Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* **6:** 807–828.

Hirotsune, S., Takahara, T., Sasaki, N., Hirose, K., Yoshiki, A., Ohashi T., Kusakabe, M., Murakami, Y., Muramatsu, M., Watanabe, S., et al. 1995. The reeler gene encodes a protein with an EGF-like motif expressed by pioneer neurons. *Nat. Genet.* **10:** 77–83.

Kato, S., Sekine, S., Oh, S., Kim, N., Umezawa, Y., Abe, N., Yokoyama-Kobayashi, M., and Aoki, T. 1994. Construction of a human full-length cDNA bank. *Gene* **150:** 243–250.

Itoh, M., Carninci, P., Nagaoka, S., Sasaki, N., Okazaki, Y., Ohsumi, T., Muramatsu, M., Hayashizaki, Y. 1997. Simple and rapid preparation of plasmid template by a filtration method using microtiter filter plates. *Nucleic Acids Res.* **25:** 1315–1316.

Itoh, M., Kitsunai, T., Akiyama, J., Shibata, K., Izawa, M., Kawai, J., Tomaru, Y., Carninci, P., Shibata, Y., Ozawa, Y. et al. 1999. Automated filtration-based high-throughput plasmid preparation system. *Genome Res.* **9:** 463–470

Marra, M., Hillier, L., Kucaba, T., Allen, M., Barstead, R., Beck, C., Blistain, A., Bonaldo, M., Bowers, Y., Bowles, L., et al. 1999. An encyclopedia of mouse genes. *Nature Genet.* **21:** 191–194.

Maruyama, K. and Sugano, S. 1994. Oligo-capping: A simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene* **138:** 171–174.

Rubin, G., Hong, L., Brokstein, P., Evans-Holm, M., Frise, E., Stapleton, M., and Harvey, D. 2000. A *Drosophila* complementary cDNA resource. *Science* **287:** 2222–2224.

Sambrook, J., Fritsch, E., and Maniatis, T. 1989. *Molecular Cloning.*

Cold Spring Harbor laboratory press, Cold Spring Harbor, NY.
Sasaki, Y., Ayusawa, D., and Oishi, M. 1994. Construction of a normalized cDNA library by introduction of a semi-solid mRNA-cDNA hybridization system. *Nucleic Acids Res.* **22:** 987–992.
Sasaki, N., Nagaoka, S., Itoh, M, Izawa, M., Konno, H., Carninci, P., Yoshiki, A., Kusakabe, M., Moriuchi, T., Muramatsu, M., et al. 1998. Characterization of gene expression in mouse blastocyst using single-pass sequencing of 3995 clones. *Genomics* **49:** 167–79.
Soares, M., Bonaldo, M., Jelene, P., Su, L., Lawton, L., and Efstratiadis, A. 1994. Construction and characterization of a normalized cDNA library. *Proc. Natl. Acad. Sci.* **91:** 9228–9232.
Takahashi, N. and Ko, M. 1994. Toward a whole cDNA catalog: Construction of an equalized cDNA library from mouse embryos. *Genomics* **23:** 202–210.
Tanaka, T., Ogiwara, A., Uchiyama, I., Takagi, T., Yazaki, Y., and Nakamura, Y. 1996. Construction of a normalized directionally cloned cDNA library from adult heart and analysis of 3040 clones by partial sequencing. *Genomics* **35:** 231–235.

Received April 20, 2000; accepted in revised form July 24, 2000.