An 'equalized cDNA library' by the reassociation of short double-stranded cDNAs

Minoru S.H.Ko
Furusawa MorphoGene Project, Exploratory Research for Advanced Technology (ERATO), Research Development Corporation of Japan (JRDC), 5-9-6 Tohkohdai, Tsukuba 300-26, Japan

Received July 5, 1990; Revised and Accepted September 12, 1990

ABSTRACT

The total number of genes in higher organisms is estimated to be under one hundred thousand. However, constructing a cDNA library containing a full set of genes expressed throughout the life time of an organism, without redundancy, is a major challenge for modern biology. Towards this goal, I have tried to make a library of mouse fibroblastoid Ltk$^-$ cells with nearly equal representations of cDNA clones. Double-stranded cDNAs (ds-cDNAs) are synthesized from mRNA using an oligo(dT)-*NotI* primer. After shearing to 200 – 400 bp, a synthetic linker-primer, which has one blunt and one sticky end and an internal *EcoRI* site, is ligated to the cDNAs. The cDNAs are amplified by the polymerase chain reaction (PCR) using the ligated linker-primer sequence. After denaturation and reassociation of the ds-cDNAs, and isolation of single-stranded cDNAs (ss-cDNAs) by hydroxylapatite chromatography, the ss-cDNAs are again amplified by PCR. The cDNAs are digested with *EcoRI* and *NotI*, and inserted into a plasmid vector. Colony hybridization with eight probes of different abundance showed a reduction in 'abundance variation' from at least 20,000-fold in the original library to 40-fold in the library constructed after three cycles of equalization. This indicates the usefulness of the current procedure for making equalized cDNA libraries.

INTRODUCTION

Cells contain individual mRNA species at different abundance. For example, 16% of the mRNA in chicken liver is a single species and is thought to be a serum albumin (1). Therefore, ordinary cDNA libraries contain redundant copies of some species of cDNA as well as nearly unique copies of many other species. The result of this is a very low fractional representation of some cDNA species and makes it difficult to clone the cDNAs of these rare mRNAs. This is particularly a problem in expression cloning when clones must be screened individually, since it is common to have to check very large numbers of clones, many of these containing the same cDNA species (2). Furthermore, in the subtraction technique for cloning differentially expressed genes, the mass of very abundant cDNA species may impair the efficiency of subtraction. To overcome these problems, the equalization of the abundance of individual cDNAs in a cDNA library is one of the best solutions. More importantly, the technique should allow a novel use of cDNA libraries, i.e., as a catalogue of individual cDNA species for reference. Therefore, the achievement of an 'equalized cDNA library' is expected to result in great advances in the identification of tissue specific mRNA species and for estimating the expression levels of many interesting genes simultaneously, as well as for molecular cloning of rare mRNA species.

In the construction of an equalized cDNA library, it should be possible to take advantage of the high sequence specificity of nucleic acid hybridization. This property has been made use of previously for subtraction cloning (3–5), where 'tracer' cDNAs, synthesized on mRNA from one source, A, is hybridized to sequences of 'driver' mRNA, isolated from a different, but usually, related source, B. The tracer cDNAs which do not become hybridized with driver mRNA represent an enriched population of sequences expressed only in A cells and these are used for constructing an A-cell specific cDNA library. For subtraction cloning, the ratio of tracer and driver nucleic acids should be high, to deplete completely the mRNA species present in both A cells and B cells. In contrast, to *equalize* the abundance of individual cDNAs, the ratios of tracer and driver nucleic acids should be one to one (see Theoretical considerations in Results). Therefore, I chose to use a self-subtraction procedure involving the denaturation and reassociation of double-stranded cDNAs (ds-cDNAs), because it is an easy and accurate way to hybridize equal amounts of tracer and driver.

It was anticipated that this idea would require some modifications in order to avoid several likely problems. Firstly, cDNA clones having very similar coding regions, but derived from different genes could cross-hybridize with each other and be eliminated. Since it is known that the 3'untranslated region of mRNA is usually specific to individual transcripts, the fragmentation of cDNAs and the cloning of the most 3'region was expected to avoid this difficulty. Secondly, the single-stranded cDNA (ss-cDNA), collected after reassociation, has to be converted to ds-cDNA for cloning into a vector plasmid. For this purpose, a linker-primer was attached to both ends of the ds-cDNAs before the denaturation step and used for primer annealing sites. Thirdly, the removal of reassociated cDNAs will make the cloning efficiency low due to the loss of most cDNAs. In order to avoid this, the polymerase chain reaction (PCR) was used to amplify the cDNA.

By using these procedures, I have constructed an equalized cDNA library of mouse Ltk$^-$ cells. I describe the procedure in detail, show how such libraries are evaluated and discuss the applications of this library.

MATERIALS AND METHODS

Construction and *in vitro* transcription of neo and tk genes

A plasmid pSP64Aneo was constructed by inserting a 1.3 kb *Hin*dIII-*Ava*I fragment containing the entire protein coding region of pSV2neo (6) into the *Hin*dIII-*Ava*I site of the pSP64(polyA) plasmid (Promega, USA). A plasmid pSP64Atk was constructed by inserting a 1.8 kb *Bgl*II-*Pvu*II fragment containing the entire protein coding region of the Herpes simplex virus thymidine kinase gene into the *Bam*HI site of the pSP64(polyA) plasmid after *Bam*HI linker ligation. For *in vitro* transcription, both pSP64Aneo and pSP64Atk plasmids were digested with *Eco*RI and transcribed by SP6 RNA polymerase according to the supplier's guide (Riboprobe system, Promega, USA). Both transcripts had 30 bp of polyA sequence at their 3′ends.

Cell cultures, RNA extraction and poly(A)$^+$ RNA purification

Ltk$^-$ cells were cultured in medium supplemented with 10% foetal calf serum and harvested in the growth phase. Total RNA was isolated according to a standard method (2). Poly(A)$^+$ RNA was purified using a mRNA purification kit (Pharmacia, Sweden). Neo and tk transcripts were added to the poly(A)$^+$ mRNA sample at 10% and 0.0005% (weight/weight) respectively.

cDNA synthesis and shearing to 200–400 bp fragments

Five μg of mRNA was used for synthesis of cDNA using the oligo(dT)-*Not*I primer (5′-AATTCGCGGCCGCTTTTTTTTT-TTTTTT-3′, Promega, USA) according to an established method (7). Synthesized ds-cDNAs were sheared to fragments of 200–400 bp using a Branson Sonifier 250 for 20 sec, at strength 2 (Branson Ultrasonics, USA). After agarose gel electrophoresis, fragments of 200–400 bp were cut out, purified and end-polished with T4 DNA polymerase.

'Lone linker' ligation and PCR amplification

Blunt ended cDNAs were ligated, for 12 hr at 18°C, to a specially designed linker-primer, 'LL-RI' (8):

LL-RIA: 5′- pGAGATATTAGAATTCTACTC -3′
LL-RIB: 3′- TATAATCTTAAGATGAGp -5′.

Since the protruding end is not adhesive, a single molecule of this linker is attached to each end of the cDNA in an orientation-specific manner (thus called 'lone linker') (8). Amplification of the cDNA fragments was then performed by PCR (9) using a Thermal Cycler (Perkin-Elmer/Cetus) with the LL-RIA oligomers as a primer. One ng of DNA was included in 100 μl of the reaction mixture and amplified through 25 cycles (at 94°C for 2 min, at 50°C for 2 min, at 72°C for 2 min). After extraction with an equal volume of TE-saturated phenol-chloroform-isoamyl alcohol, the free primer was removed from the amplified cDNAs using a Centricon-100 microconcentrator (Amicon, USA).

Table 1. Numerical simulation for a comparison between pseudo-first order kinetics (A) and second order kinetics (B) in solution hybridization

(A) Tracer(D):Driver(R) = 1:100, $k = 1 \times 10^{-8}$
$S = De^{-kRt}$ (*)

	Concentration of unhybridized tracer (S)								Abundance variation[a]
Time (t)	A	B	C	D	E	F	G	H	(fold)
0	1000000	900000	1000	900	100	90	10	1	1000000
1×10^3	369233	367034	999	899	100	90	10	1	369233
5×10^3	0	0.1	19	25	67	63	10	1	–
1×10^4	0	0	0.1	0.3	41	40	9	1	–
5×10^4	0	0	0	0	1	1	6	1	–
1×10^5	0	0	0	0	0	0	4	1	–

(B) Tracer(D):Driver(R) = 1:1, $k = 1 \times 10^{-5}$
$S = D/(1 + kDt)$ (*)

	Concentration of unhybridized tracer (S)								Abundance variation
Time (t)	A	B	C	D	E	F	G	H	(fold)
0	1000000	900000	100000	90000	100	90	10	1	1000000
1×10^3	90909	90000	50000	47368	100	90	10	1	90909
2×10^3	100	100	100	100	50	47	9	1	100
1×10^4	11	11	11	11	10	9	5	0.9	12
4×10^4	2.6	2.6	2.6	2.6	2.5	2.5	2.0	0.7	3.7
1×10^5	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.5	2

This simulation shows the kinetic behaviours of eight species of tracer (A-H) with various initial abundance. In a typical example of pseudo-first order kinetics (A), excess mRNA ('driver') is hybridized with its complementary ss-cDNA ('tracer'). In a typical example of second order kinetics (B), after the denaturation of ds-DNA, each complementary strand ('tracer' and 'driver') is hybridized with each other. (*) Both equations are well established (13–15).

The symbols used here are as follows:

D: Initial concentration of tracer nucleic acids (arbitrary unit)
R: Initial concentration of driver nucleic acids (arbitrary unit)
k: Reaction rate constant (arbitrary)
t: Time after initiation of reaction (arbitrary unit)
S: Concentration of unhybridized tracer at time t (arbitrary unit)

[a] The term 'abundance variation' is defined here as the index of the equality of the constituents in the system. This is calculated by dividing the abundance of the most abundant constituent by that of the least abundant constituent in the system. In a cDNA library, the calculation is arbitrarily made by using the available abundance data in that system, since the abundance of all the cDNA species is usually unknown. In a perfectly equalized cDNA library, the abundance variation should be 1.

Equalization step

The denaturation and reassociation of cDNA fragments in solution was principally performed according to an established method (5). Three successive equalization cycles were carried out. For the first (EI) and second (EII) libraries, 20 μg of amplified cDNA was dissolved in 10 μl of distilled water in a 1.5 ml Eppendorf tube and combined with 10 μl of 2×hybridization solution (0.24 M NaH_2PO_4 [pH 6.8], 1.64 M NaCl, 2 mM EDTA, 0.2% SDS). For the EIII library, 100 μg of amplified cDNA was used in the same volume. In each case the sample was overlaid with light mineral oil and boiled for 5 min. Reassociation was then performed at 65°C for 12 hr (for EI and EII) or for 24 hr (for EIII). Separation of ss-cDNA from ds-cDNA was performed by a standard method (10) in a water-jacketed column maintained at 60°C, containing hydroxylapatite (HPT grade, Bio-Rad, USA). The single strand fraction was desalted by Centricon-100 and was then amplified and converted into ds-cDNAs by PCR (see above).

*Not*I/*Eco*RI digestion, vector ligation and transformation

The ds-cDNAs were double-digested with *Eco*RI and *Not*I and ligated overnight to the plasmid vector pBluescript SK(−) (Stratagene, USA) digested with *Eco*RI and *Not*I. MAX Efficiency DH5α competent cells (BRL, USA) were transformed by the ligated products.

Colony Hybridization

Colonies were grown directly on the surface of 20 cm×20 cm GeneScreenPlus nylon membrane filters (NEN, USA). An isotype specific probe for mouse cytoskeletal β-actin gene (11) was a kind gift from Dr. K. Tokunaga. A 0.8 kb 3′untranslated region of this gene was cut out from the plasmid pSPMβA-3′ut and used as a probe. For neo and tk gene probes, sequences identical to the *in vitro* transcripts were used. These three fragments were radiolabelled using a random priming method (12). The other probes were radiolabelled by Klenow filling methods (10) using oligo(dT) (Promega, USA) as a primer. Colony hybridization was performed by a standard method (10). After washing filters with 1% SDS and 0.1×SSC at 65°C for 1 hr, the filters were exposed on Imaging Plates (Fuji Film, Japan) overnight. Recorded images were scanned with a BAS2000 Imaging Plate Scanner (Fuji Film, Japan) and printed out by Pictrography (Fuji Film, Japan). The number of positive colonies was counted manually.

Sequence analysis

DNA sequencing of cDNA inserts was performed using Sequenase Version 2.0 (U.S. Biochemical, USA) and double-stranded miniprep DNA templates. Sequence analysis was performed by the computer program, GENETYX (Software Development Co.,LTD., Japan). Data base searches were performed using EMBL-GDB (release 20.0, August, 1989) supplied as GENETYX-CD.

RESULTS

Theoretical considerations

In order to establish the optimal hybridization conditions for equalization, computer simulations were set up using well-established formulae describing hybridization kinetics in solution (13−15). Table 1 illustrates how the concentration of unhybridized tracers with various levels of initial abundance decrease with time. When the driver concentration is in large excess over the tracer concentration (Table 1A), the hybridization obeys pseudo-first order kinetics. Since depletion of the highly abundant species of tracer occurs rapidly, the abundances of individual species of tracers are not equalized at any time. By contrast, when the concentrations of driver sequences are equal to those of the tracer (Table 1B), the hybridization obeys second order kinetics and the abundances of individual species of tracers become equalized with time. These simulation results indicate that the ratio between tracer and driver should be one to one for the equalization of a cDNA library. This condition can be met by using the complementary strands of ds-cDNA as tracers and drivers, since the ratio of these is exactly one to one. It is worth noting that in this self-subtraction procedure, the specific loss of any species of cDNA, regardless of its abundance, does not occur at any C_0t value (Table 1B).

The simulation result showing second order kinetics (Table 1B) indicates the difficulty of attaining perfect equalization of any library. While a ten-fold increase in C_0t value can cause a 7500-fold reduction in 'abundance variation' from 90909 to 12, a further ten-fold increase in C_0t value can only cause another 6-fold reduction in 'abundance variation' from 12 to 2 (see Table 1 legend for definition of 'abundance variation'). Since a C_0t value sufficiently high for the equalization to become completed,

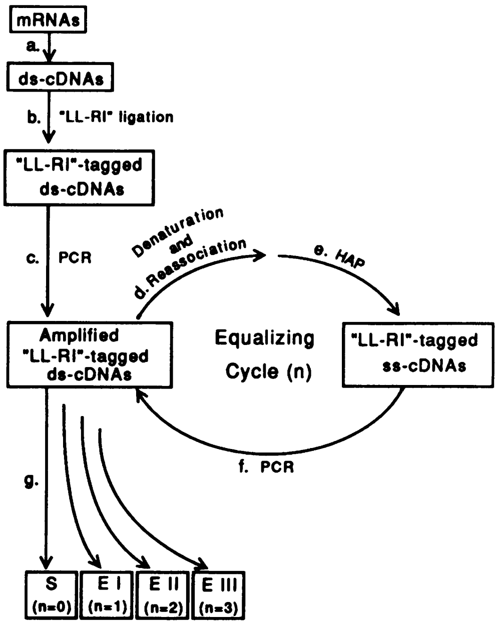

Figure 1. Flow diagram for cDNA equalization. Double-stranded cDNAs (ds-cDNAs) are synthesized using an oligo(dT)-*Not*I primer, sheared to 200−400 bp fragment size and end-polished with T4 DNA polymerase (step a). The shortened ds-cDNAs are ligated to a 'LL-RI' lone linker which has an internal *Eco*RI site for cloning (step b). This linker, attached to both ends of the ds-cDNAs, is used as a 'LL-RIA' primer annealing site for PCR (step c). The amplified 'LL-RI'-tagged ds-cDNAs are heat-denatured and reassociated to an appropriate C_0t value (step d). The unhybridized, single-stranded cDNAs (ss-cDNAs) are separated from the hybridized ds-cDNAs by hydroxylapatite (HAP) chromatography (step e). The collected ss-cDNAs are amplified and also converted to ds-cDNAs by PCR using the 'LL-RIA' primer (step f). After every 'equalizing cycle' (steps d-f), the form of the ds-cDNAs is unchanged, but the abundance of individual species of ds-cDNA becomes equalized. The amplified ds-cDNAs are digested with *Not*I and *Eco*RI, and cloned into *Not*I-*Eco*RI double digested plasmid vectors (step g). The straight library (S) is constructed without going through an 'equalizing cycle'. An Equalized library I (EI) is constructed after undergoing one 'equalizing cycle', EII, two cycles and EIII, three cycles.

is difficult to obtain for practical reasons, i.e., the limits of dissolved DNA concentration and of incubation time, procedures for carrying out repeated equalization hybridization steps and for augmenting the overall C_0t value were devised (see below).

Experimental strategy

Certain cDNAs such as members of multigene families contain very homologous protein coding regions. Although the equalization procedure using the denaturation and reassociation of full length ds-cDNAs may subtract these gene family members mutually, the fragmentation of cDNAs and cloning of unique regions will overcome this potential problem. Since it is known that the 3′untranslated region of mRNA is usually unique, as in the actin gene family (16), I chose a strategy of shearing ds-cDNAs to 200−400 bp and cloning the most 3′fragments of cDNA for library construction. By using oligo(dT)-*Not*I (see Materials and Methods) as a primer for first strand synthesis, a *Not*I site was formed after the second strand synthesis at the 3′side of the cDNA polyA sequence. The resulting ds-cDNA fragments were tagged with a specially designed *Eco*RI linker 'LL-RI' (see Materials and Methods) for cloning cDNA into vectors. This 'lone linker' (8) has both a non-palindromic protruding end and a blunt end and it has an internal *Eco*RI site, which is used both as the *Eco*RI cloning site and the primer annealing site for converting ss-cDNAs into ds-cDNAs after hydroxylapatite chromatography. Since the vector was digested with *Eco*RI and *Not*I, the most 3′fragments of cDNA which have both the *Eco*RI and *Not*I sites can be cloned selectively.

Since each cycle of the equalization procedure naturally results in a diminution in the amount of cDNA, repeated equalizations are difficult to carry out. In order to compensate for this decrease, the collected unhybridized cDNAs were amplified by PCR using the LL-RI sequence attached to both ends of cDNA as a primer annealing site. The use of the 'lone linker' for PCR is shown to allow the precise amplification of any complex mixture of DNA fragments such as mouse whole genome DNA (8) and thus the possibility of mosaic clones in the PCR steps is expected to be low. The fragmentation of cDNA inserts is also useful at this stage for avoiding any difference in amplification efficiency with fragment size. The overall amplification efficiency of DNA fragments under 1000 bp is known to be close to 100% due to high efficiency denaturation and priming at each cycle (17).

Fig. 1 is a flow diagram for the construction of an equalized cDNA library. In this procedure, it is possible to perform any number of 'equalizing cycles' until equalization is complete.

Construction of cDNA libraries and their evaluation

Firstly, a straight cDNA library ('S library') was constructed, in which cDNA inserts were expected to be 3′-specific short DNA fragments and to be present at the same abundance as in an ordinary cDNA library. As external controls, *in vitro* transcripts of neo and tk genes (see Materials and Methods) were added

Table 2. Evaluation for equalization of libraries

(A) Equalizing cycle (n)	1	2	3
C_0t value (mol.litre^{-1}.sec.)	130	130	1100
Recovered ss-cDNAs /Input ds-cDNAs (%)	16	38	37

(B)	No.positive/No.examined (%)			
Probe	S	EI	EII	EIII
neo[a] (10)	208/12810 (1.6)	19/12000 (0.16)	2/31720 (0.0063)	8/17200 (0.047)
tk[b] (0.0005)	0/45670 (<0.002)	0/8000 (<0.013)	19/50520 (0.038)	2/17200 (0.012)
mouse β-actin[c]	117/21350 (0.55)	69/12000 (0.58)	19/31720 (0.06)	27/17200 (0.16)
eIF-4A[d]	6/6900 (0.087)	N.D.[l]	1/9400 (0.011)	4/17200 (0.023)
vimentin[e]	44/6900 (0.64)	N.D.	20/9400 (0.21)	26/17200 (0.15)
IAP[f]	106/6900 (1.5)	N.D.	34/9400 (0.36)	85/17200 (0.49)
EF-1α[g]	126/6900 (1.8)	N.D.	58/9400 (0.62)	29/17200 (0.17)
ATPase 6[h]	221/6900 (3.2)	N.D.	23/9400 (0.24)	33/17200 (0.19)
b1 repeat[i]	953/6900 (14)	N.D.	222/9400 (2.4)	363/17200 (2.1)
b2 repeat[j]	70/6900 (1.0)	N.D.	59/9400 (0.63)	372/17200 (2.2)
Abundance variation[k]	>7000-fold (>1600-fold)	– (–)	380-fold (100-fold)	180-fold (40-fold)

[a] See Materials and Methods, 10% of total mRNA.

[b] See Materials and Methods, 0.0005% of total mRNA.

[c] An isotype specific 3′region of mouse cytoskeletal β-actin transcripts (11).

[d] A 600 bp insert excised from the S-049 clone (S library) which was found to be the 3′end of a mouse mRNA for the translation initiation factor eIF-4A short form (23) by EMBL data base search.

[e] A 400 bp insert excised from the S-379 clone which was found to be almost identical to the 3′end of hamster vimentin gene transcripts (24) by EMBL data base search.

[f] A 400 bp insert excised from the S-003 clone which was found to have strong homology to a mouse intracisternal-A particle gene (25−27) by EMBL data base search.

[g] A 400 bp insert excised from the S-165 clone which was found to be the 3′end of a mouse mRNA for translation elongation factor 1 alpha (28) by EMBL data base search.

[h] A 400 bp insert excised from the S-214 clone which was found to be the 3′end of ATPase 6 gene transcripts from the mouse mitochondria genome (29) by EMBL data base search.

[i] A 280 bp insert excised from the EII-012 clone (EII library) which was found to have homology to the mouse b1 repetitive consensus sequence (19) by EMBL data base search.

[j] A 230 bp insert excised from the EII-169 clone which was found to have strong homology to the mouse b2 repetitive consensus sequence (20) by EMBL data base search.

[k] Abundance variation in parenthesis was calculated without b1 and b2 repeat data.

[l] N.D.; not determined.

to the mRNA sample at 10% and 0.0005% of poly(A)$^+$ mRNA respectively in order to obtain a 20,000-fold difference in their abundance. The starting material was 260 ng of ds-cDNAs. The total number of independent bacterial colonies obtained was about 280,000. Examining the inserts of randomly picked colonies showed that 27% of those colonies contained the insert. Thus, the real size of this straight library was estimated to be about 76,000. The average inserts size was 300 bp.

A portion of amplified ds-cDNAs was processed through one cycle of the equalization procedure (Fig. 1). The total colony number of this equalized cDNA library I ('EI library') was about 49,000 and the real size of the library was estimated (as described above) to be about 16,000. The average insert length was 250 bp. The extent of equalization was checked by colony hybridization using three probes; neo, tk and mouse β-actin (Table 2B) and examples of the colony hybridization data are shown in Fig. 2. Since a reduction in abundance of mouse β-actin cDNA was not observed, I thought that the extent of equalization in the EI library was not sufficient and thus, tried to construct an 'EII library'.

For constructing an EII library, a portion of the amplified ds-cDNAs used for the EI library was processed through a further cycle of the equalization procedure (Fig. 1). The total colony number of this EII library was about 230,000 and the real number of colonies was estimated to be about 120,000. Since the total number of expressed genes in mouse L cells is estimated to be about 8000 by a C_0t analysis (18), the abundance of individual cDNA clones in a completely equalized cDNA library is calculated as about 0.013%. Colony hybridization with the three probes neo, tk and β-actin showed satisfactory results (Table 2B). In particular, the difference in abundance of neo and tk was significantly reduced, from approximately 20,000-fold to 6-fold.

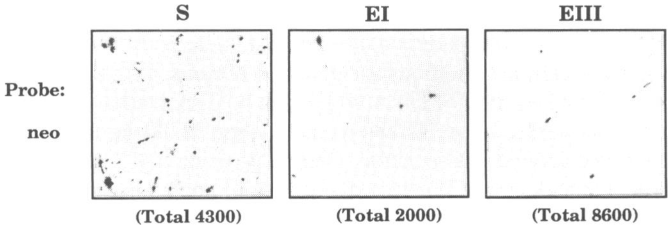

Figure 2. Examples of colony hybridization data. Each filter was hybridized with the ^{32}P-labelled neo probe (See Materials and Methods) and autoradiogrammed using a Fuji Bioimage analyzer. Representative autoradiographs from the 'S library' (S), 'EI library' (EI) and 'EIII library' (EIII) are shown. The estimated number of colonies in each filter is also shown at the bottom of each autoradiograph.

To further characterize the EII library, cDNA clones were selected randomly from both the S library and the EII library, and sequence analyses were carried out. The results showed that 74% of clones in the S library and 78% of clones in the EII library had both a polyA sequence and a *Not*I site at their 3′end, and an *Eco*RI site at their 5′end. Therefore, I assumed that the inserts of most clones represented the 3′end of a cDNA. The sequence analyses of these clones using the EMBL sequence data base indicates that the reduction in the number of redundant clones did not result from the abnormal increment of any specific sequences (Table 3). Although a completely equalized cDNA library should only contain unique species of cDNAs, the EII library still contained redundant clones. In order to check the abundance of identified clones, filter hybridizations were carried out using the cDNA inserts as probes. The abundance of these clones in each cDNA library is shown in Table 2B.

One further equalization step was carried out using essentially the same procedure as before, resulting in an 'EIII library' (Fig. 1). The total colony number of this EIII library was about 280,000 and the real size of this library was estimated to be about 110,000. Colony hybridization with the available probes showed a slight improvement in the extent of equalization (Table 2B).

Unexpectedly, the abundance of b1 and b2 repeats, which are known as a major group of mouse interspersed repetitive elements (19, 20), remained high relative to other genes, though a reduction of abundance was observed compared to the S library. Comparison of the sequences of these repetitive clones shows that they have medium to high homology with each other, but are not identical (Table 4). Therefore, the current hybridization

Table 3. The results of data base searching

S library	No.identified clones		
Name	Almost identical	Homologous	Total No.
b2 repeat family[a]	4	1	5
mitochondrial genome[b]	2	2	4
IAP[c]		2	2
ribosomal protein L18a[d]	2		2
EF-1α[e]	2		2
hamster NF1[f]		1	1
eIF-4a[g]	2		2
ferritin light chain[h]	1		1
ribosomal protein L27[i]	1		1
Y-box binding protein[j]		1	1
vimentin[k]	1		1
L1Md-A2 repeat family[l]		1	1
unknown1	2		2
unknown2	2	1	3
unknwon3		2	2
unknown, others			36
Total			66

EII library	No.identified clones		
Name	Almost Identical	Homologous	Total No.
b1 repeat family[m]		5	5
b2 repeat family	3	4	7
mitochondrial genome[n]		1	1
IAP		1	1
vimentin	2		2
L1Md-A2 repeat family		1	1
unknown4	2		2
unknown5	2		2
unknown6		2	2
unknown, others			51
Total			74

The sequences of 89 clones from the 'S library' and those of 95 clones from the 'EII library' were determined. A number of clones which had a polyA and a *Not*I site at their 3′end, and an *Eco*RI site at their 5′end were picked up and searched for homology with DNA sequences deposited in the EMBL sequence data base. The homology score (%) was calculated using GENETYX software based on a standard algorithm (37). The clones were classified into three groups; almost identical (>90%), homologous (70–90%), unknown (<70%). Unknown 1–6 were not identified as known sequences in the data base but forming several distinct groups. [a] Mouse b2 repeats (20). [b] Each sequence corresponded to a different region of the mouse mitochondrial genome (29). [c] Mouse intracisternal-A particle sequence (25–27). [d] Rat mRNA for ribosomal protein L18a (30). [e] Mouse elongation factor 1 alpha for translation (28). [f] Hamster nuclear factor 1-like protein (NF1) mRNA (31). [g] Mouse mRNA for protein synthesis initiation factor eIF-4A (23). [h] Rat ferritin light chain gene (32). [i] Rat mRNA for ribosomal protein L27 (33). [j] Human Y box binding protein-1 mRNA (34). [k] Hamster vimentin gene (24). [l] Mouse L1Md-A2 repeats (35). [m] Mouse b1 repeats (19). [n] Rat mitochondria ATP synthase β subunit (36).

Table 4. Sequence homology matrix of clones containing the b1 or b2 repeats

										Homology score (%)								
	Size						EII									S		
Clone	(bp)	012	030	034	051	065	103	117	152	165	169	226	232	094	157	262	321	414
EII-012	280																	
EII-030	230	60																
EII-034	134	52	–															
EII-051	194	71	68	55														
EII-065	242	73	63	–	73													
EII-103	170	50	53	84	52	50												
EII-117	86	–	64	83	64	–	75											
EII-152	139	51	54	75	52	51	78	70										
EII-165	159	79	69	58	62	70	50	51	54									
EII-169	230	–	51	84	59	55	80	90	75	54								
EII-226	70	68	54	87	68	63	73	76	71	61	75							
EII-232	114	–	–	92	–	–	87	91	79	–	90	89						
S-094	59	82	88	97	77	71	88	78	73	63	97	86	59					
S-157	182	50	–	91	55	–	86	84	80	50	90	84	95	97				
S-262	179	–	–	90	51	–	85	83	77	–	85	84	95	90	92			
S-321	177	–	–	93	–	–	84	86	77	–	87	52	93	100	91	93		
S-414	200	50	51	87	50	58	82	78	75	–	79	85	88	86	87	89	89	
Consensus																		
b1	130	85	72	–	75	81	–	–	–	89	–	–	51	–	–	–	–	–
b2	190	–	–	94	–	–	86	89	80	–	91	83	95	100	93	95	95	89

The homology score (%) was calculated as described in Table 3, legend. The consensus sequences of the b1 repeat or b2 repeat were derived from (38). The mark, (–) means that the homology score was under 50%.

conditions may be sufficiently stringent to fail to subtract these clones with subtle sequence differences.

DISCUSSION

I have reported here an attempt to construct an equalized cDNA library by denaturation and reassociation of shortened ds-cDNAs. In order to test the effectiveness of this procedure, four cDNA libraries denoted as S, EI, EII and EIII were constructed. The S (straight) library was constructed without any equalizing procedure and represented an ordinary cDNA library, though the size of the inserts was short. Successive applications of the equalizing procedure produced EI (equalized I), EII and EIII libraries. For evaluating the extent of equalization in each cDNA library, I proposed two analysis methods. One was based on colony hybridization data to estimate the abundance of specific cDNAs. The other was a direct sequence analysis of a limited number of randomly picked clones to investigate the redundancy of the constituents. When the redundancy of the sample is estimated, the redundancy of the population will be inferred. In general, for statistical inference, the sample size should be large and the sampling should be random. In the current sampling of clones in each library, these criteria did not seem to be met. For example, the sequencing analysis (Table 3) showed that the abundance of a vimentin cDNA was 1.5% in the S library and 2.7% in the EII library, while the colony hybridization analysis (Table 2B) showed it to be present in 0.64% of S library colonies and 0.21% of EII library colonies. Although the sequence analysis ensures that a reduction in abundance of the clone tested does not result from the amplification of specific clones, I conclude that colony hybridization data with many probes is more useful for estimating the efficiency of equalization for each library, unless more clones can be sequenced.

Colony hybridization data using ten different probes showed that there was a reduction in 'abundance variation' (for a definition of this term, see the legend of Table 1) from more than 7000-fold in the original S library to 180-fold in the three cycle-equalized EIII library (Table 2B). Since in a perfectly equalized cDNA library the abundance variation should be about 1, 180-fold in the EIII library was disappointing. Some of this redundancy was due to the high proportion (2% even in the EIII library) of colonies hybridizing with b1 and b2 repetitive sequences. However, the sequencing analysis showed that not all the cDNAs which belong to the b1 or b2 repetitive family were identical and that the homology score with each other was not so high (Table 4). More importantly, most clones possessed totally different flanking sequences outside the b1 or b2 repeat consensus sequence (data not shown). This implies that most of the clones cross-hybridized with b1 or b2 repeats by filter hybridization are derived from different transcripts. This shows that the stringent conditions for solution hybridization seem to be able to differentiate between subtly different cDNA sequences and suggests that equalized libraries may be truly representative. When the contribution of the b1 and b2 repeat family is omitted, the abundance variations are more than 1600-fold in the S library, 100-fold in the EII library and 40-fold in the EIII library (Table 2B). While the C_0t value of the EII-EIII step is 8.5 times higher than that of the S-EI or EI-EII steps (Table 2A), the improvement in abundance variation was just 2.5 times in the EII-EIII step (Table 2B). Further augmentation of C_0t value is difficult using my reassociation conditions for practical reasons, i.e., the limits on the concentration of DNA which can be dissolved and on the time for reassociation. Greater equalization could be achieved under accelerated reassociation conditions, using e.g., the phenol emulsion technique (21). However, 40-fold abundance variation is a very substantial improvement over approximately 20,000-fold variation, though it is difficult to prove that the current equalized cDNA library contains all the transcripts in the mRNA sample. In this respect, it is worth noting that the external control tk mRNA was detected in the EII and EIII libraries, while it was not detected in the S library even after screening about 45,000 colonies. This indicates that very rare species of mRNA, present at under 1 copy per cell, can be retained in the procedure and increased in relative abundance. This is presumably due to the highly efficient PCR amplification of very rare cDNA species and the enrichment of remaining unhybridized ss-cDNAs after hydroxylapatite chromatography.

The shortening of cDNA inserts has both advantages and disadvantages for library equalization. It makes it unlikely that clones having very similar coding sequences will be eliminated and ensures an unbiased amplification by PCR, and thus, allows multiple equalization cycles and efficient conversion from ss-cDNAs to ds-cDNAs after hydroxylapatite chromatography. However, the use of only 3′sequences also limits some applications of this library, since most of the coding region of individual cDNAs was excluded. The current equalized cDNA library is not directly applicable to the expression cloning technique or for obtaining information about the encoded proteins. However, once an equalized cDNA library is constructed, it should be possible to make an 'equalized and full-length' cDNA library using the equalized mixtures of cDNAs as primers for full-length cDNA synthesis.

The equalized cDNA library in its current state already has some important potential applications. Firstly, in the subtraction technique for cloning differentially expressed genes, the equalization of each cDNA sample before subtraction should improve the chance of cloning very rare mRNA species. Secondly, the equalized mixture of cDNAs could be used as a mixed probe to detect the transcribed regions of genomic sequences or cosmid clones. It is known that when an ordinary cDNA mixture is used as a probe, very rare transcripts cannot be detected (22). In such a 'Reverse Northern' approach, the nearly equal representation of each cDNA in the probe would be expected to avoid this problem. Thirdly, I expect the current procedure to allow the production of a catalogue of a full set of the genes expressed throughout the life time of any organism. This catalogue could be used to estimate simultaneously the expression levels of most genes using radiolabelled cDNA copies of mRNA at the original abundance as mixed probes and make it possible to identify tissue-specifically expressed genes. It is worth noting that it is more appropriate to use the current short cDNA fragments than full length cDNA species in these applications, since the high sequence specificity of the 3′fragments should eliminate cross-hybridization of different cDNA species.

ACKNOWLEDGEMENTS

I am most grateful to Project director, Dr. Mitsuru Furusawa for discussion and continuous support, and Ms. Naomi Takahashi and Mr. Shigeru B.H.Ko for discussion and technical assistance. I am also grateful to our colleagues for discussion and reading the manuscript, especially to Dr. Kuniya Abe for suggesting the usefulness of the modified PCR method and to Dr. Susan Godsave for critical reading of the manuscript.

REFERENCES

1. Axel,R., Feigelson,P. and Schutz,G. (1976) *Cell*, **7**, 247–254.
2. Okayama,H., Kawaichi,M., Brownstein,M., Lee,F., Yokota,T. and Arai,K. (1987) *Methods Enzymol.*, **154**, 3–28.
3. Sargent,T.D. and Dawid,I.B. (1983) *Science*, **222**, 135–139.
4. Hedrick,S.M., Cohen,D.I., Nielsen,E.A. and Davis,M.M. (1984) *Nature*, **308**, 149–153.
5. Sargent,T.D. (1987) *Methods Enzymol.*, **152**, 423–432.
6. Southern,P.J. and Berg,P. (1982) *J. Mol. Appl. Genet.*, **1**, 327–341.
7. Gubler,U. and Hoffman,J.B. (1983) *Gene*, **25**, 263–269.
8. Ko,M.S.H., Ko,S.B.H., Takahashi,N., Nishiguchi,K. and Abe,K. (1990) *Nucleic Acids Res.*, **18**, 4293–4294.
9. Saiki,R.K., Scharf,S., Faloona,F., Mullis,K.B., Horn,G.T., Erlich,H.A. and Arnheim,N. (1985) *Science*, **230**, 1350–1354.
10. Sambrook,J., Fritsch,E.F. and Maniatis,T. (1989) Molecular Cloning:A Laboratory Manual (second edition). Cold Spring Harbor University Press, Cold Spring Harbor.
11. Tokunaga,K., Taniguchi,H., Yoda,K., Shimizu,M. and Sakiyama,S. (1986) *Nucleic Acids Res.*, **14**, 2829.
12. Feinberg,A.P. and Vogelstein,B. (1983) *Anal. Biochem.*, **132**, 6–13.
13. Britten,R.J., Graham,D.E. and Neufeld,B.R. (1974) *Methods Enzymol.*, **XXIX**, 363–418.
14. Britten,R.J. and Davidson,E.H. (1985) Hybridization strategy. In Hames,B.D. and Higgins,S.J. (ed.), Nucleic Acid Hybridization-A Practical Approach. IRL Press, Oxford, pp. 3–15.
15. Young,B.D. and Anderson,M.L.M. (1985) Quantitative analysis of solution hybridization. In Hames,B.D. and Higgins,S.J. (ed.), Nucleic Acid Hybridization-A Practical Approach. IRL Press, Oxford, pp. 47–71.
16. Mohun,T.J., Brennan,S., Dathan,N., Fairman,S. and Gurdon,J.B. (1984) *Nature*, **311**, 716–721.
17. Jeffreys,A.J., Wilson,V., Neumann,R. and Keyte,J. (1988) *Nucleic Acids Res.*, **16**, 10953–10971.
18. Ryffel,G.U. and McCarthy,B.J. (1975) *Biochemistry*, **14**, 1385–1389.
19. Kalb,V.F., Glasser,S., King,D. and Lingrel,J.B. (1983) *Nucleic Acids Res.*, **11**, 2177–2184.
20. Krayev,A.S., Markusheva,T.V., Kramerov,D.A., Ryskov,A.P., Skryabin,K.G., Bayev,A.A. and Georgiev,G.P. (1982) *Nucleic Acids Res.*, **10**, 7461–7475.
21. Kohne,D.E., Levison,S.A. and Byers,M.J. (1977) *Biochemistry*, **16**, 5329–5341.
22. Abe,K., Wei,J.-F., Wei,F.-S., Hsu,Y.-C., Uehara,H., Artzt,K. and Bennett,D. (1988) *EMBO J.*, **7**, 3441–3449.
23. Nielsen,P.J., McMaster,G.K. and Trachsel,H. (1985) *Nucleic Acids Res.*, **13**, 6867–6880.
24. Quax,W.J., Egberts,W.V., Hendriks,W., Quax-Jeuken,Y., Bloemendal,H. (1983) *Cell*, **35**, 215–223.
25. Martens,C.L., Huff,T.F., Jardieu,P., Trounstine,M.L., Coffman,R.L., Ishizaka,K. and Moore,K.W. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 2460–2464.
26. Burt,D.W., Reith,A.D. and Brammar,W.J. (1984) *Nucleic Acids Res.*, **12**, 8579–8593.
27. Leuders,K.K. and Mietz,J.A. (1986) *Nucleic Acids Res.*, **14**, 1495–1510.
28. Lu,X. and Werner,D. (1989) *Nucleic Acids Res.*, **17**, 442.
29. Bibb,M.J., van Etten,R.A., Wright,C.T., Walberg,M.W. and Clayton,D.A. (1981) *Cell*, **26**, 167–180.
30. Aoyama,Y., Chan,Y.L., Meyuhas,O. and Wool,I.G. (1989) *FEBS Lett.*, **247**, 242–246.
31. Gil,G., Smith,J.R., Goldstein,J.L., Slaughter,C.A., Orth,K., Brown,M.S. and Osborne,T.F. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 8963–8967.
32. Leibold,E.A. and Munro,H.N. (1987) *J. Biol. Chem.*, **262**, 7335–7341.
33. Tanaka,T., Kuwano,Y., Ishikawa,K. and Ogata,K. (1988) *Eur. J. Biochem.*, **173**, 53–56.
34. Didier,D.K., Schiffenbauer,J., Woulfe,S.L., Zacheis,M. and Schwartz,B.D. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 7322–7326.
35. Loeb,D.D., Padgett,R.W., Hardies,S.C., Shehee,W.R., Comer,M.B., Edgell,M.H. and Hutchison III,C.A. (1986) *Mol. Cell. Biol.*, **6**, 168–182.
36. Garboczi,D.N., Fox,A.H., Gerring,S.L. and Pedersen,P.L. (1988) *Biochemistry*, **27**, 553–560.
37. Lipman,D.J. and Peason,W.R. (1985) *Science*, **227**, 1435–1441.
38. D'Amore,M.A., Gallagher,P.M., Korfhagen,T.R. and Ganschow,R.E. (1988) *Biochemistry*, **27**, 7131–7140.