Description of the entire mRNA population by a 3' end cDNA fragment generated by class IIS restriction enzymes

Kikuya Kato

Okayama Cell Switching Project, ERATO, JRDC, 103-5 Tanakamonzencho, Sakyo-ku, Kyoto 606, Japan

Received June 16, 1995; Revised and Accepted August 17, 1995

ABSTRACT

A novel means of recording the expression status of the total gene population is described. Digestion of cDNA by class IIS restriction enzymes produces a fragment with a poly (A) stretch and a 5' overhang with an unknown sequence. This fragment contains information such as the class IIS enzyme that cuts cDNA nearest to the poly (A) stretch, the sequence of the 5' overhang, and the size of the fragment. Expressed genes can be discriminated and displayed by the fragment as follows: (i) cut the cDNA with one class IIS restriction enzyme; (ii) ligate the digested cDNA to that from a pool of 64 biotinylated adaptors cohesive to all possible overhangs; (iii) digest by other two class IIS enzymes; (iv) recover the ligated molecules with streptavidin-coated paramagnetic beads; (v) perform PCR with the adaptor-primer and an anchored oligo-dT primer; (vi) separate the amplified fragments by denaturing polyacrylamide gel electrophoresis. Repeat the experiment with 64 adaptors, three enzymes and three anchored oligo-dT primers displays most of the expressed genes. Because redundancy is minimized, this technique is also ideal for generating tags for expressed genes, with which to construct a transcript map of the genome.

INTRODUCTION

Since the introduction of recombinant DNA technology, many genes have been isolated and their structures determined. Soon, the human genome project and other related projects should uncover all the genes in the genome. An efficient method is desperately needed to record the expression states of a large number of expressed genes.

So far, expressed genes have been analyzed by Northern hybridization or RT-PCR. These techniques have been scaled up to describe part of an entire mRNA population (1). However, these procedures are laborious, and usually cannot be applied to unknown genes, since both methods are based on cloned fragments of each gene or their sequence information.

Here, I describe a new technique based upon another principle. Each expressed gene was discriminated and displayed by 3' end cDNA fragments generated by class IIS restriction enzyme digestion and PCR. Because the redundancy is minimized, this

method is also ideal to generate tags for expressed genes to construct transcript maps of the genome.

MATERIALS and METHODS

Oligonucleotides

C1T adaptor: 5'-biotinylated-GTACATATTGTCGTTAGAACGCT-3'
5'-OH-NXYZAGCGTTCTAACGACAATATGTAC-3'

C1G adaptor: 5'-biotinylated-GTACATATTGTCGTTAGAACGCG-3'
5'-OH-NXYZCGCGTTCTAACGACAATATGTAC-3'

N is a mixture of A, C, G and T. XYZ represents one of all possible sequences. We used C1G sequences when Y and Z were A or T. Otherwise, we used C1T.

C1S primer: 5'-OH-GTACATATTGTCGTTAGAACGC-3'

For fluorescent detection, 5' end of C1S was labelled by one of 5-carboxyfluorescein (FAM), 2',7'-dimethoxy-4',5'-dichloro-6-carboxyfluorescein (JOE) or 6-carboxy-tetramethyl-rhodamine (TAMRA), following the suppliers (Perkin-Elmer) protocol.

Oligonucleotides used as probes for cDNA detection and library survey were as follows:

AM2 receptor, 5'-OH-TTTAGTTGTAAAGGAATTCAGC-3';
apolipoprotein A-I, 5'-OH-CCTCACTGGGCAGTCAGAGT-3';
fatty acid synthase, 5'-OH-CGCTCACGTGCAGTTTAATTGT-3';
insulin-like growth factor A-I, 5'-OH-TGGGAGGCTCCTCTACA-TTCT-3';
lamin B, 5'-OH-GCTCCCCTCATGAGTCAACAAT-3';
asialoglycoprotein receptor, 5'-OH-TTTCCTACTCCATTCAGAAG-GC-3';
purine nucleotide phosphorylase, 5'-OH-ACAACATCTGGAAGT-GAAGGG-3';
prothrombin, 5'-OH-TGAGGATTCCATAGTGTAGGTC-3';
xanthine dehydrogenase, 5'-OH-CCCTTGGTGGCTTGGCAGAT-AT-3';
cytochrome p-450 IIE, 5'-OH-TAGCTGTCTTCAGACACTCCAG-3'.

Preparation of cDNA

Total RNA was prepared from Balb/c mouse liver and kidneys as described (2), and cDNA was produced as described by Gubler and Hoffman (3), using M-MLV reverse transcriptase (Gibco-BRL), human placental RNase inhibitor (Toyobo), *E.coli* DNA polymerase I (Toyobo) and RNase H (Toyobo). The primer for the first strand cDNA synthesis was 5'-amino d(T)₁₈. Usually,

Table 1. Specificity of the adaptor and anchored primer

gene name	enzyme for the first digestion	predicted adaptor	predicted anchored base	predicted fragment size	abundance (%) ¹⁾	adaptor specificity	observed anchored base	observed fragment size ²⁾
AM2 receptor	Bsm FI	NTTA	A	111	0.08	complete	A	N.D.
apolipoprotein A-I	Bsm AI	NTGG	C	136	0.5	complete	C,G	136.03
fatty acid synthase	Fok I	NATG	C	250	0.01	complete	C	249.53
insulin-like growth factor A-I	Bsm AI	NCAG	C	367	<0.01	complete	C	N.D.
lamin B	Fok I	NCTG	A	226	<0.01	complete ³⁾	A	N.D.
asialoglycoprotein receptor	Bsm AI	NGTA	C	149	<0.01	complete	C	N.D.
purine nucleotide phosphorylase	Bsm AI	NAAC	A	356	0.03	complete	A	N.D.
prothrombin	Fok I	NCCC	A	237	0.16	dominant	A,C	236.54
xanthine dehydrogenase	Bsm FI	NCCT	C	313	0.03	(-) ⁴⁾	(-)	(-)
cytochrome p450 IIE	Bsm AI	NGAA	A	288	0.6	complete	A,G	N.D

1) Calculated from appearance in 10 000 plaques of a mouse liver cDNA library.

2) Calculated by the GeneScan software. N.D., not confirmed by sequencing.

3) Amplification was detected only after a 5 day exposure with an intensifying screen.

4) No amplification was detected even after a 5 day exposure with an intensifying screen.

3 µg RNA was converted to double-stranded cDNA with 200 U reverse transcriptase, and digested with 5 U of either of three class IIS restriction enzymes according to the supplier's protocol for 45 min. We used *FokI* (TaKaRa), *BsmAI* (NEB) and *BsmFI* (NEB). After phenol extraction and ethanol precipitation, the product was dissolved in 70 µl of distilled water.

Ligation of adaptor

The ligation mixture contained 10 mM Tris-HCl pH 8.0, 4 mM MgCl₂, 10 mM (NH₄)₂SO₄, 0.5 mM NAD (Sigma), 1.2 mM EDTA, 0.005% BSA, 3 U *E.coli* DNA ligase (TaKaRa), 1 pmol adaptor and 1 µl of the above cDNA in 10 µl. The reaction proceeded at 16°C for at least 14 h, then the product was digested by two class IIS restriction enzymes other than that used in the previous digestion.

Recovery of cDNA and PCR amplification

The ligated sample was recovered with 0.1 mg of streptavidin-coated paramagnetic beads (Dynabeads-streptavidin, Dynal), and treated with 0.1 N NaOH according to the supplier's protocol. The PCR mixture consisted of 1× PCR buffer for Stoffel fragment, 2.5 mM MgCl₂, 2 mM deoxynucleotide triphosphates, 5 pmol fluorescent-C1S, 20 pmol of an anchored oligo(dT) primer, one third of the absorbed DNA and 0.5 U Stoffel fragment (Perkin-Elmer) in a 10 µl of reaction mixture. FAM-labelled C1S with d(T)₂₅A, JOE-labelled C1S with d(T)₂₅C and TAMRA-labelled C1S with d(T)₂₅G were used. Petroleum jelly separated the primers and other components until the amplification cycle was started (4). PCR amplification consisted of 35 cycles of 94°C for 30 s, 50°C for 1 min and 72°C for 1 min. After combining 1 µl of the product with FAM-labelled C1S and d(T)₂₅A, 2 µl of that with JOE-labelled C1S and d(T)₂₅C, and 3 µl of that with

TAMRA-labelled C1S and d(T)₂₅G, the sample was mixed with 10 µl 10 mM Tris-HCl (pH 8.0), 10 mM MgCl₂, 50 mM NaCl, 0.2 mM dNTP containing 0.1 U T4 DNA polymerase and incubated at 37°C for 50 min. One fifth of the sample was loaded onto a denaturing polyacrylamide gel.

Data collection and analysis

Separation by denaturing polyacrylamide gel electrophoresis was monitored by a 373A sequencer (Perkin-Elmer). Data were collected and analyzed with the 672 GeneScan software. Usually, data from peaks >35 were collected. GS-500 ROX (Perkin-Elmer) was used as the size marker, and fragment sizes were calculated by the local Southern method.

Small programs for data processing were written using the programming language, *awk* (5).

Adaptor specificity

Amplified cDNA fragments were separated by electrophoresis through a 2% agarose gel, then transferred to Hybond-N+ (Amersham). For the adaptor assay, 1 µl from each anchored primer product was combined and applied to lanes. For the primer assay, 1 µl each of the product was applied. Oligonucleotides were labelled using T4 polynucleotide kinase. Hybridization proceeded in 10 mM Tris-HCl (pH 7.5), 1 M NaCl, 2× Denhardt's solution, 0.5% SDS, 3–8 × 10⁵ c.p.m./ml of a probe at 55°C, and the membrane was washed twice with 10 mM Tris-HCl (pH 7.5), 1 M NaCl, 0.1% SDS at room temperature, then once with the same solution at 55°C. The membranes were exposed to Kodak XAR-5 films for 3–5 h without intensifying screens.

Northern hybridization

Mouse liver or kidney RNA (1 µg) was separated by formaldehyde 1% agarose gel electrophoresis and transferred to Hybond-

Figure 1. Outline of the technique. This figure shows a situation where *FokI* was used as the first enzyme digestion.

N (Amersham). Amplified cDNA fragments were separated and purified by 4% MetaPhor (FMC BioProducts) agarose gel electrophoresis, and labelled by random priming. Hybridization proceeded in $5\times$ SSC, $2\times$ Denhardt's solution, 0.5% SDS, $3-8\times 10^5$ c.p.m./ml of a labelled fragment at 65°C , and the membrane was washed with $0.1\times$ SSC, 0.1% SDS at 65°C . The membranes were exposed to Kodak XAR-5 films for 9 or 10 days with intensifying screens.

RESULTS

Outline of the technique

Class IIS restriction enzymes, a subgroup of class II restriction enzymes, cleave DNA at a precise distance outside their recognition sites, and produce overhangs of unknown sequences (6). A total of 64 biotinylated adaptors were prepared for all

Figure 2. Autoradiograms of the specificity test for adaptors (A-D) and anchored primers (E-H). (A) Apolipoprotein A-I. Lane 1, NTAG adaptor; 2, NTGC; 3, NTTG; 4, NTGA; 5, NTGC; 6, NTGT; 7, NAGG; 8, NCGG; 9, NGGG; 10, NTGG. (B) Fatty acid synthase. Lane 1, NAAG; 2, NACG; 3, NAGG; 4, NATG; 5, NCTG; 6, NGTG; 7, NTTG; 8, NATA; 9, NATC; 10, NATT. (C) Asialoglycoprotein receptor. Lane 1, NATA; 2, NCTA; 3, NGTA; 4, NTTA; 5, NGAA; 6, NGCA; 7, NGGA; 8, NGTC; 9, NGTG; 10, NGTT. (D) Prothrombin. Lane 1, NACC; 2, NCCC; 3, NGCC; 4, NTCC; 5, NCAC; 6, NCGC; 7, NCTC; 8, NCCA; 9, NCCG; 10, NCCT. (E) Apolipoprotein A-I. (F) Fatty acid synthase. (G) Asialoglycoprotein receptor. (H) Prothrombin. (I) Purine nucleotide phosphorylase. Lanes A, C and G represent d(T)₂₅A, d(T)₂₅C and d(T)₂₅G, respectively.

possible overhangs created by class IIS restriction enzymes such as *FokI*. Each adaptor had a 5' four nucleotide overhang, of which the outermost base was a mixture of A, C, G and T, and the inner three bases were one of all possible sequences. First, the cDNA was digested by one of class IIS restriction enzymes (*FokI*, *BsmAI* and *BsmFI*) and ligated to one of the 64 adaptors. The cDNA was then digested by two class IIS enzymes not used in the previous digestion. This experiment was repeated with three enzymes and 64 adaptors, which divided the expressed genes into 192 subpopulations. Following the recovery of the ligation sample with streptavidin-coated paramagnetic beads, the cDNA was treated with a dilute alkali, which removed the oligonucleotide complementary to the adaptor-primer, a potential inhibitor of the PCR reaction. PCR amplification proceeded with this cDNA sample, using the adaptor-primer and one of the anchored oligo-dT primers, d(T)₂₅A, d(T)₂₅C and d(T)₂₅G. The products were separated by denaturing polyacrylamide gel electrophoresis and the sizes of fragments were automatically recorded by a 373A sequencer and analyzed by the 672 GeneScan software. Since a single run of the autosequencer processes 108 samples, six runs were sufficient for the total experiment. *FokI*, *BsmAI* and *BsmFI* were used in this study, and from the database survey, 97% of the total mRNA population had recognition sequences for at least one of these enzymes. The strategy is schematically summarized in Figure 1.

Specificity of adaptor ligation and anchored primers

There were 35 mRNA sequences with the poly(A) site description, of which RNA sources were mouse liver, in the GenBank database, release 30. To test the specificity of the adaptor ligation,

Figure 3. Examples of electropherograms. (left) Top electropherograms are of mouse liver RNA, and those at the bottom are of mouse kidney RNA. All results used cDNA cut with *FokI* in the first digestion. Adaptors and primers are as follows: (A) NGAA adaptor, TAMRA-C1S and d(T)₂₅G; (B) NAGC adaptor, TAMRA-C1S and d(T)₂₅G; (C) NGCG adaptor, TAMRA-C1S and d(T)₂₅G; (D) NGAT adaptor, JOE-C1S and d(T)₂₅C. (right) Northern hybridization. Each probe corresponds to the peak marked by an arrow head. Lane L, total mouse liver RNA; lane K, total mouse kidney RNA.

10 genes were selected, covering a variety of adaptor sequences. The cDNA fragments generated from mouse liver RNA were amplified with adaptors with proper sequences or those with sequences with one base replacement. Products were separated by agarose gel electrophoresis, and after Southern transfer, each membrane was probed with a labelled oligonucleotide corresponding to each gene. The results are summarized in Table 1, and some of autoradiograms are shown in Figure 2A–D. Eight out of ten tested genes were amplified to the level detectable by the DNA sequencer, and except for prothrombin, amplification proceeded only with the correct adaptors. The intensity of the prothrombin signal with NACC was 7% of that with NCCC, according to electropherograms analyzed by the DNA sequencer (data not shown). These results demonstrated *E.coli* DNA ligase discriminated the three nucleotides adjacent to the ligation site,

Figure 4. 4% MetaPhor agarose gel electrophoresis of amplified products. The cDNA was cut with *FokI* in the first digestion and the NGAA adaptor and d(T)₂₅G were used. A non-fluorescent C1S primer was used for PCR, and 10 μ l of reaction mixture was applied to each lane. Lane L, mouse liver; K, mouse kidney. This figure corresponds to Figure 3A. The arrow head indicates the fragment used as the probe in the Northern hybridization. Right bars indicate size markers (Φ X174/*Hae*III). From above, 603, 310, 281, 271, 234, 194, 118 base pairs.

and each 3' end cDNA fragment was amplified only with the correct adaptor, except for some very abundant species.

The specificity of anchored oligo-dT primers was also examined (Table 1 for summary, Fig. 2E–I for autoradiograms). Except for abundant species (prothrombin, apolipoprotein A-I and p-450 IIE), amplification occurred only with the primer predicted from the database sequence. Although both the adaptor and anchored oligo-dT primer were less specific for some of abundant species, they are a limited population, and would not complicate the total results.

Three fragments were sequenced to confirm identity of the genes (Table 1). The sizes of these fragments calculated by the software were the same as those predicted from the database.

Expressed gene analysis of liver and kidney

All 64 adaptors were tested with mouse liver cDNA. During the course of this study, I found that some adaptors based on the C1T sequence were not effectively ligated. They were replaced with those based on the C1G sequence. Amplification profiles of mouse liver and kidney RNA were compared using part of the adaptor pool. Some examples of electropherograms are shown in Figure 3 (left), and an image of conventional agarose gel electrophoresis is shown in Figure 4. Each pair of electropherograms had many common peaks, but some peaks were tissue-specific. To test whether electropherogram patterns really reflect expression states, fourteen liver- or kidney-specific peaks were examined by Northern hybridization. Thirteen produced the results predicted by electropherograms, and one did not emit any detectable signal. The expression level was likely to be too low for detection by Northern hybridization. Some of the autoradiograms are shown in Figure 3 (right).

Since an enormous amount of data was produced by this method, a comprehensive presentation was necessary. First, data of satellite peaks, i.e. small peaks around a large peak, were

Figure 5. A scatter plot of the major fragments. The cDNAs were cut with *FokI* as the first enzyme digestion. Horizontal axis, adaptors and anchored primers. Triplets of each adaptor sequence represents d(T)₂₅A, d(T)₂₅C and d(T)₂₅G, from left to right. Vertical axis, fragment size in base pairs. Fragments of mouse liver are plotted as +, those of mouse kidney as x.

removed from those exported from the GeneScan software. Thereafter fragments with peak heights larger than a defined value (in the case of Fig. 5, 150), were selected, and, if there were fragments of the same size (± 1.0 bp, considering the accuracy of the sequencer) in the other tissue, they were also selected. These fragments were displayed as a scatter plot. Some of the results are shown in Figure 5. Since common fragments are represented as asterisks, the major specific fragments were easily identified. However, the electropherograms must be read to identify small specific peaks.

DISCUSSION

Sequence information of 5' overhangs generated by class IIS restriction enzymes is useful for characterizing DNA fragments. It was first used for DNA fingerprinting physical maps (7). Sequencing of these ends using a dye terminator increases the information content, making comparisons more reliable than those by fragment size alone. Several researchers have suggested that fragments generated by class IIS restriction enzymes can be amplified with a library of 256 adaptors (indexers) (8,9). These techniques, named 'molecular indexing' or 'indexing', have so far been applied only to genomic or cloned DNA.

Here, I described the application of molecular indexing to the analysis of expressed genes. The technique is based on the finding that *E.coli* DNA ligase discriminates the three nucleotides adjacent to the joining site. Each expressed gene is displayed by

a 3' end cDNA fragment generated by class IIS restriction enzyme digestion. The technique is analogous to classical two dimensional protein gel electrophoresis (10), but genes of interest can be easily isolated and sequenced. In particular, it is very powerful for identifying and isolating differentially expressed genes.

Differential display (11) and a related technique (12) can also reveal expressed genes as PCR-amplified fragments. The mRNA subpopulations defined by arbitrary primers and anchored oligo-dT primers are amplified after reverse transcription and resolved on a denaturing acrylamide gel. Comparing fingerprinting patterns from different sources of RNA, differentially expressed genes are identified. The aim of differential display is to identify differentially expressed genes, not to describe the expression status of each gene. Although this technique is becoming popular, there remain technical problems to be solved (13). Furthermore, subpopulations amplified by different sets of primers theoretically overlap, and most of the genes would be displayed more than once when multiple sets of primers are used. For these reasons, it is only used to detect the differential expression of similar cell types: fingerprinting patterns are too complicated for comparing distant tissues such as liver and kidney.

Molecular indexing has several technical advantages. In differential display, fragment sizes are determined and sequences are recognized only by arbitrary primers. Here, it is replaced by more accurate restriction enzyme digestion and the use of *E.coli* DNA ligase. Molecular indexing theoretically divides the entire

mRNA population into 576 non-overlapping subpopulations. This makes a clear contrast to differential display, which has no means of avoiding redundancy. PCR proceeds under standard conditions, and band patterns are not sensitive to condition of thermal cycling. Reamplification is not necessary to recover the fragments.

Partial sequencing of a large number of cDNA pieces and subsequent mapping onto the genome is regarded as a short cut for those who involved in the human genome project (14,15). One of the main problems of this approach is redundancy of the cDNA clone pool (16). Normalized cDNA libraries have been constructed to solve this problem (17), but redundancy is inevitable even from such libraries. Molecular indexing almost completely solves this problem, because the majority of genes are amplified only with one adaptor and anchored primer set.

Several technical limitations should be noted. Those cDNA sequences lacking enzyme recognition sites will not be displayed. This may be complemented by using other enzymes such as *Sfa*NI and *Bbv*I. Some sequences may not be efficiently ligated to the adaptor. Artifacts intrinsic to PCR, such as by-products or inefficient amplification, are inevitable. The current protocol can discriminate expressed genes in several tissues, because the number of mRNA species in a single type of cell is estimated at 10 000. However, it is not enough for all the genes in the genome, which number about 50 000–100 000. Four-nucleotide recognition by *E.coli* DNA ligase will solve this problem, but we have to identify conditions required to do this.

Despite these limitations, molecular indexing is the first technique for scanning global gene expression in a tissue or a cell. This procedure will be widely applicable to various biological systems.

ACKNOWLEDGEMENT

The author thanks Professor Hiroto Okayama for support and reading of the manuscript.

REFERENCES

- 1 Kato, K. (1990) *Eur. J. Neurosci.*, **2**, 704–711.
- 2 Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual*, 2nd edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor.
- 3 Gubler, U. and Hoffman, B.J. (1983) *Gene*, **25**, 263–269.
- 4 Horton, R.M., Hoppe, B.L. and Conti-Tronconi, B.M. (1994) *Biotechniques*, **16**, 42–43.
- 5 Kernigan, B.W. and Pike, R. (1984) *The UNIX Programming Environment*. Prentice-Hall, Inc., New Jersey.
- 6 Szybalski, W., Kim, S.C., Hasan, N. and Podhajski, A.J. (1991) *Gene*, **100**, 13–26.
- 7 Brenner, S. and Livak, K.J. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 8902–8906.
- 8 Smith, D.R. (1992) *PCR Methods and Applications*, **2**, 21–27.
- 9 Unrau, P. and Deugau, K.V. (1994) *Gene*, **145**, 163–169.
- 10 Dunbar, B.S. (1987) *Two-Dimensional Electrophoresis and Immunological Techniques*. Plenum Press, New York.
- 11 Liang, P. and Pardee, A. B. (1992) *Science*, **257**, 967–971.
- 12 Welsh, J., Chada, K., Dalal, S.S., Cheng, R., Ralph, D. and McClelland, M. (1992) *Nucleic Acids Res.*, **20**, 4965–4970.
- 13 Liang, P., Averboukh, L. and Pardee, A.B. (1993) *Nucleic Acids Res.*, **21**, 3269–3275.
- 14 Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F., Kerlavage, A.R., McCombie, W.R., Venter, J.C. (1991) *Science*, **21**, 1651–1656.
- 15 Southern, E.M. (1992) *Curr. Opin. Genet. Dev.*, **2**, 412–416.
- 16 Kato, K. (1993) In Davies, K.E. and Tilghman, S.M. (eds), *Genome Analysis: Genome Maps and Neurological Disorders*. Cold Spring Harbor laboratory Press, Cold Spring Harbor, Vol. VI, pp. 1–18.
- 17 Ko, M.S.H. (1990) *Nucleic Acids Res.*, **18**, 5705–5711.