Short Technical Reports

Removal of PolyA Tails from Full-Length cDNA Libraries for High-Efficiency Sequencing

BioTechniques 31:1042-1049 (November 2001)

ABSTRACT

We have developed a method to overcome sequencing problems caused by the presence of homopolymer stretches, such as polyA/T, in cDNA libraries. PolyA tails are shortened by cleaving before cDNA cloning with type IIS restriction enzymes, such as *GsuI*, placed next to the oligo-dT used to prime the polyA tails of mRNAs. We constructed four rice Cap-Trapper-selected, full-length normalized cDNA libraries, of which the average residual polyA tail was 4 bases or shorter in most of the clones analyzed. Because of the removal of homopolymeric stretches, libraries prepared with this method can be used for direct sequencing and transcriptional sequencing without the slippage observed for libraries prepared with currently available methods, thus improving sequencing accuracy, operations, and throughput.

INTRODUCTION

We are systematically isolating and sequencing mouse and rice full-length cDNAs to create the mouse and rice full-length cDNA encyclopedias. Full-length cDNA candidates for complete sequencing have been selected by clustering more than 1.1 million mouse and 85 000 rice one-pass sequencing reads from 3' ends. To speed up and reduce the costs of such large-scale sequencing operations, direct sequencing would be extremely useful because it allows skipping the plasmid preparation step. Alternatively, the use of the transcriptional sequencing (12,13) would have similar advantages because it requires only a minimal quantity of template, isothermal amplification, and no template concentration adjustment.

However, when preparing cDNA libraries, homopolymeric sequences flanking the cDNA are invariably created, such as the G/C-tails (1–3,5) at the

5' end and the polyA/T stretches at the 3' ends. The latter is invariably originated by priming with oligo-dT the polyA of mRNAs. Unfortunately, homopolymeric stretches affect the overall sequencing efficiency, accuracy, cause severe slippage during the PCR amplification (9), and, limited to A/T stretches, cause slippage also during the transcriptional sequencing, making impossible the use of these convenient techniques with traditionally prepared cDNA libraries. Long polyA stretches have also been hypothesized to interfere with the normalization/subtraction of cDNA libraries (18) because of cross hybridization of polyA/T sequences.

This report contains a universally applicable method to remove homopolymers from cDNA libraries. Trimming of polyA tails was obtained by using type IIS restriction enzymes, which digest sites that are apart from the recognition sites. We have been using *GsuI*, which does not cleave hemimethylated cDNA (1,3,5), to construct Cap-Trapper-selected, normalized/subtracted full-length cDNA libraries (1–3,5), including four rice cDNA libraries. Beside of the removal of 3' A/T stretches, the method is applicable to remove other undesired sequences such as dG/C tails stretches at the 5' end of cDNAs (2,3,5). This technique, which can be applied to any cDNA library preparation method, is particularly beneficial for high-throughput sequencing gene discovery.

MATERIALS AND METHODS

Oligonucleotides and Linkers

The oligonucleotides were custom-synthesized (Invitrogen, Tokyo, Japan) and then purified by using denaturing PAGE (11). We prepared the BG primer 5'-(GA)₅GGATCCTTCTGGA-GAG(T)₁₆VN-3' (V = A, G, or C), where N represents any nucleotide. An adapter was prepared by mixing the oligonucleotides 5'-CAGTCCG-3' (phosphorylated at the 5' end) and 5'-GATCCGGACTGTT-3' in NaCl solution (final concentration, 100 mM), followed by incubation of this mixture at 65°C for 5 min, 45°C for 5 min, 37°C for 10 min, and 25°C for 10 min.

RNA Preparation

We modified a standard procedure (2) to prepare total mRNA from the brains of C57B6/J mice and from rice flowers, shoots, calluses, and seed (1, 16). We purified the mRNA by using the MACS mRNA isolation kit (Miltényi Biotec, Bergisch Gladbach, Germany).

Preparation of Full-length cDNA Libraries Lacking PolyA Stretches

Cap-Trapper-selected, full-length, first-strand cDNA was prepared from 10 µg mRNA as described (3), with the exceptions that the first-strand cDNA primer was the BG primer and hot start was performed to prime at the beginning of the polyA stretch (3,10). The first-strand cDNA was hemimethylated by using 5-methyl-dCTP (2,3,5) to avoid internal cleavage with the enzymes used in this study. The first-strand cDNAs underwent Cap-Trapper selection by using the single-strand linker ligation method (15), and the resulting primed cDNAs were normalized (5). Then, the second-strand cDNA synthesis was performed (Figure 1A), and the products were extracted by using phenol/chloroform and chloroform and then ethanol-precipitated according to standard procedures (3). The double-stranded cDNAs were digested with *GsuI* (10 U/µg cDNA; MBI Fermentas, Vilnius, Lithuania) (Figure 1B) at 30°C for 1 h in a volume of 50 µL according to the manufacturer's recommendations. We then added 1 µL 10 mg/mL proteinase K to the samples in 10 mM EDTA/0.2% SDS (reaction volume, 55 µL) and incubated them at 45°C for 15 min. They were extracted by using phenol/chloroform and chloroform and then ethanol-precipitated (3).

The samples were dissolved in sterilized water and mixed with 200 ng adapter, 1 µL 10× ligation buffer, and 160 U T4 DNA ligase (New England Biolabs, Beverly, MA, USA) in a total volume of 10 µL. The samples were incubated overnight at 16°C, followed by inactivation of ligase by heating to 65°C for 20 min (Figure 1C). After the adapter ligation, cDNA was further cleaved at the 5' end by using *XhoI* to provide the 5' cloning site. Five microliters of 10× NEB buffer 2, 0.5 µL 10

Short Technical Reports

mg/mL bovine serum albumin, and *Xho*I (25 U/μg cDNA) (all from New England Biolabs) were added to the samples in a reaction volume of 50 μL. The samples were incubated at 37°C for 1 h (Figure 1D). Next, 1 μL 10 mg/mL proteinase K was added to the samples in 10 mM EDTA/0.2% SDS (reaction volume, 55 μL) and incubated at 45°C for 15 min. The reaction mixture was extracted with phenol/chloroform and chloroform, and 20 μL 0.1 M MgCl₂ and 80 μL 20% polyethylene glycol (PEG)-8000 were added to the aqueous phase in a final volume of 200 μL. cDNA was precipitated by centrifugation at 13000× g for 10 min. The cDNA pellet was dissolved in 10 μL sterilized water, purified by using CL-4B Sepharose® Miniprep Spun Columns (Amersham Pharmacia Biotech, Tokyo, Japan), and centrifuged at 400× g; the excluded fraction was precipitated with isopropanol in the presence of 1 μg glycogen according to standard procedures (11).

We ligated overnight 10 ng cDNA to 190 ng of the λ-FLC-I-B cloning vector, which has two loxP sites (4,6), pre-

viously cleaved with *Sa*II (ends compatible with *Xho*I) and *Bam*HI. The ligated samples were packaged (MaxPlax™; Epicentre, Madison, WI, USA) and used to infect C600 cells (Figure 1E) to amplify the libraries (11). We isolated the phage DNA from 1 mL amplified phage library solution by using the Wizard® Lambda Preps DNA Purification System (Promega, Madison, WI, USA). We converted the phage DNA to plasmid by using Cre-recombinase (Novagen, Madison, WI, USA) as described (4). We used the bulk-excised plasmids to transform DH10B cells (Invitrogen) according to the manufacturer's instructions.

Analysis of the 3' End Sequence

Sequencing operations with the RIKEN integrated sequence analyzer system were previously described (7,14). We used the reverse primer 5'-AGCGGATAACAATTTTCACACAGG A-3' to sequence cDNAs from the 3' end to determine the lengths of the polyA/T stretches.

Transcriptional Sequencing and Direct Sequencing

cDNA clones prepared from two mouse brain and kidney libraries were sequenced by using both transcriptional sequencing and direct sequencing of colony PCR product. The brain Cap-Trapped cDNA library was deprived of the polyA as described here, while the kidney library was prepared with the standard Cap-Trapper method (3); therefore, the clones have regular polyA stretches. Transcriptional sequencing reactions were carried out as reported previously (13). For direct sequencing, bacterial colonies were picked and resuspended in 100 μL sterilized water; 1 μL of this cell suspension was added to the PCR mixture containing 2 μL 10× ExTaq™ buffer, 0.5 μL 2.5 mM dNTP mixture, 2.5 U ExTaq (Takara, Tokyo, Japan), 0.5 μM of the BSF primer (5'-CTGCAAGGC-GATTAAGTTGG-3'), and 0.5 μM of the BSR primer (5'-CTTCCGGCTCG TATGTTGTG-3'). cDNAs were denatured initially at 94°C for 3 min, ampli-

Figure 1. Overview of the preparation of Cap-Trapper-selected, full-length cDNA libraries from which polyA stretches are removed. Cap-Trapper-selected, full-length cDNAs underwent second-strand synthesis (A), *Gsu*I digestion (B), ligation of a *Bam*HI compatible adapter at the 3' end (C), *Xho*I restriction digestion at the 5' end (D), and cloning into the λ vector (E).

Short Technical Reports

fied by using 39 cycles of 95°C for 15 s, 50°C for 15 s, and 72°C for 2 min, then extended at 72°C for 5 min. The PCR products were incubated with 1 U exonuclease I (USB, Tokyo, Japan) and 1 U shrimp alkaline phosphatase (Invitrogen) at 37°C for 30 min, followed by inactivation at 80°C for 15 min. DNA sequencing was then carried out by using the BSF primer and BigDye® Terminators (Applied Biosystems, Foster City, CA, USA). Both types of sequencing reactions were analyzed by using an ABI PRISM® 377 DNA sequencer (Applied Biosystems).

RESULTS AND DISCUSSION

Critical Factors to Prepare cDNAs Devoid of PolyA Stretches

To remove polyA stretches, we chose the type IIS restriction enzyme *GsuI* because it conveniently cleaves 16/14 bp apart from its recognition site (Figure 1B). After cleavage of the cDNA by *GsuI*, 2-base cohesive ends are produced (Figure 1C), which are suitable for attaching a *Bam*HI linker (Figure 1D) for cloning in the λ -FLC-I vector (4). *GsuI* does not cleave hemimethylated cDNAs (2,3,5), but only the BG primer sequence, as tested by using double-stranded hemimethylated (5-methyl cytosine) DNA oligonucleotides (data not shown). In contrast to other enzymes such as *BbvI*, *GsuI* was sufficiently purified from DNA exonuclease or endonuclease activities (data not shown), allowing high-efficiency re-ligation after cleavage. It was important to prepare full-length cDNAs by using hot-start first-strand cDNA synthesis in the presence of trehalose (3,10). This increases the priming specificity of the BG primer, which contains a *GsuI* recognition sequence next to a dT₁₆VN sequence. At this condition, most of the mRNAs are primed exactly at the beginning of the polyA, next to the 3' untranslated regions (3' UTRs). Another essential step was the PEG precipitation to eliminate small fragments produced by restriction digestion and unligated excess of adapters. At the described condition, we obtained about 1×10^6 primary phage plaques when cloning only 10 ng cDNA.

Table 1. Length of the Residual T-Stretch

Classifications	Sequence	Number of Clones	Rate (%)
Group I	(vector)-(adapter)-TTTT-(cDNA)	63283	92.40%
Group II	(vector)-(adapter)-TTT-(cDNA)	4157	6.10%
Group III	(vector)-(adapter)-TT-(cDNA)	608	0.90%
Group IV	(vector)-(adapter)-T-(cDNA)	405	0.60%
Group V	(vector)-(adapter)-V-(cDNA)	38	0.10%
		68491	
The 3' end sequence of the rice cDNA libraries (68491 clones) were classified. It showed that 63283 out of 68491 clones (92.4%) were cut at the precise position of polyT. Clones of group II–V might be cut by <i>GsuI</i> incorrectly.			

Analysis of Length of PolyA Stretches

To confirm the *GsuI* specificity, the rice cDNA libraries were sequenced and classified with regard to the number of dTs, which corresponds to the remaining dAs (Table 1). We classified 62283 out of the 68491 successfully sequenced clones (92.4%) in group I (4 dTs), as expected if *GsuI* would cut exactly at the described site. Some clones (7.6%) had less than 4 dTs at their 3' ends and were allocated into groups II (3 dTs), III (2 dTs), IV (1 dT), and V (no dT), which anyway allow successful sequencing. Finding dTs stretches shorter than 4 nucleotides suggests that some *GsuI* cleavages occurred at 17/15, 18/16, 19/17, and 20/18 bp from the recognition site. Although the cDNAs in groups IV (0.6%) and V (0.1%) do not have AA-protruding ends after *GsuI* digestion, they seem to have been ligated to the adapters having TT-protruding ends. Group I contains also polyA stretches (>4 dTs; 12.3% of group I clones), the sequence of which is (adapter)-TTTTVTn-(cDNA), $n \geq 1$. This is probably caused by the failure of the first-strand cDNA primer to anneal exactly at the next to the 3' UTR, thus underscoring the importance of accurate priming (10). The number of clones with n5, which cannot be sequenced efficiently, is 6.4% only.

Transcriptional Sequencing and Direct Sequencing

Selection of new cDNA clones for

full-length sequencing is made by 3' end sequencing and clustering of 5' end validated cDNA libraries (5). After large-scale sequencing and clustering, only single clones selected from different clusters are then used for full-insert sequencing. Thus, the high efficiency of 3' end sequencing by direct or transcriptional sequencing (13) is important in a full-length cDNA sequencing approach. However, in clones with polyA stretches, slippage always occurred after the polyA stretches with 28 samples prepared by transcriptional sequencing and 36 prepared by direct sequencing, thus making these approaches impossible. In direct sequencing, the observed slippage would be caused by the *Taq* DNA polymerase during PCR amplification, as observed for microsatellites PCR (8). In transcriptional sequencing, the low melting temperature A_T/U_n ($n > 7$) DNA-RNA hybrid can be momentarily denatured during elongation and improperly rehybridize, causing slippage. This causes random insertions into most of the transcripts of additional Us if compared to the template sequence (8).

Remarkably, slippage was never observed with either sequencing methods in clones in which polyA stretches had been removed and that were shorter than 7 As. Therefore, this technique greatly facilitates direct and transcriptional sequencing, thus overcoming the template preparation hurdles for large-scale DNA sequencing (13). Additionally, slippage-free direct sequencing after PCR amplification of bacterial colonies of large plasmids will circum-

vent the difficulty of isolating cDNAs recalcitrant to plasmid preparation. In fact, many long, full-length cDNA cloned in high copy number plasmids do not allow for efficient plasmid preparation because of the inherent instability of long-insert, high copy plasmids (17) and/or product toxicity for bacterial cells. Such cDNAs are likely to be more stable as a single-copy plasmid. This protocol can be generally used for any cloning purpose in which unwanted flanking/homopolymeric DNA sequences should be removed from the DNA target.

REFERENCES

- Carninci, P., A. Westover, Y. Nishiyama, T. Ohsumi, M. Itoh, S. Nagaoka, N. Sasaki, Y. Okazaki et al. 1997. High efficiency selection of full-length cDNA by improved biotinylated cap trapper. *DNA Res.* 28:61-66.
- Carninci, P., C. Kvam, A. Kitamura, T. Ohsumi, Y. Okazaki, M. Itoh, M. Kamiya, K. Shibata et al. 1996. High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics* 37:327-336.
- Carninci, P. and Y. Hayashizaki. 1999. High-efficiency full-length cDNA cloning. *Methods Enzymol.* 303:19-44.
- Carninci, P., Y. Shibata, N. Hayatsu, M. Itoh, T. Shiraki, T. Hirozane, A. Watahiki, K. Shibata et al. 2001. Balanced size and long-size cloning of full-length, Cap-trapped cDNAs into vectors of the novel λ -FLC family allows enhanced gene discovery rate and functional analysis. *Genomics* 77:79-90.
- Carninci, P., Y. Shibata, N. Hayatsu, Y. Sugahara, K. Shibata, M. Itoh, H. Konno, Y. Okazaki et al. 2000. Normalization and subtraction of Cap-Trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. *Genome Res.* 10:1617-1630.
- Hamilton, D.L. and K. Abremski. 1984. Site-specific recombination by the bacteriophage P1 lox-Cre system. Cre-mediated synopsis of two lox sites. *J. Mol. Biol.* 178:481-486.
- Itoh, M., T. Kitsunai, J. Akiyama, K. Shibata, M. Izawa, J. Kawai, Y. Tomaru, P. Carninci et al. 1999. Automated filtration-based high-throughput plasmid preparation system. *Genome Res.* 9:463-470.
- Macdonald, L.E., Y. Zhou, and W.T. McAllister. 1993. Termination and slippage by bacteriophage T7 RNA polymerase. *J. Mol. Biol.* 232:1030-1047.
- Mekus, F., T. Dork, T. Deufel, N. Morral, and B. Tummli. 1995. Analysis of microsatellites by direct blotting electrophoresis and chemiluminescence detection. *Electrophoresis* 10:1886-1888.
- Mizuno, Y., P. Carninci, Y. Okazaki, M. Tateno, J. Kawai, H. Amanuma, M. Muramatsu, and Y. Hayashizaki. 1999. Increased specificity of reverse transcription priming by trehalose and oligo-blockers allows high-efficiency window separation of mRNA display. *Nucleic Acids Res.* 27:1345-1349.
- Sambrook, J., E.F. Fritsch, and T. Maniatis. 1989. *Molecular Cloning: A Laboratory Manual*, 2nd ed. CSH Laboratory Press, Cold Spring Harbor, NY.
- Sasaki, N., M. Izawa, Y. Sugahara, T. Tanaka, M. Watahiki, K. Ozawa, E. Ohara, H. Funaki et al. 1998. Identification of stable RNA hairpins causing band compression in transcriptional sequencing and their elimination by use of inosine triphosphate. *Gene* 222:17-23.
- Sasaki, N., M. Izawa, M. Watahiki, K. Ozawa, T. Tanaka, Y. Yoneda, S. Matsuura, P. Carninci et al. 1998. Transcriptional sequencing: a method for DNA sequencing using RNA polymerase. *Proc. Natl. Acad. Sci. USA* 95:3455-3460.
- Shibata, K., M. Itoh, K. Aizawa, S. Nagao, N. Sasaki, P. Carninci, H. Konno, J. Akiyama et al. 2000. RIKEN Integrated Sequence Analysis (RISA) System-384-format sequencing pipeline with 384 multicapillary sequencer. *Genome Res.* 10:1757-1771.
- Shibata, Y., P. Carninci, A. Watahiki, T. Shiraki, H. Konno, M. Muramatsu, and Y. Hayashizaki. 2001. Cloning full-length, Cap-Trapper-selected cDNAs by using the single-strand linker ligation method. *BioTechniques* 30:1250-1254.
- Shirzadegan, M., P. Christie, and J.R. Seemann. 1991. An efficient method for isolation of RNA from tissue cultured plant cells. *Nucleic Acids Res.* 19:6055.
- Summers, D.K. 1991. The kinetics of plasmid loss. *Trends Biotechnol.* 9:273-278.
- Wang, S.M., S.C. Fears, L. Zhang, J.J. Chen, and J.D. Rowley. 2000. Screening poly(dA/dT)-cDNAs for gene identification. *Proc. Natl. Acad. Sci. USA* 97:4162-4167.

We thank M. Nakamura, Y. Hodoyama, C. Sakai, T. Tanaka, S. Akahira, A. Yasunishi, K. Imotani, T. Akimura, A. Tomaru, R. Numazaki, F. Takahashi, D. Sasaki, K. Nomura, A. Tagawa, T. Toya, T. Yamamura, M. Ohno, H. Kato, M. Kouda, H. Sano, K. Hiramoto, K. Higo, Y. Otake, and K. Sato for their technical assistance. This study has been supported by a research grant for the RIKEN Genome Exploration Research Project from the Science and Technology Agency of the Japanese government. This work also was supported as the Rice Genome Full-Length cDNA Library Construction Project by BRAIN (Bio-Oriented Technology Research Advancement Institution) and as the Human Genome Program by the Ministry of Education, Science, and Culture to Y.H. Address correspondence to Dr. Piero Carninci, Genome Science Laboratory, RIKEN 2-1 Hirosawa, Wako, Saitama, 351-0198, Japan. e-mail: rgscerg@gsc.riken.go.jp

Received 29 January 2001; accepted 30 April 2001.

Y. Shibata^{1,2,3}, P. Carninci^{1,2}, K. Sato^{1,2,4}, N. Hayatsu^{1,2}, T. Shiraki^{1,2}, Y. Ishii^{1,2}, T. Arakawa^{1,2}, A. Hara^{1,2}, N. Ohsato¹, M. Izawa^{1,2,3}, K. Aizawa^{1,2}, M. Itoh^{1,2}, K. Shibata^{1,2}, A. Shinagawa^{1,2}, J. Kawai^{1,2}, Y. Ota^{1,2}, S. Kikuchi⁵, N. Kishimoto⁵, M. Muramatsu¹, and Y. Hayashizaki^{1,2,4}

¹Genome Exploration Research Group
RIKEN Genomic Sciences Center
Kanagawa

²Genome Science Laboratory
RIKEN 2-1, Hirosawa
Saitama

³Nippon Gene
Toyama

⁴Institute of Basic Medical Sciences
University of Tsukuba
Ibaraki

⁵National Institute of Agrobiological Resources
Ibaraki, Japan

For reprints of this or
any other article, contact
Reprints@BioTechniques.com