Construction of cDNA Libraries for Highly Efficient DNA Sequencing from the 3' End of Expressed Genes

BioTechniques 34:758-766 (April 2003)

ABSTRACT

The majority of expressed sequence tag (EST) sequences available today have been derived from the 5' ends of cDNA clones. Obtaining high-quality DNA sequences from the 3' ends of oligo(dT)-primed cDNA on a large scale has been difficult because of slippage of the DNA polymerase enzyme used in direct PCR and cycle sequencing. With the completion of whole genome sequencing for more and more organisms, mRNA 3'-UTR sequences can be particularly useful for clustering large numbers of ESTs for the effective discrimination of individual genes and gene families. We have identified a flaw in the widely used oligo(dT) primers for cDNA synthesis, and here we describe an improved priming approach to effectively synthesize cDNA devoid of homopolymeric nucleotide stretches from mRNA poly(A) tails to enable highly efficient and reliable DNA sequence determination from 3' mRNA ends. Using this method, we produced a rat lung cDNA library and successfully sequenced the 3' ends of 98% of all attempted clones.

INTRODUCTION

In eukaryotes, the vast majority of processed genes contains a poly(A) tail addition at the 3' terminus of the mRNA (1). These stretches of poly(A) residues have been extensively exploited for the construction of cDNA libraries where an oligonucleotide primer of 15 or more deoxythymidine triphosphate [oligo(dT)] residues is conveniently hybridized to the poly(A) tail to initiate first-strand cDNA synthesis. Often, these primers may include rare restriction endonuclease sites, such as *NotI* or *PacI*, to facilitate

the directional cloning of cDNAs into the appropriate vector. In the past decade, several large-scale cDNA sequencing projects have been undertaken to determine the expressed sequence tags (ESTs) of humans and model animals. Even though nearly all of these cDNA libraries have been constructed using oligo(dT) priming, the majority of EST sequencing has been accomplished only from the 5' ends of these clones. This is in part because the 5'-end regions are where coding segments are likely to be represented. Obtaining accurate sequences from the 3' ends of these clones is usually less successful because of slippage of the sequencing enzyme across the homopolymer stretch of (dT)s (2). An anchored sequencing primer of 15–25 T residues followed by a degenerate 3' base of G, C, and A is often used in lieu of the vector primer to generate reliable DNA sequences through poly(A) tails (3,4). However, this requires the more expensive and tedious process of plasmid DNA preparation from bacterial colonies. As part of our high-throughput DNA sequencing operation, we routinely sequence the 5' ends of clones using direct PCR of individual colonies. Unfortunately, 3'-end sequencing of oligo(dT) clones has been much less successful even with the use of anchored primers because slippage can also occur during direct PCR amplification. Two methods for the preparation of cDNA libraries without these poly(A/T) stretches have been recently described. One strategy is based on nested deletions created at the 3' ends of the cDNA using exonuclease to remove the poly(A) tails within linearized plasmid clones followed by religation and transformation (5). With this procedure, however, variable lengths of 3'-UTR upstream of the poly(A) tail are also inevitably removed. Another method involves the use of *GsuI*, a type IIS restriction enzyme, to shorten the poly(A) tail of cDNA primed with an oligo(dT) primer harboring the restriction site (6). Unlike the first method, this procedure precisely trims the poly(A) tail to a defined length before cloning, avoiding over- or under-trimming. We have tested this method and were surprised to find long stretches of poly(A/T) in 10%–15% of

our clones despite precise restriction by the type IIS enzyme to remove the poly(A/T) stretch on the primer. As a result, the 3'-end insert sequences of these clones could not be reliably determined. Here we report on a novel method devised to remove these homopolymer stretches completely to allow for highly efficient 3' EST DNA sequencing. The improvements described can contribute greatly to the generation of superior cDNA libraries built for cost-effective and reliable high-throughput clone sequencing.

MATERIALS AND METHODS

Primers for First-Strand cDNA Synthesis

Figure 1 lists the names and sequences of oligonucleotides (Operon Technologies, Alameda, CA, USA) that were synthesized with a 5' phosphate group and purified using PAGE. These oligonucleotides were mixed in the following ratios (18.85% BpDocT1, 4.95% BpDocT2, 1.17% BpDocT3, 0.03% BpDocT4, 75% BpDocT5) and annealed to an equimolar amount of BpDocMas (7).

cDNA Synthesis

Total RNA was isolated from rat lungs using TRIzol® (Invitrogen, Carlsbad, CA, USA). mRNA was isolated using Oligotex™ beads (Qiagen, Valencia, CA, USA). For first-strand cDNA synthesis, 1 µg DNase I-treated mRNA in 12 µL water was incubated at 70°C for 5 min and quickly transferred to an ice bucket. To this tube, 4 µL 5× first-strand buffer were added (250 mM Tris-HCl, pH 8.3, 375 mM KCl, 15 mM MgCl₂, 50 mM DTT) along with 1 µL 10 mM mixture of dATP, dGTP, dTTP, and methyl-dCTP (Amersham Biosciences, Piscataway, NJ, USA), 1 µL ³²P-dCTP at 400 µCi/mmol (Amersham Biosciences), 200 ng primer mixture, and 1 µL (100 U) RNase H-minus Moloney murine leukemia virus (MMLV) reverse transcriptase (Promega, Madison, WI, USA). The reaction was incubated at 37°C for 1 h, after which 1 µL of the reaction was removed for incorporation analysis. Fol-

Clones attempted for sequencing	192	100%
Successful readable sequences	188	97.9%
Removals		
Incorrect insert flanking sequence	1	0.5%
Mitochondrial contamination	2	1.1%
Insert smaller than 50 bp	6	3.2%
Ribosomal contamination	0	0.0%
Poly(A/T) stretches	0	0.0%
Bacterial sequence	1	0.5%
Remaining usable sequences	178	94.7%

192 clones were picked randomly and sequenced. The values in the table represent the number and percentage of clones sequenced successfully (readable), the percentage of poly(A/T)-containing clones, the level of mitochondrial or ribosomal contamination, and any short or no-insert (<50 bp) clones.

England Biolabs, Beverly, MA, USA), 4 μ L (40 U) DNA polymerase I (New England Biolabs), and 1 μ L (2 U) RNase H (Invitrogen)] were added. The reaction was incubated at 16°C for 2 h before adding 10 U T4 DNA polymerase (New England Biolabs) and allowed to further incubate for another 5 min at 16°C. The residual primer and cDNA smaller than 400 bp was then removed from the sample using a CR-

Short Technical Reports

1000 spin column (BD Biosciences Clontech, Palo Alto, CA, USA) before precipitation with ethanol.

Cloning of cDNA into Plasmid Vector

The blunt-ended cDNA was ligated to *NotI*-*XhoI* adaptors (Operon Technologies) overnight at 16°C. The DNA was then extracted with phenol:chloroform and precipitated with ethanol. The DNA pellet was resuspended and digested with 60 U *BpmI* (New England Biolabs) at 37°C for 3 h. cDNA larger than 500 bp was size-selected using a 2-mL Sepharose CL-2B drip column. The cDNA was ethanol-precipitated, and approximately 10 ng cDNA were ligated overnight at 16°C to 25 ng vector DNA cut with *BsgI* and *NotI*. The ligation was then transformed into electrocompetent DH10B *E. coli* cells (Invitrogen). Average insert sizes were determined with PCR using vector primers on 48 randomly picked colonies.

Vector Construction and Preparation

pBluescript® II KS+ (Stratagene, La Jolla, CA, USA) was digested with *NotI* and *EcoRI*. The vector backbone DNA was purified using agarose gel

electrophoresis and ligated to an *EcoRI*/*NotI* adaptor containing restriction sites for *Bam*HI and *BsgI* (5'-GGCCGCTTTTTTCGGATCCGGG-GCTGCACG-3' annealed to 5'-AAT-TCGTGCAGCCCCGGATCCGAAA-AAAAGC-3') (Figure 2). The resulting vector was named KS+/NotI_*BsgI*. To prepare the vector for insert ligation, plasmid DNA was digested with *BsgI*, and the linearized DNA was purified using agarose gel electrophoresis. Next, the purified DNA was treated with *Bam*HI and *NotI* and purified again using agarose gel electrophoresis.

DNA Sequencing and Sequence Analysis

cDNA was amplified with PCR directly from colonies to generate template DNA for cycle sequencing. Automated DNA sequencing was performed with MegaBACE DNA sequencers (Amersham Biosciences) using Big-Dye™ terminator chemistry (Applied Biosystems) with an M13 Reverse primer. Base calling was performed using Phred, and vector sequences were removed with cross_match (9). Sequence comparisons were performed using BLAST (10). Polyadenylation signal sequences used in searches were based on a previous report (11).

RESULTS

cDNA Generation and Library Construction

From 1 µg starting mRNA, we obtained a cDNA conversion rate of 27.4%. The cDNA synthesized ranged from 300 bp to more than 9 kb (data not shown). In our experience, this is compatible with cDNA synthesis rates and sizes obtained from conventional single-stranded cDNA primers containing a string of Ts or from random hexamers. Neither the use of a double-stranded primer nor the replacement of dCTP with methyl-dCTP seemed to have affected the ability of MMLV to reverse-transcribe. The primary titer we obtained when 10 ng size-selected cDNA was ligated to vector was 6×10^6 cfu. Colony PCR performed on 48 randomly picked clones showed that 47 had cDNA inserts that were at least 500 bp long.

DNA Sequence Analysis

Of the 192 clones in the library attempted in cycle sequencing, successful sequences were obtained for 188 clones (98%). The average read length was 890 bp. Contamination from residual *E. coli* genomic or rat ribosomal RNA and mitochondrial DNA was extremely low

Figure 2. Modified multiple cloning site of the modified pBluescript II KS+ plasmid where a *BsgI* recognition site has been engineered in between *NotI* and *EcoRI*. The junction of cDNA insert ligation to the *BsgI* and *NotI*-restricted plasmid is as indicated.

Short Technical Reports

Table 2. Frequency of Occurrence of Non-A Nucleotides

mRNA	No. of ESTs	Percentage
UUaaaaaaa	11	6.2
GUaaaaaaa	4	2.2
CUaaaaaaa	10	5.6
AUaaaaaaa	9	5.1
U	34	19.1
UGaaaaaaa	18	10.1
GGaaaaaaa	3	1.7
CGaaaaaaa	15	8.4
AGaaaaaaa	11	6.2
G	47	26.4
UCaaaaaaa	17	9.6
GCaaaaaaa	15	8.4
CCaaaaaaa	61	34.4
ACaaaaaaa	4	2.2
C	97	54.5
Total	178	100.0

The frequency of occurrence of the two non-A nucleotides immediately preceding the poly(A) tail is shown. The poly(A) sequence is represented by the italicized lowercase letters in the table.

(Table 1). Six clones had either no insert or insert that was smaller than 50 bp, and one clone did not have the expected flanking sequence representing the trimmed first-strand cDNA synthesis primer. None of the clones had detectable stretches of poly(A/T). The remaining 178 sequences have been deposited at GenBank® under accession nos. BU946518–BU946695. The frequency of occurrence of a non-A nucleotide at the end of the mRNA just before the start of the poly(A) stretch appears to be nonrandom, with over half of the clones sequenced having cytosine as the final nucleotide (Table 2). The rate of occurrence of the second-to-last nucleotide is also shown. While this assessment is useful for first-strand primer design, the true frequency of each nucleotide as the end sequence of mRNAs cannot be determined because the procedures used here do not distinguish naturally occurring adenosine nucleotides at mRNA ends from the subsequently added poly(A) tails. We next determined

Table 3. Putative Hexameric Polyadenylation Signals

Poly(A) Signal	No. of Clones	Percentage
AAUAAA	86	67.2
AUUAAA	21	16.4
AAAAAG	6	4.7
AAUAUA	4	3.1
AGUAAA	4	3.1
AAUACA	3	2.3
AAUAGA	1	0.8
ACUAAA	1	0.8
CAUAAA	1	0.8
UAUAAA	1	0.8
Total	128	100.0

Putative polyadenylation signals present within the last 30 nucleotides on the cDNA are identified. Fifty of the 178 successfully sequenced clones did not have a recognizable hexameric polyadenylation signal.

how many of our sequences represented true 3' mRNA ends by comparison with the available full-length rat mRNA sequences and 3' clone end ESTs contained in Unigene's #106 build (<http://www.ncbi.nlm.nih.gov/UniGene>). Of the 178 successful sequences, 91 had matches in Unigene with a minimum identity of 95% over 300 bp. Of the 91 matching sequences, 69 (75.8%) were either longer (i.e., more 3') or at least within 20 bp of the end of the Unigene entry. A search for polyadenylation signal sequences (11) on the last 30 nucleotides of our cDNA clones was performed. Out of the 178 successful sequences, 128 (71.2%) had polyadenylation signals (Table 3). As expected, the majority of the signals are from "AAUAAA", with the second common signal being "AUUAAA". A similar search for polyadenylation signals on the 51 750 rat 3'-end clone entries in Unigene found a lower frequency of 62%.

DISCUSSION

Here we describe the development of an improved cloning strategy for the construction of cDNA libraries for high-

Short Technical Reports

ly efficient DNA sequencing from the 3' ends of unidirectionally cloned gene transcripts. The key to the success of this strategy is to ensure that annealing of the oligo(dT) primer-adaptor occurred precisely at the junction of the poly(A) tail addition, as downstream priming would result in long homopolymer stretches on the cDNA despite trimming of the oligo (dT) primer. Typical strategies for cDNA synthesis for library construction are based on the use of an un-anchored primer containing a string of Ts (e.g., TT...TTT), or an anchored primer with a single non-T nucleotide at the end (e.g., TT...TTTV), or one with an additional end base (e.g., TT...TTVN). It was recently reported that the widely used un-anchored oligo(dT) primer for cDNA library construction and EST sequencing has generated a high frequency of truncated cDNAs through internal priming (12). Furthermore, as human mRNAs contain

an average of 200 adenosine residues at their 3' ends (13), un-anchored oligo(dT) priming is also likely to generate various lengths of poly(A/T) sequences at the 3' ends of cDNA templates. It has been suggested that hybrids formed between these poly(A) and poly(T) sequences of unrelated cDNA templates lead to the loss of many genes during the normalization/subtraction reactions (14). The use of an anchored oligo(dT) primer, however, will not direct all of the primers to anneal at the beginning of the poly(A) tail. One reason is that the dC-anchored primer apparently does not provide a discriminatory function for the synthesis of poly(A/T)-free cDNA in reverse transcription because of the retroviral enzyme having a high mispairing extension capacity during RNA-dependent DNA synthesis (15–17). This can be improved by the addition of an additional nucleotide to the dC anchor, except in the case of dCdT anchors,

which cannot be improved further because of the nonspecificity of dCdT itself (14). Thus, the use of typical anchored oligo(dT) primers for cDNA synthesis and 3'-end sequencing will lead to the systematic loss of ESTs from mRNAs ending with "AG". Assuming a random distribution of A, G, C, and T in the last and second-to-last nucleotides before poly(A) tail addition, this represents a loss in sequence coverage of 8.3% of all cloned cDNAs. In our random sampling of 178 sequences, the observed frequency of an "AG" terminus was 6.2% (Table 2). Therefore, we would not have gained EST sequences from these clones without the use of our special oligo(dT) primer mixture. Another possibility why stretches of poly(A/T) longer than the number of Ts on the oligo(dT) primer are often found on the 3' end of cDNA clones is because of imperfections (most likely single non-A nucleotides) in the poly(A) tail it-

Short Technical Reports

self (see Figure 1). Indeed, the poly(A) polymerase from at least one organism has been shown to be capable of tailing the 3' end of RNAs with GTP, UTP, and CTP in addition to ATP (18). During the course of library construction and sequencing across multiple organisms, we have noticed that the length of poly(A/T) stretches on cDNA varied from one species to the next, with some organisms being particularly difficult to sequence because of the long homopolymeric tails. We suspect this to be caused by imperfections in the poly(A) tail, so we initially designed primers to contain two Vs at the 3' end. However, we were concerned that the second V would lead to the exclusion of mRNA populations ending with A as the second last nucleotide, and so we mixed in enough Ts at this position to account for all genes that naturally contain a complementary A at this site. Operating under the same assumption, we continue the process additionally for three more rounds until a pool of five primers is designed. This formulation of primer mixture in place of 2-base anchored oligo(dT) primers completely eliminated the presence of these poly(A/T) tails and made it possible to generate accurate sequences for most clones attempted for sequencing. We chose to use a complementary strand to anneal to the *BpmI* recognition site of the primer so that the *BpmI* site that we introduce will not be subjected to methylation during second-strand cDNA synthesis. Furthermore, "covering up" the non-T nucleotides on the primer can only add to the specificity of its annealing to poly(A) mRNA. When our sequences were compared to the full-length mRNAs and 3'-end clone sequences in rat Unigene, we found that the majority (76%) were at least as 3'. Some clones appear to be much longer than the entry in Unigene, while others (24%) appeared to end prematurely. Because of mRNA transcripts generated from alternative polyadenylation and alternative splicing, we could not conclude whether or not these apparently shorter (or longer) cDNAs represented true 3' mRNA ends. We therefore searched for canonical eukaryotic polyadenylation signals within the last 30 nucleotides on the cloned cDNA. Our results indicate that cDNA generated from this primer

are more likely to harbor polyadenylation signals than cDNAs generated from the more traditional and widely used oligo(dT)-V or -VN primers. Our results lead us to conclude that this procedure is a simple and reliable method to clone and generate representative and highly successful EST sequences from the 3' ends of gene transcripts.

REFERENCES

1. Edmonds, M., M.H. Vaughan, Jr., and H. Nakazato. 1971. Polyadenylic acid sequences in the heterogeneous nuclear RNA and rapidly-labeled polyribosomal RNA of HeLa cells: possible evidence for a precursor relationship. *Proc. Natl. Acad. Sci. USA* 68:1336-1340.
2. Odelberg, S.J., R.B. Weiss, A. Hata, and R. White. 1995. Template-switching during DNA synthesis by *Thermus aquaticus* DNA polymerase I. *Nucleic Acids Res.* 23:2049-2057.
3. Thweatt, R., S. Goldstein, and R.J. Shmookler Reis. 1990. A universal primer mixture for sequence determination at the 3' ends of cDNAs. *Anal. Biochem.* 190:314-316.
4. Khan, A.S., A.S. Wilcox, J.A. Hopkins, and J.M. Sikel. 1991. Efficient double stranded sequencing of cDNA clones containing long poly(A) tails using anchored poly(dT) primers. *Nucleic Acids Res.* 19:1715.
5. Takasuga, A., S. Hirotsune, R. Itoh, A. Jito-hzono, H. Suzuki, H. Aso, and Y. Sugimoto. 2001. Establishment of a high throughput EST sequencing system using poly(A) tail-removed cDNA libraries and determination of 36,000 bovine ESTs. *Nucleic Acids Res.* 29:E108.
6. Shibata, Y., P. Carninci, K. Sato, N. Hayatsu, T. Shiraki, Y. Ishii, T. Arakawa, A. Hara, et al. 2001. Removal of polyA tails from full-length cDNA libraries for high-efficiency sequencing. *BioTechniques* 31:1042-1049.
7. Fu, G.K., S. Starnes, and L.L. Stuve, inventors, Incyte Pharmaceuticals, assignee. 2002. Construction of unidirectionally cloned cDNA libraries from messenger RNA for improved 3' end DNA sequencing. *In* United States Patent and Trademark Office. Incyte Pharmaceuticals.
8. Gubler, U. and B.J. Hoffman. 1983. A simple and very efficient method for generating cDNA libraries. *Gene* 25:263-269.
9. Ewing, B., L. Hillier, M.C. Wendt, and P. Green. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* 8:175-185.
10. Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403-410.
11. Beaudoin, E., S. Freier, J.R. Wyatt, J.M. Claverie, and D. Gautheret. 2000. Patterns of variant polyadenylation signal usage in human genes. *Genome Res.* 10:1001-1010.
12. Nam, D.K., S. Lee, G. Zhou, X. Cao, C. Wang, T. Clark, J. Chen, J.D. Rowley, et al. 2002. Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. *Proc. Natl. Acad. Sci. USA* 99:6152-6156.
13. Lewin, B. 1997. *Gene VI*. Oxford University Press, New York.
14. 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.
15. Yu, H. and M.F. Goodman. 1992. Comparison of HIV-1 and avian myeloblastosis virus reverse transcriptase fidelity on RNA and DNA templates. *J. Biol. Chem.* 267:10888-10896.
16. Bakhanashvili, M. and A. Hizi. 1993. The fidelity of the reverse transcriptases of human immunodeficiency viruses and murine leukemia virus, exhibited by the mispair extension frequencies, is sequence dependent and enzyme related. *FEBS Lett.* 319:201-205.
17. Abbotts, J., M. Jaju, and S.H. Wilson. 1991. Thermodynamics of A:G mismatch poly(dG) synthesis by human immunodeficiency virus 1 reverse transcriptase. *J. Biol. Chem.* 266:3937-3943.
18. Martin, G. and W. Keller. 1998. Tailing and 3'-end labeling of RNA with yeast poly(A) polymerase and various nucleotides. *RNA* 4:226-230.

We thank Jonathan Wang and Elizabeth Bosch for helpful comments and Chuck Argentine for DNA sequencing. Address correspondence to Dr. Glenn K. Fu, Director, Molecular Biology, Incyte Genomics, 3160 Porter Drive, Palo Alto, CA 94304, USA. e-mail: gfu@incyte.com

Received 11 December 2002; accepted 27 January 2003.

Glenn K. Fu and Laura L. Stuve
Incyte Genomics
Palo Alto, CA, USA