Evaluation of vector-primed cDNA library production from microgram quantities of total RNA

Jonathan Kuo, Jason Inman, Michael Brownstein and Ted B. Usdin*

Laboratory of Genetics, National Institute of Mental Health Building 35, Room 1B-215, 35 Convent Drive, Bethesda, MD 20892-3728, USA

Received October 6, 2004; Revised November 16, 2004; Accepted November 29, 2004

ABSTRACT
cDNA sequences are important for defining the coding region of genes, and full-length cDNA clones have proven to be useful for investigation of the function of gene products. We produced cDNA libraries containing 3.5–5 × 10^6 primary transformants, starting with 5 μg of total RNA prepared from mouse pituitary, adrenal, thymus, and pineal tissue, using a vector-primed cDNA synthesis method. Of ~1000 clones sequenced, ~20% contained the full open reading frames (ORFs) of known transcripts, based on the presence of the initiating methionine residue codon. The libraries were complex, with 94, 91, 83 and 55% of the clones from the thymus, adrenal, pineal and pituitary libraries, respectively, represented only once. Twenty-five full-length clones, not yet represented in the Mammalian Gene Collection, were identified. Thus, we have produced useful cDNA libraries for the isolation of full-length cDNA clones that are not yet available in the public domain, and demonstrated the utility of a simple method for making high-quality libraries from small amounts of starting material.

INTRODUCTION
Identification and isolation of cDNA clones is important for the interpretation and functional investigation of sequenced genomes. cDNA sequence information facilitates the identification of exons and thus, the definition of the coding potential of genes. cDNA clones provide probes for a variety of studies of genomic expression, and clones that contain full open reading frames (ORFs) allow functional investigation of their protein products. The Mammalian Gene Collection project (MGC) was set up as a public domain program with the goal of providing full-length cDNA sequence information and physical clones for all genes, starting with the human and mouse genomes (1). The Riken Institute (2) and a number of private sector groups also have major programs aimed at collecting full-length clones. Initially, cDNA libraries in the MGC were prepared from a variety of cell lines and tissues. Clones randomly selected from these libraries were end-sequenced and unique clones thought to contain full ORFs were fully sequenced and made available for public distribution. The MGC has been highly successful. As of October 2004, >12,000 human and 10,000 mouse unique, full-length clones were reported to be sequenced and distributed, and an additional 3000 clones from each species evaluated as likely to contain full ORFs, are in the sequencing pipeline (3) (http://mgc.nci.nih.gov/). However, as initially predicted, the yield of novel clones has decreased as the project has moved forward. Genes not represented by selected cDNA clones may have low expression levels and thus, be rare transcripts. Several approaches to isolation of rare clones are being used. One is generation and random sequencing of subtracted cDNA libraries. Another is prediction of mRNA sequences based on genomic sequence, and amplification of these specific sequences from cDNA using the PCR. A third approach is based on the idea, that transcripts that are rare in cDNA libraries prepared from available cell lines and easily obtained tissues, may be present at much greater relative abundance in specific tissues that can only be obtained in small amounts. In the brain, for example, there are well known instances of genes that are highly expressed in a relatively small numbers of neurons in highly discrete regions. For example, there are about 10,000 vasopressin containing cells in the rat paraventricular nucleus (divided between two major cell types) and 7000 in suprachiasmatic nucleus (4,5). A micro dissection that would ultimately yield a very small amount of RNA is required to obtain a sample enriched in these cells. Thus, methods of cDNA library production that are effective with very small amounts of starting material are important to obtain libraries with significant representation of these types of rare transcripts.

Most procedures for generation of cDNA libraries employ an oligonucleotide containing a short poly(dT) sequence as the primer in a reverse transcription reaction that uses purified poly-adenylated RNA as the template. Following a second-strand replacement reaction, synthetic oligonucleotide linkers or adaptors are ligated to the double-stranded cDNAs. These products, in many cases following restriction enzyme digestion, are ligated to an appropriately prepared vector. The adaptor/linker ligation is a relatively inefficient bi-molecular blunt-end ligation. It is performed with an excess of adaptor/linker, which has to be removed from the cDNA product before it can be introduced into the vector. This purification step results in loss of material. The subsequent ligation of the cDNA into vector, even in the case when ‘sticky ends’ are created, is a bimolecular reaction with limited efficiency. Incorporating the oligo-dT priming sequence into the vector to produce a

*To whom correspondence should be addressed. Tel: +1 301 402 6976; Fax: +1 301 402 6976; Email: usdint@mail.nih.gov
‘vector primer’, eliminates one or both of the ligation reactions and the purification step. Okayama (6) described a highly efficient vector-primer based cDNA library construction procedure in 1982. This method was not widely adopted, partly because it is technically demanding. Simpler vector-primer methods have been described previously (7,8), but they are also not widely used. Perhaps this is due to the fact that some technical limitations were not adequately addressed, or that commercial kits using oligonucleotide priming methods became widely available and easy to use. We have returned to a vector-primer method to generate cDNA libraries from very small amounts of tissue. We now present the evaluation of several cDNA libraries prepared from microgram quantities of total RNA.

MATERIALS AND METHODS

Preparation of total RNA

The thymus, adrenals, pituitary, and pineal gland from 20 C57Bl6 male mice were placed in RNALater™ (Ambion, Austin, TX) at the time of dissection and then stored at −80°C. Total RNA from each tissue was isolated using an RNeasy Mini Kit (Qiagen, Valencia CA) and re-suspended in RNase-free H2O.

Preparation of vector primer

To prepare the vector primer, 50 µg of pBluescript II KS vector, modified by the addition of a 700 bp stuffer fragment excised and recovered by phenol extraction. Vector d(T)20 was isolated using an RNase-free Mini Kit (Qiagen, Valencia CA) and re-suspended in RNase-free H2O.

cDNA library construction

Five micrograms of total RNA were heated at 70°C for 10 min, then cooled on ice. The total RNA was added to 20 µl of reaction mix containing 50 ng vector-primer, first-strand synthesis buffer (50 mM Tris–HCl, pH 8.3, 6 mM MgCl2 and 75 mM KCl), 10 mM DTT, 1 mM dNTP and 2 µl PowerScript™ Reverse Transcriptase (BD Biosciences, Palo Alto CA). After 2 h at 42°C the reaction was cooled on ice for 5 min. Second-strand synthesis was accomplished by adding 30 µl of second-strand synthesis buffer [100 mM Tris–HCl, pH 6.9, 450 mM KCl, 23 mM MgCl2, 0.75 mM β-NAD and 50 mM (NH4)2SO4], 3 µl of 10 mM dNTP, 2 U of RNAase H (Invitrogen), 40 U of Escherichia coli DNA polymerase (New England Biolabs), 10 U of E.coli DNA ligase (New England Biolabs) and ddH2O to a final volume of 150 µl. After 2 h incubation at 16°C, the reaction was stopped by adding 10 µl of 0.5 M EDTA. The reaction was then extracted with phenol; the upper aqueous layer was precipitated with ethanol after addition of one half volume of 7.5 M NH4OAc and 20 µg of glycogen (Roche) at −20°C overnight. The vector-cDNA pellets were recovered by centrifugation at 4°C for 30 min, washed twice with 70% ethanol, air dried and dissolved in water. The vector-cDNA was circularized in a final volume of 100 µl containing 800 U of T4 DNA ligase (New England Biolabs defined units), 1 mM ATP and ligase buffer (50 mM Tris–HCl, pH 7.5, 10 mM MgCl2 and 10 mM DTT) for 16 h at 15°C.

Bacterial transformation by electroporation

The ligation product was extracted with phenol:chloroform: isoamyl alcohol; 20 µg of glycogen carrier and a half volume of 7.5 M NH4OAc were added to the aqueous phase, and the cDNA was recovered by ethanol precipitation. After extensive washing with 75% ethanol and air drying, the cDNA was dissolved in 10 µl of H2O. Electroporation was performed in a Bio-Rad apparatus at 2.5 kV, 25 µF, 200 Ω, using ElectroMAX™ DH10B™ T1 Phase-Resistant (Invitrogen) competent cell and 5 µl of cDNA. Bacteria were grown in 1 ml of SOC media for 1 h at 37°C following transformation; dilutions were grown on ampicillin containing plates. The libraries were amplified by a solid-state amplification method (9); aliquots were frozen and plasmid DNA was purified from the remainder.

cDNA library size selection

Two µg of plasmid DNA was digested with 100 U of SstI (Invitrogen) for 2 h at 37°C in 100 µl. The reaction was stopped with gel-loading buffer, and loaded onto a 0.7% low-melting-temperature agarose gel (Invitrogen) adjacent to a lane containing a DNA ladder and run at 30 V for 16 h. DNA fragments >4 kb were visualized using a hand-held long-wavelength ultraviolet lamp, and cut from the agarose gel with a scalpel. The DNA was recovered from the agarose gel by phenol extraction, and recircularized in a final volume of 50 µl, containing 400 U of T4 DNA ligase (New England Biolabs defined units), 1 mM ATP and ligase buffer (50 mM Tris–HCl, pH 7.5, 10 mM MgCl2 and 10 mM DTT) for 16 h at 15°C. The DNA was then phenol extracted, ethanol precipitated and used to transform bacteria by electroporation as described above.
Colony PCR
A small portion of a single bacterial colony was picked with a sterile toothpick and transferred to 10 μl of reaction mix containing M13 forward and reverse primers and 0.5× Advantage 2 Polymerase Mix (BD Biosciences) and buffer supplied with the enzyme. Following incubation at 94°C for 2 min, 30 cycles at 94°C for 30 s, 55°C for 30 s and 72°C for 3 min, were followed by 72°C for 5 min. PCR products were separated on a 1% agarose gel with a DNA size marker.

Sequencing
Individual clones were picked into deep well blocks containing 100 μl of LB with 8% glycerol and 50 μg/ml carbenicillin, and grown overnight at 37°C without shaking. Aliquots from these stocks were grown in 1.2 ml of LB-carbenicillin overnight at 37°C, shaking at 320 r.p.m., and plasmid DNA prepared using a Qiagen 3000 Biorobot with the QIAprep 96 Turbo Miniprep kit; the remaining material was stored at −20°C. Plasmids were sequenced by the National Institute of Neurological Diseases Intramural sequencing core facility. Sequencing was performed with an M13 forward primer to provide 5′ reads.

Sequence analysis
These sequences were loaded into Sequencher (Gene Codes Corporation, Ann Arbor, MI) and stripped of vector sequences. Sequences that were left with zero bases were flagged as ‘Vector’ sequences. The ends of the sequences were then trimmed until the first and last 25 bases contained fewer than three ambiguities each to facilitate subsequent sequence alignment. Sequences left with zero bases at this point were flagged as ‘Low Quality’. This also allowed Sequencher to identify more vector regions, so these two trim procedures were repeated until no more bases were trimmed from the sequence. BLAST was next performed locally using standalone NCBI blast 2.2.5 for MacOSX (executables available at ftp://ftp.ncbi.nlm.nih.gov/blast/executables/). Searches were made against a database created by running ‘formatdb’ to generate an Entrez criteria: (i) Organism = ‘mus musculus’, (ii) Limit: Molecule = mRNA, (iii) Limit: OnlyFrom = RefSeq. This retrieved 26,356 sequences. The RefSeq IDs were then submitted to Stanford University’s SOURCE (http://source.stanford.edu) for annotation information including Symbol, Description and Unigene IDs and other information. Sequences (trimmed and untrimmed) and annotations were loaded into a database created in FileMaker Pro 6, which contained custom scripts for sending various BLAT and BLAST requests. Annotations were verified using UCSC BLAT against the October 2003 mouse genome (UCSC Genome Browser: http://genome.ucsc.edu/). At the same time, alignments were visually inspected to verify the presence or absence of initiation codons corresponding to the CDS of the ORF at the annotated locus. In some cases, the untrimmed sequences were also checked by alignment, as the trimmed ends might contain the start codon but were eliminated due to poor base calling. Sequences that did not produce satisfactory alignments to ORFs were analyzed by BLAST against the nr database.

RESULTS
A vector-primer for cDNA synthesis was generated by ligating an oligo(dT) containing sequence to the XhoI site within pBluescript vector, essentially as described previously (8). Following pilot experiments, which suggested that complex libraries could be made from relatively small amounts of starting mRNA, we attempted to generate libraries starting with 5 μg of total RNA. We selected four mouse tissues, pituitary gland, adrenal gland, thymus and pineal gland, that are relatively small, easy to dissect, and likely to selectively express genes that are absent or much lower in abundance in other tissues. Following a vector-primer initiated reverse transcription reaction, a standard second-strand replacement reaction was performed using RNase H to nick the DNA and E.coli DNA polymerase I to synthesize the second strand of the cDNA. This material was then self-ligated to form a closed circle and aliquots were used to transform bacteria by electroporation (Figure 1). Libraries with 9.2 × 10^5 total primary transformants were created from the pituitary, 6.3 × 10^5 from adrenal, 6 × 10^5 from thymus and 7.5 × 10^5 from pineal tissue using electro-competent bacteria with a nominal transformation efficiency of 1 × 10^10 μg^-1. PCR reactions were performed using primers flanking the cDNA insert site on ~40 individual colonies from each library. A range of insert sizes from 200 bp, the size expected for a clone with no insert, to >3 kb was observed for each library. Empty clones, based on this PCR analysis represented 12, 43, 42 and 47% for the pituitary, adrenal, thymus and pineal libraries, respectively. When the number of empty clones in each library is subtracted, the number of insert containing clones was 7.6, 3.4, 3.5 and 3.5 × 10^5 for these libraries. The average insert size, excluding empty clones, was 1.2 kb. To reduce the fraction of empty clones the adrenal, thymus and pineal libraries were size-selected following amplification in semi-solid agarose, as described in Materials and Methods. Following size-selection, empty clones based on PCR analysis represented 17, 15 and 18% of the clones for the adrenal, thymus and pineal libraries. A flow chart illustrates the steps followed (Figure 2).

To evaluate the libraries further, 270 clones were picked from each library for 5′ end sequencing. Table 1 contains the analysis of these libraries. Of the 1079 clones submitted for sequencing, 173 (16%) produced ‘Low Quality’ reads (see the ‘Sequence analysis’ section in Materials and Methods). These clones were excluded from further analysis. Approximately 20% of the remaining clones were empty vectors, which is consistent with the estimate from PCR. To evaluate the success in producing clones with full ORFs, the remaining sequences were compared with annotated mouse genomic and cDNA sequences. Nearly half of all the clones sequenced could be identified this way, and about half of these had initiating methionine codons. Overall, 56% of the clones that we sequenced could be aligned to annotated sequences, and of these, 34% also contained an identifiable initiating methionine residue codon. Thus, the net percentage of ‘full-length’ clones was 19%, and it did not vary greatly between the libraries (15, 19, 22 and 21%). The majority of the remaining clones were of partial length (identifiable but lacking the initiating methionine codon—37%), with ribosomal RNA, mitochondrial genomic DNA, clones containing intronic sequence (and thus, probably representing priming on genomic DNA
or reverse-transcription of unspliced heteronuclear RNA), and difficult-to-characterize sequences representing the remainder. The libraries appeared to reflect the expected tissue expression profile. For example, in the pituitary library, of the 230 clones analyzed, growth hormone was identified 35 times, prolactin 20 times and pro-opiomelanocortin 5 times, and these three transcripts were not present in any of the other libraries. The libraries also appeared diverse: 94, 91, 83 and 55% of the clones identified as mRNAs were represented only once in the clones examined from the thymus, adrenal, pineal and pituitary libraries, respectively (Table 2). We also evaluated the potential contribution of these libraries to the MGC project.

Of the 906 clones evaluated, 84 corresponded to transcripts that were not represented in the MGC full-length clone set and 25 of these were unique full-length clones.

**DISCUSSION**

We have established a method that can produce complex, high-quality cDNA libraries from microgram quantities of total RNA. If a random sequencing strategy is used, the number of clones produced is many more than can be analyzed. The amount of starting material required is less than other methods that do not use an amplification procedure such...
as PCR. We started with 5 μg of total RNA. Purification of polyadenylated RNA from total RNA typically yields 1% of the input mass (10). Thus, our method used the equivalent of ∼50 ng of poly(A)+ RNA to produce several hundred thousand independent recombinants, and less starting material could probably be used to make a satisfactory library. While the number of clones sequenced from each library may limit the conclusions that can be drawn, the analysis that we have been able to perform, suggests that the libraries are both representational and diverse. The thymus and adrenal libraries, which were made from the most complex tissues, had >90% unique (i.e. different, single-copy) clones. In the pituitary library, which was prepared from a tissue with a relatively small number of highly specialized cell types, 55% of the clones were unique and the redundant clones represent abundant, organ-specific transcripts, as one would expect from a specialized secretory organ. Thus, the size and complexity of the libraries suggests that they are likely to be useful for a variety of applications that require good representation of the entire repertoire of genes expressed in a tissue.

Other investigators have reported production of high-quality cDNA libraries from small amounts of starting material [e.g. (11)], but it is difficult to accurately compare the libraries we made, with those described by others, because widely different analysis parameters are reported. In a recent paper on the identification of putative full-length cDNAs for 70% of Drosophila genes (12), an estimate of 84% ‘high-quality’ reads was given, which probably corresponds to our total number of reads minus those identified as vector only and those with low quality reads—66%. The fraction of full-length clones and the amount of starting material was not reported in that study, but several workers have reported full-length percentages >50%. These investigators have generally started with relatively large quantities of poly(A)+ RNA (13) and/or used a PCR step (14–16), in addition to a procedure that selects clones based on the 5′ cap structure of mRNAs. It has been reported that PCR
severely restricts the complexity of libraries produced (13), and the introduction of polymerase-induced errors has not been analyzed for these procedures. It is not entirely clear how these reports of ‘50% full-length’ compare with our overall value of 20%, because a number of clones, such as those containing no inserts or obvious contaminants, may have been stripped out before the full-length estimate was made. It is possible that our libraries contain a greater number of empty vector and ‘garbage’ containing clones than some other libraries, but our libraries can be prepared from material that would not be of sufficient quality or quality for the other methods. In fact, simply by analyzing ~1000 clones from these libraries, we identified 25 full-length sequences not currently reported to be in the MGC. The MGC website contains statistics on a number of libraries considered to be of sufficient quality for in-depth sequencing. We compared two of these with our own libraries: MGC_166, which was made from 120 mouse pituitaries, and Soares_thymus_2NbMT. From 7167 reads, the pituitary library yielded 27 unique full-length clones (0.3%) and the thymus library yielded 131 unique full-length clones from 49443 reads (0.26%). Thus, based on the moderate number of reads that we have performed, it appears that the libraries that we have made may be able to make a significant contribution to the available clone collection.

We chose the vector-primer strategy so that we could increase the frequency of clones for transcripts that have restricted tissue distributions. The question of whether these libraries contain other types of sequences that may be under-represented in conventional cDNA libraries also arises. One class of under-represented clones consists of short sequences, such as those encoding some peptides, which are removed in many size selection procedures. Size selection is not a requirement of our vector-primer method. It was performed to reduce the number of empty clones in three of the four libraries analyzed; but, depending on a cost/benefit analysis based on the screening technique to be used, it could be eliminated. It has been suggested that the representation of long mRNAs in a library may be significantly influenced by their damage during purification of the RNA and handling. RNA handling steps are dramatically reduced in our method, by eliminating the oligo(dT) chromatography steps commonly used to purify mRNA. Our method probably does not significantly differ from other cDNA library synthesis methods in regard to the representation of clones that are reverse-transcribed with low efficiency, but it is worth noting that it should be compatible with a variety of reverse transcriptase enzymes and denaturation or modification conditions. The reverse transcriptase used for the libraries presented here was selected because in our experience it produces high yields and because the manufacturer does not make any ‘pass through’ claims on the libraries and clones produced.

We used the vector-primer method because it should be very efficient from a theoretical view, and our data support this. It differs from the Okayma’s method (6,17) principally by ligation of an oligonucleotide containing the dT ‘tail’ to the vector instead of using terminal transferase to add an oligo(dT) ‘tail’ to the vector, and by a simple self-ligation reaction to circularize the product instead of a second terminal transferase reaction, enzyme digestion and ligation of a joining fragment. Thus, it is much simpler to perform. A procedure very similar to the one we used has been previously described (7,8), but its use was limited by the presence of an unidentified exonuclease activity; it was not applied to total RNA, and libraries produced with it were not analyzed by sequencing. We used pBluescript as the scaffold for this procedure, but it should be equally applicable to almost any cloning vector.

Another major difference between the vector-primer method described and standard library construction methods is size selection of the cDNA. Some type of purification is required to separate adaptor/linkers from the cDNA in standard methods and this is often used to simultaneously perform size selection. This requires that sufficient material be used to overcome losses in the purification method and to detect the cDNA of interest. Removing adaptor/linkers is not required with vector-priming and we did not have to size-select the cDNA before inserting it into the vector. This is an inefficient step that results in significant losses in cDNA. The size-selection (to reduce the number of empty clones) that we did perform was done on DNA isolated from libraries that had already been created and expanded. To do the latter, we used a semi-solid amplification technique that is thought to minimize the skewing of clone distributions of the libraries. Linearization was done with SstI in these trial libraries, but we recommend that it should be done with an enzyme with a much less common recognition site. We believe that the number of empty clones can be significantly reduced. This may require empirical optimization of the vector to RNA ratio for each sample. With very small quantities of a particular RNA sample available, it may be more ‘cost-efficient’ to perform the post-synthesis remediation.

We have described a cDNA library construction procedure that is simple to perform and highly efficient, and have presented analysis which suggests that the libraries have properties that makes them useful for a variety of applications. The procedure can be used exactly as described, but it is also highly amenable to further modifications. Additional characterization and mining of libraries produced by the method will ultimately define its usefulness.

**ACKNOWLEDGEMENTS**

We thank Jim Nagle and Debbie Kaufman of the National Institute of Neurological Diseases intramural nucleic acids core sequencing facility for their help. This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. NO1-CO-12400. The content of this publication does not necessarily reflect the views or policies of the Department of

---

**Table 2.** Estimate of the number of unique clones in each library mRNAs is the sum of the full- and partial-length sequences identified as mRNA.

<table>
<thead>
<tr>
<th>Library</th>
<th>mRNAs</th>
<th>Unique</th>
<th>Unique (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thymus</td>
<td>119</td>
<td>112</td>
<td>94</td>
</tr>
<tr>
<td>Adrenal</td>
<td>115</td>
<td>105</td>
<td>91</td>
</tr>
<tr>
<td>Pineal</td>
<td>130</td>
<td>108</td>
<td>83</td>
</tr>
<tr>
<td>Pituitary</td>
<td>146</td>
<td>81</td>
<td>55</td>
</tr>
<tr>
<td>Total</td>
<td>510</td>
<td>406</td>
<td>80</td>
</tr>
</tbody>
</table>

The values in the ‘unique’ column represent the number of different LocusLink/Gene IDs found among the set of mRNAs in each library.
Health and Human Services, nor does mention of trade names, commercial products, or organization imply endorsement by the US Government.

REFERENCES


