Review

The current status of cDNA cloning

Matthias Harbers

DNAFORM, Inc., Leading Venture Plaza 2, 75-1 Ono-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0046, Japan

Received 17 August 2007; accepted 17 November 2007

Available online 28 January 2008

Abstract

The cloning of cDNAs, copies of cellular RNA, is one of the classical technologies in molecular biology. Over the past 30 years cDNA cloning technologies have been improved to enable the cloning of large cDNA collections, which are fundamental to today’s understanding of the utilization of genetic information. With the discovery of non-coding RNAs, additional new approaches to the cloning of short RNAs have been developed. However, with the realization that much larger portions of genomes are transcribed than anticipated from genome annotations, cDNA cloning faces new challenges to uncover rare transcripts and to make the corresponding cDNAs available for functional studies. This review provides an overview on the current status of cDNA cloning and possibilities for the discovery and characterization of new RNA families.

© 2007 Elsevier Inc. All rights reserved.

Keywords: cDNA cloning; cDNA library; mRNA; Small RNA; Non-coding RNA; Expression cloning

Contents

Introduction.................................................................................................................. 232
Reverse transcriptases and first-strand cDNA synthesis.............................................. 233
Second-strand cDNA synthesis .................................................................................. 233
Cloning vectors .......................................................................................................... 234
Single-cell cDNA library preparation ......................................................................... 234
Approaches to full-length cDNA cloning .................................................................. 234
Normalized and subtracted cDNA libraries ................................................................ 234
Addressing RNA splicing ............................................................................................ 235
Large-scale cDNA cloning projects and clone collections ............................................ 235
Linking cloning to functional analysis ......................................................................... 235
The emerging new RNA world .................................................................................... 236
Cloning small RNA ..................................................................................................... 236
Cloning long ncRNA ................................................................................................... 237
Future perspectives and developments ....................................................................... 237
Acknowledgments ....................................................................................................... 239
References .................................................................................................................. 240

Introduction

cDNA cloning is one of the fundamental technologies in molecular biology, and most of our knowledge about transcripts and proteins is derived from the ability to prepare cDNA copies from RNA and to clone them into cDNA libraries. Starting with the discovery of reverse transcriptases, different protocols for cDNA library construction have been developed over time. Improvements in library preparation have been instrumental to gene discovery and the creation of large genomic resources. Recent discoveries of new classes of RNA and transcripts expressed at very low levels demand new cDNA cloning
approaches to make such RNAs available for functional analysis. Although it is beyond the scope of this article to review all the technical developments of the past 30 years, key steps in cDNA library preparation are addressed to highlight general principles of cDNA cloning (Fig. 1) and to give an overview on the current status of cDNA cloning and future directions.

**Reverse transcriptases and first-strand cDNA synthesis**

The enzymatic conversion of RNA into double-stranded cDNA has become routine since the discovery of reverse transcriptases in 1970 [1,2], and improved conditions for cDNA synthesis became available by 1975 [3]. Reverse transcriptases are RNA- and DNA-dependent DNA polymerases that can use either RNA or DNA to prime DNA synthesis. Commercial preparations of the avian leukemia virus and Moloney strain murine leukemia virus (Mo-MLV) reverse transcriptase are commonly used today, and removal of the RNase H activity from Mo-MLV reverse transcriptases further improved cDNA yields [4]. Moreover, addition of T4 bacteriophage gene 32 protein (T4gp32) can boost the synthesis of long cDNAs [5] and trehalose increases enzyme fidelity and enables cDNA synthesis at higher temperatures [6,7].

cDNA library preparation has mostly focused on cloning of mRNAs, for which generally oligo(dT) primers (dT12–18) are used to initiate cDNA synthesis from poly(A) tails at the 3’ end. Although posttranscriptional addition of poly(A) tails is restricted to RNA polymerase II-derived mRNA transcripts, oligo(dT) priming can also occur at internal A-rich sequences, including RNA polymerase III-transcribed Alu repeats [8]. It has been estimated that some 10 to 15% of the cDNA clones within oligo(dT)-primed libraries could have truncated 3’ ends due to misannealing of oligo(dT) primers [9]; this is particularly a problem for cDNAs derived from very long messages [10,11]. Alternatively random primers of 6 to 9 nucleotides can be used to drive reverse transcription reactions [12,13]. While this is sometimes useful to reach the 5’ end of very long transcripts and frequently used for analytical purposes, this approach does not allow full-length cDNA cloning. For later digestion and directional cloning of cDNAs, recognition sites for restriction endonucleases can be introduced at the 3’ end of cDNA using bifunctional primers comprising oligo(dT) and linker regions [14]. During first-strand synthesis cDNA can be further protected against digestion by methylation-sensitive enzymes by introducing 5-methylcytosine to create hemimethylated DNA [15].

**Second-strand cDNA synthesis**

Synthesis of the second cDNA strand requires a priming site at the 5’ end of cDNA. Originally, hairpin structures at the 3’ end...
of single-stranded cDNAs were used in self-priming reactions, and second strands were synthesized by the DNA polymerase activity of the reverse transcriptase followed by S1 nuclease digestion to remove the hairpin structures. However, self-priming is a poorly controlled reaction and the required S1 nuclease treatment removes 5′ end sequences from double-stranded cDNA. Various alternative approaches for priming second-strand synthesis have been developed over time such as the addition of homopolymers to the 3′ end of single-stranded cDNA [16,17], replacement synthesis after nicking the RNA in RNA/cDNA hybrids with RNase H [18], ligation of an RNA oligonucleotide to the 5′ end of RNA prior to the reverse transcription reaction [19,20], or ligation of double-stranded adaptors to the 3′ end of single-stranded cDNA [21]. For full-length cDNA cloning, the addition of a priming site should not affect the 5′ end of double-stranded cDNA and allows for directional cloning by introducing a restriction endonuclease recognition site at the 5′ end that is distinct from the recognition site used at the 3′ end. Second-strand cDNA synthesis can be combined with PCR amplification to clone cDNAs from small amounts of RNA. However, PCR amplification is biased against longer cDNAs and templates present at low concentrations [22] and is not recommended for general library preparation.

Cloning vectors

For propagation (commonly in *Escherichia coli* [23]), double-stranded cDNAs are cloned into a plasmid [24] or bacteriophage vector. For longer cDNAs plasmid libraries are difficult to preserve, whereas *in vitro* packaging into bacteriophage λ, like the classical expression vector λgt11 [25], Lambda ZAP [26], or Lambda-FLC [27], allows for a wider cloning range, higher titers, and safe long-term storage. Due to limitations in preparing λ DNA, phages with automatic subcloning vectors have been developed [26,27]. Different vectors contain asymmetric cloning sites for directional cloning, unique restriction or recombination sites for releasing entire cDNA inserts, background cutting to reduce the number of empty vectors in the library, and special promoter features for protein expression. A wide range of dedicated cloning vectors [28] is commercially available along with approaches to library screening, e.g., by random clone picking, use of antibodies, hybridization, or PCR.

Single-cell cDNA library preparation

Important advances have been made in the cloning of cDNAs from very small amounts of RNA or even a single cell [29,30] to address, for instance, the zonal expression of transcripts (a single mammalian cell contains about 20 to 40 pg total RNA including 0.5 to 1.0 pg mRNA [31]). These developments and new technologies for the isolation of individual cells by laser capture microdissection [32,33] or cell aspiration after microinjection enable the analysis of genes expressed in specific cells from heterogeneous tissues [34]. Apart from PCR amplification, novel approaches have been developed to amplify RNA directly in cells [35] by preparing antisense RNA [36]. cDNA synthesis from whole RNA is primed by an oligonucleotide containing a T7 RNA polymerase promoter, and after second-strand cDNA synthesis, T7 RNA polymerase can be used to generate antisense RNA from the cDNA. Since multiple RNA copies are obtained from a single cDNA template, the method allows for linear amplification of RNA. Modifications of the procedure have been published to enable full-length cDNA cloning after cDNA tailing by a terminal transferase [37]. An alternative approach to single-cell cDNA library preparation makes use of oligo(dT) primers linked to magnetic beads to perform reverse transcription reactions and PCR [22] and to handle small amounts of cDNA with a reduced risk of losing material. Recently a protocol for generating cDNA libraries from 1 ng of total RNA that performs all reactions on oligo(dT) magnetic beads was published [38]. It introduces a T7 RNA polymerase promoter sequence for amplification and generates double-stranded DNA by a modified switching mechanism at the 5′ end of RNA to enable PCR amplification and cloning into a vector. However, single-cell amplification reactions are limited in their reproducibility, where PCR amplification may be more reliable than linear amplification [39]. Although the necessary amplification makes such libraries very biased, these approaches still open up new prospects in tissue- or cell-specific gene regulation, such as tumor marker discovery in difficult to classify poorly differentiated cancers [40].

Approaches to full-length cDNA cloning

Most important for effective cloning approaches is the production of full-length cDNAs (cDNAs having an open reading frame or ORF) at a high rate for functional analysis of encoded proteins and for information on true 5′ ends of cDNAs to identify promoter regions in the genome. Various approaches make use of the 5′-end-specific cap structure of mRNA to enrich for full-length cDNAs, achieving full-length rates in the range of 90% or above [41–43]. The largest cDNA collections made so far used the cap-trapper [44–47] and oligo-capping [19,20,48] methods. In the cap-trapping method the cap structure is chemically biotinylated prior to selection of full-length mRNA/cDNA hybrids on streptavidin-coated beads, while in the oligo-capping process the cap structure is replaced by an RNA oligonucleotide prior to first-strand cDNA synthesis. Other approaches include a cap-binding protein [49], an antibody against the cap structure [50], and adding an oligonucleotide to the cap structure (U.S. Patent 6,022,715) or are based on a cap-switch mechanism [51].

Normalized and subtracted cDNA libraries

In addition to the full-length cDNAs, large gene discovery programs require special cloning strategies to reduce redundancy within libraries and final clone collections and to avoid overrepresentation of housekeeping genes [52,53]. Moreover, to reduce costs, projects have focused on the cloning of one representative full-length cDNA clone per gene. Although selection criteria for full-length cDNA sequencing varied between different projects, the enrichment for new cDNA clones was preferably done at the level of cDNA library construction,
keeping the cost of sequencing randomly isolated clones to a minimum [52]. The high variation in cDNA abundance within libraries can be reduced (normalized) using time-limited re-association kinetics [54,55]. Since the most abundant cDNA species hybridize faster than rare ones, the double-stranded hybrids formed by abundant transcripts can be removed from the remaining single-stranded cDNAs of less abundant transcripts. Alternative approaches use a double-strand DNA-specific endonuclease (duplex-specific nuclease), to digest DNA/DNA hybrids or the DNA portion within RNA/DNA hybrids [56], or RNase H to destroy the RNA portion in RNA/DNA hybrids (U.S. Patent 6,544,741 and [57]). In addition to normalization, known cDNAs can be removed from cDNA libraries in a subtraction step for higher discovery rates [9,52,58]. Libraries can also be enriched for short or long cDNAs by size fractionation or by removing undesired sequences by subtraction to clone differentially expressed genes [59].

Addressing RNA splicing

Most cDNA cloning projects have largely ignored alternative splice variants and focused on representative clones. This is important to note as it has been estimated that 65% of all mammalian transcripts might be alternatively spliced [45], including important regulators related directly to human disease [60]. Splicing could explain the increased complexity of higher organisms, although our present knowledge on splice variants is insufficient to show an increase in splicing with developmental complexity. New approaches have been developed to monitor alternative exon usage in different samples (U.S. Patent 6,251,590 and [61,62]). The methods of Watahiki and Thill form DNA–DNA hybrids, in which alternative exons loop out as regions of single-stranded DNA surrounded by regions of double-stranded DNA. Molecules with single-stranded DNA regions are then isolated by a single-strand DNA binding molecule and cloned for sequence analysis. This identifies individual exons, but a modified approach allows the isolation of full-length splice variants (Patent Application WO2005108608). Selective cloning of tissue-specific splice variants will be important in the future use of cDNA libraries along with progress in full-length sequencing by new sequencing technologies (see below). The majority of mammalian genes probably use both alternative splicing and transcription from multiple start sites [63,64], and studies will further explore relationships between splicing and alternative promoter usage [65,66].

Large-scale cDNA cloning projects and clone collections

Current achievements in cDNA library preparation are marked by the success of large-scale cDNA cloning projects such as the IMAGE Consortium [67]; the Mammalian Gene Collection [68]; Drosophila melanogaster [69], human [48], rice [46], and Arabidopsis [47]; and the RIKEN mouse FANTOM projects [45]. Initially cDNA libraries were prepared to sequence expressed sequence tags (ESTs) for cataloging transcripts on a genome-wide scale [70], followed by large-scale projects at Washington University and commercial entities like Incyte and Human Genome Sciences (a list of EST projects can be found at http://image.llnl.gov/image/html/projects.shtml). The cDNA libraries produced in these efforts focused on high-throughput EST sequencing, commonly from 3′ ends, but were insufficient for preparing large clone collections due to the small insert sizes and low full-length rates [71]. Progress in full-length cDNA cloning in combination with normalization and subtraction techniques allowed for the preparation of comprehensive cDNA collections, though large cDNA collections are still available only for some model organisms due to the high cost of full-length cDNA sequence; refer to Table 1 for more information on cDNA collections in the public domain.

Large cDNA clone collections are also one of the starting points for establishing ORF clone resources from human [72] or Caenorhabditis elegans [73]. These clone collections comprise sequence-validated master clones in entry vectors that allow for direct transfer of ORFs to a broad range of expression vectors [74,75]. Most commonly, ORF collections make use of site-specific recombination cloning systems for easier large-scale manipulation of cDNA inserts [76]. It is hoped that ORF resources will support functional studies on protein-coding genes using arrayed clone sets in highly parallel experiments under controlled conditions.

Linking cloning to functional analysis

Use of ORF clone collections and other genomic resources in functional studies requires further progress in the development of new screening platforms based on biochemical assays suitable to match large cDNA clone collections [77,78]. With an increasing number of functional assays available, genetic screens can now be performed in mammalian cell cultures by selecting against cellular activities like apoptosis, senescence, differentiation, or oncogenic transformation [79]. Screening assays have also benefited from the development of RNAi libraries for gene

Table 1

<table>
<thead>
<tr>
<th>Clone resource</th>
<th>Home page</th>
<th>Reference/comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMAGE Consortium</td>
<td><a href="http://image.llnl.gov/">http://image.llnl.gov/</a></td>
<td>[67]</td>
</tr>
<tr>
<td>Mammalian Gene Collection (MGC)</td>
<td><a href="http://mgc.nci.nih.gov/">http://mgc.nci.nih.gov/</a></td>
<td>[68]</td>
</tr>
<tr>
<td>FANTOM–mouse cDNA collection</td>
<td><a href="http://fantom3.gsc.riken.go.jp/">http://fantom3.gsc.riken.go.jp/</a></td>
<td>[45]</td>
</tr>
<tr>
<td>Rat EST project</td>
<td><a href="http://ratest.eng.uiowa.edu/">http://ratest.eng.uiowa.edu/</a></td>
<td>[53]</td>
</tr>
<tr>
<td>Xenopus Gene Collection (human)</td>
<td><a href="http://xgc.nci.nih.gov/">http://xgc.nci.nih.gov/</a></td>
<td>Subproject to MGC</td>
</tr>
<tr>
<td>Zebrafish Gene Collection</td>
<td><a href="http://zgc.nci.nih.gov/">http://zgc.nci.nih.gov/</a></td>
<td>Subproject to MGC</td>
</tr>
<tr>
<td>Drosophila Gene Collection (Berkeley)</td>
<td><a href="http://www.fruitfly.org/DGC/index.html">http://www.fruitfly.org/DGC/index.html</a></td>
<td>[69]</td>
</tr>
<tr>
<td>Rice full-length cDNA consortium</td>
<td><a href="http://cdna01.dna.affrc.go.jp/cDNA/and">http://cdna01.dna.affrc.go.jp/cDNA/and</a></td>
<td>[46]</td>
</tr>
<tr>
<td></td>
<td><a href="http://www.rgrc.dna.affrc.go.jp/index.html.en">http://www.rgrc.dna.affrc.go.jp/index.html.en</a></td>
<td></td>
</tr>
<tr>
<td>Full-length Arabidopsis cDNA collection</td>
<td><a href="http://www.brc.riken.jp/lab/epd/catalog/cdnaclone.html">http://www.brc.riken.jp/lab/epd/catalog/cdnaclone.html</a></td>
<td>[47]</td>
</tr>
<tr>
<td>ORFeome Collaboration (human)</td>
<td><a href="http://www.orfeomecollaboration.org/">http://www.orfeomecollaboration.org/</a></td>
<td>[72]</td>
</tr>
<tr>
<td>C. elegans ORFeome</td>
<td><a href="http://wormdb.dfcii.harvard.edu/">http://wormdb.dfcii.harvard.edu/</a></td>
<td>[73]</td>
</tr>
</tbody>
</table>
inactivation [80–82], where protocols for converting cDNAs into siDNAs by MmeI digestion and cloning into expression vectors are available [83,84]; approaches for generation of RNAi libraries have recently been reviewed [85]. Although much attention is now focusing on loss-of-function studies using RNAi resources [86–88], these must be complemented by gain-of-function studies based on cDNA and/or ORF resources to reduce the inherent rates of false positive and negative results in screening assays [89].

Expression libraries can be used as an alternative to cDNA/ORF collections as a starting point [90–93]. Expression cloning has thus far focused on protein coding transcripts to identify specific cDNAs by their biological activity in a screening assay. Such assays can include changes in cell behavior like cell death or cell survival, changes in the expression of endogenous or reporter proteins, or direct binding to exposed polypeptides in phage display [94,95], baculovirus display [96], and ribosome display experiments [97]. Phage display was originally limited in directly screening cDNA libraries due to the need to fuse cDNAs to the N-terminus of pIII and pVIII phage proteins and the lack of posttranslational protein modification. In part, these problems have been addressed in the pJuFo system linking the protein in question to a phage protein via the high-affinity interaction of Jun and Fos leucine zippers [98] and the use of baculovirus-infected insect cells [96]. However, expression libraries often suffer from 5’ and 3’ untranslated regions in cDNAs hampering their expression or translation, wrong orientation of cDNA inserts, undefined reading frames, or the use of partial cDNA fragments, e.g., in two-hybrid screens to enable production of fusion proteins [99–102]. These limitations emphasize the need for ORF cloning, where ORFs currently have to be cloned individually by PCR. Direct cloning of cDNA libraries comprising only ORF regions has not yet been achieved, although first efforts have been made by selecting expressing cDNA clones in a yeast system [103]. Such a selection system could be of great interest to distinguish experimentally between coding and noncoding RNAs based on their ability to translate in vivo or in vitro (see below).

The emerging new RNA world

Although conducted for some 15 years, large-scale cDNA cloning projects have not yet revealed all transcripts [43,104]. On the contrary, recent publications indicate that much larger portions of genomes are actually transcribed than previously estimated from whole genome annotations [105–108]. This challenges the classical view of “isolated genes” surrounded by nontranscribed regions; in particular, overlapping sense–antisense pairs seem to be a common feature of complex genomes [109–111]. It has even been suggested that all nonrepeat portions of the human genome could be transcribed [112]; similarly, 85% of the yeast genome is expressed [113]. Sometimes referred to as “TUF” (transcripts of unknown function [107]) or “dark matter” [114], there is widespread low-level expression of potentially noncoding transcripts [115–119]. It remains to distinguish between “meaningful transcripts” and “transcriptional noise” [120]. However, more and more studies suggest that noncoding RNAs (ncRNAs) play central roles in gene expression and genome organization [121]. The lower conservation found for many noncoding transcripts argues for their evolutionary importance since such transcripts can change faster during evolution than coding transcripts.

Cloning small RNA

Recently short noncoding RNAs have become a key focus in research, emphasizing the great importance of new RNA families. Such RNA molecules include miRNAs and their precursors [122,123], snoRNAs [124], rasiRNAs [125,126], piRNAs [127], and small regulatory RNAs. The discovery of this new “RNA world” revealed that standard cDNA libraries missed many transcripts, notably those of short length (commonly under 500 bp) or lacking poly(A) tails [107,108,128,129]. New approaches are being developed for targeted cloning of short ncRNAs [130–132], also referred to as experimental RNomics (Fig. 2).

Tailor-made cDNA libraries for systematic searches on ncRNAs have concentrated on specific RNA classes, where target groups were selected by RNA size fractionation, the ability to ligate a linker to phosphorylated 5’ ends of RNA, structural signature sequences, protein and RNA target binding, or subcellular location. All these approaches require that specific priming sites for first-strand cDNA synthesis be introduced at the 3’ end of RNA molecules (short RNA commonly do not have poly(A) tails for oligo(dT) priming). These priming sites can be introduced by ligating an RNA adaptor to the open 3’ end of RNA using an RNA ligase [132] or by extending the open 3’ end of RNA by adding homopolymers using a poly(A) polymerase [133]. The poly(A) polymerase can also be used for C-tailing, which can be very useful to reduce priming from polyadenylated mRNAs [133]. Specific priming sequences for second-strand cDNA synthesis can be introduced by different approaches as outlined above for the cloning of standard cDNA libraries, such as using poly(C) overhangs in the 5’-adaptor ligation step [134] or the addition of an RNA oligonucleotide to the 5’ end of RNA [135]. As mentioned for standard cDNA cloning protocols, approaches for full-length cloning are again preferable, although this is not a major issue for analytical applications. RNA or cDNA fragments having adaptors at both ends can be rapidly amplified by PCR for further analysis and/or cloning. Usually small RNAs are enriched by size fractionation prior to cloning into a library. Since size fractionation alone is not specific enough for targeted cloning approaches, such libraries usually have low discovery rates. Additional selection steps can improve discovery rates: in the case of H/ACA snoRNAs, for example, the process can be made more specific by using anchored primers for conserved triple nucleotides in the H/ACA box [136]. Mostly short RNA libraries are prepared for new ncRNA discovery and expression profiling [137], which does not require ncRNA cloning but rather relies on the power of new high-throughput sequencing methods [138–140] in brute-force deep sequencing experiments. Although well proven for the discovery of new RNAs, sequences alone will not be sufficient to elucidate the function of newly discovered transcripts. Therefore sequencing approaches should be coupled to cDNA preparation. Here the limit may be set by new high-throughput sequencing...
approaches like the 454 Genome Sequencer FLX System, which can obtain about 200 bp per read. This is sufficient for full-length sequencing of short RNAs at high throughput followed by in vitro synthesis of the corresponding cDNAs. Certainly the rapidly increasing power of high-throughput sequencing in combination with gene synthesis will make such approaches feasible. DNA fragments of some 500 to 800 bp can rapidly be prepared from 40-bp oligonucleotides and automated PCR [141], arguing for an important role for gene synthesis in future transcript analyses.

Cloning long ncRNA

Small RNA libraries do not capture all ncRNAs, and short RNA cloning often captures only mature RNA products. For instance miRNAs are expressed as pre-miRNAs that are processed by endonucleolytic cleavage in the nucleus and cytoplasm to yield mature miRNAs [142]. Those pre-miRNAs are not found in short RNA libraries, but have to be cloned by the standard library approaches outlined above. Cell fractionation in combination with inactivation of enzymes involved in the maturation process can open up interesting possibilities such as the recent global identification of noncoding RNAs in yeast [143]. Cloning such pre-miRNAs is important to understand the regulation of maturation processes and how they affect the cellular localization and function of mature RNAs in the cell. For example, ncRNAs associated with SC35 splicing domains have been identified by their nuclear localization and the fact that they are not exported into the cytoplasm [144]. These long ncRNAs, NEAT1 and NEAT2, are abundant in human and mouse tissues, and other long ncRNAs may be discovered in a similar way. Classical cloning approaches have already identified many long ncRNAs in mouse [45], including longer RNA transcripts (denoted as “macro-ncRNA” [11] or “macroRNA” [145]) often expressed in a sense–antisense orientation to other transcripts [109–111]. However, other than well-known long ncRNAs such as Xist [146] or Air [147], long ncRNAs have not yet attracted much attention, although they seem to be regulated functional transcripts [10,11,145]. Our present knowledge about their features limits the design of dedicated cloning approaches for long ncRNAs.

Future perspectives and developments

Looking back at the development of cDNA cloning technologies over the past 30 years, there is a strong basis for developing novel approaches to the cloning and characterization of new RNA species (Table 2). Only cDNA cloning will provide the necessary resources for functional studies on those new transcripts. This makes cDNA cloning a fundamental technology for future directions in gene discovery and transcriptome analysis.
Table 2

<table>
<thead>
<tr>
<th>Targeted RNA</th>
<th>Comment</th>
<th>Needs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poly(A)+ mRNA</td>
<td>Established for 500 to 15,000 bp</td>
<td>Libraries with wider cloning range.</td>
</tr>
<tr>
<td>Poly(A)− mRNA</td>
<td>Full-length cloning not established</td>
<td>Requires alternative approaches to priming reverse transcription reaction from 3’ end (Patent Application WO2006003721).</td>
</tr>
<tr>
<td>Long mRNA</td>
<td>&gt;15,000 bp</td>
<td>Better reverse transcription reactions, new cloning vectors, e.g., BAC vectors.</td>
</tr>
<tr>
<td>Short mRNA</td>
<td>&lt;500 bp</td>
<td>Better methods to remove adaptors.</td>
</tr>
<tr>
<td>Full-length mRNA</td>
<td>Established methods available</td>
<td>Full-length cloning is preferable wherever possible. Methods to capture capped mRNA should be used in all mRNA cloning approaches.</td>
</tr>
<tr>
<td>Coding mRNA</td>
<td>Selection of ORF clones</td>
<td>Experimental selection for ability to translate [103].</td>
</tr>
<tr>
<td>Noncoding mRNA</td>
<td>Selection of clones lacking any ORF</td>
<td>Experimental selection against ability to translate [103].</td>
</tr>
<tr>
<td>Splice variants of mRNA</td>
<td>Methods for selective cloning of splice variants available</td>
<td>Requires better methods to characterize splice variants and exon usage, e.g., by tiling arrays or full-length shotgun sequencing by new high throughput sequencing.</td>
</tr>
<tr>
<td>Sense–antisense pairs</td>
<td>Very common feature for many genes</td>
<td>Selection by hybridization of “sense driver” to “sense tester” (U.S. Patents 6,528,262 and 6,986,988).</td>
</tr>
<tr>
<td>Short RNA</td>
<td>Commonly selected based on short length</td>
<td>Use of conserved structures for more selective cloning; may be hard, as conserved structures do not necessarily reflect on conserved sequences. Effective protocols to select RNA by binding to proteins and/or DNA/RNA.</td>
</tr>
<tr>
<td>Precursor RNA for short RNA</td>
<td>Most short RNAs go through maturation process</td>
<td>Cloning of long cDNAs using RNA prepared from conditioned cells.</td>
</tr>
<tr>
<td>All RNA</td>
<td>Full-length cDNA sequencing</td>
<td>New approaches to high-throughput full-length cDNA sequencing combining new sequencing technologies and shotgun sequencing.</td>
</tr>
<tr>
<td>All RNA</td>
<td>Cloning of rare transcripts</td>
<td>Linking cDNA cloning to tiling arrays and tag-based approaches.</td>
</tr>
<tr>
<td>All RNA</td>
<td>Target at “universal library” for transcriptome analysis</td>
<td>Modification of 5’ and 3’ ends for unbiased cloning of all RNA transcripts.</td>
</tr>
<tr>
<td>All RNA</td>
<td>Target at expression cloning for functional screens</td>
<td>Limiting factor for coding transcripts is ORF cloning. Direct cloning of ncRNAs into expression vectors may be suitable approach. Effective expression systems with inducible promoters are welcome. Resources should enable “gain-of-function” and “loss-of-function” experiments at the same time. Progress in the development of screening assays.</td>
</tr>
<tr>
<td>All RNA</td>
<td>Small-scale or single-cell libraries</td>
<td>Improvements in amplification methods.</td>
</tr>
</tbody>
</table>

Until now computational sequence analysis alone has largely missed many transcripts [148,149], whereas unsupervised approaches like tiling arrays [151] and tag-sequencing have pushed forward the borders in gene discovery [151]. New strategies combining tiling arrays and cDNA library screening can be envisioned, in which tiling arrays could be used to analyze the complexity of RNA samples or cDNA libraries and at the same time provide the sequence information needed to isolate clones for novel transcripts. The success of such strategies would largely depend on preparing highly complex cDNA libraries with high titer and the sensitivity of the tiling arrays to identify rare transcripts/cDNAs. For tag-based approaches a link between expression profiling and cloning has already been achieved: ditags or paired-end tags comprising the end sequences of cDNAs are derived from full-length cDNA libraries, and this provides sufficient sequence information for primer design and PCR cloning of new transcripts [152]. Similar approaches are also possible for other 5’ end tag-based approaches like CAGE [153] or 5’-SAGE [154]. Consequently classical cDNA libraries/library screening may be challenged by large-scale PCR cloning [155] and gene synthesis utilizing partial or complete sequence information from high-throughput sequencing and tiling array projects. PCR amplification has been widely used in cloning new cDNAs, although PCR amplification requires information from both ends of the transcripts and suitable templates, has an inherent error rate (see above), and may lead to multiple amplicons covering different splice variants. Ditag methods have identified very long transcripts [45], and it is expected that additional long transcripts will be discovered. Present approaches in cDNA library construction enable the cloning of cDNAs of up to 15 kb at best. Improved conditions for reverse transcription reactions are needed, as are new cloning vectors, e.g., BACs, to uncover long RNAs, including macro ncRNAs predicted from cDNA fragments [11].

Especially tag-based approaches will greatly benefit from the fast development of new high-throughput sequencing methods [138–140] that allow deep sequencing of transcriptomes at low cost. These approaches not only will be important for tag-based expression profiling but also will further facilitate important functions in full-length cDNA sequencing by shotgun methods. For example, one can imagine preparing an individual shotgun library per cDNA clone, ligating the resulting DNA fragments to adaptors having clone-specific barcode sequences, and then performing a highly multiplexed sequencing reaction by pooling the barcoded DNA fragments derived from many different cDNA clones. The barcode sequences will guide clone-specific assembly of full-length cDNA sequences (for multiplexing and 454 sequencing refer to [156,157]). Such strategies are of particular interest for analyzing more splice variants and creating more sequence-verified cDNA resources.

Using new strategies, future studies may shift from the large-scale cloning projects of the past to more focused applications driving the discovery of new transcripts and RNA classes. Knowledge-driven approaches will make targeted isolation and cloning of new RNA classes possible, and this process will benefit from a better understanding of structural features of different RNA groups or their cellular localization. For instance, naturally occurring sense–antisense pairs can be isolated by hybridizing
cDNAs from sense and antisense RNA obtained from the same sample (U.S. Patents 6,528,262 and 6,986,988), and poly(A)−mRNA could be cloned by 3′ end adaptor ligation to total RNA followed by an mRNA-specific cap selection. It was suggested to use double-stranded adaptors with oligo(dT) overhangs to block the 3′ ends from polyadenylated mRNA prior to adding an adaptor to poly(A)−RNA for cloning specifically poly(A)−RNA (Patent Application WO2006003721). Such approaches will be important in understanding the large portion of nonpolyadenylated mRNAs in the cell not yet covered by any cDNA collection [107,108,128,129].

Alternatively, new high-throughput sequencing technologies may drive the development of “universal cloning strategies.” Today cDNA library strategies follow certain assumptions to direct the cloning to preferred RNA groups, preferentially by size fractionation of mRNAs with a size of over 500 bp and much shorter ncRNAs of about 25 bp. However, cloning strategies developed for short RNA detection making use of the addition of poly(C) or poly(A) tails, or adaptor ligation at 5′ ends and/or 3′ ends can be extended to develop a universal cloning strategy. Basically every RNA molecule that can be modified at its 5′ and 3′ ends to enable priming of the reverse transcription reaction and the preparation of a second cDNA strand can be converted into a cDNA comprising the entire RNA sequence and contained in a universal cDNA library. Short RNAs derived from endonucleolytic cleavage of precursor RNAs commonly have a 5′ phosphate group and open 3′ ends as needed in library preparation. However, modifications to the 5′ and 3′ ends of RNA have been described that may prevent full-length cloning by standard protocols of RNAs derived from different or thus far unknown maturation processes. For instance, precursor RNAs derived from RNA polymerase I contain an unmodified triphosphate group at their 5′ ends, whereas in transcripts derived from RNA polymerase II the 5′ triphosphate group is rapidly modified by addition of methylated guanosine triphosphate (refer to [158] for studies on 5′ ends of RNA). Also modifications at the 3′ ends of RNA have been described, such as a 2′,3′-cyclic phosphate as an intermediate for the preparation of RNA [159,160]. Hence, further manipulation of the ends of modified RNAs may be required to ensure equal cloning of different RNA species into universal libraries.

Until now universal cloning strategies have not been attractive because the resulting libraries would be dominated by a few RNA species, mostly tRNA and rRNA. Therefore any universal cloning strategy should include additional steps to remove effectively undesired RNA species that are of no interest and would hamper library analysis. New reagents containing beads presenting oligonucleotides complementary to tRNA or enzyme-digestion using RNA fragments or oligonucleotides and RNase H (see above) can selectively remove RNAs. In combination with computational prediction of RNA structures [161,162], ORFs, or other features, RNase H-mediated digestion can remove specific RNA species using oligonucleotides hybridizing to conserved RNA motifs or cleaving off priming sites from selected RNAs. Universal cDNA libraries should enable a more unbiased transcriptome analysis. Since they would cover much larger fractions of transcriptomes than classical libraries, there is a lower risk of losing RNA groups that do not match set parameters during library preparation (see above on size ranges commonly used in library preparation). Moreover, profiles from all RNA groups within one sample could be obtained rather than focusing on a few RNA groups only. This aspect is important because many small RNAs are involved in processing of other larger RNAs coexpressed within the same cell.

For functional studies on novel coding and noncoding transcripts, cDNA clones are necessary for performing in vitro and in vivo experiments. Building global cDNA collections comprising all new RNA species, including unknown splice variants, is a major challenge demonstrated by the enormous number of new transcripts identified recently by tag-base approaches, tiling arrays, and short RNA libraries. The focus of cDNA cloning projects could shift toward functional screens in biological models to identify RNA classes by function. Here I can see new applications for expression cloning, for example, where expression libraries for ncRNAs in combination with effective lentiviral expression clonal systems [163] could play an important role in elucidating ncRNA function. Using cDNA libraries in functional screens may also help to characterize ncRNAs that could not be identified in classical mutation-driven genetic screens. The lack of an an even may even allow for easier design of such expression libraries; for example, an miRNA expression library has been prepared directly from genomic DNA fragments [164]. Although expression cloning is a powerful approach, it is still unclear how well expression cloning will work for ncRNAs. Their largely unknown functions could make it difficult to select the “right” biological context for testing, and redundancy between RNAs may further reduce the yields of screening assays. However, successful genetic screens on miRNAs [164] and the functional cloning of Shirin in an expression cloning system demonstrate that expression cloning approaches can indeed work for identifying ncRNAs [165]. The 3′ untranslated region of Shirin can bind directly to the RNA-binding protein Vg1RBP and is sufficient to induce insulin expression in Xenopus embryos. Its cloning not only highlights a new embryological activity of Vg1RBP, but could stand at the beginning of finding many more RNA-protein interactions in functional screens.

Now that we have realized how little we know about the utilization of genomic information, cDNA cloning and cDNA library preparation have a long way to go in driving discoveries in the RNA world. Many of these new approaches could be “biology-driven” to link phenotypes directly to genotypes.

Acknowledgments

It was not my intention to review here the entire literature regarding cDNA cloning. Therefore I am sorry for those whose publications could not be cited. I thank M. Dushay and P. Einar for their critical comments and suggestions for improving the manuscript. My special thanks are given to the past and present members of DNAFORM’s cDNA Library Team: T. Hayashi, Y. Hodoyma, M. Kamiya, C. Kato, F. Kobayashi, C. Kurihara, A. Lezhava, Y. Shimatani-Shibata, M. Suzuki, S. Takaku, C. Tanaka, and T. Tanaka, as well as P. Carninci and Y. Hayashizaki at RIKEN and Elizabeth Bosch at FivePrime Therapeutics for our close collaboration.
References


