# Improved full-lenght cDNA production based on RNA tagging by T4 DNA ligase

C. Clepet, Isabelle Le Clainche, Michel M. Caboche

Published online January 2, 2004

*Nucleic Acids Research, 2004, Vol. 32, No. 1* ***e6***
*DOI: 10.1093/nar/gng158*

# Improved full-length cDNA production based on RNA tagging by T4 DNA ligase

**Christian Clepet*, Isabelle Le Clainche and Michel Caboche**

Unité de Recherches en Génomique Végétale, INRA/CNRS, 2 Rue Gaston-Crémieux, F-91057 Evry Cedex, France

Received September 16, 2003; Revised and Accepted October 30, 2003

## ABSTRACT

**Second-strand cDNA priming is a central problem for full-length characterization of transcripts. A new strategy using bacteriophage T4 DNA ligase and partially degenerate adapters is proposed for grafting a sequence tag to the end of polyribonucleotides. Based on this RNA tagging system and previously described protocols, a new method for full-length cDNA production has been implemented. Validation of the method is shown in *Arabidopsis thaliana* by the construction of a full-length cDNA library and the analysis of 154 clones and by 5′-RACE–PCR run on a documented experimental system.**

## INTRODUCTION

The study of full-length cDNAs remains an indispensable approach for structural and functional genome annotations (Castelli,V., Aury,J.-M., Jaillon,O., Wincker,P., Clepet,C., Menard,M., Cruaud,C., Schächter,V., Temple,G., Caboche,M., Weissenbach,J. and Salanoubat,M., submitted). The wide range of methods described in patents and the scientific literature shows how critical but difficult this research area is. The central problem every full-length cDNA method tries to resolve is grafting a known sequence at the cap site, so as to be able to prime second-strand polymerisation of the cDNA. In some methods, cap-dependent tagging is used as a way of selecting for full-length cDNAs; in other protocols the tag is added on cDNAs previously enriched for molecules extending to the 5′cap (1,2). A number of enzymatic or chemical taggings have been described, either on single-strand cDNA (see for example 3,4), double-strand cDNA (see for example 5,6), de-capped mRNA (see for example 7–9) or straight on the mRNA cap (10,11). In the strategy of Sekine and Kato (8) and of Maruyama and Sugano (9), RNAs are dephosphorylated by alkaline phosphatase, decapped by tobacco acid pyrophosphatase and ligated to an oligonucleotide by T4 RNA ligase. This method is one of the most specific and accessible for full-length cDNA production, however its efficiency is somewhat limited at the RNA ligation step. Tavitgian *et al*. (12) and Shibata *et al*. (13) used T4 DNA ligase for oligonucleotide joining on single-strand cDNA. The procedure is based on using partially degenerate adapters to create a local double-stranded structure at the junction between the oligonucleotide 5′-phosphate and the cDNA 3′-OH. This tagging does not discriminate against incomplete reverse transcription products and needs, as in Shibata *et al*. (13), to be used in conjunction with other enrichment methods for obtaining full-length cDNAs.

T4 DNA ligase (once known as polynucleotide ligase) can catalyse DNA-templated joining of RNA fragments (14,15). Based on this property, a new full-length cDNA strategy has been designed. After alkaline phosphatase inactivation of uncapped nucleic acids and cap removal by tobacco acid pyrophosphatase, an oligonucleotide is specifically grafted to the cap site of mRNAs. In contrast to previous methods (8,9), RNA tagging is performed by using T4 DNA ligase and adapters generating local double-stranded structure at the junction with the mRNA 5′-end. The ligated mRNAs can then be used for RACE–PCR or library construction. Gateway *attB* sites have been included in the 5′ RNA adapter and the 3′ reverse transcription oligo(dT) primer, enabling cDNA library construction by recombinational cloning (16).

As a proof of concept, 5′ RACE–PCR and full-length cDNA library analyses have been performed in *Arabidopsis thaliana.* The overall sensitivity of the strategy is shown in a convergent way by both experimental approaches. In particular, the 5′ RACE–PCR revealed a new upstream transcription start site (TSS) for *ats1A*. The constructed full-length cDNA library has been evaluated by comparison with the genomic annotations and the cDNA catalogue available for *A.thaliana.* In particular, amongst the 154 clones analysed, a 5′-end located upstream of the AGI (17) annotated coding sequence (CDS) was found for 130 cDNAs and, when compared to EMBL sequences, 12 inserts showed a sequence gain towards the promoter.

Besides the production of full-length cDNAs, this new RNA tagging strategy could be used in any application where an oligo needs to be grafted on one or both sides of RNA fragments of unknown sequence. In particular, it could be useful for amplifying RNA targets isolated by ribonucleoprotein immunoprecipitation experiments (18).

## MATERIALS AND METHODS

### Preparation of the cap adapter

The cap adapter was prepared by mixing 40 µM of both oligonucleotides P1 and P2, in a buffer containing 10 mM

*To whom correspondence should be addressed. Tel: +33 160874512; Fax: +33 160874510; Email: clepet@evry.inra.fr

NaCl and 10 mM Tris–HCl (pH 7.5), heating at 70°C in a beaker of water and left to hybridise by cooling down to room temperature. The adapter was dispensed into small aliquots and kept at –80°C.

### RNA

Total RNAs were extracted by the Trizol method (Invitrogen) from 4-week-old *A.thaliana* aerial vegetative tissues, according to the manufacturer's recommendations.

### Cap-site tagging of mRNAs

*Dephosphorylation.* Aliquots of 2 µg of freshly purified total RNA were treated with 2 U alkaline phosphatase (Roche) in 20 µl of buffer containing 50 mM Tris–HCl (pH 8.5), 5 mM $MgCl_2$ and 40 U ribonuclease inhibitor (RNasin; Promega) for 1 h at 37°C. The reaction was heated for 10 min at 65°C, the volume was made up to 100 µl with $H_2O$, phenol:chloroform extracted and precipitated with 300 mM sodium acetate, 20 µg of glycogen and 2 vol of absolute ethanol, for 20 min at –20°C. RNA was pelleted at 14 000 *g* for 20 min, washed with 200 µl of 75% ethanol and air dried for 5 min on the bench.

*Decapping.* The RNA was then digested for 1 h at 37°C with 2 U tobacco acid pyrophosphatase (Epicentre) in 20 µl of its accompanying buffer supplemented by 40 U RNasin. The reaction volume was made up to 100 µl with $H_2O$, phenol-chloroform extracted and ethanol precipitated as above. The RNA pellet was washed with 75% ethanol and air dried for 5 min.

*Oligonucleotide ligation.* The decapped RNA was resuspended in 6.5 µl of $H_2O$, heated at 65°C for 5 min, equilibrated at 25°C and mixed in a 10µl final volume with 50 mM Tris–HCl (pH 7.5), 10 mM $MgCl_2$, 10 mM dithiothreitol (DTT), 1 mM ATP, 25 µg/ml of bovine serum albumin (BSA), 5% polyethylene glycol 8000, 20 U Rnasin, 4 µM double-stranded cap adapter and 1000 U highly concentrated T4 DNA ligase (New England Biolabs). The reaction was incubated for 3 h at 25°C. The reaction volume was made up to 100 µl with $H_2O$, phenol-chloroform extracted and precipitated with 300 mM sodium acetate, 20 µg of glycogen and 2 vol of ethanol, overnight at –20°C. The RNA pellet was washed with 75% ethanol and air dried for 5 min.

### Reverse transcription

The RNA pellet was resuspended in 8 µl of $H_2O$, mixed with P3 primer (50 pmol) and dNTPs (10 nmol each) in a 12 µl final volume and heated at 65°C for 5 min. The solution was equilibrated at 48°C and completed with 40 U RNasin, 10 mM DTT, 50 mM Tris–HCl (pH 8.3), 75 mM KCl, 3 mM $MgCl_2$ and 200 U M-MLV Superscript III (Invitrogen), in 20 µl final volume. Reverse transcription was performed for 50 min at 48°C and stopped by heating for 15 min at 70°C; cDNAs were stored at –20°C. The design of primer P3 enables cDNA synthesis to be anchored at the polyadenylation site.

### PCR

Unless otherwise indicated, PCR was performed with 1.5 U AmpliTaq Gold (Roche) in 15 µl of buffer containing 15 mM Tris–HCl (pH 8.0), 50 mM KCl, 1.5 mM $MgCl_2$, 200 µM of each dNTP and 0.2 µM of each primer, with a thermal cycling of 8 min at 94°C, then 94°C for 10 s, 60°C for 10 s and 72°C for 1 min for 30 cycles, with a 2 min final step at 72°C. PCR screening and production of sequencing templates were done on 1 µl of each clone stored at –80°C in 4% glycerol, with primers rbc64.U and rbc402.L or M13fwd and M13rev for the RACE–PCR library and with primers univ221.U and univ221.L for the full-length cDNA library, for which only the >1 kb inserts were sequenced. As a genomic control, 60 ng of *A.thaliana* DNA was used. Presence of the *ats* transcripts was controlled by amplifying 1 µl of cDNA with primers rbc64.U and rbc402.L. Reactions were analysed by loading 10 µl on an ethidium bromide stained 2% agarose gel.

### Cloning

*5′-RACE–PCR.* One microlitre of cDNA was amplified with primer pair P1 and rbc402.L and cloned in pGEM-T (Promega). One microlitre of PCR product was ligated into the T/A cloning vector (20 ng), for 1 h at 20°C, in 5 µl of buffer containing 50 mM Tris–HCl (pH 7.5), 10 mM $MgCl_2$, 10 mM DTT, 1 mM ATP, 25 µg/ml BSA and 200 U T4 DNA ligase (Biolabs).

*Full-length cDNA library*. Three microlitres of reverse transcription products were double-stranded and amplified in 20 µl with primers P1pcr and P2pcr, for 12 cycles of 94°C for 10 s, 66°C for 20 s and 72°C for 5 min. PCR fragments >1 kb were fractionnated on a 1.2% agarose gel, electroeluted, ethanol precipitated and resuspended in 10 µl of 10 mM Tris–HCl (pH 7.5), 1 mM EDTA. The cDNAs were inserted into vector pDONR221 by recombinational cloning (16), in a 20 µl final volume, at 24°C overnight with the reagents and protocol of the Gateway system (Invitrogen).

The RACE ligation and the Gateway reaction were purified by Sepharose 4LB chromatography (Amersham Pharmacia) and one-fifth of each product was electroporated into the *Escherichia coli* DH10B strain [$F^-$ *mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) Φ80*lac*ZΔM15 Δ*lac*X74 *deo*R *rec*A1 *end*A1 *ara*Δ139 Δ(*ara,leu*)7697 *gal*U *gal*K $\lambda^-$ *rps*L ($Str^R$) *nup*G *ton*A], with a Life Technologies electroporator, following the manufacturer's conditions. Transformed cells were left to recover in SOC medium for 45 min and plated on LB agar plates supplemented with carbenicillin (50 µg/ml), 5-bromo-4-chloro-3-indolyl-β-D-galactoside (20 µg/ml) and isopropylthio-β-D-galactoside (20 µg/ml), for the pGEM-T cloning, or with kanamycin (50 µg/ml), for the pDONR221 library.

### Sequencing

cDNA inserts were isolated by PCR and purified on silica gel membrane with the Qiagen MinElute system according to the manufacturer's protocol. Sequencing was performed on an ABI fluorescent sequencer, with primer rbc402.L for the RACE–PCR clones and with primer univ221.U, located upstream the 5′-cap adapter of the inserts, for the full-length cDNA clones.

### Sequence analysis

The genomic loci corresponding to the cDNAs were identified by using the Flagdb *A.thaliana* database (19), and genomic sequences spanning 1000 bases upstream and 2000 bases downstream of the predicted CDS start (17) were isolated from the coding strand of each locus found. BLAST alignment (20)

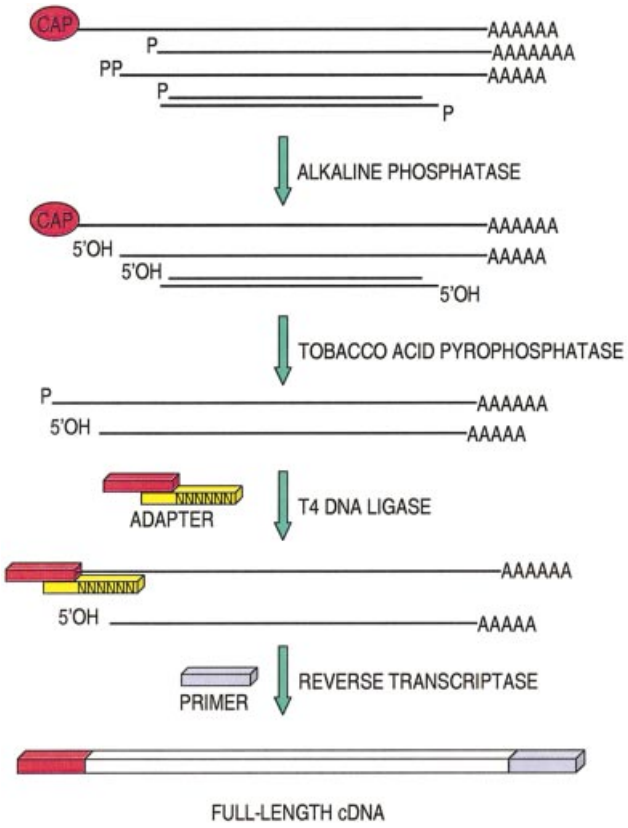


**Figure 1.** Full-length cDNA production strategy. Single- and double-stranded polynucleotides are represented by single and double bars; P indicates a 5′monophosphate moiety; PP, a 5′ di- or triphosphate.

of these 3 kb fragments against the cDNA clones and all RNA-derived *A.thaliana* EMBL sequences (8 August 2003 release) enabled a comparison of the 5′-end and overall exon structures of cDNAs produced with respect to the published data. The cDNA sequences have been released to the EMBL database under accession nos AJ609304–AJ609388.

## RESULTS AND DISCUSSION

A new full-length cDNA cloning strategy has been designed (Fig. 1). Degraded RNAs and any uncapped nucleic acids are inactivated by alkaline phosphatase. Then the cap is removed by tobacco acid pyrophosphatase, releasing a 5′-phosphate on mRNAs. In contrast to previous methods (8,9), T4 DNA ligase is used to add an oligonucleotide to the cap site of mRNAs. The strategy is based on a cap adapter generating local double-stranded RNA:RNA/DNA structure at the mRNA 5′-ends. The upper strand of the cap adapter chosen for the present study is a mixed DNA:RNA oligonucleotide (P1, Table 1) with 8 ribonucleotide residues on the 3′-end and a 3′-hydroxyl ligase acceptor site. The lower strand (P2, Table 1) presents a random protruding 5′-end of 6 nt for pairing with mRNA 5′-ends, and bears a 3′-amine group to block undesired ligations. A six random base overhang was found to be a good compromise for obtaining efficient ligation, considering the size constraint of T4 DNA ligase substrates (21) and the adapter complexity. The adapter-ligated RNA can then be reverse transcribed, with an oligo(dT) primer carrying an adapter on its 5′-end (P3, Table 1), leading to cDNAs with cloning sequences integrated at both ends. This reaction product can be used for synthesis and cloning of full-length double-stranded cDNAs or as a template for RACE–PCR experiments. The Gateway *attB1* and *attB2* sites (Invitrogen) were included in the P1/P2 and the P3 adapters, respectively, so as to permit recombinational cloning of the cDNAs, as described previously (16).

The RNA tagging strategy was first validated by running a RACE–PCR on a documented experimental system. A primer (rbc402.L, Table 1) was designed in a sequence shared by four *A.thaliana* genes; the oligonucleotide target is 100% identical for the genes *ats1B*, *ats2B* and *ats3B* and contains a single mismatch towards the 5′-end for *ats1A*. The primer is orientated towards the transcription starts of these genes and is ~300 bases distant from them on the cDNA. *Arabidopsis thaliana* RNAs were freshly prepared and the cap adapter was joined by T4 DNA ligase to the decapped 5′-ends of mRNAs. cDNAs were reverse transcribed and PCR amplified with rbc402.L and the P1 adapter primer. Gel analysis of the RACE–PCR reaction showed a band consistent with the expected ~350 bp product and absent from the controls (Fig. 2). The RACE–PCR products were cloned into the pGEM-T vector and 10 clones, positive by PCR with the primers rbc64.U and rbc402.L, were sequenced. These inserts correspond to the 5′-end regions of *ats1A*, *ats1B* and *ats3B* cDNAs, linked as expected with the P1 adapter (Fig. 3). In particular, while three *ats1A* inserts match with the EMBL annotated TSS, one *ats1A* cDNA incorporated five additional

**Table 1.** Oligonucleotides

| | |
|---|---|
| P1 | GGGGACAAGTTTGTACAAAAAAGCrArGrGrCrTrGrGrG |
| P2 | NNNNNNCCCAGCCTGCa |
| P3 | GGGGACCACTTTGTACAAGAAAGCTGGGTTTTTTTTTTTTTTTTTTVN |
| Rbc64.U | AATGGCTTCCTCTATGCTCTC |
| Rbc402.L | TGCGGAGAAGGTAGTCAACTTCCT |
| M13fwd | TGTAAAACGACGGCCAGTGA |
| M13rev | GGAAACAGCTATGACCATGAT |
| Univ221.U | TGTAAAACGACGGCCAGTCTTA |
| Univ221.L | CAGGAAACAGCTATGACCATGT |
| P1pcr | GGGGACAAGTTTGTACAAAAAAGCAGGCT |
| P3pcr | GGGGACCACTTTGTACAAGAAAGCTGGGT |

Sequences are oriented 5′→3′. N stands for dA, dG, dC or dT; V stands for dA, dG or dC; rC, rT and rG are ribonucleotides; a stands for 3′$NH_2$.

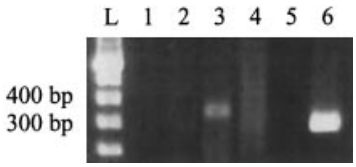


**Figure 2.** Gel analysis of RACE–PCR products. Ethidium bromide stained 2% agarose gel. Aliquots of 12 µl of PCR products (lanes 1–6), amplified with P1 and rbc402.L (lanes 1–3), P1 (lane 4), rbc402.L (lane 5) and rbc64.U and rbc402.L (lane 6) on cDNA (lanes 3–6) or on genomic DNA (lane 1). Lane 2, no DNA control. Lane L, 100 bp ladder (Life Technologies).

bases towards the promoter. The other six cDNA inserts correspond to and are one base or a few bases shorter than the 5′-ends of the *ats1A*, *ats3B* and *ats1B* transcripts reported in EMBL.

The strategy was then validated by the analysis of a full-length cDNA library. With an extensively annotated genome and more than 221 000 cDNA sequences partly released from full-length cDNA projects relying on different technologies, *A.thaliana* is a first rate system for evaluating a full-length cDNA production method. The *A.thaliana* cDNAs tagged at either end with P1 and P3 were amplified, gel fractionnated and cloned in the pDONR221 vector. One hundred and fifty-four recombinant clones were sequenced with a primer located upstream of the 5′-end of the inserts. Without a known cap site signature on the cDNA, the quality of the clones was estimated by considering the position of their 5′-ends with respect to the AGI predicted CDS (17) and the most-upstream 5′-end position of the cDNAs reported in public databases (termed the 'reference 5′-end' in the present study). A clone-by-clone sequence analysis and a summary of these data are shown in Tables 2 and 3. The 5′ sequence of the 154 cDNA constructs were assigned to 103 loci on the *A.thaliana* genome. Except for six clones the sequences produced are consistent with the annotations. One hundred and thirty AGI-consistent cDNAs contain the start ATGs. Three other cDNAs matching the inconsistently annotated regions also appear to have a satisfactory coverage: a209D1 and a209D10 span two neighboring CDS, including the upstream start ATG in both cases; a209C6 encompasses most of the predicted CDS except for the first few codons, which are part of the alternative intron 1 spanned by the clone. Overall, the 5′-end of 133 cDNAs are located upstream from the predicted start ATGs; the rate of inserts potentially spanning the full CDS is ~86% in the present study. With respect to the reference 5′-end established for the 103 loci, 12 clones showed a sequence gain towards the promoter, in particular a189D5 with a 78 base 5′ extension. Twenty clones have a 5′-end identical to the corresponding reference 5′-end and for 31 clones the 5′-ends are located at less than 10 bases downstream of the reference 5′-end. Last, despite the modest size of this cDNA inventory, a clone (a192B10) spanning the whole transcript of an 'orphan gene' was characterized.

The conclusion drawn from these results is that T4 DNA ligase can be used for joining oligonucleotides on RNA; tagging on a variety of RNAs of unknown end sequence was successful by using partially degenerate adapters. The full-length cDNA strategy relying on this new RNA tagging procedure was validated by RACE–PCR and a cDNA library analysis. The series of new upstream 5′-ends generated by both experimental approaches and the quality of the cDNAs overall attest to the efficiency of the method to characterize the 5′-end of mRNAs and to produce full-length cDNAs. Regarding the specificity of the method with respect to the cap sites, the different 5′-ends obtained for a given locus could correspond to alternative TSS or to incomplete alkaline phosphatase inactivation of partially degraded mRNAs.

This study provides an alternative strategy to the difficult issue of transcript characterization and several improvements can be expected over previous methods (8,9). The use of

| N | LOCUS ID | 5'-END SEQUENCE | CAP SITE |
|---|---|---|---|
| | *ats1A* | --GACCAAACCTCAGTCACACAAAGAGTAAAGAAGAACAATGGCTTCCTCTATGCTCTCTTCCGCTAC--- | |
| | (X13611) | >TSS >CDS | |
| 1 | | AAACCTCAGTCACACAAAGAGTAAAGAAGAACAATGGCTTCCTCTATGCTCTCTTCCGCTAC--- | -5 |
| 3 | | TCAGTCACACAAAGAGTAAAGAAGAACAATGGCTTCCTCTATGCTCTCTTCCGCTAC--- | +1 |
| 1 | | CAGTCACACAAAGAGTAAAGAAGAACAATGGCTTCCTCTATGCTCTCTTCCGCTAC--- | +2 |
| 2 | | AGTCACACAAAGAGTAAAGAAGAACAATGGCTTCCTCTATGCTCTCTTCCGCTAC--- | +3 |
| | *ats3B* | --CCAGTAGGAAAACAAGTCAGTAAGTAAACGAGCAAAAGAAGAAGAGAAACAACAAGAAGTAGTAATGGCTT--- | |
| | (X14564) | >TSS >CDS | |
| 1 | | AAGTCAGTAAGTAAACGAGCAAAAGAAGAAGAGAAACAACAAGAAGTAGTAATGGCTT--- | +6 |
| 1 | | GTCAGTAAGTAAACGAGCAAAAGAAGAAGAGAAACAACAAGAAGTAGTAATGGCTT--- | +8 |
| | *ats1B* | --CCAATAGAAAAACAATTAAGCAAAAGAAGAAGAAGAAGAAGTAATGGCTTCCTCTATGCTCTCC--- | |
| | (X14564) | >TSS >CDS | |
| 1 | | ATTAAGCAAAAGAAGAAGAAGAAGAAGTAATGGCTTCCTCTATGCTCTCC--- | +7 |

**Figure 3.** Sequence analysis of RACE–PCR clones. The sequence region next to the cap adapter is shown for 10 RACE–PCR clones. The inserts match with three genes; EMBL accession number and genomic sequence encompassing reported transcription (>TSS) and translation (>CDS) starts are shown for each gene. N, number of clones; Cap site, TSS-relative 5′ position of the RACE–PCR fragments (cloned 5′-ends identical to the EMBL TSS have a value of +1). The various cDNAs found most likely relate to alternative transcription starts; a cDNA (clone a108B3) five bases longer than the previously described *ats1A* transcript was characterized by the present protocol.

**Table 2.** Sequence analysis of the full-length cDNA library

| run | kb | locus ID | GB 5' | ATG | note |
|---|---|---|---|---|---|
| a192B10 | 1.5 | AT5G44730 | new | -143 | (6) |
| a189D5 | 1.6 | AT5G19550 | -72 | -167 | |
| a189E2 | 1.6 | AT5G25752 | -34 | -61 | |
| a209D1 | 1.1 | AT2G35450 & AT2G35440 | -19 | -8 | (1) |
| a189E12 | 1.4 | AT4G17050 | -16 | -88 | |
| a209E4 | 1.4 | | +55 | -26 | |
| a188F11 | 1.4 | | +60 | -21 | |
| a209D10 | 1.4 | AT2G42240 & AT2G42245 | -15 | -72 | (2) |
| a209A8 | 1.3 | AT1G06040 | -14 | -244 | |
| a189F12 | 1.2 | AT4G01310 | -10 | -65 | |
| a189F9 | 1.2 | AT1G71190 | -5 | -50 | |
| a209B4 | 1.3 | AT1G29930 | -5 | -71 | |
| a189A4 | 1.4 | | +1 | -66 | |
| a189E11 | 1.4 | | +1 | -66 | |
| a189A11 | 1.3 | | +1 | -66 | |
| a188G8 | 1.5 | | +1 | -66 | |
| a192C3 | 1.6 | | +1 | -66 | |
| a192C4 | 1.4 | | +1 | -66 | |
| a209E5 | 1.4 | | +1 | -66 | |
| a189C7 | 1.3 | | +1 | -66 | |
| a189D4 | 1.3 | | +1 | -66 | |
| a192B7 | 1.3 | | +2 | -65 | |
| a188B7 | 1.3 | | +5 | -62 | |
| a063B1 | 0.8 | AT1G09590 | -3 | -15 | |
| a197bG2 | 1.2 | AT1G06680 | -2 | -63 | |
| a209C5 | 1.2 | | +4 | -58 | |
| a197bB6 | 1.1 | | +4 | -58 | |
| a189D11 | 1.1 | | +41 | -21 | |
| a209D2 | 1.3 | AT1G43800 | +1 | -51 | |
| a209A7 | 1.3 | AT3G17410 | +1 | -163 | |
| a189C1 | 1.2 | AT4G05320 | +1 | -68 | |
| a189B5 | 1.2 | AT5G57850 | +1 | -39 | |
| a197bH5 | 1.5 | AT5G57930 | +1 | -86 | |
| a197bG1 | 1.2 | AT1G19570 | +1 | -42 | |
| a192C11 | 1.2 | AT3G12930 | +1 | -56 | |
| a209D12 | 1.0 | AT3G44860 | +1 | -56 | |
| a192A11 | 1.0 | | +1 | -56 | |
| a063D1 | 1.1 | AT2G34430 | +1 | -62 | |
| a192C6 | 1.3 | | +2 | -62 | |
| a189E9 | 1.2 | | +2 | -62 | |
| a209E2 | 1.7 | AT5G22340 | +1 | -36 | |
| a209C8 | 1.1 | AT1G15430 | +2 | -136 | |
| a189D12 | 1.3 | AT1G78630 | +2 | -40 | |
| a189B10 | 1.1 | AT3G54890 | +3 | -66 | |
| a197bH1 | 1.0 | | +16 | -53 | |
| a209D5 | 1.2 | AT1G14290 | +3 | -166 | |
| a209D8 | 1.6 | AT4G15545 | +3 | -10 | |
| a189C6 | 1.1 | AT1G44575 | +3 | -43 | |
| a209A1 | 1.0 | | +7 | -39 | |
| a192A9 | 1.4 | AT1G20620 | +3 | -57 | |
| a209D6 | 1.3 | | +4 | -56 | |
| a209B9 | 1.1 | AT3G08940 | +3 | -39 | |
| a189D8 | 1.1 | | +3 | -39 | |
| a189F4 | 1.1 | | +27 | -15 | |
| a209A9 | 1.1 | AT4G12730 | +4 | -64 | |
| a189D6 | 1.2 | AT4G20260 | +4 | -134 | |
| a209B3 | 1.6 | AT2G34420 | +5 | -51 | |
| a197bE1 | 1.3 | | +5 | -50 | |
| a189E10 | 1.3 | | +6 | -50 | |
| a197bG5 | 1.6 | | +9 | -46 | |
| a197bA2 | 1.3 | | +10 | -45 | |
| a197bG3 | 1.4 | AT4G34215 | +6 | -110 | |
| a209B10 | 1.1 | AT3G26740 | +7 | -60 | |
| a209B2 | 1.3 | AT4G14450 | +9 | irrelevant | (3) |
| a063A1 | 0.6 | AT3G15353 | +9 | -71 | |
| a192A6 | 2.1 | AT4G37040 | +9 | -27 | |
| a189E4 | 1.3 | AT1G17530 | +10 | -82 | |
| a192B9 | 1.9 | AT1G15820 | +10 | -86 | |
| a209E3 | 1.4 | | +25 | -71 | |
| a189F11 | 1.3 | | +32 | -64 | |
| a188H4 | 1.1 | AT4G00430 | +13 | -74 | |
| a209E6 | 1.3 | | +26 | -61 | |
| a189F10 | 1.2 | AT2G33800 | +13 | -66 | |
| a209C4 | 1.1 | AT2G42070 | +18 | -84 | |
| a192A8 | 1.5 | AT3G14210 | +22 | -42 | |
| a209A12 | 1.6 | | +32 | -32 | |
| a209B5 | 1.2 | AT1G78620 | +27 | -10 | |

| run | kb | locus ID | GB 5' | ATG | note |
|---|---|---|---|---|---|
| a197bF3 | 1.1 | AT5G54270 | +28 | -67 | |
| a209B12 | 1.3 | | +31 | -64 | |
| a197bH2 | 1.2 | | +32 | -63 | |
| a189D9 | 1.2 | AT2G19080 | +29 | -32 | |
| a197bF1 | 1.2 | AT1G20340 | +31 | -64 | |
| a209B1 | 1.1 | AT2G36830 | +33 | -60 | |
| a209C9 | 1.1 | AT5G10450 | +33 | -23 | |
| a197bH3 | 1.2 | AT1G73230 | +38 | -86 | |
| a209C3 | 1.0 | AT1G48440 | +38 | -36 | |
| a209D7 | 1.2 | AT4G25130 | +39 | -27 | (4) |
| a209C1 | 1.0 | AT3G26520 | +41 | -62 | |
| a209E8 | 1.3 | AT4G10340 | +43 | -52 | |
| a189C11 | 1.5 | | +43 | -52 | |
| a209A3 | 1.2 | | +43 | -52 | |
| a209A10 | 1.2 | AT5G04170 | +43 | -43 | |
| a189B12 | 1.6 | AT1G28200 | +47 | -20 | |
| a197bF5 | 1.3 | AT2G33040 | +48 | -86 | |
| a189D10 | 1.6 | AT5G65430 | +49 | -39 | |
| a192A7 | 1.2 | AT1G44810 | +54 | -25 | |
| a197bH4 | 1.2 | AT1G61520 | +54 | -88 | |
| a192D6 | 1.1 | | +54 | -88 | |
| a197bA1 | 1.2 | | +54 | -88 | |
| a189F1 | 1.3 | | +56 | -86 | |
| a192D5 | 1.1 | | +59 | -83 | |
| a188A9 | 1.1 | | +64 | -78 | |
| a192D8 | 1.1 | | +64 | -78 | |
| a209D3 | 1.5 | AT1G22300 | +55 | -45 | |
| a065aE8 | 0.8 | AT4G25050 | +56 | -54 | |
| a197bG4 | 1.3 | AT1G23820 | +56 | -68 | |
| a209B11 | 1.2 | AT4G05180 | +56 | -18 | |
| a189D3 | 1.2 | AT5G03240 | +61 | -31 | |
| a209C10 | 1.5 | AT1G18080 | +61 | -63 | |
| a189E8 | 1.3 | | +63 | -61 | |
| a209A5 | 1.4 | AT5G01020 | +69 | -106 | |
| a065aF7 | 0.7 | AT2G39390 | +73 | -42 | |
| a063D5 | 0.7 | AT1G72020 | +77 | -16 | |
| a197bH6 | 1.2 | AT5G48020 | +84 | -84 | |
| a209A11 | 1.1 | AT1G04170 | +85 | -116 | |
| a209C11 | 1 | AT3G26650 | +87 | -186 | |
| a197bF4 | 1 | | +87 | -186 | |
| a209D11 | 0.9 | | +87 | -186 | |
| a192A5 | 1.1 | | +88 | -185 | |
| a188E8 | 1.2 | | +108 | -165 | |
| a189B8 | 1.6 | AT5G01530 | +112 | -35 | |
| a192C5 | 1.3 | | +120 | -27 | |
| a189F7 | 1.3 | | +139 | -8 | |
| a189B11 | 0.8 | AT3G26070 | +139 | -92 | |
| a209A4 | 1.0 | AT3G16370 | +147 | -35 | |
| a192A12 | 1.2 | AT4G39090 | +148 | -9 | |
| a209E1 | 1.1 | AT2G20230 | +152 | -70 | |
| a197bA3 | 1.2 | AT2G43750 | +152 | -79 | |
| a209C7 | 1.5 | AT3G47470 | +159 | -412 | |
| a209D9 | 1.3 | | +159 | -415 | |
| a197bF2 | 1.1 | AT2G45440 | +209 | -246 | |
| a065aB1 | 0.9 | AT1G14320 | +243 | -114 | |
| a209C2 | 1.5 | AT1G60950 | +466 | -51 | |
| a209C6 | 1.4 | AT5G64250 | +202 | ~ -190 | (5) |
| a209B6 | 1.6 | AT5G03880 | +27 | +2 | |
| a192D11 | 1.6 | AT3G53750 | | +7 | |
| a209C12 | 1.1 | AT2G14750 | | +30 | |
| a192D12 | 1.3 | AT2G20420 | | +46 | |
| a192A10 | 1.7 | AT5G53480 & AT5G53485 | | irrelevant | (3) |
| a192C7 | 1.4 | AT5G16715 | | irrelevant | (3) |
| a192C12 | 1.1 | AT5G02870 | | > +100 | |
| a189E3 | 1.7 | AT1G73940 | | > +100 | |
| a189D1 | 1.3 | AT4G21190 | | > +100 | |
| a192C10 | 1.4 | AT1G56340 | | > +100 | |
| a209B7 | 1.7 | AT5G19540 | | > +100 | |
| a189C12 | 1.2 | AT2G34510 | | > +100 | |
| a209E12 | 1.2 | AT1G62750 | | > +100 | |
| a188H7 | 1.2 | | | > +100 | |
| a192B5 | 1.3 | AT1G49240 | | > +100 | |
| a192D7 | 1.1 | AT1G55490 | | > +100 | |
| a192D9 | 1.1 | AT1G65930 | | > +100 | |
| a192B12 | 1.7 | AT3G61440 | | > +100 | |
| a192D10 | 1.0 | AT2G16400 | | > +100 | |
| a209E9 | 1.1 | AT1G07230 | | > +100 | |

cDNA clones and their corresponding region on the *A.thaliana* genome (Locus ID) are shown. Kb, approximate insert length is shown in kb; the size distribution of the clones is consistent with the 1.3 kb average for *A.thaliana* mRNAs (17). GB 5′, 5′-end position of the inserts compared to the published cDNAs; for each gene, a reference 5′-end was established from the EMBL cDNA sequence stretching the furthest 5′; distances are measured in cDNA bases; negative values indicate cDNAs starting 5′ of the reference 5′-ends; 5′-ends identical to reference cDNAs are shown by +1 and a one base gain towards the promoter by –1. ATG, cloned 5′-end positions compared to predicted translation starts (17). (1) The 5′-end of this two-loci spanning cDNA is 8 bases upstream from CDS AT2G35450 and >530 bases upstream from CDS AT2G35440. (2) The 5′-end of this two-loci spanning cDNA is 72 bases upstream from CDS AT2G42240 and 661 bases upstream from CDS AT2G42245. (3) Annotations inconsistent with the cDNAs. (4) The unspliced EMBL sequence X97326 was excluded from the analysis. (5) cDNA encompassing most of the predicted CDS except for the first codons which are part of the alternative intron 1 spanned by the clone. (6) no cDNAs were previously published at this locus. cDNAs with a 5′-end upstream from the reference 5′ ends or at <10 bases downstream from the reference 5′-ends or upstream of the predicted ATG starts are in a grey box.

**Table 3.** Summary of the full-length cDNA library analysis

| | |
|---|---|
| cDNA clones | 154 |
| Genes | 103 |
| Relative position of the cloned 5′-ends | |
| (a) upstream from the reference 5′-end | 12 |
| (b) identical to the reference 5′-end | 20 |
| (c) <10 bases downstream from the reference 5′-end | 31 |
| (d) upstream from the ATG | 133 (86%) |
| Total (a + b + c) | 63 (41%) |
| New cDNA | 1 |

See Table 2 legend.

random sequence sticky ends to anchor the adapter to the 5′-end of the mRNAs should increase the time available to the ligase for finding the ends for joining and improve the tagging efficiency. For the polynucleotide substrates, the reported $K_m$ of T4 DNA ligase is several orders of magnitude smaller than the $K_m$ of T4 RNA ligase (see for example 21), although in the present protocol the overall ligation efficiency is reduced by the adapter complexity (4096 combinations), the size and the DNA/RNA hybrid nature of the duplex. With the T4 RNA ligase method, a wide range of 3′-OH polynucleotides can be joined to the 5′-phosphate of mRNAs, leading to undesired side products and loss of the less frequent transcripts; this problem is generally avoided with high molarity adapters. Because T4 DNA ligase is specific for nicks in double-stranded duplexes, the present method keeps such a background to a minimum. Accordingly, no chimeras have been found in this study. Finally, T4 RNA ligase is reported to be more active on some sequences than on others (22) and could lead to biased transcript representation in cDNA libraries.

Besides its application for full-length cDNA cloning, this RNA tagging method could be useful in other applications where an adapter has to be added onto RNAs of unknown sequence, such as in RIP experiments (18).

## ACKNOWLEDGEMENTS

The authors thank Ian Small and Abdel Bendahmane for careful reading of the manuscript and Sebastien Aubourg for discussing the analysis process.

## REFERENCES

1. Edery,I., Chu,L.L., Sonenberg,N. and Pelletier,J. (1995) An efficient strategy to isolate full-length cDNAs based on an mRNA cap retention procedure (CAPture). *Mol. Cell. Biol.*, **15**, 3363–3371.
2. Carninci,P., Kvam,C., Kitamura,A., Ohsumi,T., Okazaki,Y., Itoh,M., Kamiya,M., Shibata,K., Sasaki,N., Izawa,M. *et al.* (1996) High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics*, **37**, 327–336.
3. Frohman,M.A., Dush,M.K. and Martin,G.R. (1988) Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proc. Natl Acad. Sci. USA*, **85**, 8998–9002.
4. Dumas Milne Edwards,J.B., Delort,J. and Mallet,J. (1991) Oligodeoxyribonucleotide ligation to single-stranded cDNAs: a new tool for cloning 5′ ends of mRNAs and for constructing cDNA libraries by *in vitro* amplification. *Nucleic Acids Res.*, **19**, 5227–5232.
5. Akowitz,A. and Manuelidis,L. (1989) A novel cDNA/PCR strategy for efficient cloning of small amounts of undefined RNA. *Gene*, **81**, 295–306.
6. Chenchik,A., Diachenko,L., Moqadam,F., Tarabykin,V., Lukyanov,S. and Siebert,P.D. (1996) Full-length cDNA cloning and determination of mRNA 5′ and 3′ ends by amplification of adaptor-ligated cDNA. *Biotechniques*, **21**, 526–534.
7. Fromont-Racine,M., Bertrand,E., Pictet,R. and Grange,T. (1993) A highly sensitive method for mapping the 5′ termini of mRNAs. *Nucleic Acids Res.*, **21**, 1683–1684.
8. Sekine,S. and Kato,S. (1993) Synthesis of full-length cDNA using DNA-capped mRNA. *Nucleic Acids Symp. Ser.*, **29**, 143–144.
9. Maruyama,K. and Sugano,S. (1994) Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene*, **138**, 171–174.
10. Merenkova,N. and Dumas,J. (1996) *Patent WO 96/34981*.
11. Efimov,V., Chakhmakhcheva,O., Archdeacon,J., Fernandez,J., Fedorkin,O., Dorokhov,Y. and Atabekov,J. (2001) Detection of the 5′-cap structure of messenger RNAs with the use of the cap-jumping approach. *Nucleic Acids Res.*, **29**, 4751–4759.
12. Tavitgian,S.V., Stone,S., Jiang,P. and Kamb,A. (1998) *Patent US 5789206*.
13. Shibata,Y., Carninci,P., Watahiki,A., Shiraki,T., Konno,H., Muramatsu,M. and Hayashizaki,Y. (2001) Cloning full-length, cap-trapper-selected cDNAs by using the single-strand linker ligation method. *Biotechniques*, **30**, 1250–1254.
14. Kleppe,K., van de Sande,J.H. and Khorana,H.G. (1970) polynucleotide ligase-catalyzed joining of deoxyribo-oligonucleotides on ribopolynucleotide templates and of ribo-oligonucleotides on deoxyribopolynucleotide templates. *Proc. Natl Acad. Sci. USA*, **67**, 68–73.
15. Fareed,G.C., Wilt,E.M. and Richardson,C.C. (1971) Enzymatic breakage and joining of deoxyribonucleic acid. *J. Biol. Chem.*, **246**, 925–932.
16. Ohara,O. and Temple,G. (2001) Directional cDNA library construction assisted by the *in vitro* recombination reaction. *Nucleic Acids Res.*, **29**, e22.
17. The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, **408**, 796–815.
18. Niranjanakumari,S., Lasda,E., Brazas,R. and Garcia-Blanco,M.A. (2002) Reversible cross-linking combined with immunoprecipitation to study RNA–protein interactions *in vivo*. *Methods*, **26**, 182–190.
19. Samson,F., Brunaud,V., Balzergue,S., Dubreucq,B., Lepiniec,L., Pelletier,G., Caboche,M. and Lecharny,A. (2002) FLAGdb/FST: a database of mapped flanking insertion sites (FSTs) of *Arabidopsis thaliana* T-DNA transformants. *Nucleic Acids Res.*, **30**, 94–97.
20. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
21. Moore,M.J. (1999) Joining RNA molecules with T4 DNA ligase. *Methods Mol. Biol.*, **118**, 11–19.
22. England,T.E., Bruce,A.G. and Uhlenbeck,O.C. (1980) Specific labeling of 3′ termini of RNA with T4 RNA ligase. *Methods Enzymol.*, **65**, 65–74.