Genome annotation by high-throughput 5′ RNA end determination

Byung Joon Hwang, Hans-Michael Müller, and Paul W. Sternberg*

Howard Hughes Medical Institute and Division of Biology, 156-29, California Institute of Technology, 1200 East California Boulevard, Pasadena, CA 91125

Communicated by Philip P. Green, University of Washington School of Medicine, Seattle, WA, December 16, 2003 (received for review November 20, 2003)

Complete gene identification and annotation, including alternative transcripts, remains a challenge in understanding genome organization. Such annotation can be achieved by a combination of computational analysis and experimental confirmation. Here, we describe a high-throughput technique, trans-spliced exon coupled RNA end determination (TEC-RED), that identifies 5′ ends of expressed genes in nematodes. TEC-RED can distinguish coding regions from regulatory regions and identify genes as well as their alternative transcripts that have different 5′ ends. Application of TEC-RED to ≈10% of the *Caenorhabditis elegans* genome yielded tags 75% of which experimentally verified predicted 5′-RNA ends and 25% of which provided previously unknown information about 5′-RNA ends, including the identification of 99 previously unknown genes and 32 previously unknown operons. This technique will be applicable in any organisms that have a trans-splicing reaction from spliced leader RNA. We also describe an efficient sequential method for concatenating short sequence tags for any serial analysis of gene expression-like techniques.

Although whole genome sequences for human and many other organisms have been obtained, finding the complete set of transcripts, including alternative transcripts, has proven to be technically difficult and has become a rate-limiting step in understanding the nature of genome organization (1). Current genome annotation uses several gene prediction programs, but our limited understanding about genome complexity and diversity necessitates experimental methods to verify or correct these predictions (2). The most commonly used experimental method to identify expressed genes is EST analysis, which involves synthesizing cDNA from RNA transcripts and sequencing the resulting cDNAs one by one (3). EST analysis is biased toward the identification of the 3′ end of transcripts and is of limited utility for identifying the true 5′ end of an RNA transcript. In particular, when different RNA transcripts are produced from a single gene by alternative transcriptional initiations, EST sequencing alone cannot distinguish the shorter full-length alternative transcripts from partially degraded or incompletely extended versions of the longer transcripts. The second most widely used technique, serial analysis of gene expression (SAGE), identifies and annotates expressed genes in many organisms (4–6). In genome annotation with SAGE, sequence tags derived from the 3′ region of transcripts are obtained and matched to genome sequence data to identify expressed genomic regions.

Here, we describe a high-throughput technique that efficiently identifies 5′-RNA ends and thus can distinguish coding regions from regulatory regions and identify genes as well as their alternative transcripts that have different 5′ ends. *Caenorhabditis elegans* mRNA was chosen to develop and test this technique, trans-spliced exon coupled RNA end determination (TEC-RED), because ≈70% of mRNAs in *C. elegans* have a common first exon at their 5′ ends, generated by a trans-splicing reaction between spliced leader (SL) RNA and the 5′ outrons (intron-like sequences at 5′ ends of pre-mRNAs) (7). The *C. elegans* genome is one of the most extensively annotated by computational analysis and by ESTs, allowing the fidelity of this TEC-RED technique to be tested. Because the trans-splicing reaction from SL RNA has been identified in one unicellular eukaryotic phylum (*Sarcomastigophora*) and four metazoan phyla (cnidarians, nematodes, acoelomate flatworms, and ascidians) (8–12), this technique should be applicable to the genomes of a wide variety of organisms.

Materials and Methods

RT-PCR. Total RNA was isolated from a mixed-stage population of wild-type *C. elegans* by the method of Chomczynski and Sacchi (13). Poly(A) RNA was purified from the total RNA by using the Oligotex mRNA Maxi Kit (Qiagen, Valencia, CA). The first-strand cDNA was synthesized from 2.5 μg of poly(A) RNA in a 100-μl reverse transcription reaction with 2.5 μg of RT primer [5′-GTGATGTCTCGAGTAGTTCGAAATGGCC(T)$_{22}$-3′] according to the manufacturer's protocol (Invitrogen). The PCR reaction to label with biotin and introduce a *Bpm*I site at the 5′-cDNA end was performed in a 100-μl reaction with 30 pmol of the biotinylated upstream primer (Biotin/TEG-5′-AGACGCAAGGTTTAATTACCCAAGCTGGAG-3′) and the downstream primer (5′-GAGGTGATGTCTCGAGTAGTTCGAAATGGC-3′), using the Advantage 2 PCR enzyme system (BD Clontech). Amplification was carried out for nine cycles by using the following program: 95°C (1 min) for the first cycle; 95°C (20 sec), 40°C (10 sec), and 68°C (6 min) for the next three cycles; and 95°C (20 sec), 64°C (10 sec), and 68°C (6 min) for the next six cycles. Finally, the PCR mixtures were incubated at 68°C for 5 min and stored at 4°C.

***Bpm*I Digestion and Purification of Mono-5′ Tag.** The PCR product (300 ng) was digested with 40 units of *Bpm*I restriction enzyme (RE) (New England Biolabs). The digested DNA was applied into the six streptavidin-coated PCR tubes according to the manufacturer's protocol (Roche). The 5′-cDNA fragments captured in each PCR tube were treated with T4 DNA polymerase and then incubated overnight with T4 DNA ligase and one of the six DNA adapters below. Adapter 1 (*Kpn*I): 5′-CTATAGGGCTCAAAGATGACGAGAGGAGGTACC-3′; 3′-TGCTCTCCTCCATGG-5′. Adapter 2 (*Hin*dIII): 5′-CAAGATTCTCACGACGATGTTCGGAGTAAGCTT-3′; 3′-CAAGCCTCATTCGAA-5′. Adapter 3 (*Eag*I): 5′-TGAAGATTGCACAGAGGAGAGACCGCTCGGCCG-3′; 3′-CTCTGGCGAGCCGGC-5′. Adapter 4 (*Sac*I): 5′-CAGTTGGAATGAATGAAGCTATACCATGAGCTC-3′; 3′-GATATGGTACTCGAG-5′. Adapter 5 (*Mlu*I): 5′-CTAGTATACGTTCTAGTATCAGAGGAAACGCGT-3′; 3′-AGTCTCCTTTGCGCA-5′. Adapter 6 (*Nhe*I): 5′-TCTTGCAGTGATTAGCGTCAGTGCCTGGCTAGC-3′; 3′-GTCACGGACCGATCG-5′. The ligation products were then amplified by PCR to obtain the mono-5′ tag in Fig. 2. PCR was done with Platinum GenoTYPE *Tsp* polymerase (Invitrogen), using a common upstream primer (5′-AGACGCAAGGTTTAATTACCCAAGCTCGAG-3′) and the following downstream primers: primer 1, 5′-CTATAGGGCTCAAAGATGACGAGAGGA-3′ (for adapter 1); primer 2, 5′-CAAGATTCTCACGACGATGTTCGGAGT-3′ (for adapter 2); primer 3, 5′-TGAAGATTGCACAGAGGAGAGACCGCT-3′ (for adapter 3); primer 4, 5′-CAGTTGGAAT-

Abbreviations: RE, restriction enzyme; SAGE, serial analysis of gene expression; SL, spliced leader; TEC-RED, trans-spliced exon coupled RNA end determination.

*To whom correspondence should be addressed. E-mail: pws@its.caltech.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0308384100

GAATGAAGCTATACCAT-3′ (for adapter 4); primer 5, 5′-CTAGTATACGTTCTAGTATCAGAGGAA-3′ (for adapter 5); primer 6, 5′-TCTTGCAGTGATTAGCGTCAGTGCCTG-3′ (for adapter 6). Amplification was carried out for 22 cycles by using the following program: 94°C (1 min) for the first cycle; 94°C (20 sec), 55°C (10 sec), and 72°C (1 min) for the next six cycles; and 91°C (20 sec), 64°C (10 sec), and 72°C (1 min) for the next sixteen cycles. Finally, the PCR mixtures were incubated at 72°C for 4 min and stored at 4°C.

Concatenation. The sequential assembly protocol allows us to generate concatemers more efficiently and to analyze multiple samples at a time. At each step, PCR products were digested with the appropriate RE, ligated, and amplified by PCR. Only 10–14 PCR cycles were necessary to obtain enough PCR products to proceed to the next step, minimizing mutations during PCR amplification. In this way, we were able to eliminate both the laborious large-scale PCR reactions and the polyacrylamide gel purification step of a large quantity of small DNA fragments that have been used in SAGE. The final PCR products that contain 32-mer 5′ tag were cloned into a DNA sequencing vector.

5′ Tag Sequence Analysis. The AceDB database WS100 (frozen release of 05/10/2003; available at http://ws100.wormbase.org) was downloaded and used for the computational analysis of the 5′ tag sequences. All tags received a unique identifier. "AG plus 5′ tag sequence" was then searched against both strands of all six chromosomes, and after a match was found at a particular position on a given chromosome and strand, WS100 was queried for annotated transcripts in a window of 50,000 bp around the position of the hit. The (direction conserving) distances were computed from the match position to the closest first exon, closest exon (any), and start codon ("ATG") and further processed for classification as described in *Location of 5′ Tag Sequences in* C. elegans *Genome*. We also checked whether a tag overlapped with a *C. elegans* EST that has been aligned to the matching position by using Jim Kent's BLAT program (best match according to WormBase definition) (14). Only completely overlapping tags were considered to be confirmed by EST. The same procedure was applied for determining whether a matched tag lies within an operon. The analysis program was written in Perl and uses AcePerl (http://stein.cshl.org/AcePerl) to query AceDB.

Results

Trans-Spliced Exon Coupled RNA End Determination (TEC-RED) Technique. TEC-RED is based on two principles that have been used in the SAGE technique to quantitatively identify expressed genes by nucleotide sequence tags near the 3′ ends of the transcripts (6). First, a short nucleotide sequence tag from the 5′ end of a transcript contains sufficient information to uniquely identify a transcriptional initiation region, thus distinguishing the coding region (including 5′ UTR) from the regulatory region. Second, concatenation of short sequence tags allows a single DNA sequencing reaction to provide information about multiple transcripts. TEC-RED exploits the common anchor nucleotide sequences (SL1 or SL2 sequence of SL RNAs) that are added *in vivo* to the mRNAs of 70% of *C. elegans* genes (7). The anchor sequences are then used to excise nucleotide sequence tags from the 5′ ends of transcripts.

In this technique, cDNA synthesized from mRNA is amplified by PCR, in which the primer homologous to the SL1 or SL2 sequence contains mismatches that create a *Bpm*I RE recognition site, as well as biotin, resulting in the incorporation of biotin at the 5′ end of the cDNA and a *Bpm*I site at the 3′ end of the SL1 or SL2 sequence within the cDNA (steps 1–3 in Fig. 1*A*; Fig. 1*B*). The PCR products are then cleaved with *Bpm*I, which cleaves DNA 14 bp away from its recognition site, treated with T4 DNA polymerase to make blunt ends, and applied to streptavidin-coated PCR tubes, which purifies biotin-labeled DNA fragments (step 4). Each biotin-DNA fragment containing the modified SL1 or SL2 sequence and the 14-bp 5′-cDNA piece (referred to as a 5′ tag) is ligated with adapter DNA containing a RE recognition site (step 5). The ligated biotin-DNA fragments are purified from the free adapter DNA by using the biotin–streptavidin interaction and PCR-amplified (step 5). During the PCR, the *Bpm*I site within the SL is changed to an *Xho*I site by using a mismatched primer (G to C) (step 5 in Fig. 1*A*; Fig. 1*B*). These PCR products (mono-5′ tag) are then used to sequentially assemble concatemers of 32 5′ tags (step 6 in Fig. 1*A*; Fig. 2), which are then ligated into a plasmid vector for sequencing (step 7 in Fig. 1*A*).

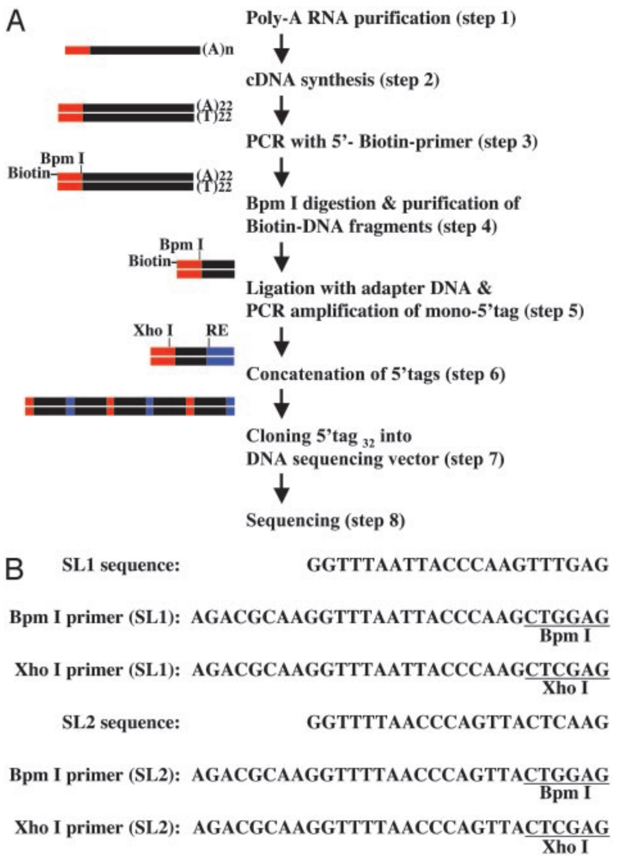

A
Poly-A RNA purification (step 1)
(A)n
cDNA synthesis (step 2)
(A)22
(T)22
PCR with 5'- Biotin-primer (step 3)
Bpm I
Biotin
(A)22
(T)22
Bpm I digestion & purification of Biotin-DNA fragments (step 4)
Bpm I
Biotin
Ligation with adapter DNA & PCR amplification of mono-5'tag (step 5)
Xho I
RE
Concatenation of 5'tags (step 6)
Cloning 5'tag 32 into DNA sequencing vector (step 7)
Sequencing (step 8)
B
SL1 sequence: GGTTTAATTACCCAAGTTTGAG
Bpm I primer (SL1): AGACGCAAGGTTTAATTACCCAAGCTGGAG
Bpm I
Xho I primer (SL1): AGACGCAAGGTTTAATTACCCAAGCTCGAG
Xho I
SL2 sequence: GGTTTTAACCCAGTTACTCAAG
Bpm I primer (SL2): AGACGCAAGGTTTTAACCCAGTTACTGGAG
Bpm I
Xho I primer (SL2): AGACGCAAGGTTTTAACCCAGTTACTCGAG
Xho I

Fig. 1. Scheme of TEC-RED. (*A*) Red boxes represent a trans-spliced SL sequence from SL RNA. The SL sequence is modified during PCR, first to introduce a *Bpm*I site (step 3), and later to change the *Bpm*I to an *Xho*I site during PCR amplification (step 5). Black boxes represent mRNA or cDNA, and blue boxes represent one of the 3′ adapters in Fig. 2. Each 5′ tag within the concatenated 5′ tag polymer (step 7) is flanked by two different anchor sequences, which indicate the orientation of each 5′ tag. (*B*) The PCR primers (*Bpm*I and *Xho*I) contain an extra 8-bp 5′ sequence (AGACGCAA) added to SL1 and SL2 sequences. *Bpm*I primer (SL1) contains a mismatch sequence to SL1 sequence at the 3′ end (TTTGAG to CTGGAG). *Bpm*I primer (SL2) contains a mismatch sequence to SL2 sequence at the 3′ end (CTCAAG to CTGGAG). Both *Xho*I primers (SL1 and SL2) contain a mismatched base at the 3′ end to *Bpm*I primers (CTGGAG to CTCGAG).

Sequential Concatenation. To efficiently concatenate 5′ tags, we developed a new method that sequentially concatenates tags, eliminating problematic procedures in SAGE protocols such as a large-scale PCR reaction and polyacrylamide gel purification of short oligonucleotides, and thus allowing more efficient analysis of multiple samples. In this method, 5′ tags released by *Bpm*I digestion are ligated with six different 3′ adapters, each of which comprises a different RE recognition site (Fig. 2). This gives a population of 5′ tags with a uniform first anchor RE recognition site (*Xho*I) in the 5′ adapter portion and six different second anchor RE recognition

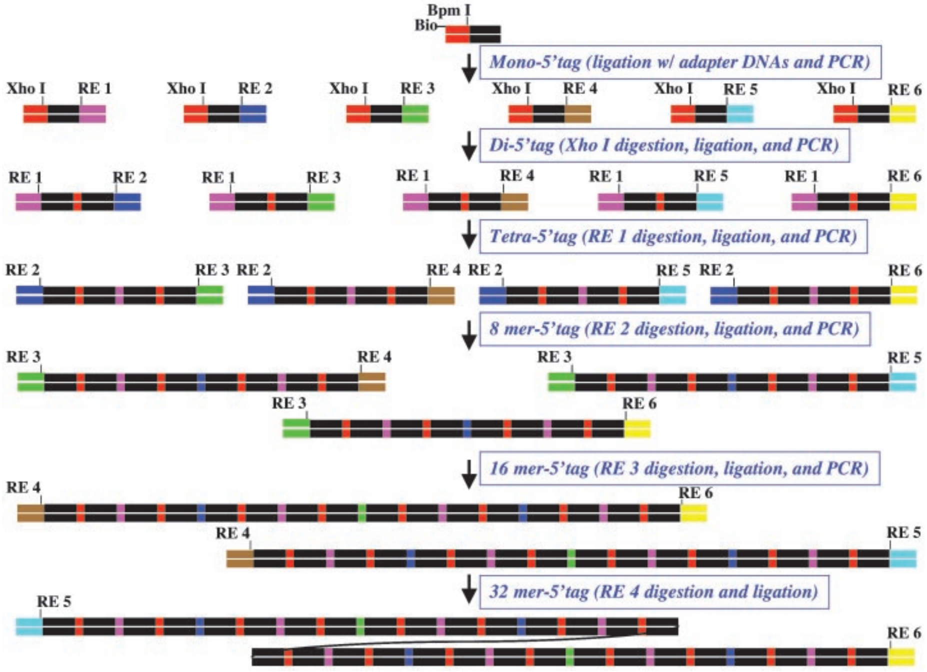

Bpm I
Bio
Mono-5'tag (ligation w/ adapter DNAs and PCR)
Xho I
RE 1
RE 2
RE 3
RE 4
RE 5
RE 6
Di-5'tag (Xho I digestion, ligation, and PCR)
Tetra-5'tag (RE 1 digestion, ligation, and PCR)
8 mer-5'tag (RE 2 digestion, ligation, and PCR)
16 mer-5'tag (RE 3 digestion, ligation, and PCR)
32 mer-5'tag (RE 4 digestion and ligation)

Fig. 2. Sequential concatenation. The biotin-labeled DNA fragments after step 4 in Fig. 1*A* are ligated with six different 3′ adapters, each of which comprises a different RE recognition site. This step gives a population of 5′ tags with a uniform first anchor RE recognition site (*Xho*I) in the 5′ adapter portion. Cleavage with *Xho*I, followed by ligation and PCR, produces head-to-head (5′ end to 5′ end) di-5′ tags, each having a uniform RE 1 recognition site in the 5′ adapter portion. The di-5′ tags are digested with RE I, ligated, and PCR-amplified to form directionally ligated tetra-5′ tags. After three more rounds of digestion (RE2, RE3, and RE4), ligation, and PCR, the concatemers are composed of 32 5′ tag units.

sites in the 3′ adapter portion. After mixing the first mono-5′ tags containing RE 1 recognition site with each of the remaining five mono-5′ tags, cleavage of the mixtures with *Xho*I, followed by ligation and PCR, produces head-to-head (5′ end to 5′ end) di-5′ tags, each having a common RE recognition site (RE 1) in the 5′ adapter portion and different RE recognition sites in the 3′ adapter portion. These di-5′ tags are mixed, digested with RE 2, and ligated to form tetra-5′ tags. These tetra-5′ tags are then mixed and digested sequentially with RE 3 and RE 4 with ligation after each digestion. After the final round of digestion, ligation and PCR, the concatemers are composed of 32 5′ tag units (Fig. 3*B*).

When 5′ tags are ligated to generate a 5′ tag polymer (concatenation), each 5′ tag is directionally ligated as illustrated in Fig. 3*A*. The 5′ end of each 5′ tag is located next to the first anchor RE (*Xho*I in Figs. 1 and 2) and the 3′ end is positioned next to the second anchor RE (RE 1–6 in Fig. 2). This directionality of the 5′ tag helps to unambiguously match the 5′ tag sequence to the transcript of origin within genome sequence. An example of a DNA sequencing chromatograph that contains 32 5′ tag sequences, in which each 5′ tag sequence is located between two anchor REs, is shown in Fig. 3*B*.

Location of 5′ Tag Sequences in *C. elegans* Genome. To analyze these 5′ tag sequence data, we developed a program that locates each 5′ tag sequence within the *C. elegans* genome sequence. Three criteria were used for this analysis. First, each 5′ tag sequence should be located following a 3′-splice acceptor site within the genome sequence. Second, the gene from the matched genome site should have the same orientation as the 5′ tag sequence. Third, when a 5′ tag sequence is located more than once within the genome sequence, the matched genome site that follows a conserved splice acceptor consensus sequence is considered to be the corresponding genomic site for the 5′ tag sequence. After the identification of the genome site matched with a 5′ tag sequence, the program identifies the annotated gene closest to the genome site. It then calculates the distances from the genome site to the first exon, to the closest exon, and to the closest ATG codon of the gene. These distance parameters are used to judge whether the 5′ tag sequence corresponds to the annotated 5′ end of a known gene. When a 5′ tag is located at the known 5′ end, all three distance parameters should be the same because the first ATG sequence of a gene model is defined as the first coding exon boundary in *C. elegans* genome annotation (Fig. 4*A*). When a 5′ tag comes from an additional transcriptional initiation site in the intron of a known gene, the distance from the 5′ tag site to the first exon is large (indicated as a negative number because the 5′ tag is located downstream of the first exon), but the distance to the closest exon is defined to be 2 because the distance is calculated from the 3′-splice acceptor site (Fig. 4*B*). When a 5′ tag is located far from any known genes, indicating an unknown gene, distance parameters to both types of exons are large and different from each other (Fig. 4*C*). When a 5′ tag indicates a new and extended 5′ end of a known gene, the distance parameters to both types of exons are also large but are the same as each other (Fig. 4*D*). This extended 5′ end might also indicate a previously unknown gene. A small number of annotated genes do not belong to any of these categories because of the effects of closely located neighboring genes, errors made during the previous annotation process, or the location of the gene within another gene (unassigned group in Table 1).

TEC-RED Analysis of *C. elegans* mRNA Containing SL1 and SL2 Sequences. To implement this TEC-RED approach, we characterized *C. elegans* mRNAs containing an SL1 or SL2 sequence at their 5′ ends. A total of 13,525 5′ tags (9,401 for mRNA with an SL1 and 4,124 for mRNA with an SL2 sequence) that are matched onto the genome sequence were obtained from 800 DNA sequencing reactions. These 5′ tags represent 2,159 different sequences (1,639 for SL1 and 520 for SL2). These numbers represent the analysis of ≈15% of the mRNA containing an SL1 sequence and ≈14% of the transcripts containing an SL2 sequence. Ninety percent of the sequences uniquely correspond to locations in the genome sequence. The remaining 10% of the sequences (194 for SL1 and 54 for SL2) match onto the genome sequence multiple times (≈90% of them match two or three times). Potential false positives in these matched 5′ tag sequences are from PCR errors and base-calling mistakes in DNA sequencing. To eliminate these false positives, we used the fact that a conserved splice acceptor consensus sequence should always be located just 5′ to the 5′ tag sequences within the genome sequence, because the trans-splicing reactions between SL RNA and the outrons of pre-mRNA are also carried out at the

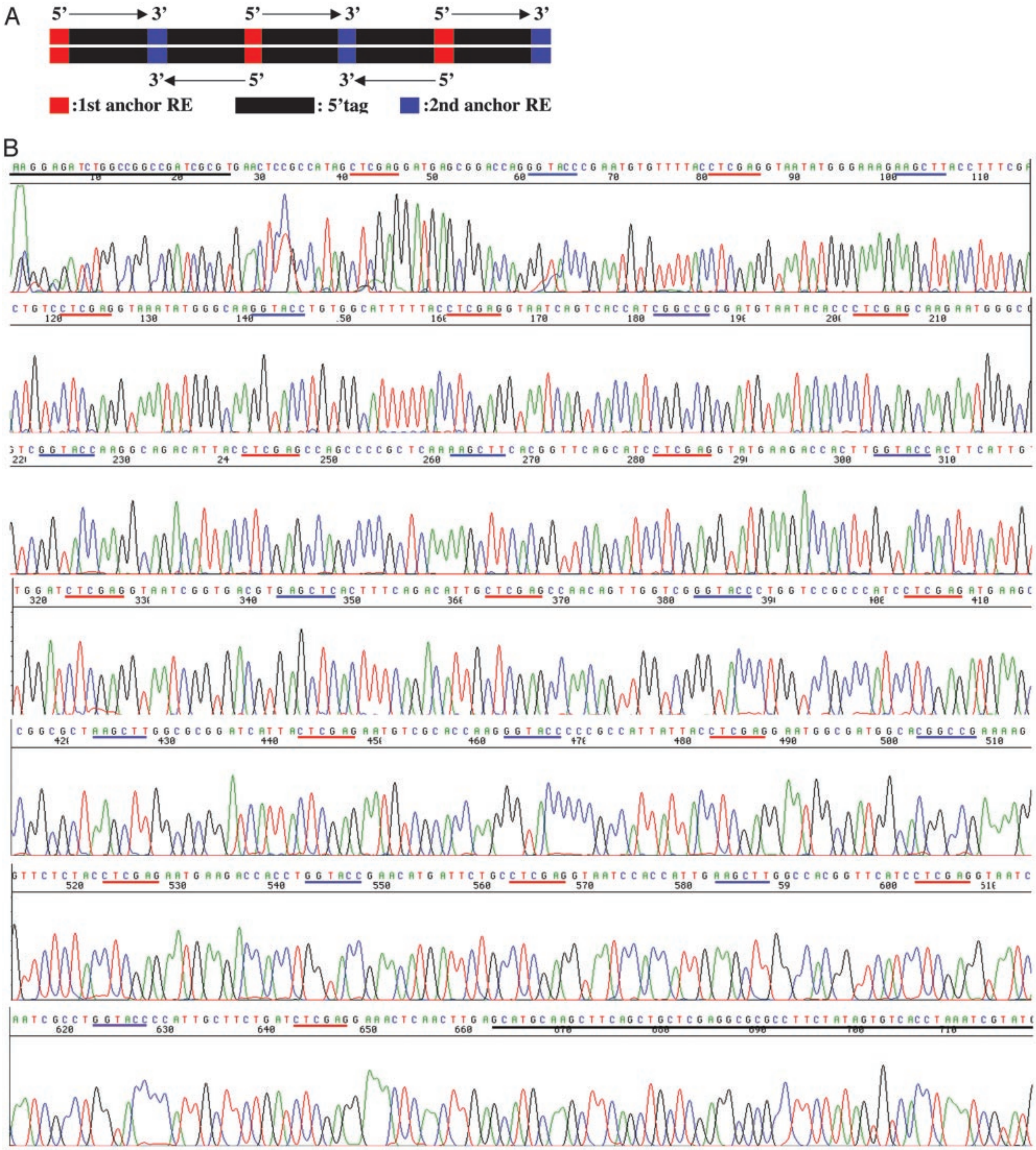
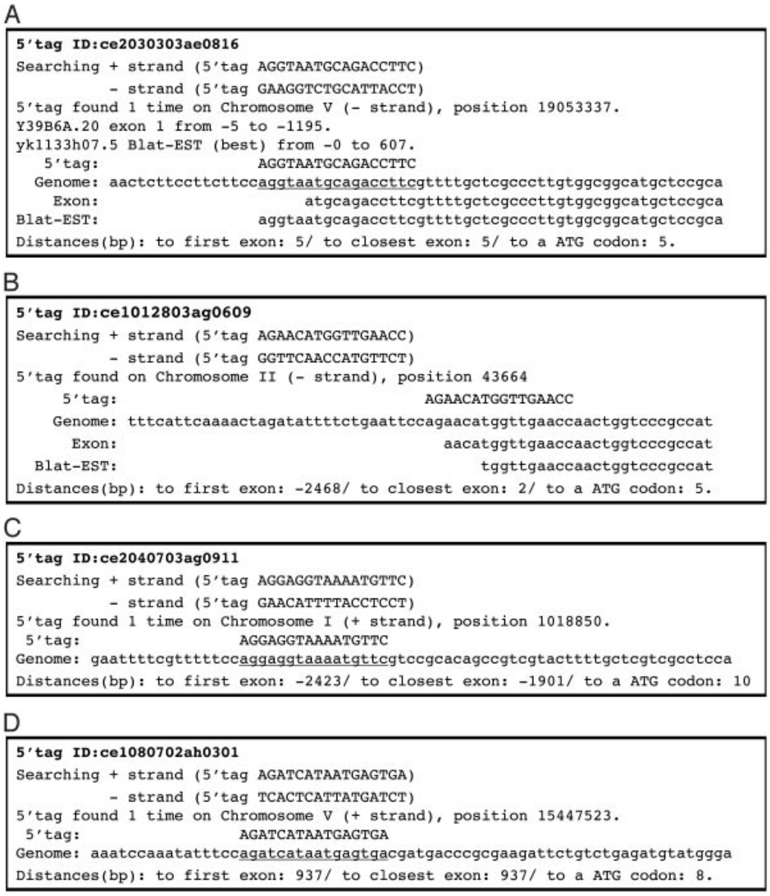


Fig. 3. TEC-RED data. (*A*) Because 5′ tags are directionally concatenated, the 5′ end of each 5′ tag is always located next to the first anchor RE. (*B*) A DNA sequencing chromatogram of the TEC-RED analysis. DNA was sequenced by using the BigDye termination method in the presence of dGTP, greatly improving the DNA sequencing quality. Each DNA sequencing reaction reads ≈700 bp and identifies 32 5′ tags. The red and blue boxes represent the first and second anchor RE sites, respectively. The anchor RE sites are located every 14 bp as described in the text.

conserved splice donor in the SL RNA and the acceptor sites in pre-mRNA (15–18). When a 5′ tag sequence is located after a sequence that can be used as a splice acceptor site, we considered the 5′ tag sequence to be a true positive. The sequences used as splice acceptor sites (7 bp including "AG") were extracted from the known 5′ ends that have been shown to carry out trans-splicing. After removing false positives in which an extended consensus sequence is not located before a 5′ tag sequence, 93% of the 5′ tag sequences (1,597 for SL1 and 415 for SL2) remained true positives that come from expressed genes (Table 1 and Table 2, which is published as supporting information on the PNAS web site).

Classification of the 5′ tag sequences reveals that 75% (1,162 for SL1 and 350 for SL2) of the sequences match known (predicted or experimentally verified by EST) 5′ ends in WormBase release WS100 (frozen 05/10/2003), and the remaining 25% indicate various kinds of differences from existing genome annotations



```
5'tag ID:ce2030303ae0816
Searching + strand (5'tag AGGTAATGCAGACCTTC)
          - strand (5'tag GAAGGTCTGCATTACCT)
5'tag found 1 time on Chromosome V (- strand), position 19053337.
Y39B6A.20 exon 1 from -5 to -1195.
yk1133h07.5 Blat-EST (best) from -0 to 607.
   5'tag:                 AGGTAATGCAGACCTTC
  Genome: aactcttccttcttccaggtaatgcagaccttcgttttgctcgcccttgtggcggcatgctccgca
    Exon:                      atgcagaccttcgttttgctcgcccttgtggcggcatgctccgca
Blat-EST:                 aggtaatgcagaccttcgttttgctcgcccttgtggcggcatgctccgca
Distances(bp): to first exon: 5/ to closest exon: 5/ to a ATG codon: 5.
```

```
5'tag ID:ce1012803ag0609
Searching + strand (5'tag AGAACATGGTTGAACC)
          - strand (5'tag GGTTCAACCATGTTCT)
5'tag found on Chromosome II (- strand), position 43664
     5'tag:                                 AGAACATGGTTGAACC
    Genome: tttcattcaaaactagatattttctgaattccagaacatggttgaaccaactggtcccgccat
      Exon:                                   aacatggttgaaccaactggtcccgccat
  Blat-EST:                                       tggttgaaccaactggtcccgccat
Distances(bp): to first exon: -2468/ to closest exon: 2/ to a ATG codon: 5.
```

```
5'tag ID:ce2040703ag0911
Searching + strand (5'tag AGGAGGTAAAATGTTC)
          - strand (5'tag GAACATTTTACCTCCT)
5'tag found 1 time on Chromosome I (+ strand), position 1018850.
 5'tag:                 AGGAGGTAAAATGTTC
Genome: gaattttcgtttttccaggaggtaaaatgttcgtccgcacagccgtcgtacttttgctcgtcgcctcca
Distances(bp): to first exon: -2423/ to closest exon: -1901/ to a ATG codon: 10
```

```
5'tag ID:ce1080702ah0301
Searching + strand (5'tag AGATCATAATGAGTGA)
          - strand (5'tag TCACTCATTATGATCT)
5'tag found 1 time on Chromosome V (+ strand), position 15447523.
 5'tag:                 AGATCATAATGAGTGA
Genome: aaatccaaatatttccagatcataatgagtgacgatgacccgcgaagattctgtctgagatgtatggga
Distances(bp): to first exon: 937/ to closest exon: 937/ to a ATG codon: 8.
```

Fig. 4. Location of a *C. elegans* genome site that matches with a 5′ tag sequence. The AG plus 5′ tag sequence and its complementary sequence were searched against the *C. elegans* genome sequence, and the data were processed as described in the text. (*A*) The 5′ tag is located at the known 5′ end. (*B*) The 5′ tag identifies an alternative transcript that is transcribed by a promoter within an intron, because the distance from the 5′ tag sequence to the first exon of this gene is 2,468 bp, and the distance to the closest exon is 2 bp. The presence of an ATG codon in the 5′ tag sequence (2 bp after the 5′ end of 5′ tag) suggests that a protein can be translated from this transcript. (*C*) The 5′ tag indicates a previously unknown gene. (*D*) The 5′ tag indicates a new extended 5′ end of a known gene.

(Table 1). Because 63% (734 for SL1 and 214 for SL2) of the known 5′ ends have not been experimentally verified by EST, this analysis experimentally demonstrates that they are true 5′ ends from expressed genes. 99 previously unknown genes including internal transcripts in known genes are identified in this analysis. We thus predict that ≈1,000 expressed genes have not been predicted in the *C. elegans* genome. About 20% (401 of 2,012) of the 5′ tag sequences indicate 5′ ends different from those annotated [350 SL1 and 51 SL2 5′ tags in the "alternative 5′ end in intron," "extended 5′ end," and "unassigned group" (A, B, and D) classes in Table 1]. The genes in these classes are potential candidates for having alternative transcripts that contain distinct 5′ ends. A more extensive collection of 5′ tags is needed to distinguish genes containing a unique 5′ end from those containing multiple 5′ ends.

Table 1. TEC-RED analysis of *C. elegans* mRNA containing SL1 or SL2 sequence

Classification	SL1 (%)	SL2 (%)
Known 5′ end	1,162 (72.8)	350 (84.4)
Prediction	734	214
Confirmed by EST	428	136
Alternative 5′ end in intron	102 (6.4)	20 (4.8)
Extended 5′ end	172 (10.8)	19 (4.6)
Peviously unknown gene	60 (3.8)	7 (1.7)
Unassigned group	101 (6.3)	19 (4.6)
(A) Minor mis-prediction of first exon (<10 bp)	18	3
(B) Major mis-prediction of first exon*	29	7
(C) Internal transcript in a known gene†	12	0
(D) Alternative 5′ end in intron/ referenced by two neighboring genes	29	2
(E) Previously unknown genes/referenced by two neighboring genes	13	7
Total	1,597	415

DNA sequencing data were processed and classified as described in the text. To ensure that the 5′ tag analysis program identified the correct genome sites, the sequences were also individually inspected by using gene prediction and annotation data/programs at WormBase (www.wormbase.org).

*The difference is larger than 10 bp and exists in the predicted first exon.

†5′ tag is located in the introns and exons other than the first exon.

Twenty of 22 SL2 5′ tags in the alternative 5′ end in intron and unassigned group (D) classes are from *C. elegans* operons, suggesting that these alternative 5′ ends containing the SL2 sequence are created by differential splicing events of polycistronic messages rather than by the alternative transcriptional initiations that create alternative 5′ ends in the messages containing the SL1 sequence. All 14 previously unknown genes containing the SL2 sequence are downstream genes in operons, suggesting that this approach is highly accurate in its identification of expressed genes. Eleven previously unknown genes are from previously unknown operons, and the remaining three genes are located in the gap regions in the middle of known operons (CEOP1724, CEOP3260, and CEOP4156). Fifteen of 19 extended 5′ ends containing the SL2 sequence are also from downstream genes in operons (11 are matched to the end of the 5′ UTR and 8 are new 5′ ends). A total of 85% (352 of 415) of the 5′ tag sequences with SL2 sequence correspond to downstream genes in operons, consistent with the previous finding that 90% of mRNAs containing an SL2 sequence are from downstream genes in operons (19). This TEC-RED analysis also identified 32 previously unknown operons, suggesting that more extensive application of TEC-RED will identify ≈250 more operons (Table 3, which is published as supporting information on the PNAS web site).

Discussion

This article describes TEC-RED, a high-throughput technique that identifies 5′-RNA ends and an efficient sequential method for concatenating short sequence tags. We used these techniques to verify *C. elegans* genome annotation done by gene prediction and EST analysis (20). The high accuracy and efficiency of this technique for identifying 5′-RNA ends may well lead to higher resolution annotation of genome sequences by distinguishing regulatory regions from coding regions. For new nematode genomes to be sequenced, this technique will also allow cost-efficient experimental genome annotation.

Applied to the *C. elegans* genome, this TEC-RED technique experimentally verified that 75% of the 5′-RNA ends annotated in WormBase are correctly predicted, and found that 25% of the 5′-RNA ends we obtained are different from the predictions in WormBase, thus providing new information about 5′-RNA ends. This new information includes alternative 5′ ends in introns, extended new 5′ ends, internal transcripts in known genes, and unpredicted new genes. We estimate that ≈1,000 expressed genes have not been predicted in the *C. elegans* genome based on this TEC-RED analysis, which agrees with the estimate of 1,300 new *C. elegans* genes from comparative analysis using the *Caenorhabditis briggsae* genome sequence (21). In a study of *C. elegans* ORFs, 29% of ORFs already touched by ESTs, and thus previously identified experimentally, could not be amplified with the PCR primers designed based on the predicted translational start and stop sites but could be amplified with internal primers (22, 23). This finding is concordant with our finding that ≈16% of the predicted 5′ ends (5′ tags in the alternative 5′ end in intron and extended 5′ end classes) are not correctly annotated.

The experimental proof for the high accuracy of this technique (lower false positive rate) came from the analysis of 5′ tags from the mRNAs containing a trans-spliced SL2 sequence. Here, 20 of 22 alternative 5′ ends in introns, all 14 previously unknown genes containing the SL2 sequence, and 15 of 19 extended 5′ ends containing the SL2 sequence are downstream genes in operons. Considering the previous finding that 90% of mRNAs containing the SL2 sequence are from downstream genes in operons (19), the fact that 49 of 55 SL2 5′ tags including all 14 5′ tags for previously unknown genes are from downstream genes in operons demonstrates the high accuracy of this technique in identifying 5′-RNA ends from expressed genes.

Several modifications will lead to a broader application of this technique. First, several type IIs and III REs such as *Bsg*I, *Mme*I, and *Eco*P15I, can replace the *Bpm*I digestion, generating different lengths of 5′ tag. In particular, when a larger genome is analyzed, generation of longer 5′ tags (18 and 27 bp) by *Mme*I and *Eco*P15I will be necessary to unambiguously match the 5′ tag sequence with the transcript of origin. Second, the sequential concatenation method can help improve the efficiency of SAGE and SELEX (systematic evolution of ligands by exponential enrichment), SAGE techniques that are based on one-step concatenation method requiring a large-scale PCR reaction and polyacrylamide gel purification of short oligonucleotides (5, 6, 24).

A trans-splicing reaction is carried out between SL RNA and the 5′ outrons (intron-like sequences at 5′ ends of pre-mRNAs) (16–18). Because outrons lack a 5′-splice donor site, trans-spliced genes should initiate transcription close to a 5′-RNA end (25). Thus, the 5′-RNA ends determined by TEC-RED will provide information on the 5′-upstream regulatory regions containing a promoter element by distinguishing coding regions from 5′-regulatory regions.

The frequency of alternative transcripts in metazoans has currently been estimated as 20–50% and has historically increased as EST sequences accumulate (22, 26, 27). To efficiently identify these large numbers of alternative transcripts, we imagine combining currently available high-throughput techniques: TEC-RED for the identification of 5′ ends, SAGE for the identification of 3′ ends (4), and ORFeome for PCR amplification of transcripts based on predicted 5′ and 3′ ends (22, 23). Thus, one can generate a high resolution genomic map of alternative transcripts in a cost-efficient way.

TEC-RED is potentially applicable outside the nematode phylum because SL trans-splicing has been reported as a major splicing event in several other phyla: one unicellular eukaryotic phylum (*Sarcomastigophora*) and metazoan phyla (cnidarians and acoelomate flatworms) (9–11). The reaction has also been reported in *Ciona intestinalis*, an ascidian protochordate (12). It is also possible that trans-splicing may happen much more generally in other organisms. Several lines of evidence support this possibility. Both trans- and cis-splicing reactions share the same splicing machinery and use the same splice donor and acceptor consensus sites (16–18). Thus, all of the organisms that carry out cis-splicing have the potential to carry out a trans-splicing reaction. Second, it has been shown that an organism's capability to carry out trans-splicing reaction depends on mRNA structure and the presence of SL RNA (25, 28, 29). Transcription initiation after a splice donor site creates pre-mRNA transcripts containing outron that can be removed only by trans-splicing, not by cis-splicing.

The TEC-RED procedure described here can be modified so that a common sequence motif can be tagged at 5′-RNA ends by an *in vitro* ligation reaction, thus making this technique applicable to organisms in which endogenous trans-splicing from SL RNA does not exist.

We thank S. Gharib for DNA preparation, T. Blumenthal for discussions about operons, and C. Bastiani, T. Blumenthal, Y. Kee, E. Schwarz, and S. Vernooy for careful reading of the manuscript. This work was supported by the Howard Hughes Medical Institute, with which P.W.S. is an Investigator and B.J.H. and H.-M.M. were Associates, by a W. M. Keck Foundation/California Institute of Technology discovery award (to B.J.H. and P.W.S.), and by National Human Genome Research Institute/National Institutes of Health Genome Scholar and Faculty Transition Award K22HG02907-01 (to B.J.H.).

1. Collins, F. S., Green, E. D., Guttmacher, A. E. & Guyer, M. S. (2003) *Nature* **422,** 835–847.
2. Pennisi, E. (2003) *Science* **301,** 1040–1041.
3. Marra, M. A., Hillier, L. & Waterston, R. H. (1998) *Trends Genet.* **14,** 4–7.
4. Pleasance, E. D., Marra, M. A. & Jones, S. J. (2003) *Genome Res.* **13,** 1203–1215.
5. Saha, S., Sparks, A. B., Rago, C., Akmaev, V., Wang, C. J., Vogelstein, B., Kinzler, K. W. & Velculescu, V. E. (2002) *Nat. Biotechnol.* **20,** 508–512.
6. Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. (1995) *Science* **270,** 484–487.
7. Zorio, D. A., Cheng, N. N., Blumenthal, T. & Spieth, J. (1994) *Nature* **372,** 270–272.
8. Krause, M. & Hirsh, D. (1987) *Cell* **49,** 753–761.
9. Rajkovic, A., Davis, R. E., Simonsen, J. N. & Rottman, F. M. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 8879–8883.
10. Sutton, R. E. & Boothroyd, J. C. (1986) *Cell* **47,** 527–535.
11. Tessier, L. H., Keller, M., Chan, R. L., Fournier, R., Weil, J. H. & Imbault, P. (1991) *EMBO J.* **10,** 2621–2625.
12. Vandenberghe, A. E., Meedel, T. H. & Hastings, K. E. (2001) *Genes Dev.* **15,** 294–303.
13. Chomczynski, P. & Sacchi, N. (1987) *Anal. Biochem.* **162,** 156–159.
14. Kent, W. J. (2002) *Genome Res.* **12,** 656–664.
15. Nilsen, T. W. (1993) *Annu. Rev. Microbiol.* **47,** 413–440.
16. Bektesh, S. L. & Hirsh, D. I. (1988) *Nucleic Acids Res.* **16,** 5692.
17. Hannon, G. J., Maroney, P. A., Denker, J. A. & Nilsen, T. W. (1990) *Cell* **61,** 1247–1255.
18. Thomas, J. D., Conrad, R. C. & Blumenthal, T. (1988) *Cell* **54,** 533–539.
19. Blumenthal, T., Evans, D., Link, C. D., Guffanti, A., Lawson, D., Thierry-Mieg, J., Thierry-Mieg, D., Chiu, W. L., Duke, K., Kiraly, M. & Kim, S. K. (2002) *Nature* **417,** 851–854.
20. *Caenorhabditis elegans* Sequencing Consortium (1998) *Science* **282,** 2012–2018.
21. Stein, L. D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M. R., Chen, N., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A., *et al.* (2003) *PLoS Biol.* **1,** E45.
22. Reboul, J., Vaglio, P., Rual, J. F., Lamesch, P., Martinez, M., Armstrong, C. M., Li, S., Jacotot, L., Bertin, N., Janky, R., *et al.* (2003) *Nat. Genet.* **34,** 35–41.
23. Vaglio, P., Lamesch, P., Reboul, J., Rual, J. F., Martinez, M., Hill, D. & Vidal, M. (2003) *Nucleic Acids Res.* **31,** 237–240.
24. Roulet, E., Busso, S., Camargo, A. A., Simpson, A. J., Mermod, N. & Bucher, P. (2002) *Nat. Biotechnol.* **20,** 831–835.
25. Conrad, R., Lea, K. & Blumenthal, T. (1995) *RNA* **1,** 164–170.
26. Misra, S., Crosby, M. A., Mungall, C. J., Matthews, B. B., Campbell, K. S., Hradecky, P., Huang, Y., Kaminker, J. S., Millburn, G. H., Prochnik, S. E., *et al.* (2002) *Genome Biol.* **3,** RESEARCH0083.
27. Modrek, B., Resch, A., Grasso, C. & Lee, C. (2001) *Nucleic Acids Res.* **29,** 2850–2859.
28. Conrad, R., Thomas, J., Spieth, J. & Blumenthal, T. (1991) *Mol. Cell. Biol.* **11,** 1921–1926.
29. Conrad, R., Liou, R. F. & Blumenthal, T. (1993) *EMBO J.* **12,** 1249–1255.