

Genome-wide characterization of methylguanosine-capped and polyadenylated small RNAs in the rice blast fungus *Magnaporthe oryzae*

Malali Gowda^{1,2}, Cristiano C. Nunes¹, Joshua Sailsbery¹, Minfeng Xue^{1,3}, Feng Chen⁴, Cassie A. Nelson⁵, Douglas E. Brown¹, Yeonyee Oh¹, Shaowu Meng¹, Thomas Mitchell⁶, Curt H. Hagedorn⁵ and Ralph A. Dean^{1,*}

¹Fungal Genomics Laboratory, Center for Integrated Fungal Research, North Carolina State University, Raleigh, NC 27606, ²Plant Biology, Michigan State University, East Lansing, MI 48824, USA, ³Department of Plant Pathology, China Agricultural University, Beijing 100094, China, ⁴US DOE Joint Genome Institute, Walnut Creek, CA 94598, ⁵University of Utah School of Medicine and Huntsman Cancer Institute, Salt Lake City, UT 84132 and ⁶Department of Plant Pathology, Ohio State University, Columbus OH 43210, USA

Received April 17, 2010; Revised June 9, 2010; Accepted June 10, 2010

ABSTRACT

Small RNAs are well described in higher eukaryotes such as mammals and plants; however, knowledge in simple eukaryotes such as filamentous fungi is limited. In this study, we discovered and characterized methylguanosine-capped and polyadenylated small RNAs (CPA-sRNAs) by using differential RNA selection, full-length cDNA cloning and 454 transcriptome sequencing of the rice blast fungus *Magnaporthe oryzae*. This fungus causes blast, a devastating disease on rice, the principle food staple for over half the world's population. CPA-sRNAs mapped primarily to the transcription initiation and termination sites of protein-coding genes and were positively correlated with gene expression, particularly for highly expressed genes including those encoding ribosomal proteins. Numerous CPA-sRNAs also mapped to rRNAs, tRNAs, snRNAs, transposable elements and intergenic regions. Many other 454 sequence reads could not be mapped to the genome; however, inspection revealed evidence for non-template additions and chimeric sequences. CPA-sRNAs were independently confirmed using a high affinity variant of eIF-4E to capture 5'-methylguanosine-capped RNA followed by 3'-RACE sequencing. These results expand the repertoire of small RNAs in filamentous fungi.

INTRODUCTION

Deep transcriptome analyses have revealed that almost the entire genome of complex eukaryotes such as mammals is transcribed (1–3). Transcript length varies from as little as 16 nt (such as tRNA) to >100 kb (such as *Xist* RNA). Mammals contain a variety of small RNAs and include tRNAs (16 nt), siRNAs/miRNAs/piRNAs (20–30 nt), tRNA halves (30–40 nt), PASR/TASR (22–200 nt) and snoRNAs (70–200 nt) that can modulate transcription, translation, replication and chromatin structure (1,3). A growing number of ncRNAs have been described in *Saccharomyces cerevisiae*. Many of these are driven by RNA Pol II and include cryptic unstable transcripts (CUTs, ~400 nt) (4) and stable unannotated transcripts (SUTs, ~700 nt) (5). Most map to transcription start sites or the 3'-end of protein-coding genes and appear to be the result of bidirectional promoter activity. In general, CUTs are degraded rapidly by the Nrd1-exosome-Trf4-Air2-Mtr4p polyadenylation (TRAMP) complex (4). Several cases have shown where these ncRNAs interact with the ribosomal complex and are translated (6,7). In the filamentous fungus *Neurospora crassa*, several new species of small RNAs have been recently described including miRNA-like small RNAs (milRNAs), Dicer-independent small interfering RNAs (disiRNAs) and qiRNAs (8,9). qiRNAs arise in response to DNA damage and map to sense and antisense strands of the rDNA array.

In this study, we undertook small RNA profiling in the ascomycete filamentous fungus, *Magnaporthe oryzae*

*To whom correspondence should be addressed. Tel: +1-919-513-0020; Fax: +1-919-513-0024; Email: ralph_dean@ncsu.edu; radean2@ncsu.edu

(anamorph *Pyricularia oryzae* Cav), which causes blast, the most destructive disease of rice worldwide. The fungus not only destroys rice leaves, panicles and roots but also infects other cereals including wheat, barley, finger millet and grasses (10–12). Due to its agronomic significance and molecular genetic tractability, *M. oryzae* has emerged as a model to study fungal pathogenesis. In 2005, the genome (40 Mb) of *M. oryzae* was sequenced and ~11 000 protein-coding genes identified (13). Studies using expressed sequence tags (EST), serial analysis of gene expression (SAGE), massively parallel signature sequencing (MPSS) and microarray expression profiling have revealed that the transcriptome is more complex than initially appreciated (13–15). Here, we conducted pyrosequencing of cDNA and describe a distinct class of small RNAs that are 5'- and 3'-modified, which we refer to as CPA-sRNAs (5'-methylguanosine-capped and 3'-polyadenylated small RNAs) (Figure 1A). CPA-sRNAs share no similarity to qiRNAs, miRNAs and disiRNAs discovered recently in *N. crassa*, which appear to possess no 5'- and 3'-modifications (8,9).

MATERIALS AND METHODS

Fungal strain and growth

Magnaporthe oryzae isolate 70–15 was used in this study because of the availability of genomic (13) and transcriptomic (14,15) resources. Conidia were germinated and mycelia cultured in a liquid medium (0.2% yeast extract and 1% sucrose) by shaking at 200 rpm, 25°C for 3 days. The mycelia were filtered through cheesecloth and used for RNA isolation.

RNA isolation, CPA-sRNA library construction and 454 sequencing

Total RNA was isolated from 2 g of mycelia using the Trizol method (15,16). PolyA⁺ RNA was purified using a PolyATtract mRNA Isolation System III (Promega) according to manufacturer's procedure. To construct the CPA-sRNA library, protocols used to generate full-length cDNA were followed, from which small molecules were size selected and sequenced (16). Briefly, the free phosphate at the 5'-ends of 1 µg polyA⁺ RNA from mycelia was removed by treating with bacterial alkaline phosphatase (BAP, Epicenter) followed by removal of the 5'-methylguanosine caps by treating with tobacco acid pyrophosphatase (Epicenter). PolyA⁺ RNA with an exposed 5'-phosphate was ligated to a 5'-RNA oligo linker (5'-AGCAUCGAGUCGGCCUUGUUGGCCUA CUGG-3') using T₄ RNA ligase (Epicenter). The ligated polyA⁺ RNA was treated with DNase I (Invitrogen) to remove contaminating genomic DNA and re-purified using the PolyATtract mRNA Isolation System III. The 3'-oligo (dT)₂₀VN linker (5'-GCGGCTGAAGACGGCC TATGTGGCC(T)₂₀VN-3') was used to synthesize cDNA using SuperScriptIII (Invitrogen) according to supplier's procedure. RNA was digested with RNase H (Invitrogen). Double-stranded cDNA was amplified with high fidelity Platinum Taq DNA polymerase (Invitrogen) using 5'-PCR primers specific for the 5'-RNA linker (5'-AGCA

TCGAGTCGGCCTTGTTG-3') and 3'-PCR primers specific for the 3'-oligo(dT)₂₀VN linker (5'-GCGGCTGAAGACGGCCTATGTG-3'). The conditions used for PCR amplification were 94°C for 2 min followed by 30 cycles of 94°C for 30 s, 60°C for 30 s and 72°C for 1 min and a final extension at 72°C for 10 min. PCR products were resolved on 3% agarose gels and cDNA between 60 and 200 nt were purified using a Gel and PCR Clean-Up System (Promega). Purified cDNA was ligated to 454 adapters and analyzed directly by 454 sequencing at the Joint Genome Institute, Walnut Creek, CA, USA.

CPA-sRNA data analysis

We obtained 127 330 raw reads in a FASTA format from a 454 sequencing run. 454 sequencing adapter and linkers at 5'- and 3'-ends were removed from raw reads and the remaining sequences were named CPA-sRNAs. Overall, we obtained a total of 80 111 CPA-sRNAs from mycelia with a size of ≥10 nts. We retained 25 389 reads with a size between 16 and 218 nts for matching to V6 *M. oryzae* genome assembly (GenBank ID: NZ_AACU00000000.2) (13). A detailed matching analysis was carried out using stringent BLASTN criteria of 80% coverage and 98% of sequence identity. We also utilized *Magnaporthe* transcriptome data (14,15) including ESTs, MPSS tags and RL-SAGE tags to annotate CPA-sRNAs. All the genomic features (contigs, genes, tRNAs, rRNAs, snRNAs, repeats, mitochondrial genome) and transcriptomic data (ESTs, SAGE, MPSS) were visualized in a genome browser based on gbrowse (17).

Defining the transcriptional unit

To define the transcriptional start and stop sites for protein-coding genes, we devised two approaches. First, we assigned a 5'-transcription start site (TSS) and 3'-transcription termination site (TTS) to gene models supported by ESTs. This provided a TSS and TTS for 2558 and 2551 genes, respectively. For the remaining annotated genes, we defined UTRs as 500 bp from start and stop codons. This is likely a slight overestimate of the average actual UTR length for protein-coding genes, but a value of 500 bp captured the vast majority of TUs. The average 5'-UTR for gene models supported by EST evidence was 327 nt. For other RNA species we defined the 5'-TSS and 3'-TTS as the first and last nucleotide of the mature RNA. For tRNAs, we used 150-nt upstream from 5'-mature tRNA for the 5'-leader and 150-nt downstream from 3'-mature tRNA for the 3'-terminator region.

Alignments, read counts and prorating data

CPA-sRNAs may align to the genome one or more times. The genomic location of each alignment may correspond to features such as genes, tRNA, rRNA or transposable elements. Thus the alignments were used to map CPA-sRNAs to genomic features. We pursued three methods for describing CPA-sRNA mapped genomic data (alignment counts, read counts and prorating) that account for ambiguity in determining the genomic origin

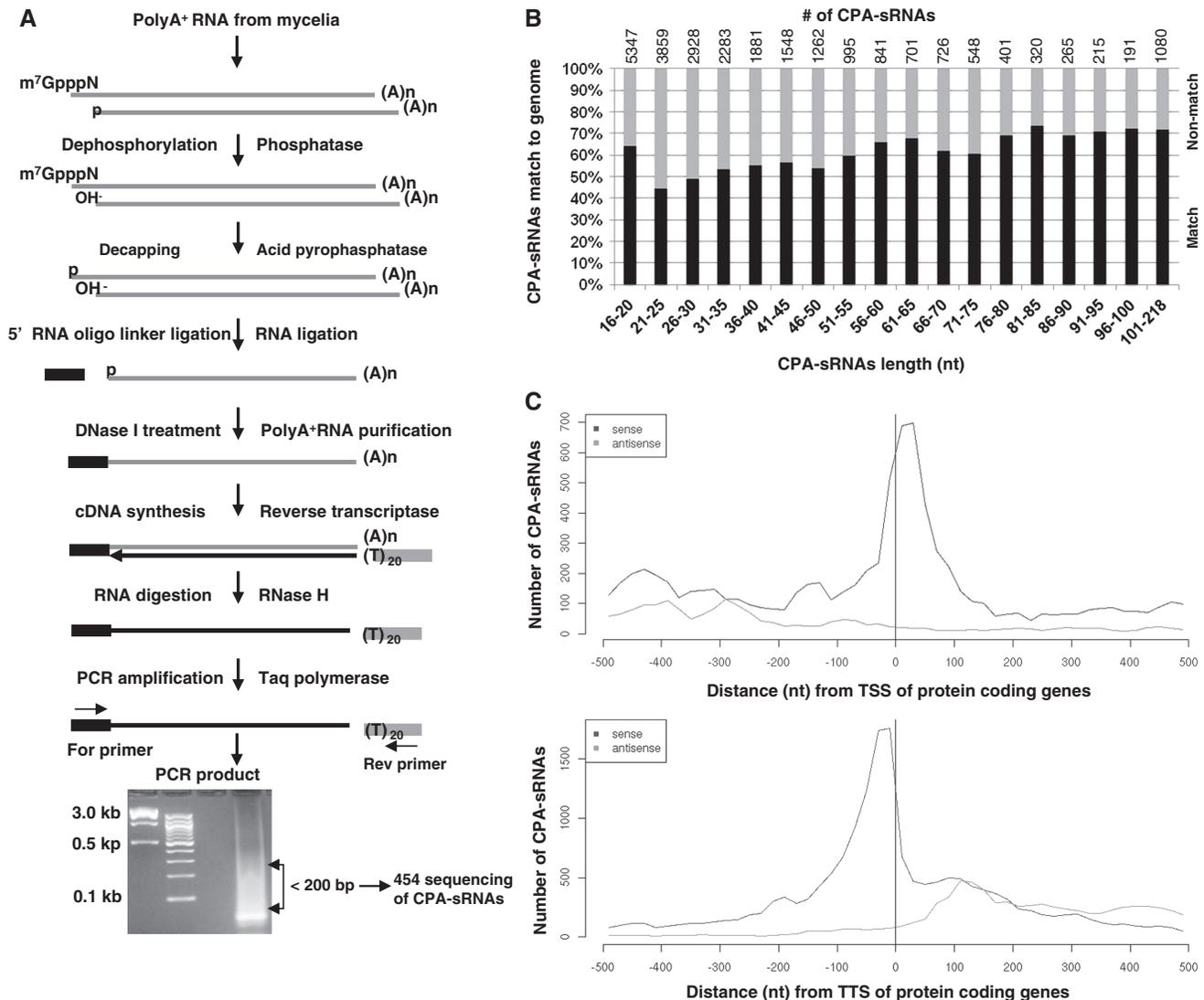


Figure 1. CPA-sRNA isolation and size distribution. (A) Strategy for CPA-sRNA preparation from mycelial total RNA. The protocol ensures capture of RNA species that possess both a 5'-cap and a 3'-polyadenylated tail. The first treatment with BAP prevents RNA containing a 5'-free phosphate from being able to ligate to the 5'-linker. The use of (dT)₂₀VN oligo for single-strand cDNA priming allows cDNA to be synthesized exclusively from RNA containing polyA. Following amplification by PCR, small cDNAs (<200 nt) were purified from a 3% agarose gel and subjected to 454 pyrosequencing. (B) Size distribution of CPA-sRNAs (≥ 16 nt) that matched to the *M. oryzae* genome (BLAST criteria: $\geq 80\%$ coverage and $\geq 98\%$ sequence identity). (C) CPA-sRNAs mapped to annotated protein-coding TUs. A vertical line represents the TSS and TTS for protein-coding genes.

of each CPA-sRNA. Alignment counts are the simple summation of all CPA-sRNA alignments to a given genomic feature. Since CPA-sRNAs may align multiple times to the genome, use of alignment counts alone might result in over counting. This is most evident with CPA-sRNAs that map to transposable elements—there are 20 671 alignments to transposable elements that originate from only 325 CPA-sRNAs. We addressed this issue by defining read counts, such that each CPA-sRNA is counted only once for a given genomic feature to which it maps. As CPA-sRNAs may map to multiple features, the use of read counts does not directly reflect the CPA-sRNAs origin (multiple mappings arise from either multiple alignments to multiple features or a single alignment spanning adjacent

features). We further refined our approach by taking into consideration multiple mappings. We prorated the counts for CPA-sRNAs by apportioning the counts across any multiple alignments and any features associated with that alignment (prorating). This was done as an iterative process: first, each CPA-sRNA was assigned a weight based on the number of copies found in sequencing data. Second, the weight from a given read was divided evenly between its genomic alignments. Third, each feature within a given alignment was given an equal portion of that alignment's weight. Last, each sub-feature divided the weight of the parent feature (sub-features exist as components of a feature—i.e. an exon is a sub-feature of gene). Summation of the apportioned CPA-sRNA weights for a given feature yields a balanced portrayal of

CPA-sRNA coverage for that feature and summation of values for sub-features equals that of its feature. Supplementary Figure S1 provides a visualization of prorating using hypothetical examples.

Purification of 5'-methylguanosine-capped RNA

5'-methylguanosine-capped transcripts were purified using recombinant eIF4E_{K119A}, which binds 5'-m⁷GpppN RNA caps with a 10- to 15-fold higher affinity than wild-type eIF-4E (18,19). GST-tagged eIF4E_{K119A} protein was bound to glutathione agarose beads (4E-beads) for 1 h at room temperature in PBS. The 4E-beads were washed in the binding buffer: 10 mM KHPO₄, pH 8.0, 100 mM KCl, 2 mM EDTA, 5% glycerol, 0.005% Triton X-100, 1.3% poly(vinyl alcohol) 98–99% hydrolyzed (Aldrich), 1 mM DTT and 20 U/ml RNase inhibitor (Ambion). About 120 µg of total RNA was heat denatured, diluted in the binding buffer, added to 200 µl (packed bead volume) of 4E-beads in a siliconized tube (Genemate, ISC BioExpress) and mixed for 1 h at room temperature. Samples were briefly centrifuged to pellet the beads with bound RNA and washed three times (5 min each) by mixing at room temperature in the binding buffer. The bound RNA on 4E-beads was phenol/chloroform extracted, precipitated and dissolved in RNase free water. The quantity of 5'-methylguanosine-capped RNA was measured by NanoDrop (Thermo Fisher) analysis and its integrity was determined with an Agilent 2100 Bioanalyzer.

3'-RACE analysis of CPA-sRNAs using 5'-capped RNA

5'-methylguanosine-capped RNA was treated with DNase I (NEB) to remove any contaminating genomic DNA. cDNA was synthesized in 20 µl reactions by adding the following reagents: 1 µg of 5'-methylguanosine-capped RNA, 50 picomole of 3'-oligo(dT)₂₀VN primer, 5 mM of dNTPs, 1 U of RNaseOut (Invitrogen) and 5 U of Superscript III (Invitrogen). Supplementary Table S1 lists all primer sequences used in this study. The reverse transcription reaction was incubated at 42°C for 2 h and heat inactivated. For evaluating CPA-sRNAs in 5'-methylguanosine-capped cDNA, PCR amplification was performed using a forward primer specific to the 5'-end of CPA-sRNAs of interest described in the 'Results' section and a reverse primer specific to 3'-oligo(dT)₂₀VN linker. PCR was done with high fidelity Platinum Taq DNA polymerase (Invitrogen) and under the following conditions 94°C for 2 min followed by 35 cycles at 94°C for 30 s, 55°C for 30 s, 72°C for 30 s and a final extension at 72°C for 5 min. PCR products were resolved on a 3% agarose gel, purified and cloned into the pGEM-T easy vector according to supplier's procedure (Promega). About 20 randomly selected white colonies were sequenced using the Sanger method.

cDNA library characterization

Northern blot analyses were conducted using total RNA, RNA purified using eIF4E_{K119A} or oligo(dT) columns and separated on 15% denaturing polyacrylamide gels.

Blots were hybridized with [³²P]ATP-labeled oligo(dT)₂₀ probes. To document that the cDNA contained long as well as short full-length cDNAs, we confirmed the presence of the full-length actin gene (MGG_03982.6) using specific PCR primers prior to size selection (Supplementary Figure S2).

Correlation analysis of CPA-sRNAs with mycelial gene expression and MPSS and SAGE tags

Correlation analyses were conducted using normalized signal intensity values of microarray data for *M. oryzae* mycelia grown in complete media for 48 h and a further 12 h in minimal media (NCBI GEO Accession #; GSE2716, Sample ID #s; GSM 52525, GSM 52524, GSM 52520) with the number of assigned CPA-sRNAs using JMP (SAS Institute) software. Analyses were conducted for both CPA-sRNAs mapping in the sense and antisense orientation with expression values for individual genes. Genes were also grouped into 100 bins based on gene expression and the relationship between the mean gene expression and the mean number of CPA-sRNAs per bin compared. Likewise, the relationship between MPSS or SAGE tags, which were both derived from RNA extracted after 72 h growth on complete media, and CPA-sRNAs was determined by comparing the mean number of tags and the mean number of CPA-sRNA per bin. GO annotations for *M. oryzae* genes were obtained from the previously published work (20).

RESULTS

CPA-sRNA discovery

Full-length cDNA was constructed from mycelial RNA, which was separated on an agarose gel and the fraction <200 nt subjected to 454 sequencing. A total of 127 330 reads were obtained, from which 25 389 CPA-sRNAs (≥16 nt; excluding 3'-polyA sequences) were further analyzed (Supplementary Table S2). Of the CPA-sRNAs, 57.4% (14 547) mapped to version 6 of the *M. oryzae* genome (BLASTN criteria of >80% coverage and >98% sequence identity). Interestingly, 84% (12 235) of CPA-sRNAs mapped to unique loci and 16% (2354) mapped to multiple locations in the genome. 10 265 (9780 prorated) CPA-sRNAs mapped to protein-coding TUs (13), and the remainder mapped to intergenic regions, transposable elements, rRNAs, tRNAs and snRNAs (Table 1 and Supplementary Table S3). 2778 (2498 prorated) CPA-sRNAs mapped to intergenic regions of *M. oryzae*. Of these, 1130 CPA-sRNAs overlapped EST or SAGE or MPSS sequences (Table 2). CPA-sRNAs ranged in length from 16 to 218 nt with a mean of 41 nt (Figure 1B).

CPA-sRNA validation

To validate CPA-sRNAs, 5'-methylguanosine-capped RNA was purified from total RNA using a high affinity variant of eIF-4E, which was previously used to prove

Table 1. Distribution of CPA-sRNAs mapped to genomic and mitochondrial features

| | Read Count ^a | | | Prorated ^b | | | Features ^c | | |
|---------------------|-------------------------|-------|-----------|-----------------------|-------|-----------|-----------------------|--------|--------------|
| | Total | Sense | Antisense | Total | Sense | Antisense | Mapped | Total | Coverage (%) |
| Genes | 10 265 | 8894 | 3507 | 9780 | 7579 | 2201 | 4327 | 11 043 | 39 |
| Introns | 681 | 456 | 313 | 378 | 239 | 139 | 467 | 19 651 | 2 |
| Exons | 9985 | 8685 | 3386 | 9401 | 7340 | 2062 | 4977 | 30 705 | 16 |
| 5'-UTR ^d | 2981 | 2526 | 705 | 2323 | 1967 | 356 | 1325 | 11 241 | 12 |
| EST Supp. | 1247 | 1201 | 46 | 1105 | 1074 | 31 | 375 | 2558 | 15 |
| Unsupp. | 1853 | 1384 | 660 | 1217 | 893 | 325 | 950 | 8683 | 11 |
| CDS | 2227 | 1603 | 779 | 1590 | 1090 | 500 | 1581 | 11 054 | 14 |
| EST Supp. | 1676 | 1233 | 518 | 1166 | 847 | 320 | 1085 | 5199 | 21 |
| Unsupp. | 731 | 426 | 340 | 423 | 243 | 180 | 496 | 5855 | 8 |
| 3'-UTR | 6260 | 5653 | 2276 | 5489 | 4283 | 1206 | 2597 | 11 076 | 23 |
| EST Supp. | 3772 | 3643 | 148 | 2982 | 2887 | 94 | 801 | 2551 | 31 |
| Unsupp. | 3636 | 2090 | 2131 | 2507 | 1397 | 1111 | 1796 | 8525 | 21 |
| tRNA ^e | 425 | 394 | 31 | 289 | 1396 | 6 | 287 | 341 | 84 |
| 5'-Leader | 261 | 237 | 24 | 192 | 191 | 1 | 151 | 341 | 44 |
| Mature | 227 | 226 | 1 | 188 | 188 | 0 | 274 | 341 | 80 |
| 3'-Term | 100 | 93 | 7 | 41 | 36 | 5 | 186 | 341 | 55 |
| rRNA | 1741 | 1740 | 1 | 1643 | 1642 | 1 | 47 | 48 | 98 |
| 5.8s | 82 | 82 | 0 | 82 | 82 | 0 | 3 | 3 | 100 |
| 8s | 66 | 66 | 0 | 46 | 46 | 0 | 41 | 41 | 100 |
| 18s | 661 | 660 | 1 | 593 | 592 | 1 | 1 | 2 | 50 |
| 28s | 932 | 932 | 0 | 922 | 922 | 0 | 2 | 2 | 100 |
| snRNA | 24 | 24 | 0 | 16 | 16 | 0 | 5 | 16 | 31 |
| Transp. Elements | 379 | 325 | 102 | 320 | 278 | 42 | 2087 | 3448 | 61 |
| Intergenic | 2778 | 2778 | – | 2498 | 2498 | – | – | – | – |
| Mitochondria | 43 | 41 | 3 | 43 | 40 | 3 | 14 | 37 | 36 |
| Genes | 11 | 8 | 3 | 8 | 6 | 3 | 8 | 15 | 53 |
| CDS | 3 | 3 | 0 | 3 | 3 | 0 | 2 | 16 | 13 |
| tRNA | 3 | 3 | 0 | 2 | 2 | 0 | 4 | 20 | 20 |
| Mature | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 20 | 5 |
| rRNA | 31 | 31 | 0 | 30 | 30 | 0 | 2 | 2 | 100 |
| Intergenic | 3 | 3 | – | 3 | 3 | – | – | – | – |

^aRead Count is the number of CPA-sRNAs that map on to particular genomic features. Note: values for genome features are typically less than the sum of component sub-features due to reads mapping to multiple locations and/or features.

^bProrating divides the weight of any given read between alignments and features. This takes into account situations where reads map to multiple locations and/or are annotated with more than one feature (see 'Materials and Methods' section and Supplementary Figure S1 for more detail). Both read counts and prorated are divided into sense/antisense with respect to the feature mapped.

^cFor each feature, the number of members mapped by a CPA-sRNA is given, followed by the total number of possible features and the percentage mapped.

^dValues indicate number of CPA-sRNAs (read count or prorated) that map to features (5'-UTR, CDS or 3'-UTR) within gene models that are supported by ESTs (EST supp.) or have no supporting EST evidence (Unsupp.).

^etRNA entries include alignments to pseudo-tRNA.

Table 2. Association of CPA-sRNAs with other transcriptional evidence

| | Genome ^a | | Intergenic | | Genes ^b | |
|-------------------------|-------------------------|-----------------------|------------|----------|--------------------|----------|
| | Read Count ^c | Features ^d | Read Count | Features | Read Count | Features |
| EST or ESS ^c | 10 277 | 7898 | 1130 | 744 | 10 149 | 4095 |
| ESS | 4747 | 2888 | 603 | 404 | 10 111 | 4035 |
| MPSS | 2293 | 1332 | 306 | 158 | 1626 | 712 |
| SAGE | 2932 | 1556 | 352 | 246 | 2150 | 1032 |
| EST | 8969 | 5010 | 756 | 340 | 8913 | 2951 |

^aCPA-sRNAs that overlap the sequence of ESS and/or ESTs at the same genome (or intergenic) location.

^bCPA-sRNAs that map to transcriptionally (ESS and/or EST) supported gene loci.

^cNumber of CPA-sRNAs that map to a given feature containing specified transcriptional evidence.

^dNumber of features containing specified transcriptional evidence associated with CPA-sRNAs.

^eESS (expressed short sequences) are either MPSS or SAGE tag annotations.

that specific miRNA precursors have 5'-methylguanosine caps (18,19). Gel blot analysis of 5'-methylguanosine-capped RNA using [γ ³²P]ATP labeled oligo(dT)₂₀ revealed a smear from 20 to 200 nt confirming diversity

of length and that CPA-sRNAs contain both a 5'-methylguanosine cap and polyA tract (Supplementary Figure S2). The presence of a 3'-polyA tail was confirmed by 3'-RACE on individual CPA-sRNAs, which were

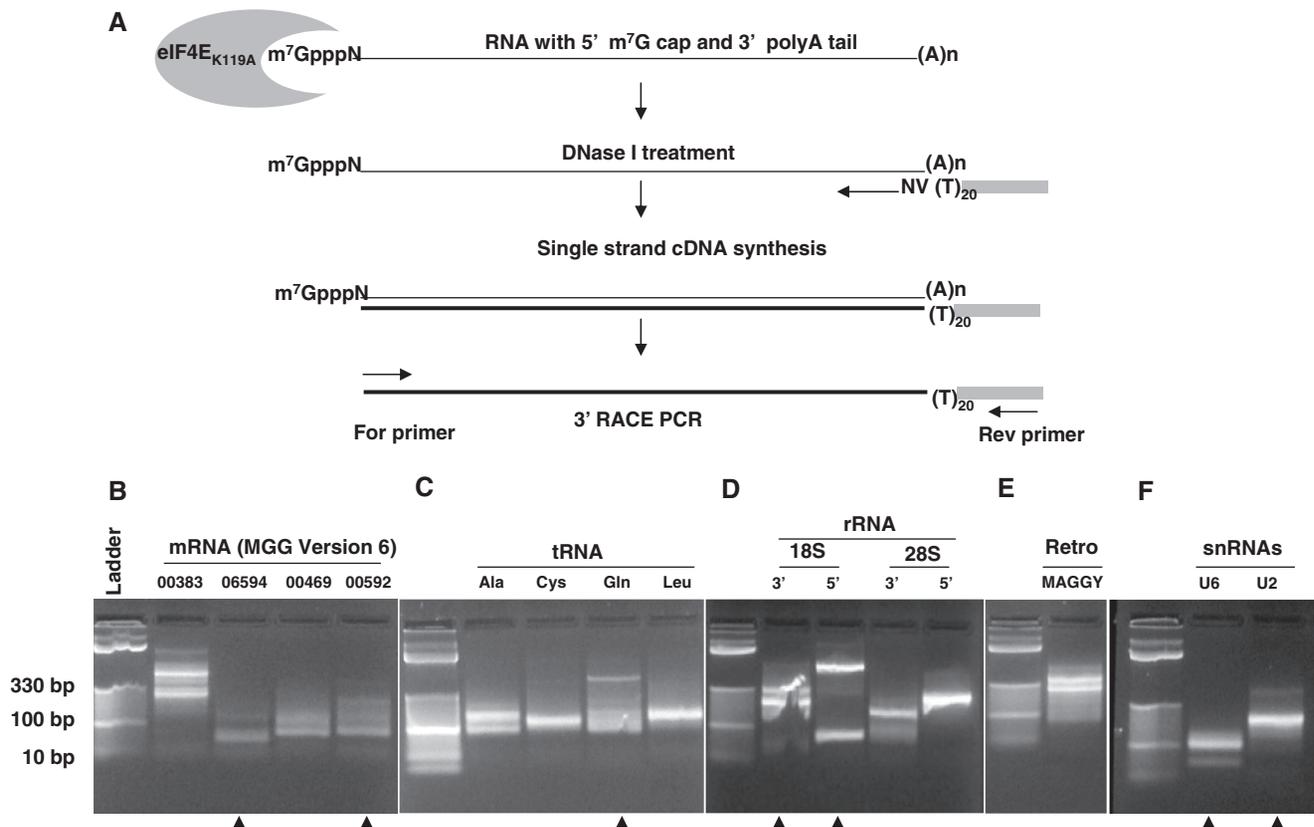


Figure 2. CPA-sRNA validation using 3'-RACE. (A) Total RNA from *M. oryzae* was used to purify 5' methylguanosine-capped RNAs using recombinant eIF4E_{K119A} bound to beads (21). 5' methylguanosine-capped RNA was treated with DNase I and single-stranded cDNA synthesized using an oligo (dT)₂₀VN primer. PCR amplification was performed using a forward primer to the 5'-end of specific CPA-sRNAs and reverse primer specific to the oligo (dT)₂₀VN linker. PCR products were analyzed on 3% agarose gels, bands eluted, cloned into pGEM-T vectors and Sanger sequenced. PCR products were resolved on a 3% agarose gels for (B) protein-coding mRNA (MGG_0383.6, MGG_6594.6, MGG_0469.6, MGG_0592.6, MGG_02597.6, MGG_07928.6, MGG_10680.6, MGG_14279.6 and MGG_01210.6); (C) tRNAs (Ala: MGG_20297.6, Cys: MGG_20209.6, Gln: MGG_20266.6 and Leu: MGG_20218.6); (D) rRNAs (18S and 28S); (E) snRNAs (U6 and U2) and (F) retroelements (MAGGY-LTR). A DNA ladder is shown on the left of each panel. Arrows indicate PCR products that were sequenced.

subsequently cloned and sequenced. Sequencing of 3'-RACE products confirmed that CPA-sRNAs mapped to protein-coding genes, transposable elements, snRNAs, tRNAs, rRNA genes and to intergenic locations and is described in more detail below (Figure 2B–F and Supplementary Figure S2).

We also examined the genomic context of CPA-sRNAs to exclude the possibility that they may have arisen from loci corresponding to longer RNAs rich in adenosine. 83% (12 096 out of 14 547) of CPA-sRNAs aligned to genome regions lacking adenosine enrichment (≥ 5 As) (data not shown), indicating that most CPA-sRNAs were not derived from internal poly-adenosine sequences of transcribed regions. We found that many CPA-sRNAs mapped to MPSS and SAGE tags derived from 3'-polyadenylated RNA located in intergenic regions and genes (Table 2). These tag associations were previously unexplained but in light of the present findings, they were likely derived from 3'-polyadenylated small RNAs. Taken together, these data from different approaches provide compelling evidence that CPA-sRNAs exist in fungal tissue and represent a distinct class of small RNAs.

CPA-sRNAs associate with transcription termini of protein-coding genes

A total of 10 265 (9780 prorated) CPA-sRNAs mapped to 4327 (39% of the total number) predicted protein-coding mRNAs (TUs), with more than a quarter (3507, 2201 prorated) mapping in the antisense orientation (Table 1). The majority of CPA-sRNAs mapped to UTRs (2981 (2323 prorated) to 5'-UTRs and 6260 (5489 prorated) to 3'-UTRs), whereas only 681 (378 prorated) mapped to introns. Examination of sense CPA-sRNAs mapping to 5'- or 3'-UTRs revealed that the vast majority associated with the transcript initiation (TSS) or termination (TTS) site, respectively (Figure 1C). CPA-sRNAs were predominantly (4095 out of 4327) associated with genes supported by ESTs, MPSS and RL-SAGE tags (Table 2). A 3'-RACE was used to confirm the presence of CPA-sRNAs for nine randomly selected protein-coding genes (Figure 2B), which included *S*-adenosyl methionine synthetase (MGG_0383.6), chitinase 18–11 (MGG_06594.6), Sad1/UNC domain-containing protein (MGG_00469.6), cell wall glucanoyl transferase Mwg1 (MGG_00592.6), yjeF-related protein (MGG_02597.6), ubiquitin (MGG_07928.6), 40S ribosomal protein S24

(MGG_10680.6), glutamine synthetase (MGG_14279.6) and nuclear encoded mitochondrial hypoxia responsive domain containing protein (MGG_01210.6). The expected 3'-RACE product size of 80–200 nt were obtained for all nine genes. Sequencing of cloned 3'-RACE products confirmed they aligned with CPA-sRNAs obtained from pyrosequencing, including the splice junction for chitinase 18–11 (MGG_06594.6) (Figure 3A).

To identify a possible role of CPA-sRNAs, we correlated their abundance with mycelia gene expression. Overall, we observed a positive correlation between CPA-sRNAs mapping in the sense orientation and mycelial gene expression, although not all individual genes followed this pattern (Figure 4A and B and Supplementary Table S4). Notably, the most highly expressed group of genes had highest numbers of mapped CPA-sRNAs. Inspection of 127 genes with ≥ 10 CPA-sRNAs mapped in the sense orientation showed that nearly all were functionally assigned with gene ontology (GO) terms involved in metabolism, with 67 (53%) being assigned to mycelial development and 43 (34%) to translation [(20) Figure 4E and Supplementary Table S5]. Of the latter, most (42) were assigned to structural components of the ribosome (Figure 4F). Further analysis confirmed a strong positive correlation between CPA-sRNAs and gene expression for all (65) annotated structural ribosomal proteins (see asterisk in Figure 4A and Supplementary Table S4). In contrast, antisense CPA-sRNAs did not map primarily to TSS and TTS, nor was there evidence supporting a correlation with gene expression (Figures 1C and 4C and D). In addition, we also observed a strong positive correlation between both sense-mapping MPSS or SAGE tags and sense-mapped CPA-sRNAs (Supplementary Figure S3).

CPA-sRNAs derived from RNA Pol I- and Pol III-transcribed genes

We detected 425 (289 prorated) CPA-sRNAs that mapped to 287 tRNA loci, with 31 (6 prorated) that mapped in the antisense orientation (Table 1, Supplementary Table S3). Of the 425 CPA-sRNAs, 114 mapped to pseudo-tRNAs, of which there are 141 essentially identical copies in the *M. oryzae* genome. Most CPA-sRNAs mapped around the beginning or end of the mature tRNA (Figure 3B and Supplementary Figure S4A). Several CPA-sRNAs corresponded to the entire tRNA, whereas others were shorter or longer than the corresponding tRNA. A number of CPA-sRNAs mapped to positions ~ -50 and $\sim +50$ nt from 5'- and 3'-ends, respectively, of the mature tRNA locus and likely correspond to the pre-tRNA transcript. A 3'-RACE confirmed CPA-sRNAs for five tRNAs (Figure 2C and Supplementary Figure S4B). The expected 3'-RACE PCR and sequence products were obtained for Ala tRNA (MGG_20297.6), Cys tRNA (MGG_20209.6), Gln tRNA (MGG_20266.6) and Leu tRNA (MGG_20218.6). The 3'-RACE and sequencing also revealed CPA-sRNAs corresponding to all three Pro tRNA paralogs (MGG_20065.6; MGG_20044.6 and MGG_20298.6). Analysis of CPA-sRNAs mapping to

rRNA revealed 1675 (1597 prorated) that mapped to 18S-5.8S-28S rRNA repeat locus (Table 1 and Supplementary Figure S5A and B). We obtained diverse CPA-sRNAs for 5.8S rRNA, many of which were supported by SAGE tags (Supplementary Figure S5C). We found 66 (46 prorated) CPA-sRNAs for the 8S rRNA locus, which has multiple copies dispersed throughout the *M. oryzae* genome (Table 1 and Supplementary Table S3). CPA-sRNAs associated with the 5'-end of 18S rRNA and 28S rRNA were validated by 3'-RACE analysis and sequencing (Figures 2D and 3C).

CPA-sRNAs derived from repetitive elements

More than 10% of the *M. oryzae* genome consists of repetitive elements (13). We found 379 (320 prorated) CPA-sRNAs mapping to transposable elements, including 102 (42 prorated) antisense mappings (Table 1 and Supplementary Table S3). Fifty (46 prorated) CPA-sRNAs mapped specifically to the LTR region of MAGGY, a gypsy-like element linked with pathogenicity (21), which were validated by 3'-RACE analysis (Figure 2E and Supplementary Figure S6). The LTR (250 nt) of MAGGY is structurally similar to retroviruses, which acts as a transcription initiator and terminator (21). We also identified a number of CPA-sRNAs that mapped to other retro-elements including MGR583 or SINE element (218; 155 prorated), PYRET (40; 29 prorated), OCCAN (11; 8 prorated) and MOLLY (5; 3 prorated), as well as to DNA transposon, POT2 (64; 20 prorated). Overall, we observed that CPA-sRNAs mapped primarily to the LTR (putative TSS and TTS) of retro-transposons while CPA-sRNAs were distributed across the entire transcript of DNA-transposons.

Few CPA-sRNAs map to mitochondrial TUs

Only 43 (43 prorated) CPA-sRNAs aligned to the mitochondria genome (~ 35 kb) (Table 1; Supplementary Table S3 and Figure S7). Of these, 31 (30 prorated) mapped to rRNA genes (*rrnL*, large subunit ribosomal RNA and *rrnS*, small subunit ribosomal RNA). In contrast to nuclear genes, only 11 (8 prorated) CPA-sRNAs were found to align to the coding sequence of two mitochondrial genes (MGG_21013.6, cytochrome b and MGG_21007.6, ATP synthase subunit 6). We did not detect CPA-sRNAs for mitochondrial tRNAs except for the SeC (MGG_21117.6) and Arg tRNA (MGG_21120.6) loci. Mitochondrial transcripts are typically transcribed by a simpler RNA polymerase, homologous to the bacteriophage T7/T3 RNA polymerase subunit as compared to the more complex nature of nuclear RNA Pol I, II and III (22). Although, we have limited knowledge of the mitochondrial RNA polymerase in filamentous fungi, our data suggest that CPA-sRNAs are biased toward nuclear RNA polymerase-derived transcripts.

Analysis of CPA-sRNAs not matching the *M. oryzae* genome sequence

Deep sequencing projects of small RNA species typically reveal many sequences that do not align to the genome

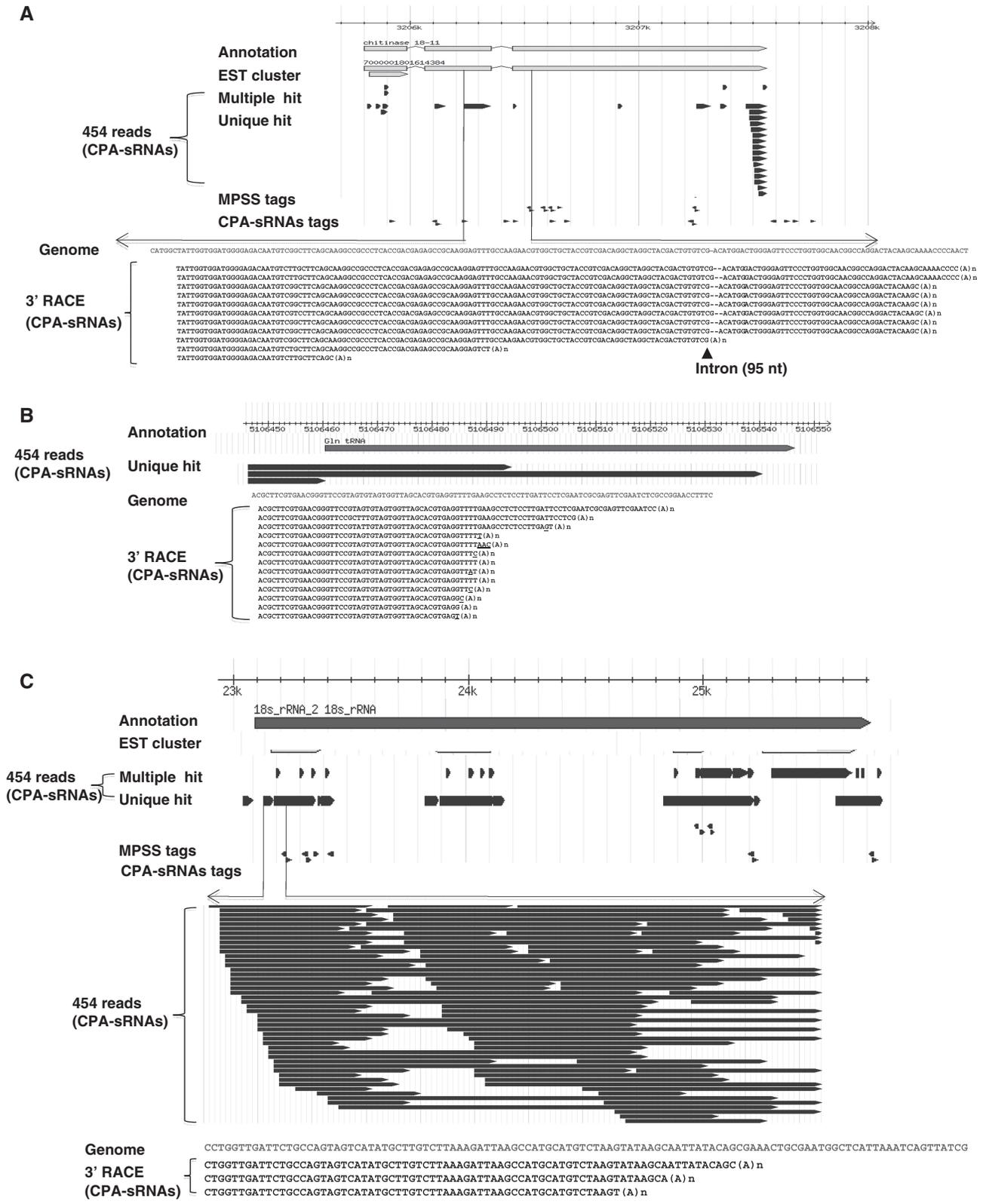


Figure 3. CPA-sRNA sequences validated for mRNA, tRNA and rRNA loci. (A) 454 and 3' RACE PCR clone sequence location at the chitinase 18–11 gene (MGG_06594.6). A dashed line in the 3'–RACE sequence data represents the absence of intronic sequence (95 nt) for the chitinase gene. (B) CPA-sRNAs associated with Gln tRNA locus (MGG_20266.6). Underlined sequence data represent non-templated nucleotide. (C) CPA-sRNAs clustering at the 5'–end of 18S rRNA locus.

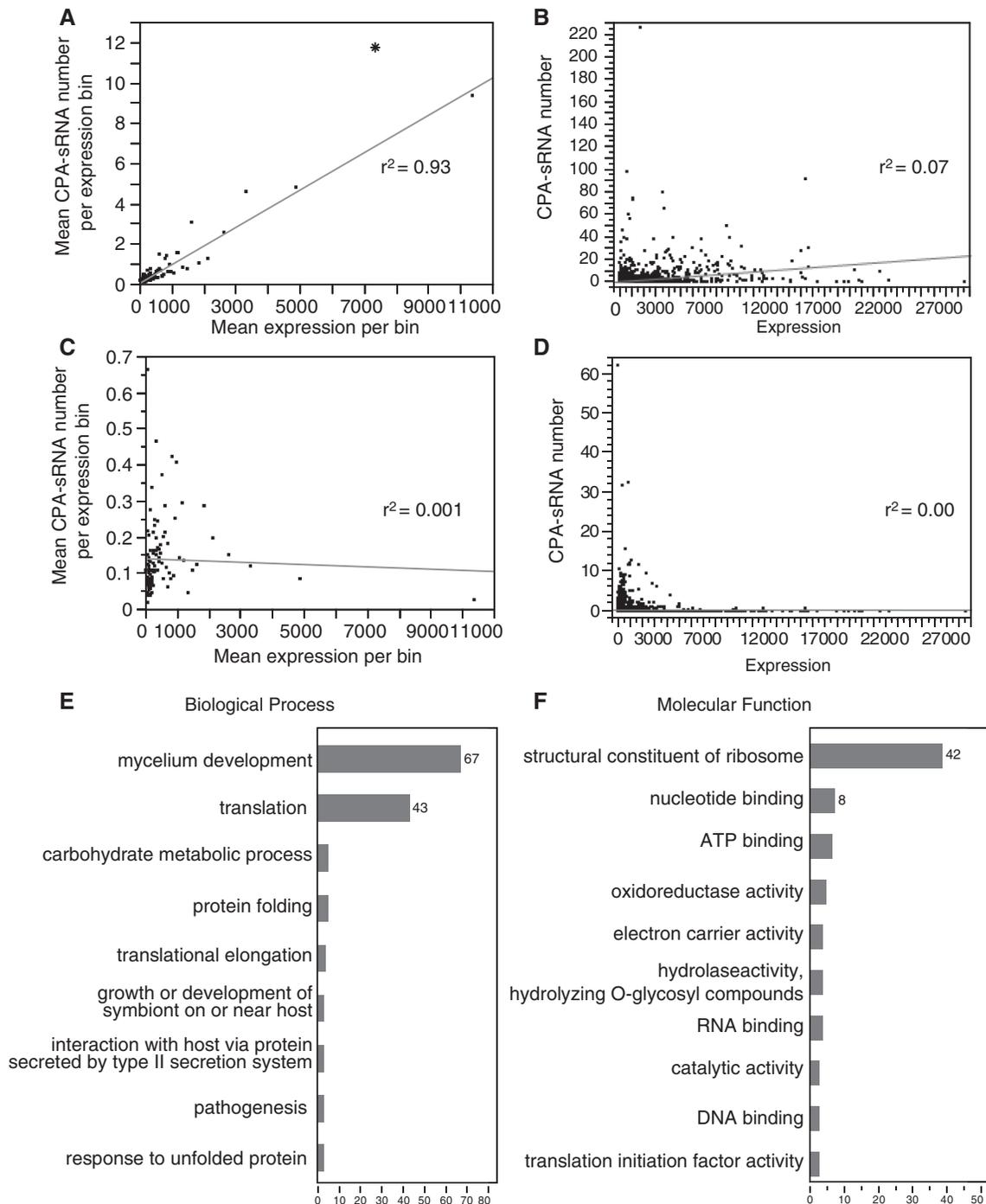


Figure 4. Correlation of expression and GO annotation of genes with mapped CPA-sRNAs. Correlation analysis of gene expression and number of mapped CPA-sRNAs based on (A, sense mapping; B, antisense mapping) bins and (C, sense mapping; D, antisense mapping) individual genes. Genes were grouped into 100 bins based on mycelial gene expression; each bin contains 99 genes. Average gene expression per bin was plotted versus the average number of CPA-sRNAs for each bin. Asterisks indicates the average number of CPA-sRNAs and expression for 65 genes annotated as ribosomal structural proteins. (E) Biological process and (F) molecular function gene ontology annotation for 127 genes with ≥ 10 mapped CPA-sRNAs. Numbers on top of vertical bars indicate number of genes per category.

and are often disregarded as sequencing errors. However, evidence now suggests that these molecules result from post-transcriptional RNA modifications (23–25). Gowda *et al.* (15) reported previously that a large fraction of MPSS and RL-SAGE tags did not match the *M. oryzae* genome. Similarly, we found that $\sim 42\%$ of CPA-sRNAs

(10 824 out of 25 389) could not be aligned to the genome. Interestingly, we identified 388 CPA-sRNAs that matched only to *M. oryzae* ESTs but not to the genome (data not shown). In addition, 1585 CPA-sRNAs matched SAGE and/or MPSS tags but were unaligned to the genome. Further examination revealed that a small number (284)

of unaligned CPA-sRNAs aligned to the genome sequence of strains P123 and Y34 (Y. Peng *et al.* unpublished data) suggesting that some CPA-sRNAs map to gaps in the 70–15 reference strain genome sequence. Manual inspection revealed evidence of 46 chimeric CPA-sRNAs (Supplementary Table S6), some of which were possibly derived from the fusion of RNAs from two or more non-contiguous genomic locations (3). We observed chimeric RNA fusions in head-to-head or head-to-tail orientations of the same exon, two exons of the same gene, exon–exon junctions of two genes, protein-coding region–intergenic or rRNA, intergenic–intergenic, rRNA–intergenic, rRNA–rRNA and rRNA–tRNA. Twenty-five percent of chimeric CPA-sRNAs were derived from exonic regions, where fusion points coincide with the canonical splicing sites (Supplementary Table S6). We also found evidence for short homologous sequences (SHS) for 44% of chimeric CPA-sRNAs, similar to reports of animal (26) and plant (27) chimeric RNAs. Further examination of these chimeric CPA-sRNAs revealed a high rate (96%) of non-template nucleotide additions, 20% (9 out of 45) of which had 1–3 non-template nucleotides internally at the point of RNA fusion. We also observed non-templated nucleotides in many genome-matched CPA-sRNAs. For example, 54% of Gln tRNA-associated CPA-sRNAs (7 out of 13) had non-templated nucleotides (G, AAC, A, C, C, T) at the 3'-end (Figure 3B). Similarly for 5.8S rRNA CPA-sRNAs, 81% (13 out of 16) and 88% (14 out of 16) had non-templated nucleotides at the 5'- and 3'-regions, respectively. Non-template sequence diversity at the 3'-end of 5.8S rRNA is supported by 26 RL-SAGE tags that matched the 3'-ends of CPA-sRNAs (15). Thus chimeric and non-templated nucleotide additions may explain why at least some CPA-sRNAs do not align to the genome.

DISCUSSION

Magnaporthe oryzae has emerged as a model to study fungal pathogenesis due to its agronomic significance, genetic tractability, availability of genomic sequences and expression datasets including expressed sequence tags (EST), serial analysis of gene expression (SAGE) tags, massively parallel signature sequencing (MPSS) tags and microarray data (13–15). During the course of analyzing a full-length cDNA library, we identified CPA-sRNAs. These are a distinct class of small RNAs because these contain both 5'-methylguanosine-capped and 3'-polyadenylated and associate with TUs of RNA Pol I (rRNA), Pol II (mRNA/retrotransposons) and Pol III (snRNA/tRNA).

RNA Pol II transcripts have been intensively studied with respect to 5'- and 3'-end modifications (28). Nuclear capping occurs co-transcriptionally by adding a 5'-*N*-methyl guanosine with an inverted 5'-5'-triphosphate bridge to the first gene-encoded nucleotide of RNA Pol II transcripts. A PolyA tail is added to 3'-ends of RNA Pol II transcripts by post-transcriptional events. These modifications protect RNAs from nucleases, and signal for RNA export and translation. Additional capping in the

cytoplasm (29) and polyadenylation linked to nonsense-mediated mRNA decay (6) have been reported in eukaryotes. Recently, small ncRNAs such as CUTs and SUTs have been described in *S. cerevisiae* that appear to be the products of RNA Pol II and thus are likely capped and polyadenylated (6). This suggests that research on capping and polyadenylation is far from complete and further studies may shed further functions of capping and polyadenylation of small and long transcripts.

Although the ends of RNA Pol I and III transcripts are less well defined as compared to RNA Pol II, it has been shown in yeast recently that some of these elements contain 3'-polyA tails (30). Furthermore, small RNAs (21–400 nt) in animals (31,32) and plants (33) possess short stretches of adenosine (1–7 nt). The snRNA U6, which is transcribed by Pol III, carries a methylguanosine cap at the 5'-end (34). In our study, we obtained CPA-sRNAs for both U6 and U2 (Figure 2F). Recently small RNAs from humans have been reported to possess cap structures at the 5'-ends (35,36), many of which were associated with the TSS. Several reports have shown that some classes of small RNA (<200 nt) associate with RNA Pol II TSS and TTS (1,37,38). Our study provides detailed evidence that CPA-sRNAs are not only associated with TSS and TTS of protein-coding genes but are also associated with snRNAs, tRNAs, rRNAs and retrotransposons. qiRNAs, miRNAs and disiRNAs reported recently in *N. crassa*, however, do not appear to possess 5'- and 3'-modified ends (8,9).

We have no knowledge of the mechanisms of CPA-sRNA biogenesis; however, it is likely they are either derived directly from the genome or are processed from longer transcripts. In the former case, they could be the product of an uncharacterized RNA polymerase, such as Pol IV, a homolog of DNA-dependent RNA polymerase II (36,39), or Pol III (40), which are then processed, if necessary, by adding methylguanosine to the 5'-end and polyA to the 3'-ends, possibly by a cytosolic mechanism (29). Alternatively, CPA-sRNAs could be derived by the action of an endo-ribonuclease/splicing complex on long mRNAs releasing short RNA fragments, which might be spliced together, undergo addition of untemplated nucleotides before adding a 5'-methylguanosine cap and 3'-polyA tail.

Currently, we have no direct evidence for a biological function for CPA-sRNAs. However, a growing body of research has shown that non-protein-coding small RNAs modulate many biological processes, including chromosome replication, chromatin remodeling, transcription regulation, RNA processing and stability as well as protein stability and translocation (2,3,41). Several studies have shown that ncRNA transcription is most predominant at the promoters (or TSSs), and also occurs at intergenic regions as well as within genes. Our finding of a strong positive correlation between CPA-sRNAs and gene expression suggests CPA-sRNAs may play a positive role in gene regulation. Small RNAs complementary to promoter regions have been shown to activate gene expression in several different cellular contexts (42–44). These RNAs are typically biased toward genes with higher levels of expression (38). Similarly CUTs are also

typically positively correlated with gene expression (4). It is possible that CPA-sRNAs associate with transcriptionally active regions forming RNA–DNA hybrids, which create transcriptional bubbles (nucleosome-free single-stranded DNA) at the TSS and TTS (45). At TSS, CPA-sRNAs may facilitate transcription factor and RNA polymerase binding or act as primers for RNA synthesis, whereas CPA-sRNAs at TTS may block further transcription and facilitate the release of the RNA polymerase. On the other hand, as has been pointed out for CUTs, it is possible they simply reflect inefficient start site initiation and represent ‘transcriptional noise’ (6).

Finally, we provide an explanation of why at least some small RNAs do not align with the genome sequence. While we cannot discount that many may be the result of sequencing error, our analyses reveal that some are the products of fragment fusion or contain non-templated additions at their termini or point of fusion such that they no longer align to the genome using strict criteria. We suggest that RNA editing and/or posttranscriptional modification may be involved in generating CPA-sRNA diversity, thus increasing the complexity of the small RNA transcriptome. However, their origin and significance remain to be determined. While there remains much to be learned about CPA-sRNA biogenesis and function, their discovery and characterization add another fascinating chapter in genome and RNA biology.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We appreciate all the members in the fungal genomics laboratory for their feedback on this project. We also thank Malathy Krishnamurthy for help editing this manuscript.

FUNDING

Funding for open access charge: United States Department of Agriculture (Award #2005-04936 to R.A.D.); National Institutes of Health grant (CA63640 to C.H.H.).

Conflict of interest statement. None declared.

REFERENCES

- Kapranov,P., Cheng,J., Dike,S., Nix,D.A., Duttgupta,R., Willingham,A.T., Stadler,P.F., Hertel,J., Hackermuller,J., Hofacker,I.L. *et al.* (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, **316**, 1484–1488.
- Mercer,T.R., Dinger,M.E. and Mattick,J.S. (2009) Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.*, **10**, 155–159.
- Wilusz,J.E., Sunwoo,H. and Spector,D.L. (2009) Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev.*, **23**, 1494–1504.
- Neil,H., Malabat,C., d’Aubenton-Carafa,Y., Xu,Z.Y., Steinmetz,L.M. and Jacquier,A. (2009) Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature*, **457**, 1038–1042.
- Xu,Z.Y., Wei,W., Gagneur,J., Perocchi,F., Clauder-Munster,S., Camblong,J., Guffanti,E., Stutz,F., Huber,W. and Steinmetz,L.M. (2009) Bidirectional promoters generate pervasive transcription in yeast. *Nature*, **457**, 1033–1037.
- Harrison,B.R., Yazgan,O. and Krebs,J.E. (2009) Life without RNAi: noncoding RNAs and their functions in *Saccharomyces cerevisiae*. *Biochem. Cell Biol.*, **87**, 767–779.
- Thompson,D.M. and Parker,R. (2007) Cytoplasmic decay of intergenic transcripts in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.*, **27**, 92–101.
- Lee,H.C., Chang,S.S., Choudhary,S., Aalto,A.P., Maiti,M., Bamford,D.H. and Liu,Y. (2009) qiRNA is a new type of small interfering RNA induced by DNA damage. *Nature*, **459**, 274–277.
- Lee,H.-C., Li,L., Gu,W., Xue,Z., Crosthwaite,S.K., Pertsemliadis,A., Lewis,Z.A., Freitag,M., Selker,E.U., Mello,C.C. *et al.* (2010) Diverse pathways generate microRNA-like RNAs and dicer-independent small interfering RNAs in fungi. *Mol. Cell*, **38**, 1–12.
- Park,J.Y., Jin,J.M., Lee,Y.W., Kang,S. and Lee,Y.H. (2009) Rice blast fungus (*Magnaporthe oryzae*) infects Arabidopsis via a mechanism distinct from that required for the infection of rice. *Plant Physiol.*, **149**, 474–486.
- Sesma,A. and Osbourn,A.E. (2004) The rice leaf blast pathogen undergoes developmental processes typical of root-infecting fungi. *Nature*, **431**, 582–586.
- Wilson,R.A. and Talbot,N.J. (2009) Under pressure: investigating the biology of plant infection by *Magnaporthe oryzae*. *Nat. Rev. Microbiol.*, **7**, 185–195.
- Dean,R.A., Talbot,N.J., Ebbole,D.J., Farman,M.L., Mitchell,T.K., Orbach,M.J., Thon,M., Kulkarni,R., Xu,J.R., Pan,H.Q. *et al.* (2005) The genome sequence of the rice blast fungus *Magnaporthe grisea*. *Nature*, **434**, 980–986.
- Ebbole,D.J., Jin,Y., Thon,M., Pan,H.Q., Bhattarai,E., Thomas,T. and Dean,R. (2004) Gene discovery and gene expression in the rice blast fungus, *Magnaporthe grisea*: analysis of expressed sequence tags. *Mol. Plant – Microbe Interact.*, **17**, 1337–1347.
- Gowda,M., Venu,R.C., Raghupathy,M.B., Nobuta,K., Li,H.M., Wing,R., Stahlberg,E., Couglan,S., Haudenschield,C.D., Dean,R. *et al.* (2006) Deep and comparative analysis of the mycelium and appressorium transcriptomes of *Magnaporthe grisea* using MPSS, RL-SAGE, and oligoarray methods. *BMC Genomics*, **7**, 310.
- Gowda,M., Li,H.M., Alessi,J., Chen,F., Pratt,R. and Wang,G.L. (2006) Robust analysis of 5′-transcript ends (5′-RATE): a novel technique for transcriptome analysis and genome annotation. *Nucleic Acids Res.*, **34**, e126.
- Stein,L.D., Mungall,C., Shu,S.Q., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al.* (2002) The Generic Genome Browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
- Cai,X.Z., Hagedorn,C.H. and Cullen,B.R. (2004) Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA*, **10**, 1957–1966.
- Choi,Y.H. and Hagedorn,C.H. (2003) Purifying mRNAs with a high-affinity eIF4E mutant identifies the short 3′poly(A) end phenotype. *Proc. Natl Acad. Sci. USA*, **100**, 7033–7038.
- Meng,S., Brown,D.E., Ebbole,D., Torto-Alalibo,T., Oh,Y.Y., Deng,J., Mitchell,T. and Dean,R.A. (2009) Gene Ontology annotation of the rice blast fungus, *Magnaporthe oryzae*. *BMC Microbiol.*, **9**, S8.
- Nakayashiki,H., Kiyotomi,K., Tosa,Y. and Mayama,S. (1999) Transposition of the retrotransposon MAGGY in heterologous species of filamentous fungi. *Genetics*, **153**, 693–703.
- Tang,G.-Q., Paratkar,S. and Patel,S.S. (2009) Fluorescence mapping of the open complex of yeast mitochondrial RNA polymerase. *J. Biol. Chem.*, **284**, 5514–5522.
- Ebhardt,H.A., Tsang,H.H., Dai,D.C., Liu,Y., Bostan,B. and Fahlman,R.P. (2009) Meta-analysis of small RNA-sequencing

- errors reveals ubiquitous post-transcriptional RNA modifications. *Nucleic Acids Res.*, **37**, 2461–2470.
24. Iida, K., Jin, H. and Zhu, J.-K. (2009) Bioinformatics analysis suggests base modifications of tRNAs and miRNAs in *Arabidopsis thaliana*. *BMC Genomics*, **10**, 155.
 25. Pantano, L., Estivill, X. and Marti, E. (2010) SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells. *Nucleic Acids Res.*, **38**, e34.
 26. Li, X., Zhao, L., Jiang, H. and Wang, W. (2009) Short homologous sequences are strongly associated with the generation of chimeric RNAs in eukaryotes. *J. Mol. Evol.*, **68**, 56.
 27. Zhang, G., Guo, G., Hu, X., Zhang, Y., Li, Q., Li, R., Zhuang, R., Lu, Z., He, Z., Fang, X. *et al.* (2010) Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Res.*, **20**, 646–654.
 28. Shatkin, A.J. and Manley, J.L. (2000) The ends of the affair: capping and polyadenylation. *Nat. Struct. Biol.*, **7**, 838–842.
 29. Schoenberg, D.R. and Maquat, L.E. (2009) Re-capping the message. *Trends Biochem. Sci.*, **34**, 435–442.
 30. Ozsolak, F., Platt, A.R., Jones, D.R., Reifenger, J.G., Sass, L.E., McInerney, P., Thompson, J.F., Bowers, J., Jarosz, M. and Milos, P.M. (2009) Direct RNA sequencing. *Nature*, **461**, 814–818.
 31. Chen, Y.H., Sinha, K., Perumal, K., Gu, J. and Reddy, R. (1998) Accurate 3' end processing and adenylation of human signal recognition particle RNA and Alu RNA in vitro. *J. Biol. Chem.*, **273**, 35023–35031.
 32. Sinha, K.M., Gu, J., Chen, Y.H. and Reddy, R. (1998) Adenylation of small RNAs in human cells—development of a cell-free system for accurate adenylation on the 5'-end of human signal recognition particle RNA. *J. Biol. Chem.*, **273**, 6853–6859.
 33. Lu, S.F., Sun, Y.H. and Chiang, V.L. (2009) Adenylation of plant miRNAs. *Nucleic Acids Res.*, **37**, 1878–1885.
 34. Kiss, T. (2004) Biogenesis of small nuclear RNPs. *J. Cell Sci.*, **117**, 5949–5951.
 35. Fejes-Toth, K., Sotirova, V., Sachidanandam, R., Assaf, G., Hannon, G.J., Kapranov, P., Foissac, S., Willingham, A.T., Duttagupta, R., Dumais, E. *et al.* (2009) Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature*, **457**, 1028–1032.
 36. Nechaev, S., Fargo, D.C., dos Santos, G., Liu, L.W., Gao, Y. and Adelman, K. (2010) Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in *Drosophila*. *Science*, **327**, 335–338.
 37. Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W., Rahl, P.B., Flynn, R.A., Young, R.A. and Sharp, P.A. (2008) Divergent transcription from active promoters. *Science*, **322**, 1849–1851.
 38. Taft, R.J., Glazov, E.A., Cloonan, N., Simons, C., Stephen, S., Faulkner, G.J., Lassmann, T., Forrest, A.R.R., Grimmond, S.M., Schroder, K. *et al.* (2009) Tiny RNAs associated with transcription start sites in animals. *Nature Genet.*, **41**, 572–578.
 39. Mosher, R.A., Melnyk, C.W., Kelly, K.A., Dunn, R.M., Studholme, D.J. and Baulcombe, D.C. (2009) Uniparental expression of PolIV-dependent siRNAs in developing endosperm of *Arabidopsis*. *Nature*, **460**, 283–286.
 40. Roberts, D.N., Stewart, A.J., Huff, J.T. and Cairns, B.R. (2003) The RNA polymerase III transcriptome revealed by genome-wide localization and activity-occupancy relationships. *Proc. Natl Acad. Sci. USA*, **100**, 14695–14700.
 41. Storz, G. (2002) An expanding universe of noncoding RNAs. *Science*, **296**, 1260–1263.
 42. Janowski, B.A., Younger, S.T., Hardy, D.B., Ram, R., Huffman, K.E. and Corey, D.R. (2007) Activating gene expression in mammalian cells with promoter-targeted duplex RNAs. *Nat. Chem. Biol.*, **3**, 166–173.
 43. Li, L.C., Okino, S.T., Zhao, H., Pookot, D., Place, R.F., Urakami, S., Enokida, H. and Dahiya, R. (2006) Small dsRNAs induce transcriptional activation in human cells. *Proc. Natl Acad. Sci. USA*, **103**, 17337–17342.
 44. Morris, K.V., Santoso, S., Turner, A.M., Pastori, C. and Hawkins, P.G. (2008) Bidirectional transcription directs both transcriptional gene activation and suppression in human cells. *PLoS Genet.*, **4**, e1000258.
 45. Hovsepian, J.A. and Frenster, J.H. (2002) RNA-induced melting of DNA during selective gene transcription. *Mol. Biol. Cell*, **13**, 1343.