9720–9730 *Nucleic Acids Research, 2011, Vol. 39, No. 22* *Published online 31 August 2011*
doi:10.1093/nar/gkr684

Direct cloning of double-stranded RNAs from RNase protection analysis reveals processing patterns of C/D box snoRNAs and provides evidence for widespread antisense transcript expression

Manli Shen[1], Eduardo Eyras[2,3], Jie Wu[4], Amit Khanna[1], Serene Josiah[5], Mathieu Rederstorff[6], Michael Q. Zhang[7,*] and Stefan Stamm[1,*]

[1]Department of Molecular and Cellular Biochemistry, University of Kentucky, Lexington, KY, 40536, USA, [2]Universitat Pompeu Fabra, Dr Aiguader 88, E08003, [3]Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 23, E08010, Barcelona, Spain, [4]Cold Spring Harbor Laboratory, 11724, Cold Spring Harbor, NY, USA, [5]Shire Human Genetic Therapies, Lexington, MA 02421, USA, [6]Nancy Université/Biopôle, UMR 7214 AREMS CNRS-UHP, 9 avenue de la Forêt de Haye, 50500 Vandoeuvre-lès-Nancy, France and [7]MCB, UT Dallas, Richardson, TX 75080, USA And TNLIST, Tsinghua University, Beijing, China

Received May 4, 2011; Revised and Accepted August 4, 2011

ABSTRACT

We describe a new method that allows cloning of double-stranded RNAs (dsRNAs) that are generated in RNase protection experiments. We demonstrate that the mouse C/D box snoRNA MBII-85 (SNORD116) is processed into at least five shorter RNAs using processing sites near known functional elements of C/D box snoRNAs. Surprisingly, the majority of cloned RNAs from RNase protection experiments were derived from endogenous cellular RNA, indicating widespread antisense expression. The cloned dsRNAs could be mapped to genome areas that show RNA expression on both DNA strands and partially overlapped with experimentally determined argonaute-binding sites. The data suggest a conserved processing pattern for some C/D box snoRNAs and abundant expression of longer, non-coding RNAs in the cell that can potentially form dsRNAs.

INTRODUCTION

RNA:RNA interactions play an important role in gene regulation, as shown by the recognition of pre-mRNA splice sites by snRNPs, and the regulation of mRNA function by miRNAs (1). All RNAs undergo extensive processing and are typically generated from longer precursor molecules. Recent high-throughput sequencing (HTS) data showed that RNAs previously viewed as metabolically stable, such as C/D and H/ACA snoRNAs as well as tRNAs undergo further processing resulting in shorter RNA forms (2–8).

To fully understand how these RNAs are formed, it is necessary to clone them. One of the most precise ways to identify RNAs generated from a precursor RNA is to employ RNase protection analysis using a radioactively labelled antisense probe against the precursor. Hybridization of the probe to its target strand generates a dsRNA that is separated from other RNAs by removing all single-stranded RNAs using RNases. This method is well suited to study the processing of a defined larger RNA into smaller fragments, as these fragments can be detected by the shortening of the protected RNAs. RNase protection experiments are well established to give quantitative results.

Although RNase protection experiments are highly sensitive and selective, their use is hampered by the inability to directly clone the protected RNA fragments, which is due to the lack of appropriate double-stranded RNAs (dsRNA) modifying enzymes. Previously, only dsRNAs from viruses that can be produced in large quantities could be cloned (9) and cloning has been demonstrated as a proof of principle using *in vitro* transcribed RNAs and model viral dsRNAs (10).

To overcome this problem, we devised a technique to clone dsRNAs from standard RNase protection reactions. An overview of the method is given in Figure 1a. The

*To whom correspondence should be addressed. Tel: +01 859 3230896; Fax: +01 859 2572283; Email: stefan@stamms-lab.net

method allows the identification of RNAs that are generated by processing of precursor RNAs. We were able to establish the processing pattern of RNAs derived from a C/D box snoRNA, MBII-85 (SNORD 116 in humans). Unexpectedly, we found evidence for abundant expression of endogenous RNAs that could form double strands *in vivo*. These endogenous RNAs overlap with genome regions that show evidence for expression of

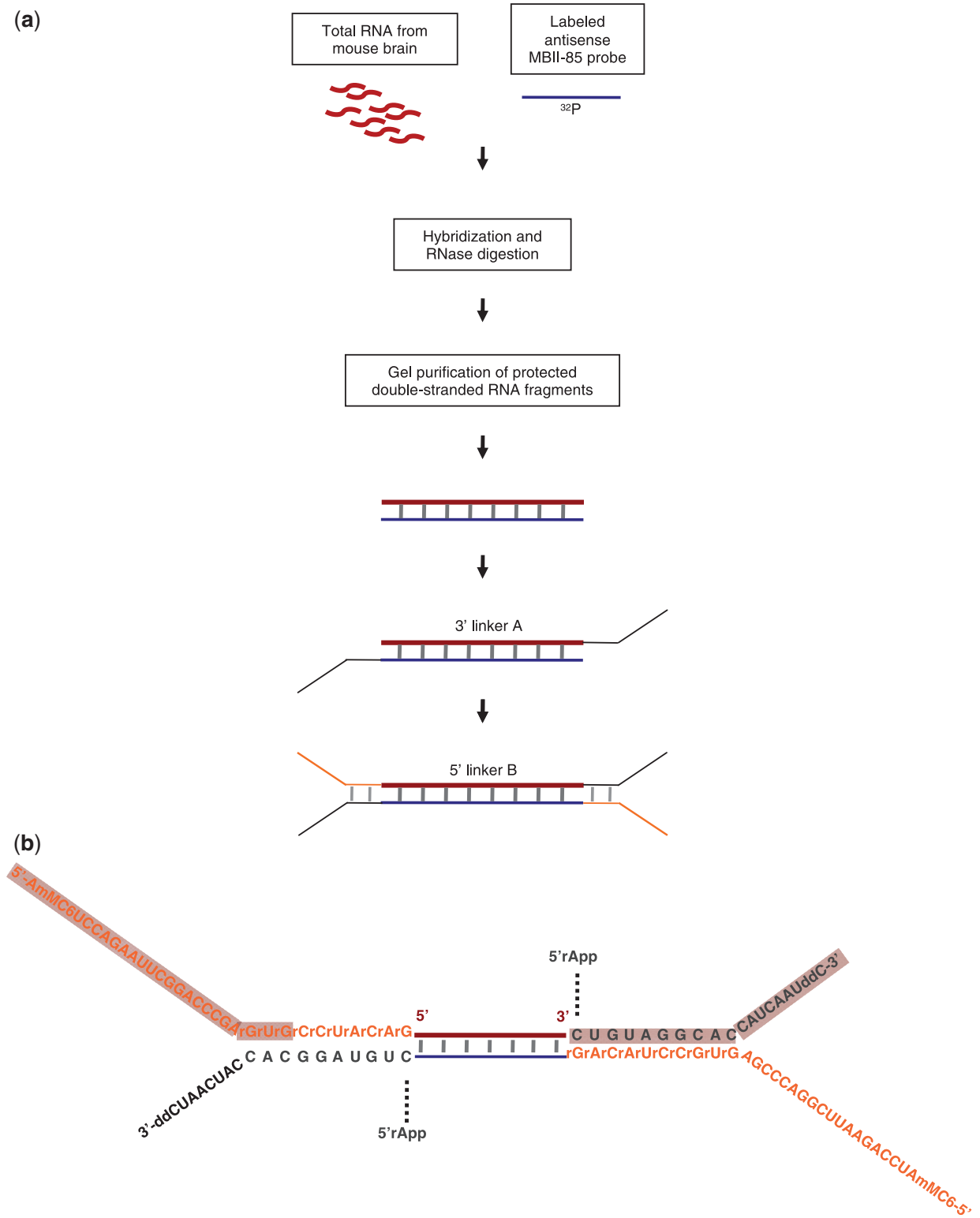

Figure 1. Overview of the method. (**a**) Cloning steps. RNAs are shown as red lines, the synthetic antisense probe as a blue line. Short vertical lines indicate hybridizations. (**b**) Overview of the primer locations prior to RT–PCR. Note that the 5′ rApp residue is removed after ligating the linkers.

RNAs from both DNA strands. Some of these potentially dsRNA parts overlap with experimentally verified RNA:miRNA interaction sites. The high abundance of potential dsRNA sites indicates a biological role of RNA expression derived from opposite DNA strands that can be detected with this method.

MATERIAL AND METHODS

Cloning of dsRNAs

Probe synthesis. We synthesize two RNA probes, a low-specific activity probe for cloning purposes and a high-specific activity probe for detection of the RNAs. For cloning purposes, we synthesize an antisense RNA using all four cold NTPs at a concentration of 1 μM. To visualize the RNA, we spike this RNA with 1 μl of ^{32}P α-UTP (800 mCi/mM) in a 20 μl reaction. This generates a low-specific activity probe of $\sim 0.6 \times 10^8$ cmp/μg. To detect protected fragments, we synthesize one probe with high-specific activity using radioactive ^{32}P α-UTP as the only source of UTP (specific activity 1.4×10^9 cmp/μg). We use the megascript kit (Ambion) for RNA synthesis.

RNase protection. We incubate 100 μg total brain RNA with 500 ng spiked antisense probe. The RNA is prepared using trizol, to avoid loss of small RNAs. In a parallel experiment, we incubate 10 μg total brain RNA with 50 000 cpm of high-specific probe. Probes and cold RNAs were precipitated, dissolved in hybridization buffer and denatured at 95°C for 3 min. Hybridization is carried out overnight, in 10 μl of hybridization buffer at 42°C. Single-stranded RNA is digested by RNase T1 and A1 in 150 μl of RNase digestion buffer. Since both RNase A and RNase T1 leaves a 3′ phosphate, we treat the reaction with shrimp nuclease for 30′ in the same buffer to generate a free 3′OH group. Prior to proteinase K digestion, 15 μg glycoblue (Ambion) is added.

Hybridization buffer: 40 mM PIPES, 1 mM EDTA, 400 mM NaCl, 80% formamide, pH 6.4.

RNase digestion buffer: 300 mM NaCl, 10 mM Tris–Cl, 5 mM EDTA, pH 7.4.

RNases removal. RNases are removed by adding 15 μl of 10 mg/ml proteinase K and 15 μl of 10% SDS, followed by one hour incubation at 37°C and phenol/chloroform extraction.

Removal of free nucleotides. Free nucleotides are removed by running the protected RNAs over a 15–18% 8 M Urea, 1× TBE gel. To later visualize the protected bands, we combine the reactions made with the high- and low-specific probes. After overnight autoradiography, the bands are cut out from the gel and the fragments are recovered by soak–crush in 3 M NH_4Ac, 1% SDS solution overnight at 37°C. Fragments are recovered by adding 2.5 volume ethanol and 1 μl of glycoblue.

Addition of the 3′ linker A. The linker A sequence is: 5′rAppCTGTAGGCACCATCAAT/3ddC. The rAppC moiety at its 5′-end allows its ligation without ATP to the 3′ OH of a nucleic acid. Its 3′-end is blocked by inclusion of ddCTP. The first ligation is carried out in a 20 μl volume. The final concentration of linker A is 4 μM. The ligation is carried out for 2 h in a 20 μl reaction in 50 mM HEPES pH 8.3, 10 mM $MgCl_2$, 3.3 mM DTT, 10 μg/ml BSA, (1× RNA ligation buffer, NEB), 8.3% (v/v) glycerol, 10% PEG 5000 and 20 U RNA ligase (NEB).

5′ phosphorylation and removal of linker. We phosphorylate the 5′-ends using polynucleotide kinase (NEB) and 1 mM ATP for 30 min, followed by precipitation. The phosphorylated RNA with the linker is again purified over a 15% 8 M Urea, 1× TBE gel. This step is necessary, as otherwise, there are free nucleotides in the RNA solution that react with the first linker and will result in clones with 2- to 3-nt inserts. RNA is again recovered by crush–soak and precipitation using glycoblue and NH_4Ac.

Addition of the 5′ linker B. The linker B sequence is 5′-AmMC6/GCTCCAGAATTCGGACCCGArGrUrG rCrCrUrArCrArG. The 5′-end is blocked using AmMC6.

The second linker is added to the RNA at a concentration of 4 μM. Ligation is performed using T4 DNA ligase at 20°C overnight in 1× ligase buffer (NEB) that contains ATP. The ligation mixture is precipitated and dissolved in 10 μl.

RT–PCR. Reverse transcription is done using monster script (Epicenter), after denaturing the template for three minutes at 95°C. The reverse transcription is carried out at 60°C for 40 minutes, using primer reverse #1: 5′ gattgatggtgcctacag. An aliquot (one-tenth) of the reaction is spiked with 2 μl α-32P dCTP. The aliquot of the cDNA is separated on a 15% acrylamide/8 M urea gel to monitor cDNA synthesis. The reverse transcription mixture is then amplified by PCR using primers reverse #1 and forward #1 (5′GCTCCAGAATTCGGACCCGA GTG-3′).

PCR validation of potential dsRNA transcripts. Two sets of primers were designed to test Dlgap2, Maml3, Kirrel3 and 2610528E23Rik genes, respectively. (Dlgap2: primer1 forward: TGTGCCTTCCGTGGTGTGCAGTTTAG, reverse:TCTTCATTTCACCGAGAGTCACACAGG; primer2 forward: GGGTGGCCGTCCATGACGTCCAGC TTTG; reverse: CATCCACATACAGAACCTTCTGTCAC C. Maml3: primer1 forward: ACACTGAGTTCTAGGAG ATTCTCAGG, reverse: GTATCATATATTAGGCTGG GTAAG; primer2 forward: TTGATCTTATCTTGTCAC ATGATTTC, reverse: CAGTGCTCACAGACCCTG AAGA. Kirrel3 primer1 forward: ACTCTCCTCTCCTC TCCTCTCTCTG, reverse: CTCTCACTGAGGATCAA GCTTGG; primer2 forward: CTAAGAACCACTCACA GGCAGCCAC, reverse: GTGGACTGACTGGAGAGG GAAGTC. 2610528E23Rik primer1 forward: CCCTGTA ACGTGTCCCTGAATGTTACTC, reverse: GAGAGAT TGTGCGTGACTGTTGGTAGG; primer2 forward: CT TCCACCTATCCTACAGATACTACTGC, reverse: GC CCTGTCTCGCAGACAGCGTATATTAC.). PCR was performed using PlatinumTaq polymerase and SuperScript III Reverse transcriptase (Invitrogen).

Bioinformatic analysis

Mapping of SOLiD reads to the mouse genome. We started with 47 755 454 Sequencing by Oligonucleotide Ligation and Detection (SOLiD™) reads of length 50. We used bowtie (version 0.12.3) (http://bowtie-bio.sourceforge.net/) (11) to align them to the mouse genome (NCBI37/mm9). In order to account for the 5′ adapter, which could appear with variable length, we adopted a sequential trimming strategy: we mapped the reads using various trimmings at the 5′-end: 23, 24,..., 29 nt, i.e. we mapped reads of length 27, 26,..., 21. In each run, we requested best match with no more than two mismatches over the length of the read. Subsequently, for each original read we kept the longest successful mapping, i.e. we keep the first mapping such that the read had not been found in any of the previous (shorter) trimming steps. This produced a total of 22 216 961 uniquely mapped reads on both strands (Supplementary Figure S2a). We then clustered independently forward and reverse strand reads using at least 1 nt overlap in the same strand (Supplementary Figure S2b), obtaining 7 084 100 clusters in forward and 7 087 115 clusters in reverse; 1 501 137 and 1 498 850 clusters with more than one read, respectively.

Comparison of SOLiD reads with genomic regions that could express dsRNAs. To compare the read clusters with the RefSeq annotation we used mm9, downloaded from UCSC. For the initial 14 171 215 read clusters, we found 3 212 152 (22.6%) to locate in a genic region, considered to be the extension from transcription start site (TSS) to transcription termination site (TTS), which were taken from the RefSeq annotation. As a comparison, we found 80 672 (49.18%) of the candidate dsRNAs to overlap with a genic region (Supplementary Figure S5). This represents ~2-fold enrichment. We also calculated the expected gene loci positions as the union of forward and reverse positions in gene loci (898 381 160 bp) and compared it with the total number of base pairs in the genome removing gaps (2 654 917 900 bp), which would give an estimate of 35% of the base pairs in the genome covered by genic regions. The total base pairs of candidate dsRNA in genic regions (11 732 808 bp) represents 50% of the total base pairs in candidate dsRNAs (23 430 566 bp). Comparing the counts, we find a significant enrichment ($\chi^2 = 983\,478$, df = 1, $P < 2.2e^{-16}$).

Calculation of dsRNA-expressing candidate regions. We first calculated the pre-mRNA regions from the RefSeq annotation that overlap with antisense transcripts, i.e. other RefSeqs and spliced ESTs. We considered only spliced ESTs, as intronless ESTs cannot always be accurately assigned to the correct strand. We defined each region as the genomic range of a pre-mRNA that has evidence of transcription from the opposite strand. We obtained a total of 35 916 different regions. We then calculated the read-clusters that overlap with any of these dsRNA-expressing candidate (DEC) regions. For this, we removed all singletons, i.e. clusters with one single read.

Mapping of AGO HITS-CLIP reads. We downloaded from http://ago.rockefeller.edu/ the reads from the HITS-CLIP experiment for Ago (12) for the mouse neocortex samples (A–E) separated into two fractions: 110 kDa (miRNA fraction) and 130 kDa (mRNA target fraction). We then carried out, as above for the SOLiD samples, a sequential trimming but this time from the 3′-end, using trimmings from 0 to 11 bases in samples A and B; and from 4 to 15 bases in samples C–E; hence, we mapped lengths from 32 to 21 nt. We only kept unique best matches allowing up to two mismatches. For each original read, we kept the longest successful mapping, i.e. the first map found in the sequential trimming. This resulted in a total of 5 374 490 mapped reads for the 110-kDa fraction, forming 556 723 clusters; and 22 109 188 mapped reads for the 130-kDa fraction, forming 1 091 734 clusters. These clusters were compared with the 106 486 dsRNA candidates derived above that do not fall in repeats, using the program fjoin (13). This produced an overlap of 10 454 dsRNA clusters with 12 562 Ago target clusters and of 5665 dsRNA clusters with 6029 Ago miRNA clusters.

RESULTS

Processing of MBII-85 and MBII-52

We recently found that the C/D box snoRNA MBII-52/SNORD 115 is processed into smaller RNAs (3), which were named psnoRNAs. MBII-52 is expressed from the SNURF–SNRPN transcription unit that has been implicated in the Prader-Willi syndrome. Figure 3a and b shows a schematic overview of this unit for human and mouse. Three patients with microdeletions in the SNURF–SNRPN have been reported (14–16). Since these patients exhibit a Prader-Willi syndrome-like phenotype, it is likely that the disease-causing genes reside in this region. Two of the microdeletions encompassed only the expression units of snoRNAs HBII-85 (human SNORD 116) and half of the HBII-52 copies. Each expression unit consists of two non-coding exons that flank the intron that hosts the snoRNA (schematically shown in Figure 2a). The comparison of the three microdeletions (thick line in Figure 3a) indicated that the absence of the human ortholog of MBII-85 is likely the most important molecular cause for the Prader-Willi like phenotype.

We therefore investigated the processing pattern of the mouse MBII-85 snoRNA. We analysed the copy that is most closely related to human copy number 27 (genomic coordinates of the expression unit consisting of snoRNA and flanking exons is chr7:66 838 324–66 844 003 on NCBI37/mm9). We used a probe complementary for MBII-85 in an RNase protection assay and we detected six fragments (a–f) in brain tissue (Figure 2B), indicating that this snoRNA is processed, similar to MBII-52 (3).

Cloning of dsRNA

A graphic overview of the method is given in Figure 1a. To identify the processing sites, we clone the fragments, as outlined here and described in detail in the methods. First, we perform a standard RNase protection assay using a uniformly labelled antisense RNA probe. For cloning, a low-specificity probe is used (0.6×10^8 cmp/μg). An

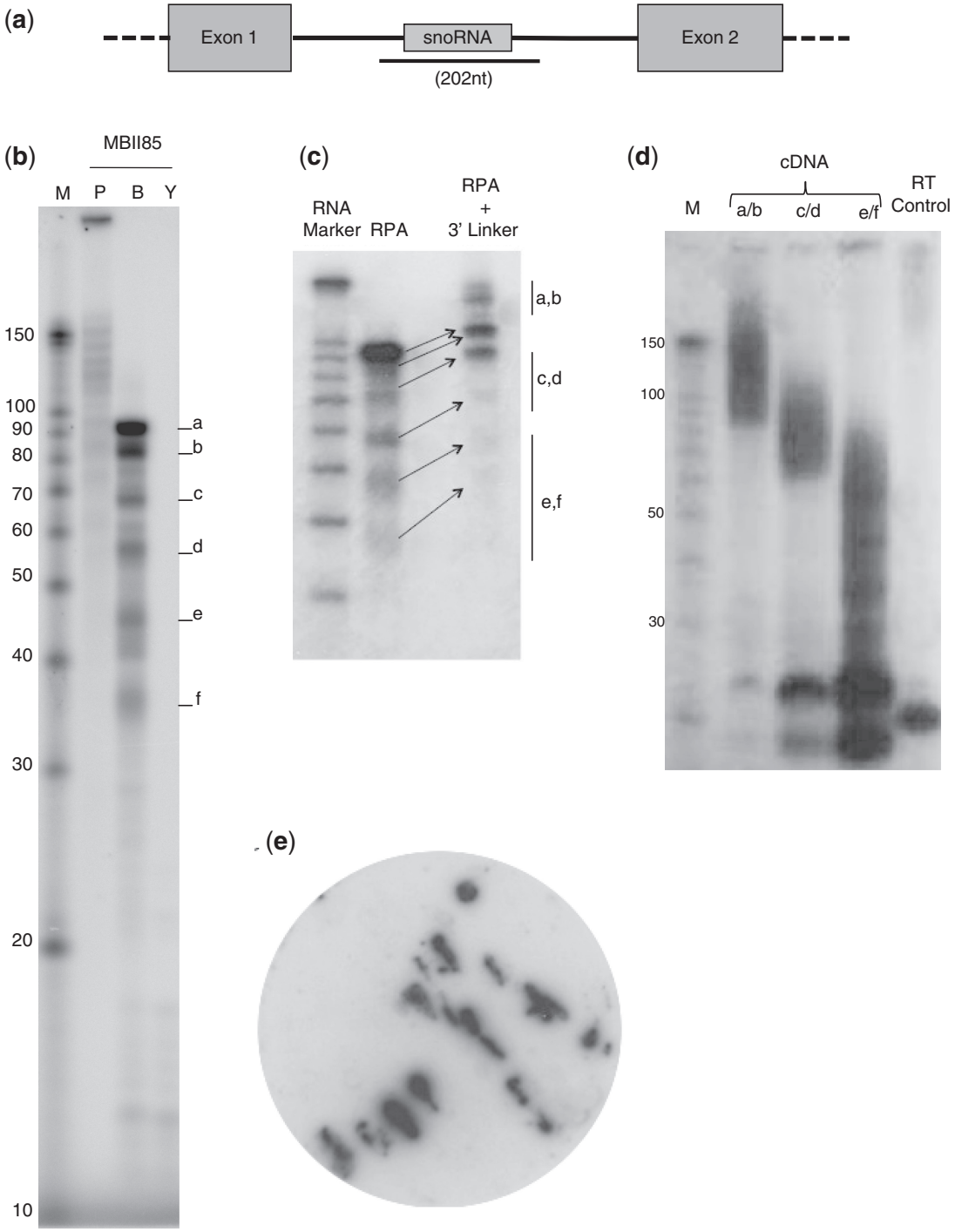

Figure 2. Cloning steps in cloning dsRNAs. **(a)** Schematic representation of the MBII-85 expression cassette and the location of the antisense probe. **(b)** RNase protection using MBII-85 as a probe and total mouse brain RNA as a target. The obtained bands are labelled a–f. M: 10-bp ladder marker, P: probe, B: protection of brain RNA, Y: protection of yeast RNA. **(c)** Addition of the 3′ linker_A. An aliquot of the protection reaction was run in lane RPA, and the total reaction with the 3′-end linker is run in lane RPA+ 3′ linker. The addition of the linker caused a shift in the dsRNA bands that is indicated by arrows. **(d)** cDNA synthesis. The representative gel shows the cDNA synthesis after ligation of the second linker. a/b, c/d and e/f correspond to the cut out areas in panel c. **(e)** Colony hybridization of bacteria transformed with cloned dsRNAs from step (d). A total of 100 bacterial clones were streaked out on an agar plate, lifted with a nitrocellulose filter and hybridized. Four end-labelled oligonucleotides covering the MBII-85 sequence were used for hybridization. As shown here, only 10–20% of the obtained clones contained antisense probe sequences.

amount of 100 μg total mouse brain RNA were hybridized overnight with 500 ng antisense probe using a commercial RNase protection kit (Ambion). To better visualize the protected fragments, we perform a parallel hybridization with an antisense probe that was generated with radiolabelled α-UTP as the only UTP source. After hybridization, both reactions are mixed and further processed together. Single-stranded RNA is removed by

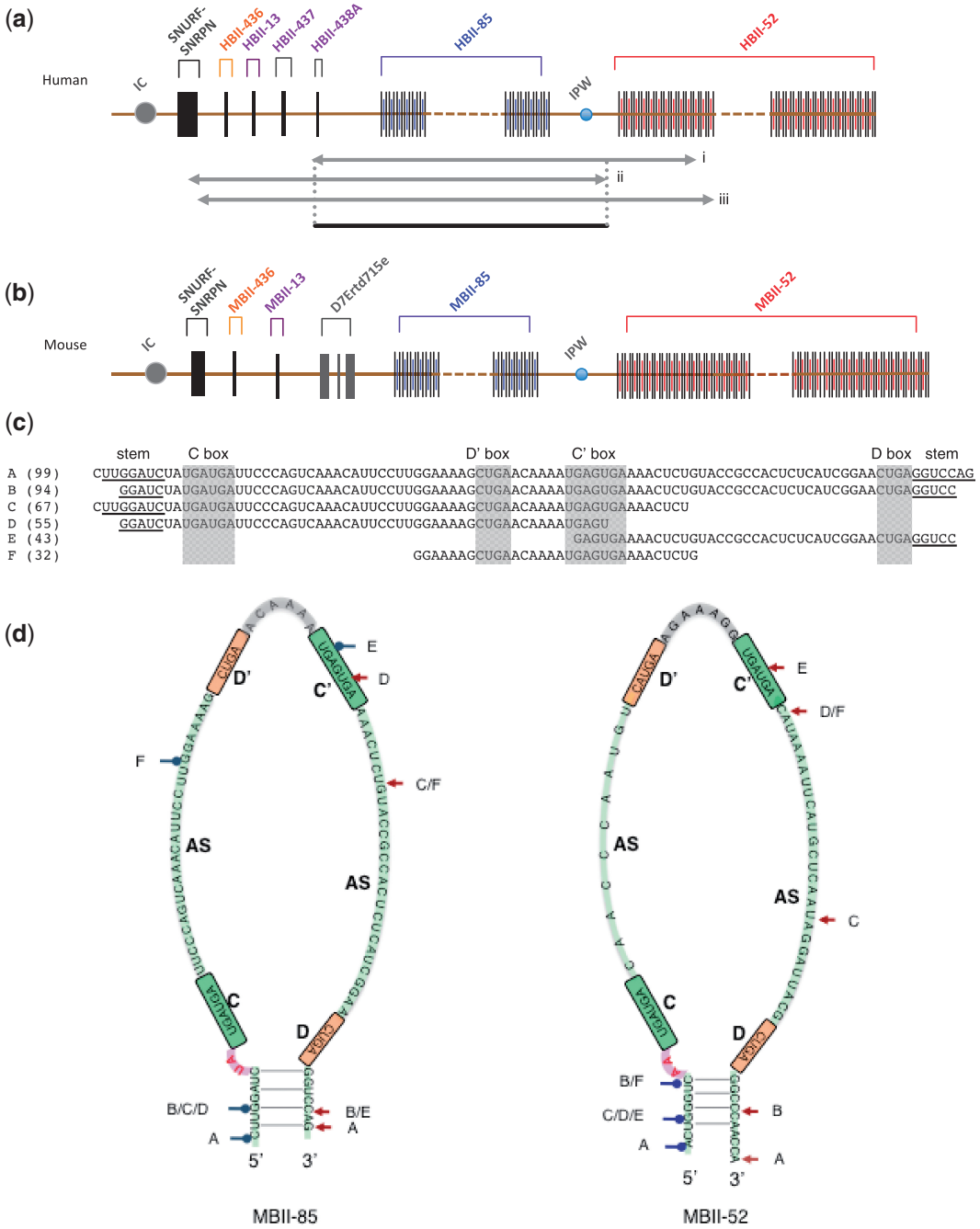

Figure 3. Processing patterns of MBII-85 and MBII-52. (**a** and **b**) Schematic overview of the SNURF–SNRPN transcription unit. Exons are indicated as vertical lines, snoRNAs are shorter vertical lines. The SNURF–SNRPN genes contains at least 11 exons and introns, which are shown as one box for simplicity. The three arrows underneath the gene structure indicate three reported microdeletions that cause Prader-Willi syndrome in humans (14–16). The thick black line indicates the deleted region common to all patients (i) reported in (15), ii reported in (14), and iii reported in (16). (a) SNURF–SNRPN transcription unit in human, (b) SNURF–SNRPN transcription unit in mouse. (**c**) Sequences of the clones obtained for MBII-85, expression unit chr7:66 838 324–66 844 003 on NCBI37/mm9. C/D box snoRNA elements are indicated by colored boxes and terminal stems underlined. The notation a–f refers to Figure 2b. (**d**) Comparison of the processing sites in the structures of MBII-52 and MBII-85. The C/D box typical sequence elements are indicated (AS: antisense box, C, D, C′, D′ boxes). Arrows point to the major processing sites. Letters next to the arrows indicate the psnoRNA forms generated by this processing, which could be either an endonuclease cleaving site or a protecting point for an exonuclease.

RNase A1 and T1 digestion. For cloning, we dephosphorylate the RNAs with shrimp nuclease and deproteinate the reaction with proteinase K, followed by phenol/chloroform extraction. This step is necessary to prevent self ligation of the products. The reactions are separated on 15% acrylamide 8 M Urea/TBE gels and visualized by autoradiography overnight (Figure 2b). The RNAs obtained in the RNase protection reaction are then excised from the gel and purified by the soak-crush method.

Next, we add 3′ linkers to the dsRNAs obtained during the RNase protection reaction. Each RNase-protection signal consists of two dsRNA strands: one strand is derived from the endogenous RNA and the other one comes from the added *in vitro* generated antisense RNA. Although the strands are separated in the 8 M urea acrylamide gel used for the protection analysis, due to their small size the RNAs will reanneal during elution from the gel. We could not add conventional RNA oligonucleotides to these RNAs using different RNA and DNA ligases (data not shown), which could account for previous failures to clone protected fragments. However, the use of pre-activated 5′ adenylated (rApp) oligonucleotides turned out to be highly efficient. 5′ adenylated (rApp) oligonucleotides are similar to the intermediate oligonucleotide during a ligation reaction, where the ligase adds rATP to the 5′-end of the nucleic acid, which is subsequently removed during the ligation (17). The 3′ linker added to the dsRNA contains a ddC at the 3′-end to prevent self-ligation (Figure 1b).

As shown in Figure 2c, the addition of the 5′ adenylated linker is highly efficient, as the protected fragments show the expected shift in mobility, caused by the additional linker sequence. After linker ligation, the RNAs are phosphorylated using polynucleotide kinase, purified by phenol–chloroform reaction and precipitated. The 5′ phosphorylated RNAs with the ligated 3′ linker are purified a second time using an 8 M urea/TBE polyacrylamide gel. We found that this step is essential, as otherwise residual short (1–5 nt) oligonucleotides ligate to the second linker and are preferentially amplified in the following PCR steps.

Next, we ligate the 5′-linker using T4 DNA ligase at 20°C overnight. To aid in ligation, the 5′-linker is partially complementary to the 3′ linker, and generates a Y-shaped structure in the final product (Figure 1b). The RNA with linkers is then reverse transcribed using a reverse transcriptase that works at 60°C. The synthesis of the first cDNA strand is monitored by incorporation of ^{32}P α-dCTP in the reverse transcription reaction, as shown in Figure 2d. The obtained cDNA reflects the length distribution of the input RNA, indicating that full-length cDNA can be obtained by reverse transcription of dsRNAs, when a reverse transcriptase acting at higher temperature is used.

The obtained cDNA fractions are amplified by PCR and subcloned into topoisomerase-activated vectors (Invitrogen). Bacteria containing MBII-85 fragments were identified by colony hybridization. As shown in Figure 2e, 10–20% of the clones contained inserts derived from MBII-85.

Structure of the processed MBII-85 snoRNAs

The sequences of the clones containing MBII-85 dsRNA fragments are shown in Figure 3c. They account for the protection pattern that we observed (Figure 2b). We next compared the processing patterns of MBII-85 with the patterns of MBII-52 that we determined earlier (3). Classic C/D box snoRNAs have characteristic sequence elements that determine the association of the full-length snoRNA with protein complexes (18). The elements are the C, D, C′ D′ boxes, a terminal stem with an RNA kink and two antisense sequences. As shown in Figure 3d the sites of processing are similar between MBII-52 and MBII-85. The processing events occur in the stem of the snoRNA, cut once in the antisense elements, and cut in the C and D box. In both snoRNAs, the nuclease action leads to specific processed RNA molecules, which suggests RNase action, either by a selective endonuclease or by exonucleases arrested at specific sites.

Although there is evidence from HTS data for abundant processing of C/D box snoRNAs (19), it should be mentioned that not all C/D box snoRNAs are processed. For example, we cannot detect fragmentation of the U90 snoRNA using RNase protection (Supplementary Figure S1a).

Comparison with traditional cloning methods for small RNAs

Cloning efforts for miRNAs identified several fragments that were derived from C/D box snoRNAs. Interestingly, >80% of these shorter RNAs derived from C/D box snoRNAs were flanked by D boxes (4). We observe similar processing sites downstream of the D box in the stems and upstream of the D′ box in the antisense region, indicating that D-box sequences act during processing or have a functional role in the resulting psnoRNAs (Figure 3D). The RNA used to clone shorter C/D box snoRNA-derived RNAs in HTS experiments was size selected to cover the miRNA range (4,19). Therefore, larger C/D box snoRNA fragments might have been missed. Regarding MBII-85, we did not find any RNA in the 22-nt range typical for miRNAs (Figure 2b). This indicates that the generation of miRNAs is not the major function of this MBII-85 expression unit. It also suggests that psnoRNAs are not made by dicer, which would generate smaller RNA fragments.

In contrast to the fairly similar HBII-52 copies (3,20), the 29 human HBII-85 snoRNA copies can be subgrouped into three clusters based on their sequences. A deep-sequencing analysis that counted overlapping reads for each snoRNA showed different expression levels of HBII-85 sequences in each cluster (21). The mouse snoRNA copy that we analysed falls into the third cluster, which shows the lowest expression. RNAs from the MBII-85 cluster have been recently cloned from isolated RNPs (22). In this method, non-coding RNAs are enriched by isolating their protein particles. The single-stranded RNA is then isolated and cloned. The cloning is performed by traditional polyC tailing of the single-stranded RNAs and by addition of 5′-end linkers, followed by RT–PCR. Sequences derived from this method showed a strong bias towards highly expressed RNAs, corresponding to the first human cluster (Supplementary Figure S1c). Although several hundred reads corresponded to MBII-85 psnoRNAs in the dataset, the MBII-85 copy used in our study was not present in the dataset, most likely as it is less abundantly expressed. A bioinformatics analysis of the all MBII-85 RNAs generated by cloning from RNPs shows that the vast majority is derived from cutting MBII-85 in

the second antisense box (Supplementary Figure S1b) (C/F position, Figure 3c) and in the first antisense box (f position, Figure 3c).

These data indicate that cloning RNA directly from RNase protection fragments can be more sensitive than deep-sequencing complete libraries of RNAs made by conventional methods, especially when one particular RNA is investigated.

Abundant expression of potentially dsRNAs in cellular RNA samples

As shown in Figure 2e, even when we enriched for protected RNAs by cutting out the bands representing the protected fragments, we obtained a high number of clones containing non-related RNA sequences. To obtain a complete picture of RNAs identified by our cloning method, we analysed the obtained cDNAs by HTS analysis using the SOLiD™ deep-sequencing platform (23). The PCR fragments were directly ligated to the solid support and subjected to sequencing.

HTS gave a total of 47 755 454 SOLiD™ reads of length 50. We used used bowtie (11) to align them to the mouse genome. The obtained reads could be mapped to 22 216 961 unique sites in the mouse genome, forming 14 171 215 clusters in both strands. Selecting those clusters that are significantly close to each other and overlap by >50% in both DNA strands, we found 164 033 double-stranded clusters (see 'Materials and Method' section in Supplementary Data S2.1 and Supplementary Figure S2).

Analysis of the distance between read clusters of more than one read showed that they tend to occur at distances <100 nt (Supplementary Figure S3). We therefore considered superclusters that were defined as read clusters separated in the same strand by <100 nt (Supplementary Figure S2c). Within these 4 800 821 forward and 4 801 075 reverse superclusters, there were 1 502 330 and 1 502 952 superclusters, respectively, containing more than one cluster. From these, a total of 164 033 overlap in both strands in >50% of their length (Supplementary Figure S2d). These represent candidate dsRNA molecules, with lengths of ~129 nt (Supplementary Figure S4). We consider these to be a conservative estimate of candidate dsRNA regions.

These clusters that potentially represent naturally occurring dsRNAs were approximately equally distributed between known genes, (n = 80 672) and intergenic regions (n = 83 361). Known genes were defined by RefSeq annotation. The data indicate a strong enrichment of dsRNAs in known gene regions ($P < 2.2e{-}16$ using the chi square test; $\chi^2 = 530\,938.4$, df = 1), (Supplementary Figures S3–S5).

We next evaluated the possibility that the dsRNAs that we cloned reflect naturally occurring dsRNA regions. We first calculated which regions in the genome could give rise to dsRNA sequences. We determined the pre-mRNA regions from the RefSeq annotation that overlap with antisense transcripts indicated by other RefSeqs and spliced ESTs. We considering only ESTs with introns, as EST sequences may appear reversed in the database. Therefore, intronless ESTs may align to the wrong strand. However, ESTs with introns can be correctly assigned to a strand using as a guide the conserved splice-site sequences.

Using these constrains, we obtained a total of 35 916 genomic regions that could generate dsRNAs. We called these genome areas dsRNA expressing candidate (DEC) regions (Supplementary Figure S6). The DEC regions are visualized on the UCSC genome browser at http://regulatorygenomics.upf.edu/Data/Cloning/.

We then determined how many of these DEC regions contain cloned protected fragments. We found that 88.4% of DECs overlap with the read clusters (70.5% with clusters of more than one read). Moreover, 21% of DECs overlap with 2.4% of our 164 033 candidate dsRNAs. An example of this analysis is shown in Figure 4A for the Dlgap2 gene that partially overlaps with the AU016332/AU016667 transcript. The DEC regions were not uniformly spread over all genes, but are concentrated in certain genes. The top 10 genes with the highest amount of cloned dsRNAs overlapping with DEC regions are listed in Table 1. The genes with the highest number of cloned dsRNAs are shown in Table 1.

Validation of dsRNA regions by RT–PCR

We next tested the existence of dsRNA transcripts with RT-PCR as a different method. We selected primers complementary to the end of dsRNAs for PCR. As templates, we reverse transcribed the RNA with either the forward or reverse primer. As shown in Figure 4B, we could amplify cDNA with the expected size using either primer for most of the tested dsRNAs. We did not observe amplification without reverse transcription, indicating that we amplified RNA. Since both the reverse and forward primers could prime cDNA synthesis, the substrate was dsRNA. The observed PCR products showed a weak intensity, suggesting that the RNAs are not highly abundant, which is expected, as these RNAs have not been cloned by traditional methods. This experiment demonstrates that our method detected novel double-stranded cellular RNAs.

Since we used total mouse brain RNA for the hybridization, we cannot rule out that these double-stranded clusters represent hybridization of two RNA strands from different cells. However, several of the genes with DEC regions are ubiquitously expressed in mouse brain, for example Dlgap2, Maml3 and 2610528E23Rik/Filip1l. It is therefore possible that within a single cell RNA can be expressed from different DNA strands at the same locus. We also do not know whether the RNA is derived from the same or different alleles. However, since we observe dsRNAs from the Y chromosome, it is possible that we detected RNA expression from polymerase II acting in different directions on a single DNA molecule.

It also remains to be determined whether these RNAs interact in a physiological context, as they are likely coated with proteins that are removed in the RNA preparation. In all examples, we see gaps between the cloned dsRNA within the longer bioinformatically predicted DECs (Figure 4). This could indicate nuclease action on

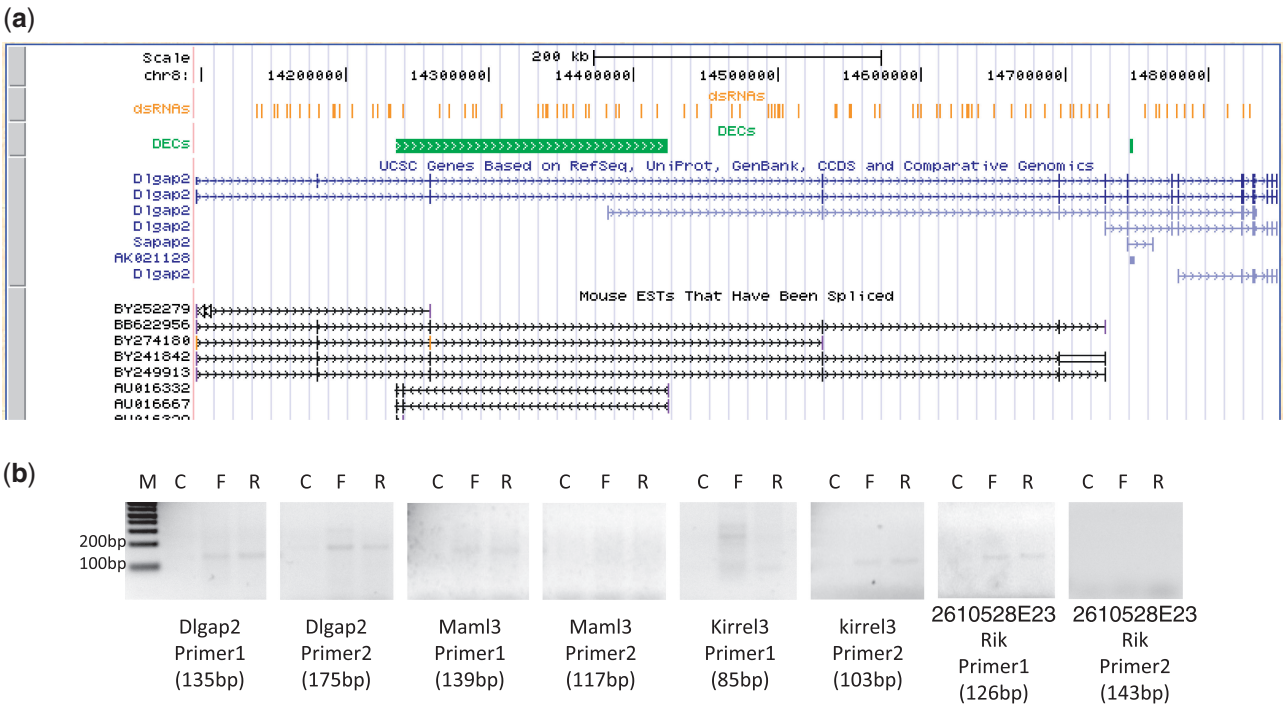

Figure 4. Visualization and detection of cloned protected fragments in putative dsRNA regions. (**a**) The protected fragments are visualized as a tract in the UCSC genome browser. The Dlgap2 gene partially overlaps with the AU016332/AU016667 transcript, which defines a DEC region. (**b**) RT–PCR analysis of dsRNAs in DEC regions. C, no reverse transcription; F, only the forward primer was used for reverse transcription; R, only the reverse primer was used for reverse transcription. Note that we could not detect amplification with Maml3, primer2 and Rik, primer 2 indicating low dsRNA expression. The tracks can be seen at http://regulatorygenomics.upf.edu/Data/Cloning/.

dsRNAs that would destroy long dsRNAs formed in the cell, analogous to dicer acting on pre-miRNAs.

DEC regions overlap with known argonaute-binding sites

Pre-miRNAs contain dsRNA regions. The dsRNAs are cleaved by dicer and are loaded onto argonaute proteins by the RISC complex. To test whether the experimentally confirmed dsRNAs could be dicer substrates, we compared our dsRNA clusters with experimentally confirmed argonaute-binding sites (12). Excluding reads in repeat regions, we found that 10 454 of dsRNA clusters overlap with Ago targets and 5665 overlap with Ago miRNA reads. Supplementary Table S1 lists examples of candidate dsRNA regions that overlap with known miRNAs. This overlap strongly suggests that some of the longer dsRNAs are processed by dicer and the resulting fragments loaded on argonautes. It is also possible that dsRNA regions are markers for so far unknown miRNA expressing regions.

Together, these data strongly suggests that within a cell RNA:RNA interactions occur that can be determined by our method. However, at this point we cannot discriminate between RNA:RNA interaction that occur within a single cell and artificial interactions caused by the homogenization of the tissue and the mixing of all cellular RNAs. Nevertheless, the candidate list of dsRNA regions allows further detailed investigations of the physiological role of dsRNAs.

DISCUSSION

Although RNase protection methods have been used for over 30 years (24), there is no protocol available to clone dsRNAs generated by these experiments. Typically, the structure of the protected fragments is deduced indirectly by their length. The method described here allows cloning of small dsRNA amounts derived from such experiments. The major problem in cloning protected fragments was that the sense and antisense stands cannot be separated, as they migrate identically through denaturing acrylamide gels. Every subsequent biochemical manipulation requires the removal of urea, which leads to re-annealing of the RNA strands. As there are no known dsRNA ligases, cloning of these fragments proved difficult.

There are two essential factors that allowed cloning of dsRNAs: the use of adenylated linkers and rigorous gel purification of intermediate products. DsRNAs with blunt ends are poor substrates for RNA or DNA ligases. However an adenylated linker that represents the activated form of a single-stranded nucleic acid in a ligation reaction can be efficiently ligated to the dsRNAs. It is important that all free nucleotides are removed prior to ligation as adenylated linkers preferentially react with their free 3′ hydroxyl groups.

The method has several advantages over current cloning procedures. First, there is no knowledge of priming sites necessary, which allows it to be applied to non-coding RNAs or mRNA fragments. Secondly, the antisense probe strand that was generated *in vitro* contains no

Table 1. Genes with the highest number of cloned dsRNAs

Name	Description	Number of cloned dsRNAs
Genes with the highest number of cloned dsRNAs overlapping with DECs in mouse brain		
Maml3	Mastermind-like 3	49
Kirrel3	Adult male diencephalon cDNA	38
Tns1	Tensin 1	28
Dlgap2	Disks large-associated protein 2 isoform 1	24
2610528E23Rik	Hypothetical CMS1 family protein	23
Filip1l	Filamin A-interacting protein 1-like isoform 1	22
Cacna1b	Neuronal type calcium channel α-1 subunit (α1B)	21
Fam190a	Hypothetical protein LOC232035	19
Ccdc91	Coiled-coil domain-containing protein 91	19
Shb	SH2 domain-containing adapter protein B	19
Col5a1	Collagen α-1 (V) chain precursor	18
Genes with the highest number of cloned dsRNAs in mouse brain RNA		
Camta1	Calmodulin-binding transcription activator 1	138
Accn1	Neuronal amiloride-sensitive cation channel 1	127
Park2	Parkinson protein 2, E3 ubiquitin protein ligase	123
Rbfox1	RNA-binding protein, fox-1 homolog	120
Macrod2	MACRO domain containing 2	117
Cdh13	Putative mediator of cell–cell interaction in the heart	111
Wwox	WW domain containing oxidoreductase	98
Dpp6	Dipeptidyl-peptidase 6	98
Auts2	Autism susceptibility candidate 2	97
Dlgap2	Disks large-associated protein 2	97

modified nucleotides, such as a 5′ cap or 2′,3′-cyclic phosphates that prevent linker ligation. Together, this allows cloning of highly modified unknown RNAs.

As the method detects a specific RNA that is enriched through the RNase protection procedure, it can be more sensitive for a specific intermediate than HTS.

Unexpectedly, we found that dsRNAs generated by natural antisense RNAs are very abundant in RNase protection experiments. A cross-interaction with such endogenous dsRNAs could account for RNase protection artifacts that are sometimes observed. Our bioinformatic analysis indicated that the majority of these dsRNAs are derived from genomic regions that show evidence for antisense expression. As some of these regions show widespread expression throughout the brain, it is likely that some antisense expression occurs in the same cells. We observed that dsRNAs from these regions are processed into smaller fragments. A cleavage of such longer dsRNAs would be similar to the cleavage of stems–loop structures found in conventional pre-miRNAs by dicer. Since RNAs derived from these dsRNAs could be mapped to argonaute-binding sites, it is possible that dsRNAs generated by antisense expression give rise to miRNAs. It was shown that dsRNAs can be exported from the nucleus to the cytosol by HIV-REV (25). It is thus possible that an endogenous cellular pathway exists that exports dsRNAs into the cytosol, where they form the substrate of miRNA formation.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Institutes of Health (RO1 GM083187 to S.S.; NIH GM074688 to M.Z.); the Prader-Willi Research Foundation USA and Shire Human Genetic Therapies; European Comission project EURASNET-LSHG-CT-2005-518238 (to S.S. and E.E.); Spanish Ministry of Science grant BIO2008-01091 (to E.E.). Funding for open access charge: National Institutes of Health (RO1 GM083187).

REFERENCES

1. Amaral,P.P., Dinger,M.E., Mercer,T.R. and Mattick,J.S. (2008) The eukaryotic genome as an RNA machine. *Science*, **319**, 1787–1789.
2. Lee,Y.S., Shibata,Y., Malhotra,A. and Dutta,A. (2009) A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes Dev.*, **23**, 2639–2649.
3. Kishore,S., Khanna,A., Zhang,Z., Hui,J., Balwierz,P., Stefan,M., Beach,C., Nicholls,R.D., Zavolan,M. and Stamm,S. (2010) The snoRNA MBII-52 (SNORD 115) is processed into smaller RNAs and regulates alternative splicing. *Hum. Mol. Genet.*, **19**, 1153–1164.
4. Brameier,M., Herwig,A., Reinhardt,R., Walter,L. and Gruber,J. (2011) Human box C/D snoRNAs with miRNA like functions: expanding the range of regulatory RNAs. *Nucleic Acids Res.*, 675–686.
5. Scott,M.S., Avolio,F., Ono,M., Lamond,A.I. and Barton,G.J. (2009) Human miRNA precursors with box H/ACA snoRNA features. *PLoS Comput. Biol.*, **5**, e1000507.
6. Burroughs,A.M., Ando,Y., Hoon,M.L., Tomaru,Y., Suzuki,H., Hayashizaki,Y. and Daub,C.O. (2011) Deep-sequencing of human Argonaute-associated small RNAs provides insight into miRNA sorting and reveals Argonaute association with RNA fragments of diverse origin. *RNA Biol.*, **8**, 158–177.
7. Ono,M., Yamada,K., Avolio,F., Scott,M.S., van Koningsbruggen,S., Barton,G.J. and Lamond,A.I. (2010) Analysis

of human small nucleolar RNAs (snoRNA) and the development of snoRNA modulator of gene expression vectors. *Mol. Biol. Cell*, **21**, 1569–1584.
8. Ono,M., Scott,M.S., Yamada,K., Avolio,F., Barton,G.J. and Lamond,A.I. (2011) Identification of human miRNA precursors that resemble box C/D snoRNAs. *Nucleic Acids Res.*, **39**, 3879–3891.
9. Cashdollar,L.W., Esparza,J., Hudson,G.R., Chmelo,R., Lee,P.W. and Joklik,W.K. (1982) Cloning the double-stranded RNA genes of reovirus: sequence of the cloned S2 gene. *Proc. Natl Acad. Sci. USA*, **79**, 7644–7648.
10. Faridani,O.R., McInerney,G.M., Gradin,K. and Good,L. (2008) Specific ligation to double-stranded RNA for analysis of cellular RNA::RNA interactions. *Nucleic Acids Res.*, **36**, e99.
11. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
12. Chi,S.W., Zang,J.B., Mele,A. and Darnell,R.B. (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, **460**, 479–486.
13. Richardson,J.E. (2006) fjoin: simple and efficient computation of feature overlaps. *J. Comput. Biol.*, **13**, 1457–1464.
14. de Smith,A.J., Purmann,C., Walters,R.G., Ellis,R.J., Holder,S.E., Van Haelst,M.M., Brady,A.F., Fairbrother,U.L., Dattani,M., Keogh,J.M. *et al.* (2009) A deletion of the HBII-85 class of small nucleolar RNAs (snoRNAs) is associated with hyperphagia, obesity and hypogonadism. *Hum. Mol. Genet.*, **18**, 3257–3265.
15. Sahoo,T., del Gaudio,D., German,J.R., Shinawi,M., Peters,S.U., Person,R.E., Garnica,A., Cheung,S.W. and Beaudet,A.L. (2008) Prader-Willi phenotype caused by paternal deficiency for the HBII-85 C/D box small nucleolar RNA cluster. *Nat. Genet.*, **40**, 719–721.
16. Duker,A.L., Ballif,B.C., Bawle,E.V., Person,R.E., Mahadevan,S., Alliman,S., Thompson,R., Traylor,R., Bejjani,B.A., Shaffer,L.G. *et al.* (2010) Paternally inherited microdeletion at 15q11.2 confirms a significant role for the SNORD116 C/D box snoRNA cluster in Prader-Willi syndrome. *Eur. J. Hum. Genet. EJHG*, **18**, 1196–1201.
17. Lau,N.C., Lim,L.P., Weinstein,E.G. and Bartel,D.P. (2001) An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science*, **294**, 858–862.
18. Brown,J.W., Marshall,D.F. and Echeverria,M. (2008) Intronic noncoding RNAs and splicing. *Trends Plant Sci.*, **13**, 335–342.
19. Taft,R.J., Glazov,E.A., Lassmann,T., Hayashizaki,Y., Carninci,P. and Mattick,J.S. (2009) Small RNAs derived from snoRNAs. *RNA*, **15**, 1233–1240.
20. Kishore,S. and Stamm,S. (2006) Regulation of alternative splicing by snoRNAs. *Cold Spring Harb. Symp. Quant. Biol.*, **LXXI**, 329–334.
21. Castle,J.C., Armour,C.D., Lower,M., Haynor,D., Biery,M., Bouzek,H., Chen,R., Jackson,S., Johnson,J.M., Rohl,C.A. *et al.* (2010) Digital genome-wide ncRNA expression, including SnoRNAs, across 11 human tissues using polyA-neutral amplification. *PLoS One*, **5**, e11779.
22. Rederstorff,M., Bernhart,S.H., Tanzer,A., Zywicki,M., Perfler,K., Lukasser,M., Hofacker,I.L. and Huttenhofer,A. (2010) RNPomics: defining the ncRNA transcriptome by cDNA library generation from ribonucleo-protein particles. *Nucleic Acids Res.*, **38**, e113.
23. Mardis,E.R. (2008) Next-generation DNA sequencing methods. *Ann. Rev. Genomics Hum. Genet.*, **9**, 387–402.
24. Berk,A.J. and Sharp,P.A. (1977) Sizing and mapping of early adenovirus mRNAs by gel electrophoresis of S1 endonuclease-digested hybrids. *Cell*, **12**, 721–732.
25. Zhang,Z. and Carmichael,G.G. (2001) The fate of dsRNA in the nucleus: a p54(nrb)-containing complex mediates the nuclear retention of promiscuously A-to-I edited RNAs. *Cell*, **106**, 465–475.