Methods 65 (2014) 302–309

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

High-resolution profiling of protein occupancy on polyadenylated RNA transcripts ☆

Mathias Munschauer, Markus Schueler, Christoph Dieterich, Markus Landthaler *

Max-Delbrück-Center for Molecular Medicine, Berlin Institute for Medical Systems Biology, Robert-Rössle Str. 10, 13125 Berlin, Germany

ARTICLE INFO

Article history:
Available online 1 October 2013

Keywords:
CLIP
PAR-CLIP
mRNA
lincRNA
Protein occupancy profiling
Non-coding RNA
Protein–RNA interaction
RNA-binding protein
Next-generation sequencing

ABSTRACT

A key prerequisite to understand how gene regulatory processes are controlled by the interplay of RNA-binding proteins and ribonucleoprotein complexes with RNAs is the generation of comprehensive high-resolution maps of protein–RNA interactions. Recent advances in next-generation sequencing technology accelerated the development of various crosslinking and immunoprecipitation (CLIP) approaches to broadly identify RNA regions contacted by RNA-binding proteins. However these methods only consider single RNA-binding proteins and their contact sites, irrespective of the overall *cis*-regulatory sequence space contacted by other RNA interacting factors. Here we describe the application of protein occupancy profiling, a novel approach that globally displays the RNA contact sites of the poly(A)+ RNA-bound proteome. Protein occupancy profiling enables the generation of transcriptome-wide maps of protein–RNA interactions on polyadenylated transcripts and narrows the sequence search space for transcript regions involved in *cis*-regulation of gene expression in response to internal or external stimuli, altered cellular programs or disease.

1. Introduction

Besides transcriptional regulation, gene expression in higher eukaryotes is extensively controlled and regulated at multiple co- and posttranscriptional levels [1,2]. RNA binding proteins (RBPs) and non-coding RNAs form dynamic ribonucleoprotein complexes (RNPs) that control the fate of an RNA molecule throughout its entire lifecycle and affect every aspect of RNA metabolism ranging from splicing to cellular localization and decay [3–5]. While RNP complexes are essential components of the splicing machinery and other messenger RNA (mRNA) processing pathways, they appear equally important towards the physiological function of long intervening noncoding RNAs (lincRNAs). Similar to mRNAs, lincRNAs are abundant and stable RNA polymerase II products that are typically capped, spliced and polyadenylated, but lack protein coding potential and differ in primary sequence conservation from classical mRNAs [6–11]. Experimental methods to identify lincRNAs frequently involve poly(A)+ purification of RNA, thus to date most annotated lincRNAs are polyadenylated [12]. However examples of alternative 3′ end topologies exist: few lincRNAs such as the most abundant isoform of Malat1 form a triple-helical RNA structure at their 3′ end [13,14], while other lincRNAs are stabilized by small nucleolar RNA (snoRNA) interactions at both ends [12] or exist as circular isoforms [15,16]. LincRNAs typically engage in complex RNP networks with numerous chromatin regulators, thereby influencing gene expression and chromatin state [17–23].

In order to foster our understanding of how dynamic protein–RNA interactions, as they occur in RNP complexes, alter gene-expression programs, the global identification of RNA regions bound and regulated by RBPs and other regulatory protein factors remains a central task. Two recent studies provide compelling evidence that the mammalian genome encodes around 800 RBPs, potentially engaging in various RNP complexes that bind and regulate defined sequence or structural elements of polyadenylated transcripts [24,25]. These and other studies suggest the existence of a large number of mRNA binders with diverse molecular functions participating in combinatorial posttranscriptional gene expression networks [2–4]. Importantly, numerous RBPs have been implicated in human disease and pathology and are thus subject of active research [26,27]. Similarly, dozens of lincRNAs were shown to have altered expression profiles in human cancers and appear to be regulated by prominent oncogenes and tumor-supressors [23].

Accelerated by recent developments in next-generation sequencing technology, a variety of methods for broad identification of RBP target sites are continuously emerging. Crosslinking and immunoprecipitation (CLIP) and its adaptations such as High-Throughput Sequencing CLIP (HITS-CLIP), crosslinking and analysis of cDNAs (CRAC), individual-nucleotide resolution CLIP

* Corresponding author. Fax: +49 30 9406 49160.
E-mail address: markus.landthaler@mdc-berlin.de (M. Landthaler).

1046-2023/$ - see front matter © 2013 The Authors. Published by Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ymeth.2013.09.017

(iCLIP) and photoactivatable-ribonucleoside-enhanced CLIP (PAR-CLIP) thus far enabled the characterization of more than three dozens of RNA-binding proteins and their target transcripts [24,28–38]. While HITS-CLIP, CRAC, and iCLIP use high-energy 254 nm wave length UV irradiation for crosslinking, PAR-CLIP relies on 365 nm UV irradiation in combination with non-perturbing metabolic labeling of transcripts with photoreactive nucleoside analogs, such as 4-thiouridine (4SU) to enhance crosslinking efficiency [31]. One of the hallmarks of the PAR-CLIP methodology is the occurrence of a prominent diagnostic T–C transition in cDNA sequence reads, marking the protein–RNA crosslinking site on the target transcript [31]. Despite every method bearing certain advantages and limitations, all approaches share the basic principle of crosslinking a single RBP to its target regions, followed by immunoprecipitation of the protein of interest and sequencing of associated RNA fragments [32–34]. Thus these methods are designed to study individual RNA binding proteins in isolation, without considering the overall context of protein–RNA interactions occurring in an expressed transcriptome at a given time. Using photoactivatable ribonucleoside enhanced crosslinking in combination with oligo (dT) affinity purification of polyadenylated transcripts, we were able to globally identify RNA contact sites of the poly(A)+ RNA-bound proteome by profiling diagnostic T–C nucleotide transitions that occur at direct protein–RNA interaction sites (Fig. 1) [24]. Protein occupancy profiling on polyadenylated RNA enabled the generation of the first transcriptome-wide catalog of potential *cis*-regulatory mRNA regions of a human cell line and revealed protein–RNA contacts throughout large sequence stretches in UTRs and coding sequences. Functional relevance of captured interactions is supported by elevated evolutionary conservation and decreased single-nucleotide-polymorphism (SNP) frequency of crosslinked nucleotides when compared to non-crosslinked nucleotides in the same set of sequences [24].

2. Protein occupancy profiling on long noncoding and messenger RNA protocol

2.1. Material

2.1.1. Buffers

Buffer	Composition
4-Thiouridine stock (1 M)	260.27 mg 4-thiouridine in 1 ml H_2O
D-MEM growth medium	D-MEM high glucose 10% (v/v) fetal bovine serum 2 mM L-glutamine 100 U/ml penicillin 100 U/ml streptomycin
Dephosphorylation buffer	50mM Tris–HCl pH 7.9 at 25 °C 100 mM NaCl 10 mM $MgCl_2$ 1 mM DTT
DNA loading dye (5×)	0.2% (w/v) Bromophenol blue 0.2% (w/v) Xylene cyanol FF 50 mM EDTA, pH 8.0 20% (w/v) Ficoll type 400
Elution buffer	10 mM Tris-HCl pH 7.5 at 25 °C
Lysis/binding buffer	100 mM Tris-HCl pH 7.5 at 25 °C 500 mM LiCl 10 mM EDTA pH 8.0 at 25 °C 1% LiDS add fresh: 5 mM DTT and Complete Mini EDTA-free protease inhibitor (Roche)
NP40 washing buffer	50 mM Tris-HCl pH 7.5 at 25 °C 140 mM LiCl 2 mM EDTA pH 8.0 at 25 °C 0.5% NP40 add fresh: 0.5 mM DTT
PCR buffer (10×)	100 mM Tris–HCl pH 8.0 at 25 °C 500 mM KCl 20 mM $MgCl_2$ 10 mM β-mercaptoethanol 1% (v/v) Triton X-100
PNK buffer	50 mM Tris–HCl pH 7.5 at 25 °C 50 mM NaCl 10 mM $MgCl_2$ 5 mM DTT
Proteinase K buffer	100 mM Tris–HCl pH 7.5 at 25 °C 150 mM NaCl 12.5 mM EDTA 2% (w/v) SDS
RNA ligase buffer with ATP (10×)	500 mM Tris–HCl pH 7.5 at 25 °C 100 $MgCl_2$ 10 mM DTT 10 mM ATP
RNA ligase buffer without ATP (10×)	500 mM Tris–HCl pH 7.5 at 25 °C 100 $MgCl_2$ 10 mM DTT
RNA loading buffer (2×)	8 M Urea 1.5 mM EDTA 1% (w/v) Bromophenol blue
Saturated ammonium sulfate solution	At 25 °C add 766.80 g ammonium sulfate to 1000 g H_2O
SDS–PAGE loading buffer (2×)	100 mM Tris–HCl pH 6.8 at 25 °C 200 mM DTT 4 mM EDTA 4% (w/v) SDS 20% Glycerol 0.2% (w/v) Bromophenol blue
Transfer buffer	25 mM Tris–HCl pH 8.5 at 25 °C 192 mM Glycine 20% (v/v) Methanol

2.2. Oligo(dT) beads, enzymes and oligonucleotides

2.2.1. Oligo(dT) beads

Material	Manufacturer
Dynabeads mRNA DIRECT	Life Technologies

2.2.2. Enzymes and other material

Material	Manufacturer
[γ-^{32}P]-ATP, 3000 Ci/mmol, 10 mCi/ml	Perkin Elmer
ATP (100 mM)	Fermentas
Benzonase	Merck Millipore
Calf intestine alkaline phosphatase (CIP)	New England Biolabs

(continued on next page)

(*continued*)

Material	Manufacturer
dNTP mix (10 mM)	Fermentas
Glycoblue	Ambion
NuPAGE MOPS SDS running buffer (20×)	Life Technologies
NuPAGE Novex 4-12% BT Midi 1.0 gel	Life Technologies
NuSieve GTG agarose	Lonza
Phusion high-fidelity DNA polymerase	Thermo Scientific
Proteinase K	Roche
ProteoSilver Silver Stain Kit	Sigma-Aldrich
Protran nitrocellulose membranes	Whatman
Qiaquick gel extraction kit	Qiagen
RNase I	Ambion
SequaGel UreaGel System	National Diagnostics
Superscript III reverse Transcriptase	Life Technologies
T4 polynucleotide kinase (T4 PNK)	New England Biolabs
T4 RNA ligase 2, truncated K227Q	New England Biolabs
T4 RNA-ligase 1	New England Biolabs

2.2.3. Oligonucleotides

Oligonucleotide	Sequence
5′-Adapter (RNA)	5′-rGrUrUrCrArGrArGrUrUrCrUrArCrArGrUrCrCrGrArCrGrArUrC
3′-Adapter NBC1 (pre-adenylated)	5′-AppTCTAAAATCGTATGCCGTCTTCTGCTTG-InvdT
3′-Adapter NBC2 (pre-adenylated)	5′-AppTCTCCCATCGTATGCCGTCTTCTGCTTG-InvdT
3′-Adapter NBC3 (pre-adenylated)	5′-AppTCTGGGATCGTATGCCGTCTTCTGCTTG-InvdT
3′-Adapter NBC4 (pre-adenylated)	5′-AppTCTTTTATCGTATGCCGTCTTCTGCTTG-InvdT
3′-Adapter NBC5 (pre-adenylated)	5′-AppTCTCACGTCGTATGCCGTCTTCTGCTTG-InvdT
3′-Adapter NBC6 (pre-adenylated)	5′-AppTCTCCATTCGTATGCCGTCTTCTGCTTG-InvdT
3′-Adapter NBC7 (pre-adenylated)	5′-AppTCTCGTATCGTATGCCGTCTTCTGCTTG-InvdT
3′-Adapter NBC8 (pre-adenylated)	5′-AppTCTCTGCTCGTATGCCGTCTTCTGCTTG-InvdT
3′ PCR primer	5′-CAAGCAGAAGACGGCATACGA
5′ PCR primer	5′-AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA

2.3. Procedure

2.3.1. Cell culture and UV-crosslinking

Protein occupancy profiling is readily applicable to a variety of cellular systems and experimental conditions. We successfully used the described procedure to profile the protein occupancy on polyadenylated transcripts in HEK293, HeLa, MCF7, HUVEC and mouse embryonic stem cells. Critical to the success of the experiment is the efficient incorporation of the photoreactive nucleoside, 4-thiouridine, into nascent cellular RNAs. We recommend testing labeling efficiencies and compare it to 4SU incorporation rates observed in HEK293 cells by thiol-specific biotinylation and dot blot assay [39] or LC-MS analysis [40]. In HEK293 cells, 4SU substitution rates between 1% and 4% yield sufficient protein–RNA crosslinking to characterize global binding preferences of RNA binding proteins [31,33].

Cells are grown in D-MEM high glucose growth medium and expanded to 15 cm tissue culture plates. Typically five 15 cm tissue culture plates ($\sim 10^8$ cells) are labeled at 80% confluency with 100 μM 4SU for 12–16 h. If desired an optional 100 μM 4SU labeling pulse can be performed shortly before harvesting (1–2 h) to ensure labeling of short lived transcripts. To induce formation of a covalent bond between 4SU labeled RNA and bound proteins, culture media is removed and cells are crosslinked with 365 nm UV light (0.2 J/cm^2) on ice using a Stratalinker 2400 (Stratagene) [41]. Crosslinked cells are scraped off with a rubber policeman and collected by centrifugation (235 RCF, 4 °C, 5 min). The cell pellet is washed twice with ice-cold PBS followed by centrifugation and flash-freezing of collected cells in liquid nitrogen for long-term storage (alternatively proceed to Section 2.3.2 immediately).

2.3.2. Preparation of cell lysate and affinity purification of mRNA

Cells are lysed in 10 pellet volumes of lysis/binding buffer by gentle pipetting and incubation at room temperature for 10 min. Genomic DNA is sheared by passing the lysate 6 times through a 21 gauge needle. The desired volume of oligo (dT) Dynabeads is briefly washed in 1 ml lysis/binding buffer. For five 15 cm tissue culture plates, an equivalent of 2 ml Dynabead suspension, as supplied by the manufacturer (concentrated bed volume ~15 μl), is added to cell lysates and incubated for 1 h at room temperature on a rotating wheel. Following incubation, beads are concentrated on a magnet and supernatant is stored on ice for multiple rounds of oligo (dT) affinity purification. Beads are washed three times in 10 pellet volumes of lysis/binding buffer, containing 1% lithium dodecyl sulfate (LiDS) to ensure stringent isolation of protein-mRNA complexes. Note that decreasing LiDS concentrations can improve yields, but might reduce specificity. Beads are subsequently washed three additional times in 10 pellet volumes of NP40 washing buffer and crosslinked poly(A)+ RNA–protein complexes are heat-eluted in 200–500 μl low-salt elution buffer by a 2 min incubation at 80 °C. Eluate is placed on ice and beads are re-incubated with lysate for a total number of 3 oligo (dT) hybridizations, repeating the described procedure. Eluates are combined and stored at −80 °C or subjected to nuclease treatment (see Section 2.3.3).

From $\sim 10^8$ crosslinked cells, we typically obtain 10–30 μg of poly(A)+ purified RNA. Note that we did not evaluate the minimal RNA concentration necessary for successful experiments. To control for efficient protein–RNA crosslinking and specific enrichment of mRNA bound proteins in oligo (dT) precipitates, ~20 μl of the eluate are incubated with with RNase I (10 U/ml) and Benzonase (125 U/ml) for 2 h at 37 °C in elution buffer containing 1 mM $MgCl_2$. Following RNA digestion, proteins are analyzed by SDS–PAGE. Fig. 2 shows a representative silver stained SDS–PAGE gel of three sequential oligo (dT) purifications with and without 365 nm UV crosslinking. A defined pattern of protein bands in all three elutions (E1–E3) indicates specific isolation of RNA bound proteins. Multiple rounds of poly(A)+ RNA depletion by oligo (dT) affinity purification are beneficial to increase the yield of lowly expressed transcripts or transcripts with short poly(A) tails and their bound proteins. However, the amount of co-purifying contaminating RNA species such as ribosomal RNA increases with each additional hybridization. We found that three consecutive oligo (dT) affinity purifications yield 80–90% of mRNA mapping sequence reads after following the subsequently described procedures [24].

2.3.3. Nuclease treatment and isolation of protein protected mRNA fragments

To generate protein protected RNA fragments of 20–60 nucleotides (nts) in length, RNase I (100 U/μl) is added to eluates at a dilution of 1:1000–1:5000 and samples are incubated 10 min at

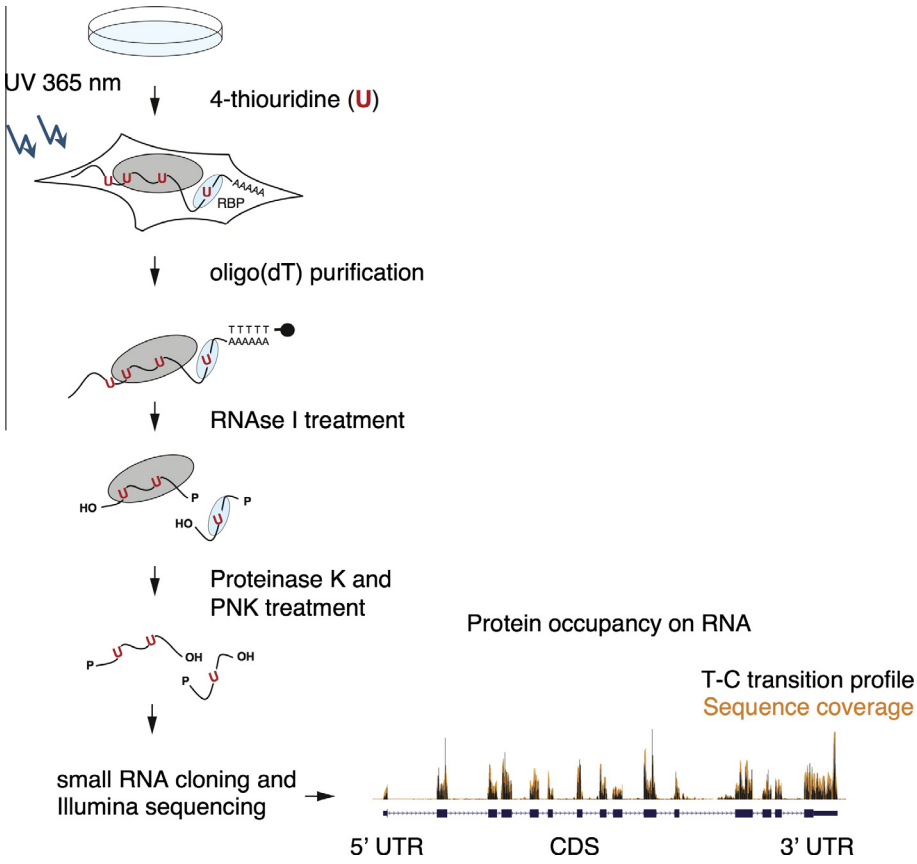

Fig. 1. Graphical representation of the protein occupancy profiling methodology. Culture media is supplemented with 4SU, which is taken up by cells and incorporated into nascent transcripts. UV irradiation at 365 nm induces covalent photocrosslinking between 4SU labeled RNA and associated RBPs. Following crosslinking, polyadenylated transcripts are purified using oligo (dT) beads, protein protected RNA fragments are generated by RNase I digestion and proteins are removed by proteinase K treatment. RNA fragments are converted into a cDNA library and next-generation sequenced. Mapping of sequencing reads to the respective reference sequence reveals diagnostic T–C transitions as a signature of protein–RNA crosslinking. T–C transitions are used to globally profile protein interaction sites across the transcriptome.

37 °C. Depending on the cell types used, careful optimization of RNase treatment in pilot experiments is recommended. Use of RNase I does not introduce a nucleotide bias [42]. Following incubation, 4 volumes of saturated ammonium sulfate solution are added and samples are incubated 30 min on ice for precipitation. We found that salting out with ammonium sulfate is superior to other protein precipitation methods (e.g. ethanol or trichloroacetic acid precipitation), as it efficiently removes non-protein bound RNA from the sample. Note that the presence or absence of diagnostic T–C transitions, as a signature of protein–RNA crosslinking will be used during data analysis to discriminate crosslinked from non-crosslinked RNA sequences.

Precipitated protein–RNA complexes are collected by centrifugation (20,000 RCF, 4 °C, 30 min) and supernatant is removed carefully. Resulting protein pellets are air-dried, resuspended in a desired volume of SDS-loading buffer and denatured for 3 min at 95 °C. Protein–RNA complexes are separated on a NuPAGE Novex 4–12% BT Midi 1.0 gel and transferred to a nitrocellulose membrane (1 h at 20 V) to further remove non-protein bound RNA. Protein-containing lanes are excised and sliced membrane pieces are incubated in 1× proteinase K buffer containing 4 mg/ml proteinase K for 30 min at 55 °C. RNA is recovered by phenol/chloroform extraction and ethanol precipitation.

2.3.4. Generation of adapter ligation compatible, radiolabeled RNA fragments

RNase I treated protein protected RNA fragments carry 5′ hydroxyl groups and 3′monophosphate termini, which require conversion into 5′ phosphates and 3′ hydroxyl groups prior to adapter ligation, respectively. Following ethanol precipitation, recovered RNA fragments are re-hydrated in 42.5 µl H_2O and 5 µl 10× dephosphorylation buffer. Calf intestinal alkaline phosphatase is added to a final concentration of 0.5 U/µl and reaction is incubated 30 min at

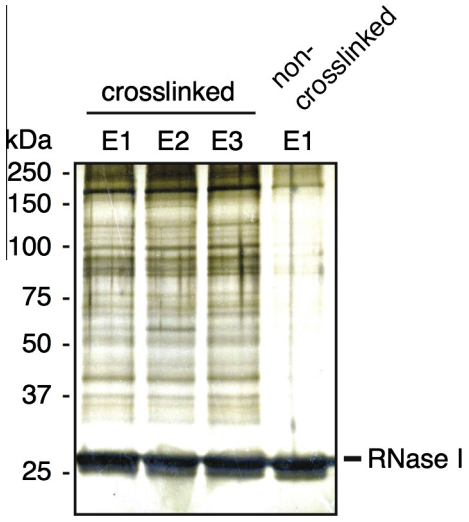

Fig. 2. Oligo (dT) affinity purification of crosslinked protein–RNA complexes. Silver-stained SDS–PAGE gel of representative experiment is shown. 20 µl of oligo (dT) affinity purified 4SU labeled RNA with or without 365 nm crosslinking were treated with RNase I (10 U/ml) and Benzonase (125 U/ml) for 2 h at 37 °C. Proteins were analyzed by SDS–PAGE and gel was silver stained. E1–E3 represents elutions 1–3 of crosslinked protein–RNA complexes, respectively. Elution E1 of non-crosslinked polyA+ RNA is shown as negative control.

37 °C. Dephosphorylated RNA is phenol/chloroform extracted, ethanol precipitated and collected by centrifugation. Recovered RNA is prepared for radiolabeling at the 5′ end by re-hydrating in 39.8 µl H_2O, 0.2 µl [γ-^{32}P]-ATP (0.2 µCi/µl final) and 5 µl 10x PNK buffer. T4 PNK is added to a final concentration of 1 U/µl. After 30 min of incubation at 37 °C, non-radioactive ATP is added to a final concentration of 1 mM and incubated for 5 min at 37 °C to complete phosphorylation of RNA. Radiolabeled RNA is again phenol/chloroform extracted, ethanol precipitated and collected by centrifugation.

2.3.5. 3′ Adapter ligation

To identify the desired size population of RNA during adapter ligation steps, radiolabeled 21 and 50 nt RNA size markers are used as ligation controls. 40 fmol of 5′-^{32}P-labeled RNA size markers are diluted 1:100 (in equimolar ratios) and adjusted to a volume of 6 µl. Radiolabeled RNA samples are re-hydrated in 6 µl H_2O and 2 µl 10× RNA ligase buffer, 1 µl of 100 µM pre-adenylated 3′ adapter and 10 µl 24% PEG 8000 are added to size markers and samples. The use of pre-adenylated adapters eliminates the need for ATP during ligation, and thus minimizes the problem of pool RNA adenylation at the 5′ phosphate that can lead to circularization of RNA fragments. We are using barcoded 3′ adapters that allow multiplexed Illumina sequencing (see Section 2.2.3). RNA is incubated at 95 °C for 30 s to reduce secondary structure and placed on ice immediately. 1 µl T4 RNA ligase 2, truncated K227Q is added to each sample and ligation reaction is incubated overnight at 16 °C. The truncated K227Q version of RNA ligase 2 minimizes adenylate transfer from the 3′ adapter 5′ phosphate to the 5′ phosphate of small RNA fragments and thus reduces RNA circularization. Following overnight incubation, one volume RNA loading buffer is added to the samples and RNA is separated on a 15% denaturing 7.5 M urea polyacrylamide gel (75 min., 30 W), running in 1× TBE. Radiolabeled RNA is visualized by exposure to a phosphorimager screen and 3′ ligated size markers are used to approximate the length of ligated RNA fragments (shown in Fig. 3 A for biological replicate experiments). Successfully 3′ ligated RNA and size markers are excised from the gel and eluted in 400 µl 0.3 M NaCl shaking on a thermomixer at ~1000 rpm and 4 °C overnight. Following overnight gel elution, RNA is ethanol precipitated and collected by centrifugation.

2.3.6. 5′ Adapter ligation

5 µl H_2O, 2 µl 10× RNA ligase buffer with ATP, 1 µl of 100 µM 5′ adapter and 10 µl 24% PEG 8000 are added to 3′ ligated RNA samples and size markers. RNA is denatured at 95 °C for 30 s, put on ice and 2 µl T4 RNA ligase 1 is added to the reaction. Ligation is performed for 1 h at 37 °C. Following incubation, RNA is separated on a 12% denaturing 7.5 M urea polyacrylamide gel (75 min., 30 W) and RNA is visualized by exposure to a phosphorimager screen. Again, 3′ and 5′ ligated size markers are used to approximate the length of the successfully 3′ and 5′ ligated RNA fragments (shown in Fig. 3 B for biological replicate experiments). RNA migrating at the desired size is excised from the gel and eluted in 400 µl 0.3 M NaCl shaking on a thermomixer at ~1000 rpm and 4 °C overnight. As a carrier, 1 µl 100 µM 3′ PCR primer is added to gel pieces during elution. Following overnight elution, the RNA is ethanol precipitated and collected by centrifugation.

2.3.7. Reverse transcription and pilot PCR

To convert recovered 3′ and 5′ ligated RNA into a cDNA library, RNA is taken up in 5.6 µl H_2O, 4.2 µl dNTPs (2 mM each), 3 µl 5× first strand buffer and 1.5 µl 0.1 M DTT. Following a 30 s incubation at 95 °C, RNA is placed on ice for 3 min, before primer annealing is performed at 50 °C for 3 min. 0.75 µl Superscript III Reverse Transcriptase is added to reaction and incubated 30 min at 42 °C. After cDNA synthesis, sample volume is adjusted to 100 µl using H_2O.

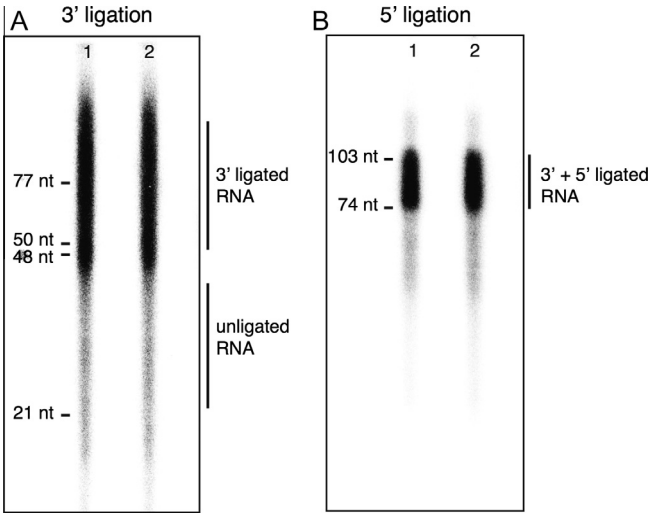

Fig. 3. Small RNA cloning of protein protected RNA fragments from two biological replicate experiments in HEK293 cells. (A) Autoradiogram of 3′ adapter ligated RNA fragments separated on a 15% denaturing polyacrylamide urea gel. RNase I treatment generates fragments of 20–60 nt in length (unligated RNA). Upon ligation of the 29 nt 3′ adapter, the majority of RNA shifts to a 50–100 nt size range. The RNA population migrating between 50 and 77 nt size markers is excised from gel and subjected to 5′ ligation. Numbers 1 and 2 indicate different biological replicate experiments. (B) Autoradiogram of 3′ and 5′ adapter ligated RNA fragments separated on a 12% denaturing polyacrylamide urea gel. Ligation of the 27 nt 5′ adapter leads to a size shift of 3′ ligated RNA fragments. The RNA population migrating between 74 and 103 nt size markers is excised from gel and reverse-transcribed. Numbers 1 and 2 indicate different biological replicate experiments.

10 µl diluted cDNA are added to 40 µl PCR mastermix (10 µl 5× Phusion buffer, 1.25 µl dNTPs (10 mM), 0.25 µl 100 µM 5′ PCR primer, 0.25 µl 100 µM 3′ PCR primer, 0.5 µl Phusion High-Fidelity DNA Polymerase, 27.75 µl H_2O) and subjected to PCR amplification after activation at 94 °C for 2 min using the following cycle conditions: 94 °C 45 s, 50 °C 85 s, 72 °C 60 s. To determine the number of cycles necessary for linear amplification, 5 µl of sample are removed every 2 cycles starting from cycle number 8–22. Samples are analyzed on a 2.5% agarose gel containing 0.4 µg/ml of ethidium bromide to check for linear amplification. As shown in Fig. 4 the PCR products should appear as a single band migrating at the expected size of 125–150 nt. In some cases a weaker band that corresponds to adapter-dimer or template switch products might be observed at ~100 nt. Asterisks indicate the number of PCR cycles that were used during large scale PCR amplification (Fig. 4).

2.3.8. Final PCR amplification and purification of cDNA library

After determining the optimal cycle number for linear PCR amplification, 30 µl cDNA are added to 120 µl PCR mastermix (see Section 2.3.7) and split into three equal fractions of 50 µl each. PCR amplification for all three 50 µl reactions is performed using the same cycle conditions as described above. Following amplification, the PCR product is ethanol precipitated and separated on a 2.5% low melting agarose gel, containing 0.4 µg/ml of ethidium bromide for 2.5 h at 150 V. cDNA is visualized using a UV transilluminator and a band migrating at 125–150 nt is excised. cDNA is purified using the Qiaquick gel extraction kit according to the manufacturer's instructions. cDNA is eluted in 30 µl RNase free water and yield as well as fragment size are determined using a Qubit flurometric quantification instrument (Life Technologies) and Agilent Bioanalyzer DNA chip. Multiplexed cDNA libraries are sequenced using 100 cycles on a Illumina HiSeq 2500 instrument.

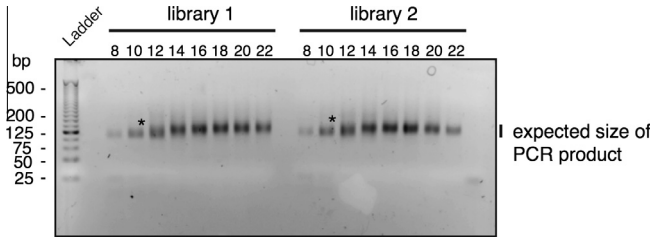

Fig. 4. Small scale PCR amplification of cDNA library. Agarose gel after small scale pilot PCR is shown. 5 µl of sample were removed at indicated cycle numbers and analyzed on a 2.5% low melting agarose gel, containing 0.4 µg/ml of ethidium bromide. Bands migrating between 125 and 150 bp correspond to the expected size of the PCR product. Asterisks indicate cycle numbers chosen for large scale PCR amplification.

2.3.9. Computational data analysis

When mapping protein occupancy profiling reads obtained from Illumina sequencing to the respective reference genome, diagnostic T–C transitions are observed at high frequency (sequencing read mapping statistics of representative cDNA libraries are provided as Supplementary Table S1). Importantly in the absence of UV 365 nm crosslinking, 4SU labeled RNA does not lead to significant accumulation of spontaneously occurring T–C transitions. As shown in Fig. 5A, T–C transitions detected in non-crosslinked 4SU labeled RNA occur at low frequency and represent only 5.15% of the T–C transition frequency that is observed upon crosslinking. Thus the frequency of spontaneously occurring T–C transitions is similar to the background mutation rate of unsubstituted nucleotides in protein occupancy profiling data (Fig. 5A). Note that the ratio of perfect matching to edited sequence reads is reversed upon 365 nm UV irradiation, indicating effiecient protein–RNA crosslinking. As mentioned earlier, T–C transitions are the signature of protein–RNA crosslinking in 4SU labeled RNA. Thus we use the consensus T–C signature observed in at least two replicate experiments to globally profile protein–RNA interactions across the entire poly(A)+ RNA transcriptome. More specifically, we use TopHat2 (version 2.06) [43,44] for spliced alignment of strand-specific protein occupancy profiling reads to the human reference genome sequence (hg18). Prior knowledge on candidate splice junctions from EnsEMBL (release 54, www.ensembl.org) is used to increase the sensitivity of the mapping process. We separate all reads by strand and generate two strand-specific mpileup read coverage files with samtools (version 0.1.18, [45]). These file are subsequently used to produce a separate bedgraph file for each strand (Watson/Crick). Additionally, a single bedgraph file for strand-specific T–C conversions is produced in a similar manner. T–C transition sites are only included in the final file if at least two transitions from independent reads are observed on average. Bedgraph files can be conveniently loaded into UCSC hg18 genome browser for visualization purposes. Fig. 5B shows the distribution of sequence normalized T–C transition event counts mapping to different exonic transcript regions as well as introns, lincRNAs [46] and ncRNAs for two occupancy profiling libraries (also see Supplementary Table S2).

To streamline the described analysis process, we have developed "poppi", the protein occupancy profiling pipeline (unpublished). Poppi performs all the described analysis steps and allows quality assessment of protein occupancy profiling experiments including screening of the diagnostic T–C transition counts. In addition, it allows the correlation of protein occupancy profiles to annotated features, between replicates as well as between different experimental conditions to define significant local changes in protein occupancy. Figs. 6A–C exemplify protein occupancy profiles on polyadenylated mRNA and lincRNA for biological replicate experiments performed in HEK293 cells. In Fig. 6A, the genomic region encoding the full length EEF2 transcript is visualized. Fig. 6B shows a zoom-in to the 3′UTR region of EEF2. The HEK293 occupancy profile of the long intervening noncoding RNA DANCR (also known as ANCR or KIAA0114), which is required to maintain the undifferentiated state in human somatic tissue progenitor cells [47] is shown in Fig. 6C.

3. Concluding remarks

The broad discovery and characterization of protein binding sites on RNA that control distinct posttranscriptional processes was facilitated by application of various CLIP based approaches to comprehensively map protein–RNA interaction sites. However, until now all available methods focus only on the binding specificity of individual RNA-binding proteins [32–34]. Protein occupancy profiling provides a snapshot view of all protein–RNA interactions occurring in a transcriptome at a given time without limiting the captured interactions to a specific RBP. Thus investigators can now take the sequence space of potential *cis*-regulatory elements into consideration and monitor differential changes in protein occupancy on specific transcript regions in response to intra- and extracellular signals or stimuli. Importantly such studies can be performed in an unbiased manner, without prior knowledge of

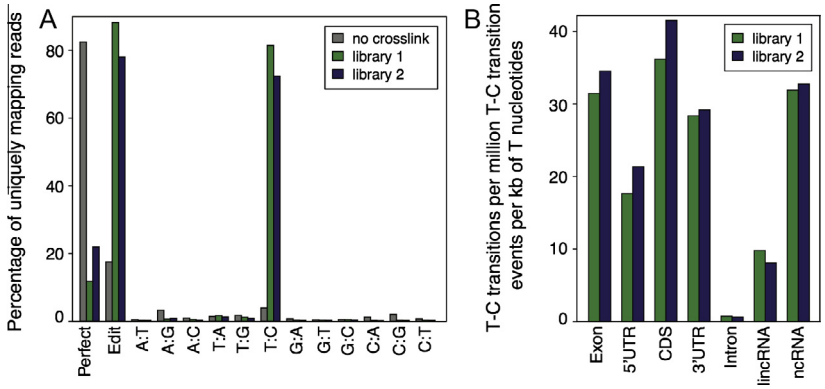

Fig. 5. T–C transition frequency in presence or absence of UV crosslinking and distribution of T–C transitions to different transcript and genomic regions. (A) Percentage of uniquely mapping sequence reads containing specific nucleotide mismatches in occupancy profiling data. T–C transition frequency of 4SU labeled and crosslinked RNA (libraries 1 and 2) is compared to that of non-crosslinked 4SU labeled RNA. T–C transitions detected in non-crosslinked 4SU labeled RNA represent ~5% of the T–C transition frequency that is observed upon crosslinking. T–C mismatches are the signature of efficient crosslinking of 4SU-labeled RNA to protein. (B) Distribution of T–C transitions to different transcript and genomic regions shown for occupancy profiling libraries 1 and 2. The human lincRNA catalog published by Cabili et al. was used as reference to estimate mapping of T–C transitions to lincRNAs [46].

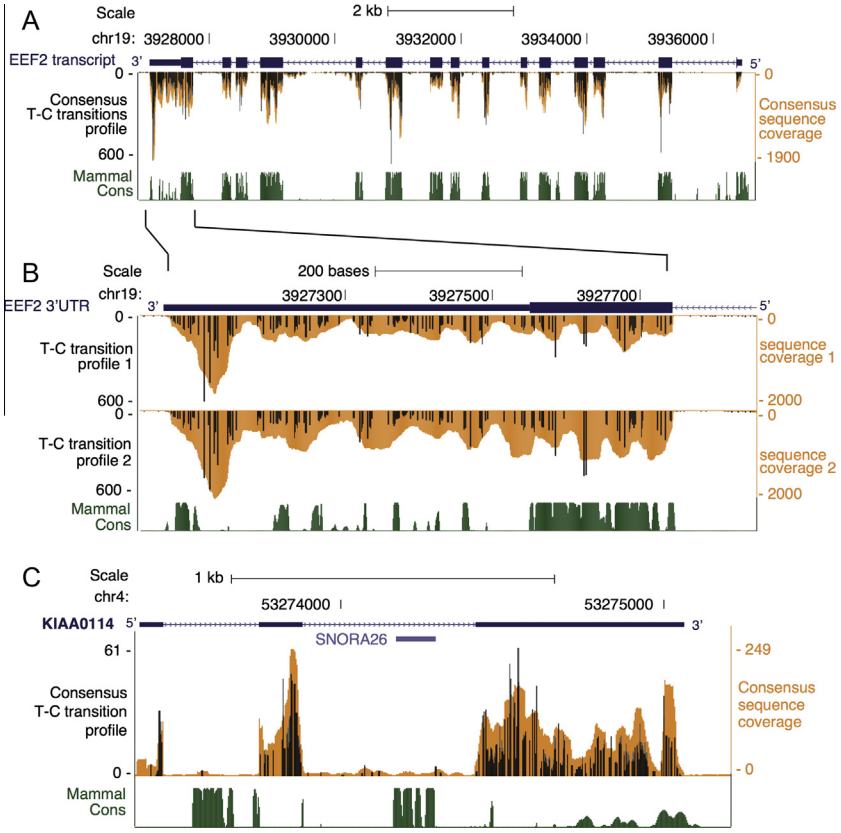

Fig. 6. Browser view of protein occupancy profiles on mRNA and lincRNA. (A) Browser view of genomic region encoding full length EEF2 transcript. Protein–RNA crosslinking events represented by diagnostic T–C transitions are shown in black, as consensus T–C transition profile. Sequence coverage is shown in orange, PhyloP mammalian conservation score is indicated in green. (B) As in (A), but zoom-in to 3′UTR of EEF2 transcript is shown. Diagnostic T–C transitions are shown as individual T–C transition profiles obtained from biological replicate experiments (libraries 1 and 2, respectively). (C) As in (A), but genomic region encoding the long noncoding RNA DANCR (also known as ANCR or KIAA0114) is shown.

the protein component mediating the observed effects. By using oligo (dT) affinity purification of RNA, the dynamics in protein occupancy can readily be captured for protein coding transcripts as well as many lincRNAs. Global mapping of protein–RNA contact sites on lincRNAs can provide insights into the postulated modular design of these noncoding RNAs and determine the individual lincRNA–protein interaction domains. Such research can shed valuable light on the fascinating question of how lincRNAs are functionally assembled and how they interact with various protein components to mediate their regulatory tasks.

Protein occupancy profiling makes use of a diagnostic mutation signature present in 4SU labeled RNA that is crosslinked to proteins in binding distance. It is important to note that the presence or absence of the diagnostic mutation in sequence reads is used to separate crosslinking signal from background noise. Efficient labeling of cellular RNA with 4SU is essential to the success of the experiment and might require optimization. Use of stringent and partly denaturing mRNA purification conditions ensure specific isolation of directly interacting protein–RNA complexes, while RNase I is used for generation of protein protected RNA fragments without nucleotide bias. Protein precipitation by ammonium sulfate and transfer of separated complexes onto nitrocellulose are key steps towards the reduction of free RNA and ensure high signal to noise ratios when comparing T–C containing versus perfect matching reads (Fig. 5A).

In the future a central task will be to overlap occupied regions with evolutionary constrained sequences and RNA candidate structures [48] as well as with RNA interaction data of individual proteins [49] to identify specific RNA regulatory elements and their structural contexts. Capturing the dynamics in protein occupancy on functional RNA elements holds potential to reveal unappreciated aspects of how posttranscriptional gene regulation contributes to complex biological processes in developmental or disease. In addition, we envision the identification of differentially occupied mRNA sites to be highly valuable towards the examination of rapidly emerging data on genetic variation between individuals.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ymeth.2013.09.017.

References

[1] G. Dreyfuss, V.N. Kim, N. Kataoka, Nat. Rev. Mol. Cell Biol. 3 (2002) 195–205.
[2] J.D. Keene, Nat. Rev. Genet. 8 (2007) 533–543.
[3] J.D. Keene, S.A. Tenenbaum, Mol. Cell 9 (2002) 1161–1167.
[4] J.D. Keene, P.J. Lager, Chromosome Res. 13 (2005) 327–337.
[5] T. Glisovic, J.L. Bachorik, J. Yong, G. Dreyfuss, FEBS Lett. 582 (2008) 1977–1986.
[6] J. Ponjavic, C.P. Ponting, G. Lunter, Genome Res. 17 (2007) 556–565.
[7] M. Guttman, M. Garber, J.Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, M.J. Koziol, A. Gnirke, C. Nusbaum, J.L. Rinn, E.S. Lander, A. Regev, Nat. Biotechnol. 28 (2010) 503–510.
[8] M. Guttman, I. Amit, M. Garber, C. French, M.F. Lin, D. Feldser, M. Huarte, O. Zuk, B.W. Carey, J.P. Cassady, M.N. Cabili, R. Jaenisch, T.S. Mikkelsen, T. Jacks, N. Hacohen, B.E. Bernstein, M. Kellis, A. Regev, J.L. Rinn, E.S. Lander, Nature 458 (2009) 223–227.
[9] B. Banfai, H. Jia, J. Khatun, E. Wood, B. Risk, W.E. Gundling Jr., A. Kundaje, H.P. Gunawardena, Y. Yu, L. Xie, K. Krajewski, B.D. Strahl, X. Chen, P. Bickel, M.C. Giddings, J.B. Brown, L. Lipovich, Genome Res. 22 (2012) 1646–1657.
[10] S.A. Slavoff, A.J. Mitchell, A.G. Schwaid, M.N. Cabili, J. Ma, J.Z. Levin, A.D. Karger, B.A. Budnik, J.L. Rinn, A. Saghatelian, Nat. Chem. Biol. 9 (2013) 59–64.

[11] M. Guttman, P. Russell, N.T. Ingolia, J.S. Weissman, E.S. Lander, Cell 154 (2013) 240–251.
[12] I. Ulitsky, D.P. Bartel, Cell 154 (2013) 26–46.
[13] J.E. Wilusz, C.K. JnBaptiste, L.Y. Lu, C.D. Kuhn, L. Joshua-Tor, P.A. Sharp, Genes Dev. 26 (2012) 2392–2407.
[14] C.J. Brown, B.D. Hendrich, J.L. Rupert, R.G. Lafreniere, Y. Xing, J. Lawrence, H.F. Willard, Cell 71 (1992) 527–542.
[15] T.B. Hansen, T.I. Jensen, B.H. Clausen, J.B. Bramsen, B. Finsen, C.K. Damgaard, J. Kjems, Nature 495 (2013) 384–388.
[16] S. Memczak, M. Jens, A. Elefsinioti, F. Torti, J. Krueger, A. Rybak, L. Maier, S.D. Mackowiak, L.H. Gregersen, M. Munschauer, A. Loewer, U. Ziebold, M. Landthaler, C. Kocks, F. le Noble, N. Rajewsky, Nature 495 (2013) 333–338.
[17] S. Bertani, S. Sauer, E. Bolotin, F. Sauer, Mol. Cell 43 (2011) 1040–1046.
[18] K.C. Wang, Y.W. Yang, B. Liu, A. Sanyal, R. Corces-Zimmerman, Y. Chen, B.R. Lajoie, A. Protacio, R.A. Flynn, R.A. Gupta, J. Wysocka, M. Lei, J. Dekker, J.A. Helms, H.Y. Chang, Nature 472 (2011) 120–124.
[19] M.C. Tsai, O. Manor, Y. Wan, N. Mosammaparast, J.K. Wang, F. Lan, Y. Shi, E. Segal, H.Y. Chang, Science 329 (2010) 689–693.
[20] G.D. Penny, G.F. Kay, S.A. Sheardown, S. Rastan, N. Brockdorff, Nature 379 (1996) 131–137.
[21] J.L. Rinn, M. Kertesz, J.K. Wang, S.L. Squazzo, X. Xu, S.A. Brugmann, L.H. Goodnough, J.A. Helms, P.J. Farnham, E. Segal, H.Y. Chang, Cell 129 (2007) 1311–1323.
[22] S. Schoeftner, A.K. Sengupta, S. Kubicek, K. Mechtler, L. Spahn, H. Koseki, T. Jenuwein, A. Wutz, EMBO J. 25 (2006) 3110–3122.
[23] J.L. Rinn, H.Y. Chang, Annu. Rev. Biochem. 81 (2012) 145–166.
[24] A.G. Baltz, M. Munschauer, B. Schwanhausser, A. Vasile, Y. Murakawa, M. Schueler, N. Youngs, D. Penfold-Brown, K. Drew, M. Milek, E. Wyler, R. Bonneau, M. Selbach, C. Dieterich, M. Landthaler, Mol. Cell 46 (2012) 674–690.
[25] A. Castello, B. Fischer, K. Eichelbaum, R. Horos, B.M. Beckmann, C. Strein, N.E. Davey, D.T. Humphreys, T. Preiss, L.M. Steinmetz, J. Krijgsveld, M.W. Hentze, Cell 149 (2012) 1393–1406.
[26] K.E. Lukong, K.W. Chang, E.W. Khandjian, S. Richard, Trends Genet. 24 (2008) 416–425.
[27] A. Castello, B. Fischer, M.W. Hentze, T. Preiss, Trends Genet. 29 (2013) 318–327.
[28] J. Ule, K.B. Jensen, M. Ruggiu, A. Mele, A. Ule, R.B. Darnell, Science 302 (2003) 1212–1215.
[29] D.D. Licatalosi, A. Mele, J.J. Fak, J. Ule, M. Kayikci, S.W. Chi, T.A. Clark, A.C. Schweitzer, J.E. Blume, X. Wang, J.C. Darnell, R.B. Darnell, Nature 456 (2008) 464–469.
[30] S. Granneman, G. Kudla, E. Petfalski, D. Tollervey, Proc. Natl. Acad. Sci. USA 106 (2009) 9613–9618.
[31] M. Hafner, M. Landthaler, L. Burger, M. Khorshid, J. Hausser, P. Berninger, A. Rothballer, M. Ascano Jr., A.C. Jungkamp, M. Munschauer, A. Ulrich, G.S. Wardle, S. Dewell, M. Zavolan, T. Tuschl, Cell 141 (2010) 129–141.
[32] J. Konig, K. Zarnack, N.M. Luscombe, J. Ule, Nat. Rev. Genet. 13 (2012) 77–83.
[33] M. Ascano, M. Hafner, P. Cekan, S. Gerstberger, T. Tuschl, Wiley Interdiscip. Rev. RNA 3 (2012) 159–177.
[34] M. Milek, E. Wyler, M. Landthaler, Semin. Cell Dev. Biol. 23 (2012) 206–212.
[35] N. Mukherjee, D.L. Corcoran, J.D. Nusbaum, D.W. Reid, S. Georgiev, M. Hafner, M. Ascano Jr., T. Tuschl, U. Ohler, J.D. Keene, Mol. Cell 43 (2011) 327–339.
[36] S. Lebedeva, M. Jens, K. Theil, B. Schwanhausser, M. Selbach, M. Landthaler, N. Rajewsky, Mol. Cell 43 (2011) 340–352.
[37] R. Graf, M. Munschauer, G. Mastrobuoni, F. Mayr, U. Heinemann, S. Kempa, N. Rajewsky, M. Landthaler, RNA Biol. 10 (2013).
[38] A.C. Jungkamp, M. Stoeckius, D. Mecenas, D. Grun, G. Mastrobuoni, S. Kempa, N. Rajewsky, Mol. Cell 44 (2011) 828–840.
[39] L. Dolken, Z. Ruzsics, B. Radle, C.C. Friedel, R. Zimmer, J. Mages, R. Hoffmann, P. Dickinson, T. Forster, P. Ghazal, U.H. Koszinowski, RNA 14 (2008) 1959–1972.
[40] A. Andrus, R.G. Kuimelis, Curr. Protoc. Nucleic Acid. Chem., (2001) (Chapter 10, Unit 10 5).
[41] M. Hafner, M. Landthaler, L. Burger, M. Khorshid, J. Hausser, P. Berninger, A. Rothballer, M. Ascano, A.C. Jungkamp, M. Munschauer, A. Ulrich, G.S. Wardle, S. Dewell, M. Zavolan, T. Tuschl, J. Vis. Exp. (41) (2010).
[42] P.F. Spahr, B.R. Hollingworth, J. Biol. Chem. 236 (1961) 823–831.
[43] C. Trapnell, L. Pachter, S.L. Salzberg, Bioinformatics 25 (2009) 1105–1111.
[44] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, S.L. Salzberg, Genome Biol. 14 (2013) R36.
[45] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, Bioinformatics 25 (2009) 2078–2079.
[46] M.N. Cabili, C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev, J.L. Rinn, Genes Dev. 25 (2011) 1915–1927.
[47] M. Kretz, D.E. Webster, R.J. Flockhart, C.S. Lee, A. Zehnder, V. Lopez-Pajares, K. Qu, G.X. Zheng, J. Chow, G.E. Kim, J.L. Rinn, H.Y. Chang, Z. Siprashvili, P.A. Khavari, Genes Dev. 26 (2012) 338–343.
[48] K. Lindblad-Toh, M. Garber, O. Zuk, M.F. Lin, B.J. Parker, S. Washietl, P. Kheradpour, J. Ernst, G. Jordan, E. Mauceli, L.D. Ward, C.B. Lowe, A.K. Holloway, M. Clamp, S. Gnerre, J. Alfoldi, K. Beal, J. Chang, H. Clawson, J. Cuff, F. Di Palma, S. Fitzgerald, P. Flicek, M. Guttman, M.J. Hubisz, D.B. Jaffe, I. Jungreis, W.J. Kent, D. Kostka, M. Lara, A.L. Martins, T. Massingham, I. Moltke, B.J. Raney, M.D. Rasmussen, J. Robinson, A. Stark, A.J. Vilella, J. Wen, X. Xie, M.C. Zody, J. Baldwin, T. Bloom, C.W. Chin, D. Heiman, R. Nicol, C. Nusbaum, S. Young, J. Wilkinson, K.C. Worley, C.L. Kovar, D.M. Muzny, R.A. Gibbs, A. Cree, H.H. Dihn, G. Fowler, S. Jhangiani, V. Joshi, S. Lee, L.R. Lewis, L.V. Nazareth, G. Okwuonu, J. Santibanez, W.C. Warren, E.R. Mardis, G.M. Weinstock, R.K. Wilson, K. Delehaunty, D. Dooling, C. Fronik, L. Fulton, B. Fulton, T. Graves, P. Minx, E. Sodergren, E. Birney, E.H. Margulies, J. Herrero, E.D. Green, D. Haussler, A. Siepel, N. Goldman, K.S. Pollard, J.S. Pedersen, E.S. Lander, M. Kellis, Nature 478 (2011) 476–482.
[49] Y. Zheng, C. Zhang, D.R. Croucher, M.A. Soliman, N. St-Denis, A. Pasculescu, L. Taylor, S.A. Tate, W.R. Hardy, K. Colwill, A.Y. Dai, R. Bagshaw, J.W. Dennis, A.C. Gingras, R.J. Daly, T. Pawson, Nature 499 (2013) 166–171.