



Published in final edited form as:

ACS Chem Biol. 2020 August 21; 15(8): 2247–2258. doi:10.1021/acscchembio.0c00445.

## Enzymatic RNA Biotinylation for Affinity Purification and Identification of RNA–Protein Interactions

**Kayla N. Busby,**

Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, California 92093, United States

**Amitkumar Fulzele,**

Division of Biological Sciences, University of California, San Diego, La Jolla, California 92093, United States

**Dongyang Zhang,**

Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, California 92093, United States

**Eric J. Bennett,**

Division of Biological Sciences, University of California, San Diego, La Jolla, California 92093, United States

**Neal K. Devaraj**

Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, California 92093, United States

### Abstract

Throughout their cellular lifetime, RNA transcripts are bound to proteins, playing crucial roles in RNA metabolism, trafficking, and function. Despite the importance of these interactions, identifying the proteins that interact with an RNA of interest in mammalian cells represents a major challenge in RNA biology. Leveraging the ability to site-specifically and covalently label an RNA of interest using *E. coli* tRNA guanine transglycosylase and an unnatural nucleobase substrate, we establish the identification of RNA–protein interactions and the selective enrichment of cellular RNA in mammalian systems. We demonstrate the utility of this approach through the identification of known binding partners of 7SK snRNA via mass spectrometry. Through a minimal 4-nucleotide mutation of the long noncoding RNA HOTAIR, enzymatic biotinylation enables identification of putative HOTAIR binding partners in MCF7 breast cancer cells that

---

**Corresponding Author: Neal K. Devaraj** – Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, California 92093, United States; Phone: 1-858-534-9539; ndevaraj@ucsd.edu; Fax: 1-858-246-5891.

Accession Codes

Plasmids for TGT enzyme (#138201) and Control-TAG RNA (#138209) are available on Addgene. RNA sequencing data has been deposited with GEO with accession number GSE149274.

The authors declare no competing financial interest.

Complete contact information is available at: <https://pubs.acs.org/10.1021/acscchembio.0c00445>

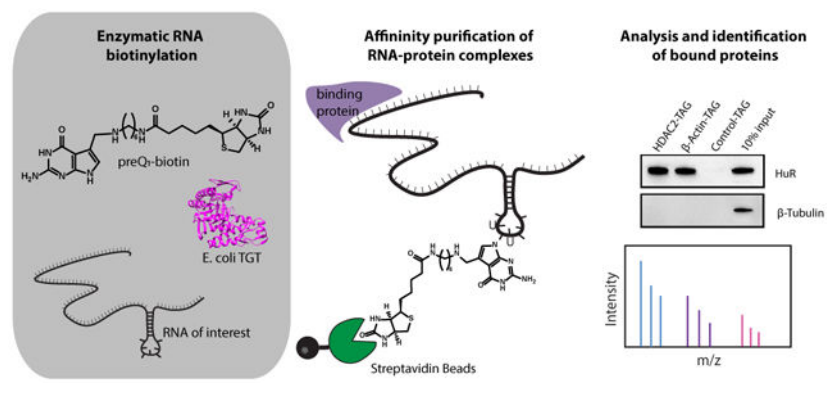
Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acscchembio.0c00445>.

Supplementary tables, figures and additional experimental details; supplementary data containing mass spectrometry hitlists (PDF)

suggest new potential pathways for oncogenic function. Furthermore, using RNA sequencing and qPCR, we establish that an engineered enzyme variant achieves high levels of labeling selectivity against the human transcriptome allowing for 145-fold enrichment of cellular RNA directly from mammalian cell lysates. The flexibility and breadth of this approach suggests that this system could be routinely applied to the functional characterization of RNA, greatly expanding the toolbox available for studying mammalian RNA biology.

## Graphical Abstract



## INTRODUCTION

Proper regulation of RNA metabolism, with elaborate networks of RNAs and their associated RNA-binding proteins (RBPs), underlies cellular homeostasis. Post-transcriptional gene regulation is dictated by the interactions of coding and noncoding RNAs with RNA-binding proteins to ensure the accurate coordination of RNA processing, transport, translation, and degradation. For example, the 3' untranslated regions (3' UTRs) of mRNAs are known to be regulatory hubs for RNA-binding proteins that control mRNA expression level and stability.<sup>1</sup> Furthermore, long noncoding RNAs, through their interactions with proteins, have been shown to be involved in various regulatory processes such as chromatin remodeling and transcriptional activation.<sup>2</sup> These networks of RNA-protein interactions have been implicated in numerous human diseases, ranging from cancer to neurodegeneration.<sup>3,4</sup> Therefore, the characterization of these RNA-protein interactions is crucial to our understanding of RNA biology and the development of innovative RNA-centric therapeutic strategies.

With increasing appreciation for the role of RNA-protein interactions in gene regulation and disease, there is a growing need for methods enabling their identification via protein-centric and RNA-centric approaches. Protein-centric approaches, which allow for identification of RNA transcripts bound to a given protein, have progressed rapidly in the past decade as next-generation sequencing technologies have become more widely available. Techniques such as RNA immunoprecipitation (RIP-seq) and cross-linking and immunoprecipitation (CLIP-seq) are widely used.<sup>5</sup> While conceptually simple, RNA-centric approaches that allow for identification of proteins bound to a given RNA transcript are not as common as protein-centric approaches for two related reasons: First, a typical mRNA is about 3

orders of magnitude less abundant than a typical protein in mammalian cells,<sup>6</sup> and second, mass spectrometry is needed to identify cognate proteins, which, unlike sequencing, is not amenable to signal amplification. Because of these challenges, RNA affinity purification and cognate protein identification remain challenges in RNA biology. Despite these obstacles, RNA-centric studies are essential to studying the function and regulation of targeted RNA transcripts.

Existing noncovalent approaches for RNA-centric interaction mapping include biotinylated antisense oligonucleotides (ASOs) and naturally occurring RNA–protein partners, such as MS2 tagging. Because cellular RNA exhibits intricate secondary and tertiary structure and is typically coated with various RNA binding proteins, it has historically been challenging to design antisense probes that efficiently purify an RNA of interest. The use of RNA hybridizing antisense probes to selectively purify the long-noncoding RNA *Xist* from mammalian cells after cell cross-linking was recently achieved through two similar approaches, RAP-MS and CHIRP-MS.<sup>7–9</sup> To attain adequate selectivity and purification efficiency, both strategies utilized biotinylated antisense probes that tile the length of the 18-kb RNA. While this work represents an important advance in the field, the use of such large libraries of biotinylated antisense probes is very expensive. Antisense oligonucleotides have also been demonstrated using a two-step purification of mRNA transcripts with a preliminary polyA purification, in addition to the analysis of rRNA.<sup>10,11</sup> While these approaches extend the utility of antisense oligonucleotides, rRNA is extremely abundant, and the two-step approach relies on polyA purification, and thus is not applicable to nonpolyadenylated RNAs. Furthermore, optimization of ASO sequence and consideration of RNA base pairing availability is required. In addition to ASO approaches, naturally occurring RNA–protein interactions, such as the interaction between the MS2 coat protein and its cognate RNA hairpin, have been developed as a method for RNA affinity purification.<sup>12–14</sup> To facilitate affinity purification, MS2 cognate hairpins are appended to the RNA of interest, and the MS2 coat protein is expressed as a fusion with another affinity protein. However, limited by the relatively low affinity of the MS2 protein with the target RNA (~3 nM), this approach typically involves the incorporation of 12–24 MS2 hairpins for efficient purification, and the performance of such systems is dependent on the expression level of the RNA of interest.<sup>12,15</sup> Thus, a stronger affinity handle is ideal in studying RNA–protein interactions, especially for less abundant RNA species. For instance, the affinity between biotin and streptavidin is ~1 fM, making biotin an excellent affinity handle to tag an RNA of interest and capture its interactome.<sup>16</sup>

Enzymatic covalent biotinylation of RNA has various potential advantages over noncovalent approaches for affinity purification. First, due to the covalent interaction, it would have the potential to facilitate recovery of less abundant RNA transcripts. Furthermore, by providing a robust covalent linkage, stringent purification conditions of the RNA–protein complexes can be utilized, making such purifications much more accessible. Common approaches for biotinylation of RNA transcripts include *in vitro transcription* with biotinylated precursors or ligation of a biotin to the end of an RNA transcript.<sup>17,18</sup> However, these existing methods for covalent biotinylation of RNA do not target a specific RNA transcript and would therefore not be used for the purification of a target cellular RNA. While enzymatic methods for RNA modification have been developed, many of these methods are not selective for a target

transcript, and none of these methods have been demonstrated to achieve selective covalent labeling of a target RNA against the transcriptome.<sup>19</sup> Few examples exist whereby enzymes can be utilized to have been harnessed to accomplish site-specific covalent modification of RNA.<sup>19–22</sup> In previous approaches, reactive precursors such as primary amines and common metabolites were utilized that are likely not orthogonal to mammalian systems.<sup>23</sup> Additionally, these approaches have required large stretches of RNA sequence, such as an entire tRNA structure, for recognition and modification.

Recently, we have shown that a bacterial RNA modifying enzyme, tRNA guanine transglycosylase (TGT), can be utilized to covalently label an RNA of interest bearing a single short, 17–25 nucleotide encoded hairpin recognition motif.<sup>24,25</sup> On the basis of the minimal recognition motif, direct covalent modification, and potential for selectivity in mammalian systems, we sought to apply this methodology, called RNA Transglycosylation at Guanosine (RNA-TAG) to the study of RNA–protein complexes in mammalian cells (Figure 1).

Herein, we demonstrate the use of RNA-TAG as a novel and effective approach to site-specifically biotinylate and isolate RNAs for the analysis of RNA–protein interactions using immunoblotting and quantitative mass spectrometry. We also demonstrate that native RNA structures can be transformed into enzymatic labeling sites in order to minimize perturbation of the native RNA sequence. Furthermore, while previous imaging studies have suggested that TGT has some degree of selectivity against mammalian transcriptomes,<sup>24,26</sup> we demonstrate that the use of dimerized TGT enables efficient and substantially more selective labeling and affinity purification of an expressed RNA of interest within cell lysates, demonstrating the potential breadth of this approach to include the study of cellular RNA interactions.

## RESULTS AND DISCUSSION

### Detection of RNA–Protein Interactions by Immunoblot.

We first sought to apply RNA-TAG to study the protein binding partners of human mRNA targets. As a proof-of-concept, we selected a fragment of HDAC2 mRNA known to bind the important RNA binding protein HuR,<sup>27</sup> as well as full-length  $\beta$ -actin mRNA, which is also a known binding partner to HuR.<sup>28,29</sup> We designed RNA constructs for each of these sequences, termed HDAC2-TAG and  $\beta$ -actin-TAG, which had an appended 25-nucleotide TGT recognition hairpin at the 3' end surrounded by a short, unstructured linker at either side. Furthermore, we designed an RNA construct, termed Control-TAG, that encoded the TGT recognition hairpin and linker regions but did not have a specific RNA target encoded (Figure 2A,B).

To test whether RNA transcripts biotinylated using RNA-TAG could be utilized for enrichment of cognate RNA binding proteins, we first labeled each transcript through treatment with *E. coli* TGT and a preQ<sub>1</sub>-biotin conjugate. The resulting purified RNA product was then folded and incubated directly with cellular lysates prepared from HeLa cells in order to allow binding of the labeled RNA with cognate RNA binding proteins. Streptavidin-mediated affinity purification was used to enrich the target RNA–protein

complexes, which were then analyzed via immunoblot (Figure 2C). Gratifyingly, we found that HuR was enriched in the HDAC2-TAG and  $\beta$ -actin-TAG samples but not in the Control-TAG sample. Furthermore,  $\beta$ -tubulin, which is not expected to interact with these RNAs, was not enriched in any samples, indicating specificity of the enrichment.

One potential advantage of this approach is the strength of the streptavidin–biotin interaction, potentially allowing for RNA–protein complexes to be detected at extremely low concentrations. In order to examine this, we carried out RNA–protein complexation, affinity purification, and immunoblotting with decreasing amounts of labeled HDAC2-TAG RNA (Figure 2D). The HuR binding partner was detected with as little as 1.5 pg of biotinylated RNA, corresponding to a concentration in the subpicomolar range ( $10^{-13}$  M). For comparison, an average mRNA transcript has approximately 20 copies per cell, corresponding to concentrations in the picomolar range ( $10^{-11}$  M), based on a cell volume of 2.6 pL.<sup>6,30</sup> We also tested whether RNA biotinylation could be carried out at lower concentrations of RNA and found that RNA was labeled with good efficiencies at concentrations down to 1 pM (Supporting Information Figure S1).

### Identification of RNA–Protein Interactions by Quantitative Proteomics.

Encouraged by our immunoblot data, we sought to determine whether this methodology for the purification of enzymatically biotinylated RNA–protein complexes could be applied to the discovery of RNA–protein interactions using quantitative mass spectrometry (Figure 3A). In addition to the model RNA HDAC2-TAG, we used 7SK as a target for proteomics experiments. 7SK is an abundant small nuclear RNA (snRNA) that is approximately 330 nucleotides in length and is involved in transcription regulation. 7SK snRNA has been extensively studied due to its role in the regulation of the positive transcription elongation factor b (P-TEFb).<sup>31</sup> This activity is regulated through its interaction with proteins such as lupus antigen-related protein 7 (LARP7), methylphosphate capping enzyme (MePCE), and HMBA-induced proteins 1 and 2 (HEXIM1/2). Proteins interacting with 7SK snRNA have recently been reported using the CHIRP-MS method,<sup>8,32</sup> which utilizes tiling biotinylated antisense probes for RNA affinity purification; therefore, 7SK is an ideal target for validation of RNA-TAG affinity purification. The protein binding sites of 7SK snRNA have been previously studied, with MePCE and HEXIM1 thought to bind near the 5' end, and LARP7 binding near the 3' end;<sup>31</sup> informed by this work, we encoded a TGT recognition element on the 5' end (7SK-TAG, Figure 3B).

In order to ascertain which proteins were enriched by binding these target RNAs, we carried out quantitative mass spectrometry using a reductive demethylation-peptide labeling strategy.<sup>33</sup> To control for background proteins and nonspecific proteins, three samples were compared for each target RNA, consisting of biotinylated target RNA (HDAC2-TAG or 7SK-TAG), unlabeled target RNA (HDAC2-TAG or 7SK-TAG), and biotinylated Control-TAG RNA. As before, biotinylation of HDAC2-TAG, 7SK-TAG, and Control-TAG was achieved through treatment with preQ<sub>1</sub>-biotin and *E. coli* TGT. Each RNA sample was incubated with cell lysate to allow cognate binding proteins to bind, and complexes were affinity purified, with five biological replicates for each target. In order to verify the biotinylation of the RNA constructs, Northern blots were performed, and transferred

RNA was developed directly using Streptavidin-HRP (Supporting Information Figure S2). Digested and labeled peptides were pooled and analyzed by liquid chromatography–mass spectrometry. Ratios between the experimental samples and the two control samples were calculated and used to identify highly enriched proteins. To generate lists of interacting proteins, the experimental sample (biotinylated target RNA) was compared to both controls (unlabeled target RNA and biotinylated control-TAG RNA). Proteins detected in at least three replicates with an average enrichment greater than 3-fold against both controls were considered putative binding proteins.

As we expected, HuR was highly enriched (>49 fold) in the HDAC2-TAG affinity purification, compared to the Control-TAG affinity purification. Thirteen additional proteins were enriched by HDAC2-TAG, which can be categorized as proteins involved in mRNA stabilization and splicing (Supporting Information Data). In the 7SK-TAG affinity purification, we observed enrichment of canonical 7SK binding partners, including HEXIM1, CDK9, DDX21, MEPCE, and LARP7 (Figure 3C). Other known binding partners, such as members of the BAF complex, were also identified.<sup>32</sup> These results establish that our RNA-TAG approach can successfully identify RNA-binding proteins that are associated with specific RNAs.

#### **Application of RNA-TAG to Target HOTAIR through Minimal Mutations.**

Use of RNA-TAG requires that a TGT recognition element, typically 17–25 nucleotides, be encoded into an RNA of interest in order to facilitate its labeling.<sup>25</sup> Previous work has demonstrated that the minimal requirements for TGT recognition include a 7-nucleotide loop with the consensus sequence YUGUNNN, where Y represents a pyrimidine (C or U) and N represents any base.<sup>34,35</sup> Furthermore, it has been hypothesized that recognition of an RNA substrate by TGT through charge complementarity requires adoption of a zigzag conformation of the loop region.<sup>36</sup> Because many RNAs, especially long noncoding RNAs, are highly structured,<sup>37</sup> we next sought to explore whether existing stem loop structures within an RNA of interest could be leveraged to enable RNA-TAG labeling. HOTAIR, a 2148 nt lncRNA, is overexpressed in several cancer tissues and has been implicated in various oncogenic processes including cell invasion, tumor development, and metastasis.<sup>38</sup>

Structural studies of HOTAIR have found it to be highly structured, with over 50% of the nucleotides involved in base-pairing interactions; furthermore, 56 helical segments were identified, with 38 terminal loops.<sup>39</sup> We hypothesized that perhaps some of these existing stem-loops could be mutated in order to transform them into potential substrates for TGT. Therefore, we selected four candidate stem loops within HOTAIR that had a loop region of seven nucleotides, H12, H22, H23, and H51 (Figure 4A). The loop regions of each of these stem loops were mutated to match the loop region of the TGT recognition element. These mutated HOTAIR transcripts were subjected to TGT labeling conditions, alongside a wild-type HOTAIR transcript, and control transcripts bearing a full TGT recognition element. Detection of biotinylation via Northern blot showed that mutation of hairpin H22 (HOTAIR-H22-TAG) enabled labeling efficiencies similar to 5′TAG-HOTAIR, a transcript encoding a full 25-nucleotide TGT recognition element (Figure 4B, Supporting Information Figure S3). This result was confirmed using labeling with preQ<sub>1</sub>-BODIPY and gel analysis

(Supporting Information Figure S4). The preferential recognition of the modified hairpin 22 compared to the other modified hairpins H12, H23, and H51 may be attributed to the increased length of the stem, contributing to greater stability and shape recognition of the stem-loop motif. Importantly, the wild type HOTAIR transcript showed very low levels of detectable biotinylation.

Despite extensive studies on the biological function of HOTAIR, its role in chromatin remodeling is still unclear and is a subject of debate.<sup>40–44</sup> Therefore, we sought to employ our modified HOTAIR construct, along with the quantitative proteomics approach described above, in order to identify putative HOTAIR-binding proteins using MCF7-cells, a breast-cancer cell line known to express HOTAIR at high levels.<sup>45,46</sup> Biotinylated HOTAIR-H22-TAG and 5'TAG-HOTAIR were compared to unlabeled controls, as well as a biotinylated transcript encoding a 5'TAG sequence and the antisense sequence corresponding to HOTAIR (5'TAG-antisense-HOTAIR), across five biological replicates. Because the protein binding sites of HOTAIR are not well understood, analysis of two HOTAIR constructs with different points of attachment (HOTAIR-H22-TAG and 5'TAG-HOTAIR) was pursued. Biotinylation of the RNA constructs was verified by Northern blot (Supporting Information Figure S5). To generate lists of putative interacting proteins, the experimental sample (biotinylated target RNA) was compared to both controls (unlabeled target RNA and biotinylated 5'TAG-antisense-HOTAIR RNA). Proteins detected in at least three replicates with an average enrichment greater than 3-fold against both controls were considered putative binding proteins.

Comparison of the samples enriched by the differentially labeled transcripts, 5'TAG-HOTAIR and HOTAIR-H22-TAG, identified 55 proteins that were enriched by both constructs, independent of the site of attachment. We observed a lower number of enriched proteins with the HOTAIR-H22-TAG RNA compared to the 5'TAG-HOTAIR, which could be attributed to the less perturbing RNA modification, differences in RNA–protein purification efficiency, or disruption of the RNA structure. Several proteins involved in RNA processing, ribonucleoprotein biogenesis, and ribosome biogenesis were identified as putative binding proteins for both HOTAIR-TAG RNAs (Figure 4C,D, Supporting Information Data).<sup>47</sup> Interestingly, some of these proteins have been previously reported to have RNA-dependent functions in cancer. The long-noncoding RNA GSEC was previously shown to inhibit DHX36 through its g-quadruplexes, leading to cancer cell migration;<sup>48</sup> importantly, an alternative g-quadruplex structure of HOTAIR has been reported.<sup>49</sup> Furthermore, the oncogenic lncRNA MALAT1 has been proposed to influence alternative splicing through interactions with serine-arginine (SR) proteins, such as SRSF2.<sup>50,51</sup> It has also been proposed that the long-noncoding RNA SLERT plays a role in the dysregulation of ribosome biogenesis in cancer, which may indicate a potential pathway for HOTAIR's oncogenic function.<sup>52</sup> While further studies are needed to validate and characterize these binding partners, these findings suggest that HOTAIR may play alternative roles in oncogenesis that go beyond regulation of chromatin remodeling complexes.

## Enrichment of Cellular RNA with RNA-TAG.

Having demonstrated the successful purification of various RNA–protein complexes using RNA-TAG labeling, we sought to determine, as a proof of principle, whether the RNA-TAG labeling approach was suitable for the purification of cellular RNA directly from cell lysates. In principle, this could allow for the study of cellular RNA–protein or RNA–RNA interactions. Using viral transduction, we prepared HeLa cells stably expressing the HDAC2-TAG RNA transcript. We reasoned that the addition of TGT enzyme and preQ<sub>1</sub>-biotin directly to the corresponding cell lysate may enable selective biotinylation of the expressed transcript, which would allow for subsequent affinity purification with streptavidin-coupled beads. In this context, the selectivity of the TGT enzyme to label the targeted RNA over other RNAs present in the cell would be essential. It has been hypothesized that bacterial TGT forms a homodimer, where the first unit is catalytic and the second unit plays a role in recognition and proper orientation of the bound RNA substrate.<sup>53</sup> Therefore, we postulated that a stable TGT dimer may have improved labeling selectivity and prepared an obligate TGT dimer, connected by a 16 amino acid XTEN linker, to test this hypothesis.<sup>54</sup>

We first sought to determine whether the expressed RNA could be efficiently and selectively labeled within a cellular lysate. TGT enzyme variants and preQ<sub>1</sub>-biotin were added to cell lysates from HeLa cells with stable HDAC2-TAG RNA expression, allowed to react, and total RNA was purified from the cell lysate to allow for RNA quantification. Streptavidin-mediated affinity purification was performed, and on-bead reverse transcription, followed by qPCR, was utilized to quantify affinity-purified, biotinylated RNA transcripts. As a control, the RNA input to the streptavidin affinity purification was also analyzed by RT-qPCR, so that the fraction of RNA recovered from a sample could be calculated (Figure 5A). From this analysis, we were gratified to find that 75% of the target HDAC2-TAG transcript was recovered using the *E. coli* TGT enzyme, and 44% was recovered using the TGT obligate dimer (Figure 5B). For comparison, an antisense oligonucleotide approach demonstrated ~70% recovery of the long noncoding RNA XIST using RT-qPCR.<sup>7</sup> Other RNA transcripts were also quantified by qPCR in order to test the selectivity of this method. Two RNAs, GAPDH mRNA and LBR mRNA, were selected due to their sequences, which contain a potential UGU-bearing stem-loop structure (Supporting Information Table 1), with GAPDH also being a very abundant mRNA. Other abundant RNAs, including  $\beta$ -Actin mRNA, U1-snRNA, and 18s-rRNA, were also tested. From this analysis, significant selectivity was observed against these RNAs using *E. coli* TGT, with GAPDH having a 1% recovery and other RNAs having recoveries of less than 0.2% (Figure 5B, Supporting Information Tables 2 and 3). An improvement in selectivity was observed in samples labeled with the obligate TGT dimer, specifically with the GAPDH transcript (0.02% recovery).

We then sought to further examine the selectivity of the monomeric and dimeric forms of TGT in the context of the human transcriptome through the preparation of RNA sequencing libraries. To accomplish this, we carried out lysate labeling, followed by streptavidin-mediated affinity purification of the extracted total RNA, and library preparation, using three replicates for each condition. Approximately 4–5 million reads were collected for each replicate, with reads aligning to greater than 45,000 RNA transcripts. Analysis of



these libraries showed that the samples prepared with dimeric TGT had much more selective enrichment of the desired HDAC2-TAG-Puro transcript (Figure 5C,D, Supporting Information Figure S6). Furthermore, in each of the three replicates prepared with the obligate TGT dimer, the HDAC2-TAG-Puro transcript had the highest number of aligned reads compared to any other transcript, with an average of 10.95% of all aligned reads mapping to the HDAC2 construct. For comparison, RNA sequencing was previously reported with the alternative RAP (RNA Antisense Purification) technique, with 68% of reads aligned to the targeted Xist transcript.<sup>55</sup> Because the number of reads is dependent on transcript length, and Xist (~17kb) is more than 8 times longer than the HDAC2-TAG-Puro transcript (~2kb) studied here, these data are consistent with a similar degree of RNA selectivity among these techniques. On the basis of the observed improvement in selectivity for the obligate TGT dimer, we sought to further explore the selectivity of these two enzyme variants by further analysis of potential off-targets observed in the enriched sequencing libraries (FTH1, RPS6, RPL41, MRPL51, MALAT1). We examined whether these transcripts were enriched due to labeling by either monomeric or dimeric TGT through qPCR, as described above. This confirmed that the *E. coli* TGT enzyme displayed off-target labeling of several of these transcripts (5–21%), whereas the obligate TGT dimer off-target labeling was substantially reduced (Figure 5E, Supporting Information Tables 4 and 5). This supports the hypothesis that the TGT homodimer aids in RNA recognition.

On the basis of the highly selective and efficient labeling of RNA with dimeric TGT, we hypothesized that this technique could be utilized to affinity purify an RNA of interest directly from cell lysates in the presence of remaining preQ<sub>1</sub> biotin probe, which could potentially be applied to the affinity purification of cellular RNA–protein complexes. Treatment of cell lysates with enzymatic labeling conditions, followed directly with affinity purification and stringent washes, allowed for direct enrichment of HDAC2-TAG RNA, approximately 130-fold, compared to a sample lacking the TGT enzyme. The obligate dimer showed similar levels of enrichment, approximately 145-fold, with significantly lower enrichment of the nontarget GAPDH mRNA (Figure 6, Supporting Information Tables 6 and 7). These data suggest that RNA labeling in cellular lysates using the obligate TGT dimer allows for selective, stringent purification of the expressed RNA and could potentially be extended to the purification of cellular RNA–protein complexes directly from cell lysates on a larger scale to enable proteomic analysis.

## CONCLUSION

RNA–protein interactions regulate a broad range of critical cellular processes; however, identifying the proteins that interact with an RNA of interest in mammalian cells represents a major challenge in biology. While RNA-centric interaction mapping has been increasingly employed to elucidate RNA function, the noncovalent nature of existing approaches limits their utility to highly expressed RNA transcripts and represents a barrier to entry to many researchers due to the specialized and complex techniques employed. Direct covalent modification of an RNA of interest with biotin has many potential advantages for functional characterization. Due to the strength of the streptavidin–biotin interaction, RNA can be affinity purified when present at low concentrations and using stringent wash conditions. In this work, we present RNA-TAG as a novel methodology for the purification of RNA–

protein complexes. By encoding a 25-nucleotide hairpin into mRNA constructs, analysis of RNA–protein complexes using immunoblotting was demonstrated using very low RNA concentrations. Furthermore, by combining the RNA-TAG methodology with quantitative proteomics, the identification of several known binding partners of the model 7SK snRNA was achieved.

In many cases, the addition of exogenous sequence elements to enable RNA labeling and isolation may be disadvantageous. In order to overcome this challenge, we mutated HOTAIR, a highly structured long noncoding RNA, to transform naturally occurring RNA hairpins into substrates for the TGT enzyme. Using this approach, the labeling of HOTAIR through the mutation of four nucleobases was utilized to enable proteomic analysis. This identified novel putative binding partners involved in pathways such as RNA processing, ribonucleoprotein biogenesis, RNA splicing, and ribosome biogenesis that may play a role in HOTAIR's oncogenic functions.

Last, the ability of the RNA-TAG system to selectively label RNA expressed in mammalian cells was explored. Unlike other techniques for RNA enzymatic labeling, which do not enable targeted biotinylation of a specific transcript in a complex mixture,<sup>19</sup> we demonstrated that RNA can be labeled selectively with good efficiency in cellular lysates, potentially allowing for this technique to be applied on a larger scale to isolate cellular RNA–protein complexes. While the bacterially expressed *E. coli* TGT was found to be fairly promiscuous, engineering of a TGT dimer significantly improved RNA labeling selectivity, as demonstrated through RNA sequencing and RT-qPCR. Bacterial TGTs are thought to form a homodimer in cells, where one subunit is catalytic and the other subunit plays a role in binding and orienting the RNA substrate properly for catalysis;<sup>53</sup> this observation of improved selectivity through the use of dimeric TGT supports this hypothesis. Using the dimeric TGT, the expressed RNA was labeled in cell lysates with good efficiency (44%), with selectivity similar to competitive affinity purification methods. For comparison, the use of tiling antisense oligonucleotides to purify the cellular lncRNA XIST enabled about ~70% RNA recovery<sup>7</sup> and similar selectivity by RNA sequencing,<sup>55</sup> as discussed above. This approach is unique in its ability to efficiently and selectively biotinylate an RNA of interest in the presence of the human transcriptome, thus expanding the potential to affinity purify low abundance RNA transcripts. Because this technique relies on covalent biotinylation of an RNA of interest, we also anticipate that this technique would be compatible with UV RNA–protein cross-linking and stringent purification approaches that are commonly employed for the study of RNA–protein interactions.

Together, this report demonstrates a simple and highly flexible platform for the analysis of RNA–protein interactions through direct enzymatic biotinylation of a target RNA. This approach has several advantages, including the selectivity of RNA labeling against the human transcriptome, the ability to reduce the perturbation of the RNA recognition element through the mutation of existing structural elements, and the use of covalent biotinylation to achieve a strong and stringent affinity purification. The simplicity and robustness of the RNA-TAG approach creates opportunities for widespread utilization in studying RNA biology; we anticipate that this technology will aid in providing new insights into the functional characterization of RNA and its broad range of roles in cell biology.

## METHODS

### Synthesis of preQ<sub>1</sub>-biotin.

PreQ<sub>1</sub>--C<sub>6</sub>H<sub>12</sub>--NH<sub>2</sub> (2-amino-5-(((6-aminohexyl)amino)methyl)-3,7-dihydro-4H-pyrrolo[2,3-*d*]-pyrimidin-4-one), prepared as previously described<sup>56</sup> (2.5 mg, 9.0 μmol), was dissolved in 200 μL of anhydrous DMF, followed by the addition of anhydrous diisopropylethylamine (4.6 μL, 27 μmol). A solution of biotin-NHS ester (3.2 mg, 9.3 μmol) in 200 μL of DMF was then slowly added to the mixture. The reaction was stirred for 2 h at RT. The reaction was concentrated, and the residue was purified by HPLC to give a white solid (2.2 mg, 49% yield).

### *In Vitro* Transcription.

All plasmids used for *in vitro* transcription were prepared using standard cloning techniques. The plasmid encoding Control-TAG is available on Addgene (Addgene #138209, pcDNA3.1-(empty)-TAG). This vector backbone was used to prepare HDAC2-TAG, β-actin-TAG, and 7SK-TAG. The plasmid encoding HOTAIR was provided by Anna Pyle.<sup>39</sup> T7 RNA polymerase was expressed and purified as previously reported.<sup>24</sup> Templates were linearized via restriction digest for Control-TAG (XbaI), HDAC2-TAG(XbaI), β-actin-TAG (XbaI), and HOTAIR constructs (*SaI*). PCR amplification was used to linearize 7SK-TAG and append a T7 promoter with the forward primer AAGCTGTAATACGACTCACTATAGGGGATCCCCGGGAGCAGAC and the reverse primer AAAAGAAAGGCAGACTGCCACATG. Transcription reactions were set up with RNA NTPs (5 mM each ATP, CTP, UTP, 9 mM GTP; NEB, Ipswich, MA), 0.004 U/μL Thermostable Inorganic Pyrophosphatase (NEB, Ipswich, MA), 0.15 μg/μL T7-RNAP, and 0.05% Triton X-100 (Sigma, St. Louis, MO) in T7 reaction buffer. Each transcription reaction was set up with approximately 4 μg of cut plasmid or 1 μg of purified PCR product as a template in a 100 μL transcription reaction. The transcription reaction was run at 37 °C for 3–4 h. RNA was then purified via lithium chloride precipitation by the addition of LiCl Precipitation Solution (Invitrogen) to reach a final concentration of 2.5 M LiCl, followed by incubation at –20 °C for 1 h or overnight. The RNA was quantified at 260 nm, confirmed as a single observable UV shadowing band by 4% denaturing PAGE (4% polyacrylamide in TBE with 8 M urea) and kept frozen at –20 °C until used.

### TGT Labeling of RNA Transcripts.

*E. coli* tRNA Guanine transglycosylase (TGT) was expressed and purified as previously described.<sup>24</sup> TGT labeling reactions were carried out in 1× TGT reaction buffer with 5 mM DTT, 1 μM RNA transcript, 1 μM *E. coli* TGT, 10 μM preQ<sub>1</sub>-biotin, and 1 U/μL Murine RNase Inhibitor (NEB). Labeling reactions were incubated for 2 h at 37 °C and purified via ethanol precipitation.

### Northern Blot Biotinylation Assay.

Northern blot biotinylation assays were carried out with 2 pmol of RNA for HDAC2-TAG, 7SK-TAG, and Control-TAG transcripts and 0.1 pmol of RNA for all HOTAIR constructs. RNA was electrophoretically separated by 4% denaturing PAGE (4% polyacrylamide in

TBE with 8 M urea), stained with SYBR green II (Invitrogen), and imaged. Subsequently, RNA was transferred via electroblot to a positively charged nylon membrane (Invitrogen) in 0.5× TBE. Blot was developed for biotin detection using the Chemiluminescent Nucleic Acid Detection Module Kit (Thermo Scientific)

### Cell Culture.

Human HeLa-S3 cells were grown in complete DMEM media (Gibco) supplemented with 10% fetal bovine serum and 1% penicillin/streptomycin. Human MCF7 HTB-22 cells were grown in Eagle's MEM (Gibco) supplemented with 10% fetal bovine serum, 1 mM pyruvate, 0.01 mg mL<sup>-1</sup> human recombinant insulin, and 1% penicillin/streptomycin. Cells were cultured at 37 °C in a humidified incubator under 5% CO<sub>2</sub>.

### RNA-Protein Analysis by Immunoblot.

Biotinylated RNA was folded in 1× RNA folding buffer (10 mM Tris pH 7, 0.1 M KCl, 10 mM MgCl<sub>2</sub>) by heating for 2 min at 90 °C, followed by 20 min at 25 °C. HeLa cell pellets (~3 × 10<sup>6</sup>) were lysed with 500 μL of Mammalian Protein Extraction Reagent (Thermo Scientific), supplemented with 2× HALT protease inhibitor (Thermo Scientific), according to manufacturer's instructions. Protein concentration was quantified using a Pierce BCA Protein Assay Kit (Thermo Scientific). Folded RNA (1 pmol, unless otherwise noted) was incubated with 40 μg of HeLa cell lysate in 1× Immunoblot Binding and Washing Buffer (1× IBW, 25 mM Tris-HCl pH 7.5, 150 mM KCl, 2 mM MgCl<sub>2</sub>, 1 mM DTT, 0.25% Nonidet-P40, 1% Tween 20), supplemented with 1 U/μL Murine RNase Inhibitor (NEB) for 1 h at RT to allow binding of cognate proteins. Ten microliters of washed Dynabeads M-280 Streptavidin (Invitrogen) was added to each binding reaction and further incubated at RT for 1 h. Beads were washed 3 times in 1× IBW Buffer, followed by 2 washes with PBS. Protein was eluted from beads by heating for 5 min at 95 °C in 1× Laemmli Buffer (Bio-Rad). Retrieved protein was detected by standard immunoblotting techniques using the following antibodies: mouse anti-HuR (3A2, Santa Cruz Biotechnology) and mouse anti-β-tubulin (D-10, Santa Cruz Biotechnology).

### RNA-Protein Enrichment, Trypsinization, and Stable Isotope Labeling for Mass Spectrometry.

HeLa cell pellets (~2 × 10<sup>7</sup>) or MCF7 cell pellets (~4 × 10<sup>7</sup>) were lysed with 3 mL of Mammalian Protein Extraction Reagent (Thermo Scientific), supplemented with 2× HALT protease inhibitor (Thermo Scientific), according to the manufacturer's instructions. Protein concentration was quantified using Pierce BCA Protein Assay Kit (Thermo Scientific). Biotinylated RNAs were labeled as described above, while unlabeled RNA controls were used directly. RNA was folded in 1× RNA folding buffer by heating for 2 min at 90 °C, followed by 20 min at 25 °C. For HDAC2-TAG and 7SK-TAG experiments, each folded RNA (5 pmol) was incubated with 200 μg of HeLa cell lysate in 1× Proteomics Binding and Washing Buffer (1× PBW, 20 mM Potassium Phosphate pH 7.5, 150 mM KCl, 2 mM MgCl<sub>2</sub>, 1 mM DTT, 0.25% Nonidet P40, 1% Tween 20), supplemented with 1 U/μL Murine RNase Inhibitor (NEB, Ipswich, MA) for 1 h at RT to allow binding of cognate proteins. For HOTAIR experiments, each folded RNA (10 pmol) was incubated with 400 μg of MCF7 cell lysate in 1× PBW, supplemented with 1 U/μL Murine RNase Inhibitor (NEB) for 1 h

at RT to allow binding of cognate proteins. Twenty-five microliters of washed Dynabeads M-280 Streptavidin (Invitrogen) was added to each binding reaction and further incubated at RT for 1 h. Beads were washed three times in 1× PBW buffer, followed by five washes with PBS. Beads were then resuspended in 50  $\mu\text{L}$  of trypsinization solution consisting of 10 ng/ $\mu\text{L}$  Trypsin (Promega) in 100 mM tetramethylammonium bicarbonate (TEAB), at pH 8.0. Trypsinization was carried out overnight at 37 °C with sample rotation. The supernatant containing trypsinized peptides was then transferred to a new tube and vacuum-dried. Stable isotope dimethyl labeling was carried out according to previously published protocols.<sup>33</sup> The samples were labeled as indicated in the Supplementary Data.

### Mass Spectrometry.

The labeled, vacuum-dried peptides were resuspended in 5% formic acid/5% acetonitrile buffer and added to the vials for mass spectrometry analysis. Samples were analyzed with triplicate injections by LC-MS-MS using EASY-nLC 1000 liquid chromatography connected with Q-Exactive mass spectrometer (Thermo Scientific) as described previously<sup>57</sup> with some modification as follows. The peptides were eluted using the 60 min acetonitrile gradient (45 min 2%–30% ACN gradient followed by 5 min 30%–60% ACN gradient, a 2 min 60–95% ACN gradient, and a final 8 min isocratic column equilibration step at 0% ACN) at a 250 nL/min flow rate. All the gradient mobile phases contained 0.1% formic acid. The data dependent analysis (DDA) was done using the top 10 method with a positive polarity, scan range of 400–1800  $m/z$ , a 70,000 resolution, and an AGC target of 3e6. A dynamic exclusion time of 20 s was implemented and unassigned; singly charged and charge states above 6 were excluded for the data dependent MS/MS scans. The MS2 scans were triggered with a minimum AGC target threshold of 1e5 and with a maximum injection time of 60 ms. The peptides were fragmented using a normalized collision energy (NCE) setting of 25. Apex trigger and peptide match settings were disabled.

### Mass Spectrometry Analysis.

RAW files were processed, searched, and analyzed essentially as described previously.<sup>58</sup> Converted mzXML files were searched using SEQUEST (version 28) using a target-decoy database containing reviewed UniProtKB/Swiss-Prot Human protein sequences and common contaminants. Each mzXML file was searched in triplicate with the following parameters: 50 ppm precursor ion tolerance and 0.01-Da fragment ion tolerance; trypsin (1 1 KR P) was set as the enzyme; up to three missed cleavages were allowed; a dynamic modification of 15.99491 Da on methionine (oxidation); and a static modification of 57.02146 Da on cysteine for iodoacetamide alkylation. For searches with light and medium reductive dimethyl labels, additional dynamic modifications of 4.0224 Da on lysine and peptide N termini and static modifications of 28.0313 Da on lysine and peptide N termini were included. For searches with light and heavy reductive dimethyl labels, additional dynamic modifications of 8.04437 Da on lysine and peptide N termini and static modifications of 28.0313 Da on lysine and peptide N termini were included. For searches with medium and heavy reductive dimethyl labels, additional dynamic modifications of 4.02193 Da on lysine and peptide N termini and static modifications of 32.05374 Da on lysine and peptide N termini were included. Peptide matches were filtered to a peptide false discovery rate of 2% using the linear discriminant analysis. Proteins were further filtered to a

false discovery rate of 2. Peptides were assembled into proteins using maximum parsimony, and only unique and razor peptides were retained for subsequent analysis. All peptide heavy/light, medium/light, and heavy/medium ratios with a signal-to-noise ratio of above 5 were used for assembled protein quantitative ratios.

### Generation of Cell Lines Stably Expressing HDAC2-TAG.

Stable cell lines expressing HDAC2-TAG were generated using lentiviral transduction and subsequent puromycin selection as follows. 293T cells were transfected with the following helper plasmids pHAGE - GAG-POL; pHAGE - VSVG; pHAGE - tat1b; pHAGE - rev; and pHAGE-HDAC2-TAG-IRES-Puro using Mirus TransIT 293 transfection reagent. After 24 h, fresh medium was added to the cells. At 48 h post transfection, the medium containing the infectious virions was collected and filtered using a 0.45 mm sterile syringe filter. Polybrene (6 $\mu$ g/mL) was added to the filtered virus-containing medium. The mixture was added to HeLa cells seeded at 30% confluency and infected for 24 h, prior to selection with 500 ng/mL of Puromycin to obtain stable expression clones.

### Expression and Purification of Obligate Dimeric TGT.

A plasmid encoding obligate dimeric TGT was cloned from the TGT-His plasmid (Addgene #138201) using a DNA HiFi Assembly (New England Biolabs) to add a 16-amino acid XTEN linker (SGSETPGT-SESATPES) between two identical coding sequences for *E. coli* TGT. The enzyme was then expressed in BL21-DE3 cells with pG-KJE8 chaperone plasmid (Takara Bio) in media supplemented with 0.5 mg mL<sup>-1</sup> arabinose and 5 ng/mL doxycycline to induce chaperone expression and 1 mM IPTG to induce enzyme expression. Expression was allowed to proceed overnight at 18 °C, and dimeric TGT was then purified using standard His-tag purification, as described previously.<sup>24</sup>

### RNA Labeling in Cell Lysate.

Stable HeLa cells expressing HDAC2-TAG, grown to ~80% confluency in a 10 cm plate, were lysed directly upon the addition of 600  $\mu$ L of Mammalian Protein Extraction Reagent (Thermo Scientific), supplemented with 2 $\times$  HALT protease inhibitor (Thermo Scientific) and 1 U/ $\mu$ L Murine RNase, according to manufacturer's instructions. To this cell lysate, 1  $\mu$ M preQ<sub>1</sub>-biotin, 100 nM *E. coli* TGT or 500 nM obligate dimeric TGT (as noted), and 1 U/ $\mu$ L of murine RNase inhibitor were added. Reactions were incubated for 2 h at RT prior to analysis.

### qPCR Analysis of RNA Recovery.

Following RNA labeling in cell lysate, as described above, total RNA was extracted from the lysate upon the addition of three volumes of Trizol-LS (Invitrogen) and purification with the Zymo Direct-Zol kit (Zymo Scientific) according to manufacturer's instructions. Ten microliters of washed Dynabeads MyOne C1 Streptavidin (Invitrogen) was incubated with 200 ng of isolated total RNA sample in 1 $\times$  RNA binding and washing buffer (1 $\times$  RBW, 5 mM Tris-HCl, pH 7.5, 0.5 mM EDTA, 1 M NaCl, 0.05% Tween) supplemented with 2 U/ $\mu$ L of murine RNase inhibitor for 1 h at RT. Beads were washed three times in 1 $\times$  RBW buffer, followed by two washes with ultrapure water. On-bead reverse transcription was performed

in a volume of 20  $\mu\text{L}$  using 200 U of Maxima Reverse Transcriptase (Thermo Scientific) according to the product instructions. For input controls, 200 ng of the corresponding total RNA sample was used. Following reverse transcription, cDNA was diluted 6-fold, and qPCR was performed using 4  $\mu\text{L}$  of a diluted cDNA template in a 20  $\mu\text{L}$  reaction volume using qPCRBIO SyGreen Blue Mix (PCR Biosystems). Quantitative PCR was performed on a CFX-96 Real-Time System with a C100 Thermocycler (Bio-Rad). RNA recovery values were determined by comparison of the input and recovered RNA  $C_T$  values by computing the  $C_T$  (purified-input) according to the following equation:  $\text{RNA recovery} = 2^{-\Delta C_T}$ , based on the assumptions of the  $C_T$  method.<sup>59</sup> Primer sequences are available in the Supporting Information.

### Enrichment of Labeled RNA from Cell Lysate.

Following RNA labeling in cell lysate, as described above, 25  $\mu\text{L}$  of washed Dynabeads MyOne C1 Streptavidin (Invitrogen) was added directly to the lysate labeling reaction and allowed to incubate for 1 h at RT. Beads were washed three times in 1 $\times$  Stringent Wash Buffer I (20 mM Tris-HCl, pH 8, 2 mM EDTA, 150 mM NaCl, 1% Triton-X, 0.1% SDS), followed by three washes with 1 $\times$  Stringent Wash Buffer II (20 mM Tris-HCl, pH 8, 2 mM EDTA, 500 mM NaCl, 1% Triton-X, 0.1% SDS), and finally two washes with ultrapure water. On-bead reverse transcription was performed in a volume of 20  $\mu\text{L}$  using 200 U of Maxima Reverse Transcriptase (Thermo Scientific) and a cocktail of 15 pmol each of gene-specific RT primers, according to the product instructions. Following on-bead reverse transcription, we performed qPCR and determined the  $C_T$  (gene-reference) for each condition, using  $\beta$ -actin as a reference. We then calculated the  $C_T$  between the enzyme treated sample and control sample lacking enzyme treatment and then calculated enrichment according to the following equation:  $\text{enrichment} = 2^{-\Delta\Delta C_T}$ , based on the assumptions of the  $C_T$  method.<sup>59</sup>

### RNA Sequencing.

In order to identify transcripts biotinylated using this method, RNA was biotinylated in cell lysate, as described above. Total RNA was extracted from the lysate upon the addition of three volumes of Trizol-LS (Invitrogen) and purification with the Zymo Direct-Zol kit (Zymo Scientific) according to the manufacturer's instructions. Library preparation was conducted using a TruSeq Stranded mRNA kit (Illumina), using a modified protocol. In lieu of polyA purification, 25  $\mu\text{L}$  of washed Dynabeads MyOne C1 Streptavidin (Invitrogen) was incubated with 10  $\mu\text{g}$  of isolated total RNA sample in 1 $\times$  RBW buffer supplemented with 2 U/ $\mu\text{L}$  of murine RNase inhibitor for 1 h at RT. Beads were washed three times in 1 $\times$  RBW buffer, followed by two washes with ultrapure water. Subsequent to RNA binding and washing the streptavidin beads, 19.5  $\mu\text{L}$  of Fragment-Prime-Finish mix was added to each bead sample to elute and fragment the RNA. Library preparation was carried out in subsequent steps according to the manufacturer's protocol. All libraries were then sequenced using a MiSeq (Illumina) to produce paired end 75-bp reads.

## RNA Sequencing Data Analysis.

The raw RNA-seq data were first separated in paired fragment mates. Sequence AGATCGGAAGAGCACACGTCTGAACTCCAGCA or AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT was used to cut potential sequencing adaptor reads from raw RNA-seq read\_1 or read\_2, respectively. Sequencing reads were aligned using Bowtie2 in end-to-end mapping mode with an index built from the human hg38 assembly and the transgene sequence. Salmon *et al.*<sup>60</sup> was used to map RNA-seq reads to human transcriptome and quantify transcript expression.

*DESeq2* (R package via Bioconductor)<sup>61</sup> was used to perform the differential expression analysis. *tximport* (R package via Bioconductor) was used to import and summarize Salmon RNA-seq quantification results for downstream *DESeq2* analysis.<sup>62</sup> HDAC-TAG-Puro gene was manually added to the index to map each transcript to its corresponding gene. Scatter plot of log<sub>2</sub> fold change and heatmap are generated using built-in *DESeq2* functions.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## ACKNOWLEDGMENTS

The authors gratefully acknowledge A. Pyle for providing the plasmid for transcription of HOTAIR and UCSD Molecular Mass Spectrometry Facility for small molecule high resolution mass spectrometry support. This work was supported by the National Institutes of Health [R01 GM123285 to N.D. and E.B. and T32 GM112584-03 to K.B.].

## REFERENCES

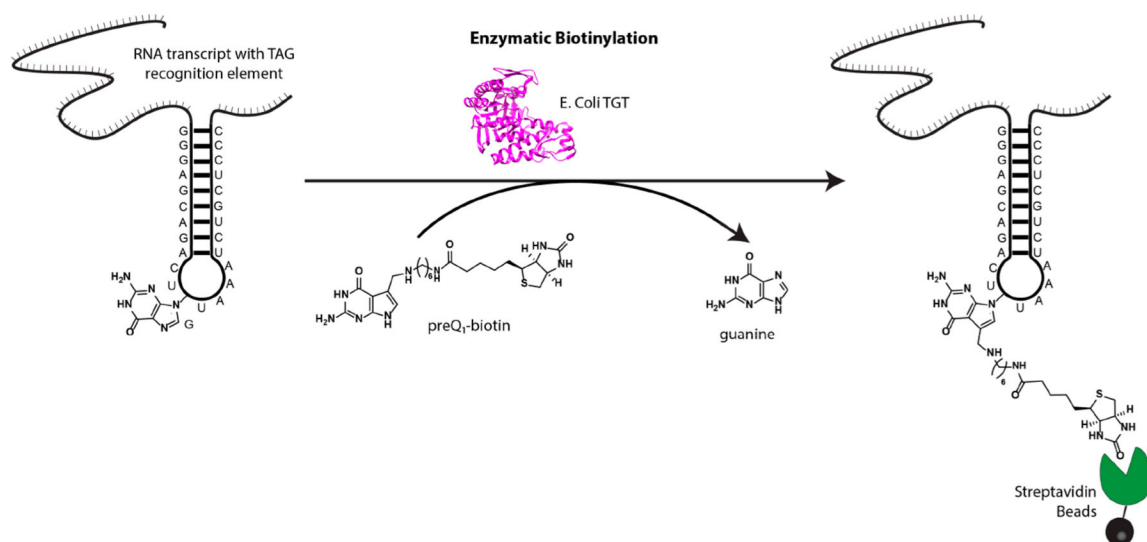
- (1). Singh G, Pratt G, Yeo GW, and Moore MJ (2015) The Clothes Make the MRNA: Past and Present Trends in MRNP Fashion. *Annu. Rev. Biochem* 84 (1), 325–354. [PubMed: 25784054]
- (2). Morris KV, and Mattick JS (2014) The Rise of Regulatory RNA. *Nat. Rev. Genet* 15 (6), 423–437. [PubMed: 24776770]
- (3). Lukong KE, Chang K, Khandjian EW, and Richard S (2008) RNA-Binding Proteins in Human Genetic Disease. *Trends Genet.* 24 (8), 416–425. [PubMed: 18597886]
- (4). Cooper T. a., Wan L, and Dreyfuss G (2009) RNA and Disease. *Cell* 136 (4), 777–793. [PubMed: 19239895]
- (5). McHugh C. a, Russell P, and Guttman M (2014) Methods for Comprehensive Experimental Identification of RNA-Protein Interactions. *Genome Biol.* 15 (1), 203. [PubMed: 24467948]
- (6). Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, and Selbach M (2011) Global Quantification of Mammalian Gene Expression Control. *Nature* 473 (7347), 337–342. [PubMed: 21593866]
- (7). McHugh C. a., Chen C-K, Chow A, Surka CF, Tran C, McDonel P, Pandya-Jones A, Blanco M, Burghard C, Moradian A, Sweredoski MJ, Shishkin A. a., Su J, Lander ES, Hess S, Plath K, and Guttman M (2015) The Xist LncRNA Interacts Directly with SHARP to Silence Transcription through HDAC3. *Nature* 521, 232. [PubMed: 25915022]
- (8). Chu C, Zhang QC, da Rocha ST, Flynn RA, Bharadwaj M, Calabrese JM, Magnuson T, Heard E, and Chang HY (2015) Systematic Discovery of Xist RNA Binding Proteins. *Cell* 161 (2), 404–416. [PubMed: 25843628]
- (9). Roth A, and Diederichs S (2015) Rap and Chirp about X Inactivation. *Nature* 521, 170–171. [PubMed: 25971508]



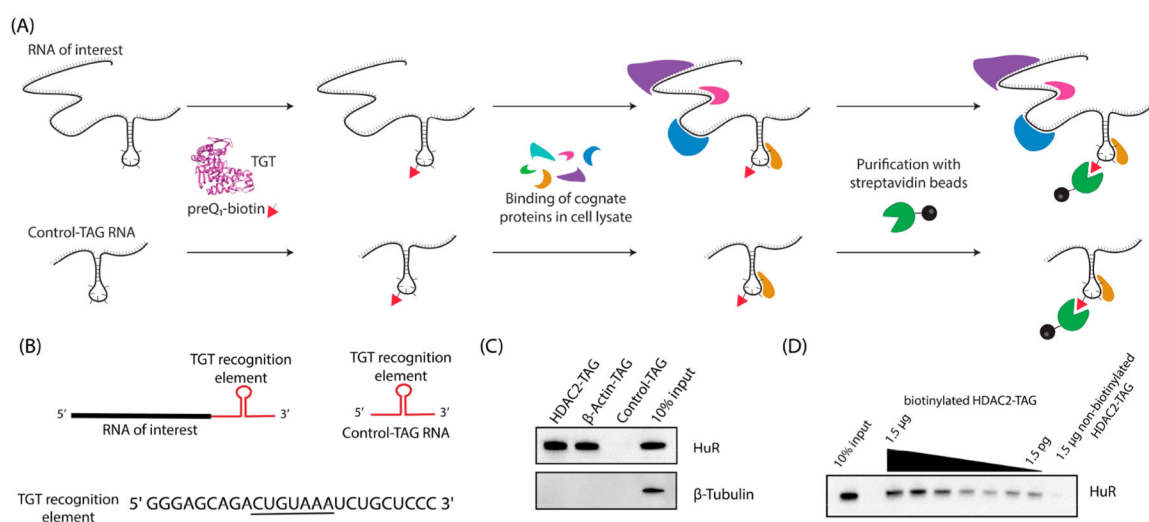
- Author Manuscript
- Author Manuscript
- Author Manuscript
- Author Manuscript
- (10). Rogell B, Fischer B, Rettel M, Krijgsveld J, Castello A, and Hentze MW (2017) Specific RNP Capture with Antisense LNA/DNA Mixmers. *RNA* 23 (8), 1290–1302. [PubMed: 28476952]
  - (11). Matia-González AM, Iadevaia V, and Gerber AP (2017) A Versatile Tandem RNA Isolation Procedure to Capture in Vivo Formed mRNA-Protein Complexes. *Methods* 118–119, 93.
  - (12). Yoon JH, Srikantan S, and Gorospe M (2012) MS2-TRAP (MS2-Tagged RNA Affinity Purification): Tagging RNA to Identify Associated MiRNAs. *Methods* 58 (2), 81–87. [PubMed: 22813890]
  - (13). Zhou Z, and Reed R (2003) Purification of Functional RNA-Protein Complexes Using MS2-MBP. *Curr. Protoc. Mol. Biol* 63, 27.3.1–27.3.7.
  - (14). Liu S, Zhu J, Jiang T, Zhong Y, Tie Y, Wu Y, Zheng X, Jin Y, and Fu H (2015) Identification of LncRNA MEG3 Binding Protein Using MS2-Tagged RNA Affinity Purification and Mass Spectrometry. *Appl. Biochem. Biotechnol* 176 (7), 1834–1845. [PubMed: 26155902]
  - (15). Slobodin B, and Gerst JE (2010) A Novel mRNA Affinity Purification Technique for the Identification of Interacting Proteins and Transcripts in Ribonucleoprotein Complexes. *RNA* 16 (11), 2277–2290. [PubMed: 20876833]
  - (16). Johansson HE, Dertinger D, Lecuyer KA, Behlen LS, Greef CH, and Uhlenbeck UC (1998) A Thermodynamic Analysis of the Sequence-Specific Binding of RNA by Bacteriophage MS2 Coat Protein. *Proc. Natl. Acad. Sci. U. S. A* 95 (16), 9244–9249. [PubMed: 9689065]
  - (17). Vázquez AE, Nie L, and Yamoah EN (2009) Synthesis of Biotin-Labeled RNA for Gene Expression Measurements Using Oligonucleotide Arrays. *Methods Mol. Biol* 493, 21–29. [PubMed: 18839339]
  - (18). Moritz B, and Wahle E (2014) Simple Methods for the 3′ Biotinylation of RNA. *RNA* 20 (3), 421–427. [PubMed: 24448448]
  - (19). Muthmann N, Hartstock K, and Rentmeister A (2020) Chemo-Enzymatic Treatment of RNA to Facilitate Analyses, *Wiley Interdisciplinary Reviews: RNA*, Blackwell Publishing Ltd, DOI: 10.1002/wrna.1561.
  - (20). Schulz D, Holstein JM, and Rentmeister A (2013) A Chemo-Enzymatic Approach for Site-Specific Modification of the RNA Cap. *Angew. Chem., Int. Ed* 52 (30), 7874–7878.
  - (21). Holstein JM, Schulz D, and Rentmeister A (2014) Bioorthogonal Site-Specific Labeling of the 5′-Cap Structure in Eukaryotic MRNAs. *Chem. Commun. (Cambridge, U. K.)* 50 (34), 4478–4481.
  - (22). Li F, Dong J, Hu X, Gong W, Li J, Shen J, Tian H, and Wang J (2015) A Covalent Approach for Site-Specific RNA Labeling in Mammalian Cells. *Angew. Chem., Int. Ed* 54 (15), 4597–4602.
  - (23). Alexander SC, and Devaraj NK (2017) Developing a Fluorescent Toolbox to Shed Light on the Mysteries of RNA. *Biochemistry* 56 (39), 5185–5193. [PubMed: 28671838]
  - (24). Alexander SC, Busby KN, Cole CM, Zhou CY, and Devaraj NK (2015) Site-Specific Covalent Labeling of RNA by Enzymatic Transglycosylation. *J. Am. Chem. Soc* 137 (40), 12756–12759. [PubMed: 26393285]
  - (25). Busby KN, and Devaraj NK (2020) Enzymatic Covalent Labeling of RNA with RNA Transglycosylation at Guanosine (RNA-TAG). *Methods Enzymol.* 641, 373. [PubMed: 32713531]
  - (26). Zhou CY, Alexander SC, and Devaraj NK (2017) Fluorescent Turn-on Probes for Wash-Free mRNA Imaging: Via Covalent Site-Specific Enzymatic Labeling. *Chem. Sci* 8 (10), 7169–7173. [PubMed: 29081948]
  - (27). de Silanes IL, Zhan M, Lal A, Yang X, and Gorospe M (2004) Identification of a Target RNA Motif for RNA-Binding Protein HuR. *Proc. Natl. Acad. Sci. U. S. A* 101 (9), 2987–2992. [PubMed: 14981256]
  - (28). Dormoy-Raclet V, Menard I, Clair E, Kurban G, Mazroui R, Di Marco S, von Roretz C, Pause A, and Gallouzi I-E (2007) The RNA-Binding Protein HuR Promotes Cell Migration and Cell Invasion by Stabilizing the  $\beta$ -Actin mRNA in a U-Rich-Element-Dependent Manner. *Mol. Cell Biol* 27 (15), 5365–5380. [PubMed: 17548472]
  - (29). Lebedeva S, Jens M, Theil K, Schwänhauser B, Selbach M, Landthaler M, and Rajewsky N (2011) Transcriptome-Wide Analysis of Regulatory Interactions of the RNA-Binding Protein HuR. *Mol. Cell* 43 (3), 340–352. [PubMed: 21723171]

- (30). Zhao L, Kroenke CD, Song J, Piwnica-Worms D, Ackerman JJH, and Neil JJ (2008) Intracellular Water-Specific MR of Microbead-Adherent Cells: The HeLa Cell Intracellular Water Exchange Lifetime. *NMR Biomed.* 21, 159–164. [PubMed: 17461436]
- (31). Peterlin BM, Brogie JE, and Price DH (2012) 7SK SnRNA: A Noncoding RNA That Plays a Major Role in Regulating Eukaryotic Transcription. *Wiley Interdiscip. Rev. RNA* 3 (1), 92–103. [PubMed: 21853533]
- (32). Flynn RA, Do BT, Rubin AJ, Calo E, Lee B, Kuchelmeister H, Rale M, Chu C, Kool ET, Wysocka J, Khavari PA, and Chang HY (2016) 7SK-BAF Axis Controls Pervasive Transcription at Enhancers. *Nat. Struct. Mol. Biol* 23 (3), 231–238. [PubMed: 26878240]
- (33). Boersema PJ, Raijmakers R, Lemeer S, Mohammed S, and Heck AJR (2009) Multiplex Peptide Stable Isotope Dimethyl Labeling for Quantitative Proteomics. *Nat. Protoc* 4 (4), 484–494. [PubMed: 19300442]
- (34). Nakanishi S, Ueda T, Hori H, Yamazaki N, Watanabe N, and Okada K (1994) A UGU Sequence in the Anticodon Loop Is a Minimum Requirement for Recognition by Escherichia Coli TRNA-Guanine Transglycosylase. *J. Biol. Chem* 269, 32221–32225. [PubMed: 7528209]
- (35). Curnow AW, and Garcia GA (1995) TRNA-Guanine Transglycosylase from Escherichia Coli: Minimal TRNA Structure and Sequence Requirements for Recognition. *J. Biol. Chem* 270 (29), 17264–17267. [PubMed: 7615526]
- (36). Xie W, Liu X, and Huang RH (2003) Chemical Trapping and Crystal Structure of a Catalytic tRNA Guanine Transglycosylase Covalent Intermediate. *Nat. Struct. Mol. Biol* 10 (10), 781–788.
- (37). Spitale RC, Flynn R. a., Zhang QC, Crisalli P, Lee B, Jung J-W, Kuchelmeister HY, Batista PJ, Torre E. a., Kool ET, and Chang HY (2015) Structural Imprints in Vivo Decode RNA Regulatory Mechanisms. *Nature* 519 (7544), 486–490. [PubMed: 25799993]
- (38). Hajjari M, and Salavaty A (2015) HOTAIR: An Oncogenic Long Non-Coding RNA in Different Cancers. *Cancer Biol. Med* 12, 1–9. [PubMed: 25859406]
- (39). Somarowthu S, Legiewicz M, Chillon I, Marcia M, Liu F, and Pyle AM (2015) HOTAIR Forms an Intricate and Modular Secondary Structure. *Mol. Cell* 58 (2), 353–361. [PubMed: 25866246]
- (40). Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, and Chang HY (2007) Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs. *Cell* 129 (7), 1311–1323. [PubMed: 17604720]
- (41). Tsai M, Manor O, Wan Y, Mosammamaparast N, Wang JK, Lan F, Shi Y, Segal E, and Chang HY (2010) Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes. *Science (Washington, DC, U. S.)* 329 (5992), 689–693.
- (42). Amândio AR, Necsulea A, Joye E, Mascrez B, and Duboule D (2016) Hotair Is Dispensable for Mouse Development. *PLoS Genet.* 12 (12), e1006232. [PubMed: 27977683]
- (43). Selleri L, Bartolomei MS, Bickmore WA, He L, Stubbs L, Reik W, and Barsh GS (2016) A Hox-Embedded Long Noncoding RNA: Is It All Hot Air? *PLoS Genet.* 12 (12), e1006485. [PubMed: 27977680]
- (44). Davidovich C, Zheng L, Goodrich KJ, and Cech TR (2013) Promiscuous RNA Binding by Polycomb Repressive Complex 2. *Nat. Struct. Mol. Biol* 20 (11), 1250–1257. [PubMed: 24077223]
- (45). Bhan A, Hussain I, Ansari KI, Kasiri S, Bashyal A, and Mandal SS (2013) Antisense Transcript Long Noncoding RNA (LncRNA) HOTAIR Is Transcriptionally Induced by Estradiol. *J. Mol. Biol* 425 (19), 3707–3722. [PubMed: 23375982]
- (46). Gupta R. a, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai M-C, Hung T, Argani P, Rinn JL, Wang Y, Brzoska P, Kong B, Li R, West RB, van de Vijver MJ, Sukumar S, and Chang HY (2010) Long Non-Coding RNA HOTAIR Reprograms Chromatin State to Promote Cancer Metastasis. *Nature* 464 (7291), 1071–1076. [PubMed: 20393566]
- (47). Ge SX, Jung D, and Yao R (2020) ShinyGO: A Graphical Enrichment Tool for Animals and Plants. *Bioinformatics* 36 (8), 2628–2629. [PubMed: 31882993]
- (48). Matsumura K, Kawasaki Y, Miyamoto M, Kamoshida Y, Nakamura J, Negishi L, Suda S, and Akiyama T (2017) The Novel G-Quadruplex-Containing Long Non-Coding RNA GSEC

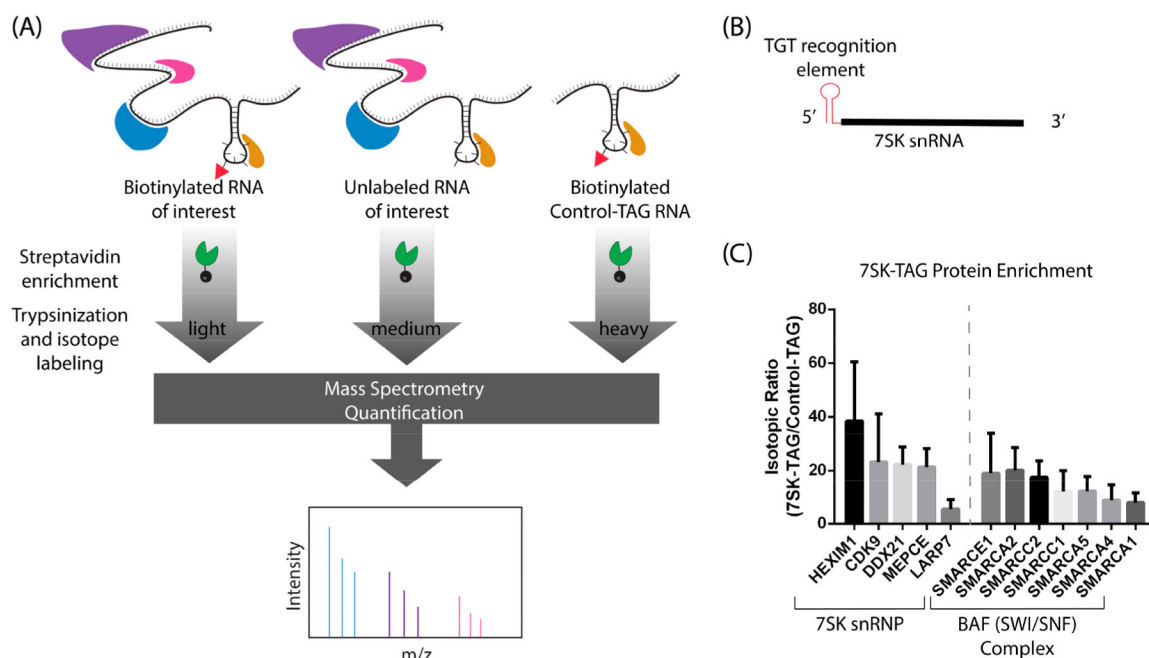
- Antagonizes DHX36 and Modulates Colon Cancer Cell Migration. *Oncogene* 36 (9), 1191–1199. [PubMed: 27797375]
- (49). Wu L, Murat P, Matak-Vinkovic D, Murrell A, and Balasubramanian S (2013) The Binding Interaction between Long Non-Coding RNA HOTAIR and PRC2 Proteins. *Biochemistry* 52 (52), 9519–9527. [PubMed: 24320048]
- (50). Tripathi V, Ellis JD, Shen Z, Song DY, Pan Q, Watt AT, Freier SM, Bennett CF, Sharma A, Bubulya PA, Blencowe BJ, Prasanth SG, and Prasanth KV (2010) The Nuclear-Retained Noncoding RNA MALAT1 Regulates Alternative Splicing by Modulating SR Splicing Factor Phosphorylation. *Mol. Cell* 39 (6), 925–938. [PubMed: 20797886]
- (51). Romero-Barrios N, Legascue MF, Benhamed M, Ariel F, and Crespi M (2018) Survey and Summary Splicing Regulation by Long Noncoding RNAs. *Nucleic Acids Res.* 46 (5), 2169–2184. [PubMed: 29425321]
- (52). Xing YH, Yao RW, Zhang Y, Guo CJ, Jiang S, Xu G, Dong R, Yang L, and Chen LL (2017) SLERT Regulates DDX21 Rings Associated with Pol I Transcription. *Cell* 169 (4), 664–678 e16. [PubMed: 28475895]
- (53). Stengl B, Meyer E. a., Heine A, Brenk R, Diederich F, and Klebe G (2007) Crystal Structures of TRNA-Guanine Transglycosylase (TGT) in Complex with Novel and Potent Inhibitors Unravel Pronounced Induced-Fit Adaptations and Suggest Dimer Formation Upon Substrate Binding. *J. Mol. Biol* 370, 492–511. [PubMed: 17524419]
- (54). Schellenberger V, Wang CW, Geething NC, Spink BJ, Campbell A, To W, Scholle MD, Yin Y, Yao Y, Bogin O, Cleland JL, Silverman J, and Stemmer WPC (2009) A Recombinant Polypeptide Extends the in Vivo Half-Life of Peptides and Proteins in a Tunable Manner. *Nat. Biotechnol* 27 (12), 1186–1190. [PubMed: 19915550]
- (55). Engreitz JM, Pandya-Jones A, McDonel P, Shishkin A, Sirokman K, Surka C, Kadri S, Xing J, Goren A, Lander ES, Plath K, and Guttman M (2013) The Xist LncRNA Exploits Three-Dimensional Genome Architecture to Spread across the X Chromosome. *Science* 341 (6147), 1237973. [PubMed: 23828888]
- (56). Ehret F, Zhou CY, Alexander SC, Zhang D, and Devaraj NK (2018) Site-Specific Covalent Conjugation of Modified mRNA by TRNA Guanine Transglycosylase. *Mol. Pharmaceutics* 15 (3), 737–742.
- (57). Gendron JM, Webb K, Yang B, Rising L, Zuzow N, and Bennett EJ (2016) Using the Ubiquitin-Modified Proteome to Monitor Distinct and Spatially Restricted Protein Homeostasis Dysfunction. *Mol. Cell. Proteomics* 15 (8), 2576–2593. [PubMed: 27185884]
- (58). Reinke AW, Mak R, Troemel ER, and Bennett EJ (2017) In Vivo Mapping of Tissue- and Subcellular-Specific Proteomes in *Caenorhabditis Elegans*. *Sci. Adv* 3 (5), e1602426. [PubMed: 28508060]
- (59). Livak KJ, and Schmittgen TD (2001) Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2<sup>-</sup>CT Method. *Methods* 25 (4), 402–408. [PubMed: 11846609]
- (60). Patro R, Duggal G, Love MI, Irizarry RA, and Kingsford C (2017) Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression. *Nat. Methods* 14 (4), 417–419. [PubMed: 28263959]
- (61). Love MI, Huber W, and Anders S (2014) Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. *Genome Biol.* 15 (12), 550. [PubMed: 25516281]
- (62). Sonesson C, Love MI, and Robinson MD (2015) Differential Analyses for RNA-Seq: Transcript-Level Estimates Improve Gene-Level Inferences [Version 2; Referees: 2 Approved]. *F1000Research* 4, 1521. [PubMed: 26925227]



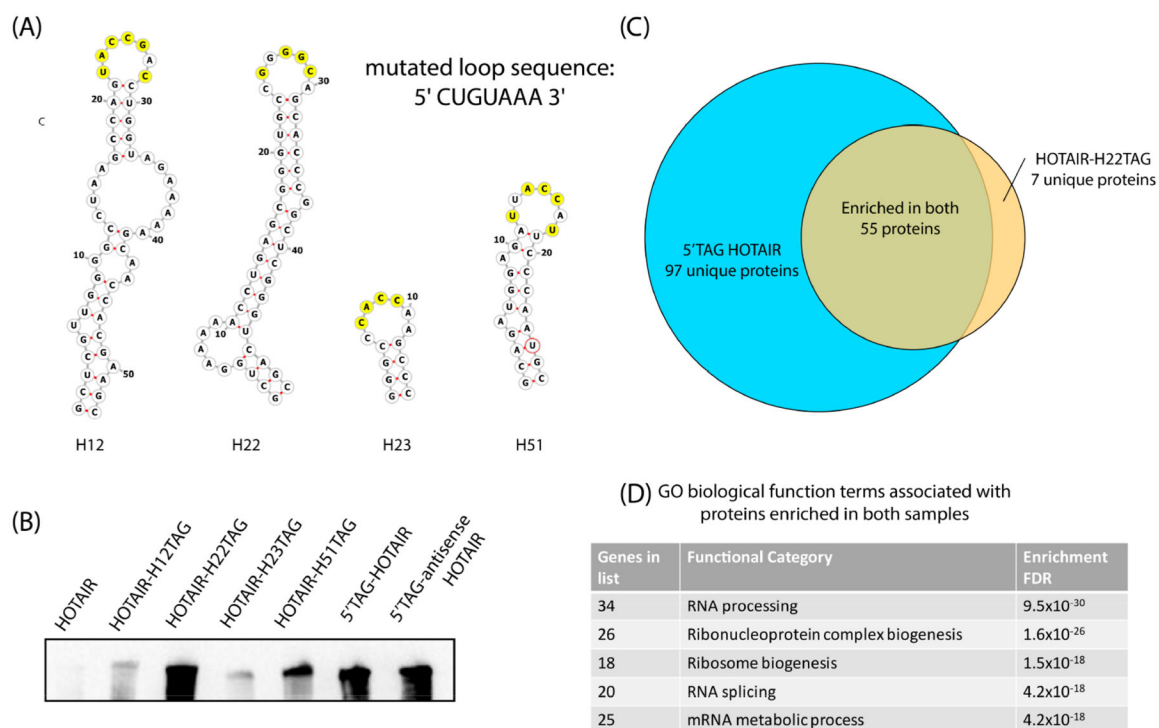
**Figure 1.** Enzymatic RNA labeling with RNA-TAG. Using bacterial tRNA Guanine Transglycosylase (TGT), an RNA of interest containing a 25 nucleotide TGT recognition hairpin is labeled with a derivative of the natural substrate, preQ<sub>1</sub>.

**Figure 2.**

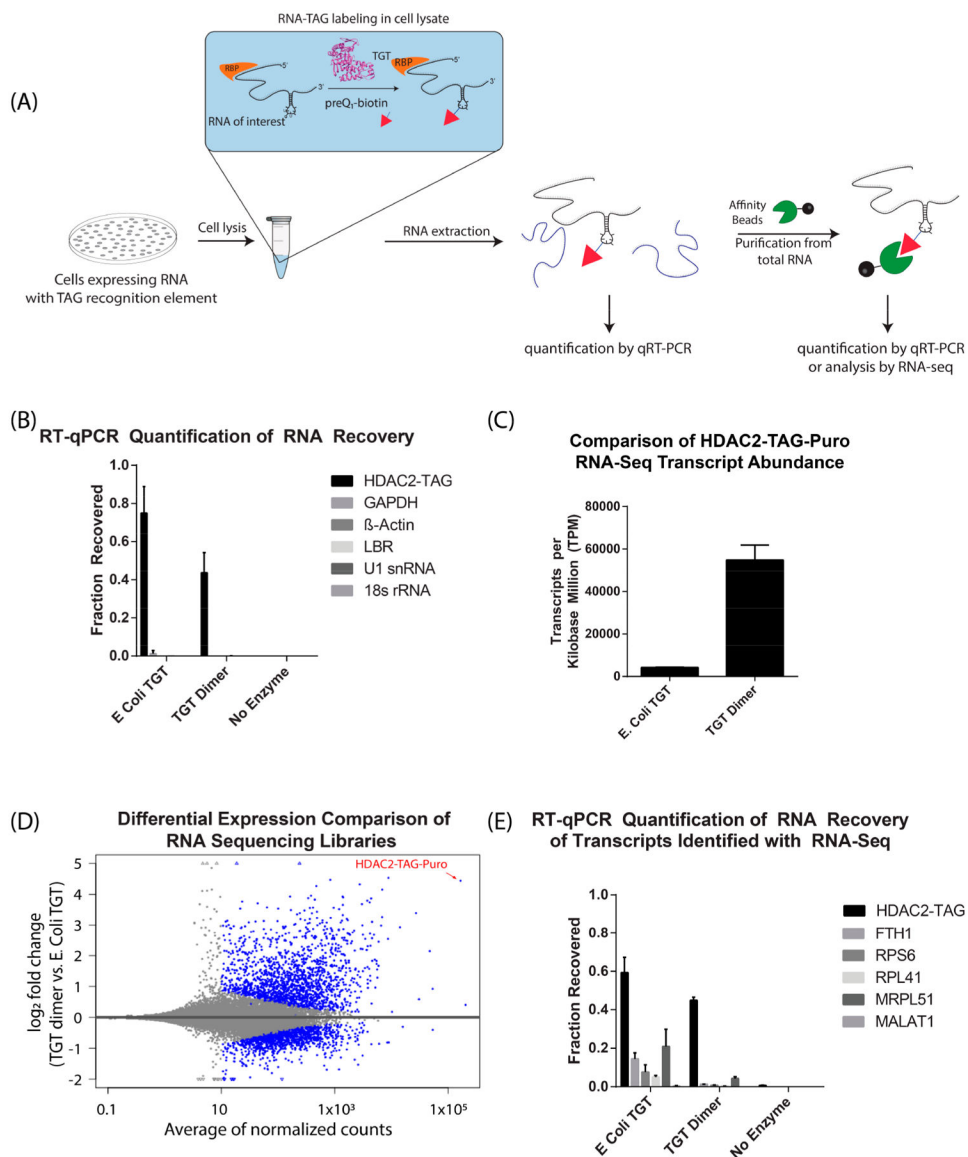
Analysis of RNA–protein interactions using RNA-TAG. (A) An RNA of interest with an appended TAG sequence is biotinylated via treatment with *E. coli* TGT and preQ<sub>1</sub>-biotin and allowed to bind cognate RNA-binding proteins, alongside a control RNA for analysis of bound proteins. (B) RNA constructs used for Western blotting experiments contain an RNA of interest (HDAC2-HuR binding domain or  $\beta$ -Actin mRNA) followed by a 25-nt TGT recognition element with the indicated sequence, where the loop region is underlined, surrounded by short spacer regions. Control-TAG RNA has no RNA of interest inserted. (C) RNAs biotinylated with RNA-TAG were used directly to analyze cognate proteins via Western blotting by incubation of 1 pmol labeled RNA with 40  $\mu$ g of cell lysate. (D) HDAC2-TAG-HuR complex can be detected using reduced RNA concentrations, as shown by a titration of RNA from 1.5  $\mu$ g to 1.5pg.

**Figure 3.**

Quantitative proteomics to identify RNA–protein interactions. (A) The affinity purification of the biotinylated RNA of interest was carried out alongside two controls, an unlabeled RNA of interest and a biotinylated Control-TAG transcript. Isotope labeling was carried out using reductive dimethylation; samples were pooled and analyzed with LCMS/MS. (B) 7SK-TAG RNA construct with TAG sequence appended directly to the 5' end. (C) Isotopic ratios (biotinylated 7SK-TAG/biotinylated Control-TAG) for enriched proteins that are known interactors with 7SK snRNA or part of the BAF complex known to interact with 7SK snRNP (mean  $\pm$  S.D.). Numbers of replicates were as follows:  $n = 5$  (CDK9, DDX21, MEPCE, LARP7, SMARCA5),  $n = 4$  (SMARCE1, SMARCA2, SMARCC1),  $n = 3$  (HEXIM1, SMARCA4, SMARCA1, SMARCC2).

**Figure 4.**

Analysis of HOTAIR-associated proteins with alternative labeling strategies. (A) Stem loop structures derived from previous structural characterization<sup>39</sup> with yellow residues indicating bases that required mutation to match the loop (CUGUAAA) of the TGT recognition element. (B) Northern blot analysis of RNA transcripts subjected to TGT biotinylation conditions, developed with Streptavidin-HRP. (C) Venn diagram showing proteins that were enriched from MCF7 cell lysates >3 fold compared to both unlabeled and 5'TAG antisense-HOTAIR controls. (D) Top gene ontology biological function terms enriched in both samples.<sup>47</sup>



**Figure 5.** RNAs expressed in cells can be biotinylated efficiently and selectively in cell lysates. (A) Schematic of RNA labeling in cell lysate and RT-qPCR analysis. Cell lysates are supplemented with TGT enzyme and preQ<sub>1</sub>-biotin. Subsequently, RNA is extracted for affinity purification and analysis (B) Fraction of RNA recovered after cell lysate labeling was calculated (mean ± S.D.,  $N=3$ ) by comparison of purified RNA to an input RNA sample, as described in methods. (C) Transcripts per kilobase million (TPM) of target HDAC2-TAG-Puro transcript in RNA sequencing libraries (mean ± S.D.,  $N=3$ ). (D) Scatter plot of log<sub>2</sub> fold change between TGT dimer and *E. coli* TGT versus mean expression (average of normalized counts). The blue color marks genes detected as differentially expressed at 10% false discovery rate with Benjamini-Hochberg testing ( $p$ -value < 0.1). The triangle symbols at the upper and lower plot border indicate transcripts with very large or infinite log<sub>2</sub> fold change. The HDAC-TAG-Puro transcript is located at the top right corner



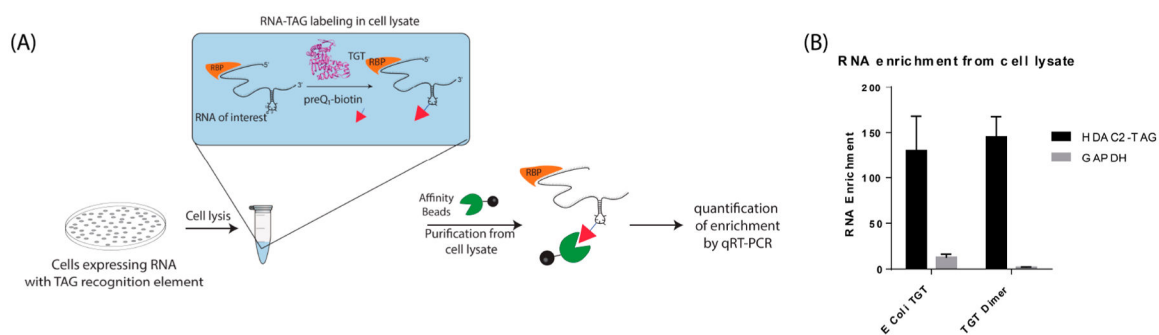
of the plot and pointed by a red arrow. (E) Fraction of RNA recovered after cell lysate labeling was calculated for RNAs detected in the sequencing libraries (mean  $\pm$  S.D,  $N=3$ ).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript



**Figure 6.** Labeling and purification of RNA directly from cell lysates. (A) RNA was labeled in cell lysates, and immediately purified using streptavidin beads and stringent washes, allowing for quantification of enrichment of RNA directly from cell lysates by RT-qPCR. (B) RT-qPCR enrichment as measured relative to no enzyme control, using  $\beta$ -actin as a reference gene, representing mean  $\pm$  S.D.,  $N=3$ . Supporting data can be found in the Supporting Information.