

# RIP-PEN-seq identifies a class of kink-turn RNAs as splicing regulators

Received: 2 August 2022

Accepted: 13 March 2023

Published online: 10 April 2023

 Check for updates

Bin Li<sup>1,4</sup>, Shurong Liu<sup>1,4</sup>, Wujian Zheng<sup>1,2,4</sup>, Anrui Liu<sup>1,2,4</sup>, Peng Yu<sup>1,4</sup>, Di Wu<sup>1,2</sup>, Jie Zhou<sup>1,2</sup>, Ping Zhang<sup>1</sup>, Chang Liu<sup>1</sup>, Qiao Lin<sup>1,2</sup>, Jiayi Ye<sup>1,2</sup>, Simeng He<sup>2</sup>, Qiaojuan Huang<sup>1</sup>, Hui Zhou<sup>1</sup>, Jianjun Chen<sup>3</sup>, Lianghu Qu<sup>1</sup>✉ & Jianhua Yang<sup>1,2</sup>✉

A kink-turn (K-turn) is a three-dimensional RNA structure that exists in all three primary phylogenetic domains. In this study, we developed the RIP-PEN-seq method to identify the full-length sequences of RNAs bound by the K-turn binding protein 15.5K and discovered a previously uncharacterized class of RNAs with backward K-turn motifs (bktRNAs) in humans and mice. All bktRNAs share two consensus sequence motifs at their fixed terminal position and have complex folding properties, expression and evolution patterns. We found that a highly conserved bktRNA1 guides the methyltransferase fibrillarin to install RNA methylation of U12 small nuclear RNA in humans. Depletion of bktRNA1 causes global splicing dysregulation of U12-type introns by impairing the recruitment of ZCRB1 to the minor spliceosome. Most bktRNAs regulate the splicing of local introns by interacting with the 15.5K protein. Taken together, our findings characterize a class of small RNAs and uncover another layer of gene expression regulation that involves crosstalk among bktRNAs, RNA splicing and RNA methylation.

RNA structural motifs confer diversity of RNA architecture to regulate various biological processes<sup>1</sup>. The kink-turn (K-turn) is the most prevalent three-dimensional (3D) RNA structural motif in messenger RNAs (mRNAs) and non-coding RNAs (ncRNAs)<sup>2–4</sup>. The K-turn structures are typically characterized by a canonical stem (C-stem) followed by an asymmetric 3-nucleotide (3-nt) bulge and G•A and A•G base pairs on the 3' non-canonical stem (NC-stem)<sup>2,3</sup>. The naturally occurring K-turn often serves as a specific binding site for RNA-binding proteins (RBPs), including the protein L7Ae<sup>5</sup> in archaea and the other homologues YbxF, YlxQ<sup>6</sup> and 15.5-kDa protein (15.5K)<sup>7</sup>. Notably, the unique conformational property of the K-turn, as well as its binding proteins, enables the K-turn to play pivotal roles in maintaining the proper structure and biological function of various types of RNAs<sup>2–4</sup>.

The K-turn structural motifs are recurrently observed in various types of ncRNA families and play important roles in RNA metabolism<sup>2–4</sup>. In eukaryotes, the well-known K-turn ncRNAs are the box

C/D small nucleolar RNAs (snoRNAs) and U4/U4atac small nuclear RNAs (snRNAs). The C/D boxes are located at the box C/D RNA terminal pair and form the terminal K-turn motif<sup>8</sup>. Recognition of the K-turn structural motif of C/D box RNAs by the 15.5K protein initiates small nucleolar ribonucleoprotein (snoRNP) assembly to carry out site-specific 2'-O-methylation of ribosomal RNAs (rRNAs) and snRNAs<sup>9</sup>, which is essential for the functional fidelity of ribosomes and gene expression<sup>10,11</sup>. The 15.5K protein also binds to the identical K-turn structural motifs in the 5' stem-loop of U4 and U4atac, promoting the assembly of the major and minor spliceosomes<sup>12,13</sup>. Mutations in the U4atac K-turn structural motif have been shown to impair the binding of the 15.5K protein<sup>14</sup> and cause various diseases, such as the Taybi–Linder syndrome (TALS/MOPD1)<sup>15,16</sup>, Roifman syndrome (RFMN)<sup>17</sup> and Lowry–Wood syndrome (LWS)<sup>18</sup>. However, the prevalence, mechanism and function of K-turn structures in the transcriptome remain largely unknown.

<sup>1</sup>MOE Key Laboratory of Gene Function and Regulation, State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, China. <sup>2</sup>The Fifth Affiliated Hospital, Sun Yat-sen University, Zhuhai, China. <sup>3</sup>Department of Systems Biology, Beckman Research Institute of City of Hope, Monrovia, CA, USA. <sup>4</sup>These authors contributed equally: Bin Li, Shurong Liu, Wujian Zheng, Anrui Liu, Peng Yu. ✉ e-mail: [lssqlh@mail.sysu.edu.cn](mailto:lssqlh@mail.sysu.edu.cn); [yangjh7@mail.sysu.edu.cn](mailto:yangjh7@mail.sysu.edu.cn)

In this study, we developed a method for RNA immunoprecipitation coupled with sequencing of paired ends of ncRNAs (RIP-PEN-seq) to identify the full-length sequences of ncRNAs bound by the 15.5K protein and discover a class of backward K-turn RNAs (bktRNAs; Fig. 1a) with two consensus motifs. In contrast to all well-known K-turn RNAs (ktRNAs) in humans, whose internal loops are located at the 5' ends of the RNAs (forward ktRNAs (fktRNAs); Fig. 1a), the internal loops of the ktRNAs identified in this study are located at the 3' termini; we, therefore, named them bktRNAs (Fig. 1a). These bktRNAs have a strong preference for a 5' motif (CUGA) and a 3' motif (UGAUG) at the fixed terminal position of RNA. Moreover, we identified a functional bktRNA1 that interacts with the 15.5K and fibrillar (FBL) proteins to regulate RNA modification of U12 snRNA. Loss of bktRNA1 results in aberrant RNA splicing of hundreds of U12-type introns. In addition, we reveal bktRNA1 and that 80% of randomly selected bktRNAs are involved in the local regulation of intron splicing. Together, our results reveal that bktRNAs form a previously undescribed class of post-transcriptional regulators.

## Results

### bktRNAs with consensus motifs identified by RIP-PEN-seq

To identify ncRNAs (20–500 nt) with K-turn structural motifs, we developed an efficient RNA cloning scheme called RIP-PEN-seq that uses dual RNA adapters and size selection, and a series of advanced experimental strategies to capture both ends of any ncRNA bound by the 15.5K protein (Fig. 1b, Extended Data Fig. 1a and Methods). We applied RIP-PEN-seq to HEK293T cells expressing FLAG-tagged 15.5K (Extended Data Fig. 1b,c), and a total of approximately 520 million paired-end reads were obtained, of which approximately 282 million could be uniquely mapped to the human genome. Notably, more than 45% of the uniquely aligned reads mapped to previously defined fktRNAs. The start and end sites of these known fktRNAs could be precisely identified at single-nucleotide resolution (Extended Data Fig. 1d,e). For example, the precise 5' and 3' ends of 10 known C/D box fktRNAs with K-turn motifs located within the introns of *GASS5* genes were identified by our RIP-PEN-seq method (Extended Data Fig. 1f). Together, these results suggest that our RIP-PEN-seq approach not only shows high specificity and accuracy in enriching ktRNAs but can also capture their full-length sequences.

Although almost all known ktRNAs are fktRNAs in humans, as the 15.5K protein has no preference for binding forward or backward K-turn structural motifs<sup>2,19</sup>, we hypothesized that the 15.5K RIP-PEN-seq data may contain hidden bktRNAs. We then developed a computational tool, called kturnSeeker, to identify potential ktRNAs based on sequence and secondary structure information from the RIP-PEN-seq data (Extended Data Fig. 1g,h). As a result, we identified 605 previously undescribed fktRNAs as well as 118 canonical fktRNAs (Extended Data Fig. 1i and Supplementary Table 1). Notably, most previously undescribed fktRNAs are located within introns (Extended Data Fig. 1i and Supplementary Table 1). Furthermore, after removing fktRNAs and known annotations and keeping the ktRNAs present in at least two sequencing libraries, we discovered 84 ktRNAs (alignment to 93 genomic loci) with backward K-turn structural motifs, which we named bktRNA1 to bktRNA84 (Supplementary Table 1).

Intriguingly, the K-turn structural motifs of most bktRNAs are located at 4 nt and 2 nt from the RNA 5' end and 3' end, respectively (Fig. 1c,d). Moreover, we observed highly substantial enrichment of a CUGA motif often 4 nt downstream of the 5' end and a UGAUG motif 2 nt upstream of the 3' end by de novo motif enrichment analysis (Fig. 1e and Extended Data Fig. 2a). Notably, all these enrichment consensus motifs were located within the K-turn structural motifs of bktRNAs (Fig. 1f–i and Extended Data Fig. 2b).

To further confirm that these two conserved motifs can truly adopt the backward K-turn structure in vivo, we developed 15.5K RIP-PEN-SHAPE-MaP, an approach combining SHAPE-MaP<sup>20,21</sup> and our

RIP-PEN-seq experiments to probe intact RNA structures of bktRNAs (Extended Data Fig. 2c and Methods). We first constructed reactivity scores for the known fktRNAs and found that paired nucleotides in the NC-stem and C-stems exhibited lower SHAPE reactivity and that unpaired nucleotides in the loop region (Loop) had relatively high SHAPE reactivity (Extended Data Fig. 2d,e and Supplementary Table 2), demonstrating the accuracy of the 15.5K RIP-PEN-SHAPE-MaP method. For example, the SHAPE reactivity signal agreed well with the structure of fktRNA SNORD102 (Extended Data Fig. 2e). We then mapped the SHAPE reactivity signals to the predicted secondary structures of bktRNAs and found that the paired nucleotides in backward K-turn structures had lower reactivity signals than the unpaired nucleotides (Fig. 1g–i and Supplementary Table 2). Therefore, these data confirmed that the 5' and 3' motifs identified in bktRNAs are likely to adopt the backward K-turn structure in vivo.

We next applied RIP-PEN-seq to mouse Hepa1-6 cells stably expressing FLAG-tagged 15.5K (Supplementary Fig. 1a,b), and it also detected mouse fktRNAs with high specificity and accuracy (Supplementary Fig. 1c–e). We identified 43 bktRNAs in mice that shared common genomic and motif characteristics with human bktRNAs (Supplementary Fig. 1f–k and Supplementary Table 3). Overall, by developing the RIP-PEN-seq approach and the kturnSeeker algorithm, we uncovered a class of ktRNAs with a backward K-turn structural motif composed of two consensus sequence elements in fixed positions of RNAs in humans and mice.

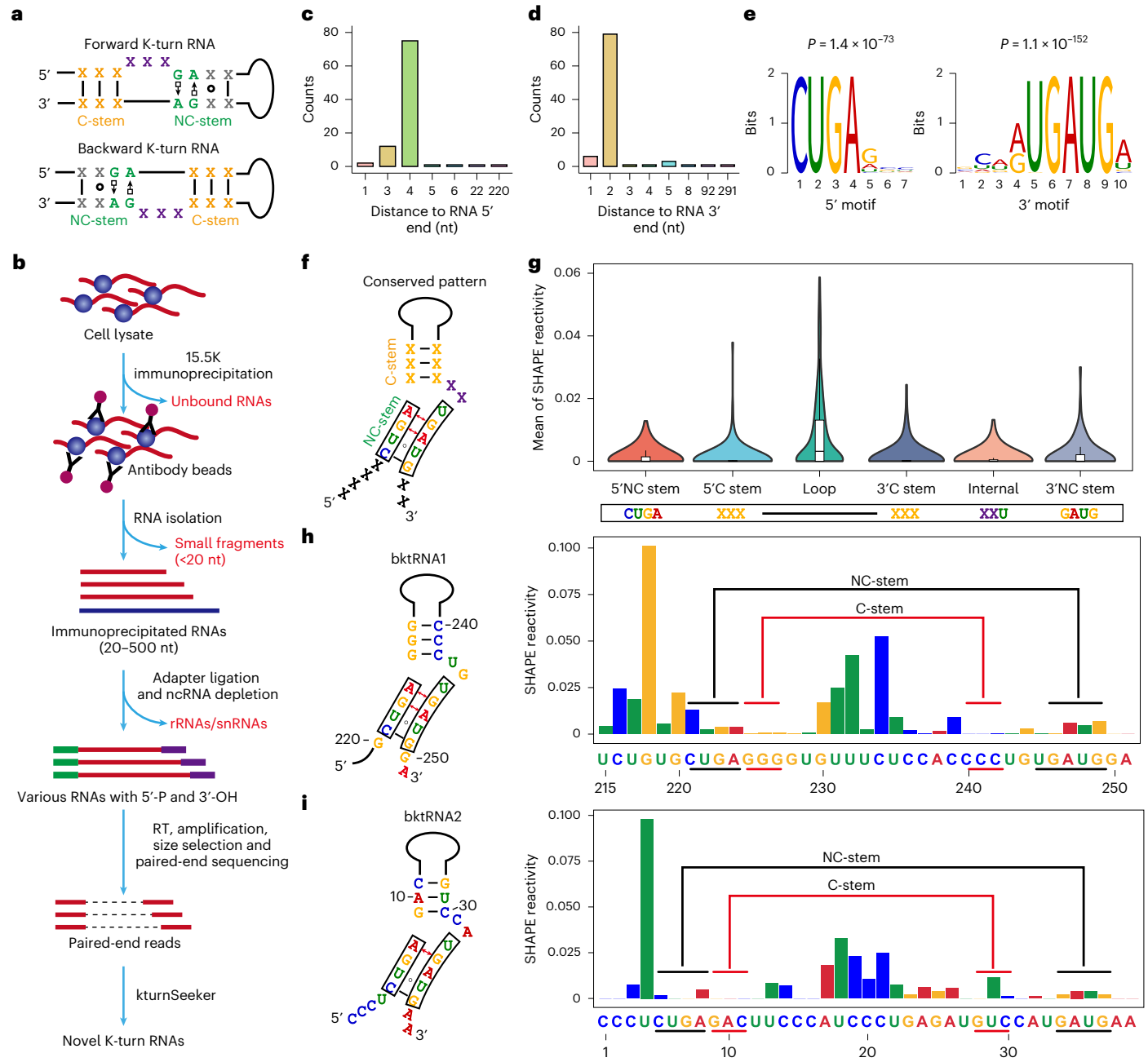
### bktRNAs have complex folding, expression and evolution patterns

To identify bktRNAs across human tissues and cells, we further applied our kturnSeeker tool to 28 small RNA sequencing (sRNA-seq) datasets generated by our PEN-seq method (Methods) and 234 public sRNA-seq datasets produced by the ENCODE consortium. This method revealed 379 high-confidence candidate bktRNAs (Fig. 2a), 72 of which overlapped with the abovementioned bktRNAs bound by the 15.5K protein (~86%, 72/84; Fig. 2a). By integrating the RIP-PEN-seq dataset and the sRNA-seq datasets, we identified 391 bktRNAs (407 genomic loci; Fig. 2b and Supplementary Table 4). As described for the 15.5K-bound bktRNAs, these bktRNAs had substantial enrichment of CUGA motifs 4 nt downstream of the 5' end and UGAUG motifs 2 nt upstream of the 3' end (Fig. 2c,d).

To determine the folding characteristics<sup>22,23</sup> of these K-turn sequences, we first devised a nomenclature for the nucleotide positions in the backward K-turn structure according to the nomenclature for the forward K-turn structure<sup>23,24</sup> (Extended Data Fig. 3a). We further examined the distribution of 3b:3n and –1b:–1n sequences in bktRNAs (Extended Data Fig. 3b). We found that almost all bktRNAs have a strong propensity for the 3b:3n = U:U (99.5%, 405/407), which leads to an N3 conformation and the weak folding in response to metal ions<sup>22–25</sup>. We also discovered that approximately 69% (281/407) of bktRNAs have –1b:–1n = C:G or A:U, which leads to good folding in response to metal ions<sup>22–25</sup> (Extended Data Fig. 3c). Moreover, bktRNAs in mice showed similar folding properties (Extended Data Fig. 3d,e). Notably, the 3b:3n = U:U identified in bktRNAs is considerably different from the distribution of ribosomal Kt-7, SAM-I riboswitch and U4 snRNAs<sup>22–25</sup>.

We next explored the m<sup>6</sup>A modification of bktRNAs<sup>26,27</sup> by intersecting our bktRNAs with the annotated m<sup>6</sup>A sites from RMBase<sup>28</sup> and found that 12 and 1 bktRNAs contained m<sup>6</sup>A modifications in humans and mice (Extended Data Fig. 3f), respectively. Moreover, neither human nor mouse bktRNAs contain m<sup>6</sup>A modifications at the 1n position (Extended Data Fig. 3g). Together, these findings suggest that the sequence composition of the backward K-turn structure as well as the interaction with RBPs might contribute to the folding of bktRNAs.

We further characterized bktRNAs and found that the sizes of these bktRNAs ranged from 20 nt to ~400 nt, with a median length of 40 nt (Fig. 2e). Notably, more than 84% of bktRNAs (342/407) originated from

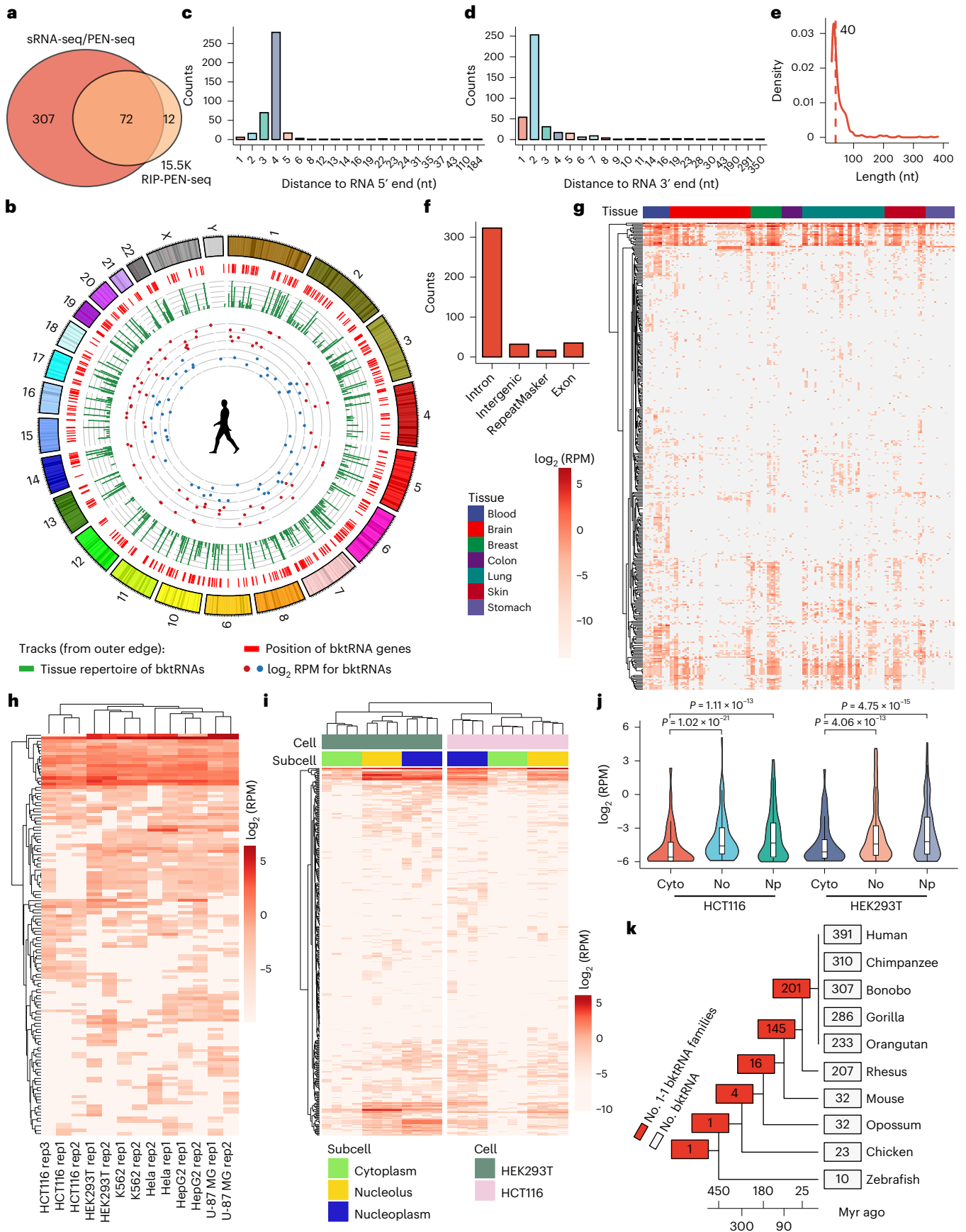


**Fig. 1 | A class of ncrRNAs with a backward K-turn structure.** **a**, Secondary structure of the consensus fktRNA and bktRNA. X represents any nucleotide. The helix 5' to the loop is called the C (canonical) stem (C-stem), and the G•A and A•G base pairs at the 3' side are called the NC-stem. There is a 3-nt internal loop between the C-stem and the NC-stem. In the upper panel, the internal loop and the C-stem located at the 5' end of the RNA compose the fktRNA, whereas the internal loop and the C-stem located at the 3' terminus compose the bktRNA. **b**, Procedure for the construction of RIP-PEN-seq libraries. **c,d**, Histogram showing the numbers of bktRNAs with different distances between the backward K-turn structure and the bktRNA 5' end (c) or 3' end (d). **e**, Enriched motifs identified in bktRNAs by MEME software<sup>48</sup>. Two significant motifs were identified in the bktRNAs. One motif (CUGA) is near the 5' end of the bktRNA, and one motif

(UGAUG) is near the 3' end of the bktRNA. The motif *P* value is estimated based on the *E* value of the log-likelihood ratio. **f**, Gene model of bktRNA1 identified from RIP-PEN-seq data. **g**, Violin plots displaying the SHAPE reactivity across the backward K-turn structure (including 5' NC stem, 5' C stem, Loop, 3' C stem, internal (internal loop) and 3' NC stem), averaged across all bktRNAs (*n* = 22) in the RIP-PEN-SHAPE-MaP data. The box plots indicate the median and the upper and lower quartiles. **h,i**, The predicted secondary structure of the backward K-turn (left panel) and SHAPE reactivity signal (right panel) on bktRNA1 (h) and bktRNA2 (i). The 5' and 3' motifs are indicated with black boxes in the structure figures. The NC-stem and C-stem are marked with black and red underlines in the bar plot, respectively. The SHAPE reactivity is calculated from merged *n* = 4 biological replicates. RT, reverse transcription.

the intron locus (Fig. 2f). Intriguingly, the host genes of the bktRNAs were significantly enriched in various Gene Ontology terms, such as the actin filament-based process and regulation of cell adhesion terms (Extended Data Fig. 3h).

To explore the expression patterns of bktRNAs, we analyzed bktRNA expression levels across cells and tissues as well as subcellular localization. Numerous bktRNAs appeared to be specifically expressed across human tissues or cells (Fig. 2g,h and Extended Data Fig. 3i).



**Fig. 2 | Complex expression patterns of bktRNAs in tissues, cell lines and subcellular compartments and their evolution patterns.** **a**, Overlap analysis of bktRNAs identified from 15.5K RIP-PEN-seq and PEN-seq experiments or published data. **b**, Circos plot illustrating the genomic positions, tissue expression repertoires, expression levels and evolutionary conservation of bktRNAs. The plot legend is shown in the lower panel. Red,  $\log_2$  reads per million reads (RPM) > 0; blue,  $\log_2$  RPM < 0. **c,d**, Histogram showing the numbers of bktRNAs with different distances between the backward K-turn structure and the bktRNA 5' end (**c**) or 3' end (**d**). **e**, Length distribution of all bktRNAs. The dashed line shows the median length of bktRNAs. **f**, Numbers of bktRNAs in different genomic annotation categories. **g**, Heat map showing the expression profiles of bktRNAs in various human tissues using sRNA-seq data from ENCODE.

The expression levels of bktRNAs in cells were categorized into corresponding tissues. Cells/tissues with at least 10 datasets were reserved. **h**, Heat map of bktRNA expression in PEN-seq libraries. **i**, Expression profiles of bktRNAs in different subcellular regions (cytoplasm, nucleolus and nucleoplasm) in HCT116 and HEK293T cells. **j**, Violin plots displaying the expression intensity distribution of bktRNAs ( $n = 355$ ) in different subcellular regions in HCT116 and HEK293T cells. The box plots indicate the median and the upper and lower quartiles. Cyto, cytoplasm; No, nucleolus; Np, nucleoplasm. The  $P$  value between the two categories were determined by the two-sided Mann–Whitney–Wilcoxon test. **k**, Simplified phylogenetic trees of human bktRNAs. Internal branches and roots, numbers of 1-1 orthologous bktRNA families for the indicated species. Tree tips, bktRNA numbers for each species.

For example, the *bktRNA198* and *bktRNA126* genes were substantially expressed in lung tissues (Extended Data Fig. 3i). Moreover, we applied our PEN-seq method in subcellular RNA fractions (sub-PEN-seq) and discovered that most bktRNAs exhibited nuclear-specific localization patterns (Fig. 2i,j and Supplementary Table 4).

We next reconstructed homologous families based on sequence similarity, and we inferred a stringent minimum evolutionary age of bktRNAs. We discovered that most (145/391, 37.1%) bktRNA families were primate specific, but 16 (4.1%) families likely originated more than 90 million years (Myr) ago, and one (0.26%) family likely originated more than 450 Myr ago (Fig. 2k). Interestingly, we found that *bktRNA1*, which is located in an intron of *CWF19L1*, was the most conserved among all bktRNAs (Fig. 2k) and is an endogenous chimeric ncRNA composed of a backward K-turn domain and an H/ACA domain (SNORA12) (Extended Data Fig. 4a–c, Supplementary Figs. 2 and 3 and Supplementary Note 1). Notably, the SHAPE reactivity signals from RIP-PEN-SHAPE-MaP experiments also confirmed this conservative secondary structure of *bktRNA1* (Extended Data Fig. 4b,c). Moreover, we discovered a K-turn-like structure that formed a helix–internal-loop–helix and did not have G•A and A•G base pairs in *bktRNA1* (Extended Data Fig. 4b,c and Supplementary Note 1). In summary, we have confidently identified a large number of bktRNAs with complex folding, expression patterns and sequence conservation.

### U12 snRNA is the direct target of *bktRNA1*

To identify direct targets of bktRNAs, we performed ultraviolet cross-linking, ligation and sequencing of hybrids (CLASH) experiments for 15.5K (Fig. 3a). By analyzing the chimeric reads in the 15.5K CLASH sequencing data (Supplementary Fig. 4a), we found that *bktRNA1* interacted with U12 snRNA and formed a duplex containing 14 perfect base pairs (Fig. 3b,c and Supplementary Fig. 4a,b). Moreover, PARIS data in HEK293T cells<sup>29</sup> also showed that *bktRNA1* and U12 formed an RNA duplex in vivo (Fig. 3d and Supplementary Fig. 5a,b). We also found that the evolutionary conservation of the pairing region between *bktRNA1* and U12 is very highly conserved across 100 vertebrates (Fig. 3e,f and Extended Data Fig. 5a). In addition, the fluorescence in situ

hybridization (FISH) and immunofluorescence (IF) experiments confirmed that *bktRNA1* and U12 snRNA showed co-localization with 15.5K proteins (Extended Data Fig. 5b,c). These results reveal that U12 snRNA is the target of *bktRNA1* bound by the 15.5K protein.

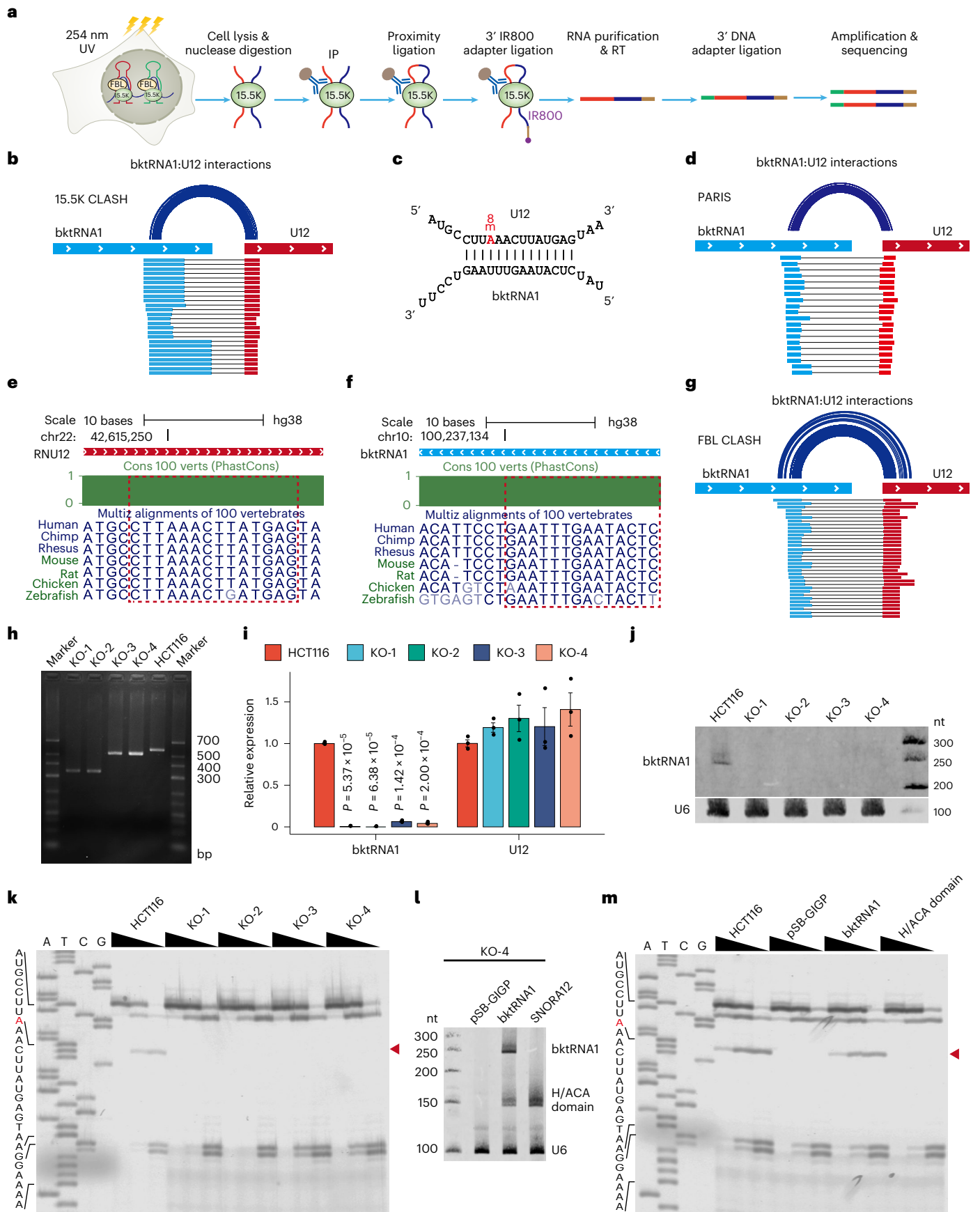
### *bktRNA1* is indispensable for 2'-O-methylation of U12 snRNA

As a known 2'-O-methylation site<sup>30</sup> is located within the complementary region between U12 snRNA and *bktRNA1*, we hypothesized that *bktRNA1* guides the 2'-O-methyltransferase FBL to introduce 2'-O-methylation of U12 snRNA. To test this hypothesis, we first developed an infrared primer extension (irPE) method (Methods) and confirmed that the A8 residue in U12 snRNA is 2'-O-methylated in human cells (Supplementary Fig. 6a–c). The known 2'-O-methylated G22 site and two new sites (Um2 and Am21) were also identified by the irPE method (Supplementary Fig. 6b,c). In addition, we identified 34 *bktRNA1*:U12 chimeras in the FBL CLASH sequencing data (Fig. 3g and Supplementary Fig. 7a,b). IF and FISH experiments further confirmed that *bktRNA1* co-localized with FBL (Supplementary Fig. 8a,b). Moreover, knockdown of FBL in HCT116 cells decreased the level of 2'-O-methylation of A8 (Am8) in U12 snRNA as well as the level of 2'-O-methylation of G22, which is installed by FBL<sup>30</sup> (Supplementary Fig. 8c–g). These results suggest that *bktRNA1* might guide FBL to introduce a conserved 2'-O-methylation of the A8 residue in U12 snRNA.

To investigate *bktRNA1* participation in the 2'-O-methylation of U12 snRNA, we knocked out *bktRNA1* in human HCT116 cells via CRISPR–Cas9. We generated four knockout (KO) clones, including the KO-1 and KO-2 clones with deletion of full-length *bktRNA1* and the KO-3 and KO-4 clones with deletion of the guide region and K-turn domains at the 3' end of *bktRNA1* (Fig. 3h and Supplementary Fig. 9a–c). Quantitative polymerase chain reaction (qPCR) and northern blotting analyses failed to detect accumulation of *bktRNA1* in any of the four KO cell lines, and the expression of the U12 snRNA was not affected (Fig. 3i,j). Likewise, correct processing and efficient accumulation of *CWF19L1* mRNA in *bktRNA1*-KO cells were verified by RT–PCR and Sanger sequencing (Supplementary Fig. 9d,e). The 2'-O-methylation state of U12 at A8 was monitored by the irPE method with RNA extracted

**Fig. 3 | *bktRNA1* guides the FBL protein to introduce 2'-O-methylation in U12 snRNA.** **a**, Schematic of 15.5K infrared-CLASH technology. **b**, 15.5K-CLASH technology was used to identify the interactions between *bktRNA1* and U12 snRNA. The corresponding chimeric reads are marked with different colors; the blue parts of the chimeric reads are located within *bktRNA1*, and the red parts of chimeric reads are located within U12 snRNA. The curved lines connecting the two RNAs indicate the base pairings of the corresponding nucleotides. **c**, Base pairing between *bktRNA1* and U12 snRNA identified from CLASH data. A8 is the 2'-O-methylated site (red) in U12. **d**, High-throughput sequencing data from PARIS experiments were used to identify the interactions between *bktRNA1* and U12 snRNA. **e,f**, Evolutionary conservation (from human to zebrafish) of the interaction region in U12 (**e**) and *bktRNA1* (**f**). The base-pairing regions are indicated with red dashed boxes. **g**, The FBL-CLASH approach was used to identify the interactions between *bktRNA1* and U12 snRNA. **h**, PCR analyses

of genomic DNA obtained from *bktRNA1* KO cells to validate *bktRNA1* biallelic deletion clones. The DNA size marker is shown in base pairs. **i**, qPCR for the relative expression of *bktRNA1* and U12 in *bktRNA1*<sup>-/-</sup> and WT HCT116 cells. Data are presented as mean values  $\pm$  s.e.m. ( $n = 3$  biological replicates), two-tailed, paired  $t$ -test. **j**, Northern blotting of *bktRNA1* in *bktRNA1*<sup>-/-</sup> and WT HCT116 cells. U6 snRNA served as a loading control. **k**, irPE analysis of U12 in *bktRNA1*<sup>-/-</sup> and WT HCT116 cells. The black triangle represents the presence of decreasing amounts (0.5 mM, 0.05 mM and 0.01 mM) of dNTPs. The representative RNA sequence is shown in the left panel, and the red nucleotides denote 2'-O-methylation sites detected by irPE. The red triangle in the right panel indicates the primer extension stop signal. **l**, irNorthern blotting of *bktRNA1* in *bktRNA1*<sup>-/-</sup> and *bktRNA1*-re-expressing cells. U6 snRNA served as a loading control. **m**, irPE analysis of U12 in *bktRNA1*<sup>-/-</sup> and *bktRNA1*-re-expressing cells. Cons, conservation; RT, reverse transcription; verts, vertebrates.



from wild-type (WT) and KO cells (Fig. 3k). Comparing with WT HCT116 cells, the primer extension stop signal was absent from A8 in KO cells, indicating that 2'-O-methylation of A8 was completely inhibited in the absence of bktRNA1 (Fig. 3k).

We further restored bktRNA1 expression in KO-4 cells (Supplementary Fig. 9f,g) and found that stable expression of full-length bktRNA1, but not partial bktRNAs, in the KO cells (Fig. 3l) largely restored the primer extension stop signals at 1 nt downstream of A8 (Fig. 3m), indicating that ectopically expressed bktRNA1 fully restored A8 methylation of U12 snRNA. Additionally, bktRNA1 mutants with mutation or truncation of the backward K-turn domain, as well as substitution in the residue that directly interacts with the A8 residue of U12 snRNA, could completely inhibit or significantly diminish A8 methylation of U12 snRNA (Supplementary Figs. 10 and 11 and Supplementary Note 2). Moreover, artificial guide RNAs that contained the guide sequences of bktRNA1 could induce 2'-O-methylation of A8 in U12 snRNA (Supplementary Fig. 12 and Supplementary Note 2). Collectively, these results demonstrate that bktRNA1 is required for efficient 2'-O-methylation of U12 snRNA.

### Depletion of bktRNA1 causes dysregulation of U12-type introns

Our finding that bktRNA1 is responsible for A8 2'-O-methylation in U12 snRNA raised the question of whether bktRNA1 and its modification are required for the splicing of U12-type introns in human cells. To explore this, we performed strand-specific mRNA-seq in all four bktRNA1-KO cell lines (Extended Data Fig. 6a). Global analysis of aberrant intron retention revealed substantially greater retention of U12-type introns in bktRNA1<sup>-/-</sup> cells than in WT cells (Fig. 4a–d and Supplementary Table 5). Depletion of bktRNA1 affected more than 75% of U12-type introns (Fig. 4e); at least 37% of the retained introns had significant changes ( $P < 0.05$ ; Extended Data Fig. 6b). In contrast, depletion of bktRNA1 had a negligible effect on U2-type introns (Extended Data Fig. 6b). Moreover, compared to U2-type introns, U12-type introns showed global and considerable increases in inclusion level in all four bktRNA1<sup>-/-</sup> cell lines (Fig. 4f–i). Furthermore, these retained U12-type introns exhibited striking overlap across the four bktRNA1-KO cell lines (Fig. 4j). Intriguingly, differential expression analysis revealed that the aberrant splicing of U12-type introns did not affect the expression levels of mature RNAs, consistent with the findings of a previous study<sup>31</sup> (Extended Data Fig. 6c). These results suggest that loss of bktRNA1 has a global effect on the splicing of U12-type introns.

To validate the aberrant U12-type intron retention identified by RNA-seq, we applied qPCR and RT-PCR to measure splice products in bktRNA1-KO and control cells. As expected, the abundance of all 17 randomly selected U12-type introns was upregulated in bktRNA1<sup>-/-</sup> cells as compared to WT cells and the abundance of the U2-type intron in *GAPDH* genes (Fig. 4k–m and Extended Data Fig. 6d). Notably, re-expression of bktRNA1 or artificial guide RNAs (ART7 and ART10) in KO cells restored correct U12-type splicing (Extended Data Fig. 6e,f). In contrast, restoration of the H/ACA domain (SNORA12) of bktRNA1 did not rescue the deficiency in minor splicing (Extended Data Fig. 6g).

These data strongly suggest that bktRNA1 and Am8 in U12 snRNA are important for the fidelity of U12-type splicing in human cells.

### bktRNA1 modulates the recruitment of ZCRB1 to minor spliceosome

To examine whether the loss of bktRNA1 interferes with the assembly of minor spliceosome components, we performed a RIP assay in which RNA–protein complexes were immunoprecipitated using antibodies against U12-associated proteins (ZCRB1, PDCD7, RNPC3 and SNRNP48)<sup>32</sup> (Fig. 5a and Extended Data Fig. 7a). The amount of snRNAs (U11, U12, U4atac, U5 and U6atac) involved in minor spliceosomes in the co-precipitate were then measured by qPCR and northern blotting in WT and bktRNA1 KO cells. Compared to that in WT cells, we found that the ZCRB1-antibody-bound complex in KO cells had obviously lower amounts of U11 and U12 snRNAs (Fig. 5b and Extended Data Fig. 7b,c). Additionally, these differences were not due to the relative amounts of proteins because no changes in the protein levels of ZCRB1 were observed upon knockout of bktRNA1 (Fig. 5c). Overexpression of bktRNA1 in KO cells largely increased the enrichment of the U11 and U12 snRNAs in the ZCRB1 complex (Fig. 5d,e and Extended Data Fig. 7d), indicating that ectopically expressed bktRNA1 fully restored the interaction of ZCRB1 with U11 and U12. We next performed U12 chromatin immunoprecipitation by RNA purification (ChIRP) assays followed by western blotting to verify the interaction of ZCRB1 and U12 in vivo (Fig. 5f and Supplementary Fig. 13a,b). In ChIRP analyses, U12 probes retrieved more ZCRB1 in bktRNA1 WT cells than in KO cells (Fig. 5f, upper panel). Conversely, re-expression of bktRNA1 in KO cells markedly increased the amounts of ZCRB1 retrieved by U12-specific probes (Fig. 5f, lower panel).

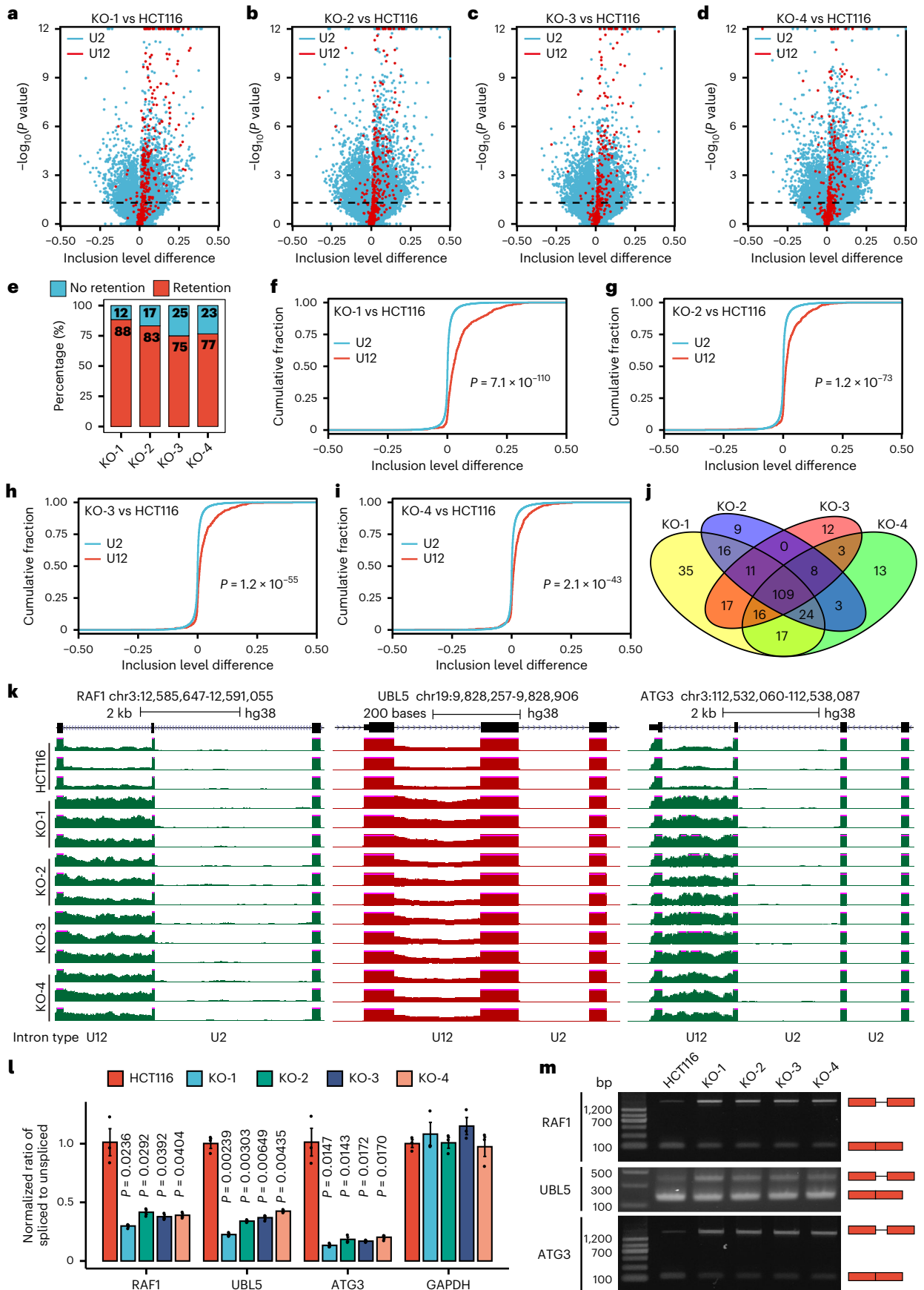
We further performed RNA affinity chromatography followed by immunoblotting to validate the proteins that interact with methylated U12 probes. We discovered that ZCRB1 selectively bound to 2'-O-methylated U12 RNA baits compared to the unmethylated control (Fig. 5g). Similarly to endogenous proteins, the purified recombinant ZCRB1 proteins (Supplementary Fig. 13c,d) also preferentially bound to the 2'-O-methylated U12 probe (Methyl-U12,  $K_d = 340.3 \pm 48.9$  nM) over the unmethylated one (Unmethyl-U12,  $K_d = 2,293.0 \pm 358.4$  nM) in vitro (Fig. 5h). We further performed an RNA EMSA analysis with mutant U12 probes, which have a change of A to G (UUGA) in the consensus sequence recognized by ZCRB1 (refs. 33,34), and found that recombinant ZCRB1 exhibited higher affinity toward the mutated U12 probe with methylated guanine ribonucleotide (Methyl-mutU12,  $K_d = 1,787.0 \pm 211.4$  nM) than unmethylated probes (Unmethyl-mutU12,  $K_d = 13,029.0 \pm 4,687.0$  nM) (Supplementary Fig. 13e).

We then investigated the minor splicing dysregulation and phenotypic changes upon ZCRB1 knockdown in HCT116 cells (Extended Data Fig. 8a). The high similarity between the bktRNA1-KO and ZCRB1-knockdown cells for inducing minor intron retention and functional phenotypes further indicates a close connection between bktRNA1 and ZCRB1 (Extended Data Fig. 8b–m, Supplementary Table 6 and Supplementary Note 3). Together, these results suggest that the 2'-O-methylation at A8 in U12 snRNA guided by bktRNA1 is crucial

### Fig. 4 | Dysregulation of U12-type intron splicing in bktRNA1-depleted cells.

**a–d**, Dot plots displaying the intron retention levels in a representative pairwise analysis of bktRNA1 WT HCT116 cells versus bktRNA1-deficient cells, including KO-1 (**a**), KO-2 (**b**), KO-3 (**c**) and KO-4 (**d**) cells. The red dots represent U12-type introns and the blue dots represent U2-type introns. **e**, Proportion of aberrantly retained U12-type introns in bktRNA1-deficient cells. The red boxes represent retained introns, and the blue boxes represent unretained introns. **f–i**, Cumulative fraction of the inclusion level difference between U12-type and U2-type introns in WT and bktRNA1-deficient cells, including KO-1 (**f**), KO-2 (**g**), KO-3 (**h**) and KO-4 (**i**) cells. The  $P$  values on the cumulative plots of inclusion level differences were calculated using a two-sided Mann–Whitney–Wilcoxon test. **j**, Venn diagram showing the numbers of overlapping retained introns across

four bktRNA1-deficient HCT116 cell lines. **k**, Coverage of RNA-seq reads from WT HCT116 and bktRNA1-deficient cells in the corresponding U12-type and U2-type introns of protein-coding genes (*RAF1*, *UBL5* and *ATG3*). The read counts were normalized using the reads per kilobase per million mapped reads (RPKM) values and are shown using an identical scale for all samples. The intron types are indicated in the bottom panel. **l**, The ratio of spliced to unspliced pre-mRNA for U12-type (*RAF1*, *UBL5* and *ATG3*) and U2-type (*GAPDH*) introns in bktRNA1-deficient and WT HCT116 cells was determined by qPCR. Data are presented as mean values  $\pm$  s.e.m. ( $n = 3$  biological replicates), two-tailed, paired  $t$ -test. **m**, RT-PCR analysis of *RAF1*, *UBL5* and *ATG3* U12-type intron retention in bktRNA1-deficient and WT HCT116 cells. The schemes of the spliced forms are shown at the right. The boxes indicate exons flanking the minor intron.





for the recruitment of ZCRB1 to the U11–U12 di-snRNP complex as well as for the splicing of U12-type introns (Fig. 5i).

### bktRNAs regulate the splicing of local introns

Given that bktRNAs are universally located within introns and that their binding partner 15.5K can promote the assembly of minor and major spliceosomes<sup>7</sup>, we hypothesized that bktRNAs might alternatively be involved in regulating intron processing and RNA splicing. To test this hypothesis, we performed PEN-seq and mRNA-seq in 15.5K-knockdown cells (Supplementary Fig. 14a,b). As expected, the expression of bktRNAs was markedly decreased in 15.5K-knockdown cells compared to control cells (Supplementary Fig. 14c,d and Supplementary Table 7). Notably, intron retention analysis from the mRNA-seq datasets showed that the silencing of 15.5K resulted in the global alteration of intron splicing (Supplementary Fig. 14e,f and Supplementary Table 8). Moreover, the inclusion level was considerably different between bktRNAs and the remaining introns upon 15.5K knockdown (Supplementary Fig. 14g,h).

We further validated the splicing efficiency of local introns containing bktRNA1 in WT and KO cells by qPCR and found that the depletion of bktRNA1 resulted in downregulation of local intron splicing but did not affect the distant intron splicing (Fig. 6a). Moreover, the impaired splicing efficiency of local introns could not be rescued by re-expression of WT bktRNA1 in either bktRNA1 KO-1 or KO-4 cells (Fig. 6b), implying that bktRNA1 might regulate the local splicing of introns by sequence elements of bktRNA1 but not the specific bktRNA1 transcript, which may be similar to some long non-coding RNAs (lncRNAs)<sup>35</sup>.

To test the role of bktRNAs in the local regulation of intron splicing, we constructed GFP reporters that harbor bktRNAs with WT or backward K-turn motif mutants (mutated from CUGA to CUAG, Mut) (Fig. 6c). Consistent with the local regulation of intron splicing by bktRNA1, WT bktRNA1 in the GFP reporter intron also promoted the splicing of local introns, whereas mutation of the 5' motif (bktRNA1-MM) or deletion of the backward K-turn motif significantly decreased the splicing of local introns (Fig. 6d). By applying this GFP reporter to ten randomly selected bktRNAs, we found that eight bktRNAs can regulate the splicing of local introns (Fig. 6d). For example, WT bktRNA168 enhanced the local splicing in comparison to the mutant type (Fig. 6d), whereas bktRNA2 with a backward K-turn-mutant increased the splicing of local introns (Fig. 6d). These results suggest that the backward K-turn motif of bktRNAs is important for the splicing of local introns.

To confirm the local regulation of bktRNAs in intron splicing *in vivo*, we mutated the 5' motif (from CTGA to CTAG) of the backward K-turn structure in *bktRNA161* with prime editing<sup>36</sup> (Supplementary Fig. 15a). We generated seven homozygous mutants (from mut161-1 to mut161-7) based on Sanger sequencing (Supplementary Fig. 15b). In line with the GFP reporter assays (Fig. 6d), the splicing of the local

intron containing bktRNA161, compared to the distant intron, was enhanced upon mutation of the backward K-turn structural motif of bktRNA161 in all mutants (Fig. 6e). Moreover, the binding affinity of the bktRNA161-containing intron with the 15.5K protein was substantially reduced in bktRNA161 mutant HEK293T cells (Fig. 6f and Supplementary Fig. 15c). Together, these results strongly support that bktRNAs are involved in the local regulation of intron splicing in a backward K-turn motif-dependent manner (Fig. 6g).

## Discussion

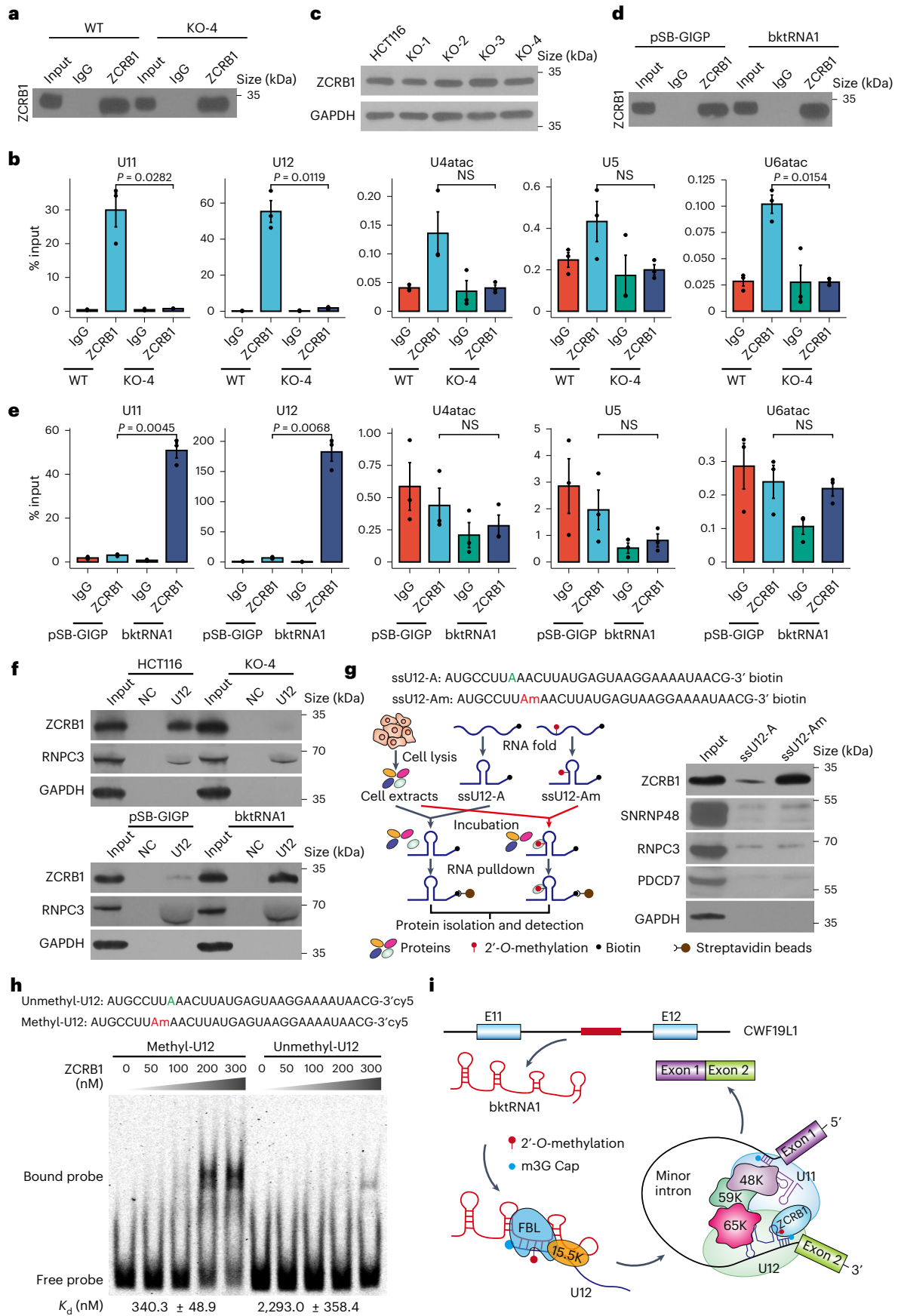
In this study, we discovered a class of bktRNAs that act as binding partners for the K-turn binding protein 15.5K by developing the RIP-PEN-seq method, and we revealed that bktRNA1 can act as a post-transcriptional regulator of minor splicing and that bktRNAs can regulate the splicing of local introns. Although U12-type introns constitute ~0.5% of the introns in the human genome, they exist in approximately 700–800 genes involved in various cellular processes<sup>37,38</sup>. Mutations in genes encoding the components of the U12 spliceosome, including both snRNAs and interacting proteins, have been shown to cause a series of developmental disorders in humans<sup>39–43</sup>. For example, mutation in U12 is associated with cerebellar ataxia<sup>44</sup>. These findings raise an interesting question: whether aberrant expression of bktRNA1 could lead to some of the abovementioned developmental disorders in humans. Coincidentally, mutation of *CWF19L1*, the host gene of *bktRNA1*, has been considered as cause of autosomal recessive cerebellar ataxia (ARCA)<sup>45–47</sup>, consistent with the phenotype caused by the mutation in U12 snRNAs<sup>44</sup>. Given that the loss of bktRNA1 leads to global splicing dysregulation of U12-type introns, we hypothesized that mutation of *CWF19L1* might affect the expression and processing of bktRNA1 to cause ARCA disease. Therefore, future works might address whether *bktRNA1* is associated with ARCA.

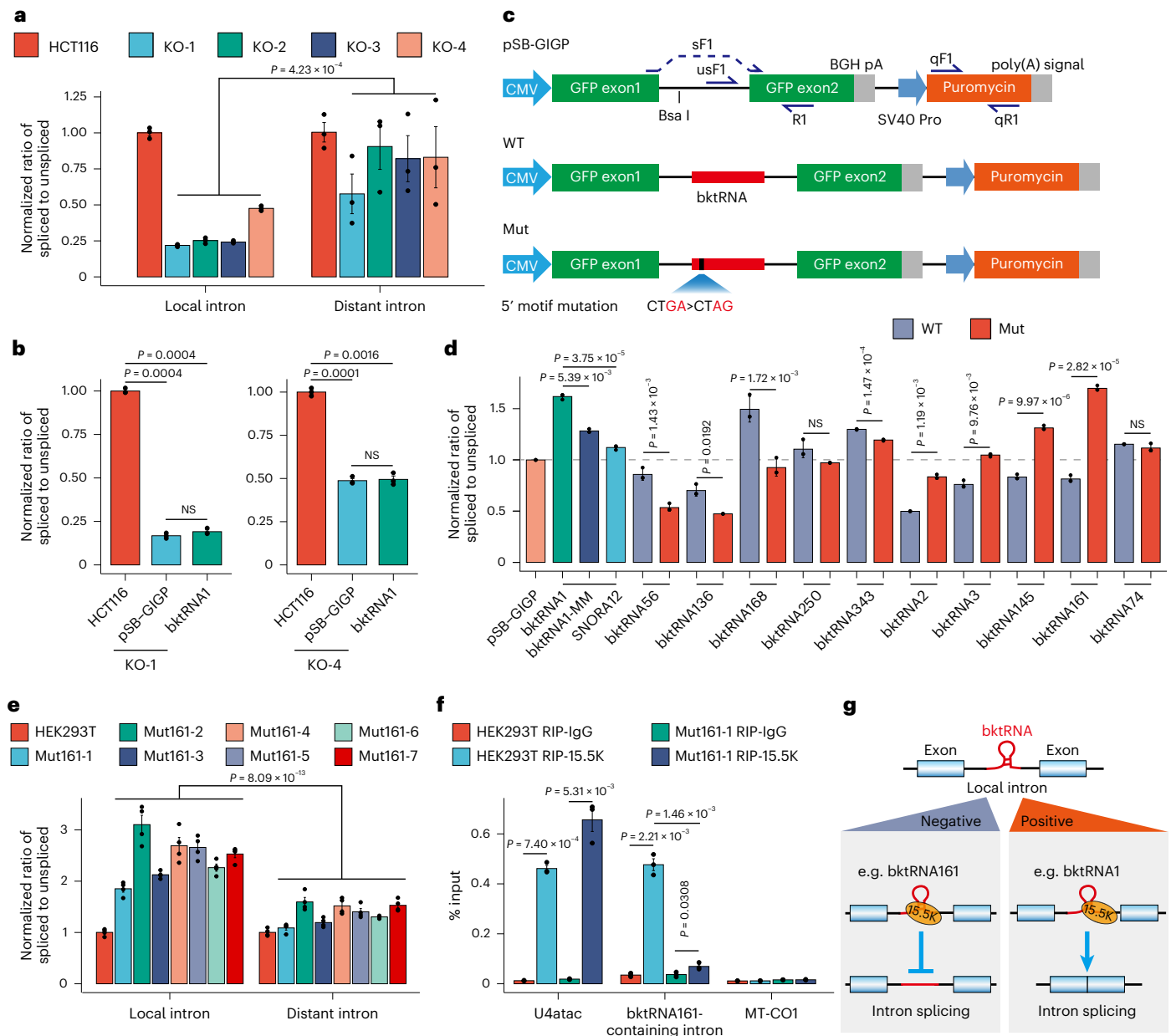
Although backward K-turn structural motifs have been identified in few ncRNAs (for example, 23S rRNA from *Archaea*)<sup>2</sup>, these motifs are not composed of consensus sequences and are located within different positions of ncRNAs. In contrast to these few ncRNAs<sup>2</sup>, our bktRNAs identified from mammals all have two consensus sequence elements located within backward K-turn motifs. Importantly, these two consensus sequence elements are located at fixed positions in bktRNAs (Fig. 1c–e). Moreover, we found that these backward K-turn structural motifs of bktRNAs are indispensable for avoiding degradation by exonucleases, the processing and maturation of bktRNAs (Supplementary Fig. 10 and Supplementary Note 2) and the local regulation of intron splicing (Fig. 6).

We used the bktRNA1 gene as a model to explore the regulatory function of bktRNAs (Fig. 5i). Moreover, the consensus structural motifs of bktRNAs and their dynamic, cell-specific/tissue-specific expression patterns have shed light on their roles in gene expression regulation. We found that bktRNA1 and other bktRNAs are involved in the local regulation of intron splicing (Fig. 6g). Notably, the backward

**Fig. 5 | Depletion of bktRNA1 affects U12 interactions with ZCRB1.** **a**, Western blots showing precipitation with an anti-ZCRB1 antibody in WT and bktRNA1-deficient KO-4 cells. **b**, Native RIP was performed in WT and KO-4 cells using an anti-ZCRB1 antibody or normal IgG antibody, after which qPCR was performed with primers recognizing snRNAs involved in minor spliceosomes. The percentage of RIP-enriched snRNAs was calculated relative to the input RNA. Data are presented as mean values  $\pm$  s.e.m. ( $n = 3$  biological replicates), two-tailed, paired  $t$ -test. NS, no significance. **c**, Western blotting analysis of ZCRB1 in HCT116 WT and bktRNA1<sup>-/-</sup> clones. GAPDH served as the loading control. **d**, Detection of ZCRB1 in immunoprecipitates from KO-4 and bktRNA1-rescued (bktRNA1) cells by western blotting analysis. **e**, Native RIP was performed in bktRNA1-deficient and bktRNA1-rescued cells using an anti-ZCRB1 antibody or normal IgG antibody followed by qPCR with primers for snRNAs involved in minor spliceosomes. The percentage of RIP-enriched snRNAs was calculated relative to the input RNA. Data are presented as mean values  $\pm$  s.e.m. ( $n = 3$  biological replicates),

two-tailed, paired  $t$ -test. **f**, Western blotting analyses of proteins retrieved by the U12-ChIRP experiments. RNPC3 was used as a positive control, and GAPDH was used as a negative control. NC, negative control probes. **g**, Validation of 2'-*O*-methylated U12-specific binding proteins by RNA affinity chromatography using single-stranded RNA probes with 2'-*O*-methylated (red) or unmethylated (green) adenosine. A schematic workflow (left panel) and western blotting analyses (right panel) of RNA affinity chromatography were used to validate U12-interacting proteins (ZCRB1, RNPC3, PDCD7 and SNRNP48). GAPDH was used as a negative control. **h**, RNA EMSA was used to assess the binding of recombinant ZCRB1 and the indicated U12 RNA probes. The dissociation constants ( $K_d$ , nM) of recombinant ZCRB1 proteins with methylated (Methyl-mutU12) or unmethylated (Unmethyl-mutU12) U12 ssRNA probes are indicated in the lower panel. **i**, Proposed model showing the functions and mechanisms of bktRNA1. 65K is RNPC3; 59K is PDCD7; and 48K is SNRNP48.





**Fig. 6 | Local regulation of intron splicing by bktRNAs.** **a**, The ratio of spliced to unspliced introns containing bktRNA1 (local intron) and another intron from the same host gene (distant intron in CWF19L1) was determined by qPCR in bktRNA1-deficient and WT HCT116 cells. Data are presented as mean values  $\pm$  s.e.m. ( $n = 3$  biological replicates), two-tailed, paired  $t$ -test. **b**, Ratio of spliced to unspliced intron (local) in bktRNA1-rescued HCT116 KO-1 (right panel) and KO-4 (left panel) cells. Data are presented as mean values  $\pm$  s.e.m. ( $n = 3$  biological replicates), two-tailed, paired  $t$ -test. **c**, Schematic of the GFP splicing reporter design. bktRNA cloning and primers were used for splicing efficiency detection. We used the primer pairs sF1/R1 and usF1/R1 to detect mature and precursor GFP, respectively. The analysis of puromycin with primer pairs qF1/qR1 acted as an internal control. **d**, qPCR analysis of the ratio of spliced to unspliced GFP RNA in HEK293T cells transfected with WT and mutant bktRNAs (mutated from CTGA to CTAG). Data are presented as mean values  $\pm$  s.e.m. ( $n = 3$  biological replicates), two-tailed,

paired  $t$ -test. **e**, The ratio of spliced to unspliced intron containing bktRNA161 (local intron) and another intron from the same host gene (distant intron in TEAD1) was determined by qPCR in bktRNA161-mutant and WT HEK293T cells. Data are presented as mean values  $\pm$  s.e.m. ( $n = 4$  biological replicates), two-tailed, paired  $t$ -test. **f**, Quantification of introns containing bktRNA161 by qPCR in 15.5K immunoprecipitates from WT and mutant bktRNA161 HEK293T cells. U4atac and MT-CO1 were kept as positive and negative controls, respectively. Data are presented as mean values  $\pm$  s.e.m. ( $n = 3$  biological replicates), two-tailed, paired  $t$ -test. **g**, Model of local regulation of intron splicing by bktRNAs. The interaction between the backward K-turn structure of bktRNAs and 15.5K negative (for example, bktRNA161) or positive (for example, bktRNA1) regulates the splicing of the local intron containing bktRNAs. In contrast, disruption of the backward K-turn structure in bktRNA161 or bktRNA1 relieves the local regulation of intron splicing.

K-turn structural motifs are indispensable for the local regulation of intron splicing by bktRNAs (Fig. 6). Interestingly, bktRNAs can act not only as positive regulators but also as negative regulators of the local splicing of introns (Fig. 6g), suggesting that bktRNAs may interact with different factors to regulate the splicing of local introns.

Thus, it would be very interesting to explore any cofactors (for example, splicing factors) that control the local regulation of intron splicing by these bktRNAs.

In summary, our data strongly suggest that bktRNAs form a class of post-transcriptional regulators that probably help RBPs recognize

substrates through base-pairing interactions (for example, bktRNA1) and generally function in modulating their binding partner 15.5K to regulate the splicing of local introns. Our studies reveal that a single methylation alteration caused by the loss of bktRNA1 contributes to global splicing dysregulation and that bktRNAs can act as local regulators in the splicing of introns. Overall, these discoveries add another layer of complexity to the control of gene expression that involves crosstalk among RNA splicing, RNA methylation and bktRNAs.

## Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41587-023-01749-0>.

## References

- Butcher, S. E. & Pyle, A. M. The molecular interactions that stabilize RNA tertiary structure: RNA motifs, patterns, and networks. *ACC Chem. Res.* **44**, 1302–1311 (2011).
- Klein, D. J., Schmeing, T. M., Moore, P. B. & Steitz, T. A. The kink-turn: a new RNA secondary structure motif. *EMBO J.* **20**, 4214–4221 (2001).
- Lilley, D. M. The K-turn motif in riboswitches and other RNA species. *Biochim. Biophys. Acta* **1839**, 995–1004 (2014).
- Schroeder, K. T., McPhee, S. A., Ouellet, J. & Lilley, D. M. A structural database for k-turn motifs in RNA. *RNA* **16**, 1463–1468 (2010).
- Rozhdetsvensky, T. S. et al. Binding of L7Ae protein to the K-turn of archaeal snoRNAs: a shared RNA binding motif for C/D and H/ACA box snoRNAs in Archaea. *Nucleic Acids Res.* **31**, 869–877 (2003).
- Baird, N. J., Zhang, J., Hamma, T. & Ferré-D'Amaré, A. R. YbxF and YlxQ are bacterial homologs of L7Ae and bind K-turns but not K-loops. *RNA* **18**, 759–770 (2012).
- Nottrott, S. et al. Functional interaction of a novel 15.5kD [U4/U6.U5] tri-snRNP protein with the 5' stem-loop of U4 snRNA. *EMBO J.* **18**, 6119–6133 (1999).
- Szewczak, L. B. W., DeGregorio, S. J., Strobel, S. A. & Steitz, J. A. Exclusive interaction of the 15.5 kD protein with the terminal box C/D motif of a methylation guide snoRNP. *Chem. Biol.* **9**, 1095–1107 (2002).
- Szewczak, L. B. W., Gabrielsen, J. S., Degregorio, S. J., Strobel, S. A. & Steitz, J. A. Molecular basis for RNA kink-turn recognition by the h15.5K small RNP protein. *RNA* **11**, 1407–1419 (2005).
- Chawla, M., Oliva, R., Bujnicki, J. M. & Cavallo, L. An atlas of RNA base pairs involving modified nucleobases with optimal geometries and accurate energies. *Nucleic Acids Res.* **43**, 6714–6729 (2015).
- Polikanov, Y. S., Melnikov, S. V., Söll, D. & Steitz, T. A. Structural insights into the role of rRNA modifications in protein synthesis and ribosome assembly. *Nat. Struct. Mol. Biol.* **22**, 342–344 (2015).
- Liu, S., Ghalei, H., Lührmann, R. & Wahl, M. C. Structural basis for the dual U4 and U4atac snRNA-binding specificity of spliceosomal protein hPrp31. *RNA* **17**, 1655–1663 (2011).
- Vidovic, I., Nottrott, S., Hartmuth, K., Lührmann, R. & Ficner, R. Crystal structure of the spliceosomal 15.5kD protein bound to a U4 snRNA fragment. *Mol. Cell* **6**, 1331–1342 (2000).
- Jafarifar, F., Dietrich, R. C., Hiznay, J. M. & Padgett, R. A. Biochemical defects in minor spliceosome function in the developmental disorder MOPD I. *RNA* **20**, 1078–1089 (2014).
- Edery, P. et al. Association of TALS developmental disorder with defect in minor splicing component U4atac snRNA. *Science* **332**, 240–243 (2011).
- He, H. et al. Mutations in U4atac snRNA, a component of the minor spliceosome, in the developmental disorder MOPD I. *Science* **332**, 238–240 (2011).
- Merico, D. et al. Compound heterozygous mutations in the noncoding *RNU4ATAC* cause Roifman Syndrome by disrupting minor intron splicing. *Nat. Commun.* **6**, 8718 (2015).
- Farach, L. S. et al. The expanding phenotype of *RNU4ATAC* pathogenic variants to Lowry Wood syndrome. *Am. J. Med. Genet. A* **176**, 465–469 (2018).
- Lapinaite, A. et al. The structure of the box C/D enzyme reveals regulation of RNA methylation. *Nature* **502**, 519–523 (2013).
- Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A. E. & Weeks, K. M. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat. Methods* **11**, 959–965 (2014).
- Luo, Q.-J. et al. RNA structure probing reveals the structural basis of Dicer binding and cleavage. *Nat. Commun.* **12**, 3397 (2021).
- McPhee, S. A., Huang, L. & Lilley, D. M. A critical base pair in k-turns that confers folding characteristics and correlates with biological function. *Nat. Commun.* **5**, 5127 (2014).
- Huang, L. et al. Structure and folding of four putative kink turns identified in structured RNA species in a test of structural prediction rules. *Nucleic Acids Res.* **49**, 5916–5924 (2021).
- Liu, J. & Lilley, D. M. The role of specific 2'-hydroxyl groups in the stabilization of the folded conformation of kink-turn RNA. *RNA* **13**, 200–210 (2007).
- Huang, L., Wang, J. & Lilley, D. M. A critical base pair in k-turns determines the conformational class adopted, and correlates with biological function. *Nucleic Acids Res.* **44**, 5390–5398 (2016).
- Ashraf, S., Huang, L. & Lilley, D. M. J. Effect of methylation of adenine N<sup>6</sup> on kink turn structure depends on location. *RNA Biol.* **16**, 1377–1385 (2019).
- Huang, L., Ashraf, S., Wang, J. & Lilley, D. M. Control of box C/D snoRNP assembly by N<sup>6</sup>-methylation of adenine. *EMBO Rep.* **18**, 1631–1645 (2017).
- Xuan, J. J. et al. RMBase v2.0: deciphering the map of RNA modifications from epitranscriptome sequencing data. *Nucleic Acids Res.* **46**, D327–D334 (2018).
- Zhipeng, L. et al. RNA duplex map in living cells reveals higher-order transcriptome structure. *Cell* **165**, 1267–1279 (2016).
- Tycowski, K. T., Aab, A. & Steitz, J. A. Guide RNAs with 5' caps and novel box C/D snoRNA-like domains for modification of snRNAs in metazoa. *Curr. Biol.* **14**, 1985–1995 (2004).
- Cologne, A. et al. New insights into minor splicing—a transcriptomic analysis of cells derived from TALS patients. *RNA* **25**, 1130–1149 (2019).
- Will, C. L. et al. The human 18S U11/U12 snRNP contains a set of novel proteins not found in the U2-dependent spliceosome. *RNA* **10**, 929–941 (2004).
- Dominguez, D. et al. Sequence, structure, and context preferences of human RNA binding proteins. *Mol. Cell* **70**, 854–867 e859 (2018).
- Jolma, A. et al. Binding specificities of human RNA-binding proteins toward structured and linear RNA sequences. *Genome Res.* **30**, 962–973 (2020).
- Engreitz, J. M. et al. Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* **539**, 452–455 (2016).
- Anzalone, A. V. et al. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149–157 (2019).
- Alioto, T. S. U12DB: a database of orthologous U12-type spliceosomal introns. *Nucleic Acids Res.* **35**, D110–D115 (2007).
- Olthof, A. M., Hyatt, K. C. & Kanadia, R. N. Minor intron splicing revisited: identification of new minor intron-containing genes and tissue-dependent retention and alternative splicing of minor introns. *BMC Genomics* **20**, 686 (2019).
- Madan, V. et al. Aberrant splicing of U12-type introns is the hallmark of ZRSR2 mutant myelodysplastic syndrome. *Nat. Commun.* **6**, 6042 (2015).

40. Reber, S. et al. Minor intron splicing is regulated by FUS and affected by ALS-associated FUS mutants. *EMBO J.* **35**, 1504–1521 (2016).
41. Verberne, E. A., Faries, S., Mannens, M., Postma, A. V. & van Haelst, M. M. Expanding the phenotype of biallelic *RNPC3* variants associated with growth hormone deficiency. *Am. J. Med. Genet. A* **182**, 1952–1956 (2020).
42. Argente, J. et al. Defective minor spliceosome mRNA processing results in isolated familial growth hormone deficiency. *EMBO Mol. Med.* **6**, 299–306 (2014).
43. Martos-Moreno, G. et al. Response to growth hormone in patients with *RNPC3* mutations. *EMBO Mol. Med.* **10**, e9143 (2018).
44. Elsaid, M. F. et al. Mutation in noncoding RNA *RNU12* causes early onset cerebellar ataxia. *Ann. Neurol.* **81**, 68–78 (2017).
45. Burns, R. et al. Homozygous splice mutation in *CWF19L1* in a Turkish family with recessive ataxia syndrome. *Neurology* **83**, 2175–2182 (2014).
46. Evers, C. et al. Exome sequencing reveals a novel *CWF19L1* mutation associated with intellectual disability and cerebellar atrophy. *Am. J. Med. Genet. A* **170**, 1502–1509 (2016).
47. Nguyen, M. et al. Pathogenic *CWF19L1* variants as a novel cause of autosomal recessive cerebellar ataxia and atrophy. *Eur. J. Hum. Genet.* **24**, 619–622 (2016).
48. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28–36 (1994).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2023

## Methods

### Construction of RIP-PEN-seq and PEN-seq libraries

For conventional RIP-seq, the immunoprecipitated RNAs (especially for RNAs with length  $\geq 50$  nt) were often fragmented and then subjected to RNA-seq library construction with random primer-based reverse transcription<sup>49</sup>. As a result, the conventional RIP-seq method cannot identify the full length of immunoprecipitated RNA, and, thus, conventional RIP-seq cannot discover the precise positions of motifs at RNAs. In this study, we used dual RNA adaptors to capture both ends of ncRNAs. This strategy and the following strategies allowed us to identify the full-length sequences of ncRNAs (from 20 nt to 500 nt) and facilitated the discovery of bktRNAs with characteristic motifs and structures presented at RNA ends.

RIP-PEN-seq and PEN-seq (Fig. 1b and Extended Data Fig. 1a) employed several strategies to achieve high-performance ncRNA full-length sequencing. First, to enable the high ligation efficiency and quantitative full-length library, specific randomized barcode-containing (4N, N represents for A, T or U, G, C) 5' adapters and 3' adapters were synthesized and directly ligated to RNAs. Second, to increase the diversity of RNA species and discover low-expression RNAs, we used RNase H or Ribo-Zero strategies to remove high-abundance rRNAs and/or snRNAs<sup>50,51</sup>. It is worth noting that the Ribo-Zero kit from Illumina had been discontinued, and the nuclease-mediated depletion displays huge biases in sRNA-seq following the traditional procedure<sup>52</sup>. Here, we performed rRNA/snRNA depletion after adaptors ligation and used single-stranded DNA (ssDNA) exonuclease RecJf rather than DNase I to remove the probes. Third, to avoid the widespread mispriming artifacts<sup>53</sup> and obtain full-length cDNAs, we used a nested reverse transcription primer during reverse transcription, along with a full-length reverse PCR primer for amplification. Finally, to overcome the obstacles of RNA modifications and stable RNA secondary structures in cDNA synthesis, we used SuperScript IV Reverse Transcriptase, which has a high level of processivity and thermostability, to generate cDNAs.

RIP experiments were performed using the protocol outlined in a previous report<sup>54,55</sup>. In brief, for each immunoprecipitate,  $1 \times 10^7$  cells stably expressing FLAG-15.5K were pelleted; resuspended in an equal volume of ice-cold polysomal lysis buffer (10 mM HEPES pH 7.0, 100 mM KCl, 5 mM MgCl<sub>2</sub>, 0.5% NP-40) supplemented with 1 mM DTT, 100 U ml<sup>-1</sup> RNase inhibitor (Takara), 1 $\times$  protease inhibitor cocktail (Roche) and 0.4 mM RVC (New England Biolabs (NEB)); incubated on ice for 15 minutes; snap frozen in liquid nitrogen; and stored at  $-80$  °C. After the lysate was thawed on ice and centrifuged at 15,000g for 15 minutes, the supernatant was pre-cleared with Dynabeads Protein G (Invitrogen) at 4 °C for 1 hour. Ten microliters of the supernatant was saved as input and mixed with 1 ml of RNAzol. Subsequently, a 100- $\mu$ l aliquot of the cell extract was diluted with 900  $\mu$ l of NT2 buffer (50 mM Tris pH 7.4, 150 mM NaCl, 1 mM MgCl<sub>2</sub>, 0.05% NP-40) supplemented with 1 mM DTT, 100 U ml<sup>-1</sup> RNase inhibitor (Takara), 1 $\times$  protease inhibitor cocktail (Roche) and 20 mM EDTA. The extracts were mixed with 5  $\mu$ g of an anti-FLAG antibody (Proteintech) or a normal IgG antibody (Millipore) and incubated overnight at 4 °C with end-over-end rotation. The next day, the RNP/antibody complex was precipitated by incubation with Dynabeads Protein G at 4 °C for 3 hours. The beads were collected and washed five times in NT2 buffer. After the final wash, the beads were resuspended in 1 ml of RNAzol and saved as RIP RNA or IgG RNA.

For PEN-seq, total RNA was isolated from cell lysis with RNAzol, treated with RQ1 DNase (Promega) and purified by RNA Clean & Concentrator-5 (Zymo Research). The samples were then ligated to 3' randomized RNA adaptor with T4 RNA ligase 2 truncated KQ (NEB) in 1 $\times$  T4 RNA ligase reaction buffer supplemented with 12.5% PEG 8000 at 16 °C for 18 hours. The excessive adaptors were digested with 100 U of 5' deadenylase (NEB) at 30 °C for 1 hour, incubated with 2  $\mu$ g of *Escherichia coli* single-strand DNA-binding protein (Promega) on ice for 30 minutes and subjected to ssDNA digestion with 60 U of RecJf (NEB) at 37 °C for another hour. The ligated RNA was

ligated to a 5' randomized RNA adaptor using T4 RNA Ligase 1 (NEB) in 1 $\times$  T4 RNA ligase reaction buffer supplemented with 1 mM ATP at 16 °C for 18 hours. The ligated RNA was column-purified by RNA Clean & Concentrator-5 (Zymo Research) and subjected to RNase H-based ribosomal RNA depletion according to a standard protocol with some modifications<sup>51</sup>. For rRNA/snRNA probe digestion, DNase I was replaced by RecJf exonuclease to prevent digestion of the 3' randomized RNA adaptor. After column-based purification, the rRNA/snRNA-depleted RNA was reverse transcribed using SuperScript IV Reverse Transcriptase (Invitrogen) in 1 $\times$  reverse transcription buffer with truncated reverse transcription primer to eliminate mispriming artifacts. cDNA was amplified using Phusion High-Fidelity DNA Polymerase (Thermo Fisher Scientific) in 1 $\times$  Phusion HF buffer (Thermo Fisher Scientific) with primers (the RP1 forward primer and RPI-X reverse primer). The PCR-amplified cDNA was gel purified using a 4% NuSieve 3-1 Agarose gel (Lonza) to remove adapter dimers and sequenced on an Illumina HiSeq X Ten platform with paired-end reads (150 bp or 125 bp) at Annoroad Gene Technology Company. The adaptors and primers are listed in Supplementary Table 9.

### RIP-PEN-SHAPE-MaP library construction

RIP-PEN-SHAPE-MaP was performed on HEK293T-F-15.5K cells as previously described<sup>21</sup> with some modifications. In brief, RIP experiments were carried out as described above until the last NT2 buffer wash. After the last wash, the beads were incubated with SHAPE modification buffer (100 mM HEPES pH 8.0, 6 mM MgCl<sub>2</sub>, 150 mM NaCl, 50 mM NAI-N<sub>3</sub> (MedChemExpress, HY-103006))<sup>56</sup> at 37 °C for 12 minutes on a thermomixer at 1,000 r.p.m. The control group was prepared using DMSO following the same procedure as described above. RNA was extracted using RNAzol, treated with DNase I and purified with RNA Clean & Concentrator-5 (Zymo Research) in accordance with the manufacturer's protocol. Then, approximately 300 ng of NAI-N<sub>3</sub>-treated or DMSO-treated RNA sample was subjected to the PEN-seq procedure with some modifications. (1) The 3' randomized RNA adaptor and 5' randomized RNA adaptor were replaced with randomized adaptors containing six degenerate nucleotides. (2) To increase the read coverage on bktRNAs, we designed probes<sup>57</sup> for RNase H-based high-abundance snoRNA subtraction in addition to rRNA and snRNA depletion. The high-abundance snoRNAs were determined according to the top 1,000 reads in 15.5K RIP-PEN-seq datasets. (3) Reverse transcription was performed using SuperScript II Reverse Transcriptase (Thermo Fisher Scientific) in MnCl<sub>2</sub>-containing 1 $\times$  reverse transcription buffer (50 mM Tris-HCl pH 8.0, 75 mM KCl, 6 mM MnCl<sub>2</sub>, 10 mM DTT)<sup>20</sup>. The whole reaction system in the tube was incubated at 42 °C for 3 hours. (4) For cDNA purification, Exo I (NEB) was used to digest excess reverse transcription primers at 37 °C for 15 minutes. Then, 7  $\mu$ l of 1 M NaOH and 5  $\mu$ l of 0.5 M EDTA per 20- $\mu$ l reaction volume were added to remove RNA templates at 70 °C for 12 minutes, after which cleanup of the cDNA was performed with Oligo Clean & Concentrator (Zymo Research). (5) Library amplification was performed using NEBNext Ultra II Q5 Master Mix (NEB) and was programmed as follows: stage I: 98 °C for 30 s; stage II: 98 °C for 10 s, 65 °C for 75 s, with a limited cycle; stage III: 65 °C for 5 minutes. The number of cycles was 16–20. The adaptors, primers and probes are listed in Supplementary Table 9.

### Mapping of 2'-O-methylation residues

The positions of 2'-O-methyls were identified by dNTP concentration-dependent primer extension (PE) assays. The primer used for U12 primer extension (irPE-U12) was synthesized by Sangon. To label the azide-modified oligonucleotides with IR dye, 0.4 nmol of oligonucleotides was mixed with 10 nmol IRDye 680RD DBCO (LI-COR Biosciences) in 1 $\times$  PBS (Thermo Fisher Scientific) at 25 °C overnight. The IR dye-labeled oligonucleotides were then purified with Oligo Clean & Concentrator kits (Zymo Research). The 5' end labeled primers were extended by SuperScript III Reverse Transcriptase (Invitrogen) in the

presence of decreasing amounts (0.5 mM, 0.05 mM and 0.01 mM) of dNTPs using 30 µg of total RNA to determine modifications in U12 snRNA. Sequence analysis was carried out according to the dideoxy method with a Sequenase Version 2.0 DNA Sequencing Kit (USB). The products were separated on a 10% urea-polyacrylamide gel and visualized on a LI-COR Odyssey Infrared Imager with Odyssey software version 3.0.

### ChIRP

ChIRP was performed according to previous studies<sup>58–60</sup> with some modifications. HCT116 cells were crosslinked with 3% formaldehyde at room temperature for 30 minutes, followed by 125 mM glycine quenching at room temperature for 5 minutes. Cells were centrifuged, and the pellet was washed with ice-cold PBS and then snap frozen in liquid nitrogen and stored at –80 °C. Crosslinked cells ( $\sim 2 \times 10^7$ ) were resuspended with 1 ml of ChIRP pre-digestion buffer (20 mM Tris-HCl pH 7.5, 5 mM MgCl<sub>2</sub>, 0.5 mM CaCl<sub>2</sub>, 0.5% Triton X-100) supplemented with 1× protease inhibitor cocktail (Roche), 1 mM DTT, 100 U ml<sup>-1</sup> RNase inhibitor, 10 mM RVC (NEB) and 12 U of DNase I (Promega) and rotated at 37 °C for 10 minutes. The reaction was stopped with 20 mM EDTA, and cells were pelleted, washed once with 1× PBS and resuspended in ChIRP cell lysis buffer (50 mM Tris-HCl pH 7.0, 10 mM EDTA, 1% SDS) with 1× protease inhibitor cocktail (Roche), 1 mM DTT and 100 U ml<sup>-1</sup> RNase inhibitor. Cells were sonicated using Sonic Vibra-Cell VCX130 for 20 cycles with the following conditions: 30% amplitude, 10 s on, 10 s off. After centrifugation at 16,000g for 10 minutes at 4 °C, the supernatant was collected and pre-cleared with Dynabeads MyOne Streptavidin C1 beads (Thermo Fisher Scientific) for 30 minutes at 37 °C. For hybridization, the pre-cleared cell lysates were diluted in two times volume of ChIRP hybridization buffer (50 mM Tris-HCl, pH 7.0, 750 mM NaCl, 1 mM EDTA, 1% SDS, 15% formamide) with 1× protease inhibitor cocktail (Roche), 1 mM DTT and 100 U ml<sup>-1</sup> RNase inhibitor and incubated with 100 pmol of U12-targeted biotinylated probe mix (ChIRP-U12-1, 2, 3) or negative control probe (ChIRP-NC) at 37 °C overnight. The next day, 100 µl of pre-washed Dynabeads MyOne Streptavidin C1 beads were added and rotated for an additional 30 minutes. The beads were washed five times with ChIRP washing buffer (2× SSC, 0.5% SDS) at 37 °C. A quarter of the ChIRP products were used for RNA isolation, and the remaining were subjected to protein elution. For RNA isolation, input and ChIRP samples were resuspended in ChIRP RNA-isolation buffer (10 mM Tris-HCl pH 7.0, 100 mM NaCl, 1 mM EDTA, 0.5% SDS) supplemented with 1,000 mg ml<sup>-1</sup> proteinase K and incubated at 50 °C for 45 minutes, followed by 95 °C for 15 minutes and RNAzol extraction. Eluted RNA was subjected to ChIRP-qPCR for detection of related transcripts. For protein extraction, input and ChIRP samples were resuspended in ChIRP biotin elution buffer (7.5 mM HEPES pH 7.5, 12.5 mM D-biotin, 1.5 mM EDTA, 75 mM NaCl, 0.15% SDS, 0.075% Sarkosyl, 0.02% Na-deoxycholate), mixed at 25 °C for 20 minutes and at 65 °C for 10 minutes. After collecting the eluents, the beads were eluted again. The two eluents were pooled and precipitated with 20 µg ml<sup>-1</sup> protein carrier (BSA) and four times volume of ice-cold acetone at –20 °C overnight. The next day, proteins were pelleted and dissolved in 1× SDS loading buffer (Beyotime) and boiled at 95 °C for 30 minutes. Final protein samples were subjected to SDS-PAGE for western blotting. The probes used for ChIRP are listed in Supplementary Table 9.

### Expression and purification of ZCRB1 protein

We use the *E. coli* system to produce and purify specific ZCRB1 protein following the HaloTag protein purification system (Promega) with pieces of modifications. In brief, pH6HTN-ZCRB1 vectors that expressed HaloTag fusion ZCRB1 protein were transformed into *E. coli* *Transetta DE3* chemically competent cells (TransGen Biotech). A fresh colony was inoculated into 2YT medium containing the antibiotic and grown at 37 °C for overnight. On the second day, the starter

culture was diluted as 1:100 in fresh 2YT medium with antibiotic and cultured at 37 °C for about 6 hours, reaching the logarithmic growth phase. Then, 1 ml of culture was seeded into 100 ml of 2YT medium with IPTG at 0.4 nM final concentration and grown for 4 hours at 25 °C for harvesting. According to the HaloTag protein purification system instructions, cells in 100 ml of medium were harvested by centrifugation at 4,000g for 10 minutes at 4 °C, and then the cell pellets were resuspended and lysed by sonication (Sonic Vibra-Cell VCX130) in 10 ml of Halo purification buffer (50 mM HEPES pH 7.5, 150 mM NaCl, 0.5 mM EDTA, 0.005% NP-40) plus with 1 mM DTT, complete protease inhibitor cocktail (Roche), 1 mg ml<sup>-1</sup> lysozyme and 0.01 U of DNase I for 5 s on, 5 s off, for 36 cycles with 35% amplitude. After centrifugation at 10,000g for 15 minutes at 4 °C, the supernatant was transferred for HaloTag protein purification by HaloTag Resin, which had been pre-washed using HaloTag purification buffer. With binding the HaloTag-ZCRB1-fused protein and centrifugation, the supernatant (the sample flowthrough) was discarded. By washing with HaloTag purification buffer for a total of three times, the HaloTag Resin bound with ZCRB1 was clove with TEV protease cleavage solution (60 µl of TEV protease supplied in 1 ml of HaloTag purification buffer), which was removed from ZCRB1 elution by HisLink Resin. Finally, 2 ml of eluted ZCRB1 was concentrated with an Amicon Ultra-4 Centrifugal Filter device (10 K<sub>d</sub>) for 30 minutes up to 100-µl volume. The final ZCRB1 protein and the reservations at every purified step were analyzed by Coomassie blue staining and western blot, simultaneously.

### RNA electrophoretic mobility shift assays

RNA oligonucleotides labeled Cy5 were synthesized by RiboBio. The REMSA was carried out per the instructions of the LightShift Chemiluminescent RNA EMSA Kit (Thermo Fisher Scientific) with some modifications. In brief, RNA probes in annealing buffer (20 mM Tris pH 7.5, 200 mM KCl) were denatured and annealed by incubation at 95 °C for 4 minutes and then slowly cooled down to 25 °C at 0.05 °C per second. Recombinant ZCRB1 was diluted to concentration series of 0 nM, 50 nM, 100 nM, 200 nM and 300 nM in Halo purification buffer. For each reaction, 1 µl of RNA probes with 6.25 nM final concentration and 1 µl of ZCRB1 protein (0 nM, 50 nM, 100 nM, 200 nM and 300 nM) were incubated in 2 µl of 10× REMSA Binding Buffer (supplied in the RNA EMSA Kit) plus 5% glycerol and 2 µg of transfer RNA (tRNA) at room temperature for 30 minutes after adding 5 µl of 5× loading buffer (50 mM HEPES pH 7.5, 80% glycerol, 0.25% bromophenol blue) after separation by native PAGE. The fluorescence signal was visualized by Odyssey Imaging Systems and quantified by Image Studio, and then the dissociation constant (K<sub>d</sub>) was calculated with nonlinear curve fitting (function one-site specific binding) using GraphPad Prism with  $Y = B_{\max} \times X / (K_d + X)$ , where Y is the ratio of [RNA-protein] / ([free RNA] + [RNA-protein]), X is the input protein concentration and B<sub>max</sub> is set to 1.

### Identification of ktRNAs from RIP-PEN-seq or PEN-seq data

We clustered the overlapping paired-end reads that mapped to the genomes. Specifically, we defined the start and end sites with the highest abundance in the cluster as the transcription start site (TSS) and transcription terminal site (TTS), respectively. The sequences between TSSs and TTSs co-occurring within 500 bp were extracted as candidate transcripts. Each candidate transcript with a precise TSS and TTS was input to kturnSeeker. A new software program, kturnSeeker, was developed to discover RNAs with fktRNAs and bktRNAs. kturnSeeker searches A•G, G•A and G•C base pairs (the NC-stem), the 3-nt bulge and a section with regular base pairing (the C-stem) step by step in candidate sequences and scores the C-stem with a strict scoring scheme (a G•C pair is bonus 3, an A•U pair is bonus 2 and mismatch is minus 5). Only those with a total score (the C-stem) greater than 7 were retained as candidate ktRNAs. kturnSeeker can, for the first time, screen bktRNAs as well as fktRNAs by reverse searching the K-turn structure.

To identify novel ktRNAs, ktRNAs overlapping with known annotations were discarded. Moreover, all novel ktRNAs must be present in at least two sequencing libraries.

### Annotation of ktRNAs

The genome sequence of humans (hg38) was downloaded from the UCSC Genome Browser site<sup>61</sup>. Human gene annotations were acquired from GENCODE<sup>62</sup> in August 2018. The repeat elements in RepeatMasker were downloaded from the UCSC Genome Browser site<sup>61</sup>. The sequences and annotation data for canonical human and mouse box C/D snoRNAs were downloaded from snoRNA-LBME-db<sup>63</sup>, deepBase<sup>64</sup>, GENCODE<sup>62</sup>, snoRNAome<sup>65,66</sup> and RefSeq<sup>67</sup> in August 2018. All C/D RNAs were intersected with canonical C/D box snoRNAs using BEDTools software<sup>68</sup>, and only C/D box RNAs that did not overlap with canonical C/D box snoRNAs were identified as novel ktRNAs. All novel ktRNAs were further intersected with canonical gene annotations using BEDTools software<sup>68</sup>.

### Identification of homologous bktRNA families in other species with different evolutionary clades

We constructed homologous bktRNA families based on DNA sequence similarity. To identify evolutionary conservation of human bktRNA families, we used the UCSC liftOver tool (version 396)<sup>61</sup> and BLAST+ (version 2.11.0)<sup>69</sup> to obtain the homologous regions and sequences of the human bktRNAs in other species with different evolutionary clades. The homologous sequences from other species were input into kturn-Seeker software to identify candidate bktRNAs. The genomes included the *Pan troglodytes* reference genome (chimpanzee, panTro6), the *Pan paniscus* reference genome (bonobo, panPan3), the *Gorilla gorilla gorilla* reference genome (gorilla, gorGor6), the *Pongo pygmaeus abelii* reference genome (orangutan, ponAbe3), the *Macaca mulatta* reference genome (rhesus, rheMac10), the *Mus musculus* reference genome (mouse, mm10), the *Monodelphis domestica* reference genome (opossum, monDom5), the *Gallus gallus* reference genome (chicken, galGal6) and the *Danio rerio* reference genome (zebrafish, danRer10).

### Processing of 15.5K and FBL CLASH sequencing data

First, Cutadapt (version 2.8)<sup>70</sup> was used to cut the adapters of the paired-end reads. The paired-end reads without adapters were further merged using FLASH software (version 1.2.11)<sup>71</sup> with the following parameters: -z -m 10 -M 135 -x 0.25. fastx\_collapser software was used to collapse the reads with the default parameters for removal of PCR duplicates. Reads with fully identical sequences, including the same barcode and insert fragment, were marked as PCR duplicates, and only one of the identical sequences was retained for subsequent analysis. Reads shorter than 30 nt were removed.

### Differential splicing analysis

Adaptor sequences were trimmed from raw RNA-seq data using Cutadapt (version 2.8)<sup>70</sup>. The clean reads were mapped to the reference genome (hg38) using STAR software (2.7.1a)<sup>72</sup> with the genome index built from GENCODE version 32 (ref. 73) annotation and with the following additional parameters: -alignEndsType EndToEnd -outSAMstrandField intronMotif -outFilterMismatchNmax 5. The genomic coordinates of introns were extracted from GENCODE version 32 (ref. 73) annotation, and duplicates were further removed. For each intron, the numbers of reads that supported either the spliced isoform or the retained isoform were counted. Reads that skipped the intron and spanned at least 10 bp in both exons were thought to support the spliced isoform, and reads that included the intron with at least 10 bp overlapping between the exon side and the intron side were thought to support the retained isoform. The read counts were normalized by effective length, defined as the number of possible positions for supporting reads, to estimate the abundance of the spliced isoform and the retained isoform. Then, the inclusion level was calculated as the abundance of the retained isoform divided by the abundance of both

isoforms. The statistical method and codes from rMATS<sup>74</sup> were used to look for introns with significant inclusion level differences between WT and KO cells. For each pair of compared groups, introns with a sum of the number of reads supporting spliced isoforms or retained isoforms less than 20 in either sample were filtered before statistical testing to remove potential false positives. The annotation of minor introns was downloaded from MIDB<sup>38</sup>. The *P* value on the cumulative plots of inclusion level differences were calculated using two-sided Mann–Whitney–Wilcoxon tests.

### Statistics and reproducibility

Data are presented as the mean values  $\pm$  s.e.m. We used paired Student's *t*-tests for comparisons between the two experimental groups. All statistics were performed using GraphPad Prism 6 or R (3.6.3). The numbers of biological replicates for the experiment are indicated in the figure legends. At least four independent experiments of RIP-PEN-seq were performed for both HEK293T-FLAG-15.5K and Hepa1-6-FLAG-15.5K cells. Four independent experiments of RIP-PEN-SHAPE-MaP were performed for HEK293T-FLAG-15.5K cells. The PEN-seq analysis for cellular fractions in HEK293T and HCT116 cells was performed four times, respectively. Two or three independent sets of HEK293T, HCT116, U-87 MG, HepG2, HeLa and K562 RNA samples were used for PEN-seq analysis. All images of polyacrylamide gels, agarose gels and northern blots are representative of at least two biological replicates. No statistical method was used to predetermine the sample size.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

### Data availability

All sequencing data that support the findings of this study have been deposited in the National Center for Biotechnology Information's Gene Expression Omnibus with the following accession numbers: [GSE160970](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE160970) for all HEK293T 15.5K RIP-PEN-seq; [GSE182757](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182757) for all Hepa1-6 RIP-PEN-seq; [GSE160636](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE160636) for FBL and 15.5K CLASH-seq in HEK293T cells; [GSE160887](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE160887) for PEN-seq in HCT116, U-87 MG, HeLa, HEK293T, HepG2 and K562 cells; [GSE186849](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE186849) for PEN-seq in 15.5K knockdown HEK293T cells; [GSE182843](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182843) for PEN-seq in HEK293T and HCT116 cell fractions; [GSE160515](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE160515) for RNA-seq in bktRNA1 KO HCT116 cells; [GSE182830](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182830) for RNA-seq in ZCRB1 knockdown HCT116 cells; [GSE182759](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182759) for RNA-seq in 15.5K knockdown HEK293T cells; and [GSE220470](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE220470) for RIP-PEN-SHAPE-MaP in HEK293T cells. All data are available in the manuscript and in Supplementary Information and Source data files. There are no restriction on data availability. Source data are provided with this paper.

### Code availability

The program kturnSeeker was written in the C++ programming language and is available from GitHub with no restrictions or conditions on access: <https://github.com/sysu-software/kturnSeeker>.

### References

- Kramer, K. et al. Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins. *Nat. Methods* **11**, 1064–1070 (2014).
- Morlan, J. D., Qu, K. & Sinicropi, D. V. Selective depletion of rRNA enables whole transcriptome profiling of archival fixed tissue. *PLoS ONE* **7**, e42882 (2012).
- Adiconis, X. et al. Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nat. Methods* **10**, 623–629 (2013).
- Zinshteyn, B., Wangen, J. R., Hua, B. & Green, R. Nuclease-mediated depletion biases in ribosome footprint profiling libraries. *RNA* **26**, 1481–1488 (2020).



53. Gillen, A. E., Yamamoto, T. M., Kline, E., Hesselberth, J. R. & Kabos, P. Improvements to the HITS-CLIP protocol eliminate widespread mispriming artifacts. *BMC Genomics* **17**, 338 (2016).
54. Keene, J. D., Komisarow, J. M. & Friedersdorf, M. B. RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. *Nat. Protoc.* **1**, 302–307 (2006).
55. Deng, B. et al. An LTR retrotransposon-derived lncRNA interacts with RNF169 to promote homologous recombination. *EMBO Rep.* **20**, e47650 (2019).
56. Spitale, R. C. et al. Structural imprints in vivo decode RNA regulatory mechanisms. *Nature* **519**, 486–490 (2015).
57. Phelps, W. A., Carlson, A. E. & Lee, M. T. Optimized design of antisense oligomers for targeted rRNA depletion. *Nucleic Acids Res.* **49**, e5 (2021).
58. Chu, C., Qu, K., Zhong, F. L., Artandi, S. E. & Chang, H. Y. Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol. Cell* **44**, 667–678 (2011).
59. Percharde, M. et al. A LINE1-nucleolin partnership regulates early development and ESC identity. *Cell* **174**, 391–405 (2018).
60. Chu, C. et al. Systematic discovery of Xist RNA binding proteins. *Cell* **161**, 404–416 (2015).
61. Haeussler, M. et al. The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res.* **47**, D853–d858 (2019).
62. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
63. Lestrade, L. & Weber, M. J. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res.* **34**, D158–D162 (2006).
64. Yang, J. H., Shao, P., Zhou, H., Chen, Y. Q. & Qu, L. H. deepBase: a database for deeply annotating and mining deep sequencing data. *Nucleic Acids Res.* **38**, D123–D130 (2010).
65. Jorjani, H. et al. An updated human snoRNAome. *Nucleic Acids Res.* **44**, 5068–5082 (2016).
66. Kishore, S. et al. Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small RNA sequencing. *Genome Biol.* **14**, R45 (2013).
67. Pruitt, K. D. et al. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* **42**, D756–D763 (2014).
68. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
69. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
70. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 (2011).
71. Magoc, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
72. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
73. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
74. Shen, S. et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-seq data. *Proc. Natl Acad. Sci. USA* **111**, E5593–E5601 (2014).

## Acknowledgements

We thank L. Huang from Sun Yat-sen Memorial Hospital for valuable suggestions on this manuscript. We thank Y. Zhang for sharing Sleeping Beauty transposon system. We thank all the staff from Sun Yat-sen University for their support and hard work during the COVID-19 pandemic. This work was supported, in part, by the National Key R&D Program of China (2019YFA0802202 (to J.Y.) and 2022YFA1303300 (to J.Y.)); the National Natural Science Foundation of China (32225011 (to J.Y.), 91940304 (to J.Y.), 31971228 (to J.Y.), 31770879 (to J.Y.), 31970604 (to L.Q.), 31900903 (to B.L.) and 32100467 (to S.L.)); the Youth Science and Technology Innovation Talent of Guangdong TeZhi Plan (2019TQ05Y181 (to J.Y.)); funds from Guangzhou City (202002030351 (to J.Y.)); and Fundamental Research Funds for the Central Universities, Sun Yat-sen University (20lgpy112 (to B.L.) and 2021qntd26 (to B.L.)).

## Author contributions

J.Y., B.L. and L.Q. conceived and designed the entire project. J.Y. and L.Q. designed and supervised the research. B.L., S.L., W.Z., A.L., P.Y., D.W., J.Z., P.Z., C.L., Q.L., J.Y., S.H., Q.H., H.Z. and J.Y. performed the experiments and/or data analyses. J.Y., B.L. and A.L. performed the genome-wide or transcriptome-wide data analyses. J.Y., B.L., S.L. and L.Q. contributed reagents/analytic tools and/or grant support. J.C. provided helpful discussions. J.Y., L.Q., B.L., S.L. and A.L. wrote and revised the paper. All authors discussed the results and commented on the manuscript.

## Competing interests

J.C. is a scientific advisory board member of Race Oncology. The remaining authors declare no competing interests.

## Additional information

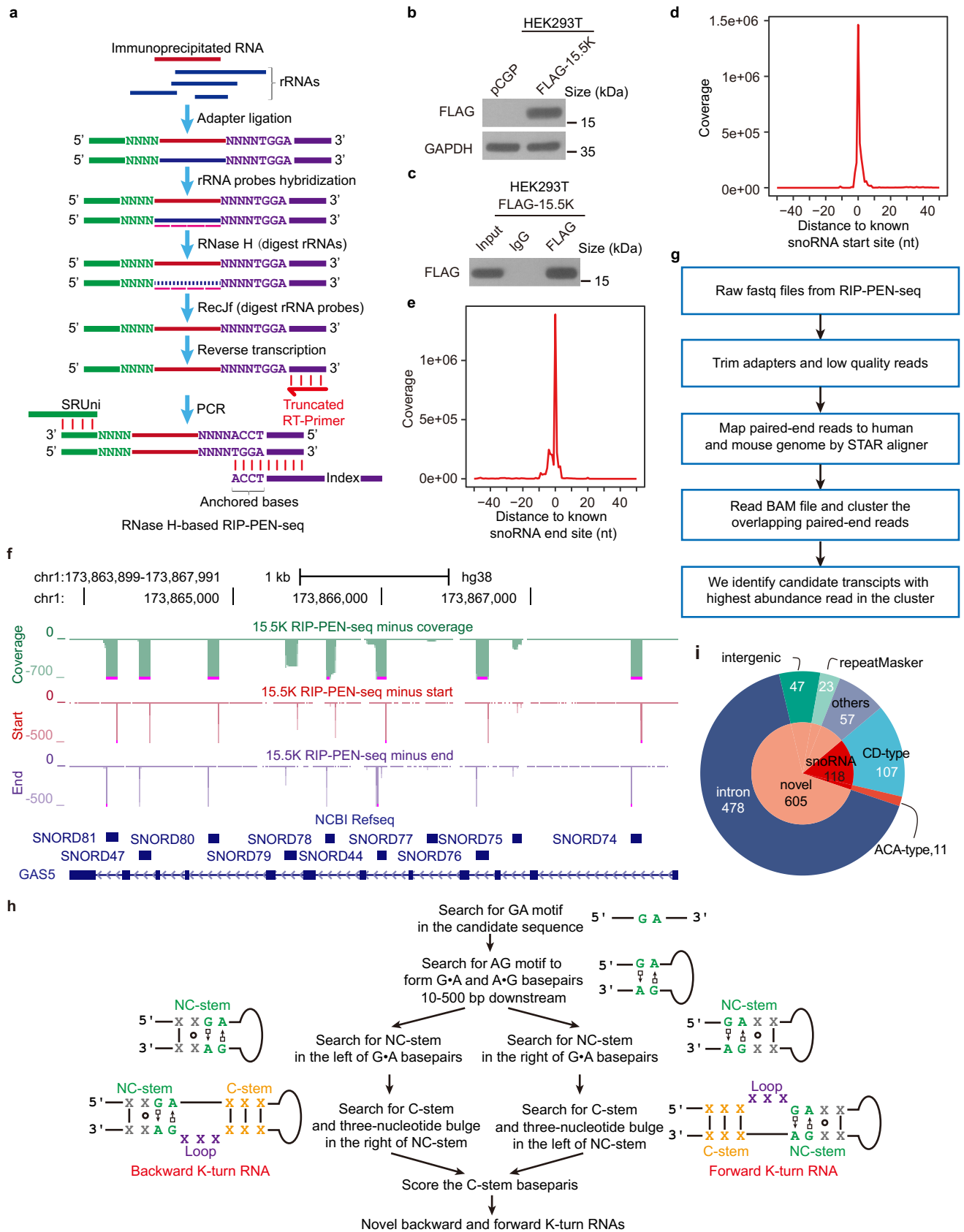
**Extended data** is available for this paper at <https://doi.org/10.1038/s41587-023-01749-0>.

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41587-023-01749-0>.

**Correspondence and requests for materials** should be addressed to Lianghu Qu or Jianhua Yang.

**Peer review information** *Nature Biotechnology* thanks the anonymous reviewers for their contribution to the peer review of this work.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

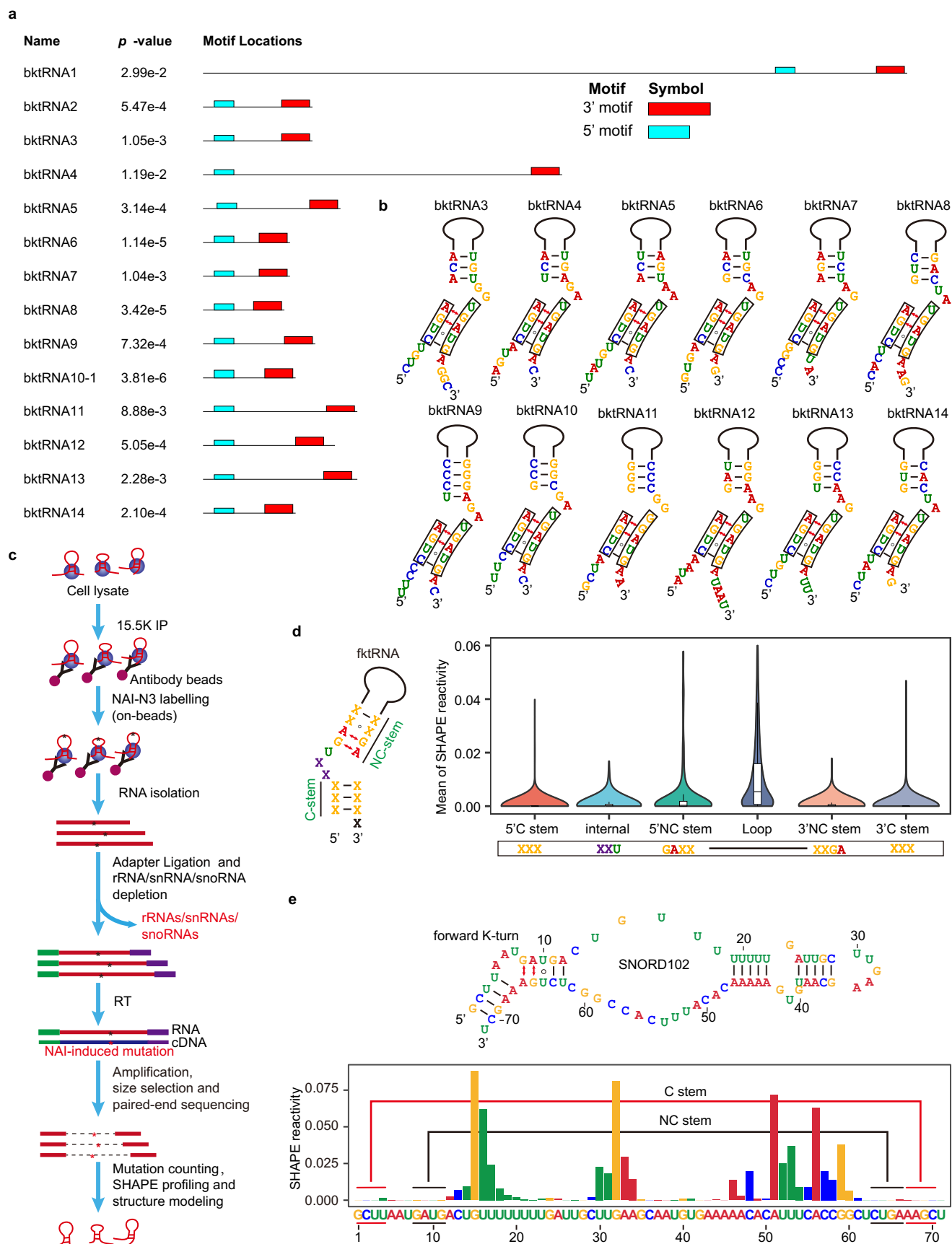


Extended Data Fig. 1 | See next page for caption.

**Extended Data Fig. 1 | Characterization of the RIP-PEN-seq technique.**

**a**, Diagram of RNase H-based rRNA depletion for the construction of RIP-PEN-seq library. **b**, Western blotting analysis of the overexpression of 15.5K-FLAG protein in HEK293T cells. GAPDH serves as the loading control. **c**, RIP from HEK293T cells was performed using anti-FLAG and IgG. **d**, **e**, Meta-analyses of the RIP-PEN-seq results for the start (**d**) and end (**e**) sites of the forward K-turn RNAs (box C/D ncRNAs) in HEK293T cells. **f**, Genome-browser plot of RIP-PEN-seq (coverage, green; 5'-start, red; 3'-end, purple) for representative forward K-turn RNAs (box C/D snoRNAs) in the introns of GASS. **g**, Computational workflow for

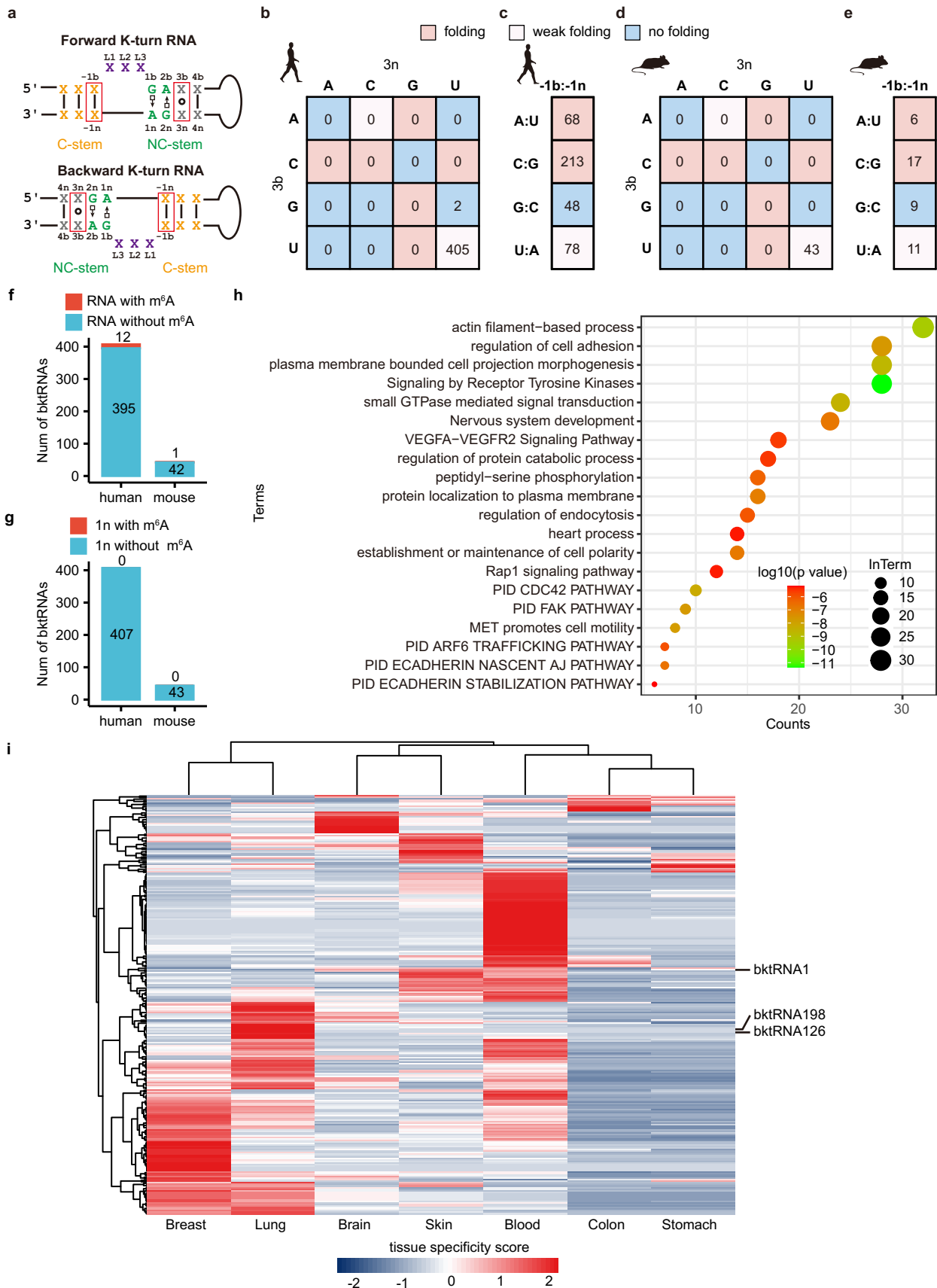
analysis of the RIP-PEN-seq sequencing data and identification of candidate transcripts. **h**, KturnSeeker core algorithm workflow. KturnSeeker was developed to identify and quantify forward (fktRNAs) and backward ktRNAs (bktRNAs) from RIP-PEN-seq data. KturnSeeker can screen bktRNAs as well as fktRNAs by reverse searching the K-turn structure. **i**, Gene type distribution of forward ktRNAs identified from RIP-PEN-seq. CD-type represents C/D box-containing snoRNAs and scaRNAs. ACA-type represents snoRNAs or scaRNAs that only contain H/ACA boxes.



Extended Data Fig. 2 | See next page for caption.

**Extended Data Fig. 2 | Structural characterization of bktRNAs. a,** The position and significance of motifs located within bktRNAs. The enriched motifs were identified by MEME software. The K-turn structural motif of bktRNAs is composed of two conserved sequence motifs: a CUGA motif often 4 nt downstream of the 5' end and a UGAUG motif often 2 nt upstream of the 3' end. The position p-value is defined as the probability that a random sequence would have a motif match score greater or equal to the sequence under test. **b,** Another twelve novel bktRNAs discovered from RIP-PEN-seq data. All novel bktRNAs had a CUGA motif that was often 4 nt downstream of the 5' end and a UGAUG motif that was often 2 nt upstream of the 3' end. These two sequence motifs were located within the K-turn structural motifs of bktRNAs. The 5' motif (CUGA) and 3' motif (UGAUG) are marked with black rectangles. The non-canonical A•G and G•A

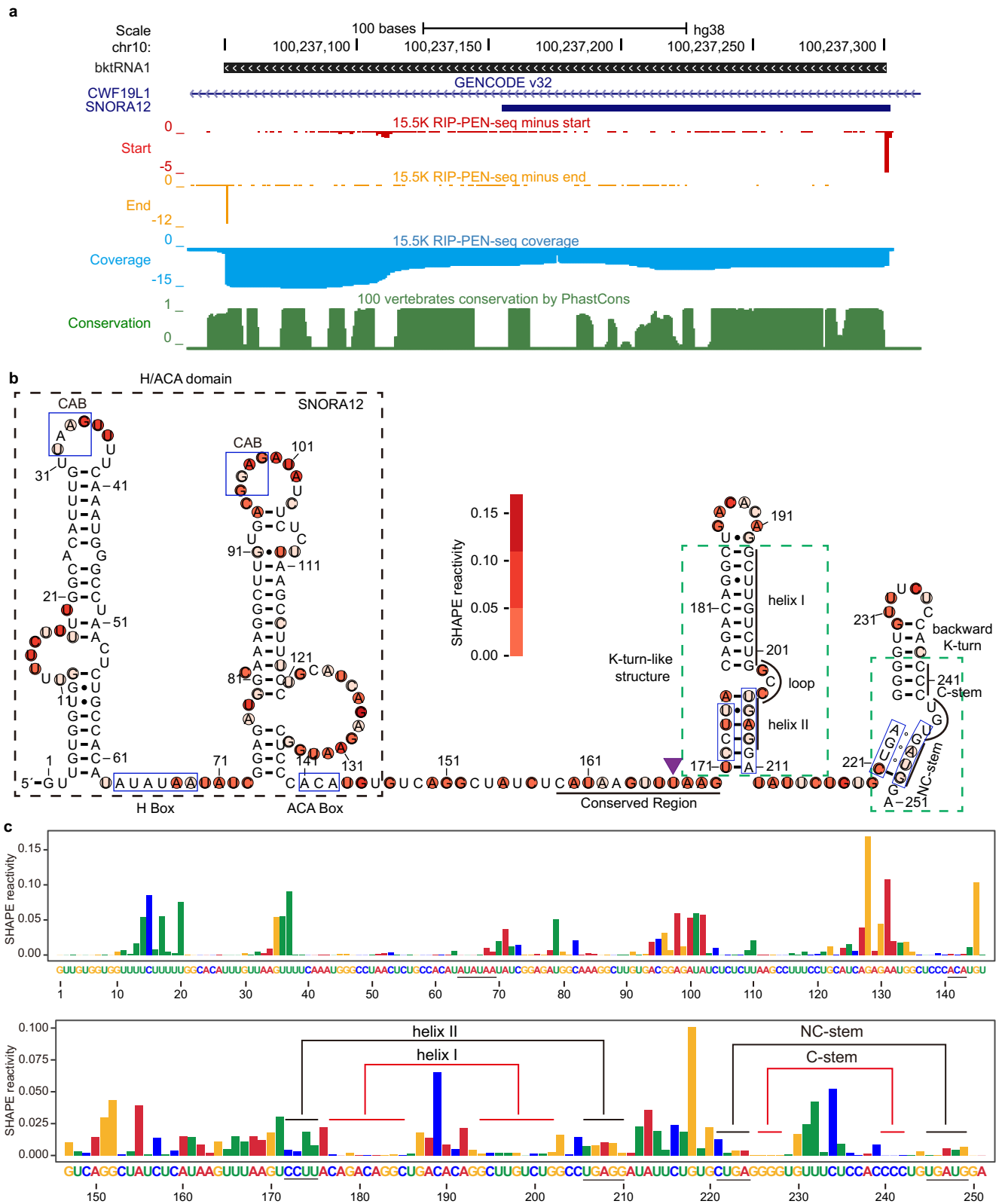
base pairs and the mismatch in the backward K-turn structure are also marked. **c,** Schematic overview of 15.5 K RIP-PEN-SHAPE-MaP. **d,** Secondary structure of consensus fktRNA (left panel, n = 98). Violin plots displaying the SHAPE reactivity across the forward K-turn structure (including 5' C stem, internal (internal loop), 5' NC stem, Loop, 3' NC stem, and 3' C stem), averaged across all known box C/D snoRNAs (right panel). The boxplots indicate the median and the upper and lower quartiles. **e,** The predicted secondary structure (upper panel) and SHAPE reactivity signal (lower panel) on fktRNA SNORD102 (also known as U102). The forward K-turn structure is indicated in the structure figure. The NC-stem and C-stem are marked with black and red underlines in the bar plot, respectively. The SHAPE reactivity signal was determined by RIP-PEN-SHAPE-MaP in this study.



Extended Data Fig. 3 | See next page for caption.

**Extended Data Fig. 3 | The backward K-turn sequence composition, predicted functions and tissue-specific expression profile of bktRNAs.** **a**, The secondary structure of consensus forward K-turn RNA (fktRNA) and backward K-turn RNA (bktRNA). The nucleotide positions in the K-turn structure are named according to the nomenclature rules for the forward K-turn structure. **b**, Matrix plot showing the number of human bktRNAs with the indicated nucleotide in the 3b:3n sequences. **c**, Number of human bktRNAs with the four possible Watson-Crick base pairs in the -1b:-1n position. **d**, Matrix plot showing the number of mouse bktRNAs with the indicated nucleotide in the 3b:3n sequences. **e**, Number of mouse bktRNAs with the four possible Watson-Crick base pairs in the -1b:-1n

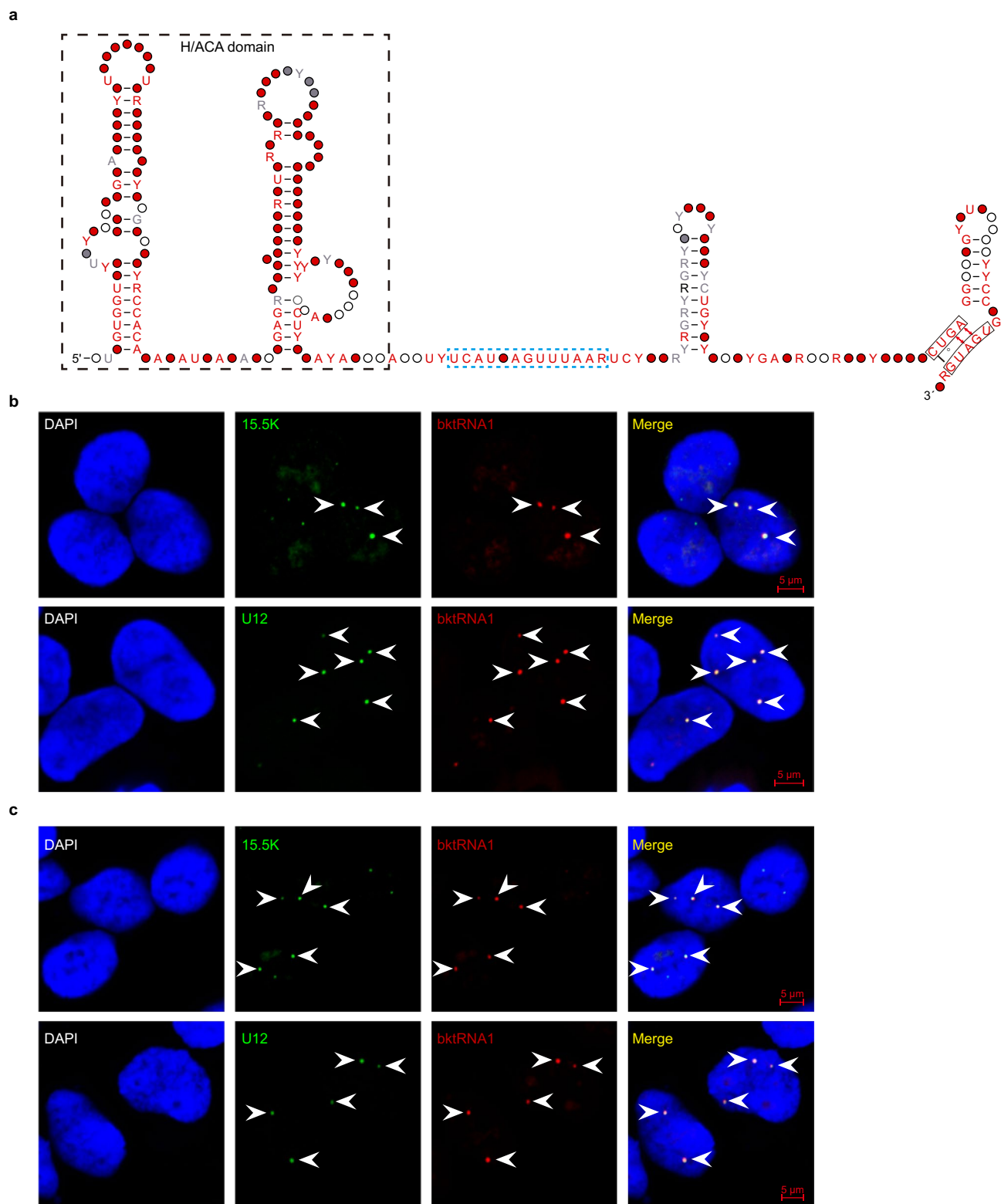
position. **f**, Number of bktRNAs with or without m<sup>6</sup>A modification in humans and mice. **g**, Number of bktRNAs with or without m<sup>6</sup>A modification at the 1n position in humans and mice. **h**, Enrichment analysis of the bktRNA host protein-coding genes by Metascape software. **i**, Tissue-specific expression profiles of bktRNAs. The expression levels of bktRNAs are displayed in the rows and the tissues are shown in the columns. The rows and columns are sorted based on k-means clustering of bktRNA genes. The colour intensity represents the tissue-specific score (JS score) as calculated for each bktRNA using the csSpecificity function. Representative bktRNAs are indicated in the right panel.



**Extended Data Fig. 4 | Genomic characterization, expression, conservation, and secondary structure of bktRNA1.** **a**, Genome-browser plot of RIP-PEN-seq (coverage, blue; 5'-start, red; 3'-end, yellow) for bktRNA1, as well as the evolutionary conservation across 100 vertebrates (green). **b**, Secondary structure of bktRNA1 in the human genome was predicted by R-scape software. The SHAPE reactivities for each nucleotide were mapped to secondary structures using R2R software. The box H/ACA domain is indicated with a black dashed box,

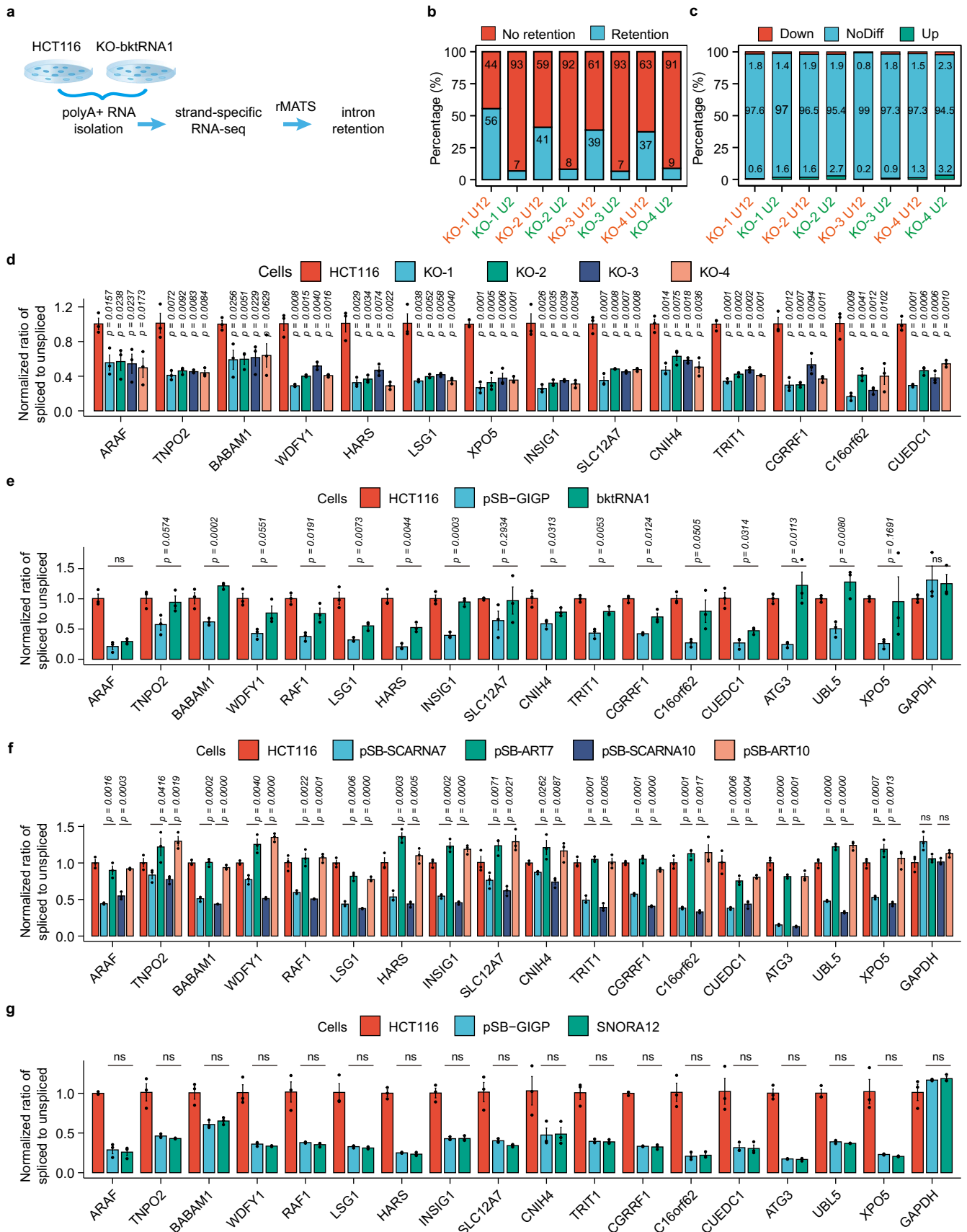
and the backward K-turn structure and the potential K-turn-like structure are marked with green dashed boxes. The NC stem and C stem are indicated with black lines. The blue boxes show the representative motifs. CAB, Cajal body box. **c**, The SHAPE reactivity signal on bktRNA1. The representative motifs are underlined in the bar plot. The SHAPE reactivity signal was determined by RIP-PEN-SHAPE-MaP in this study.





**Extended Data Fig. 5 | Secondary structures and subcellular localization of bktRNA1 and its interacting partner U12 snRNA. a,** Predicted conserved RNA structure of bktRNA1 determined by measuring pairwise covariations with R-scape software. The H/ACA domain is indicated with a black dashed box. The functional region paired with U12 snRNA is indicated with a blue dashed box. **b,** In situ co-localization of bktRNA1 with 15.5 K proteins and U12 snRNAs in HEK293T

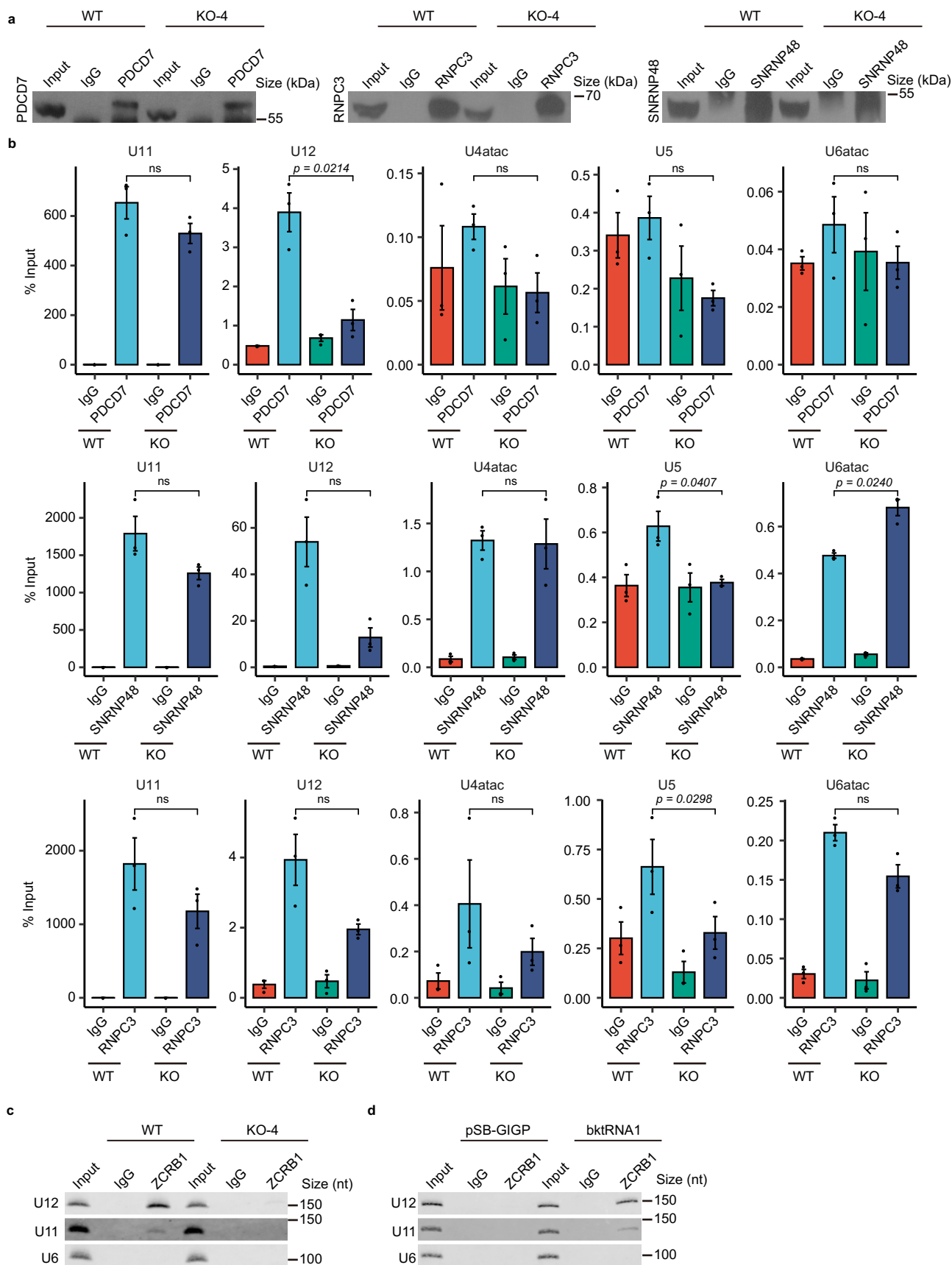
cells by fluorescent in situ hybridization (FISH) and immunofluorescence (IF) microscopy. White arrows indicate the signal detected by probes or antibody. **c,** In situ co-localization of bktRNA1 with 15.5 K proteins and U12 snRNAs in HCT116 cells by fluorescent in situ hybridization (FISH) and immunofluorescence (IF) microscopy. White arrows indicate the signal detected by probes or antibody.



Extended Data Fig. 6 | See next page for caption.

**Extended Data Fig. 6 | Splicing efficiency analysis for wild-type and knockout bktRNA1.** **a**, Workflow for intron retention analysis in HCT116 and KO-bktRNA1 cells. **b**, Proportion of aberrantly retained U12- and U2-type introns (filtered by  $p < 0.05$ ) in bktRNA1-deficient cells. **c**, Proportion of statistically significant changes (filtered by  $p < 0.05$ ) in U12- and U2-type genes in bktRNA1-deficient cells. **d**, The ratio of spliced to unspliced pre-mRNA for U12-type introns was determined by qPCR in bktRNA1-deficient cells. Data are presented as mean values  $\pm$  SEM ( $n = 3$ , biological replicates), two-tailed, paired t-test. **e**, The ratio of spliced to unspliced pre-mRNA for U12-type and U2-type (GAPDH) introns

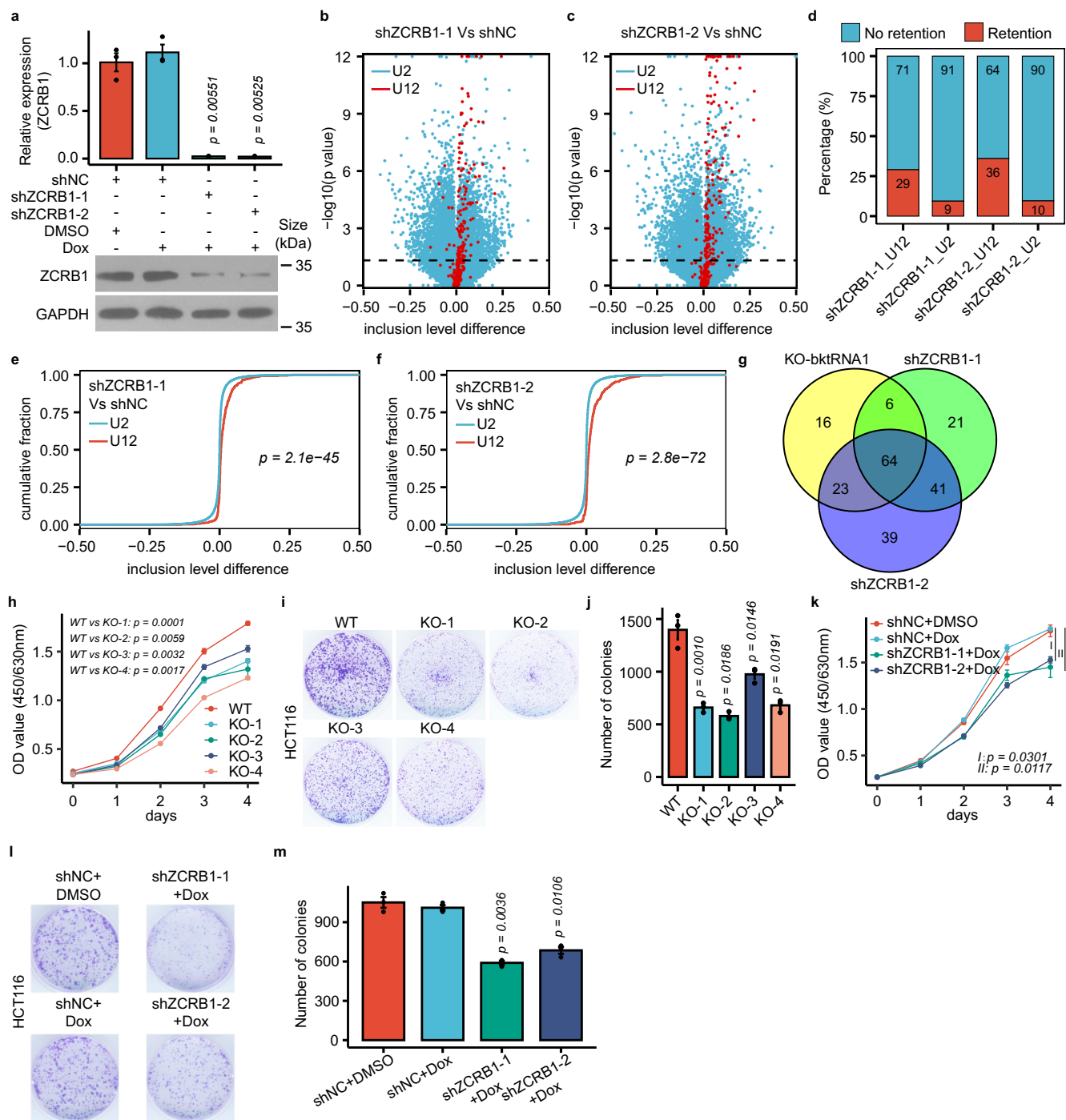
was determined by qPCR in bktRNA1-rescued HCT116 KO-4 cells. **f**, The ratio of spliced to unspliced pre-mRNA for U12-type and U2-type (GAPDH) introns was determined by qPCR in SNORA12-rescued HCT116 KO-4 cells. Data are presented as mean values  $\pm$  SEM ( $n = 3$ , biological replicates), two-tailed, paired t-test. ns, no significance. **g**, The ratio of spliced to unspliced pre-mRNA for U12-type and U2-type (GAPDH) introns was determined by qPCR in artificial scaRNA-overexpressing HCT116 KO-4 cells. Data are presented as mean values  $\pm$  SEM ( $n = 3$ , biological replicates), two-tailed, paired t-test. ns, no significance.



Extended Data Fig. 7 | See next page for caption.

**Extended Data Fig. 7 | Depletion of bktRNA1 affects the interaction between U12 and ZCRB1.** **a**, Western blots showing precipitation with each indicated antibody in wild-type (WT) and bktRNA1-deficient KO-4 (KO) cells. **b**, Native RIP was performed in wild-type (WT) and bktRNA1-deficient KO-4 (KO) cells using each indicated antibody or normal IgG antibody, after which qPCR was performed with primers recognizing minor splice snRNAs (U11, U12, U4atac, U5, U6atac). The percentage of RIP-enriched snRNAs was calculated relative

to the input RNA. Data are presented as mean values  $\pm$  SEM (n = 3, biological replicates), two-tailed, paired t-test. ns, no significance. **c**, ZCRB1 RIP-enriched snRNAs were detected by Northern blotting in wild-type (WT) and bktRNA1-deficient KO-4 (KO) cells. U6 snRNA served as a negative control. **d**, ZCRB1 RIP-enriched snRNAs were detected by Northern blotting in bktRNA1-deficient and bktRNA1-rescued cells. U6 snRNA served as a negative control.

**Extended Data Fig. 8 | ZCRB1 knockdown affects U12-type intron splicing.**

**a**, qPCR (upper panel) and western blotting analysis (lower panel) of Dox-inducible ZCRB1 knockdown in HCT116 cells. GAPDH was used as an internal reference gene for qPCR, and GAPDH served as the loading control for western blotting. Data are presented as mean values  $\pm$  SEM ( $n = 3$ , biological replicates), two-tailed, paired t-test. **b, c**, Dot plots displaying the intron retention levels in a representative pairwise analysis of ZCRB1 knockdown and negative control cells. The red dots represent U12-type introns, and the blue dots represent U2-type introns. **d**, Proportion of aberrantly retained U2- and U12-type introns in ZCRB1 knockdown cells. The red boxes represent retained introns, and the blue boxes represent unretained introns. **e, f**, Cumulative fraction of the inclusion level difference between U12-type and U2-type introns in ZCRB1 knockdown and negative control cells. The  $P$  value on the cumulative plots of inclusion level

differences were calculated using a two-sided Mann-Whitney-Wilcoxon test. **g**, Venn diagram showing the numbers of overlapping retained introns across four bktRNA1-deficient HCT116 cell lines and ZCRB1 knockdown cells. **h**, CCK-8 assay of HCT116 cells with bktRNA1 knockout. Data are presented as mean values  $\pm$  SEM ( $n = 3$ , biological replicates), two-tailed, paired t-test. **i**, Colony formation assay of HCT116 cells with bktRNA1 knockout. **j**, Quantitative analysis of colony formation assay in the indicated lines. Data are presented as mean values  $\pm$  SEM ( $n = 3$ , biological replicates), two-tailed, paired t-test. **k**, CCK-8 assay of HCT116 cells with ZCRB1 knockdown. Data are presented as mean values  $\pm$  SEM ( $n = 3$ , biological replicates), two-tailed, paired t-test. **l**, Colony formation assay of HCT116 cells with ZCRB1 knockdown. **m**, Quantitative analysis of colony formation assay in the indicated lines. Data are presented as mean values  $\pm$  SEM ( $n = 3$ , biological replicates), two-tailed, paired t-test.

## Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

#### Data collection

High throughput sequencing data was collected by Illumina X-ten sequencer. Primer extension assay images were collected by Odyssey software v3.0. ZEN lite 3.1 from ZEISS microscopy was used to acquire cell images. The amplification signal of qPCR data was acquired by QuantStudio Real-time PCR software v1.7.2. Images of agarose gel were collected by Uvtech1D.

#### Data analysis

cutadapt v2.8, FLASH v1.2.11, STAR aligner 2.7.8a, UCSC Genome Browser, UCSC liftOver tool v396, BLAST+ v2.11.0, rMATS v4.0.2, Metascape v3.5, GraphPad Prism 6, R v3.6.3, snoSeeker v1.0. The program kturnSeeker was written in the C++ programming language and is available from GitHub ( <https://github.com/sysu-software/ktturnSeeker> ).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

## Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

All sequencing data that support the findings of this study have been deposited in NCBI's Gene Expression Omnibus (GEO) with the following accession numbers: GSE160970 for all HEK293T 15.5K RIP-PEN-seq; GSE182757 for all Hepa1-6 RIP-PEN-seq; GSE160636 for FBL and 15.5K CLASH-seq in HEK293T cells; GSE160887 for PEN-seq in HCT116, U-87 MG, HeLa, HEK293T, HepG2 and K562 cells; GSE186849 for PEN-seq in 15.5K knockdown HEK293T cells; GSE182843 for PEN-seq in HEK293T and HCT116 cell fractions; GSE160515 for RNA-seq in bktRNA1 knockout HCT116 cells; GSE182830 for RNA-seq in ZCRB1 knockdown HCT116 cells; GSE182759 for RNA-seq in 15.5K knockdown HEK293T cells and GSE220470 for RIP-PEN-SHAPE-MaP in HEK293T cells. Publicly available datasets used in this study are as follows: HEK293T PARIS datasets (GEO: GSE74353); human reference genome (hg38, <https://hgdownload.soe.ucsc.edu/goldenPath/hg38/chromosomes/>); human genome annotation (GENCODE v32, [https://www.gencodegenes.org/human/release\\_32.html](https://www.gencodegenes.org/human/release_32.html)); box C/D snoRNA sequences and annotation (snoRNA-LBME-db, <https://www-snoRNA.biotoul.fr/>); deepBase, <https://rna.sysu.edu.cn/deepbase3/>; RefSeq, <https://www.ncbi.nlm.nih.gov/refseq/>); the Pan troglodytes reference genome (chimpanzee, panTro6, <ftp://hgdownload.soe.ucsc.edu/goldenPath/panTro6/>), the Pan paniscus reference genome (bonobo, panPan3, <ftp://hgdownload.soe.ucsc.edu/goldenPath/panPan3/>), the Gorilla gorilla gorilla reference genome (gorilla, gorGor6, <ftp://hgdownload.soe.ucsc.edu/goldenPath/gorGor6/>), the Pongo pygmaeus abelii reference genome (orangutan, ponAbe3, <ftp://hgdownload.soe.ucsc.edu/goldenPath/ponAbe3/>), the Macaca mulatta reference genome (rhesus, rheMac10, <ftp://hgdownload.soe.ucsc.edu/goldenPath/rheMac10/>), the Mus musculus reference genome (mouse, mm10, <ftp://hgdownload.soe.ucsc.edu/goldenPath/mm10/>), the Monodelphis domestica reference genome (opossum, monDom5, <ftp://hgdownload.soe.ucsc.edu/goldenPath/monDom5/>), the Gallus gallus reference genome (chicken, galGal6, <ftp://hgdownload.soe.ucsc.edu/goldenPath/galGal6/>) and the Danio rerio reference genome (zebrafish, danRer10, <ftp://hgdownload.soe.ucsc.edu/goldenPath/danRer10/>)

## Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender	<input type="text" value="n/a"/>
Population characteristics	<input type="text" value="n/a"/>
Recruitment	<input type="text" value="n/a"/>
Ethics oversight	<input type="text" value="n/a"/>

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences       Behavioural & social sciences       Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The numbers of biological replicates for the experiment are indicated in the figure legends. At least four independent experiments of RIP-PEN-seq were performed for both HEK293T-FLAG-15.5K and Hepa1-6-FLAG-15.5K cells. Four independent experiments of RIP-PEN-SHAPE-seq were performed for HEK293T-FLAG-15.5K cells. The PEN-seq analysis for cellular fractions in HEK293T and HCT116 cells was performed four times, respectively. Two or three independent sets of HEK293T, HCT116, U-87 MG, HepG2, HeLa and K562 RNA samples were used for PEN-seq analysis. All images of polyacrylamide gels, agarose gels and northern blots are representative of at least two biological replicates. For all experiments, no sample size calculation was performed. Sample sizes were chosen empirically based on previously published work, knowledge of intra-group variation and expected effect sizes, or based on discussions with experts in their respective fields. These sample sizes were sufficient for statistical analysis. Sample sizes for these experiments were chosen based upon field standards and prior knowledge of experimental variation. (PMID: 34480159; PMID: 29476152; PMID: 30867593)
Data exclusions	<input type="text" value="No data were excluded."/>
Replication	<input type="text" value="All experiments were performed in at least triplicate unless otherwise stated. All attempts at replication were successful."/>
Randomization	<input type="text" value="For all experiments, cells or samples were randomly assigned to control or treatment groups."/>



## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Antibodies

### Antibodies used

- 1) Anti-FLAG, mouse (for RIP & WB), Proteintech, Cat#66008-2-Ig, clone#1E7B4, lot#10019027
- 2) Anti-FBL, rabbit (for IF & WB), Bethyl Laboratories, Cat#A303-891A, No clone#, lot#1
- 3) Anti-15.5K, rabbit (for RIP, IF & WB), Proteintech, Cat#15802-1-AP, No clone#, lot#00019102
- 4) Anti-IgG, mouse (for RIP), Proteintech, Cat#B900620, No clone#, lot#20030599
- 5) Anti-IgG, rabbit (for RIP), Proteintech, Cat#30000-0-AP, No clone#, lot#20010049
- 6) Anti-GAPDH, mouse (for WB), Proteintech, Cat#60004-1-Ig, clone#1E6D9, lot#10013030
- 7) Anti-CWF19L1, rabbit (for WB), CUSABIO, Cat#CSB-PA727965LA01HU, No clone#, lot#E0915A
- 8) Anti-PDCD7, rabbit (for RIP & WB), Proteintech, Cat#12485-1-AP, No clone#, lot#00022780
- 9) Anti-RNPC3, rabbit (for RIP & WB), Proteintech, Cat#25820-1-AP, No clone#, lot#00024548
- 10) Anti-SNRNP48, rabbit (for RIP & WB), Proteintech, Cat#24297-1-AP, No clone#, lot#00020388
- 11) Anti-ZCRB1, rabbit (for RIP & WB), Proteintech, Cat#25629-1-AP, No clone#, lot#00056392
- 12) Anti-Coilin, rabbit (for IF), Proteintech, Cat#10967-1-AP, No clone#, lot#00013511
- 13) Alexa-Fluor 488 donkey, rabbit (for IF), ThermoFisher, Cat#A21206, No clone#, lot#2156521
- 14) IPKine™ HRP Mouse Anti-Rabbit IgG LCS (for WB), Abbkine, Cat#A25022, No clone#, lot#ATTSE1001
- 15) IPKine HRP Goat Anti-Rabbit IgG HCS (for WB), Abbkine, Cat#A25222, No clone#, lot#ATTJL1701

### Validation

All vendors showed validation data for antibodies on their websites (including Western blot, IF, IP etc.). Manufacturer's Website:

- 1) FLAG (Cited in 60 publications: <https://www.ptglab.com/products/Flag-tag-Antibody-66008-2-Ig.htm>) In addition, we had performed RIP-WB to validate the performance of this antibody in this study.
- 2) FBL (Cited in 4 publications: <https://www.fortislife.com/products/primary-antibodies/rabbit-anti-fibrillar-antibody/BETHYL-A303-891#Citations>) In addition, we had validated this antibody in this study by FBL-specific shRNAs.
- 3) 15.5K (Cited in 4 publications: <https://www.ptglab.com/products/NHP2L1-Antibody-15802-1-AP.htm>) In addition, we had validated this antibody in this study by 15.5K-specific shRNAs.
- 4) Anti-IgG, mouse (Cited in 38 publications: <https://www.ptglab.com/products/Mouse-IgG-B900620.htm>)
- 5) Anti-IgG, rabbit (Cited in 75 publications: <https://www.ptglab.com/products/IgG-control-Antibody-30000-0-AP.htm>)
- 6) GAPDH (Cited in 6314 publications: <https://www.ptglab.com/products/GAPDH-Antibody-60004-1-Ig.htm>)
- 7) CWF19L1 (Citations not listed; validation provided on website: <https://www.cusabio.com/Polyclonal-Antibody/CWF19L1-Antibody-11107060.html>)
- 8) PDCD7 (Cited in 1 publications: <https://www.ptglab.com/products/PDCD7-Antibody-12485-1-AP.htm>)
- 9) RNPC3 (Cited in 2 publications: <https://www.ptglab.com/products/RNPC3-Antibody-25820-1-AP.htm>)
- 10) SNRNP48 (Citations not listed; validation provided on website: <https://www.ptglab.com/products/SNRNP48-Antibody-24297-1-AP.htm>)
- 11) ZCRB1 (Cited in 1 publications: <https://www.ptglab.com/products/ZCRB1-Antibody-25629-1-AP.htm>) In addition, we had validated this antibody in this study by ZCRB1-specific shRNAs and ZCRB1 overexpression assay.
- 12) Coilin (Cited in 13 publications: <https://www.ptglab.com/products/COIL-Antibody-10967-1-AP.htm>)
- 13) Alexa-Fluor 488 donkey (Cited in 4482 publications: <https://www.thermofisher.cn/cn/zh/antibody/product/Donkey-anti-Rabbit-IgG-H-L-Highly-Cross-Adsorbed-Secondary-Antibody-Polyclonal/A-21206>)
- 14) IPKine™ HRP Mouse Anti-Rabbit IgG LCS (Cited in 26 publications: <https://www.abbkine.com/product/ipkine-hrp-mouse-anti-rabbit-igg-lcs-a25022/>)
- 15) IPKine HRP Goat Anti-Rabbit IgG HCS (Cited in 7 publications: <https://www.abbkine.com/product/ipkine-hrp-goat-anti-rabbit-igg-hcs-a25222/>)

## Eukaryotic cell lines

Policy information about [cell lines and Sex and Gender in Research](#)

### Cell line source(s)

HEK293T (SCSP-502), HCT116 (TChu 99), HeLa (TChu187), U-87 MG (TChu138), HepG2 (SCSP-510), K562 (SCSP-5054) and Hepa1-6 (SCSP-512) cells were purchased from the Cell Bank of the Chinese Academy of Sciences.

Authentication

-HCT116, U-87 MG and HeLa cells were validated by Cellcook Biotech using STR profiling cell authentication services.  
-Other cells (HEK293T, HepG2, K562 and Hepa1-6) obtained from the Cell Bank of the Chinese Academy of Sciences relied on the company's certificates of analysis by using STR profiling. No other independent authentication was performed.

Mycoplasma contamination

The cells were free of mycoplasma contamination based on the MycoBlue Mycoplasma Detector (Vazyme).

Commonly misidentified lines  
(See [ICLAC](#) register)

No misidentified cell lines that we know were used for this study.