TECHNICAL REPORT
https://doi.org/10.1038/s41594-020-0436-2

nature structural & molecular biology

Chemical RNA digestion enables robust RNA-binding site mapping at single amino acid resolution

Jong Woo Bae[1,2], S. Chul Kwon[1,2], Yongwoo Na[1,2], V. Narry Kim[1,2] ✉ and Jong-Seo Kim[1,2] ✉

RNA-binding sites (RBSs) can be identified by liquid chromatography and tandem mass spectrometry analyses of the protein–RNA conjugates created by crosslinking, but RBS mapping remains highly challenging due to the complexity of the formed RNA adducts. Here, we introduce RBS-ID, a method that uses hydrofluoride to fully cleave RNA into mono-nucleosides, thereby minimizing the search space to drastically enhance coverage and to reach single amino acid resolution. Moreover, the simple mono-nucleoside adducts offer a confident and quantitative measure of direct RNA–protein interaction. Using RBS-ID, we profiled ~2,000 human RBSs and probed *Streptococcus pyogenes* Cas9 to discover residues important for genome editing.

RNA-binding proteins (RBPs) play central roles in post-transcriptional regulation, affecting every step of RNA metabolism. Unraveling the complex network of RNA–protein interactions has been the subject of intense research in recent years. Many novel RBPs have been identified via 'RNA interactome capture' approaches, in which RBPs and RNAs are crosslinked (usually by UV light), enriched in an RNA-dependent manner and analyzed by liquid chromatography and tandem mass spectrometry (LC-MS/MS)[1–3]. Several methods were further developed to identify RNA-binding residues, albeit with limited success. A pioneering study by Hentze and colleagues mapped RNA-binding regions at low resolution by protease treatment (LysC or ArgC) followed by enrichment of the peptide–RNA adducts[4]. After trypsin treatment to fragment the peptides further, the unmodified fragments adjacent to the crosslinked fragments were detected to infer the RNA-interacting regions. Another approach used the loss of MS signal from the crosslinked peptides to find the RNA-binding regions[5]. To enhance the resolution of the RNA-binding site (RBS) map, Urlaub and colleagues developed a sophisticated workflow (RNP^{xl}) that calculates masses of possible peptide–RNA conjugates and searches MS/MS spectra against a sequence database to identify the crosslinking sites directly[6]. Related approaches (CAPRI, iTRAPP) with useful modifications were further developed[7,8]. However, the coverage was still very low (36 and 280 human RBSs from RNP^{xl} and CAPRI, respectively; Fig. 1).

This low coverage occurred mainly because these methods depend on enzymatic RNA digestion. The nuclease-mediated digestion is incomplete, probably due to steric hindrance, leaving diverse RNA moieties of over 100 different types including nucleotides and oligonucleotides. This complexity causes severe problems during MS/MS analysis. To identify peptides that harbor RBSs, tens to hundreds of different RNA moieties have to be considered when constructing the spectral search space. The inflated search space compromises RBS identification by reducing the peptide-spectrum match (PSM) identification rate[9] while increasing the number of false positives[10]. Also, the RNA moieties seldom remain intact until the MS2 sequencing step[6–8,11]. So, RBSs had to be indirectly mapped using the fragmented remnants of RNA instead of the intact moiety. It is thus important to cleave the RNA into a homogeneous and minimal form to narrow the search space and directly define the RBS with the mass of intact RNA residue.

In this study, we introduce an RNA-specific chemical cleavage method to replace enzymatic digestion, which drastically improves the coverage (~2,000 sites) and resolution (single amino acid level) of RBS identification (Fig. 1a).

Results

Complete chemical digestion of RNA using hydrofluoric acid. For the chemical digestion of RNA, we tested hydrofluoric acid (HF), because HF has been shown previously to cleave phosphodiester bonds in DNA oligonucleotides and phosphopeptides, leaving an -OH group at the site where the phosphate group is removed[12,13] (Extended Data Fig. 1a). We first asked if HF can cleave phosphodiester bonds within RNA, thereby digesting the polymer into monomers. Synthetic AUGC-repeat RNA and total RNA were incubated with HF and the products were analyzed by reversed-phase LC with UV detection (Fig. 1b and Extended Data Fig. 1b,c). The RNAs were fully digested into pyrimidine nucleosides and purine nucleobases. Moreover, when total cellular protein digest was treated with HF, neither peptide bond cleavage nor peptide identification was affected, indicating that HF selectively acts on RNAs but not peptides (Extended Data Fig. 1d,e).

We next examined if HF can cleave phosphodiester bonds of the peptide-crosslinked RNAs (Fig. 1a). For this, HeLaT cells were irradiated with UVC light (254 nm) to induce crosslinking between RNA and protein. After cell lysis and tryptic digestion, the peptide–RNA conjugates were purified in an RNA-dependent manner using silica membrane[8,11] (for total RNA-RBS) or in combination with oligo-dT beads[1,2] (for mRNA-RBS). Subsequently, the peptide–RNA conjugates were treated with HF and subjected to LC-MS/MS analysis. We searched for any adduct masses on peptides using two independent open search tools (MODa[14] and MSFragger[15]). Interestingly, the mass shift corresponding to uridine was prominently detected (Fig. 1c, Extended Data Fig. 2a–c and Supplementary Table 1), followed by that of uracil, which was probably derived from in-source ribose neutral loss (Extended Data Fig. 2d

[1]Center for RNA Research, Institute for Basic Science, Seoul, Korea. [2]School of Biological Sciences, Seoul National University, Seoul, Korea.
✉e-mail: narrykim@snu.ac.kr; jongseokim@snu.ac.kr

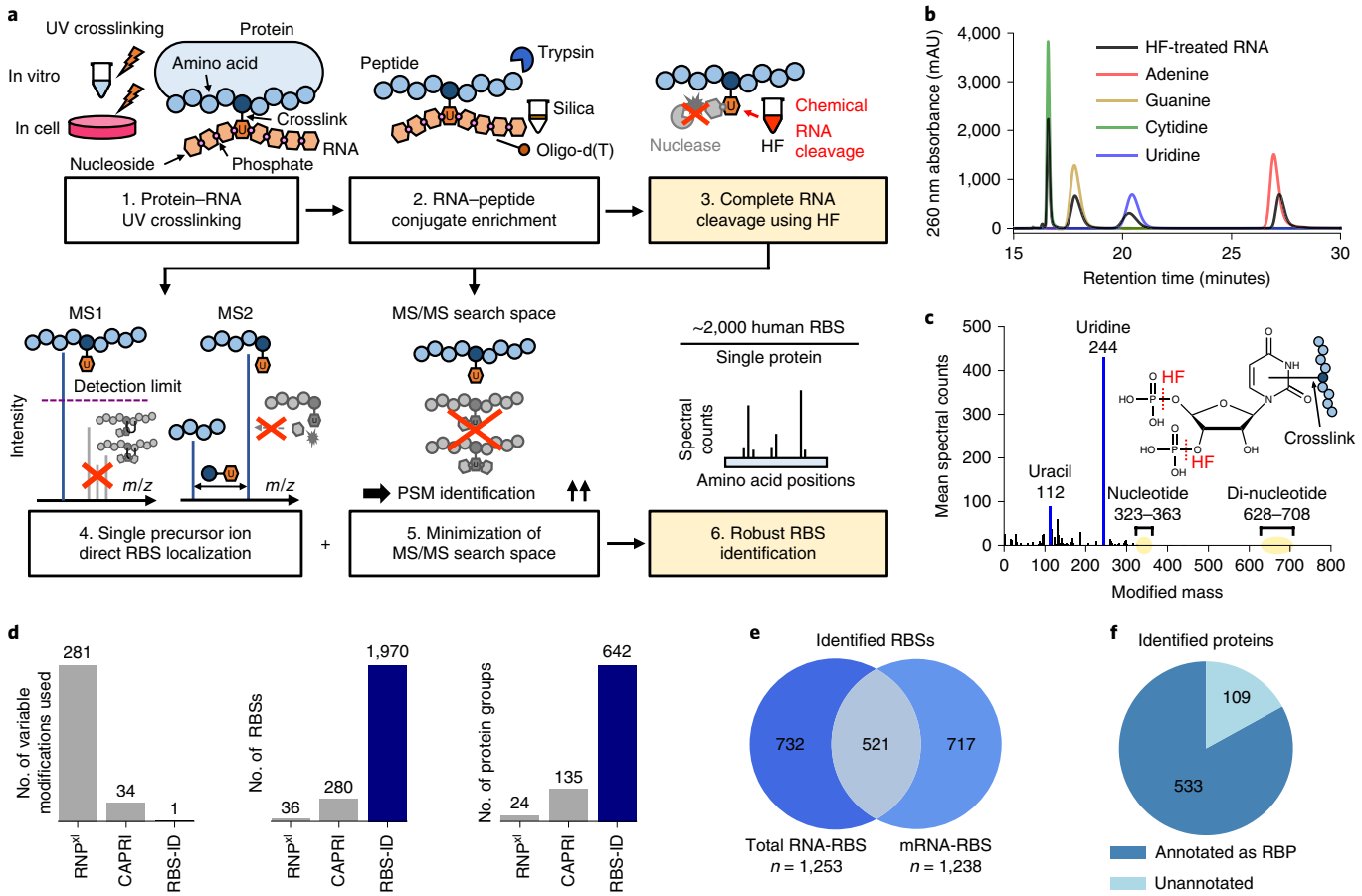


Fig. 1 | RBS-ID robustly identifies RBSs at the proteome level. a, Experimental scheme of RBS-ID. Following UV crosslinking, the crosslinked protein is digested by trypsin into peptides, and the peptide–RNA conjugates are enriched either by silica or oligo-d(T). Hydrofluoric acid (HF) digests RNA, leaving only uridine crosslinked to the peptide, which greatly simplifies the precursor ion moiety in MS1 and allows direct localization of the U-crosslinked RBS on the peptide in MS2. Because of the reduced spectral search space, the RBS is robustly identified. **b**, UV absorbance chromatogram of HeLaT total RNA digested with HF (black solid line), merged with those of reference chemicals (colored solid lines). The peaks of the reference chemicals have been re-sized for better visualization. **c**, Open search for modified mass on total RNA-RBSs using the search tool MODa. Phosphodiester bonds cleaved by HF are depicted. **d**, Comparison of RBS-ID against other methods. The number of variable modification types considered in the spectral search (left), the number of identified RBSs (center) and the number of corresponding protein groups (right) across the indicated methods are shown. For equitable comparison, only a single RBS was considered for a single peptide. **e**, Number of total RNA- and mRNA-RBSs identified by RBS-ID. **f**, Proportion of RBS-containing protein groups annotated as RBPs.

and Supplementary Fig. 1). This observation is consistent with previous reports that UVC irradiation causes exclusive crosslinking of uridine to amino acids[6–8,11]. Importantly, the mass shifts corresponding to incompletely digested products such as nucleotides or di-nucleotides were not detected or were of very low abundance (Fig. 1c and Extended Data Fig. 2a–c). This result clearly shows that HF treatment indeed digests the crosslinked RNAs completely into monomers while preserving the crosslink between peptide and RNA. In addition, we did not observe mass shifts corresponding to two crosslinked sites on a single peptide (Extended Data Fig. 2b,c). Also, a closed search allowing up to two modifications per peptide yielded negligible PSM counts with two uridine modifications, suggesting that peptides with only a single U crosslink are prevalent under our experimental conditions (Extended Data Fig. 2e).

Proteome-wide RBS identification with unprecedented coverage. The simple mono-U adduct minimized the spectral search space, bringing great advantages during tandem mass analysis in terms of the scope of RBS identification, the accuracy of RBS localization and the sensitivity to detect RNA-crosslinked peptides (Fig. 1a). We performed a closed search using MS-GF+[16] on total RNA-RBS and mRNA-RBS datasets, applying a single variable modification of uridine without amino acid specifications (Fig. 1d). After identifying RBS-containing peptides, we applied a stringent cutoff for RBS localizations within peptides, by taking RBSs with exclusive maximum spectral counts for each peptide (we also provide localization scores for all filtered-in or filtered-out RBS localizations[17,18]; for details see Methods). Collectively, we identified 1,970 RBSs in 642 protein groups at single amino acid resolution (Fig. 1d–f, Supplementary Fig. 2 and Supplementary Table 2a). Of note, we separately provide a list of candidate RBSs with non-exclusive maximum frequency in individual peptides (not included in further analysis; Supplementary Table 2b). Compared with a previous study using nuclease-based RNA cleavage[6], the majority of sites from commonly identified peptides (24 out of 25) were identical (Extended Data Fig. 3a), indicating the accuracy of the methods. RBS-ID shows remarkably improved coverage in terms of the number of both RBSs and proteins (Fig. 1d). The minimal search space helped reduce both false positives and false negatives in the spectral search, allowing robust proteome-wide identification of RBSs. It is also noted that the MS1 intensity-based quantification of RBSs shows high reproducibility between replicate experiments[19], further supporting the robustness of RBS-ID (Extended Data Fig. 4).

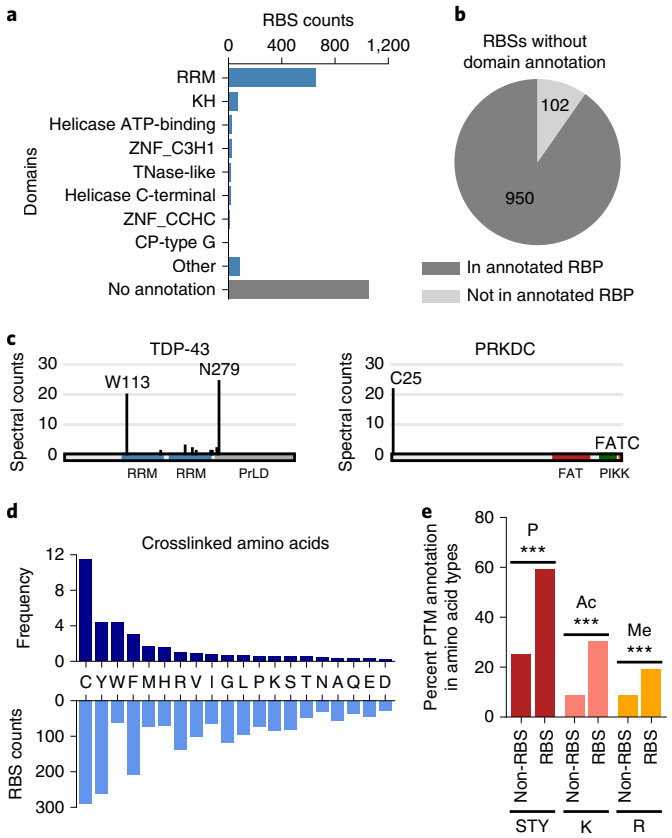


Fig. 2 | Domain and site-level analysis of identified RBSs. a, Domain annotation of identified RBSs. **b**, Among 1,052 RBSs that are not localized within known RNA-binding domains ('No annotation' group from **a**), 950 RBSs are within annotated RBPs. **c**, RBSs identified in TDP-43 and PRKDC, with domain annotations. RBSs with high spectral counts are indicated. **d**, Frequency of amino acids identified as RBSs in this study, calculated as each amino acid's proportion in identified RBSs divided by that of all sequences in proteins where RBSs were identified (top), and the corresponding RBS counts (bottom). **e**, Previously annotated post-translational modification (PTM) sites (phosphorylation (P), acetylation (Ac) and methylation (Me)) coincide with RBSs more frequently than non-RBSs of the same amino acid types within RBS-containing proteins. For each amino acid and corresponding PTM annotation, a two-sided Fisher's exact test was performed between groups of RBSs and non-RBSs. ***$P < 0.0005$ (P values are rounded up to the fourth decimal point: 0.0000 (STY-phosphorylation), 0.0000 (K-acetylation), 0.0002 (R-methylation)). Data for graphs in **a**, **d** and **e** are available online as source data.

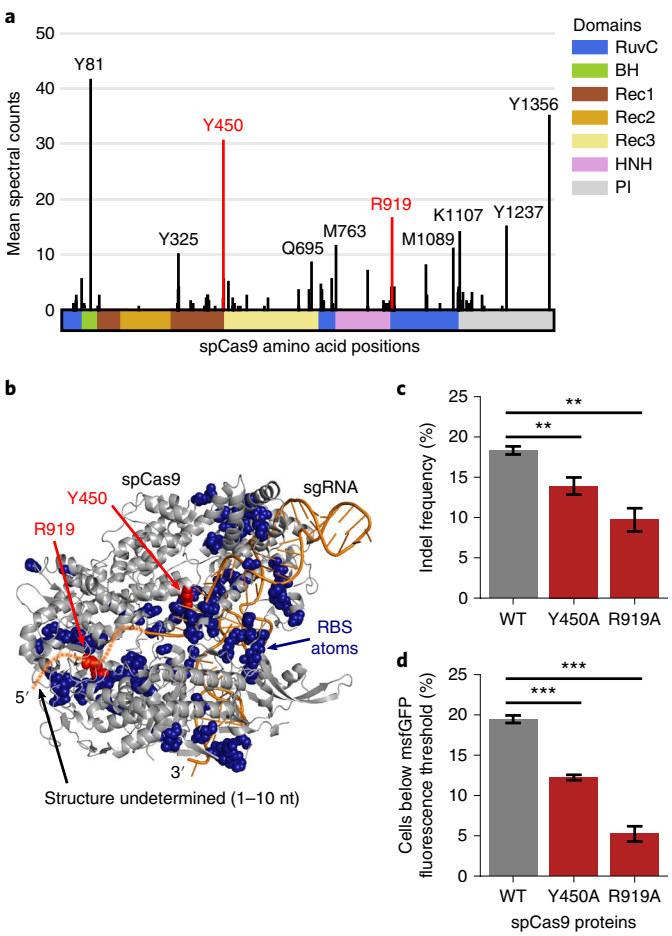


Fig. 3 | RBS-ID identifies functionally important residues in spCas9. **a**, Eighty-four RBSs mapped on spCas9. Domains and the top 10 RBSs with high spectral counts are labeled. **b**, Positions of RBSs in the previously reported crystal structure of spCas9 in complex with sgRNA (PDB 4ZT0[28]). Individual atoms of RBSs are drawn as dark blue spheres. The structurally undetermined part of sgRNA (1–10 nt) is drawn as a dashed orange line. **c**, Gene editing activity as measured by T7 endonuclease assay after transfection of wild-type (WT), Y450A and R919A spCas9 constructs[30]. Mean values of $n = 3$ biologically independent samples are depicted with error bars that indicate s.d. A two-sided unpaired Student's t-test assuming equal variance was performed. P values are rounded up to the fourth decimal point: 0.0061 (WT versus Y450A), 0.0013 (WT versus R919A). **d**, Gene editing activity as determined by quantifying the proportion of cells below the green fluorescence threshold. Fluorescence was measured by flow cytometry 96 h after transfection of WT, Y450A and R919A spCas9 constructs. Mean values are depicted with error bars that indicate the s.d. between $n = 3$ biologically independent samples. A two-sided unpaired Student's t-test was performed as described above. P values are rounded up to the fourth decimal point: 0.0001 (WT versus Y450A), 0.0000 (WT versus R919A). Data for graphs in **c** and **d** are available online as source data.

Among the proteins with RBSs identified in this study, 533 protein groups (83%) were previously annotated as RBPs in UniProtKB[20] (Fig. 1f). The unannotated ones are also enriched with gene ontology (GO) terms related to nucleic acid binding, nucleic acid processing, ribosome and translation, indicating they are likely to be authentic RBSs[21] (Extended Data Fig. 5a). As expected, RBSs identified exclusively from total RNA enrichment are found mainly in proteins related to ribosomes, while those from poly(A) + RNA enrichment are associated with mRNA processing[21] (Fig. 1e and Extended Data Fig. 5b,c). The identified sites are mostly within RBDs profiled in previous mapping studies[4,7] (Extended Data Fig. 3b,c). Many RBSs are located within known RNA interaction motifs such as RNA recognition motif (RRM), K homology (KH), helicase and zinc finger motifs (Fig. 2a). For example, we found three prominent RBSs in PABPC1 (Extended Data Fig. 6a,b). These sites are homologous to the RNA interacting residues found in yeast Pab1[22]. The structure indicates that all three residues are located in the RNA-interacting surface of RRM3 and RRM4, which are known to have low nucleotide specificity compared to RRM1 and RRM2[23], suggesting that these RBSs interact with non-poly(A) RNA.

Interestingly, a significant proportion of RBSs were mapped to regions that are not annotated as RBDs[20] (Fig. 2a, gray bar). However, ~90% of these previously unknown RBSs are from proteins known as RBPs[20] (Fig. 2b). For example, a prominent RBS on TDP-43 was identified in the intrinsically disordered region that is

known to mediate aggregation of TDP-43 in amyotrophic lateral sclerosis (Fig. 2c, left panel, N279). PRKDC, the catalytic subunit of DNA-PK—a kinase critical for DNA repair—contains a clear RBS at its N terminus that does not have a known domain or function (Fig. 2c, right panel, C25). Given the potential role for RNA in DNA repair[24], it will be interesting to investigate the RNA binding activity of PRKDC. Our data will provide a useful starting point to study the function and mechanism of newly identified RBSs and RBPs (Extended Data Fig. 6c–h).

The most frequently identified RBSs are Cys, Tyr, Trp and Phe, followed by Met, His and Arg (Fig. 2d). These amino acids are mainly RNA base-interactors (Trp, Tyr, Phe and Arg) or efficient electron acceptors (sulfur-, aromatic ring- or π-bond-containing amino acids), which could lead to efficient UV crosslinking[25]. Of note, Cys was the most frequently identified residue. This is probably because we carefully differentiated the U-crosslinked Cys from the carbamidomethylated Cys residues (see Methods for details).

The single amino acid resolution of RBS-ID allowed a comparison with the annotated post-translational modification (PTM) sites[26]. In line with previous findings[4,8], RBSs coincide with the PTM sites (phosphorylation for Ser, Thr and Tyr, methylation for Lys and acetylation for Arg) more frequently when compared with 'non-RBSs' of the same amino acid types (Fig. 2e). Because the pre-occupancy of RBS via PTMs is likely to affect RNA binding, the PTMs may serve as regulatory codes for RNA–protein interaction and post-transcriptional regulation. Indeed, Y200 of HuR, detected by RBS-ID, was previously shown to be phosphorylated, and its phosphorylation modulates the RNA-binding activity of HuR[27].

Verification of RBS-ID using the spCas9 ribonucleoprotein (RNP) complex. In addition to the proteome-wide discovery of RBSs, we verified the RBS-ID method in a structurally known RNA–protein interaction using *Streptococcus pyogenes* Cas9 (spCas9), which forms a complex interaction surface with single guide RNA (sgRNA)[28,29]. Recombinant spCas9 and synthetic sgRNA were assembled in vitro and analyzed by RBS-ID. We identified 84 RBSs in spCas9 (Fig. 3a, Extended Data Fig. 7 and Supplementary Table 3a) along with candidate RBSs as described previously (Supplementary Table 3b). The RBSs are distributed on the RNA–protein interface of the previously reported crystal structure, supporting the validity of our findings (Fig. 3b and Extended Data Fig. 8)[28,29].

To further investigate the functional importance of the identified RBSs, we selected two RBSs (Y450 and R919) and introduced alanine mutations (Supplementary Fig. 3). In the crystal structure, Y450 was reported to interact with sgRNA in the sgRNA-DNA pairing region directly, whereas R919 is located close to the pairing region but its interaction with sgRNA has not been determined[28,29]. We transiently expressed the wild-type and mutant spCas9 proteins along with sgRNA targeting green fluorescent protein (GFP) in HEK293E cells that stably express GFP. We then analyzed the indel frequencies using the T7EI assay to monitor genome editing activity[30]. Interestingly, the mutants showed lower indel occurrence compared to the wild-type (Fig. 3c and Extended Data Figs. 9a-b and 10). Orthogonally, we measured the changes in GFP fluorescence using flow cytometry (Fig. 3d and Extended Data Figs. 9c,d and 10). The mutants were less active than wild-type spCas9 in GFP gene deletion. Together, these results suggest that the two RBSs identified by RBS-ID are functionally important for the gene editing activity of spCas9.

Discussion

In this study, we introduce chemical cleavage for RBS identification. HF treatment greatly simplifies the crosslinked RNA moiety compared with nuclease-based approaches, overcoming a major hurdle in RBS mapping. This technical advance allows us to comprehensively and precisely identify thousands of RBSs. We demonstrate RBS-ID's performance at both proteome-wide (in vivo) and single-protein (in vitro) levels. Proteome-wide RBS-ID provides a valuable resource to discover RBS candidates for functional studies. Thus far, we have discovered RBSs in ~600 proteins. The coverage and depth of RBS profiling can be increased in the future by applying enhanced chromatographic techniques prior to mass analysis and by increasing the input sample amount. On the other hand, for in-depth profiling of RBSs within a single protein of interest, RBS-ID can be performed in combination with in vitro reconstitution of purified RNA–protein complex or immunoprecipitation from cell lysates. This can substantially facilitate biochemical and structural studies of an RNA–protein complex of interest.

RBS-ID is particularly powerful in studies of RNP remodeling because we directly detect the mono-U adduct on the protein rather than relying on affinity-based pulldown, which is often prone to contamination by non-specific or transient interactors. In addition to providing evidence for direct RNA-binding, our method offers a means of quantitatively probing RNP remodeling. Although a quantitative comparison of different RBSs is largely infeasible due to the amino acid preference in UV crosslinking and the differential ionization efficiency of peptide ions, one can make comparisons of the quantities of a given RBS across different conditions. The difference would reflect the condition-dependent change in RNA–protein interaction. Although spectral counts may serve as a semi-quantitative proxy for measuring the abundance of RBSs, MS1 intensity-based label-free quantification or isotopic labeling approaches can bring more precision and accuracy (Extended Data Fig. 4). Collectively, noise-free quantification of the RNA–protein interaction by RBS-ID will help in the investigation of dynamic RNP remodeling.

RBS-ID is an amenable platform that can be modified and combined with other methods. For example, to identify other RNA bases, one can apply different crosslinking reagents such as 6-thioguanidine, which crosslinks to protein following long-wave UV irradiation. The technical principle of RBS-ID can also be applied and implemented to identify DNA-binding residues in DNA-binding proteins, by using HF to simplify DNA adducts crosslinked to DNA-binding residues. This will be a useful tool to investigate the mechanism of transcriptional or epigenetic control and DNA repair. In addition, one can couple RBS-ID with affinity purification using specific antisense oligos to examine a particular RNA–protein or DNA–protein complex in depth. Hence, RBS-ID can serve as a potent and general tool to unravel nucleic acid-protein interactions.

Online content

Any methods, additional references, Nature Research 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/s41594-020-0436-2.

Received: 9 January 2020; Accepted: 15 April 2020;
Published online: 8 June 2020

References

1. Castello, A. et al. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* **149**, 1393–1406 (2012).
2. Baltz, A. G. et al. The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol. Cell* **46**, 674–690 (2012).
3. Leitner, A., Dorn, G. & Allain, F. H. T. Combining mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy for integrative structural biology of protein–RNA complexes. *Cold Spring Harb. Perspect. Biol.* **11**, a032359 (2019).
4. Castello, A. et al. Comprehensive identification of RNA-binding domains in human cells. *Mol. Cell* **63**, 696–710 (2016).

5. He, C. et al. High-resolution mapping of RNA-binding regions in the nuclear proteome of embryonic stem cells. *Mol. Cell* **64**, 416–430 (2016).
6. 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).
7. Panhale, A. et al. CAPRI enables comparison of evolutionarily conserved RNA interacting regions. *Nat. Commun.* **10**, 2682 (2019).
8. Shchepachev, V. et al. Defining the RNA interactome by total RNA-associated protein purification. *Mol. Syst. Biol.* **15**, e8689 (2019).
9. Jeong, K., Kim, S. & Bandeira, N. False discovery rates in spectral identification. *BMC Bioinformatics* **13**(Suppl. 16), S2 (2012).
10. Bogdanow, B., Zauber, H. & Selbach, M. Systematic errors in peptide and protein identification and quantification by modified peptides. *Mol. Cell Proteom.* **15**, 2791–2801 (2016).
11. Trendel, J. et al. The human RNA-binding proteome and its dynamics during translational arrest. *Cell* **176**, 391–403 (2019).
12. Crean, C., Uvaydov, Y., Geacintov, N. E. & Shafirovich, V. Oxidation of single-stranded oligonucleotides by carbonate radical anions: generating intrastrand cross-links between guanine and thymine bases separated by cytosines. *Nucleic Acids Res.* **36**, 742–755 (2008).
13. Woo, E. M., Fenyo, D., Kwok, B. H., Funabiki, H. & Chait, B. T. Efficient identification of phosphorylation by mass spectrometric phosphopeptide fingerprinting. *Anal. Chem.* **80**, 2419–2425 (2008).
14. Na, S., Bandeira, N. & Paek, E. Fast multi-blind modification search through tandem mass spectrometry. *Mol. Cell Proteom.* **11**, 010199 (2012).
15. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* **14**, 513–520 (2017).
16. Kim, S. & Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **5**, 5277 (2014).
17. Edwards, N. J. PepArML: a meta-search peptide identification platform for tandem mass spectra. *Curr. Protoc. Bioinformatics* **44**, 13.23.1–13.23.23 (2013).
18. Chalkley, R. J. & Clauser, K. R. Modification site localization scoring: strategies and performance. *Mol. Cell Proteom.* **11**, 3–14 (2012).
19. Chang, C. et al. PANDA: a comprehensive and flexible tool for quantitative proteomics data analysis. *Bioinformatics* **35**, 898–900 (2019).
20. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
21. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).
22. Schafer, I. B. et al. Molecular basis for poly(A) RNP architecture and recognition by the Pan2-Pan3 deadenylase. *Cell* **177**, 1619–1631 (2019).
23. Kuhn, U. & Pieler, T. *Xenopus* poly(A) binding protein: functional domains in RNA binding and protein–protein interaction. *J. Mol. Biol.* **256**, 20–30 (1996).
24. Hawley, B. R., Lu, W. T., Wilczynska, A. & Bushell, M. The emerging role of RNAs in DNA damage repair. *Cell Death Differ.* **24**, 580–587 (2017).
25. Shetlar, M. D., Carbone, J., Steady, E. & Hom, K. Photochemical addition of amino acids and peptides to polyuridylic acid. *Photochem. Photobiol.* **39**, 141–144 (1984).
26. Hornbeck, P. V. et al. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* **43**, D512–D520 (2015).
27. Yoon, J. H. et al. Tyrosine phosphorylation of HuR by JAK3 triggers dissociation and degradation of HuR target mRNAs. *Nucleic Acids Res.* **42**, 1196–1208 (2014).
28. Jiang, F., Zhou, K., Ma, L., Gressel, S. & Doudna, J. A. A Cas9–guide RNA complex preorganized for target DNA recognition. *Science* **348**, 1477–1481 (2015).
29. Anders, C., Niewoehner, O., Duerst, A. & Jinek, M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* **513**, 569–573 (2014).
30. Ran, F. A. et al. Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).

Methods

Cell culture. All mammalian cell lines (HeLaT, HEK293E) used in this study were 2D-cultured in Dulbecco's modified Eagle's medium (Welgene) supplemented with 9% FBS (Welgene) and 1× antibiotic-antimycotic (Gibco), in a 37 °C incubator with a CO_2 concentration of 5%. HeLaT cells are modified HeLa cells with TUT4 gene deletion. The HeLaT cells tested negative for mycoplasma contamination. HEK293 cells were not tested for mycoplasma contamination. All cell lines were authenticated by ATCC by short tandem repeat profiling following ISO 9001:2008 and ISO/IEC 17025:2005 quality standards. HeLaT and HEK293 cells were generous gifts from laboratories in the School of Biological Sciences, Seoul National University.

RNA digestion using HF and liquid chromatography with UV detection. *Reference chemicals.* 1 mM adenine, guanine, uridine, cytidine and uracil (all Merck) were each dissolved in 50 mM Tris pH 8.0 (Ambion).

Total RNA digestion using HF. Total RNA was extracted from HeLaT cells using an RNeasy Maxi Kit (Qiagen) according to the manufacturer's instructions. An 800 µl volume of 48% HF (Merck) was carefully added to 160 µg of RNA dissolved in 200 µl of 50 mM Tris pH 8.0, 2 mM $MgCl_2$ and locked into a $CaCO_3$ (Merck) trap. The sample was incubated at 4 °C overnight and dried in a Concentrator plus system (Eppendorf), supplemented with $CaCO_3$ trap at room temperature (RT). The sample was reconstituted with 280 µl triple distilled water (TDW). All experimental processes handling HF in this paper were carried out in a fume hood.

20-mer RNA digestion using HF. The 20-mer RNA (5′-AUGCAUGCAUGCAUG CAUGC-3′) was chemically synthesized (BioNEER). An 80 µl volume of 48% HF was added to 10 µg of RNA dissolved in 20 µl of 50 mM Tris pH 8.0, 2 mM $MgCl_2$ (Ambion) and locked into a $CaCO_3$ trap. The sample was incubated at 4 °C overnight, and dried in a Concentrator plus system supplemented with a $CaCO_3$ trap at RT. The sample was reconstituted with 130 µl TDW. 20-mer RNA without HF treatment was also prepared under identical conditions.

Liquid chromatography with UV detection. Samples totaling 10 µl were injected per run. LC was carried out using a 1290 Infinity system (Agilent) equipped with an Acquity UPLC BEH C18 column (2.1 mm inner diameter (i.d.) × 30 cm, Waters). The LC flow rate was 300 µl min^{-1} and the 10 min, 20 min, 20 min and 10 min linear gradients (total of 60 min gradient) were from 100% solvent A (200 mM triethylammonium acetate (Merck)) to 80% solvent A, 70% solvent A, then 90% solvent B (90% methanol (Merck)), respectively. UV absorbance at 260 nm was monitored.

Recombinant spCas9 protein and sgRNA purification. *Recombinant spCas9 protein.* The recombinant spCas9 plasmid (pET28-His6-HA-NLS-TEV-Cas9) was a gift from J.-S. Kim (SNU, IBS). The protein was expressed in *Escherichia coli* BL21DE3 cells (Agilent) growing in Terrific broth (Merck). The cells were harvested and resuspended with 50 ml of lysis buffer (20 mM Tris pH 7.5, 300 mM NaCl (Ambion), 5 ml $CaCl_2$ (Ambion), 0.1 mM PMSF (Roche), 2 mM 2-mercaptoethanol (BME, Merck), 10% glycerol (Thermo Fisher Scientific), 1 µg ml^{-1} RNaseA (Sigma) and 2 µg ml^{-1} staphylococcal nuclease) per 750 ml culture. After sonication, the lysate was centrifuged for 1 h at 35,000*g*, 4 °C. The supernatant was collected and loaded onto 5 ml Ni-NTA Superflow beads (Qiagen) equilibrated with equilibration buffer (20 mM Tris pH 7.5, 300 mM NaCl and 2 mM BME). The beads were washed with wash buffer (20 mM Tris pH 7.5, 300 mM NaCl, 2 mM BME and 20 mM imidazole (Merck)) and the proteins were eluted with elution buffer (20 mM Tris pH 7.5, 300 mM NaCl, 2 mM BME and 200 mM imidazole (Merck)). The sample was desalted with HiPrep 26/10 (GE Healthcare) and concentrated in storage buffer (20 mM Tris pH 7.5, 150 mM NaCl and 1 mM 1,4-dithiothreitol (DTT, Merck)) with an Ultracel 100K cutoff filter (Amicon). Finally, the proteins were snap-frozen and stored at −80 °C before use.

Anti-CBX1 sgRNA. The sgRNA was prepared following a previous publication with some modifications[31]. A 100 µl volume of PCR mixture was set up with 1 µl Phusion polymerase (Thermo Fisher Scientific) in 1× HF (high fidelity) buffer (Thermo Fisher Scientific), 2 mM dNTP (Takara), 1 µM sgAmp-T7-F (5′-TAATACGACTCACTATAG-3′), 1 µM sgAmp-R (5′-AAAAAAAGCACCGA CTCGGTGC-3′), 100 nM CBX1-sgTem-F (5′-TAATACGACTCACTATAGGAA AGCTGGCGGGCACTATGGTTTAAGAGCTATGCTGGAAACA-3′) and 100 nM sgCommTem-R (5′- AAAAAAAGCACCGACTCGGTGCCACTTTTTCA AGTTGATAACGGACTAGCCTTATTTAAACTTGCTATGCTGTTTCCA GCATAGCTCTTA-3′). The thermocycler setting comprised 98 °C for 30 s; 30 cycles of 98 °C for 10 s, 57 °C for 30 s and 72 °C for 15 s; and 72 °C for 5 min. The PCR product was purified using a QIAquick PCR purification kit (Qiagen) according to the manufacturer's protocols. A 50 µl in vitro transcription mix was then set up using a T7 MEGAscript kit (Thermo Fisher Scientific) and 300 ng of PCR product according to the manufacturer's instructions, then incubated overnight. The remaining PCR products were digested with 2.5 µl TURBO DNase I (Thermo Fisher Scientific) for 30 min. Finally, the sgRNA was purified using an RNeasy MinElute Cleanup Kit (Qiagen) according to the manufacturer's instructions. NaCl was added to 150 mM and the sgRNA was snap-frozen and stored at −80 °C before use.

UV crosslinking. *In vivo UV crosslinking.* HeLaT cells were grown on 150 Φ dishes (SPL) to 70–80% confluency. The cells were washed twice with PBS (Amresco), then placed on ice in a Spectrolinker XL-1500 UV Crosslinker (Spectroline), 13 cm away from the UV lamp. The cells were irradiated with 450 mJ of 254 nm UV light.

In vitro UV crosslinking. A 60 µl volume of 7 g l^{-1} purified spCas9 in 20 mM Tris pH 7.5, 150 mM NaCl, 1 mM DTT, 18 µl of 7 g l^{-1} sgRNA in 150 mM NaCl and 9 µl of 10 mM $MgCl_2$ were mixed and incubated at RT for 10 min. The sample was then placed onto a tube cap and placed on ice in the Spectrolinker XL-1500 UV Crosslinker, 5.5 cm away from the UV lamp. The sample was irradiated twice with 600 mJ of 254 nm UV light.

HeLaT digest-HF treatment. *HeLaT digest preparation.* A 100 Φ dish of HeLaT cells grown to 70–80% confluency were lysed in 1 ml of 8 M urea (Merck) in 50 mM ammonium bicarbonate, pH 8.5 (ABC, Merck). The lysates were homogenized with a 21 G syringe and sonicated for 10 min. DTT was added to each sample to 10 mM and incubated for 1 h at 37 °C, and iodoacetamide (IAA, Merck) was added to 40 mM and incubated for 1 h at 37 °C. The sample was diluted with 50 mM ABC, pH 8.5 to 1 M urea. Next, 1 mM $CaCl_2$ and 40 µg trypsin (MS grade, Thermo Fisher Scientific) were added and incubated overnight at 37 °C. The sample was desalted using a C18 SPE cartridge (Supelco).

HF digestion and peptide clean-up. A 200 µl volume of 5 µg of HeLaT digest was reconstituted in 50 µl 50 mM Tris pH 8.0, then 800 µl of 48% HF was added and locked into a $CaCO_3$ trap. The sample was incubated at 4 °C overnight, and dried in a Concentrator plus system supplemented with a $CaCO_3$ trap at RT. The sample was reconstituted with TDW and cleaned up using ZipTip (Millipore). The eluted peptides were dried in the Concentrator plus at RT, and reconstituted in 50 mM ABC, pH 8.5. Following centrifugation at 13,000*g* for 5 min at RT, the supernatant was collected and one-fifth was injected for LC-MS/MS.

RBS enrichment. *Total RNA-RBS.* Three 150 Φ dishes of UV-crosslinked HeLaT cells were each lysed on-dish using 1.2 ml RIPA buffer (Thermo Fisher Scientific) supplemented with 18 µl SUPERase-In (Ambion). The lysates were homogenized with a 21 G syringe; DTT was added to each sample to 10 mM and incubated for 1 h at 37 °C, then IAA was added to 40 mM and incubated for 1 h at 37 °C. Next, 1 mM $CaCl_2$ and 20 µg trypsin (2%, wt/wt ratio) was added and incubated for 4 h at 37 °C. A 6 ml volume of buffer RLT (Qiagen) supplemented with 40 mM DTT and 4.2 ml 100% ethanol was added to each sample. Then, RNA-peptide conjugates were purified using an RNeasy Midi (Qiagen) according to the manufacturer's instructions, except that the centrifugation speed at binding was 500*g*. The conjugates were eluted in 420 µl TDW, then 50 mM Tris pH 8.0, 1 mM $CaCl_2$ and 0.2 µg trypsin were added. The samples were incubated overnight at 37 °C and dried to 200 µl in a Concentrator plus at RT. Three samples were merged and 6 ml buffer RLT supplemented with 40 mM DTT and 5.4 ml 100% ethanol was added. RNA-peptide conjugates were purified using an RNeasy Midi according to the manufacturer's instructions, except that the centrifugation speed at binding was 500*g* and RW1 buffer was omitted. The RNA-peptide conjugates were eluted in 240 µl TDW and 50 mM Tris pH 8.0 was added. Then, 1 ml of 48% HF was added and locked into a $CaCO_3$ trap. The sample was incubated at 4 °C overnight, and dried in a Concentrator plus supplemented with a $CaCO_3$ trap at RT. The remaining steps, including HF incubation and peptide clean-up, were identical to that of HeLaT digest-HF treatment and clean-up.

mRNA-RBS. Solutions were prepared following a previous publication with some modifications[1]. Ten dishes of 150 Φ of UV-crosslinked HeLaT cells were individually lysed on-dish using 3.5 ml lysis buffer (20 mM Tris pH 7.5, 500 mM LiCl (Merck), 1 mM EDTA (Ambion), 5 mM DTT, 0.5% IGEPAL-CA630 (Merck) and 0.5% LiDS (Merck)). The lysates were merged and homogenized with a 21 G syringe and added to 5 ml oligo-dT(25) beads (New England Biolabs) pre-equilibrated with lysis buffer. The sample was incubated for 1 h at RT. The supernatant was then removed and the beads were washed using wash buffer 1 (20 mM Tris pH 7.5, 500 mM LiCl, 1 mM EDTA, 5 mM DTT, 0.5% IGEPAL-CA630 and 0.1% LiDS) twice and wash buffer 2 (20 mM Tris pH 7.5, 500 mM LiCl, 1 mM EDTA and 5 mM DTT) once at RT. The beads were moved to a new tube at the initial wash with wash buffer 1. Following washes, 2.5 ml of digest buffer (50 mM Tris pH 8.0 and 1 mM $CaCl_2$) was added. Then, alkylation of cysteine residues and trypsinization were carried out in similar manner to that described in the Total RNA-RBS section. Following incubation, 20 ml buffer RLT supplemented with 40 mM DTT and 15 ml 100% ethanol was added. RNA-peptide conjugates were purified using an RNeasy Midi according to the manufacturer's instructions, except that the centrifugation speed at binding was 500*g* and RW1 buffer was omitted. The RNA-peptide conjugates were eluted in 240 µl TDW and 50 mM Tris pH 8.0 was added. The remaining steps including HF treatment and peptide clean-up were identical to those for HeLaT digest-HF treatment and clean-up.

spCas9-RBS. UV-crosslinked spCas9 RNP was moved to a new 1.5 ml tube; 50 mM Tris and 1 mM $CaCl_2$ was added to each sample, followed by 10 mM DTT, then incubated for 1 h at 37 °C. IAA (40 mM) was then added and incubated for 1 h at 37 °C, followed by 14 µg trypsin. The samples were incubated for 4 h at 37 °C, then, 700 µl of buffer RLT and 600 µl of 100% ethanol were added. The RNA-peptide conjugates were purified using an RNeasy MinElute Cleanup Kit according to the manufacturer's instructions, except that the elution was performed twice. Two samples were merged and 50 mM Tris pH 8.0 and TDW were added to a final volume of 100 µl. Then, 400 µl of 48% HF was added and locked into a $CaCO_3$ trap. The remaining steps, including HF incubation and peptide clean-up, were identical to those for HeLaT digest-HF treatment and clean-up.

LC-MS/MS and data processing. *LC-MS/MS for RBS-ID.* LC-MS/MS analysis was carried out using an Orbitrap Fusion Lumos Tribrid MS set-up (Thermo Fisher Scientific) coupled with a nanoAcquity system (Waters) equipped with an in-house packed capillary analytical column (75 µm i.d. × 100 cm) and trap column (150 µm i.d. × 3 cm) with 3 µm Jupiter C18 particles (Phenomenex). The LC flow rate was 300 nl min^{-1} and the 60 min (spCas9-RBS) or 100 min (total RNA-RBS and mRNA-RBS) linear gradient ranged from 95% solvent A (0.1% formic acid (Merck)) to 40% solvent B (100% acetonitrile, 0.1% formic acid). Full MS scans (m/z 300–1,500) were acquired at a resolution of 60k (at m/z 200). Higher-energy collisional dissociation (HCD) fragmentation was performed under 30% of normalized collision energy (NCE) via precursor isolation within 1.4 Th of window. The MS2 scans were acquired at a resolution of 15k ((maximum precursor ion injection time (ITmax) of 30 ms and automatic gain control (AGC) of 1×10^4)).

LC-MS/MS for HeLaT digest-HF treatment. LC-MS/MS analysis was carried out using a Q Exactive MS set-up (Thermo Fisher Scientific) coupled with LC, as above, using a 100 min linear gradient. Full MS scans (m/z 350–1,800) were acquired at a resolution of 70k (at m/z 200). HCD fragmentation was performed under 25% of NCE via precursor isolation 19 within 2.0 Th of window. The MS2 scans were acquired at a resolution of 17.5k (ITmax of 60 ms and AGC of 1×10^6).

Open search. Peak lists were generated from raw files as mzXML files using RawConverter[32], as data-dependent acquisition of charge states 2–7. For an MSFragger search, the mzXML files were converted into mzML files using MSConverter[33]. A MODa (v1.6.0)[14] search was performed on modification sizes 0–400 and 400–800, separately. The target-decoy database generated by MS-GF+[16] from the Swiss-Prot human database (May 2019)[20] was used. The PSM-level false discovery rate (FDR) was set to <1%. We combined moda.ptm files from the two searches to analyze the frequent modified mass. For modifications on Cys, 57 was added to the modified mass to account for fixed modification (carbamidomethylation). An MSFragger (ver.20190530)[15] open search via Fragpipe was performed with mzML files on modification sizes 0–400 and 400–800, separately. The target-decoy database used was identical to that used for MODa, and the FDR of the PSM sorted by hyperscore was set to <1%. We combined massdiff values in the tsv files from the two searches to analyze the frequently modified mass. Adjustment of modification on Cys was not performed, as MSFragger open search cannot specify the modified site within peptides. The Thermo Xcalibur Qual Browser (Thermo Fisher Scientific) and its boxcar method for peak smoothing were used for visualization of the extracted ion chromatogram of precursor ions.

Closed search for RBS-ID. Peak list generation and file conversion were performed as in open search. An MS-GF+ search was performed with mzXML files. The fixed modification considered was $C_2H_3N_1O_1$ (carbamidomethylation) on Cys at any residue. The variable modifications considered were $C_7H_9N_1O_5$ (uridine minus carbamidomethylation) on Cys and $C_9H_{12}N_2O_6$ (uridine) on the remaining 19 amino acids at any residue to distinguish crosslinked uridine on Cys residues from carbamidomethylation of uncrosslinked free Cys residues by IAA. Only one modification was allowed per peptide; for analysis of multiple RBSs on peptides, two modifications per peptide were allowed. The Swiss-Prot human database (May 2019, total RNA-RBS plus mRNA-RBS) or the UniProt *E. coli* BL21DE3 database supplemented with the spCas9 (UniProt accession Q99ZW2) sequence (spCas9-RBS) was used, and the decoy database was generated by MS-GF+. The output mzid file was converted to a tsv file for post-processing.

Further processing for stringent RBS localization with scores. For each experiment, PSMs within a ±5 ppm window from the mean precursor isotopic error of highly scored PSMs (for example, MS-GF+ $Q=0$) were collected. RBS-containing PSMs with a peptide-level FDR of <0.01 were taken. PSMs where an RBS was not uniquely assigned were discarded to remove any ambiguity (including those localized to the C-terminal Arg/Lys of peptides). The localization scores of RBSs were calculated in a similar manner as in PepArML[17]: for individual RBSs in peptides as (sum of $-\log_{10}$(SpecEValue from MS-GF+) of each RBS PSM corresponding to the individual peptide) divided by (sum of $-\log_{10}$(SpecEValue from MS-GF+) of all RBS PSMs corresponding to the individual peptide), rounded up to the second decimal point. For RBSs identified from multiple peptides (that is, missed cleavage products), the corresponding peptides were integrated. Next, for peptides with different RBS localizations in their sequence, we considered only a single RBS localization with exclusive maximum spectral counts for stringency. RBSs with non-exclusive maximum spectral counts in peptides were considered as candidate RBSs. RBS locations within proteins were mapped to the proteome databases used above. For total RNA-RBS and mRNA-RBS experiments, MS1 intensity-based label-free quantification was performed with PANDA[19], using. mzid files from MS-GF+ and Thermo.RAW files as input. Intensity value sums of differently charged peptide ions corresponding to each RBS were integrated from PeptideIons.txt files. For comparison of MS1 intensity-based label-free quantification between replicates, Pearson's correlation coefficient and Spearman's correlation coefficient were each calculated and rounded up to the second decimal point. Then, PSM counts, localization scores and MS1 intensities of four experiments (two total RNA-RBS, two mRNA-RBS) or two experiments (spCas9-RBS) were merged together. For spCas9-RBS, only RBSs mapped to the spCas9 sequence were considered.

Integration of previous studies and databases. Identified total RNA-RBS and mRNA-RBS were mapped to annotated RBPs in UniProtKB (June 2019), annotated domains in UniProt PROSITE (June 2019), annotated UniProt zinc finger and coiled coil (August 2019) if unmapped to UniProt PROSITE and annotated PTM sites in PhosphoSitePlus (June 2019)[26]. Fisher's exact test (two-sided) was performed on each modification in RBS-containing proteins, grouped by RBS or non-RBS. GO term analysis was performed with DAVID 6.8[21] using the leading protein groups to which RBSs belong. The crystal structure of yeast Pab1 in complex with poly(A) RNA was obtained from the Protein Data Bank (PDB code 6R5K)[22]. Crystal structures of the spCas9–sgRNA complex with RBS atoms were obtained from the Protein Data Bank (PDB codes 4TZ0[28] and 4UN3[29]) using PyMOL (Schrodinger, version 1.7.2.1.). Also, MS2 spectra were visualized using LcMsSpectator (Pacific Northwest National Laboratory).

Closed search for HeLaT digest-HF treatment. Peak list generation and file conversion were performed as in open search. A semi-tryptic MS-GF+ search was performed with mzXML files. The fixed modification considered was $C_2H_3N_1O_1$ (carbamidomethylation) on Cys at any residue and no variable modifications were considered. The Swiss-Prot human database (May 2019) was used, and the decoy database was generated by MS-GF+. The output mzid file was converted to a tsv file. PSM-level FDR was set to <1%, and the proportion of semi-tryptic PSMs and the proportion of PSMs below FDR were calculated.

The MS proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository[34] with the dataset identifier PXD016254.

Measurement of spCas9 gene editing efficiency. *Cloning.* The FLAG-Strep-msfGFP construct was cloned into a modified pLVX vector. Anti-msfGFP sgRNA was cloned into pSpCas9(BB) (PX459, Addgene 62988). The oligos used were sgF (5′-CACCGCTGAAGTTCATCTGCACCAC-3′) and sgR (5′-AAACGTGGTGCAGATGAACTTCAGC-3′). Plasmids before and after sgRNA integration were referred to as non-target-wild-type (NC) and wild-type (WT), respectively. Then, Y450A or R919A point mutations were introduced to the anti-msfGFP pspCas9(BB) plasmids. Sanger sequencing of the plasmids was performed (Cosmogenetech) and visualized using Benchling.

msfGFP-expressing clonal stable HEK293E cell line. For lentivirus production, 10 µg of pLVX-FLAG-Strep-msfGFP was co-transfected with 7 µg psPAX2, 3 µg pMD2.G and 4 µg pHIV-REV into HEK293E cells grown in a 100 Φ dish using Fugene HD transfection reagent (Promega). Three days after transfection, the lentivirus was harvested by passing through a 0.45 µm filter and concentrated with a Lenti-X concentrator (Clontech). The virus was resuspended in OPTI-MEM medium (Gibco) and transduced into HEK293E cells at MOI > 10. The cells were selected with puromycin for seven days and stored as stocks in liquid nitrogen. A stock was then thawed and a single clone selected.

T7EI assay. This experiment was conducted following a previous publication with some modifications[30]. msfGFP-expressing clonal stable HEK293E cells growing in 12-well plates were transfected with TDW (for western blot only) or 1 µg pSpCas9(BB) plasmids (NC, WT, Y450A, R919A). At 48 h after transfection, the cells were trypsinized and harvested in PBS. Approximately 1×10^4 cells were used for gRNA extraction using 500 µl of QuickExtract DNA extraction solution (Epicentre) according to the manufacturer's instructions for western blot. For the T7E1 assay, 50 µl PCR mixture was set up with 0.5 µl Phusion polymerase in 1× HF buffer, 200 µM dNTP, 0.5 µM msfGFP-F (5′-ATGTCTAAGGGCGAGGAACTGTTC-3′), 0.5 µM msfGFP-R (5′-GGTCACAAATTCCAGCAGCAC-3′) and 5 µl of gDNA extract. The thermocycler setting consisted of 95 °C for 30 s; 35 cycles of 95 °C for 15 s and 72 °C for 30 s; 72 °C for 5 min. A 10 µl volume of PCR amplicon was mixed with 2 µl of 10× DNA loading dye and run on 1.5% EtBr-agarose gel as input. Next, 3 µl of 5× NEBuffer 2 (New England Biolabs) was added to 10 µl of PCR amplicon. The samples were heated to 95 °C and ramped down to 25 °C at −2 °C min^{-1}. Then, 2 U of T7 endonuclease I (New England Biolabs) in 2 µl 1× NEBuffer 2 was added. The

mixture was incubated for 1 h at 37 °C. The samples were then mixed with 2 µl of 10× DNA loading dye and run on 1.5% EtBr-agarose gel. The intensities of bands corresponding to ~675 bp (uncut) and ~543 bp (cut) were calculated using ImageJ, and normalized by fragment length. Then, the background intensity corresponding to uncut fragments in NC were removed from other samples. The cut fraction (normalized for fragment length) was calculated as (cut fragment intensity × 675)/(uncut fragment intensity × 543). Finally, the indel occurrence was calculated as $100 \times (1 - (1 - \text{cut fraction})^{1/2})$.

Western blot. The remaining cells from above were pelleted and lysed with 200 µl RIPA buffer and 60 µl 5× SDS sample buffer (250 mM Tris pH 6.8, 10% SDS, 500 mM DTT, 50% glycerol, 0.6 g l^{-1} bromophenol blue). The samples were run on SDS-PAGE and transferred to a nitrocellulose membrane. A western blot was performed with anti-FLAG M2 (Merck) and anti-GAPDH (Santa Cruz) antibodies, and visualized with horseradish peroxidase conjugated anti-mouse secondary antibody (Jackson Lab). The intensities of bands corresponding to 3XFLAG-NLS-spCas9-puro and GAPDH were measured using ImageJ, and the relative band intensity of the former compared to the latter was calculated.

Flow cytometry. msfGFP-expressing clonal stable HEK293E cells growing in 12-well plates were transfected with 1 µg pSpCas9(BB) plasmids (NC, WT, Y450A, R919A). At 48 h after transfection, the cells were trypsinized and harvested in PBS. The GFP fluorescence of cells was measured using a BD Accuri C6 flow cytometer (BD Biosciences). Signals from live cells were chosen by gating cell size and shape. On measuring the GFP fluorescence of more than 20,000 cells per sample, a threshold below which 0.5% of cells in NC exhibit GFP fluorescence was defined. The proportion of cells below the threshold in WT, Y450A and R919 samples was calculated.

Statistics. For Fig. 2e, a two-sided Fisher's exact test was performed between $n = 1$ groups of RBS and non-RBS. *P* values were rounded up to the fourth decimal point: 0.0000 (STY-phosphorylation), 0.0000 (K-acetylation) and 0.0002 (R-methylation). For Fig. 3c, a two-sided unpaired Student's *t*-test assuming equal variance was performed between $n = 3$ biologically independent samples. The *t* values were rounded up to the fourth decimal point: 5.3032 (WT versus Y450A), 7.9965 (WT versus R919A). *P* values were rounded up to the fourth decimal point: 0.0061 (WT versus Y450A), 0.0013 (WT versus R919A). There were four degrees of freedom. For Fig. 3d, a two-sided unpaired Student's *t*-test was performed as described above. The *t* values rounded up to the fourth decimal point: 17.7773 (WT versus Y450A), 19.0390 (WT versus R919A). *P* values were rounded up to the fourth decimal point: 0.0001 (WT versus Y450A), 0.0000 (WT versus R919A). There were four degrees of freedom. For Extended Data Fig. 1d, a two-sided unpaired Student's *t*-test was performed as described above. The *t* value was rounded up to the fourth decimal point: −3.4612. The *P* value was rounded up to the fourth decimal point: 0.0258. There were four degrees of freedom. For Extended Data Fig. 1e, a two-sided unpaired Student's *t*-test was performed as described above. The *t* value was rounded up to the fourth decimal point: 1.8632. The *P* value was rounded up to the fourth decimal point: 0.1359. There were four degrees of freedom. For Extended Data Fig. 10b, a two-sided unpaired Student's *t*-test was performed as described above. The *t* values were rounded up to the fourth decimal point: −1.1182 (WT versus Y450A), −1.4310 (WT versus R919A). The *P* values were rounded up to the fourth decimal point: 0.3261 (WT versus Y450A), 0.2257 (WT versus R919A). There were four degrees of freedom.

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

Data availability

MS data have been deposited at the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with dataset identifier PXD016254. Source data for Figs. 2a, 2d, 2e, 3c and 3d and Extended Data Figs. 9a, 9b and 10a are available with the paper online.

References

31. Leonetti, M. D., Sekine, S., Kamiyama, D., Weissman, J. S. & Huang, B. A scalable strategy for high-throughput GFP tagging of endogenous human proteins. *Proc. Natl Acad. Sci. USA* **113**, E3501–3508 (2016).
32. He, L., Diedrich, J., Chu, Y. Y. & Yates, J. R. III. Extracting accurate precursor information for tandem mass spectra by RawConverter. *Anal. Chem.* **87**, 11361–11367 (2015).
33. Chambers, M. C. et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).
34. Vizcaino, J. A. et al. The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.* **41**, D1063–D1069 (2013).

Acknowledgements

We thank S. Shin, K. Baeg and S. Lee for insightful comments and discussion. We are also grateful to J. Kim, J. Yang, D. Choi and E. Kim for technical help, and all members of our laboratories for helpful discussion. We thank J. S. Kim (Seoul National University), Pacific Northwest National Laboratory and the OMICS.PNL.GOV for providing valuable plasmid and software. This work was supported by IBS-R008-D1 of the Institute for Basic Science from the Ministry of Science and ICT of Korea (J.W.B., S.-C.K., Y.N., V.N.K. and J.-S.K.) and BK21 Research Fellowships (J.W.B.) from the Ministry of Education, Science and Technology of Korea.

Author contributions

J.W.B., V.N.K. and J.-S.K. conceived the project and designed the experiments. J.W.B. developed the protocol and performed all biochemical experiments with the support of S.C.K. and Y.N. J.W.B. generated and analyzed all LC-MS/MS datasets. J.W.B., V.N.K. and J.-S.K. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

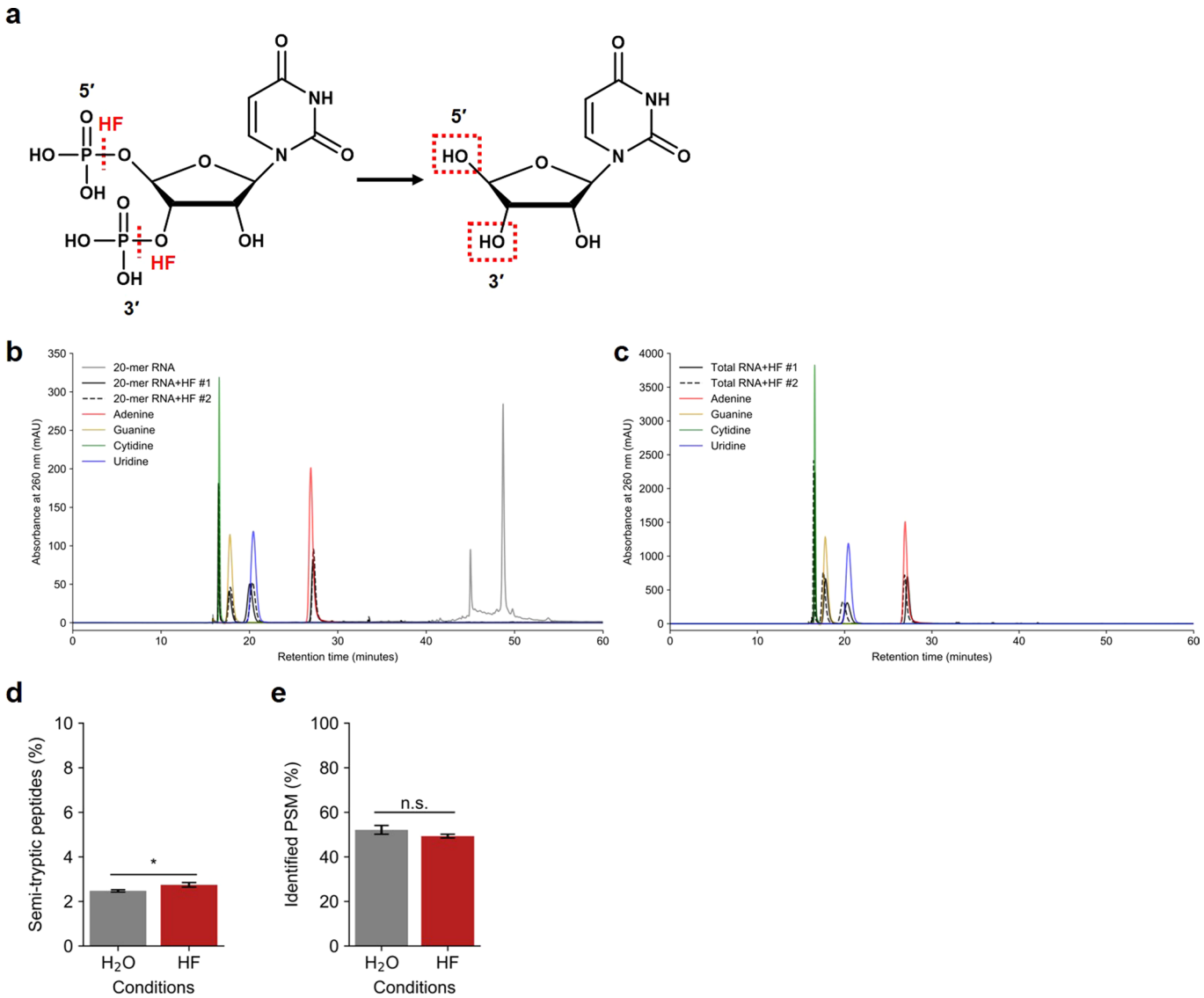
Extended data is available for this paper at https://doi.org/10.1038/s41594-020-0436-2.

Supplementary information is available for this paper at https://doi.org/10.1038/s41594-020-0436-2.

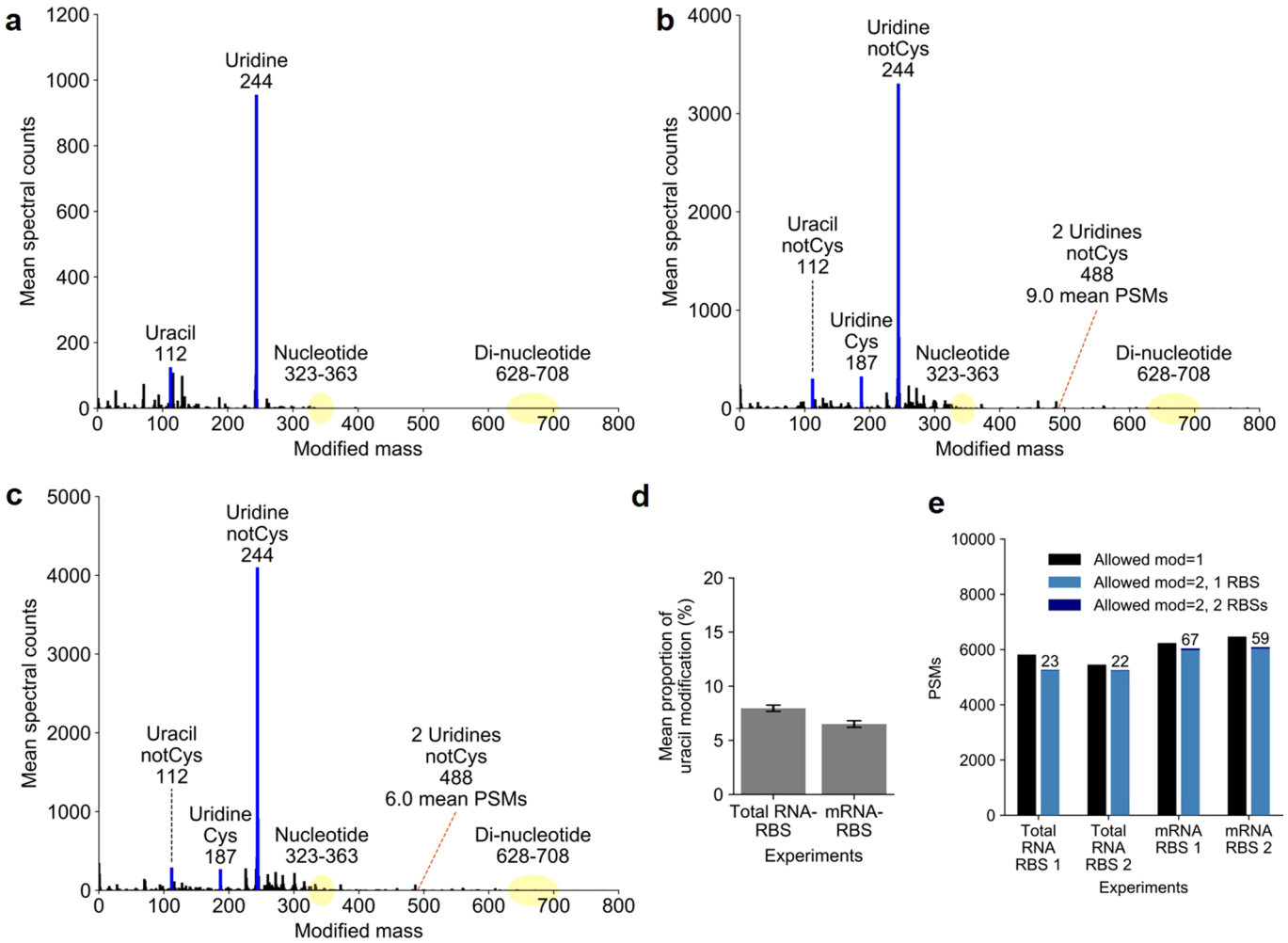
Correspondence and requests for materials should be addressed to V.N.K. or J.-S.K.

Peer review information Peer reviewer reports are available. Anke Sparmann was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

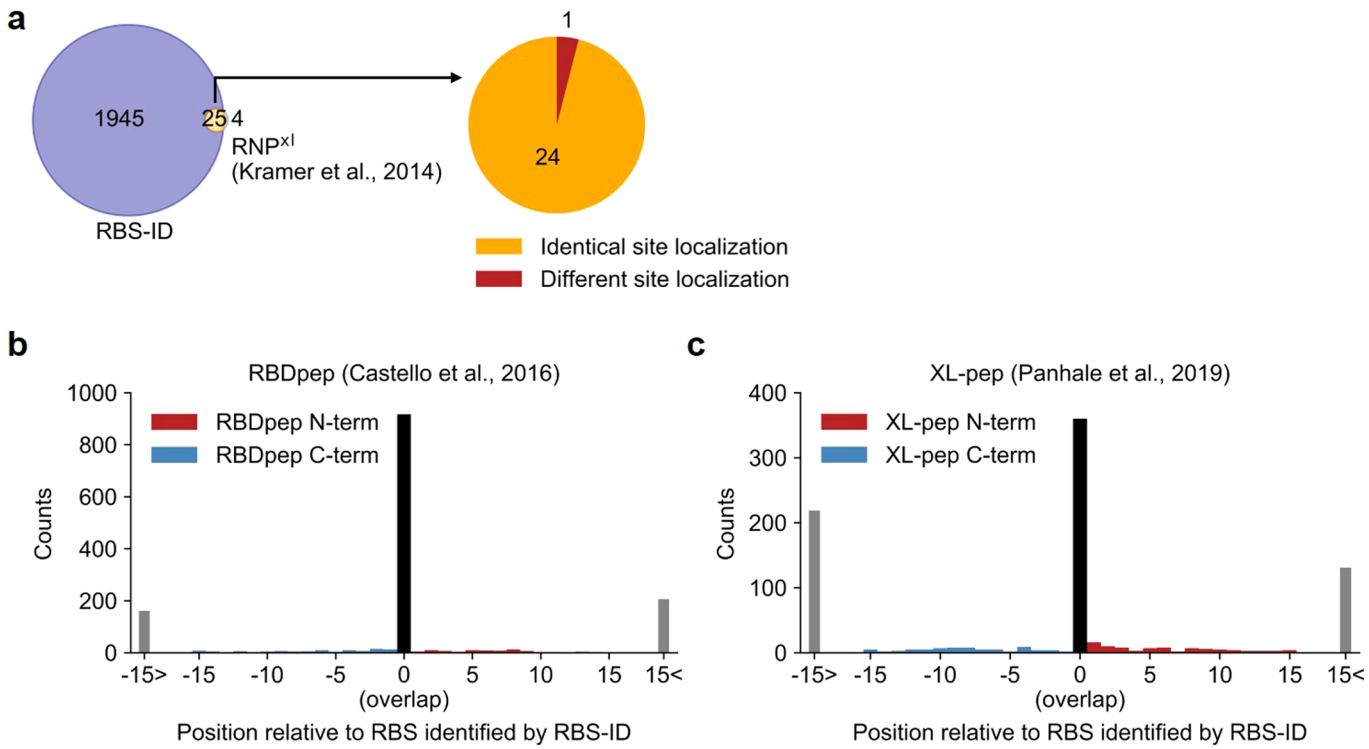
Reprints and permissions information is available at www.nature.com/reprints.



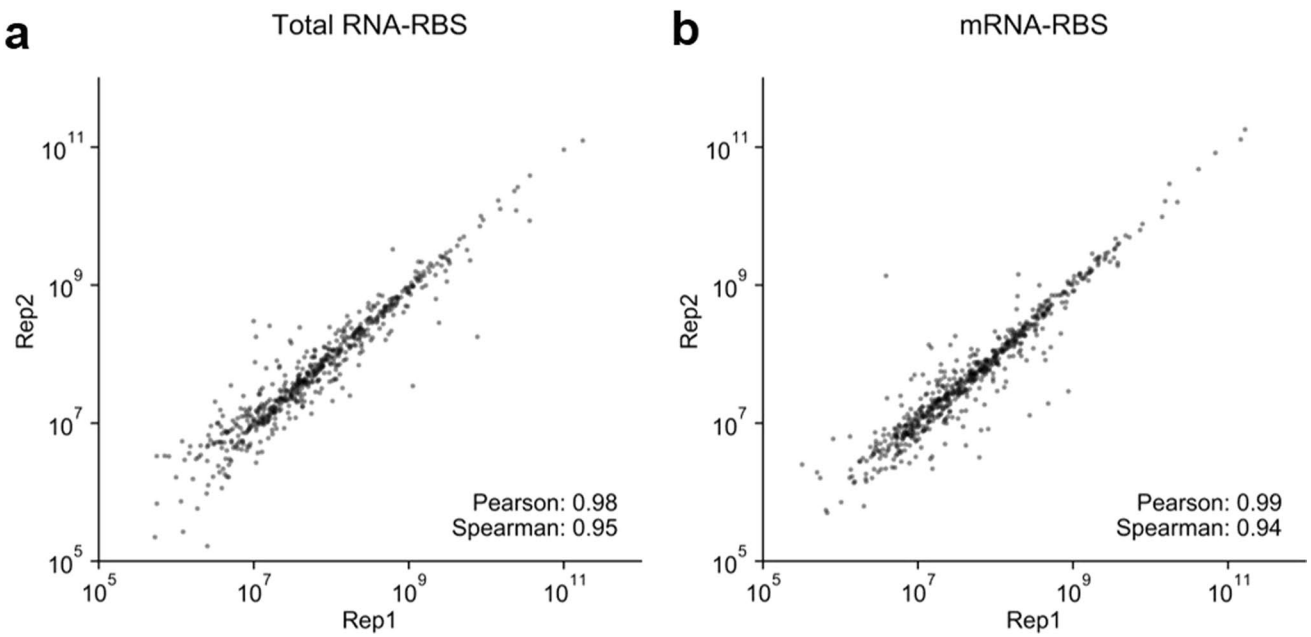
Extended Data Fig. 1 | HF treatment on RNA and peptides. a, Schematic illustration of HF treatment on RNA (left) and the resulting products (right). HF-mediated cleavage sites are highlighted. **b**, UV-absorbance chromatogram of 20-mer RNA (AUGCAUGCAUGCAUGCAUGC) digested with HF (black solid and dashed lines from duplicate experiments), merged with those of undigested RNA (gray) and reference chemicals (colored, solid). Peaks of reference chemicals were re-sized for better visualization. The black solid line is the source data for Fig. 1b. **c**, UV-absorbance chromatogram of HeLaT total RNA digested with HF (black, solid or dashed lines from duplicate experiments), merged with those of reference chemicals (colored, solid). Peaks of reference chemicals were re-sized for better visualization. **d**, Proportion of semi-tryptic PSM identified from HeLaT digest peptides upon HF treatment, compared to negative control treated with H_2O. Two-sided unpaired Student's T-test was performed between n = 3 biologically independent samples (H_2O vs. HF), assuming equal variance. The mean values were depicted with error bars that indicate standard deviation between replicates. P-value, rounded up to the fourth decimal point: 0.0258. **e**, Proportion of identified PSM (below PSM-level FDR = 0.01) upon HF treatment, compared to H_2O treatment. Two-sided unpaired Student's T-test was performed as described above. The mean values were depicted with error bars that indicate standard deviation between replicates. P-value, rounded up to the fourth decimal point: 0.1359.



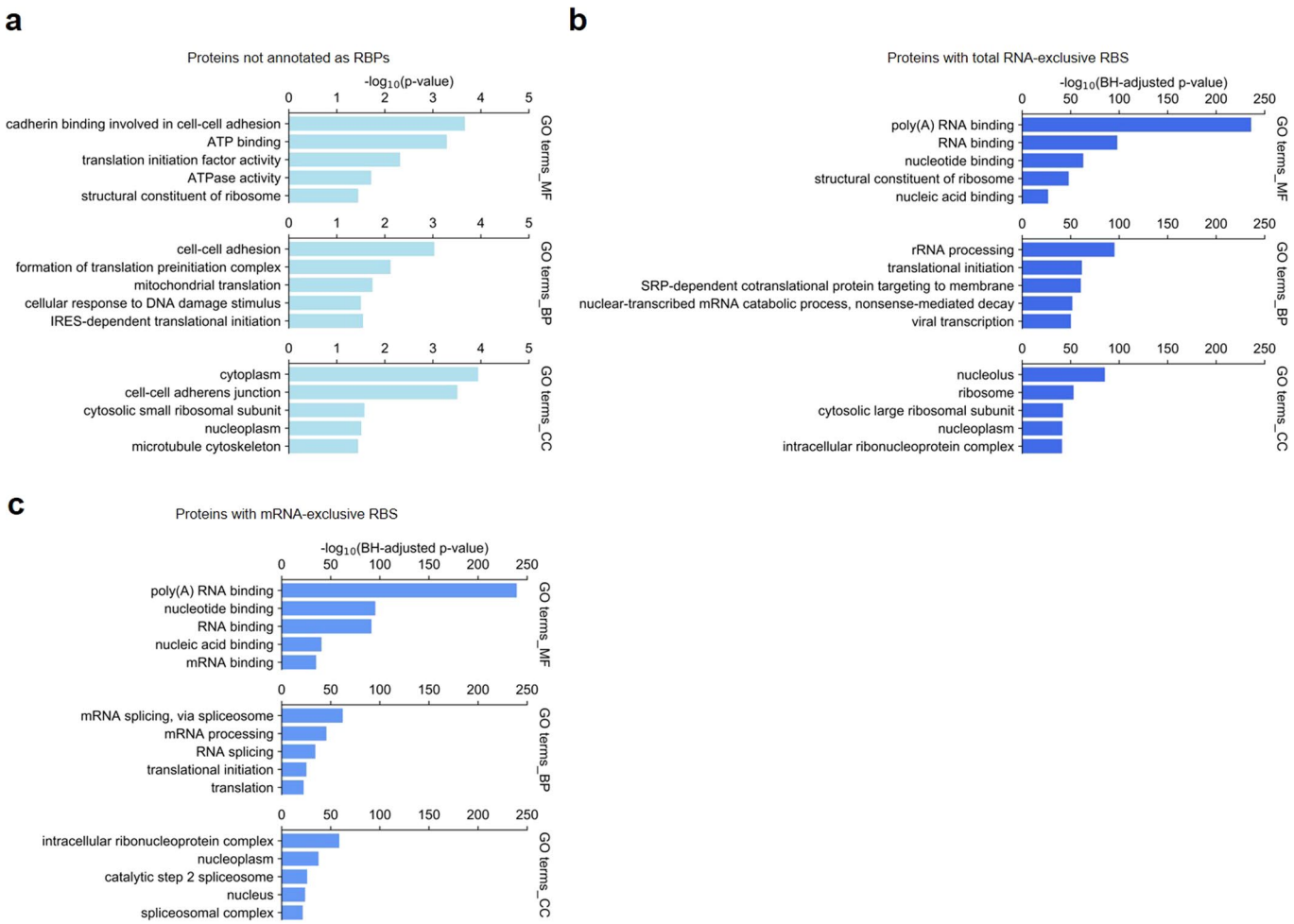
Extended Data Fig. 2 | Open search on total RNA-RBS and mRNA-RBS using MODa and MSFragger. a, MODa[14] search for modified mass on mRNA-RBS. The y-axis indicates mean spectral counts from duplicate experiments. **b**-**c**, MSFragger[15] search for modified mass on total RNA-RBS (**b**) and mRNA-RBS (**c**). **d**, Mean percentage of Uracil modification over Uridine modification, a highly likely in-source fragmentation product, between replicate experiments, calculated from MSFragger search results as (#PSM with Uracil adduct: 112 & 55 to account for Cys)/(#PSM with Uracil or Uridine adduct: 244, 112 & 187, 55 to account for Cys). The error bars indicate standard deviation. Free Cys was carbamidomethylated, so the corresponding adduct mass was used as a fixed modification. Owing to mutually exclusive U-crosslinking and carbamidomethylation on Cys, the observed conjugate mass of uridine on Cys (187) was smaller than that of other amino acids (244) by the difference of mass of carbamidomethyl group (57). Thus, modification mass on Cys was corrected by adding the mass of carbamidomethyl group. The percentages were rounded up to the second decimal point. **e**, Comparison of closed search results on total RNA and mRNA RBS-ID experiments, allowing up to one or two modifications per peptide. Modification-specific peptide-level FDR was set to 0.01. PSM counts for peptides with two modifications were depicted.



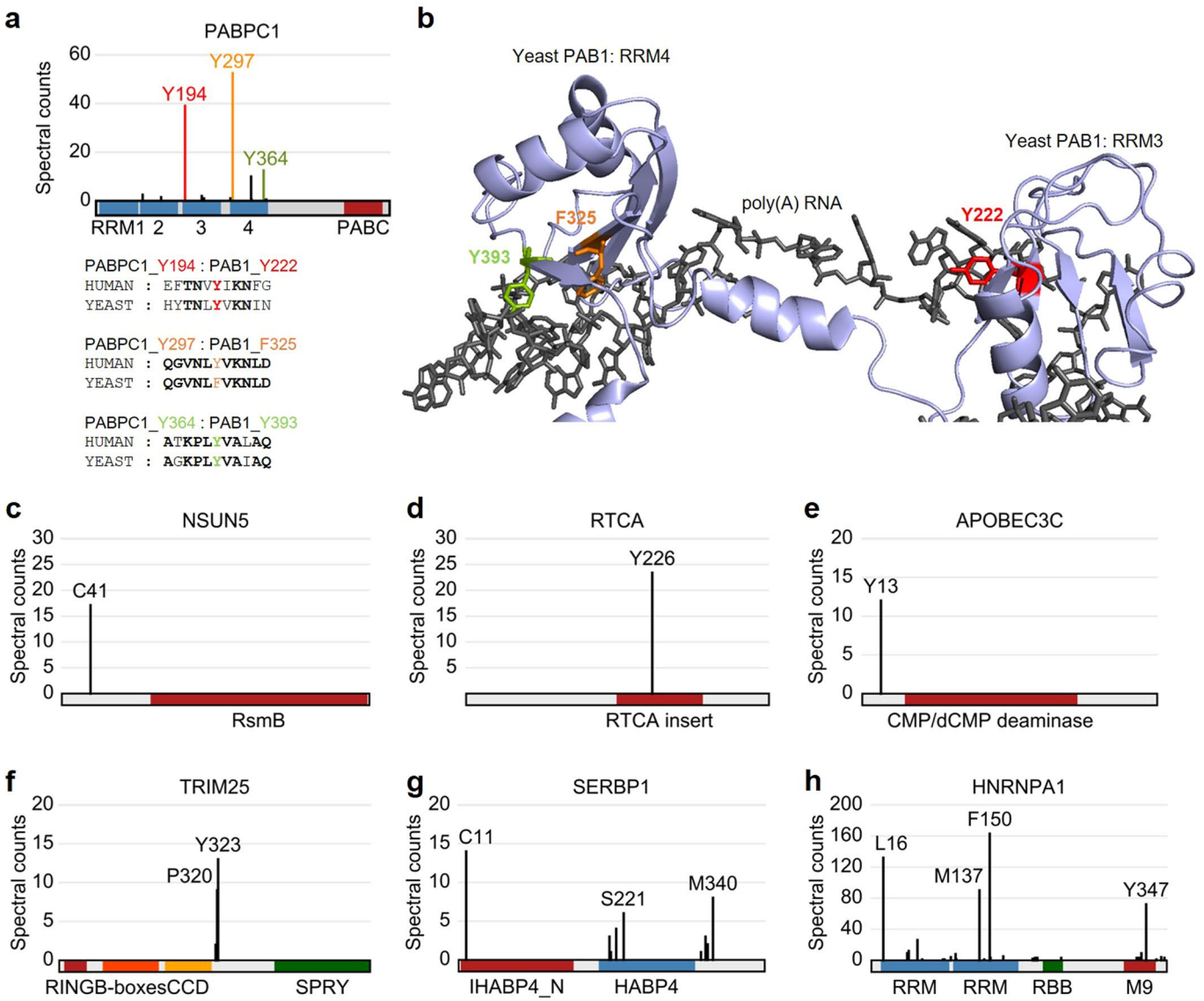
Extended Data Fig. 3 | Comparison between RBS-ID and previous RBS- or RBD-profiling studies. a, Comparison of the position of RBS in peptides shared between RBS-ID and RNP^{xl} Venn diagram (left) shows the overlap between the peptides with uniquely localized RBS in RBS-ID (1,972) and those in human proteins of RNP^{xl} (29)[6]. Please note that most peptides from RNP^{xl} are also detected by RBS-ID, demonstrating the comprehensiveness of our method. The Pie chart (right) displays that among 25 common peptides, 24 peptides show consistent localization of RBS between the datasets, indicating the accuracy of the methods. **b**, Relative position of the peptides identified as 'RBDpep'[4]. X-axis shows the position of the terminus of peptides relative to RBSs identified by RBS-ID (n = 1,478). **c**, Relative position of the peptides identified as 'XL-peptide'[7]. X-axis shows the position of the terminus of peptides relative to RBSs identified by RBS-ID (n = 869).



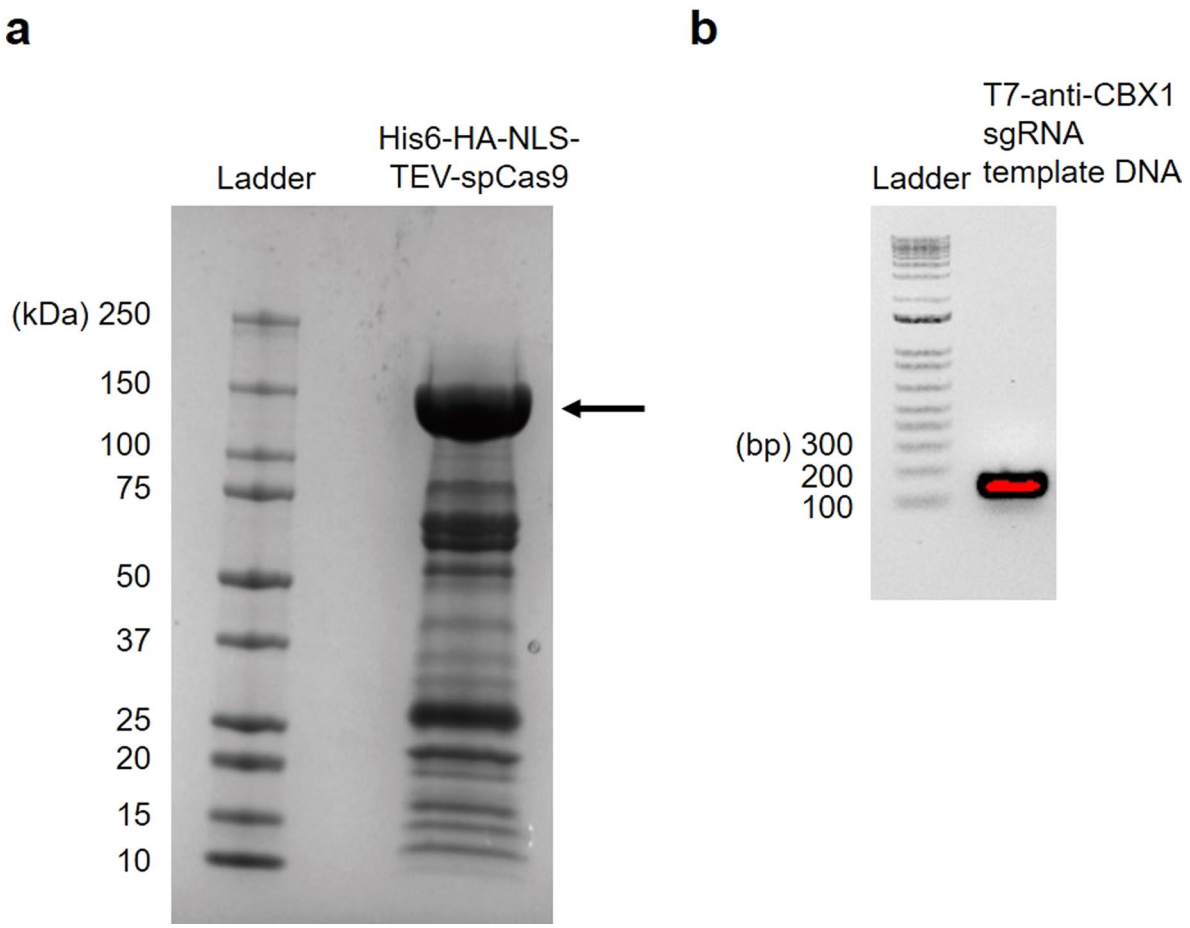
Extended Data Fig. 4 | Reproducibility of MS1 intensity-based label-free quantification. a-**b**, MS1 intensity-based label-free quantification[27] of RBS-containing peptides co-identified in Total RNA-RBS (**a**) or mRNA-RBS (**b**) replicate experiments. Pearson's correlation coefficient and Spearman's correlation coefficient were each calculated and rounded up to the second decimal point.



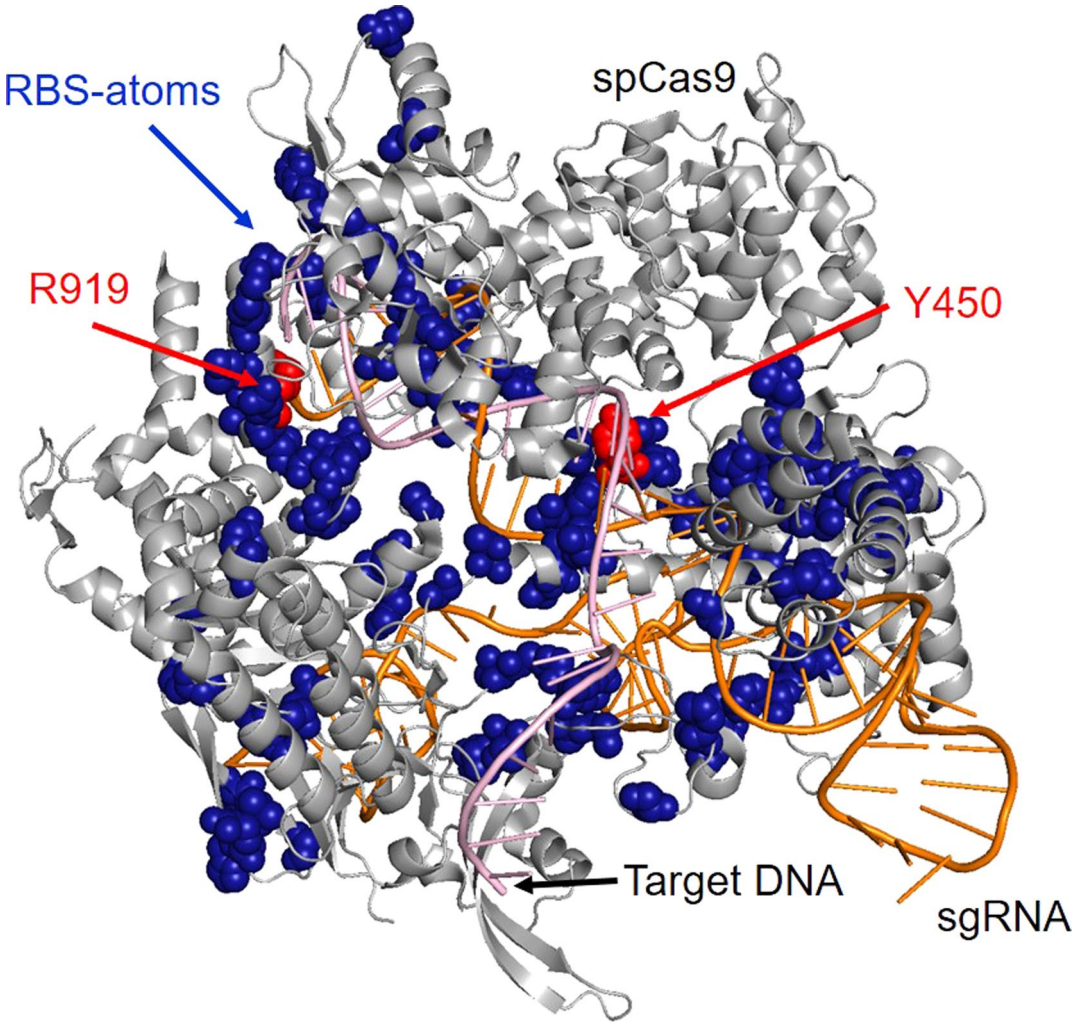
Extended Data Fig. 5 | RBS-identified protein groups and regions. a, Top 5 GO terms associated with proteins that are not annotated as RBPs (MF: molecular function, BP: biological process, CC: Cellular component, 5 each)[20,21]. **b**-**c**, Top 5 GO terms associated with proteins whose RBSs are identified exclusively in total RNA enrichment but not in poly(A) + RNA enrichment (**b**) or in poly(A) + RNA enrichment but not in total RNA enrichment (**c**) (MF, BP, CC, 5 each).



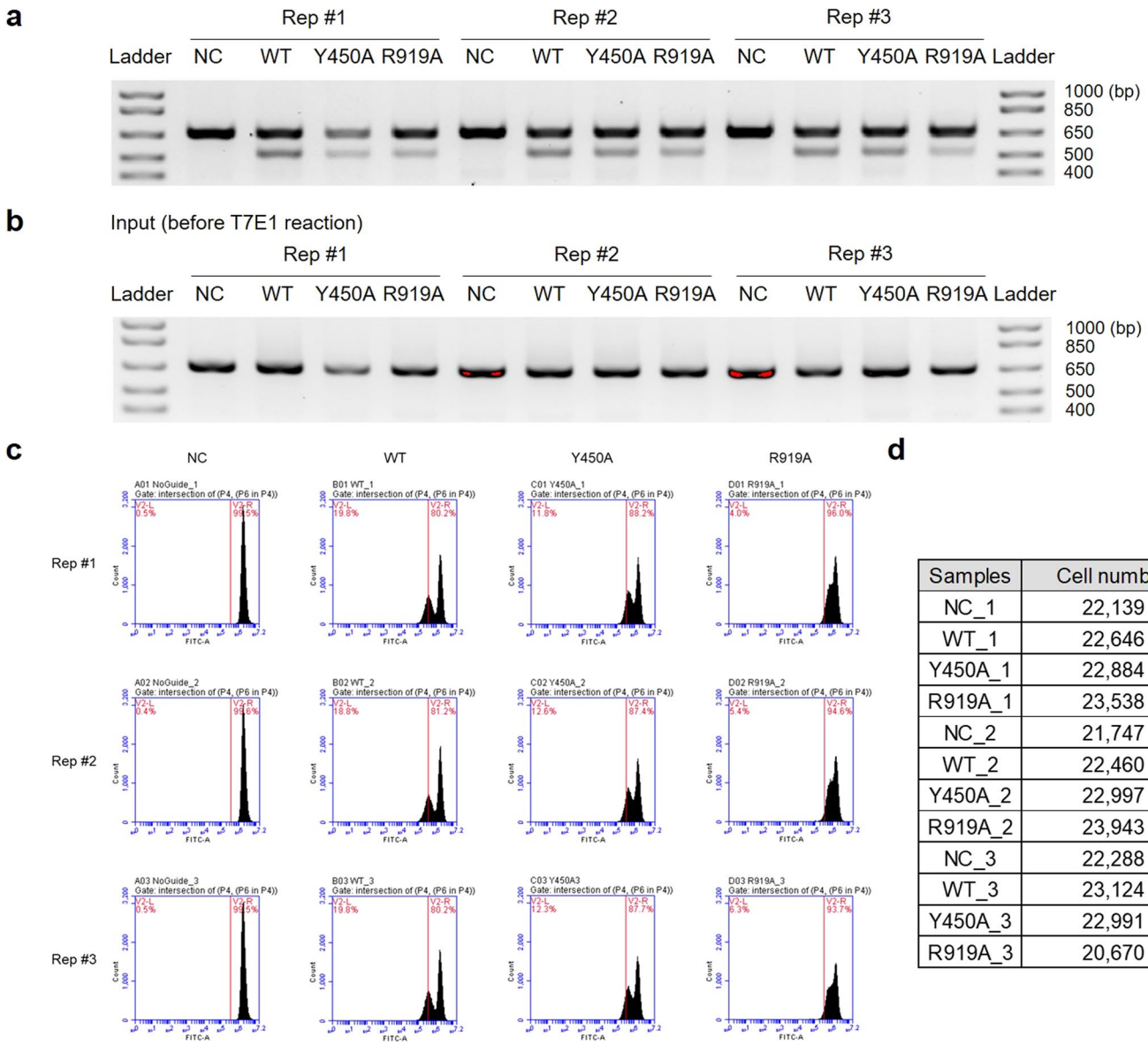
Extended Data Fig. 6 | Examples of RBS-identified proteins. a, Example of RBSs identified in distant primary sequence positions (PABPC1). Sequence homology of amino acids at -5 to +5 positions from Y194, Y297, and Y364 compared to that of Y222, F325, and Y393 in yeast Pab1 are described, respectively. Identical amino acids in the same positions are bold-faced. **b**, Partial structure of yeast Pab1 bound to poly(A) RNA (PDB 6R5K[22]). Y222, F325, and Y393 are indicated. **c**-**h**, Examples of RBSs identified in regions that are not annotated as RBDs. RBSs identified in NSUN5 (**c**), RTCA (**d**), APOBEC3C (**e**), TRIM25 (**f**), SERBP1 (**g**), and HNRNPA1 (**h**) are depicted. RBSs with high spectral counts are indicated.



Extended Data Fig. 7 | Purified spCas9 protein and template DNA for sgRNA synthesis. a, Purified His6-HA-NLS-TEV-spCas9 on SDS-PAGE gel stained with Coomassie G25. **b**, Template DNA prepared for T7 *in vitro* transcription of anti-CBX1 sgRNA.

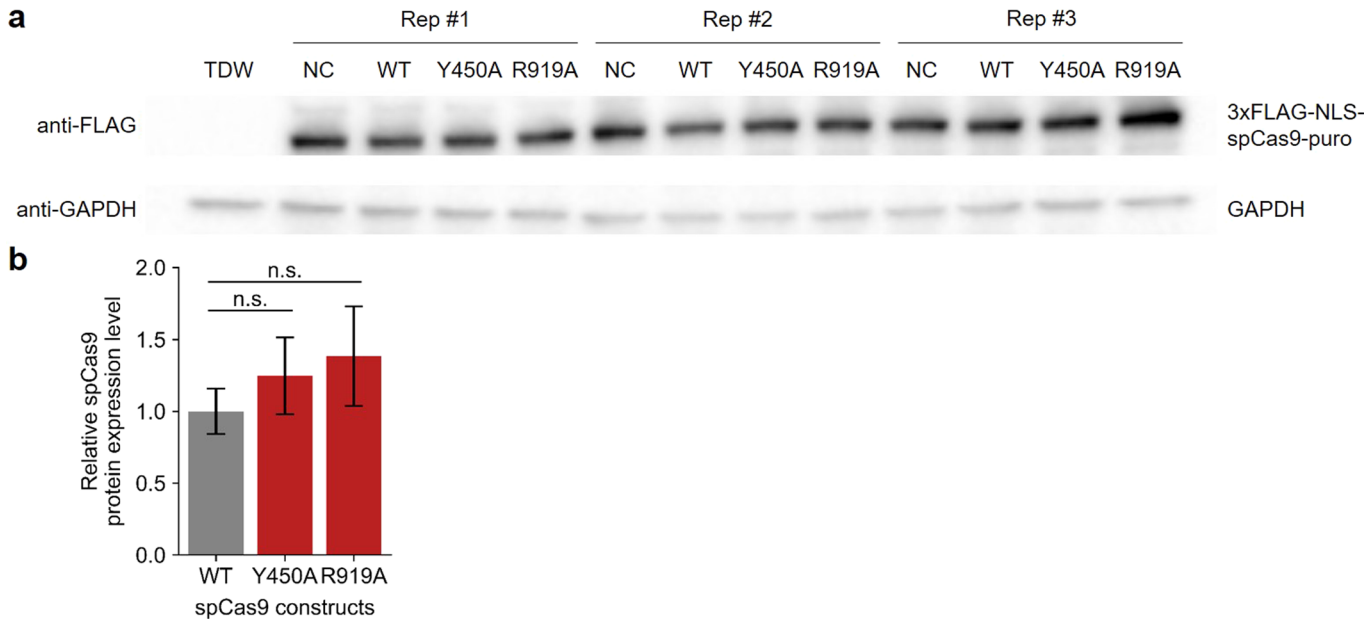


Extended Data Fig. 8 | Positions of RBS in crystal structure of spCas9 in complex with sgRNA and target DNA. Positions of RBSs in the crystal structure of spCas9 in complex with sgRNA and target DNA (PDB 4UN3[29]). Individual atoms of RBSs were drawn as dark blue spheres.



Samples	Cell number
NC_1	22,139
WT_1	22,646
Y450A_1	22,884
R919A_1	23,538
NC_2	21,747
WT_2	22,460
Y450A_2	22,997
R919A_2	23,943
NC_3	22,288
WT_3	23,124
Y450A_3	22,991
R919A_3	20,670

Extended Data Fig. 9 | Impact of RBS mutagenesis on spCas9 gene editing activity. a, EtBr-agarose gel image of cut/uncut fragments after T7E1 assay[30], 48 hours after transfection of negative control (mock), non-target-wildtype (NC), wildtype (WT), Y450A, R919A spCas9. **b**, EtBr-agarose gel image of input PCR amplicons. **c**, Histograms of GFP fluorescence of msfGFP-expressing clonal stable 293 cells 96 hours after transfection with non-target-wildtype (NC), wildtype (WT), Y450A, R919A spCas9. **d**, Cell counts per each flow cytometry run. Uncropped images for panels a-b are available as source data.



Extended Data Fig. 10 | Verification of comparable expression of spCas9. a, Verification of comparable expression of spCas9 proteins. Western blot of msfGFP-expressing HEK293E clonal stable cell 48 hours after transfection of negative control (mock), non-target-wildtype (NC), wildtype (WT), Y450A, or R919A spCas9. **b**, Quantification of spCas9 protein expression level. Two-sided unpaired Student's T-test was performed with n = 3 biologically independent samples (WT vs. Y450A; WT vs. R919A), assuming equal variance. The mean values were depicted with error bars that indicate standard deviation between replicates. P-value, rounded up to the fourth decimal point: 0.3261 (WT vs. Y450A), 0.2257 (WT vs. R919A). Uncropped image for panel a is available as source data.

nature research

Corresponding author(s): V. Narry Kim, Jong-Seo Kim

Last updated by author(s): Apr 14, 2020

Reporting Summary

Nature Research 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 Research policies, see Authors & Referees 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	Image Lab 6.0 was used to collect gel and blot data. Thermo Scientific Xcalibur was used to collect mass spectrometry data. Agilent ChemStation was used to collect HPLC-UV absorbance spectrum data.
Data analysis	RawConverter, MSConverter, MSFragger (v.20190530), MODa (v1.6.0), MS-GF+ (v2017.01.27), LcMsSpectator, and PANDA (v1.1.6-beta) were used to analyze raw files in mass spectrometry experiments. DAVID 6.8 was used for gene ontology analysis. ImageJ was used to quantify band intensities from gel and blot images. PyMOL was used to visualize previously published structural data. Python 3.7-based custom codes were used to for post-processing of the analyzed data, as described in detail in online methods.

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/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the dataset identifier PXD016254. The raw data were used to generate Figs. 1, 2 and 3.

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

Life sciences study design

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

Sample size	No statistical methods were used to predetermine sample sizes.
Data exclusions	No data were excluded from analysis.
Replication	All replicates are biological replicates obtained from biologically independent experiments. All attempts at replication were successful. The experiments number has been clearly stated in the figure legends.
Randomization	This is irrelevant to our study, as only cell lines or in vitro samples were used.
Blinding	This is irrelevant to our study, as only cell lines or in vitro samples were used.

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
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Anti-GAPDH antibody (Santa Cruz, sc-32233) ANTI-FLAG M2 antibody (Merck, F3165)
Validation	Antibody information is available on manufacturer's website.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HeLaT, HEK-293 cells were generous gifts from laboratories in School of Biological Sciences, Seoul National University. HeLaT is a modified HeLa with TUT4 gene deletion.
Authentication	All cell lines were authenticated by ATCC, via STR Profiling following ISO 9001:2008 and ISO/IEC 17025:2005 quality standards.
Mycoplasma contamination	HeLaT was tested negative for mycoplasma contamination. HEK-293 was not tested for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	msfGFP-stable expressing HEK-293 cells were grown on 12-well plates. Cells were trypsinized and reconstituted in PBS prior to flow cytometry analysis.
Instrument	BD accuri C6 Plus
Software	BD accuri C6 Plus
Cell population abundance	More than 20,000 cells that passed the gating were analyzed per experimental condition. Cells were not post-fractionated.
Gating strategy	Cells were gated using FSC-A, SSC-A, and SSC-H to include only live cells. FITC-A values of cells within the gates were analyzed. A boundary was set to include ~0.5% of cells with smallest FITC-A values in msfGFP-stable expressing HEK-293 cells transfected with wildtype Cas9 and non-targeting sgRNA.

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.