https://doi.org/10.1038/s41592-022-01629-6

Regulatory non-coding RNAs: everything is possible, but what is important?

Jimmy K. Guo and Mitchell Guttman

In recent years, the number of annotated noncoding RNAs (ncRNAs) and RNA-binding proteins (RBPs) has increased dramatically. The wide range of RBPs identified highlights the enormous potential for RNA in virtually all aspects of cell biology, from transcriptional regulation to metabolic control. Yet, there is a growing gap between what is possible and what has been demonstrated to be functionally important. Here we highlight recent methodological developments in the study of RNA–protein interactions, discuss the challenges and opportunities for exploring their functional roles, and provide our perspectives on what is needed to bridge the gap in this rapidly expanding field.

Mammalian genomes encode thousands of noncoding RNAs (ncRNAs), with ~20,000 annotated long ncRNA (lncRNA) genes — a number that rivals and may ultimately exceed the total number of protein-coding genes in the human genome[1]. Yet, most ncRNAs remain functionally uncharacterized and the diversity of biological roles that they play are largely unexplored. Identifying what proteins interact with a ncRNA can provide critical insights into its possible functions and mechanisms, enabling formation of experimentally testable hypotheses. For example, mapping various protein interactions to ncRNAs has led to proposed models whereby specific ncRNAs can: (1) guide regulatory proteins to specific genomic DNA sites[2,3]; (2) tether multiple protein components into macromolecular complexes[4,5]; (3) mediate and stabilize three-dimensional chromatin loops[6–8]; (4) activate[9] or inhibit[10] specific enzymatic function; and (5) compete proteins away from their mRNA[11] or genomic DNA targets[12–14].

Over the past decade, the development of global RNA-centric proteomics methods (Fig. 1a), such as RBR-ID (proteomic identification of RNA-binding regions)[15] and RBDmap[16], have enabled unprecedented exploration of which proteins bind to RNA. These efforts have vastly expanded the number of identified RBPs, with >4,000 human proteins (~20% of the human proteome) currently annotated as 'RNA-binding' by UniProt[17]. These RBPs include many that lack canonical RNA-binding domains, such as RRMs (RNA recognition motifs) or KH (hnRNP K homology) domains, and encompass critical chromatin and transcriptional regulators, nuclear structure proteins and metabolic enzymes[15,16]. The large number of putative RBPs representing such a diverse functional spectrum suggests vast potential for regulatory ncRNA function.

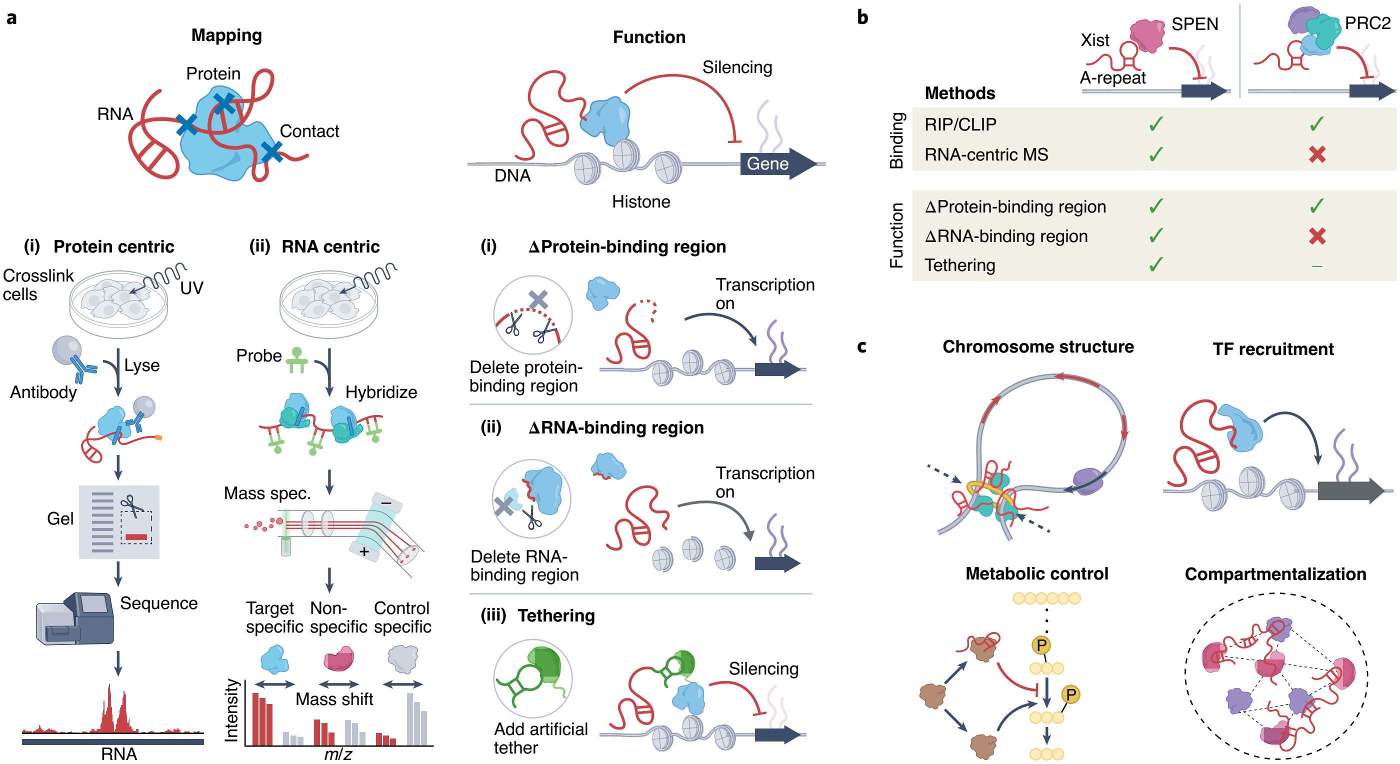
Despite this, it remains mostly unknown how many of these RBPs interact with ncRNAs, and which specific ncRNAs they might associate with. Typically, defining the RNAs that proteins bind in vivo requires protein-centric approaches, such as CLIP (cross-linking and immunoprecipitation)[18] (Fig. 1a). When paired with high-throughput sequencing[19,20], CLIP can comprehensively define specific sites on RNAs that interact with a protein of interest. This approach utilizes ultraviolet (UV) light to create a covalent photo-crosslink between a protein and its bound RNAs, but not between pairs of proteins. Because these crosslinks can be formed in a living cell, a specific protein can be purified — usually via an antibody — under stringent washing conditions that disrupt non-crosslinked RNA–protein and protein–protein interactions. However, as CLIP can only map a single protein at a time, it is technically challenging to explore the thousands of annotated RBPs. Additionally, CLIP relies on high-quality antibodies, which are not always available — especially for non-canonical RBPs. Accordingly, there have been limited efforts to map most of these proteins to specific RNAs. Moreover, even in cases where interactions between specific non-canonical RBPs and RNAs have been identified, the functional relevance of these interactions have been questionable.

A cautionary tale from Xist and PRC2

The Xist lncRNA represents a valuable case study illustrating some of the practical challenges in deciphering ncRNA–protein interactions and function. Briefly, Xist is required for initiating chromosome-wide transcriptional silencing on the X chromosome to balance gene expression between male (XY) and female (XX) mammals[21]. Although Xist was first identified in 1991 (ref. [22]), the molecular components required for initiation of chromosome-wide silencing were not identified until 2015 (refs. [23–25]).

In the intervening years, extensive characterization of Xist showed that: (1) Xist coats the inactive X[26]; (2) Xist is sufficient to initiate transcriptional silencing on the X[27]; (2) initiation of Xist corresponds to accumulation of PRC2 and its associated H3K27me3 repressive mark over the inactive X[28]; (3) the A-repeat region of Xist is required for transcriptional silencing[29]; and (4) the A-repeat of Xist interacts with PRC2 (ref. [30]). Because PRC2 was known to be involved in transcriptional silencing in other contexts[31], this led to a model where Xist binds directly to PRC2 via the A-repeat to silence transcription (Fig. 1b).

Although this model seemingly explained these observations, there was a problem: deletion of PRC2 did not impact Xist-mediated transcriptional silencing[24,32,33] (Fig. 1b). As Xist–PRC2 interactions were identified using either in vitro measurements[30] or native RIP (RNA immunoprecipitation)[34], they might represent in-solution associations rather than bona fide interactions that occur in vivo. In a classic experiment, Mili and Steitz showed that native immunoprecipitation



Methods	SPEN	PRC2
RIP/CLIP	✓	✓
RNA-centric MS	✓	✗
ΔProtein-binding region	✓	✓
ΔRNA-binding region	✓	✗
Tethering	✓	–

Fig. 1 | Identification and functional characterization of RNA–protein interactions. a, Left: methods for mapping RNA–protein interactions in vivo using UV crosslinking, including: (i) protein-centric approaches in which a specific protein is selectively purified and its associated RNAs are mapped using high-throughput sequencing; and (ii) RNA-centric approaches where a specific RNA or set of RNAs are selectively purified and the bound proteins are identified by mass spectrometry. Right: experimental framework for dissecting the functional relevance of an RNA–protein interaction illustrated by a ncRNA–protein complex that acts to repress transcription. Schematic of methods to disrupt an RNA–protein interaction via: (i) deletion (Δ) of a protein-binding region on the RNA; or (ii) deletion (Δ) of the RNA-binding region on the protein; and (iii) rescuing a phenotype through synthetically tethering the effector protein to the RNA. **b**, Experimental evidence for (green tick) or against (red cross) the functional interaction between Xist and either SPEN (left) or PRC2 (right). **c**, A range of proposed roles for RNA-mediated regulation of cellular processes, including mediation of three-dimensional DNA structure, recruitment of transcription factors (TFs) to genomic sites, feedback inhibition of metabolic pathways, and subcellular compartmentalization of proteins and RNA. *m*/*z*, mass-to-charge ratio; P, phosphate group.

methods can identify RNA–protein interactions that could not have occurred in vivo[35]. Similarly, mammalian PRC2 was shown to interact with bacterial RNAs with comparable affinity to that of other mammalian RNAs, including the A-repeat[14].

In fact, PRC2 has been reported to bind promiscuously to virtually all RNAs and the biological significance of this remains a topic of debate[36]. While most studies of RNA interactions with PRC2 rely on in vitro measurements and RIP, there have been recent attempts to map PRC2 using CLIP[12,13,37]; these reported further evidence of promiscuous PRC2 binding to RNA, including to Xist. However, while experimentally stringent, these studies highlight another critical issue related to the analysis of CLIP data. It is well documented that the number of reads mapping to an RNA is proportional to its overall abundance and therefore simply identifying reads does not indicate binding[19,38]. For this reason, it is not possible to distinguish between promiscuous binding of a protein to all RNAs and the lack of binding to any RNAs. In addition, because of the low efficiency of UV crosslinking and stringency of CLIP, the complexity of the underlying sequencing library is often extremely low, leading to read pileups at specific locations due to PCR duplications. Indeed, many of the reported interactions between PRC2 and specific RNA regions in CLIP experiments appear to be PCR duplicates rather than enrichment of true binding events[37].

Consistent with the idea that the Xist–PRC2 association might not represent an in vivo binding event, several studies purified Xist using different in vivo crosslinking strategies coupled with high stringency washes and mass spectrometry[23–25]. None of these methods identified an association between Xist and any previously reported PRC2 components. In contrast, these studies all independently identified SPEN (also known as SHARP)[23–25], a transcriptional co-repressor. Follow-up studies have demonstrated that SPEN is required for Xist-mediated transcriptional silencing in cell-based models[23,24] and in early development[39]. SPEN has been shown to bind to the A-repeat region of Xist via CLIP[40,41], congruent with the finding that Xist lacking the A-repeat cannot silence transcription[29] (Fig. 1b).

Importantly, the discrepancy between biochemical evidence supporting specific RNA–protein interactions and genetic evidence demonstrating that these same interactions are often dispensable for function is not limited to Xist and PRC2. For example, recent evidence indicates that PRC2 is dispensable for HOTAIR-mediated gene silencing[42], even though it was initially reported to bind to PRC2 using RIP[5,43]. Similarly, the YY1 transcriptional regulator was reported to bind to Xist to tether the RNA to chromatin[44], yet neither deletion of the YY1 protein[24] nor deletion of the reported YY1 binding site from Xist (F-repeat)[29] impacts the localization of Xist to chromatin or Xist-mediated transcriptional silencing. Consistent with this, stringent purification of Xist followed by mass spectrometry failed to identify YY1 as an Xist binding protein[23,24].

These examples highlight the practical issues associated with deleting an identified 'binding site' as evidence supporting the functional role of an RNA–protein interaction (Fig. 1a). Specifically, deletion of a binding site on an RNA may result in phenotypic effects due to disruption of a different protein (for example, SPEN rather than PRC2 to the A-repeat). Similar issues may occur when disrupting the RNA-binding region of a protein, which could impact its overall structure and other essential functions. For example, deletion of the RNA-binding region of CTCF impacts formation of chromatin loops[6,7]; yet, because it overlaps a zinc-finger motif (a known DNA-binding motif), it is unclear if the observed impacts are solely due to RNA binding. Because of these potential issues, alternative approaches that directly test the importance of the RNA–protein interaction are critical. One way to do this is by reconstituting the RNA–protein interaction via a synthetic fusion following disruption of the RNA-binding region and/or protein binding site and measuring whether this can rescue the expected phenotype (Fig. 1a). For example, synthetically tethering an RNA-binding mutant of SPEN directly to Xist was shown to rescue transcriptional silencing on the X[39] (Fig. 1b).

Bridging the gap between discovery and function

Many additional non-canonical RBPs, such as metabolic enzymes (for example, ENO1[45]) and various chromatin complexes including DNA methylation enzymes (for example, DNMT1[10] and TET2[15]), repressive (for example, PRC1[46]) and activating (for example, WDR5[47]) chromatin modifiers, transcription factors (for example, SOX2[48]), and three-dimensional DNA structure proteins (for example, CTCF[6,7]), have been reported to bind to RNAs. Based on these observations, chromatin regulators have emerged as central players in the mechanisms by which ncRNAs regulate gene expression (Fig. 1c). Although an attractive model due to the intrinsic high local concentrations that ncRNAs can form in the nucleus[3], the functional importance of RNA binding in chromatin regulation remains untested in most cases. As the number of proteins reported to bind to RNA continues to increase, we are faced with a growing chasm between the potential of what ncRNAs can do and the reality of what functional roles they play.

Motivated by the lessons learned from the examples discussed above, we propose a comprehensive framework – including new experimental methods – that will be useful for bridging this gap. This framework consists of: (1) stringent experimental methods to define high confidence RNA–protein interactions – including high stringency and/or denaturing purification for RNA-centric proteomic discovery and protein-centric RNA mapping; (2) scalable methods that can characterize the large numbers of putative RBPs, which will require development of new tools that utilize the stringency and binding site precision of CLIP, but with dramatically improved throughput. Moreover, we anticipate needing additional affinity reagents or alternative purification strategies to map proteins that are currently inaccessible via existing antibodies; (3) rigorous computational and statistical methods to identify meaningful regions of RNA binding that account for abundance, complexity and other sources of artifacts; (4) quantitative measurements of protein and RNA binding affinity[49] and occupancy frequency[50] in living cells – such approaches will enable more precise characterization of true binding events through establishing quantitative criteria, including accurate measurements of potentially promiscuous RBP interactions; and (5) precise functional characterization of an RNA–protein interaction through targeted disruption of the interaction and rescue through reconstitution.

With a reliable framework such as this, we anticipate being able to define classes of ncRNA and protein functions to fully understand the scale and scope of ncRNA-mediated functions. This information will allow us to explore what intrinsic properties of RNA make it such a widespread and versatile molecular regulator. Moreover, it will allow us to begin to address more global questions, such as why a large fraction of the human proteome has evolved to bind to RNA, and why the genome encodes so many distinct ncRNA species.

Jimmy K. Guo[1,2] **and Mitchell Guttman**[1] ✉

[1]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA. [2]Keck School of Medicine, University of Southern California, Los Angeles, CA, USA.

✉e-mail: mguttman@caltech.edu

Published online: 6 October 2022

References

1. Frankish, A. et al. *Nucleic Acids Res.* **49**, D916–D923 (2021).
2. Rinn, J. L. & Chang, H. Y. *Annu. Rev. Biochem.* **81**, 145–166 (2012).
3. Quinodoz, S. A. et al. *Cell* **184**, 5775–5790 (2021).
4. Kaneko, S. et al. *Mol. Cell* **53**, 290–300 (2014).
5. Tsai, M. C. et al. *Science* **329**, 689–693 (2010).
6. Hansen, A. S. et al. *Mol. Cell* **76**, 395–411 (2019).
7. Saldaña-Meyer, R. et al. *Mol. Cell* **76**, 412–422 (2019).
8. Mumbach, M. R. et al. *Nat. Methods* **16**, 489–492 (2019).
9. Zovoilis, A., Cifuentes-Rojas, C., Chu, H.-P., Hernandez, A. J. & Lee, J. T. *Cell* **167**, 1788–1802 (2016).
10. Di Ruscio, A. et al. *Nature* **503**, 371–376 (2013).
11. Lee, S. et al. *Cell* **164**, 69–80 (2016).
12. Kaneko, S., Son, J., Shen, S. S., Reinberg, D. & Bonasio, R. *Nat. Struct. Mol. Biol.* **20**, 1258–1264 (2013).
13. Beltran, M. et al. *Genome Res.* **26**, 896–907 (2016).
14. Davidovich, C., Zheng, L., Goodrich, K. J. & Cech, T. R. *Nat. Struct. Mol. Biol.* **20**, 1250–1257 (2013).
15. He, C. et al. *Mol. Cell* **64**, 416–430 (2016).
16. Castello, A. et al. *Mol. Cell* **63**, 696–710 (2016).
17. Bateman, A. et al. *Nucleic Acids Res.* **49**, D480–D489 (2021).
18. Ule, J. et al. *Science* **302**, 1212–1215 (2003).
19. Van Nostrand, E. L. et al. *Nat. Methods* **13**, 508–514 (2016).
20. Zarnegar, B. J. et al. *Nat. Methods* **13**, 489–492 (2016).
21. Heard, E. & Disteche, C. M. *Genes Dev.* **20**, 1848–1867 (2006).
22. Brockdorff, N. et al. *Nature* **351**, 329–331 (1991).
23. Chu, C. et al. *Cell* **161**, 404–416 (2015).
24. McHugh, C. A. et al. *Nature* **521**, 232–236 (2015).
25. Minajigi, A. et al. *Science* **349**, aab2276 (2015).
26. Clemson, C. M., McNeil, J. A., Willard, H. F. & Lawrence, J. B. *J. Cell Biol.* **132**, 259–275 (1996).
27. Penny, G. D., Kay, G. F., Sheardown, S. A., Rastan, S. & Brockdorff, N. *Nature* **379**, 131–137 (1996).
28. Plath, K. et al. *Science* **300**, 131–135 (2003).
29. Wutz, A., Rasmussen, T. P. & Jaenisch, R. *Nat. Genet.* **30**, 167–174 (2002).
30. Zhao, J., Sun, B. K., Erwin, J. A., Song, J.-J. & Lee, J. T. *Science* **322**, 750–756 (2008).
31. Margueron, R. & Reinberg, D. *Nature* **469**, 343–349 (2011).
32. Kalantry, S. & Magnuson, T. *PLoS Genet.* **2**, e66 (2006).
33. Schoeftner, S. et al. *EMBO J.* **25**, 3110–3122 (2006).
34. Zhao, J. et al. *Mol. Cell* **40**, 939–953 (2010).
35. Mili, S. & Steitz, J. A. *RNA* **10**, 1692–1694 (2004).
36. Davidovich, C. et al. *Mol. Cell* **57**, 552–558 (2015).
37. Rosenberg, M. et al. *Nat. Struct. Mol. Biol.* **28**, 103–117 (2021).

38. Hafner, M. et al. *Nat. Rev. Methods Primers* **1**, 20 (2021).
39. Dossin, F. et al. *Nature* **578**, 455–460 (2020).
40. Chen, C.-K. et al. *Science* **354**, 468–472 (2016).
41. Lu, Z. et al. *Nat. Commun.* **11**, 6163 (2020).
42. Portoso, M. et al. *EMBO J.* **36**, 981–994 (2017).
43. Rinn, J. L. et al. *Cell* **129**, 1311–1323 (2007).
44. Jeon, Y. & Lee, J. T. *Cell* **146**, 119–133 (2011).
45. Huppertz, I. et al. *Mol. Cell* **82**, 2666–2680 (2022).
46. Guttman, M. et al. *Nature* **477**, 295–300 (2011).
47. Wang, K. C. et al. *Nature* **472**, 120–124 (2011).
48. Holmes, Z. E. et al. *Nat. Commun.* **11**, 1805 (2020).
49. Porter, D. F. et al. *Nat. Commun.* **12**, 1569 (2021).
50. Sharma, D. et al. *Nature* **591**, 152–156 (2021).

Acknowledgements

The authors thank I.-M. Strazhnik for Fig. 1 and S. Hiley for editing. J.K.G. is supported by the University of Southern California MD/PhD program. Research in the Guttman laboratory is funded by the National Institutes of Health (NIH) 4DN program, an NIH Director's Transformative R01 award, the Chan Zuckerberg Initiative and funds from California Institute of Technology.

Author contributions

The authors contributed equally to all aspects of the article.

Competing interests

The authors declare no competing interests.